Statistical learning exam

A market analysis of the airline sector

Customer satisfaction analysis with PCA - clustering of the main airlines companies - logistic models and decision trees

Antonio De Patto

Matriculation number - 46430A

1 - Unsupervised analysis

When the variables considered are numerous it is difficult to fully understand the structures existing between data and a significant interpretation is complex to draw. So, we have to reduce the dimension of the data with a limited information loss and this is possible using a multivariate statistical method called Principal Component Analysis. The main purpose of PCA is to create a linear combination of our original variables which explain away as much variance in our data as possible so that we can lower the number of variables which explain most of the variation. PCA can be conducted starting from the covariance matrix or from the correlation matrix. In the first case each component is expressed as a linear combination of the deviations from the mean of the p variables. In the case of the correlation matrix the components are considered as a linear combination of the standardized deviations.
When the original variables are expressed in different units of measurement and/or have very different orders of magnitude, they are not directly comparable and therefore the PCA starting from the covariance matrix proves to be inappropriate. This difficulty can be overcome by considering the variables expressed in terms of standardized deviations, which is equivalent to taking the correlation matrix instead of the covariance matrix as the starting point of the PCA.
Since PCA is based on variance maximization and correlation coefficients we need to have adequate correlations between variables in order to be reduced to a smaller number of components. If the variables are correlated with each other it is possible to identify a small number of dimensions and summarize the information available starting from a covariance matrix. More precisely, starting from a data matrix of n*p dimensions with all quantitative variables, allows the p variables (correlated) to be replaced by a new set of variables, called principal components (CP), which have the following properties:

- they are uncorrelated (orthogonal);
- are listed in decreasing order of their variance.

The first CP is the linear combination of the p starting variables having maximum variance; the second CP is the linear combination of the p variables with immediately lower variance, subject to the constraint of being orthogonal to the previous component, etc. If the p variables are strongly correlated, a number k of components ($k < p$) takes into account a high share of the total variance, so we can limit ourselves to considering only these components, refusing the remaining ones ($p - k$).

The project will discuss the customer satisfaction of an airline company, whose name is not given, in order to measure and summarize what the customers did like or not about the company's flights. The main purpose of this dataset is to predict whether a future customer would be satisfied by the services provided by the company and which aspect of the services need to be emphasized more in order to generate more satisfied customers.
We firstly start with a quick descriptive analysis regarding the variables and the data of the dataset. After that we will go more deeply with a Principal Component Analysis(PCA) that will help us to measure the satisfaction of the passengers.

