UNIVERSITÀ DEGLI STUDI
DELL'INSUBRIA

TESINA DI FONDAMENTI DI DATA ANALYTICS:

"PERFORMANCE TRA LE NUVOLE: ANALISI DEI FATTORI DI RITARDO DELLE LINEE AEREE STATUNITENSI "

Autori:
Matteo Forner (744416 - mforner1@studenti.uninsubria.it)
Chiara Norcia (744716 - cnorcia@studenti.uninsubria.it)
Federico Rausa (744473 - frausa@studenti.uninsubria.it)

ANNO ACCADEMICO: 2022/2023

INDICE:

ABSTRACT

- CONTESTO: l'analisi concerne la misura e la natura dei ritardi delle linee aeree e le relative cause e si svolge, dal punto di vista degli operatori coinvolti, sotto un profilo amministrativo e gestionale.
- OBIETTIVO: si vuole ricavare un set di strumenti utili ad individuare possibili cause e strategie di mitigazione dei ritardi dei voli, a partire da strumenti di previsione dei ritardi, nonché fornire una visualizzazione sintetica dei risultati di performance prodotti dagli operatori.
- METODO: mediante l'analisi delle relazioni fra le variabili a disposizione, si costruisce una regressione che consente di identificare gli elementi caratterizzanti una gestione inefficiente e alcuni modelli di clustering degli operatori per poter selezionare i benchmark di riferimento.
- RISULTATI: i modelli e gli indici prodotti si dimostrano molto utili dal punto di vista degli operatori, ma non da quello dei clienti.
- CONCLUSIONI: il lavoro svolto offre una panoramica del fenomeno dei ritardi delle linee aeree, offrendo al contempo una spiegazione sintetica del fenomeno e una rivelazione della complessità sottostante.

INTRODUZIONE

Con il seguente studio, intendiamo fornire una panoramica dell'ambiente delle linee aeree statunitensi, grazie ai dati raccolti e pubblicati dal Dipartimento dei Trasporti degli Stati Uniti attraverso il "Bureau of Transportation Statistics"[1] che tiene traccia delle prestazioni dei voli domestici operati dalle grandi compagnie aeree. Negli ultimi decenni il trasporto aereo è diventato uno dei mezzi più comuni per viaggiare ma, purtroppo, i ritardi dei voli risultano essere un problema comune che influisce sulla comodità dei passeggeri e sull'economia delle compagnie aeree. Il dataset **Airline_Delay_Cause** contiene informazioni sui ritardi dei voli registrati negli aeroporti degli Stati Uniti, tra cui: il mese, la compagnia aerea, l'aeroporto di partenza e arrivo, il numero di voli arrivati, il numero di voli con ritardi superiori a 15 minuti e la durata dei ritardi per vari fattori, come i ritardi dovuti alla compagnia aerea, alle condizioni meteo, alla sicurezza e così via.

Il nostro obiettivo è quello di esplorare il dataset **Airline_Delay_Cause** e individuare i pattern nascosti tra le variabili inerenti al fenomeno dei ritardi. In particolare, esamineremo le tendenze dei ritardi dei voli per mese, compagnia aerea e aeroporto, così si identificheranno i fattori principali che contribuiscono ai ritardi dei voli. Si useranno tecniche statistiche per visualizzare i dati e per testare le ipotesi sui fattori che influiscono sui ritardi dei voli. Lo scopo è quello di fornire informazioni utili, agli aeroporti e alle altre imprese operanti nel settore, a fini gestionali e di ausilio nella definizione possibili interventi di miglioramento. Inoltre, vogliamo dare ai passeggeri una visualizzazione sintetica delle componenti dei ritardi che possono subire dopo aver selezionato una certa compagnia, in un dato mese e in un dato aeroporto. Ci soffermeremo, in particolar modo, sulla valutazione delle performance delle compagnie aeree, degli aeroporti e sull'identificazione dei fattori che li distinguono.

[1] Origine dati: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

DESCRIZIONE DEL DATASET

Tutte le variabili hanno come riferimento, per le righe, le diverse possibili combinazioni di anno, mese, compagnia aerea e aeroporto, e hanno per oggetto, una caratteristica di quel set di voli, gestito da una data compagnia aerea, in un dato mese e in un dato aeroporto. I dati sono temporalmente limitati, infatti il range di tempo va da Gennaio 2022 a Gennaio 2023, e riguardano esclusivamente fatti avvenuti negli aeroporti degli Stati Uniti. Delle 21 variabili presenti nel dataset, 15 sono continue e 6 sono categoriche. Delle 6 variabili categoriche 2 sono clonate: **carrier_name** con **carrier** e **airport_name** con **airport**. Sebbene la variabile **month** sia un numero, la consideriamo come categorica.

Il totale delle nostre osservazioni è 21.867.

Il nostro dataset:

- La variabile **year** e **month** rappresentano rispettivamente l'anno e il mese in cui una data compagnia aerea ha effettuato un dato numero di voli in un dato aeroporto degli Stati Uniti.
- La variabile **carrier_name** indica le diverse compagnie aeree che hanno inviato i propri aerei in un dato mese (month), in un dato anno (year) e in un dato aeroporto (airport_name). **Carrier** ne è la rappresentazione attraverso i codici alfanumerici.
- La variabile **airport_name** descrive i diversi aeroporti che hanno ricevuto gli aerei di una data compagnia aerea (carrier_name), in un dato mese (month) e in un dato anno (year). **Airport** ne è la rappresentazione attraverso i codici identificativi.
- La variabile **arr_flights** rappresenta il numero totale di voli effettuati da quella compagnia aerea (carrier_name) in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **arr_del15** indica il numero di voli arrivati in ritardo di almeno 15 minuti da quella compagnia aerea (carrier_name) in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **carrier_ct** descrive il numero medio di minuti di ritardo dei voli ritardati a causa di problemi con quella compagnia aerea (carrier_name) come equipaggiamento, manutenzione equipaggio, rifornimento di carburante… in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **weather_ct** rappresenta il numero medio di minuti di ritardo dei voli ritardati a causa di condizioni meteorologiche avverse, come temporali, neve, vento…, per una data compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **nas_ct** indica il numero medio di minuti di ritardo dei voli ritardati a causa di problemi di traffico aereo, come traffico congestionato, maltempo in altre parti del paese…, da una data compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in quel dato mese (month) e in un dato anno (year).
- La variabile **security_ct** descrive il numero medio di minuti di ritardo dei voli ritardati a causa di problemi di sicurezza, come evacuazioni dell'aereo, controlli di sicurezza supplementari…, da una data compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **late_aircraft_ct** rappresenta il numero medio di minuti di ritardo dei voli ritardati a causa di un volo precedente che è arrivato in ritardo e ha causato un ritardo nel volo successivo da quella determinata compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **arr_cancelled** indica il numero di voli cancellati da quella determinata compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **arr_diverted** descrive il numero di voli che sono stati dirottati verso un altro aeroporto da una determinata compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **arr_delay** rappresenta il numero totale dei minuti di ritardo accumulati da tutti i voli effettuati da quella determinata compagnia aerea (carrier_name), in un dato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **carrier_delay** indica il totale dei minuti di ritardi causati da problemi con una determinata compagnia aerea (carrier_name), in un determinato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **weather_delay** descrive il totale dei minuti di ritardo causati da condizioni metereologiche avverse per una determinata compagnia aerea (carrier_name), in un determinato aeroporto (airport_name), in un determinato mese (month) e in un dato anno (year).

- La variabile **nas_delay** rappresenta il totale dei minuti di ritardo causati dal sistema di gestione del traffico aereo (NAS) per una determinata compagnia aerea (carrier_name), in un determinato aeroporto (airport_name), in un determinato mese (month) e in un determinato anno (year).
- La variabile **security_delay** indica il totale dei minuti di ritardi causati da problemi di sicurezza per una determinata compagnia aerea (carrier_name), in un determinato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).
- La variabile **late_aircraft_delay** descrive il totale dei minuti di ritardi causati da un volo precedente che è arrivato in ritardo e ha causato un ritardo al volo successivo per una determinata compagnia aerea (carrier_name), in un determinato aeroporto (airport_name), in un dato mese (month) e in un dato anno (year).

La variabile **arr_delay** è la mera somma di tutte le variabili **_delay**.

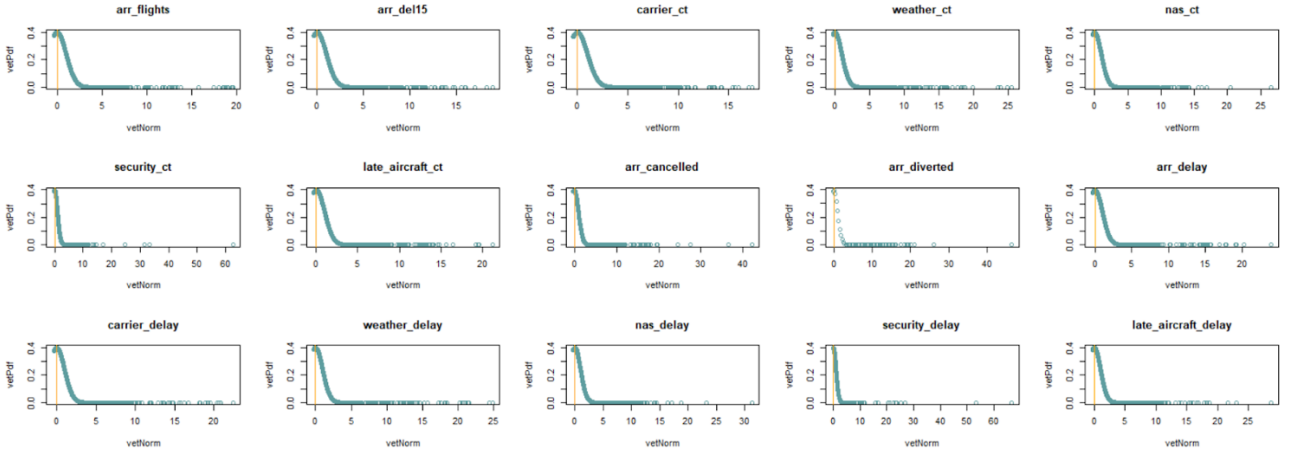
ANALISI UNIVARIATA

Dallo studio delle distribuzioni delle variabili, contenute nella Figura 1, si nota che le stesse non sembrano normali, in quanto la mediana normalizzata si discosta notevolmente dal valore 0, cioè dalla media normalizzata. Inoltre, il primo e il terzo quartile non risultano simmetrici in quanto il primo quartile, moltiplicato per -1, restituisce un valore molto diverso dal terzo. Proprio per questo i grafici delle pdf normali, contenute nella Figura 2, non appaiono come campane simmetriche. Notiamo inoltre che, dalla deviazione standard e dai quartili, emergono grandi differenze tra le distribuzioni delle variabili studiate; queste differenze non sembrano comparabili.

Figura 1: Tavole delle statistiche descrittive delle variabili continue

Variabile	minimo	primoQ	mediana	terzoQ	massimo	Nsample	media	deviazioneStd
arr_flights	1	43	90	218	18388	21867	332,3705	920,193
arr_del15	0	7	18	48	3479	21867	68,2998	180,8679
carrier_ct	0	2,69	7,41	20,385	1096,18	21867	25,4946	61,762
weather_ct	0	0	0,55	2	222,9	21867	2,5048	8,639
nas_ct	0	0,8	3,19	9,86	1391,74	21867	16,669	51,6807
security_ct	0	0	0	0	58,69	21867	0,2099	0,9354
late_aircraft_ct	0	1,35	4,75	14,545	1537,66	21867	23,4215	71,3715
arr_cancelled	0	0	1	5	1565	21867	8,7604	36,9602
arr_diverted	0	0	0	1	153	21867	0,7852	3,2822
arr_delay	0	387,5	1097	3077	323449	21867	4606,949	13296,9291
carrier_delay	0	135	449	1348	119425	21867	1816,17	5250,6549
weather_delay	0	0	24	168	24324	21867	258,6141	969,8391
nas_delay	0	24	127	424	84155	21867	790,0435	2659,8823
security_delay	0	0	0	0	3551	21867	9,754	53,0502
late_aircraft_delay	0	71	317	1103	158653	21867	1732,368	5458,0114

Figura 2: Pdf normali



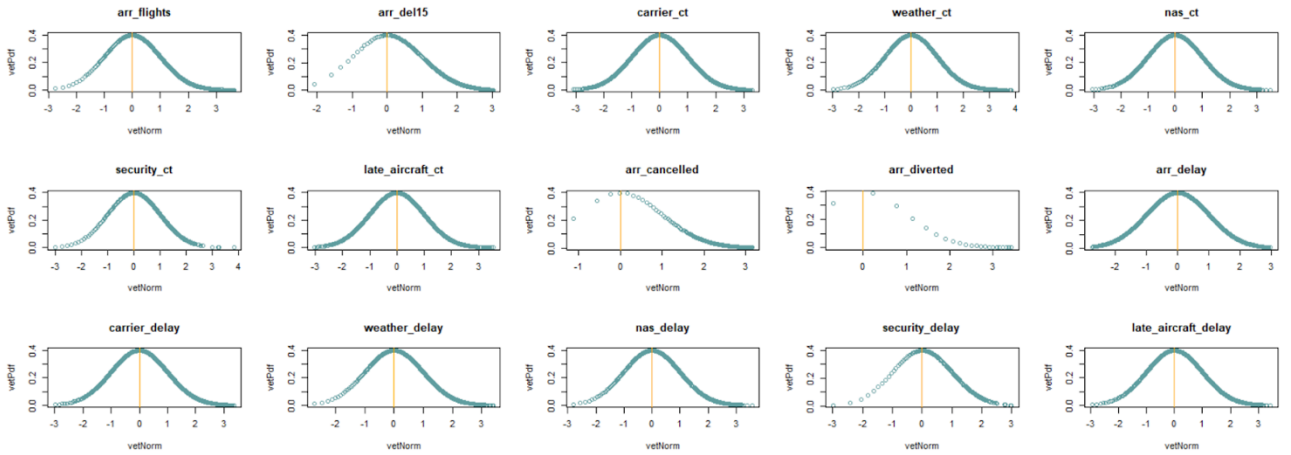
Proviamo allora ad applicare le trasformazioni logaritmiche che tendono a ridurre i range delle distribuzioni. Dopodichè proviamo anche a rimuovere gli outliers attraverso il metodo "Z-score"[2]. In seguito alle trasformazioni fatte applicando il modello "Z-score", abbiamo ottenuto delle distribuzioni molto più normali, anche se la dimensione campionaria si è ridotta notevolmente. A causa della trasformazione logaritmica, abbiamo dovuto rimuovere tutti i valori nulli, in quanto log(0)=-∞, e la rimozione degli outliers ci porta necessariamente a rimuovere i data points che risultano tali.

[2] Il metodo "Z-score" assume che la distribuzione sia gaussiana, ossia normale. Inoltre, considera outliers tutti i valori che, normalizzati come Z-score, sono, in modulo, maggiori di 3.