1.1 - The dataset

Fig. 1.1: descriptive analysis

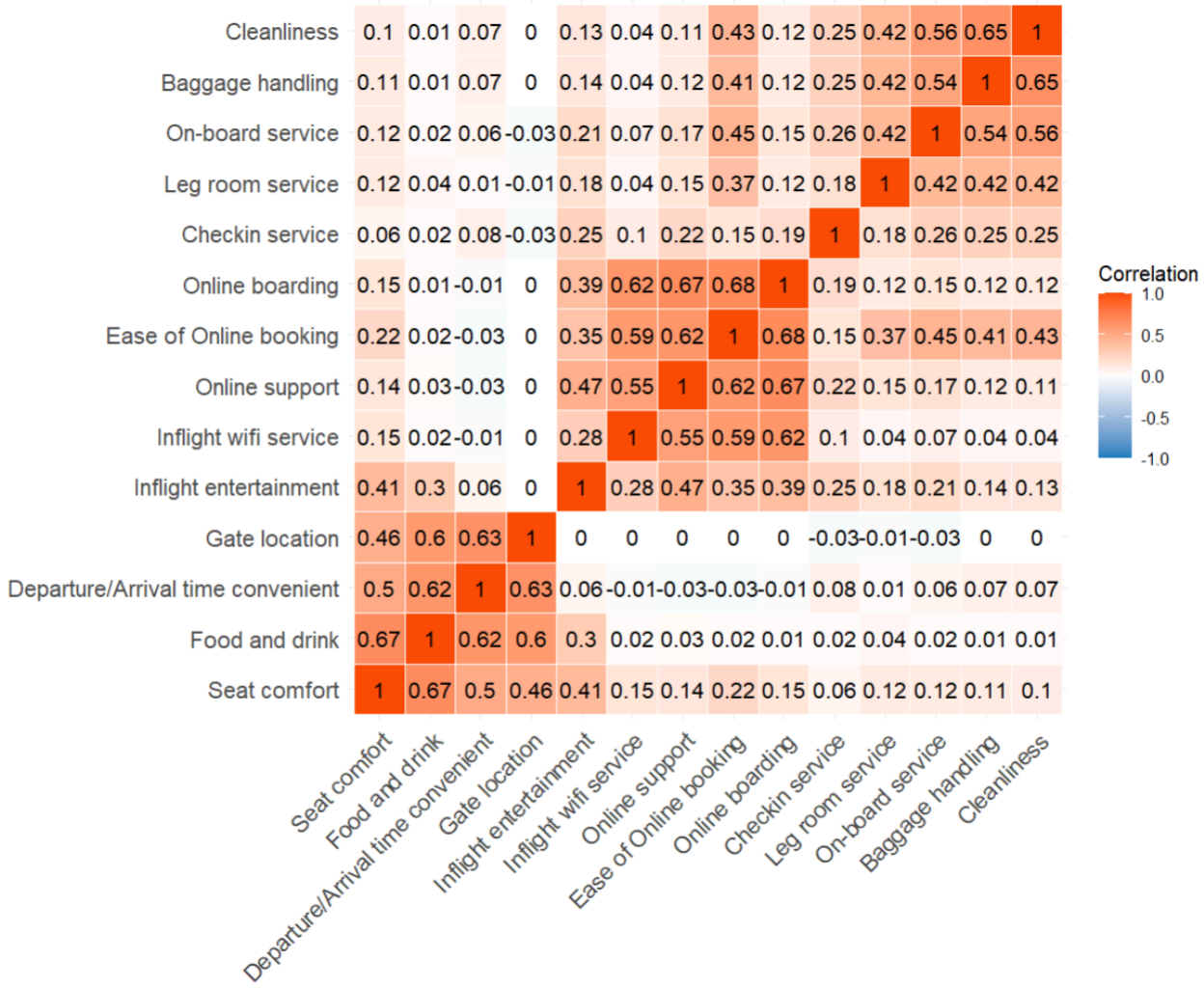
	Mean	Sigma
Seat comfort	2.95	1.30
Departure/Arrival time convenient	3.13	1.40
Food and drink	2.99	1.33
Gate location	3.00	1.31
Inflight wifi service	3.26	1.31
Inflight entertainment	3.47	1.25
Online support	3.54	1.30
Ease of Online booking	3.50	1.29
On-board service	3.48	1.27
Leg room service	3.52	1.27
Baggage handling	3.71	1.14
Checkin service	3.33	1.26
Cleanliness	3.72	1.14
Online boarding	3.37	1.29

The dataset is composed of 129.880 observations and 23 variables, which can be divided into three subgroups. One subgroup refers to the personal data of the passengers, like gender and age and the second subgroup regards mainly the characteristics of the flight taken by the passenger. For example we have data about the flight distance, the travel class in the plane(business, eco, eco plus), the purpose of the flight(personal or business travel) and information about the loyalty of the passenger(loyal customer for habitual users and disloyal customer for unusual users of the company).
The last subgroup of variables refers to the scores provided by each customer about the flight experience. Indeed all the variables belonging to this subgroup are quantitative ordinal variables with a scale from 1(min) to 5(max). Since the variables analyzed are all ordinal, it is necessary to assume that the scores of the variables, in this case from 1 to 5, are interpreted on an interval scale, i.e. the equidistance between the successive response categories must be assumed. This assumption would be difficult to support if the survey had been carried out using non-numeric labels, for example 'strongly agree', 'somewhat agree' etc. In this case the equidistance between intervals could not be considered and we would have had to use the mode or median rather than the mean to summarize our data.
Now we can finally start by adjusting our dataset and delete all the missing values. After this first step our dataset is composed of 119.255 units, so we have nearly lost 10.000 observations. Anyway, since the dataset is very large we can continue our experiment and finally get to work. As we can see in table 1.1 the mean and the standard deviation of each variable is displayed. So we have a first overview of what the passengers think about the company. Indeed, we have a general positive view of the company since the mean is alway greater than 2.5 and we have very positive scores in ‘Cleanliness’, ‘Baggage handling’ and in ‘Online support’. So we have found the strengths of the company as well as the weaknesses. What we would like to do next is study the correlation between variables, so we can understand if there is a linkage between a variable and another.