Figura 3: Tavole delle statistiche descrittive delle variabili continue

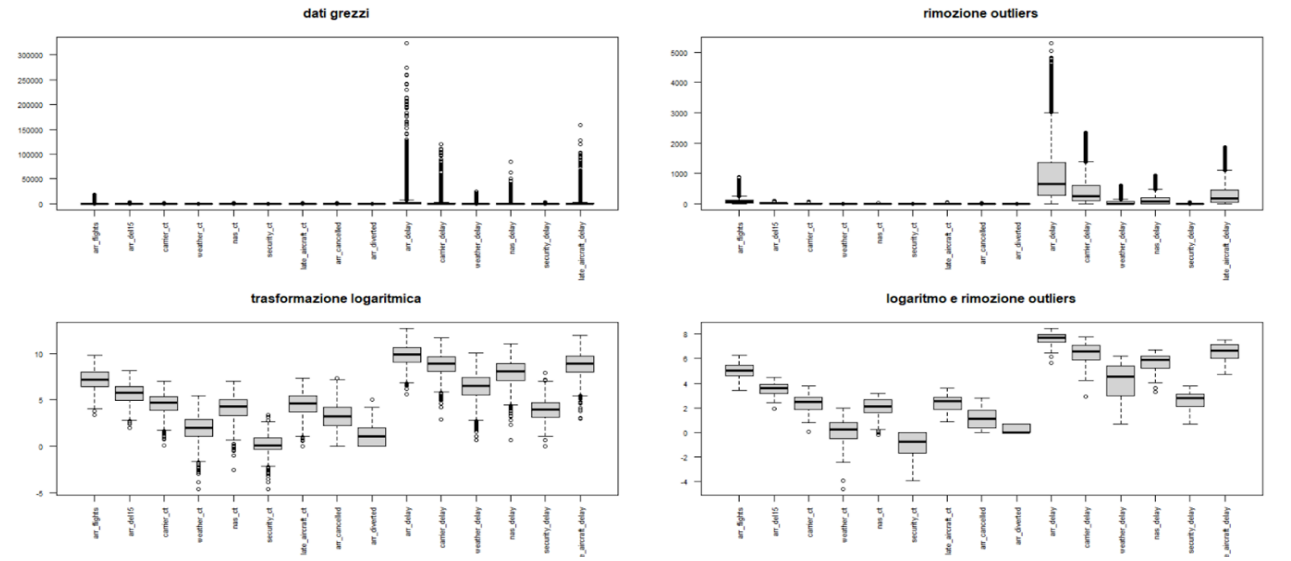
variabile	minimo	primoQ	mediana	terzoQ	massimo	Nsample	media	deviazioneStd
arr_flights	0,6931	3,7842	4,5109	5,3936	9,8195	21693	4,6246	1,4311
arr_del15	0	2,0794	2,9444	3,8918	7,4343	21147	3,0238	1,4581
carrier_ct	-2,3026	1,174	2,115	3,0994	6,9996	20487	2,1823	1,4536
weather_ct	-3,2189	-0,0513	0,5128	1,2975	5,4067	12539	0,5601	1,2612
nas_ct	-3,2189	0,6831	1,5518	2,5649	7,2383	17867	1,6453	1,5888
security_ct	-3,5066	-0,6931	0	0,5128	4,0723	3157	-0,1692	1,1095
late_aircraft_ct	-2,6593	0,8879	1,8099	2,8831	7,338	18916	1,956	1,5225
arr_cancelled	0	0	1,0986	2,1972	5,3845	13711	1,4041	1,2617
arr_diverted	0	0	0	0,6931	3,091	5503	0,5248	0,7453
arr_delay	2,7081	6,0822	7,0596	8,0678	11,8578	21159	7,0994	1,5973
carrier_delay	1,3863	5,1761	6,2403	7,2786	11,6904	20495	6,2357	1,6265
weather_delay	0,6931	3,8501	4,8752	5,8805	10,0992	12619	4,8822	1,5376
nas_delay	0,6931	4,2627	5,273	6,3439	11,3404	17879	5,3546	1,6723
security_delay	0	2,7726	3,5264	4,3438	7,0317	3178	3,5193	1,1737
late_aircraft_delay	1,0986	4,9836	6,0591	7,2064	11,9745	18945	6,115	1,7006

Figura 4: Pdf normali



Concludiamo l'analisi univariata affermando che i logaritmi delle nostre variabili risultano distribuiti normalmente, per cui anche i fenomeni osservati possono considerarsi normali.

Figura 5: Sintesi dei risultati ottenuti per ogni metodologia



ANALISI BIVARIATA

Il dataset scelto comprende sia le variabili categoriche sia le variabili continue. Vi sono 3 variabili categoriche e 15 variabili continue[3]. Se però tenessimo conto delle dummy, avremo altre 12 variabili 0-1 corrispondenti a **month**, 19 corrispondenti a **carrier** e 395 corrispondenti ad **airport**. Abbiamo quindi bisogno di un metodo sintetico per valutare i possibili rapporti di correlazione, che in totale sarebbero 98.346 e risulterebbero estremamente onerosi in termini di tempo e di calcoli.

Valutiamo il grado di correlazione tra variabili continue attraverso il coefficiente di correlazione di Pearson per relazioni lineari, ma ricorriamo anche al coefficiente di Spearman, in grado di cogliere livelli di correlazione per casi non lineari, nei quali il coefficiente di Pearson non rileva correlazione. In generale, il coefficiente di Spearman tende a essere più alto di quello di Pearson in situazioni in cui la relazione fra le due variabili non è strettamente lineare. Perciò, data una coppia di variabili, considereremo più rilevanti i coefficienti di Pearson maggiori o simili ai rispettivi coefficienti di Spearman.

Ai fini della regressione, si considerano maggiormente significativi gli R^2, ovvero i coefficienti di Pearson al quadrato, superiore al 64%, cioè gli R^2 corrispondenti a coppie di variabili continue in grado di spiegare reciprocamente oltre 64% della variabilità totale, in entrambe le direzioni. I coefficienti di correlazione infatti non implicano causalità, e pertanto non consentono di definire quale sia la variabile dipendente e quale quella esplicativa. Tali coefficienti ci consentono anche di decidere quali e quante variabili indipendenti usare nel nostro modello. Vogliamo infatti assolvere il più possibile il problema della multicollinearità, collassando le variabili indipendenti che sono significativamente correlate l'una con l'altra, riducendo il problema dell'overfitting e rispettando la seconda ipotesi alla base della regressione lineare.

Essendo interpretabile, il coefficiente di Spearman, al pari del coefficiente di Pearson, ed essendo l'indice R^2 interpretabile come il mero quadrato del coefficiente di Pearson, per analogia, è possibile elevare al quadrato il coefficiente di Spearman e utilizzarlo per descrivere la porzione di varianza spiegata vicendevolmente dalla coppia di variabili, esattamente come R^2.

In questo modo teniamo un secondo termine di confronto sulla correlazione, ovvero, se Spearman Squared fosse significativamente grande, ad esempio maggiore del 64%, sarebbe maggiore di Pearson Squared, e allo stesso tempo Pearson Squared fosse significativamente piccolo, ad esempio inferiore al 40%, probabilmente la relazione tra le variabili dovrebbe considerarsi non lineare, e di conseguenza, ai fini della regressione, si dovrebbe ricorrere ad un modello polinomiale. In caso contrario, Pearson Squared e Spearman Squared assumessero valori simili e maggiori del 64%, o se Pearson Squared fosse significativamente grande e comunque maggiore di Spearman Squared, il modello di regressione con maggior capacità previsioni va potrebbe risultare lineare.

Per quanto riguarda le variabili categoriche non siamo in grado di misurare quantitativamente il loro grado di correlazione reciproca o di correlazione con le variabili continue. Possiamo, però, testare se tale correlazione esiste, mediante il Test Chi-Quadro[4] e mediante il Test Anova[5]. Possiamo, perciò, misurare il livello di fiducia nel fatto che tali correlazioni esistono tramite i p-value[6] dei test.

[3] La gran parte delle variabili continue, trasformata con il logaritmo naturale, si presenta come distribuita normalmente, e ciò consente di misurare con maggiore accuratezza il livello di correlazione fra le stesse.

[4] Il test Chi-Quadro viene utilizzato per le correlazioni tra le variabili categoriche.

[5] Il Test Anova viene utilizzato per la correlazione tra una variabile categorica e una variabile continua.

[6] Il p-value o valore p, anche detto livello di probabilità, indica quanto è probabile che una statistica test S_n sia almeno pari al valore s_n supposta vera l'ipotesi nulla.

Tabella 1:Tavola dei quadrati dei coefficienti di correlazione maggiori del 64%

variabile.A	variabile.B	Pearson.Sq	Spearman.Sq
nas_ct	arr_delay	64,62%	62,64%
arr_flights	late_aircraft_delay	65,67%	62,06%
carrier_ct	late_aircraft_ct	65,75%	62,82%
arr_delay	nas_delay	70,15%	68,18%
security_ct	security_delay	70,49%	68,14%
arr_del15	nas_ct	71,78%	69,97%
arr_flights	carrier_delay	72,26%	69,60%
arr_del15	nas_delay	72,45%	70,44%
arr_flights	late_aircraft_ct	72,95%	68,46%
arr_del15	late_aircraft_delay	77,31%	76,28%
weather_ct	weather_delay	78,50%	73,81%
arr_flights	carrier_ct	79,71%	78,22%
late_aircraft_ct	arr_delay	79,75%	78,12%
arr_flights	arr_delay	79,98%	77,79%
arr_del15	carrier_delay	80,00%	79,28%
arr_delay	late_aircraft_delay	81,01%	80,41%
carrier_ct	arr_delay	81,95%	82,20%
arr_del15	late_aircraft_ct	83,89%	82,42%
arr_delay	carrier_delay	85,33%	86,02%
arr_del15	carrier_ct	86,93%	87,27%
arr_flights	arr_del15	87,06%	85,37%
carrier_ct	carrier_delay	91,40%	90,37%
late_aircraft_ct	late_aircraft_delay	92,33%	91,74%
nas_ct	nas_delay	92,92%	91,45%
arr_del15	arr_delay	93,52%	92,63%

Dalla Tabella 1 vediamo che tutti i coefficienti di Pearson sono molto vicini a quelli di Spearman, per cui possiamo presumere che vi siano relazioni lineari tra le variabili. Poiché nella costruzione della regressione il nostro obiettivo consiste nel prevedere il valore della variabile **carrier_delay** presenteremo anche tutte quelle inerenti la stessa.

Tabella 2: Tavola dei quadrati dei coefficienti di correlazione con carrier_delay

variabile.A	variabile.B	Pearson.Sq	Spearman.Sq
carrier_delay	security_ct	18,07%	22,25%
carrier_delay	security_delay	18,65%	20,22%
carrier_delay	arr_diverted	27,72%	25,30%
carrier_delay	weather_delay	29,88%	27,35%
carrier_delay	weather_ct	35,11%	32,86%
carrier_delay	arr_cancelled	38,24%	32,74%
carrier_delay	nas_ct	49,83%	46,84%
carrier_delay	nas_delay	52,52%	50,27%
carrier_delay	late_aircraft_delay	57,90%	56,59%
carrier_delay	late_aircraft_ct	60,52%	57,93%
carrier_delay	arr_flights	72,26%	69,60%
carrier_delay	arr_del15	80,00%	79,28%
carrier_delay	arr_delay	85,33%	86,02%
carrier_delay	carrier_ct	91,40%	90,37%

Facendo riferimento alla Tabella 2, osserviamo che dalla variabile **arr_flights** notiamo che tutti i coefficienti sono maggiori del 64% e che i coefficienti di Pearson non sono mai molto diversi da quelli di Spearman. Possiamo quindi presumere che la relazione fra le variabili significativamente correlate abbiano andamento lineare. Notiamo, però, che le variabili maggiormente correlate con **carrier_delay**, cioè: **arr_flights**, **arr_del15**, **arr_delay** e **carrier_ct**, sono molto correlati anche tra di loro. questa correlazione potrebbe far sorgere problemi di multicollinearità nella realizzazione del modello di regressione se venissero usate simultaneamente come variabili indipendenti. Per affrontare tale problema, ci affideremo anche ai risultati della PCA

Figura 6: Grafici scatterplot per carrier_delay e le altre variabili continue senza trasformazione logaritmica e senza rimozione degli outliers

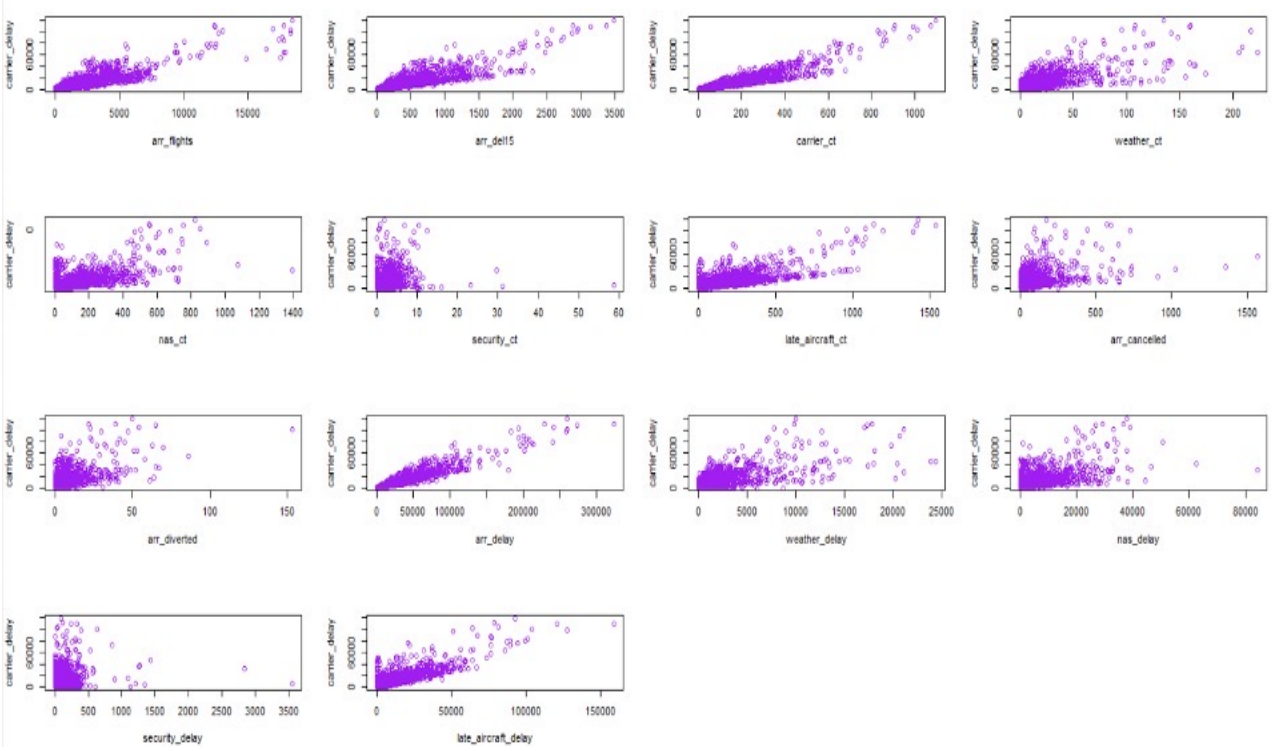
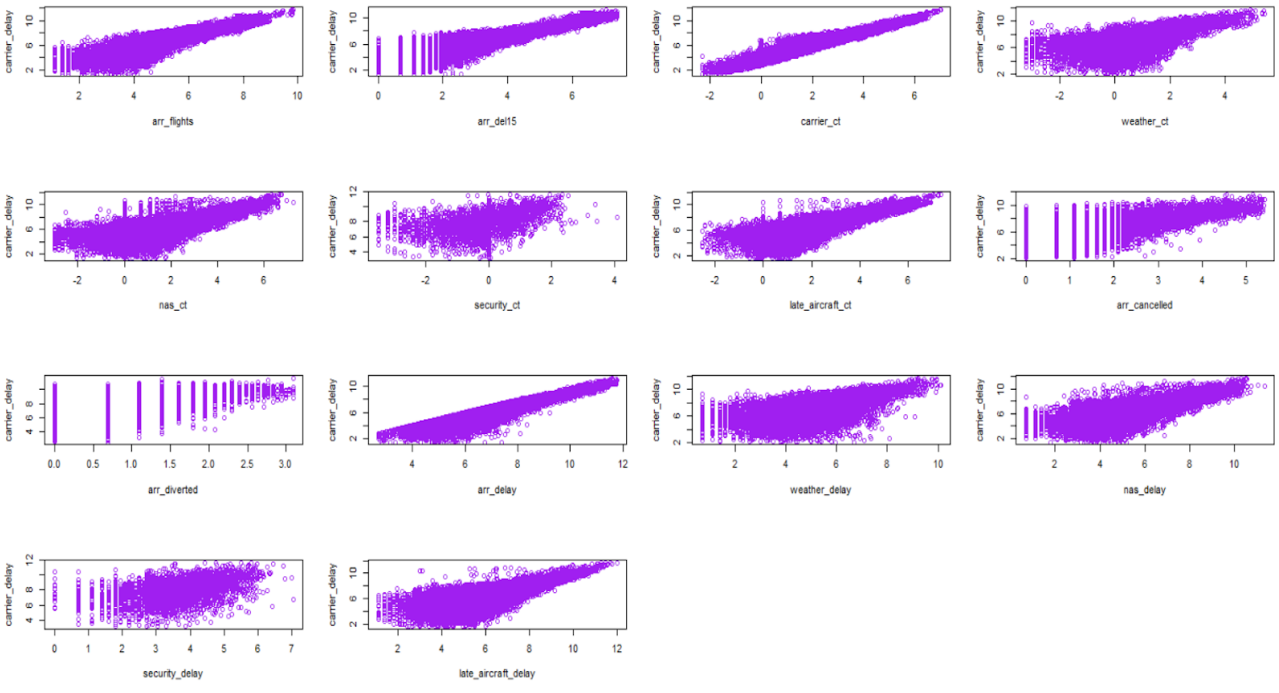