For example, between ‘Food and drink’ and ‘Seat comfort’ there is a high correlation so the two variables will probably assume the same behavior. Another example can be seen between ‘Online boarding’ and ‘Ease of Online Booking’. In this case we would expect a high correlation because the two variables measure nearly the same aspect, so we can suppose that if

passengers express a positive feeling for the former variable they will do the same with the latter. By looking at the correlation plot we can see that there are three subgroups of variables that are highly correlated. The first is in the bottom left square and includes the variables ranging from 'Seat Comfort' to 'Gate Location'. The second is located in the center and we can consider the variables ranging from ‘Inflight entertainment’ to ‘Online boarding’. Finally the last subgroup can be found in the upper right part of the graph and the variables to be considered range from ‘Checkin service’ to ‘Cleanliness’.

Fig. 1.2: correlation plot and correlation between variables



So considering all the 14 variables would lead to a low quality PCA since not all the variables are correlated together. Anyway we can start by creating a first model considering all the variables in order to begin our analysis and have a first overview of the data. After that we can proceed by removing some variables from the model and consider the ones that are more statistically and economically relevant and significant. Starting from a principal components analysis of the 14 variables, we can generate 14 principal components and see that the first component is the most important since it explains the highest amount of variability in the data,

followed by the second component till the 14th one. The variance explained by each component decreases as the number of components decreases, so we would like to find out the best number of components that could summarize the 14 variables.

1.2 - The choice of the number of components

Since all the variables taken into consideration have the same scale we could start from both the covariance matrix and the correlation matrix. However, since the correlation method is the most frequently used, our analyzes will start by exploiting the correlation matrix. Now we need to decide how many components to use in order to best summarize our dataset. There are mainly three methods that could be used:

1. A number of components is considered such that they take into account a sufficiently high percentage of the total variance. As p increases, the total variance increases and therefore it may be reasonable to consider a smaller percentage of explained variance. It can be requested that the extracted components take into account on average at least 95% of the variance of each of the p starting variables, which implies setting the minimum threshold equal to the following expression: $0.95^p * 100$, which is equal to 77.38% if p = 5; to 59.87% if p = 10. In our case, where we are studying 14 variables, we would like the principal components to take into account on average 48,77% of the variance of each of the 14 variables.

Fig. 1.3: eigenvalues and variance explained

Eigenvelues	Variance(%)	Cumulative variance(%)
3.985	28.463	28.463
2.697	19.263	47.726
2.110	15.069	62.795
0.981	7.006	69.801
0.841	6.008	75.809
0.636	4.546	80.355
0.480	3.425	83.781
0.460	3.288	87.069
0.358	2.557	89.626
0.354	2.528	92.154
0.342	2.444	94.597
0.301	2.153	96.750
0.269	1.918	98.668
0.186	1.332	100.000