Figura 7: Grafici scatterplot per carrier_delay e le altre variabili continue dopo la trasformazione logaritmica e la rimozione degli outliers



Il Test Anova valida l'ipotesi nulla secondo cui una variabile categorica non ha alcun impatto sulla varianza di una variabile continua. Ai fini della regressione, lineare, logistica e non lineare, per poter utilizzare delle variabili categoriche è necessario convertirle in matrici di dummy. È inoltre è possibile calcolare il coefficiente di correlazione tra una singola dummy e una variabile continua, ma a fronte di numerose categorie, il Test Anova fornisce risposte molto più sintetiche. Anche in questo caso presentiamo i p-value dei Test Anova per le so le corrispondenze con la variabile **carrier_delay**.

Tabella 3: Tavola dei quadrati dei coefficienti di correlazione con carrier_delay

categorica	continua	AnovaTestPV
month	carrier_delay	5,8424E-11
carrier	carrier_delay	8,17E-272
airport	carrier_delay	0

Vediamo che tutte le variabili categoriche sono significativamente correlate con **carrier_delay**. I Test Anova consentono infatti in tutti i casi di rifiutare l'ipotesi nulla per cui non vi sarebbe correlazione tra ogni variabile categorica e ogni variabile continua, con dei p-value sempre minori di 0,001.

Ai fini della regressione la variabile **month** e la variabile **carrier** assumono rispettivamente 12 e 17 valori, ossia nel nostro dataset sono presenti 12 mesi e 17 compagnie, mentre la variabile **airport** comprende circa 394 aeroporti. A questo punto dovremmo sperimentare diversi modelli con tutte le dummy correlate e le loro possibili combinazioni ma ciò risulterebbe estremamente oneroso in termini di calcoli.

Il Test Chi-Quadro valida l'ipotesi nulla secondo cui, fra due variabili categoriche, non esiste alcuna correlazione, data la tavola delle contingenze delle due variabili. La tavola delle contingenze mostra il numero di volte che due categorie di ciascuna variabile si verificano congiuntamente.

Tabella 4: Tavole dei p-value dei Test Chi-Quadro sulla correlazione tra le variabili categoriche

var.A	var.B	ChiTest.p.value
month	carrier	1
month	airport	1
carrier	airport	0

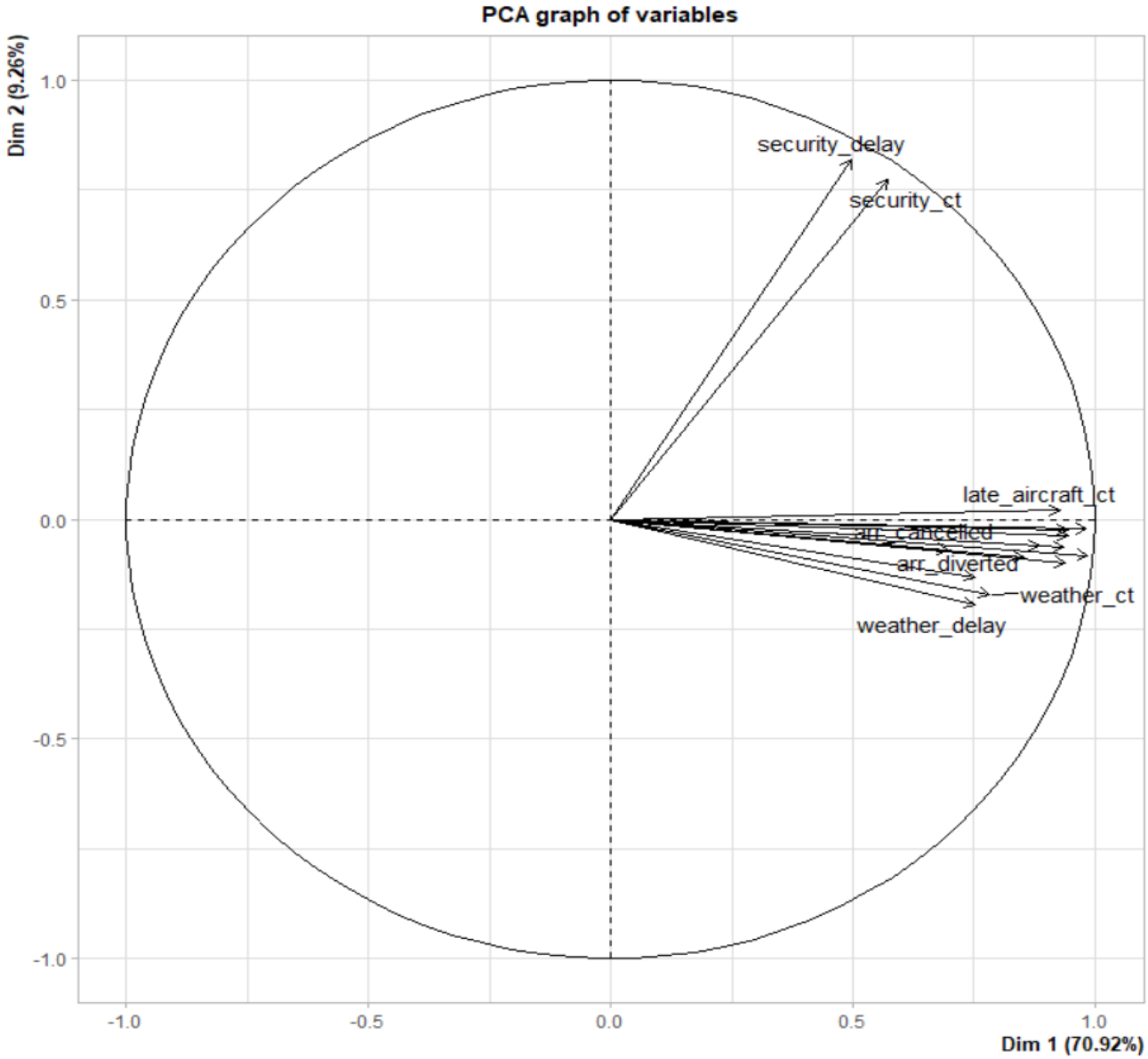
I Chi-Test ci consentono di rifiutare l'ipotesi nulla secondo cui non vi sia correlazione tra **carrier** e **airport** per tutti gli usuali livelli di significatività, ma non ci consentono di rifiutarla per le rispettive correlazioni con la variabile **month**. Ciò che questo Test ci consente di affermare è che molte compagnie aeree effettuano voli prevalentemente in determinate aeroporti, indipendentemente dal periodo dell'anno. Inoltre, nella costruzione del modello di regressione, possiamo non utilizzare la variabile **airport**, in quanto quest'ultima risulta perfettamente correlata con la variabile **carrier**. Il fatto di non utilizzare la variabile **airport** rispecchia il principio dell'assenza della multicollinearità, in quanto la sua variabilità può essere considerata già spiegata dalla variabile **carrier**. Non possiamo dire lo stesso della variabile **month** in quanto risulta priva di correlazioni significative con le altre due variabili categoriche.

A fronte di quanto sopra citato, va evidenziata la significatività degli indici trovati, in quanto si è lavorato con una dimensione campionaria pari, al più, a 21.865 osservazioni, e, almeno a circa 3.000 osservazioni, al netto di rimozioni dovute agli outliers e alla trasformazione logaritmica che esclude valori originariamente pari a 0 poiché $\log(0)=-\infty$.

ANALISI FATTORIALE

All'analisi bivariata affianchiamo anche la PCA[7], in quanto vogliamo rapidamente verificare le ipotesi fatte durante lo studio delle correlazioni. Ci aspettiamo che i suoi risultati siano coerenti con gli indici di Pearson e Spearman trovati in precedenza. La PCA ci consentirà di individuare le variabili più importanti tra le variabili continue, e quindi, di collassare quelle che ne seguono la variabilità.

Figura 8: Grafico delle variabili PCA



Dal biplot, presentato in Figura 8, si nota subito che possono aversi almeno due o al massimo tre gruppi di variabili, dette anche dimensioni rilevanti. Vediamo infatti che dei quindici vettori rappresentanti le nostre variabili, ci sono, perlopiù, due o tre direzioni comuni

- Un gruppo comprende le variabili security
- Un gruppo comprende le variabili weather
- Un gruppo, il più importante, comprende tutte le altre variabili

[7] PCA: principal component analysis

Figura 9: Scree plot

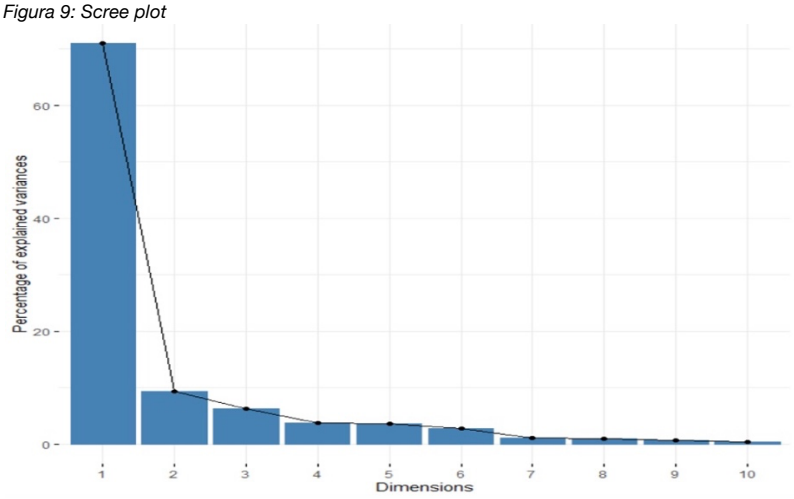


Figura 10: Summary PCA

```
Importance of components:
                          PC1     PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     3.2617 1.17833 0.96877 0.7460 0.72695 0.63545 0.40449
Proportion of Variance 0.7092 0.09256 0.06257 0.0371 0.03523 0.02692 0.01091
Cumulative Proportion  0.7092 0.80180 0.86437 0.9015 0.93670 0.96362 0.97453
                          PC8     PC9    PC10    PC11    PC12   PC13       PC14
Standard deviation     0.3695 0.31847 0.25314 0.22219 0.14239 0.1022 2.206e-05
Proportion of Variance 0.0091 0.00676 0.00427 0.00329 0.00135 0.0007 0.000e+00
Cumulative Proportion  0.9836 0.99039 0.99466 0.99795 0.99930 1.0000 1.000e+00
                          PC15
Standard deviation     6.811e-15
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

La domanda che a questo punto ci sorge spontanea è: i gruppi dovrebbero essere due o tre? Per rispondere a questo quesito ci affidiamo allo Scree plot contenuto nella Figura 9. Dallo Scree plot vediamo che la prima componente ha una proporzione della varianza del 70%, la seconda componente ha una proporzione della varianza del 9% circa e la terza del 6%. Il gomito dello Scree plot sembra più pronunciato sulla seconda componente, ma la differenza di contributo fra la seconda componente e la terza è di appena del 3%, e, se lo usassimo tutte e tre, spiegheremmo circa l'85% della variabilità complessiva.

Vogliamo ora capire quali sono le variabili più rilevanti per ciascuna di queste tre dimensioni, anche se dal biplot si potrebbero già intuire.

Figura 11: Biplot

```
Rotation (n x k) = (15 x 15):
                            PC1         PC2         PC3
arr_flights           0.2870827  0.05397354  0.08016591
arr_del15             0.3008004  0.01815229  0.13757206
carrier_ct            0.2893087  0.01974695  0.01154437
weather_ct            0.2395952  0.14489199 -0.58639235
nas_ct                0.2711234  0.05090236  0.27698568
security_ct           0.1750293 -0.65519629 -0.07741777
late_aircraft_ct      0.2843072 -0.01689704  0.21006643
arr_cancelled         0.2131044  0.05918549  0.13955551
arr_diverted          0.2307032  0.11130198 -0.03809590
arr_delay             0.3020680  0.06984000  0.03166386
carrier_delay         0.2876019  0.08405418 -0.09716551
weather_delay         0.2302906  0.16521118 -0.60764135
nas_delay             0.2620055  0.07419663  0.23949200
security_delay        0.1521197 -0.69401018 -0.13863355
late_aircraft_delay   0.2891454  0.03051531  0.16322117
```

Figura 12: Contribution of variables to Dim -1

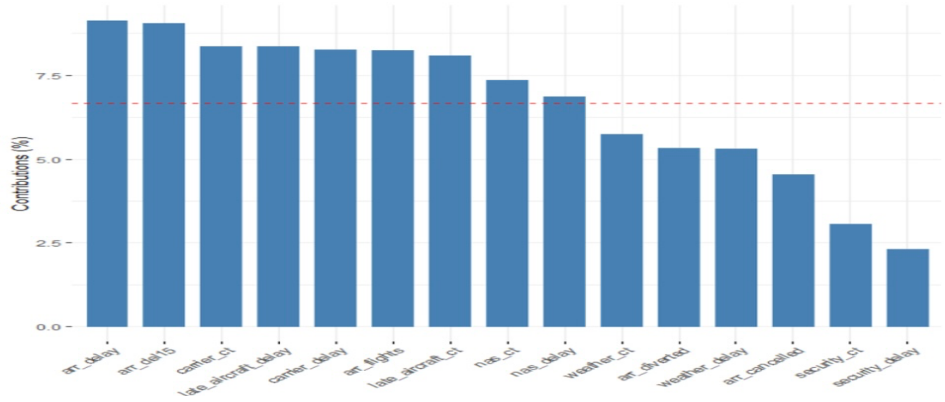


Figura 13: Contribution of variables to Dim -2

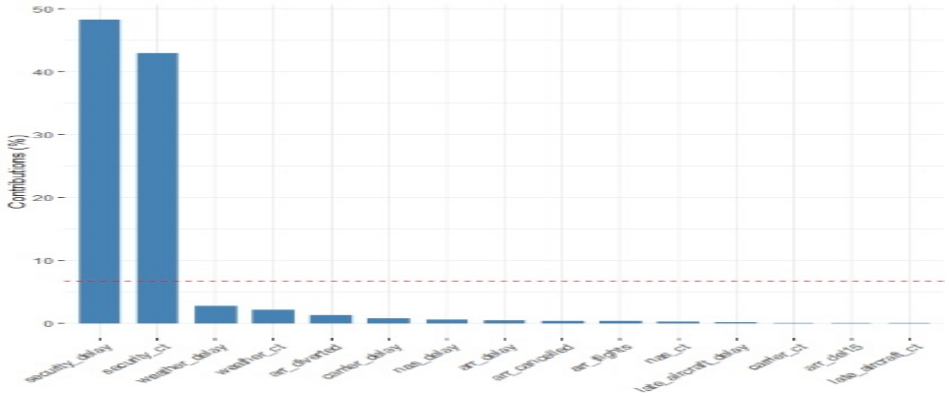
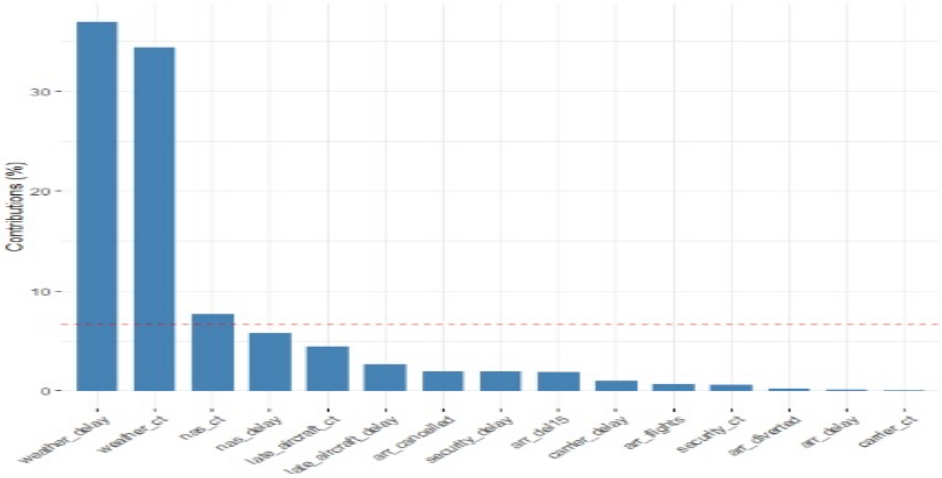


Figura 14: Contribution of variables to Dim -3



Dall'analisi si vede che le variabili che possono riassumere meglio i nostri dati sono: **arr_delay**, **security_delay** e **weather_delay**.

La PCA ci consente anche di ponderare meglio la rimozione degli outlier rimuovendo il minor numero di osservazioni, rilevando come outliers solo quelli che risultano tali per il metodo dello "Z-score" per le variabili più importanti. Se infatti utilizzassimo il metodo "Z-score" su tutte e 15 le variabili continue, troveremmo più di 5.000 outliers. Se effettuassimo la trasformazione logaritmica, a fronte del risparmio di osservazioni da escludere in quanto outliers, dovremmo escluderne altrettante che assumono valore nullo[8] e ci resterebbero appena 1.000 o 3.000 righe, delle 22.000 che erano in origine. Il metodo più economico, in termini di numero di osservazioni da rimuovere, risulta quindi considerare outliers solo quei punti che risultano tali da **arr_delay**, **security_delay** e **weather_delay**, oppure dalle coordinate delle tre componenti principali della PCA,

[8] log(0)=- ∞

decidiamo di optare per gli “Z-score” di queste ultime, in questo modo, rileviamo come outliers solo 1.053 osservazioni, su 21.867.

Figura 15: PCA prima della rimozione degli outliers

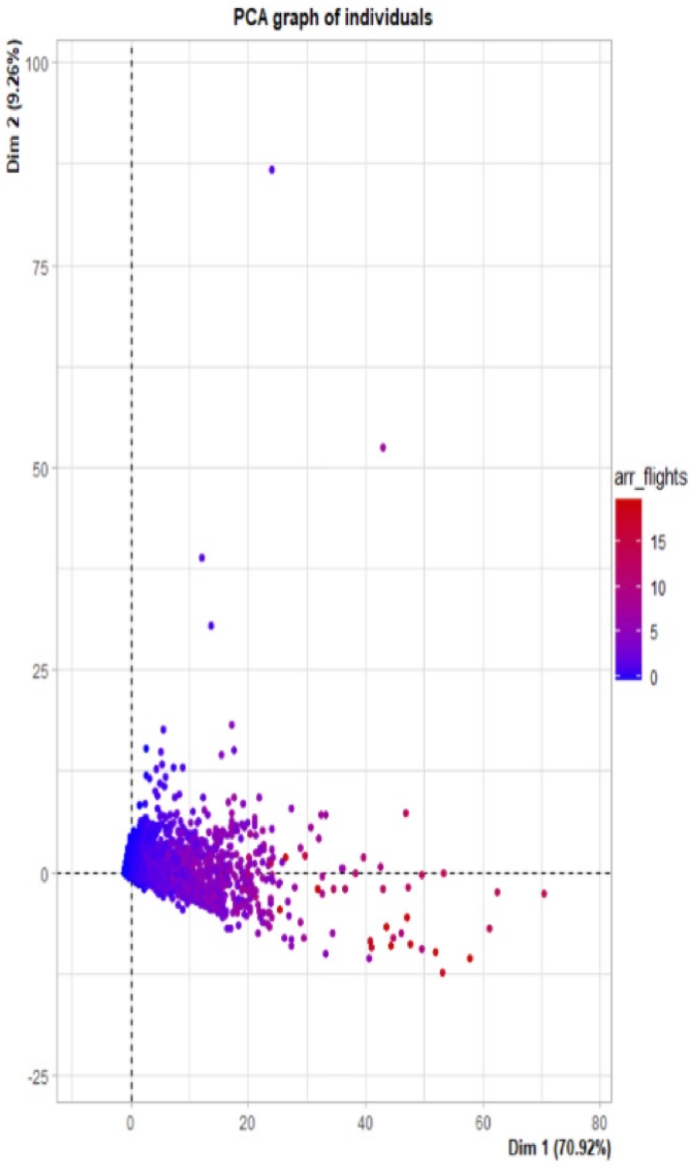
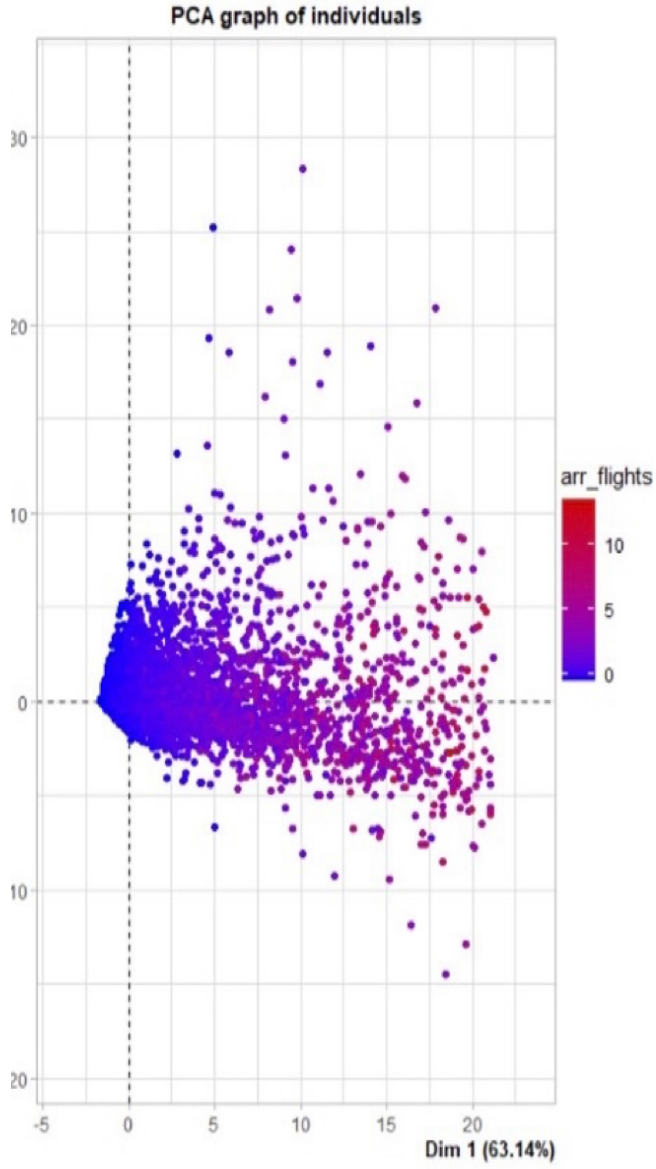


Figura 16: PCA dopo la rimozione degli outliers



REGRESSIONE LINEARE MULTIPLA

L'obiettivo del nostro modello è quello di prevedere, in minuti, il ritardo complessivo accumulato da una certa compagnia aerea in un dato mese e in un dato aeroporto, ovvero vogliamo prevedere il valore della variabile **carrier_delay**. Questo modello può cogliere quali siano le relazioni più forti connesse alla nostra variabile obiettivo e provare a spiegarne la variabilità, pur non potendo individuare la direzione di eventuali rapporti causali.

Abbiamo preso come punto di partenza, per la costruzione del modello, i risultati dell'analisi bivariata e dell'analisi fattoriale. Per poter costruire un buon modello di regressione è necessario che le variabili indipendenti non siano reciprocamente correlate tra loro.

Dall'analisi fattoriale risulta che dovremmo prendere come variabili indipendenti[9] **arr_delay**, **security_delay** e **weather_delay**. Da ciò sorgono due problemi: uno di natura tecnica e uno di buonsenso. Il problema di natura tecnica ci dice che non ha senso creare un modello di regressione per prevedere **carrier_delay** a partire da **arr_delay**, **weather_delay** e **security_delay**, in quanto equivarrebbe a cercare di prevedere il valore della somma fra **nas_delay e late_aircraft_delay**. Inoltre, il buonsenso ci dice che non sarebbe molto utile come modello, infatti, se non avessimo a disposizione i dati di una variabile **_delay**, che misura esattamente il numero di minuti di ritardo dovuti a una specifica causa, probabilmente non avremmo a disposizione neanche le altre variabili dello stesso tipo. Sarebbe più verosimile una situazione in cui vogliamo prevedere una variabile **_delay** con delle variabili esplicative di tipo diverso, come, per esempio, **arr_fligths**, **arr_del15**, una dummy o una variabile **_ct**. Immaginiamo di essere i manager di un aeroporto, e di voler stimare in un dato mese quanti minuti di ritardo si possono considerare causati dalla cattiva gestione di una nostra compagnia aerea partner. Potrebbe risultare utile per valutare la convenienza di una collaborazione con la stessa o con un'altra compagnia aerea. Delle diverse compagnie potremmo avere a disposizione il numero di voli arrivati in ritardo superiore a 15 minuti, il numero complessivo di voli e delle stime dei tempi medi di ritardo dovuti a cause specifiche. Potremmo però non avere a disposizione il conteggio esatto dei minuti di ritardo per ogni specifica causa, misurato dalle colonne **_delay**. Perciò abbiamo deciso di utilizzare le seconde variabili più importanti per la PCA e sufficientemente correlate con **carrier_delay**: **arr_del15**, **security_ct** e **weather_ct**. La variabile **weather_ct**, che indica il numero medio di minuti di ritardo dovuti alle condizioni metereologiche, risulta poco correlata con **arr_del15** ma abbastanza rilevante per **carrier_delay**.

Abbiamo perciò deciso di provare a includere anche la variabile **weather_ct**, che rappresenta le condizioni metereologiche mediamente registrate nel mese, anche se risulta essere poco correlata con **arr_flights** ma abbastanza rilevante per **carrier_delay**. Decidiamo di includere tutte le variabili dummies, dei mesi e delle compagnie, escludendo quelle degli aeroporti, in quanto significativamente correlate con quelle delle compagnie e molto più onerose da trattare in termini di calcoli.

Il nostro processo di costruzione si articola in tre fasi:

- Scelta delle variabili continue e delle varie categorie di dummies da includere
- Costruzione del modello, cercando degli outliers fra gli errori dallo stesso prodotti, per poi ricostruire il modello sulla base del dataset filtrato dalle righe che li producevano
- Rimozione delle variabili dummy poco significative utilizzando i p-value del Test-T e dei Test risultanti dalle tavole Anova. I p-value considerati sono solo quelli significativi al 5%

In totale otteniamo sei modelli. Il primo modello comprende tutte le variabili continue scelte e si basa sul dataset originario. Il secondo modello risulta molto simile al primo; l'unica differenza sta nel fatto che è stato addestrato sul dataset da cui avevamo rimosso gli indici delle osservazioni considerate outliers dalle prime tre componenti della PCA in base al metodo "Z-score". Nel terzo modello abbiamo aggiunto le variabili dummy, nel quarto modello sono state rimosse le osservazioni che producevano outliers fra gli errori, sempre in base al metodo "Z-score". Nei successivi due modelli abbiamo collassato il numero delle variabili esplicative rimuovendo iterativamente tutte quelle con un p-value maggiore del 5%, fino a quando, nel sesto modello non sono risultate tutte sufficientemente significative.

[9] Bisogna ricordare la relazione implicita tra le variabili, ossia: **arr_delay** = **carrier_delay** + **weather_delay** + **nas_delay** + **security_delay** + **late_aircraft_delay**.

Tabella 5: Principali effetti di ciascun passaggio, classificati per ogni modello

	adjRsq	IQR_residuals	Nvar
mI	0.8750899	395.4506	3
mII	0.8383297	357.9748	3
mIII	0.8567841	517.6460	32
mIV	0.8590744	515.6168	32
mV	0.8590644	514.2983	27
mVI	0.8564515	484.7806	13

Nella Tabella 5 si riassumono i principali effetti di ciascun passaggio, per ogni modello:

- adjRsq: R^2 aggiustato
- IQR_residuals: scarto interquartilico dei residui
- Nvar: numero delle variabili

Il nostro obiettivo è quello di massimizzare R^2 corretto e minimizzare gli altri due valori. Lo scarto interquartilico dei residui[10] può dire molto sia sull'ampiezza che sulla variabilità. Al crescere del numero di variabili k, inoltre, R^2 cresce artificiosamente. Dovendo scegliere fra due modelli, entrambi con oltre 10 variabili, uno con R^2 corretto poco più alto dell'altro, e l'altro con la metà del numero di variabili, probabilmente l'opzione migliore sarà l'uso del secondo, in quanto al crescere del numero di variabili cresce il rischio di overfitting e di multicollinearità. Tra più modelli, infatti, il più semplice spesso si rivela migliore.