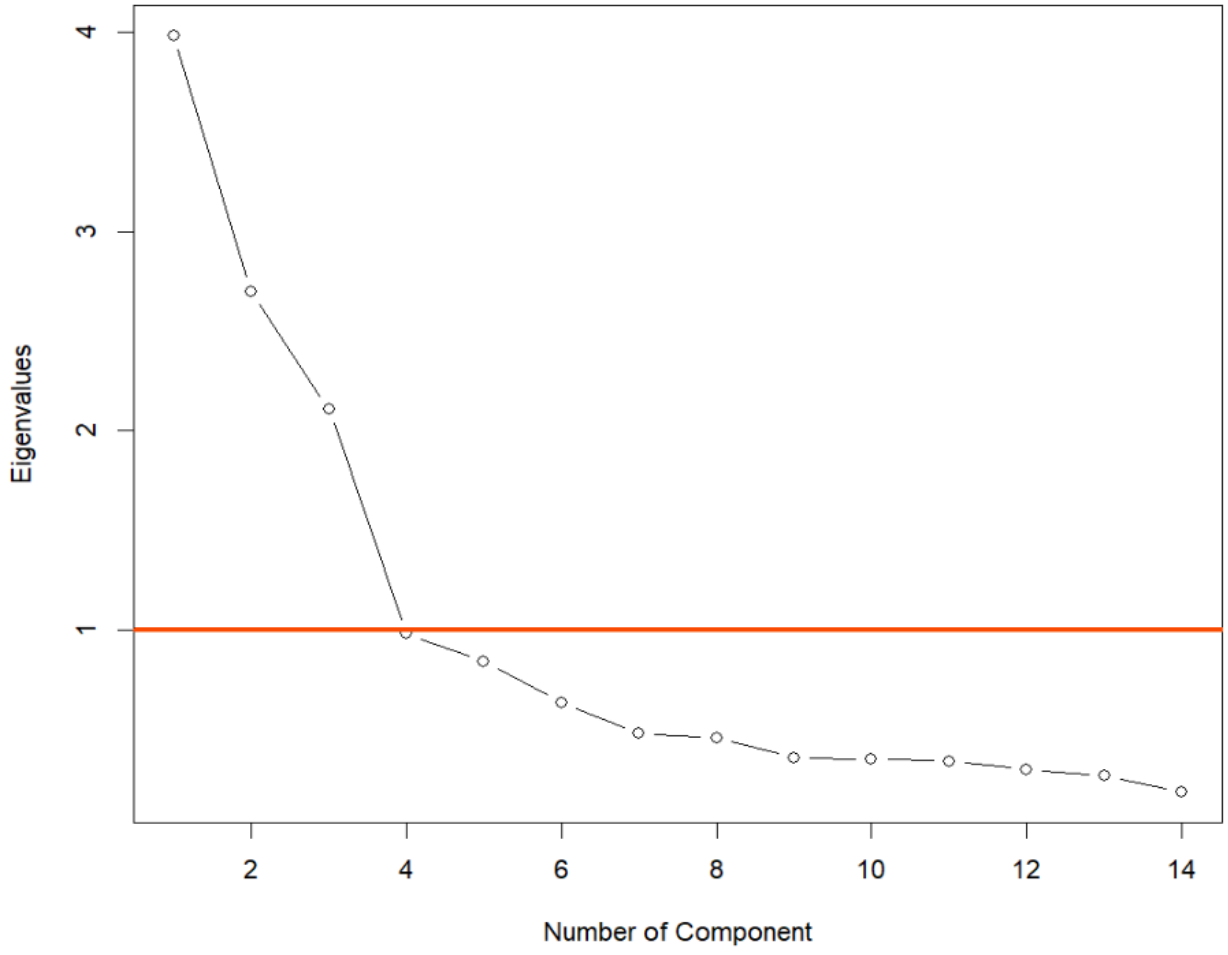
2. All components whose eigenvalue is greater than 1 are preserved. The ratio of this method derives from the fact that the eigenvalue of a component is equal to its variance and that by operating on standardized variables these have unit variance. Therefore, we decide to maintain a component (which is a linear combination of the p variables) only if it explains a greater share of the total variance than the variance explained by a single variable.

3. We can construct a scree plot of the eigenvalues as a function of ‘v’ components (v = 1, 2, ..., p). A Scree Plot is a simple line segment plot that shows the eigenvalues for each individual component. It shows the eigenvalues on the y-axis and the number of factors on the x-axis. If k components are important and the remaining (p - k) insignificant, between k and k + 1 there is a sharp variation in the slope (an "elbow"),

which signals that k is the appropriate number of components to conserve. Basically, the scree plot criterion looks for the “elbow” in the curve and selects all components just before the line flattens out.

Generally it is the joint use of the three criteria that allows identifying the number k of components to use. Starting with the first method, since we have 14 variables to consider, the number of components must be such as to explain at least 48.77% of the variance of the variables. According to this method, the best number of components to use is three, with a cumulative variance of 62.80%. Of course the power of the PCA will be limited since the remaining part of the variance, equal to 37.2%, is not explained but considering the high number of variables taken into account we can accept this limitation.

Fig. 1.4: scree plot for the choice of the number of components



Considering the second method explained before, we should consider only the components whose eigenvalue is greater than 1, since they explain more variance than a single original variable. So, by looking at table 1.3 just the first three components must be considered, even though the fourth one is very close to one and could be also considered in the PCA. Finally, considering the scree plot in figure 1.4 the first four components should be considered since we

observe a significant change in the slope at k equal 4. This point indicates that adding more components does not significantly increase the variance explained and therefore we can stop. So overall, the best number of components to be considered is 3.