```
> summary(ModelVI)

Call:
lm(formula = carrier_delay ~ weather_ct + arr_del15 + security_ct +
    carrierB6 + carrierYV + carrierDL + carrierHA + carrierUA +
    carrierF9 + carrierOO + mese1 + mese4 + mese6, data = VarNumDum)

Residuals:
     Min       1Q   Median       3Q      Max
-10595.0   -291.8     -4.5    193.0  17517.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -192.96462   11.19858 -17.231  < 2e-16 ***
weather_ct   124.84662    2.62302  47.597  < 2e-16 ***
arr_del15     20.64749    0.09603 215.004  < 2e-16 ***
security_ct -159.94974   13.71639 -11.661  < 2e-16 ***
carrierB6   1152.87255   37.29780  30.910  < 2e-16 ***
carrierYV    403.95978   32.19444  12.548  < 2e-16 ***
carrierDL    923.83917   26.59181  34.741  < 2e-16 ***
carrierHA    323.46147   61.71674   5.241 1.61e-07 ***
carrierUA    220.99370   29.36274   7.526 5.43e-14 ***
carrierF9    251.85573   31.54127   7.985 1.48e-15 ***
carrierOO    510.78785   21.54250  23.711  < 2e-16 ***
mese1         60.79053   20.07200   3.029  0.00246 **
mese4         85.93329   26.44879   3.249  0.00116 **
mese6         57.55810   26.45830   2.175  0.02961 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1007 on 20662 degrees of freedom
Multiple R-squared:  0.8565,    Adjusted R-squared:  0.8565
F-statistic:  9490 on 13 and 20662 DF,  p-value: < 2.2e-16
```

Il nostro modello spiega l'85.65% della variabilità totale, in base al valore di R^2. Le variabili esplicative continue[11] della regressione sono tutte significative a livello 0.001. La stima del coefficiente della variabile

[10] Minore è lo scarto interquartilico maggiore sarà la concentrazione degli errori intorno alla media, che assumiamo essere nulla.

[11] **Weather_ct**, **security_ct** e **arr_del15**

weather_ct è pari a 124.84 e indica l'incremento positivo in minuti della variabile **carrier_delay** che si avrebbe per ogni incremento unitario di **weather_ct**, in media e a parità di ogni altra condizione. La medesima interpretazione può essere fatta per le altre due variabili esplicative continue. Notiamo, in particolare, che **arr_del15** presenta un impatto positivo ma modesto sui minuti di ritardo causati dalle inefficienze di gestione della compagnia aerea, mentre al variare di **security_ct** vi è una variazione importante di segno opposto a **carrier_delay**. Possiamo fare analoghe considerazioni per ciascuno dei coefficienti delle variabili dummy che rappresentano, ceteris paribus, l'incremento o il decremento medio della variabile obiettivo al loro presentarsi.

Possiamo notare come i coefficienti delle dummy delle compagnie aeree riflettono variazioni molto più importanti nella variabile dipendente rispetto che a quelle dei mesi. Questo senso in quanto la variabile dipendente riflette le performance di puntualità di ogni compagnia aerea, che dipende più dalla gestione della singola compagnia che non dal periodo dell'anno. Anche i p-value dei relativi Test-T sono quasi sempre significativi per tutti gli usuali livelli di significatività, per le compagnie aeree, risultano invece più modesti, sebbene sempre significativi al 5%, per i mesi.

L'ipotesi nulla del Test-T, tutti significativi al 5%, è che il coefficiente di ciascuna variabile è uguale a 0, viceversa, l'ipotesi alternativa dice che il coefficiente di ciascuna variabile è diverso da 0. L'ipotesi nulla del Test-F è pari a 2,2e-16 è minore di tutti gli usuali livelli di significatività in quanto risulta praticamente nullo; proprio per questo motivo, si rifiuta l'ipotesi nulla.

La mediana dei residui è abbastanza vicina alla media dei residui, in quanto la mediana è pari a 12.4, mentre la media dei residui si assume essere pari a 0, con una deviazione standard dei residui nell'ordine del migliaio.

```
> anova(ModelVI)
Analysis of Variance Table

Response: carrier_delay
               Df     Sum Sq    Mean Sq    F value    Pr(>F)
weather_ct      1 6.6296e+10 6.6296e+10 65351.5468 < 2.2e-16 ***
arr_del15       1 5.6306e+10 5.6306e+10 55503.7708 < 2.2e-16 ***
security_ct     1 2.1315e+08 2.1315e+08   210.1133 < 2.2e-16 ***
carrierB6       1 6.9376e+08 6.9376e+08   683.8694 < 2.2e-16 ***
carrierYV       1 4.1454e+07 4.1454e+07    40.8633 1.667e-10 ***
carrierDL       1 9.5403e+08 9.5403e+08   940.4328 < 2.2e-16 ***
carrierHA       1 1.2636e+07 1.2636e+07    12.4563 0.0004175 ***
carrierUA       1 1.4555e+07 1.4555e+07    14.3479 0.0001524 ***
carrierF9       1 2.5927e+07 2.5927e+07    25.5577 4.330e-07 ***
carrierOO       1 5.7136e+08 5.7136e+08   563.2182 < 2.2e-16 ***
mese1           1 5.9130e+06 5.9130e+06     5.8287 0.0157756 *
mese4           1 9.4066e+06 9.4066e+06     9.2725 0.0023291 **
mese6           1 4.8009e+06 4.8009e+06     4.7325 0.0296094 *
Residuals   20662 2.0961e+10 1.0145e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 17: Anova

```
n   2.149100e+04
k   1.400000e+01
ssr 1.931008e+11
sse 1.634194e+10
sst 2.094428e+11
mse 7.609395e+05
msr 1.379292e+10
se  8.723185e+02
```

Dagli indici dell'Anova, contenuti nella Figura 17, desumiamo che il modello sia abbastanza buono in quanto la variabile da prevedere rappresenta i minuti di ritardo accumulati da un'intera compagnia in un aeroporto nell'arco di un mese, e, di conseguenza, l'errore standard risulta accettabile.

CLUSTERING

In questa sezione, abbiamo scelto di effettuare il clustering[12] sulle singole compagnie aeree e sui singoli aeroporti. Un'indagine sul clustering delle compagnie aeree, basata sui dati dei ritardi di volo, potrebbe essere utile per diversi scopi:

- Benchmarking delle prestazioni: il clustering delle compagnie aeree consente di effettuare un'analisi comparativa delle prestazioni tra gruppi simili. È possibile identificare le compagnie aeree con prestazioni superiori in termini di puntualità e quelle con problemi di ritardo ricorrenti.
- Ottimizzazione delle operazioni: l'analisi del clustering può offrire una visione approfondita delle relazioni tra le varie cause di ritardo e le compagnie aeree. Questo può aiutare le compagnie aeree a identificare i fattori chiave che contribuiscono ai ritardi e a prendere misure preventive.

Anche l'analisi del clustering degli aeroporti basata sui dati dei ritardi di volo potrebbe avere diverse applicazioni:

- Monitoraggio delle performance degli aeroporti: il clustering degli aeroporti permette di identificare gli aeroporti con una buona gestione dei ritardi e quelli che necessitano di miglioramenti. Ciò può fornire una base per il monitoraggio delle performance, l'individuazione delle migliori pratiche e l'implementazione di interventi correttivi mirati, dall'ampliamento delle piste di atterraggio a una diversa gestione del traffico aereo da parte della torre di controllo.
- Valutazione dell'impatto delle cause di ritardo: l'analisi del clustering può aiutare a valutare l'impatto delle diverse cause di ritardo specifiche per ogni aeroporto. Ad esempio, si potrebbero identificare cluster di aeroporti in cui le cause meteorologiche hanno un maggiore impatto sui ritardi o cluster in cui i problemi operativi sono più significativi. Queste informazioni possono essere utilizzate per sviluppare strategie di mitigazione dei ritardi adattate alle esigenze di ciascun aeroporto.

Figura 18: Codici indentificativi delle compagnie aeree del nostro campione

```
> CarrierNames
      [,1] [,2]
 [1,] "9E" "Endeavor Air Inc."
 [2,] "AA" "American Airlines Inc."
 [3,] "AS" "Alaska Airlines Inc."
 [4,] "B6" "JetBlue Airways"
 [5,] "DL" "Delta Air Lines Inc."
 [6,] "F9" "Frontier Airlines Inc."
 [7,] "G4" "Allegiant Air"
 [8,] "HA" "Hawaiian Airlines Inc."
 [9,] "MQ" "Envoy Air"
[10,] "NK" "Spirit Air Lines"
[11,] "OH" "PSA Airlines Inc."
[12,] "OO" "SkyWest Airlines Inc."
[13,] "UA" "United Air Lines Inc."
[14,] "WN" "Southwest Airlines Co."
[15,] "YX" "Republic Airline"
[16,] "QX" "Horizon Air"
[17,] "YV" "Mesa Airlines Inc."
```

Dato che la lista degli aeroporti è molto lunga, non verrà riportata in questo studio. Abbiamo suddiviso la procedura di calcolo del clustering in cinque fasi.

Nella prima fase selezioniamo come variabili di interesse quelli considerate rilevanti nella PCA, ovvero: **arr_delay**, **security_delay** e **weather_delay**. Delle 15 variabili di partenza ora ne usiamo solo tre, semplificando così notevolmente l'analisi, in quanto assumiamo che le variabili mancante siano spiegate da quelle incluse.

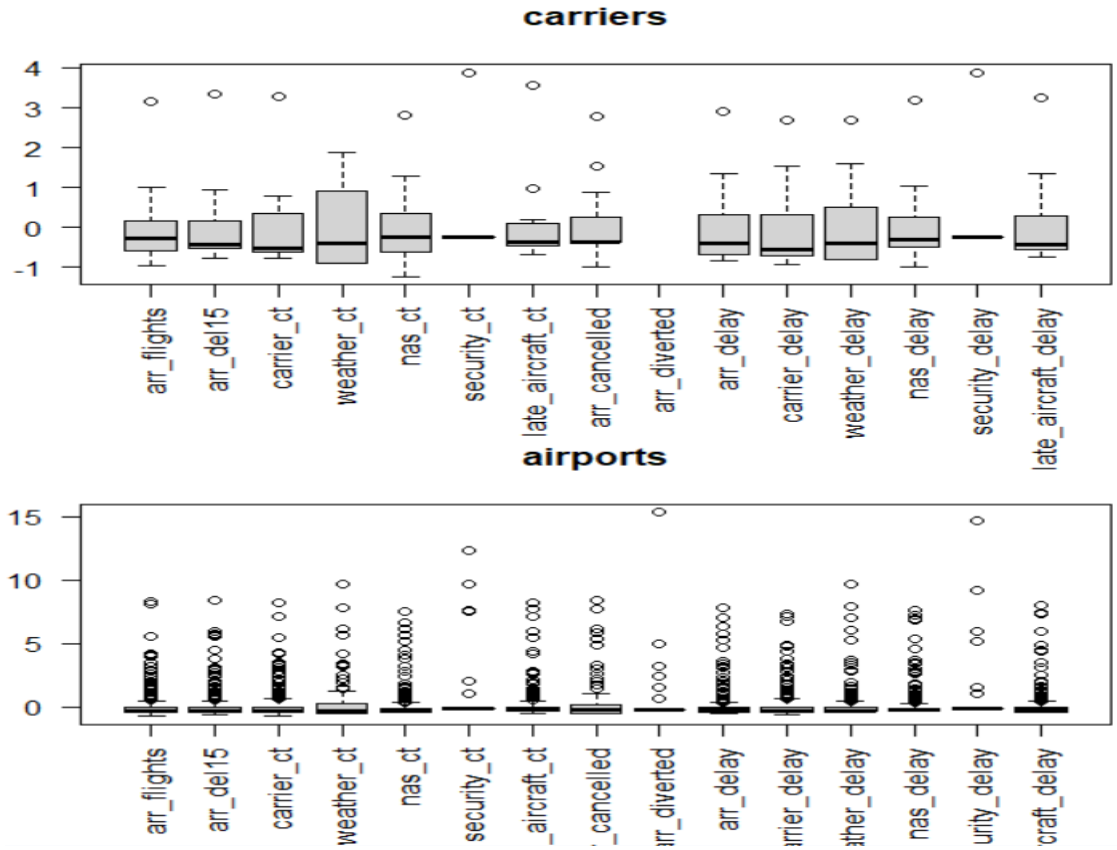
Successivamente, nel corso della seconda fase rimuoviamo dal dataset originario gli outliers selezionati attraverso la PCA e il metodo "Z-score", in modo tale da ridurre il più possibile le cause di rumore. Bisogna notare che abbiamo selezionato come outliers quelli risultanti come tali dagli "Z-score" delle coordinate delle

[12] L'analisi del clustering può fornire una panoramica dettagliata delle prestazioni, dei fattori influenti e delle sfide strategiche specifiche con cui ciascun aeroporto e ciascuna compagnia deve confrontarsi

tre componenti principali, e non da quelli delle tre variabili selezionate, che comunque le determinano nella proporzione maggiore.

Nella terza fase dobbiamo estrarre dai dati le informazioni utili a classificare la singola compagnia aerea e il singolo aeroporto. Nel nostro dataset compaiono 17 compagnie aeree e 373 aeroporti. Dobbiamo quindi costruire due diversi dataset, uno per le compagnie aeree e uno per gli aeroporti, al fine di svolgere le rispettive analisi di clustering. Prima di tutto normalizziamo i dati tramite il metodo dello "Z-score" in modo tale da rendere comparabili le variabili. Dato che vogliamo studiare le compagnie aeree dobbiamo ridurre tutte le righe che le contengono ad una sola riga, in quanto necessitiamo di un'unica osservazione per ciascuna compagnia. Per fare ciò filtriamo il dataset originario di 21.884 righe per le sole righe contenenti quella specifica compagnia. Dopodiché, per ciascuna colonna del dataset filtrato, ovvero per ciascuna variabile, possiamo estrarre alternativamente la somma, la media o la mediana delle righe. Dopo diverse prove, i nostri risultati sono migliorati notevolmente con l'uso della mediana, in quanto, a differenza degli altri due valori è insensibile al rumore prodotto da alcuni outliers ancora presenti. Nella costruzione del dataset e degli aeroporti, abbiamo svolto la medesima operazione. Ci troviamo ad avere a che fare con due nuovi dataset.

Figura 19: Rappresentazione sintetica delle variabili standardizzate con il metodo "Z-score"



Dalla Figura 19 potrebbe sembrare che vi siano molti outlier, anche se, in realtà, bisogna ricordare che ogni osservazione assume i valori delle mediane delle variabili nei dati filtrati

In questa quarta fase adottiamo il metodo grafico Elbow[13] per cercare di individuare il numero ottimale di cluster k in cui raggruppare i nostri dati. Al crescere di k sia la withing sum of square dei cluster, ovvero la loro dispersione interna, sia la silhouette media del modello, in linea generale, si riducono, finché non si annullano per $k = n$. L'obiettivo di un'analisi di clustering dovrebbe essere quello di minimizzare sia k sia la withing sum of square, nel caso dei kmeans, e di minimizzare k e l'altezza dei rami nel caso dei modelli gerarchici mantenendo la silhouette media più alta possibile. Nel caso del metodo kmeans la withing sum of square per ogni livello di k viene calcolata a partire dalla costruzione di diversi modelli di clustering. Ciascun

[13] Secondo il metodo, assolutamente visivo, Elbow si considera ottimale il numero di cluster k a cui corrisponde l'angolo del gomito sul grafico che mette in relazione k con la within sum of square, nel caso di modelli kmeans, o con l'altezza dei rami del dendogramma, nel caso dei modelli di clustering gerarchico

modello viene prodotto a partire dal posizionamento randomico dei k centroidi. Ogni modello kmeans, infatti, risulta sempre diverso da un altro modello kmeans, anche se costruito nella stessa maniera e con il medesimo k. Pertanto, se si ricostruisce il grafico del metodo Elbow più volte, si può pervenire a più curve con forme diverse, con l'angolo più ripido su diversi livelli di k. Abbiamo costruito il grafico Elbow del modello kmeans per le compagnie aeree e per gli aeroporti più volte e abbiamo tenuto molto spesso, come rispettivi valori di k, 4 e 13. Nel caso del metodo di clustering gerarchico tale problema non sussiste in quanto, trattandosi di un metodo bottom-up, non vi è posizionamento randomico dei centroidi e la soluzione si ripete sempre uguale a parità di condizioni. Dal grafico risulta comunque e sempre una tendenza in diminuzione dell'altezza dei rami del dendogramma, che deve per forza annullarsi nel momento in cui ogni cluster comprende una sola osservazione[14]. Per i dati sulle compagnie aeree con il kmeans e con il modello gerarchico con distanza euclidea e complete linkage, Figura 19, notiamo che i risultati corrispondono[15]. Per i dati sugli aereoporti con il kmeans e con il modello gerarchico con distanza euclidea e complete linkage, Figura 20, invece vi è differenza.

Figura 20: Elbow method to find optimal K (Kmeans): carrier (left) and airports (right)

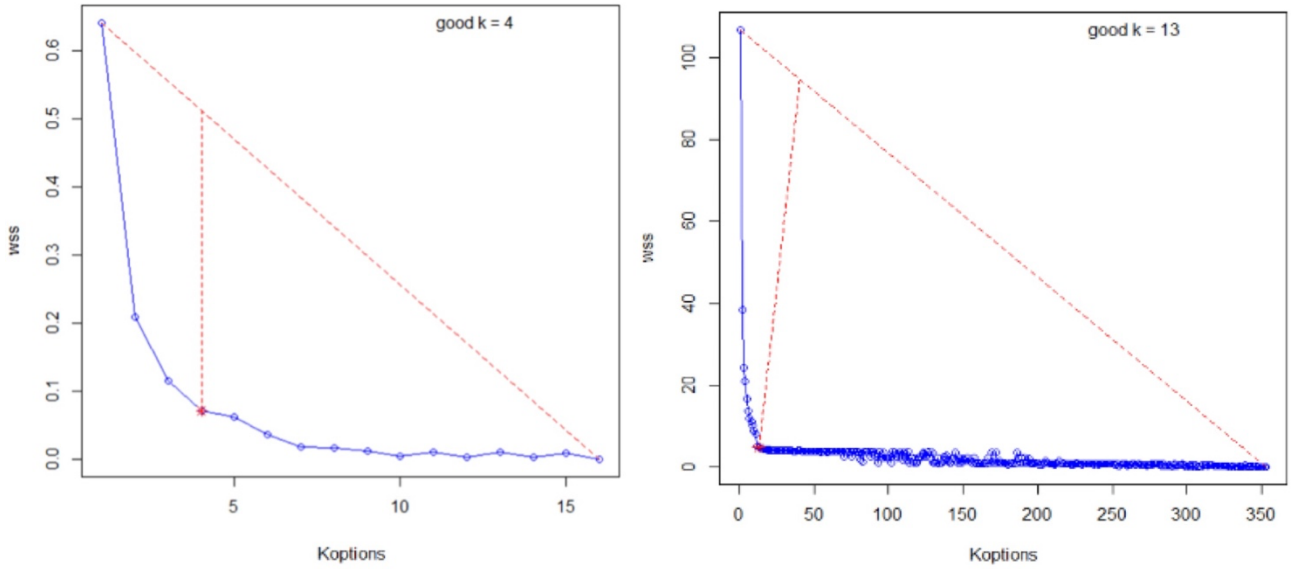
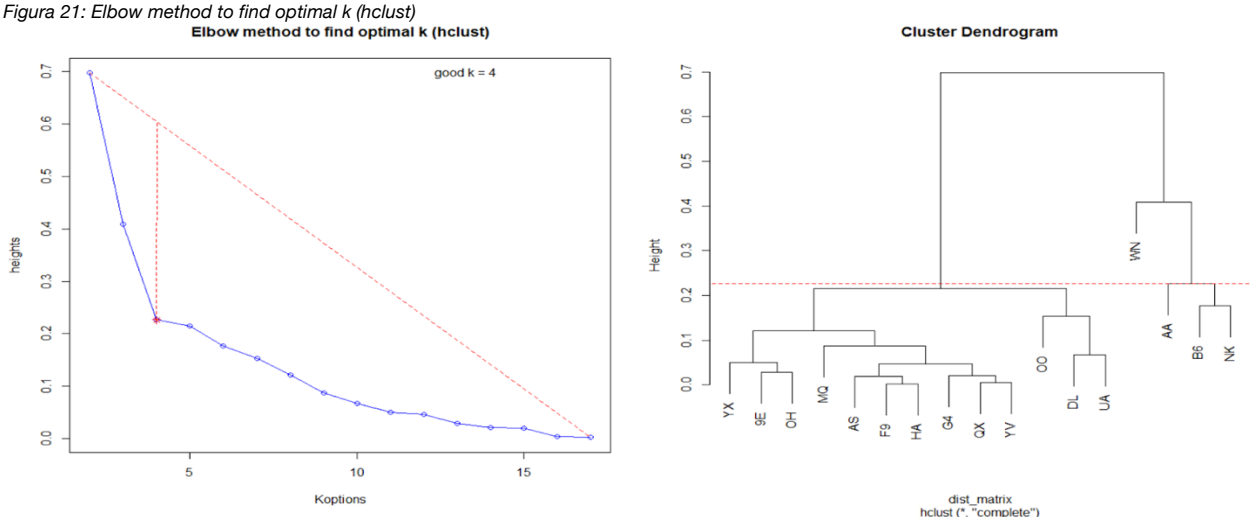
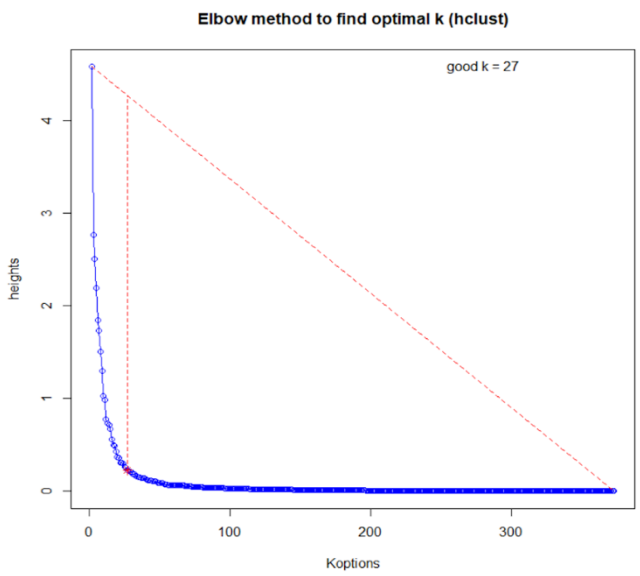


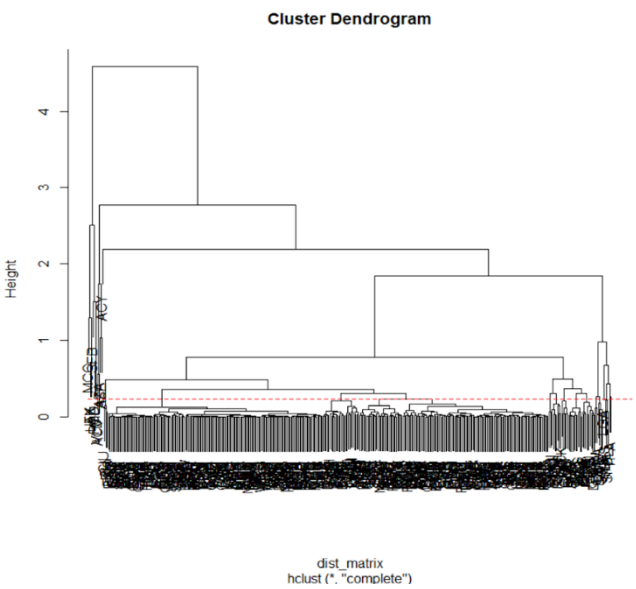
Figura 21: Elbow method to find optimal k (hclust)



[14] $k = n$ con altezza pari a 0

[15] La retta tratteggiata è la retta di lunghezza massima fra quelle passanti per la curva e perpendicolare alla retta che collega il punto $k = 1$ al punto $k = maxk$





Nella quinta ed ultima fase del lavoro, abbiamo provato a utilizzare il kmeans e il clustering gerarchico, nelle sue possibili configurazioni, ovvero con le diverse combinazioni delle metriche di distanza e dei metodi di linkage[16]. Abbiamo utilizzato k, come proposto dal metodo Elbow. In seguito, abbiamo selezionato il metodo che produceva una valutazione adeguata. Per scegliere il modello di clustering migliore ci siamo affidati a diversi indici

- k ottimo, trovato tramite il metodo Elbow, per ciascun modello
- Silhouette media, dato il k ottimo, che risulta tanto migliore quanto più si avvicina ad 1
- Coefficiente di Gini sulla disparità dimensionale dei gruppi, dato il k ottimo, in quanto consideriamo più interessanti e utili delle ripartizioni del campione in fasce il più possibile omogenei dimensionalmente, e che, pertanto, minimizzino il Gini calcolato sul vettore Cluster-Size
- Indice di Calinski-Harabasz, dato il k ottimo, che risulta tanto migliore quanto maggiore è il valore assoluto
- Indice di Davis-Bouldin, dato il k ottimo, che risulta tanto migliore quanto minore è il valore assoluto

La ragione per cui abbiamo considerato anche l'equità fra le dimensioni dei gruppi è semplice, infatti con certe tecniche di clustering, si ottengono k gruppi di cui $k-1$ di dimensioni molto piccole, spesso contenenti anche una sola osservazione. Riteniamo che un'analisi interessante ed efficace sia in grado di riconoscere diverse fasce di osservazioni nella popolazione, e che quindi, debba attuare un'equa segmentazione dei gruppi all'interno del campione, che assumiamo essere sufficientemente rappresentativo della popolazione.

[16] Metodi: complete, single e average

1. Clustering delle compagnie aeree

Il clusterung delle compagnie aeree si basa sulle categorie della variabile _**carrier**.

Figura 22: Metodi alternativi di clustering che avremmo potuto adottare

```
> CarrierCluster
model                          k   mean silhouette     Gini-ClusterSize    Calinski-Harabasz    Davis-Bouldin
"kmeans"                       "4" "0.452685824472597" "0.529411764705882" "-2.732197913081"    "0.723865794378367"
"hierarchic-euclidean-complete" "4" "0.495655353112162" "0.725490196078431" "-2.86624591842637"  "0.479724046389455"
"hierarchic-euclidean-single"  "5" "0.42853954195004"  "0.705882352941176" "-1.83446962302441"  "0.236281418531684"
"hierarchic-euclidean-average" "4" "0.495655353112162" "0.725490196078431" "-2.86624591842637"  "0.479724046389455"
"hierarchic-manhattan-complete" "7" "0.399498719455126" "0.568627450980392" "-0.905466938155061" "0.267347085314219"
"hierarchic-manhattan-single"  "5" "0.24605979313553"  "0.705882352941176" "-1.80431034550067"  "0.428807722116324"
"hierarchic-manhattan-average" "8" "0.281066853032301" "0.512605042016807" "-0.642879911996669" "0.278677706714574"
"hierarchic-maximum-complete"  "4" "0.461736323707484" "0.529411764705882" "-2.732197913081"    "0.723865794378367"
"hierarchic-maximum-single"    "6" "0.430684356170838" "0.6"               "-1.2758393331893"   "0.412764940899319"
"hierarchic-maximum-average"   "4" "0.452342311152982" "0.725490196078431" "-2.65411405198591"  "0.59402145638308"
"hierarchic-canberra-complete" "6" "0.341736436593021" "0.647058823529412" "-1.27735996930399"  "0.297955419885197"
"hierarchic-canberra-single"   "6" "0.341736436593021" "0.647058823529412" "-1.27735996930399"  "0.297955419885197"
"hierarchic-canberra-average"  "6" "0.341736436593021" "0.647058823529412" "-1.27735996930399"  "0.297955419885197"
"hierarchic-minkowski-complete" "4" "0.495655353112162" "0.725490196078431" "-2.86624591842637" "0.479724046389455"
"hierarchic-minkowski-single"  "5" "0.42853954195004"  "0.705882352941176" "-1.83446962302441"  "0.236281418531684"
"hierarchic-minkowski-average" "4" "0.495655353112162" "0.725490196078431" "-2.86624591842637"  "0.479724046389455"
```

Il modello kmeans risulta positivo per la maggior parte dei nostri indici. L'indice di Gini è relativamente alto, ma a fronte di un k relativamente altrettanto basso. L'indice di Davis-Bouldin è l'unico parametro che ne sconsiglia l'uso.