1.3 - Principal Components Analysis

We can finally start with the principal component analysis, that will consider three principal components and will help us to summarize the information available. In the component matrix, shown in figure 1.5, we can find the correlation coefficients between each variable and each of the three main components. The sign of these coefficients indicates the type of linear relationship, direct or inverse, between the component and the variable to which they refer, while the numerical value, in module, indicates the entity of the link.

Fig. 1.5: first three components and communality

	Component 1	Component 2	Component 3	Communality
Seat comfort	0.446	-0.689	-0.050	0.676
Departure/Arrival time convenient	0.203	-0.796	0.118	0.689
Food and drink	0.268	-0.846	-0.011	0.788
Gate location	0.153	-0.789	0.023	0.646
Inflight wifi service	0.557	0.146	-0.566	0.652
Inflight entertainment	0.593	-0.112	-0.242	0.423
Online support	0.675	0.173	-0.485	0.721
Ease of Online booking	0.837	0.234	-0.146	0.777
On-board service	0.588	0.157	0.515	0.636
Leg room service	0.493	0.129	0.441	0.454
Baggage handling	0.559	0.152	0.592	0.686
Checkin service	0.394	0.085	0.162	0.189
Cleanliness	0.561	0.158	0.601	0.701
Online boarding	0.678	0.185	-0.510	0.754

Starting with the first component we can see that each variable is positively correlated with the component, especially for the variables ‘Online support’, ‘Ease of online booking’ and ‘Online boarding’, so all variables that express the opinion of the customer with the online services are highly and positively correlated with the first component. Instead, the variables that are highly correlated with the second component are ‘Seat comfort’, ‘Departure/Arrival time convenient’,

‘Food and drink’ and ‘Gate location’, so all variables that are linked more with the flight experience.

Most likely the first component expresses the preferences of passengers who travel often and for short trips, for whom the speed and ease of booking is fundamental. On the other hand, the variables that are highly correlated with the second component are linked to customers who make long journeys, for which the comfort of the seat is relevant, as is the food and drink offered during the journey. It also comes to mind that long-distance trips are also more expensive, therefore they refer to people who are perhaps richer and more demanding since the two important variables linked with the second component regards the location of the gate and the convenience of the departure and arrival time. These last variables therefore correspond to another level of customer satisfaction which cannot be adequately captured by the first component. It is important to remember that the two main components are uncorrelated and orthogonal to each other, therefore customers satisfied with the services offered online by the company are little interested in premium services such as the comfort of the seat or the location of the gate.

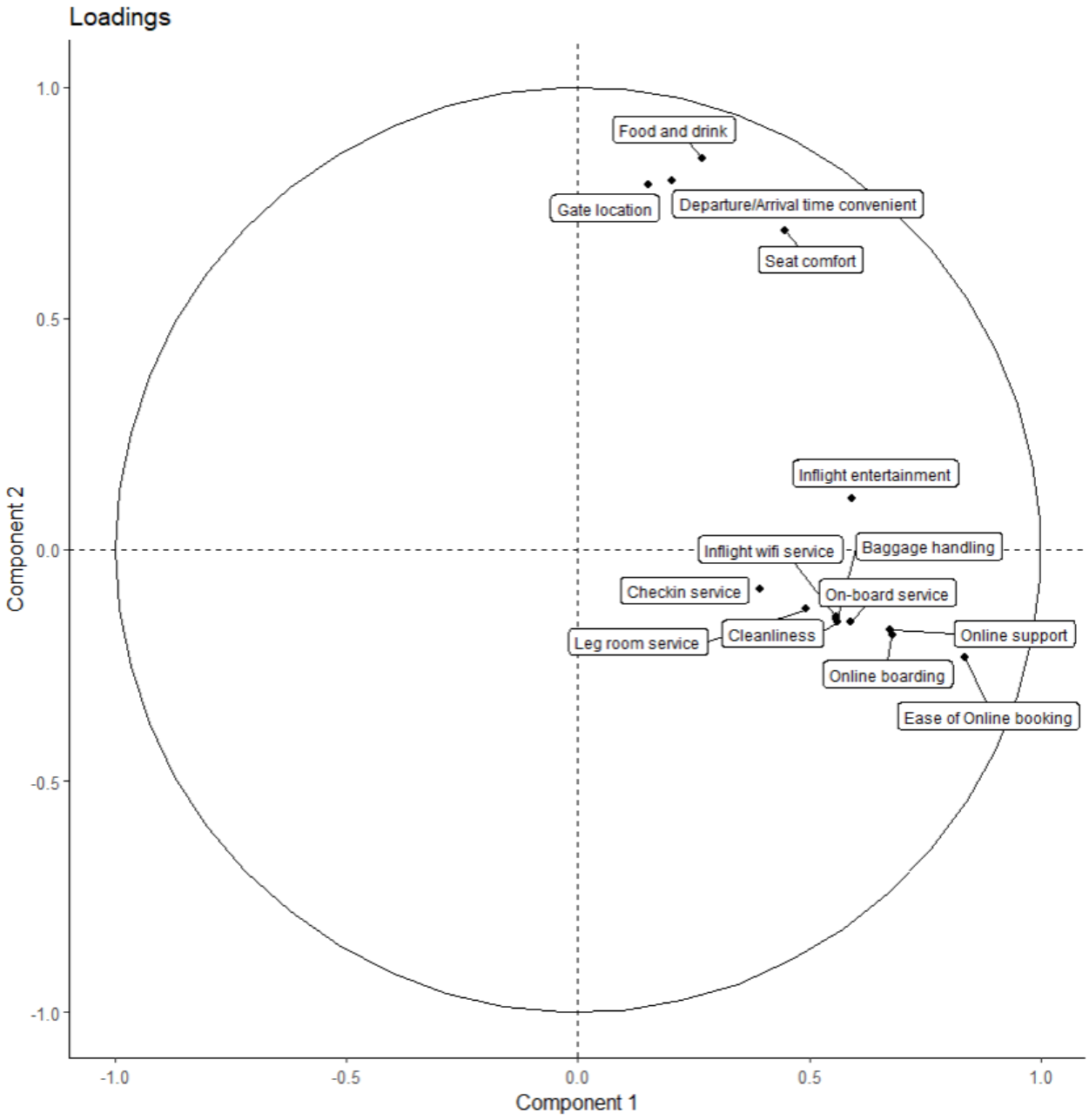
To conclude, analyzing the last component we find that the most correlated variables are 'Cleanliness', 'Baggage handling' and 'On-board service'. These characteristics could be associated with thoughtful passengers, who require high standards of cleanliness and attention to their luggage. In any case, this is the component that explains less variance than the first two, and we can see how the variables that are most correlated with this component are equally correlated with the first one. Therefore, although this component is useful for our analyses, the first two components are the most fundamental, also from an economic point of view. In fact, it is clear that customers associated with the first component have opposite interests with respect to customers associated with the second component.

Finally, we find the Communality column, which highlights how well the starting variables are explained by the components of the model. The 14 variables are not reproduced perfectly in the three-dimensional space as the Communality never exceeds 80% but the values are still high except for ‘Checkin service’ that is barely explained. Anyway the check in service depends on the airport you are in, therefore it could not depend too much on the company. All the results just obtained can be seen in the loadings plot in figure 1.6 that offers a visualization of the information previously described. However only the first two components are shown and only 47,72% of the total variance is explained. This graph always shows the correlation coefficients between each of the variables and the first two principal components. The correlation is used as the coordinates of the variable on the loadings plot. As we can see, since the correlation coefficients are between -1 and 1, all the loadings are internal to the circumference with unit radius. Furthermore, the length of each vector coincides with the share of variance of that variable explained overall by the first two components. Therefore, if the vector corresponding to a variable comes very close to the circumference of unit radius, this indicates that the variable in question is reproduced almost perfectly in the plane of the first two components.