Figura 23: Dendogramma delle Compagnie aeree, secondo la distanza euclidea e complete linkage

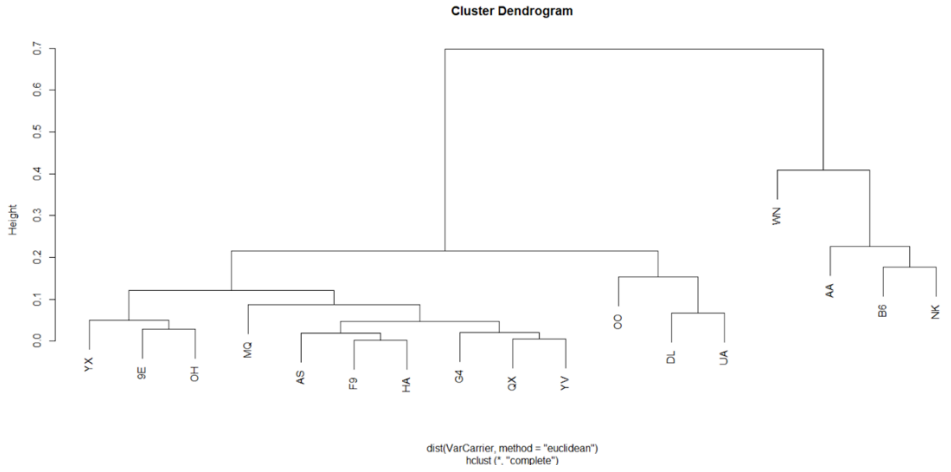


Figura 24: Cluster ottenuti dal modello gerarchico

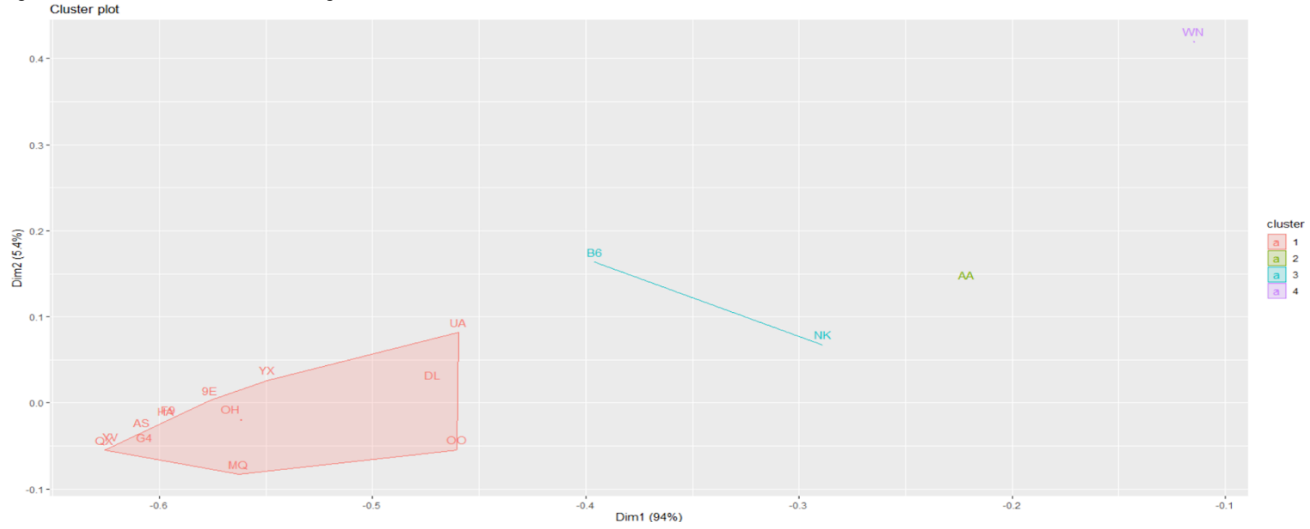


Figura 25: Silhoutte del clustering gerarchico (eucliderian-complete) sulle compagnie aeree

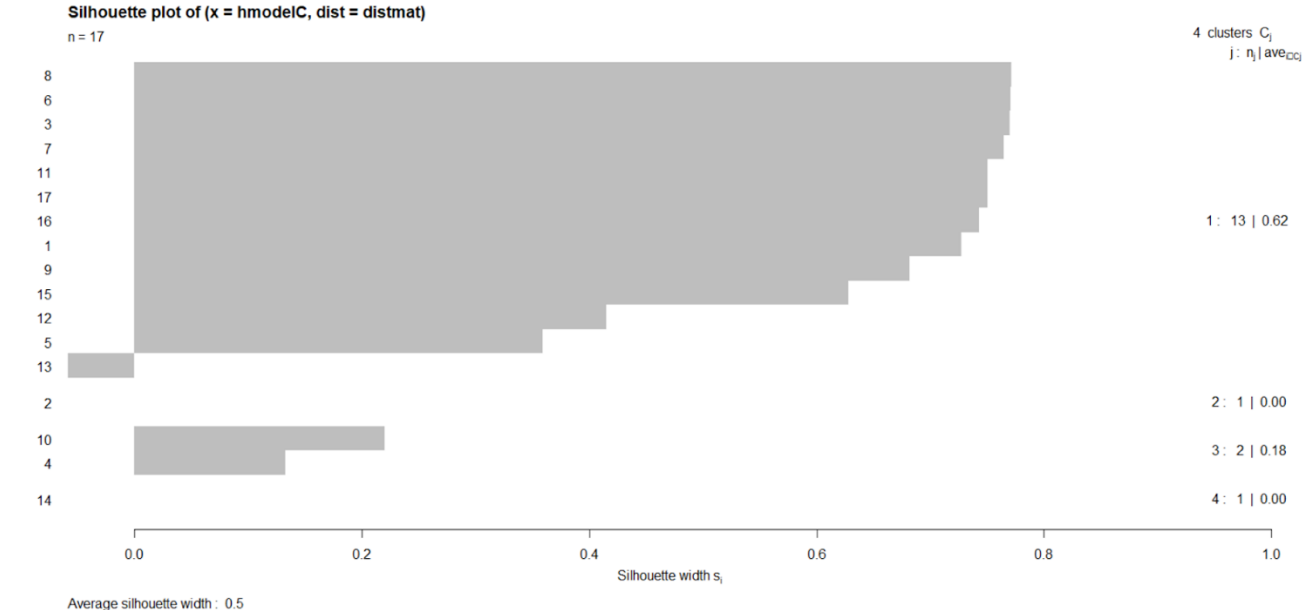
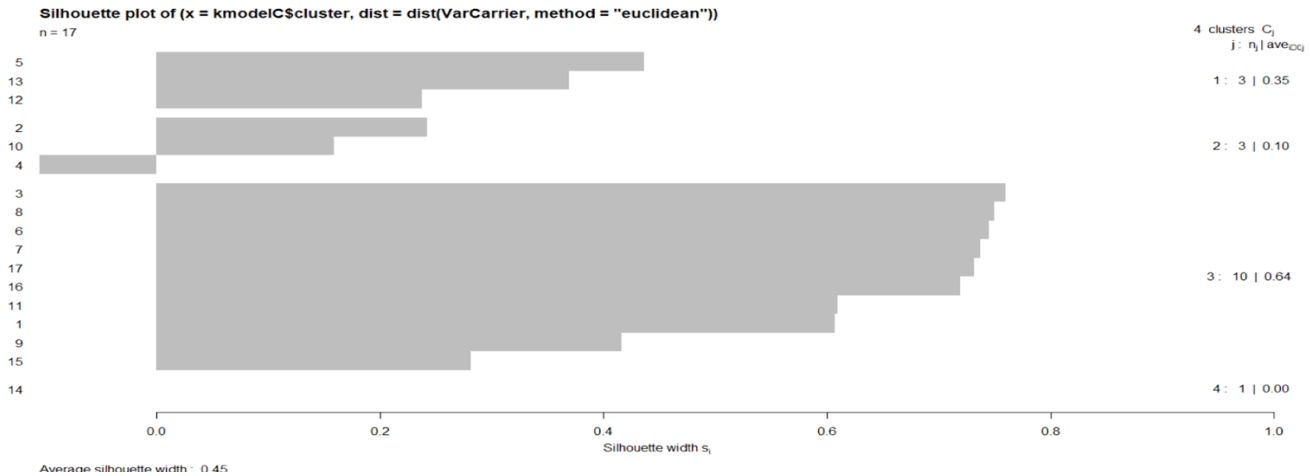


Figura 26: Cluster ottenuti dal modello kmeans sulle compagnie aeree

Figura 27: Silhouette del kmeans sulle compagnie aeree



Dal dendogramma vediamo che i gruppi costruiti dal modello gerarchico sono molto simili ai risultati del kmeans, il quale però presenta una composizione più equa dei gruppi e degli indici di valutazione accettabili.

Figura 28: Dispersione del kmeans

```
> kmodelC = kmeans(SVarCarrier, 4)
> ksilC = silhouette(kmodelC$cluster, dist(SVarCarrier, method='euclidean'))
> mean(ksilC[,3])
[1] 0.4526858
> kmodelC$size
[1] 10  1  3  3
> kmodelC$withinss
[1] 0.01735524 0.00000000 0.04132577 0.01196566
> kmodelC$tot.withinss
[1] 0.07064668
> kmodelC$betweenss/(kmodelC$tot.withinss+kmodelC$betweenss)
[1] 0.8897172
```

Notiamo una bassa within sum of square in tutti I cluster, un'elevata silhouette media e una between sum of square del 88%. I cluster sembrano ben costruiti e distanziati.

Figura 29: Centroidi del kmeans

```
> centri[1:8]
  arr_flights arr_del15 carrier_ct weather_ct   nas_ct security_ct late_aircraft_ct arr_cancelled
1     64.0000  12.35000    5.00000  0.1170000 2.889000  0.00000000         3.400500           0.9
2    330.0000  90.00000   40.15000  0.8100000 9.235000  0.00000000        37.965000           6.0
3    144.3333  37.33333   15.17667  0.8400000 9.835000  0.03666667        10.666667           3.0
4    120.0000  21.00000   11.27000  0.7716667 3.016667  0.00000000         4.976667           1.0
> centri[9:15]
  arr_diverted arr_delay carrier_delay weather_delay nas_delay security_delay late_aircraft_delay
1            0     683.1       258.600       5.90000  111.1500       0.000000               210.7
2            0    4569.0      1841.000      45.00000  393.0000       0.000000              2225.5
3            0    2499.0      1030.167      54.83333  481.1667       1.666667               808.0
4            0    1371.0       707.000      43.00000  115.5000       0.000000               364.0
```

Abbiamo tenuto i valori medi di ciascun cluster. Ma qual è il cluster migliore? Non possiamo basarci solo sui valori medi di **arr_delay**, in quanto è facile notare come sia correlato con **arr_flights**, infatti, i cluster con maggior numero di voli in arrivo presentano un accumulo di ritardi maggiore. Inoltre, vediamo che **carrier_delay** segue **arr_delay**, ossia le compagnie a cui sono attribuiti più minuti mensili di ritardo per problemi di gestione sono anche quelle con più minuti di ritardo in generale. Dati dunque minuti di ritardo complessivi, e il numero di voli effettuati, per poter individuare il cluster più performante nella capacità di minimizzare i minuti di ritardo dobbiamo prendere il rapporto tra queste due variabili.

*Figura 30: Rapporto tra **carrier_delay** e **arr_delay***

```
> matrix(centri[,'arr_delay']/centri[,'arr_flights'])
         [,1]
[1,] 10.67344
[2,] 13.84545
[3,] 17.31409
[4,] 11.42500
```

Come vediamo dalla Figura 30, il cluster 1 può considerarsi un benchmark di riferimento, in quanto minimizza il rapporto. Tale cluster può essere utilizzato come benchmark per valutare la capacità di una data compagnia aerea nel limitare i ritardi dei voli, inoltre, contiene la maggior parte delle compagnie aeree, 10 su 17, il 59% del totale. Il cluster 3 si distingue dagli altri per gli elevati ritardi per condizioni metereologiche e di sicurezza, nonché per la cattiva comunicazione con la torre di controllo, presentando il rapporto peggiore tra **arr_delay** e **arr_flights**. Lo stesso può sembrare eccessivamente disperso se si valuta la sua silhouette, ma dal suo posizionamento sulle dimensioni della PCA appare come un cluster ben costruito. Il cluster 2 contiene una sola osservazione, la compagnia WN, Southwest Airplines, che ha effettuato il maggior numero di voli, accumulando così più minuti di ritardo di tutte le altre e il maggior numero di voli in ritardo di oltre 15 minuti. Il rapporto con **arr_flights** però la rileva come una compagnia con performance più vicine a quella dei cluster di benchmark che non a quelle delle compagnie peggiori, ovvero quelle del cluster 3.

2. Clustering degli aeroporti

Abbiamo qui adottato le medesime metodologie usate nel clustering delle compagnie aeree. Abbiamo costruito un modello kmeans e l'abbiamo confrontato con i diversi modelli gerarchici. Anche in questo caso però, abbiamo optato per il modello kmeans.

Figura 31: Metodi considerati

```
> AirportCluster
                                model k    mean silhouette     Gini-ClusterSize    Calinski-Harabasz   Davis-Bouldin
"kmeans"                          "13" "0.456313155960187" "0.73503127792672"  "-29.4962697950752" "0.717971258455975"
"hierarchic-euclidean-complete"   "27" "0.413729523739904" "0.832130336151784" "-13.0656115018581" "0.426970595146488"
"hierarchic-euclidean-single"     "26" "0.52093160262556"  "0.927506702412868" "-13.6425781539313" "0.173477410645058"
"hierarchic-euclidean-average"    "20" "0.584486332411759" "0.9311415267391"   "-18.264646705065"  "0.301953005266961"
"hierarchic-manhattan-complete"   "25" "0.428142201922452" "0.854334226988382" "-14.2511545815903" "0.462677377572853"
"hierarchic-manhattan-single"     "27" "0.515627923051405" "0.923283151165189" "-13.0794909735805" "0.182072591110687"
"hierarchic-manhattan-average"    "26" "0.513945933298432" "0.905415549597855" "-13.6448163242518" "0.296018270381191"
"hierarchic-maximum-complete"     "29" "0.424686641962178" "0.839142091152815" "-12.0689682585266" "0.481429218763308"
"hierarchic-maximum-single"       "26" "0.52894005877859"  "0.923217158176944" "-13.6404715343817" "0.214251672545608"
"hierarchic-maximum-average"      "29" "0.477370704598781" "0.883761011106855" "-12.0669986342462" "0.303985113472474"
"hierarchic-canberra-complete"    "58" "0.407853771996199" "0.757866516156343" "-5.41016365085878" "0.418619534525469"
"hierarchic-canberra-single"      "61" "0.409643320691418" "0.834226988382484" "-5.09914684598626" "0.192357250541946"
"hierarchic-canberra-average"     "49" "0.488887056818136" "0.846067917783736" "-6.61960789286967" "0.325055380816571"
"hierarchic-minkowski-complete"   "27" "0.413729523739904" "0.832130336151784" "-13.0656115018581" "0.426970595146488"
"hierarchic-minkowski-single"     "26" "0.52093160262556"  "0.927506702412868" "-13.6425781539313" "0.173477410645058"
"hierarchic-minkowski-average"    "20" "0.584486332411759" "0.9311415267391"   "-18.264646705065"  "0.301953005266961"
```

Figura 32: Cluster ottenuti dal modello kmeans sugli aeroporti

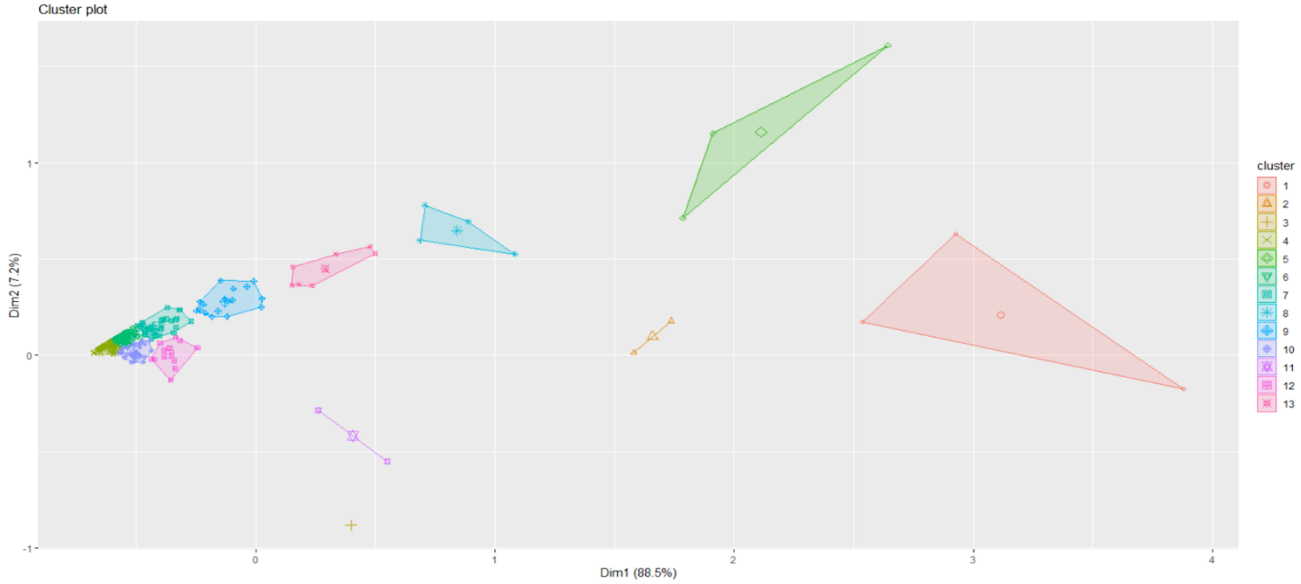
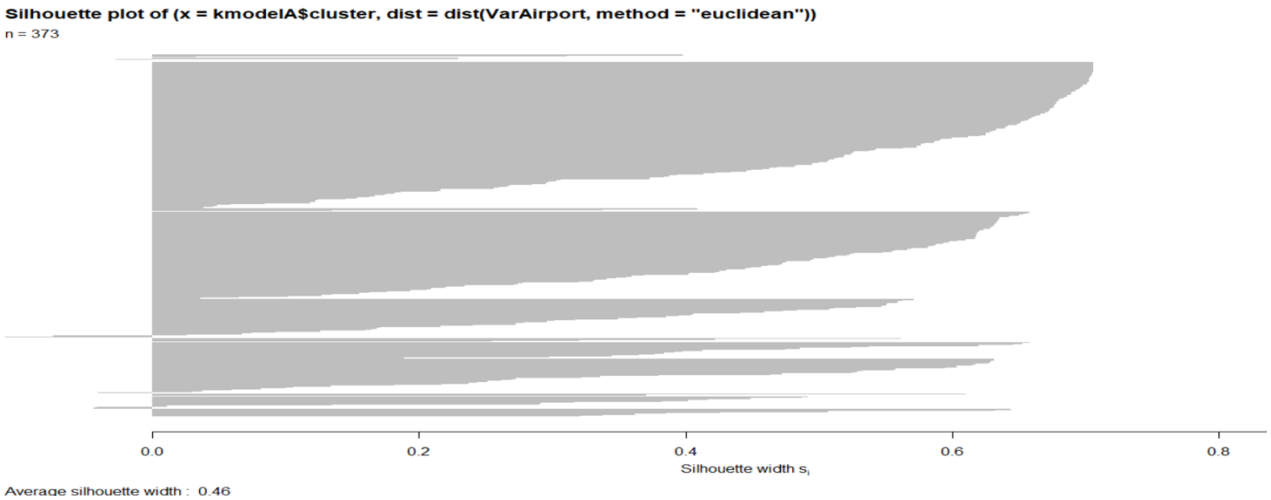


Figura 33: Silhouette plot



La silhouettes media, raffigurata in Figura 33, pari al 46%, risulta accettabile.

Figura 34: Centroidi del kmeans

```
> centri[1:8]
   arr_flights  arr_del15 carrier_ct  weather_ct     nas_ct security_ct late_aircraft_ct arr_cancelled
1    579.66667 129.750000  44.905000  3.13500000 41.1883333   0.0000000        33.949167     9.1666667
2     88.65385  16.153846   8.499231  1.72307692  1.6392308   0.0000000         3.176154     1.0000000
3    497.75000 137.000000  57.805000  6.65000000 25.6050000   0.3950000        46.785000    14.5000000
4     45.73103   6.600000   2.845517  0.09365517  0.9194483   0.0000000         1.702483     0.4965517
5   1221.50000 269.666667  85.976667  4.57000000 75.5983333   0.2050000        67.770000    14.0000000
6     66.94872   9.897436   4.396667  0.94602564  1.1253846   0.0000000         2.234103     0.8333333
7    139.17500  26.987500  10.791500  0.67737500  5.1520000   0.0000000         7.201750     1.8000000
8    246.64706  53.205882  19.803235  1.04617647 11.1591176   0.0000000        14.715000     3.5000000
9    126.50000  17.000000  11.220000  1.18500000  1.1950000   0.0000000         3.800000     3.5000000
10   382.77778  81.666667  29.558889  1.73444444 21.5116667   0.0000000        17.842778     5.3888889
11    78.57527  14.521505   5.966183  0.21279570  2.2316667   0.0000000         4.058226     0.9462366
12   632.66667 215.000000  50.466667 10.41666667 69.1266667   0.2933333        91.966667    23.6666667
13   257.00000  55.000000  19.870000  1.82000000 18.4000000   1.0000000        12.860000    10.0000000
> centri[9:15]
   arr_diverted  arr_delay carrier_delay weather_delay  nas_delay security_delay late_aircraft_delay
1    0.75000000  8529.1667     3092.1667    249.083333 2042.16667       0.000000          2641.83333
2    0.15384615  1176.6538      550.4615    131.346154   67.30769       0.000000           208.46154
3    4.50000000 10262.7500     5158.2500    717.500000 1068.50000      11.000000          3296.75000
4    0.03103448   364.6414      150.2621      3.717241   33.68276       0.000000            96.90345
5    1.33333333 17258.6667     6004.0000    423.166667 3410.16667       6.333333          5415.50000
6    0.07692308   634.2179      262.1410     61.653846   43.50000       0.000000           144.50000
7    0.01250000  1664.2750      660.9000     32.937500  212.17500       0.000000           492.35000
8    0.17647059  3217.8824     1252.8824     68.411765  473.14706       0.000000          1057.73529
9    1.00000000  1779.5000      681.0000    656.500000   48.50000       0.000000           266.50000
10   0.44444444  5143.1111     1864.2778    141.277778 1007.00000       0.000000          1392.33333
11   0.03225806   899.4301      361.5538      8.709677   87.88710       0.000000           258.36559
12   2.00000000 16691.3333     4361.0000   1182.666667 3439.66667      14.666667          6550.66667
13   0.00000000  4130.0000     1277.0000     40.000000  951.00000      54.000000          1174.00000
```

L'interpretazione su quali siano gli aeroporti migliori deriva dalle coordinate dei centroide. Una possibile misurazione delle performance può consistere nel rapporto, già descritto, fra **arr_delay** e **arr_flights**, riassunto nella Figura 35.

*Figura 35: Rapporto tra **arr_delay** e **arr_flights***

```
> matrix(centri[,'arr_delay']/centri[,'arr_flights'])
           [,1]
 [1,] 14.713916
 [2,] 13.272451
 [3,] 20.618282
 [4,]  7.973609
 [5,] 14.129076
 [6,]  9.473190
 [7,] 11.958146
 [8,] 13.046506
 [9,] 14.067194
[10,] 13.436284
[11,] 11.446733
[12,] 26.382508
[13,] 16.070039

> ksilA = silhouette(kmodelA$cluster, dist(SVarAirport, method='euclidean'))
> mean(ksilA[,3])
[1] 0.4532168
> kmodelA$size
 [1]   6  13   2 145   3  39  40  17   2   9  93   3   1
> kmodelA$withinss
 [1] 0.4170490 0.1240698 0.5331176 0.1517656 0.9886581 0.1268863 0.1850890
 [8] 0.2005974 0.1582420 0.2052260 0.1226584 1.4633629 0.0000000
> kmodelA$tot.withinss
[1] 4.676722
> kmodelA$betweenss/(kmodelA$tot.withinss+kmodelA$betweenss)
[1] 0.9562434
```

Anche in questo caso notiamo che la betweenss% è molto buona, pari al 95%, perciò deduciamo che vi sia notevole distanziamento tra i cluster a fronte di un basso livello di dispersione all'interno di ciascun cluster.

CONCLUSIONI

Con il presente lavoro abbiamo ottenuto risultati utili per poter comprendere meglio il fenomeno dei ritardi di volo nelle linee aeree statunitensi, sotto diversi punti di vista. Nella domanda di ricerca iniziale, ci siamo chiesti se fosse possibile individuare le principali cause di ritardo dei voli e se vi fossero correlazioni fra le stesse. I risultati dell'analisi bivariata e dell'analisi fattoriale dimostrano che certe cause di ritardo coesistono, come il numero di minuti di ritardo dovuti a problemi di gestione della compagnia aerea e il numero di voli effettuati nel medesimo aeroporto. Deduciamo infatti, anche dalla regressione, che molte compagnie aeree peggiorano nella performance di gestione quando si presentano fenomeni metereologici o di sicurezza avversi, o quando aumentano i tempi di ritardo attribuiti a problemi connessi alla torre di controllo. Non possiamo definire le direzioni dei rapporti di causa, ma possiamo affermare se due fenomeni sono connessi o meno.

Inizialmente ci siamo anche chiesti se fosse possibile ricavare informazioni utili per i passeggeri, in modo da poter scegliere la combinazione ottima di mese, compagnia aerea e aeroporto, per viaggiare evitando possibili ritardi.

I nostri dati sono purtroppo limitati ad un'osservazione globale del fenomeno e non facilitano l'analisi approfondita delle caratteristiche del singolo volo. I risultati ottenuti non consentono pertanto di soddisfare questo obiettivo, ma consentono comunque di svolgere un'analisi comparativa delle diverse performance di compagnie ed aeroporti.

La regressione lineare multipla ci ha consentito di definire la natura e il peso dei fattori connessi con la variabile **carrier_delay** e quindi di raccogliere informazioni che dal punto di vista strategico possono risultare molto utili per una compagnia. I coefficienti trovati sulle dummy delle compagnie e dei mesi possono inoltre aiutare i clienti a scegliere la compagnia e il mese migliore per viaggiare minimizzando i ritardi, in quanto i ritardi complessivi attribuita alla compagnia sono fortemente correlati con i ritardi complessivi in generale.

L'analisi del clustering, infine, ci ha aiutato a classificare rapidamente le diverse fasce di performance in cui ogni compagnia aerea e ogni aeroporto può inserirsi, nonché a individuare le compagnie e gli aeroporti più performanti. Ciò ha permesso anche di definire dei valori benchmark di riferimento tramite i quali è ora possibile valutare le prestazioni di altre compagnie aeree, operanti in altre aree geografiche o esistenti in altre epoche.

BIBLIOGRAFIA

Baarsch, J., & Celebi, M. E. (2012, March). Investigation of internal validity measures for K-means clustering. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, pp. 14-16). sn.

Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2016). Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(1), 1-20.

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

Dogan, R. I. (1995). Principal component analysis.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, *3*, 409-425.

Forina, M., Armanino, C., & Raggio, V. (2002). Clustering with dendrograms on interpretation variables. *Analytica Chimica Acta*, *454*(1), 13-19.

Han, J., Zhu, L., Kulldorff, M., Hostovich, S., Stinchcomb, D. G., Tatalovich, Z., ... & Feuer, E. J. (2016). Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. *International journal of health geographics*, *15*, 1-11.

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, *30*(2), 87-93.

Jeromy Anglim (https://stats.stackexchange.com/users/183/jeromy-anglim), Reporting coefficient of determination using Spearman's rho, URL (version: 2017-04-13): https://stats.stackexchange.com/q/61726

Kalliguddi, A. M., & Leboulluec, A. K. (2017). Predictive modeling of aircraft flight delay. *Universal Journal of Management*, *5*(10), 485-491.

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231-238.

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. *Korean journal of anesthesiology*, *70*(1), 22-26.

Kleine d. (2021, May 27*). Detecting Knee/Elbow Points in a Graph*. *Retrieved from Towards Data Science website.*

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299-3303.

Rupp, N. G., Holmes, G. M., & Center, C. G. S. (2004). The Flight Operations Decision Process.

Russon, M. G., & Neumann, J. J. (2018). Minimum sum regression as the optimum robust algorithm in the computation of financial beta.

Shih, J. H., & Fay, M. P. (2017). Pearson's chi-square test and rank correlation inferences for clustered data. *Biometrics*, *73*(3), 822-834.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E., Herrera, F., & López-Cózar, E. D. (2013). On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology*, *64*(7), 1468-1479.

West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162-165.

SITOGRAFIA

https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

https://www.iaeng.org/publication/IMECS2012/IMECS2012_pp471-476.pdf

https://dl.acm.org/doi/abs/10.1145/2888402?casa_token=y5P5HocNgWAAAAAA:NZcsZmSWKt3MfdA2sWwn1OHaAYrNbbFxGQhWpAzuAISgs5j6WUiEGAlCqbd67jj1lUKesNid4lU

https://www.clausiuspress.com/assets/default/article/2020/10/22/article_1603378206.pdf

https://www.cs.umd.edu/class/spring2018/cmsc644/PCA.pdf

https://link.springer.com/article/10.1023/A:1009868929893

https://www.sciencedirect.com/science/article/pii/S0003267001015173?casa_token=3n6EOV1AuucAAAAA:3VaV7kgVBKNGUClLJnVMBmUAg7f12IoNeKPETxu5SMYo9VhdX0J0WsrnI3o0jAR7ebXDr1Qb

https://link.springer.com/article/10.1186/s12942-016-0056-6

https://sciendo.com/downloadpdf/journals/quageo/30/2/article-p87.pdf

https://stats.stackexchange.com/questions/44268/reporting-coefficient-of-determination-using-spearmansrho#:~:text=Spearman's%20rho%2C%20for%20example%2C%20represents,the%20proportion%20of%20shared%20variance

https://dl.acm.org/doi/abs/10.1145/2888402?casa_token=y5P5HocNgWAAAAAA:NZcsZmSWKt3MfdA2sWwn1OHaAYrNbbFxGQhWpAzuAISgs5j6WUiEGAlCqbd67jj1lUKesNid4lU

https://www.researchgate.net/profile/ManojKuppusamy/publication/283755180_Labeling_Methods_for_Identifying_Outliers/links/5646a95008ae451880aa6d84/Labeling-Methods-for-Identifying-Outliers.pdf

https://synapse.koreamed.org/articles/1156679

https://medium.com/towards-data-science/detecting-knee-elbow-points-in-a-graph-d13fc517a63c

https://www.researchgate.net/profile/Dauda-Usman/publication/288044597_Standardization_and_Its_Effects_on_K-Means_Clustering_Algorithm/links/56b5f9b908aebbde1a79bce7/Standardization-and-Its-Effects-on-K-Means-Clustering-Algorithm.pdf

https://www.researchgate.net/profile/Anish-Kalliguddi/publication/321800887_Predictive_Modeling_of_Aircraft_Flight_Delay/links/5be20a3992851c6b27ab2f60/Predictive-Modeling-of-Aircraft-Flight-Delay.pdf

https://www.businessperspectives.org/images/pdf/applications/publishing/templates/article/assets/8106/imfi_en_2016_04cont_Russon.pdf

https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12653?casa_token=r-LUXF_vBWQAAAAA%3ATqiQl6gJx8nJsnkjmSYQrXu1skcd0aK2IoEvDJjo_IslBYntttPXroNyeV9mu_ZVju_wl2L427YW9A

https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.22837?casa_token=Hvr9uJAsW3MAAAAA%3Ao7LqB8PGN_f3RklV5v5kXzqE2l6wsjQfhDB-lzykzD0-ZRqzKu70vuqxotOQ2iaTOI5rCBSRNG8elg

https://journals.sagepub.com/doi/pdf/10.1177/00045632211050531