The angle between each vector and each of the two axes signals the correlation between the variable under examination and the first and second component respectively, since the

correlation coefficient can be interpreted as the cosine of the angle between the vectors of the standardized deviations. If the angle is very small, the cosine in magnitude approaches 1, and therefore the variable is strongly correlated, directly or inversely depending on the direction of the vector, with the corresponding CP. If the angle is close to 90° the variable and the CP are almost uncorrelated (orthogonal). The angle between two vectors, corresponding to two generic variables, signals the correlation between them, with an interpretation similar to that of the previous point: if the angle is very small the correlation is high and direct; if the angle approaches 90° the correlation approaches 0; if the angle is close to 180° the correlation is high, but inverse.

Fig. 1.6: loadings plot



So all the results obtained previously are shown in the loading plot as well. For example we can see that the variables that are best explained by the first component are ‘Ease of online booking’, ‘Online support’ and ‘Online boarding, since the angle between the variables and the horizontal axis is small. Also the variables in question are reproduced almost perfectly since the vectors or points representing the variables are close to the diameter of the circumference. Regarding the second component we can say that ‘Seat comfort’, ‘Departure/Arrival time convenient’, ‘Food and drink’ and ‘Gate location’ are the variables best explained and we can clearly see the two main clusters or types of passengers we described before. The third group instead is not identified since it is associated with the third component.

If the biplot was completed with the scores associated with each customer, in the last quadrant (bottom right) we would find customers who potentially prefer short and quick trips, for which the speed and ease of booking the flight is essential. In the first quadrant(top right) we would find the customers who are most interested in the comfort and services associated with the flight, such as the food provided by the company. Thanks to these analyses, we can therefore evaluate, using the correlation coefficients between the variables and components, which aspects of satisfaction most influence the customer's overall opinion. In this case, the aspects that most influence a category of customers are linked to online services, so the company could deepen its research and understand how it can improve these services. For example, the site could be renewed and made even simpler and more usable, or the levels of online support could be increased with the use of telephone operators. As regards the more premium aspects, such as the food served during the flight or the comfort of the seat, the company could make its flights more attractive to this category of customers by providing greater customization of the menu, providing the customer with different selections of products or renovate the seats with more comfortable ones, in order to increase luxuriousness.

2 - Clustering

Having completed the analysis relating to the main aspects of our airline company, we proceed with a market analysis. The analysis in question will be carried out using a dataset composed of 23.171 reviews relating to different airlines. Our goal will be to use these reviews to summarize the composition of the airline market and identify which companies are the most popular and why. In fact, each review is related to an evaluation regarding the comfort of the seat, the services offered by the on-board staff, relating to the 'cabin' variable, the food offered, the quality of the services offered on the ground, i.e. before departure, the quality of the entertainment and wifi and finally the quality-price ratio evaluated as last variable.

Fig. 2.1: descriptive analysis

	Mean	Sigma
seat	2.39	0.77
cabin	2.61	0.79
food	2.24	0.79
ground	2.11	0.85
entertainment	2.27	0.87
wifi	1.94	0.80
value	2.12	0.83

The variables are classified on a scale of values from 1 to 5 and are very similar to the variables used in the previous dataset, therefore it will be possible to compare our airline with others on the market. Also in this case the variables available are ordinal quantitative variables and it is assumed that the intervals between the various options are equidistant, for example the distance between 1 and 2 is the same between 4 and 5. We then proceed by grouping each review based on the airline company to which it refers and calculating the average of the variables previously described so as to have a unique value associated with each company. Since some companies are reviewed more than others, we will only consider for our analysis those companies that have more than 25 reviews, so as to be able to make our analyzes less distorted and consider the most important companies. In total, therefore, 48 airlines operating all over the world will be considered and through the clustering and k-means method we will group the airlines on the basis of common characteristics.

2.1 - Expectation

Before starting with statistical analyzes we focus on airline grouping expectations. Our sample includes both national and international airlines, companies that operate globally and national airlines that operate only locally. We must then consider the distinction between low cost airlines and the more luxurious and prestigious ones. Therefore we expect that the more luxurious airlines will be grouped together and will have very positive ratings, on the contrary the smaller companies, which only operate for short and domestic routes will have more negative ratings, as the aspects related to the comfort of the seat, of the food offered, entertainment etc. will not be as important as for international airlines that operate long journeys and for which a passenger's comfort is a main consideration.

2.2 - k-means

This algorithm leads to a classification of the statistical units into g distinct groups, with g fixed a priori, through an iterative procedure consisting of the following steps.

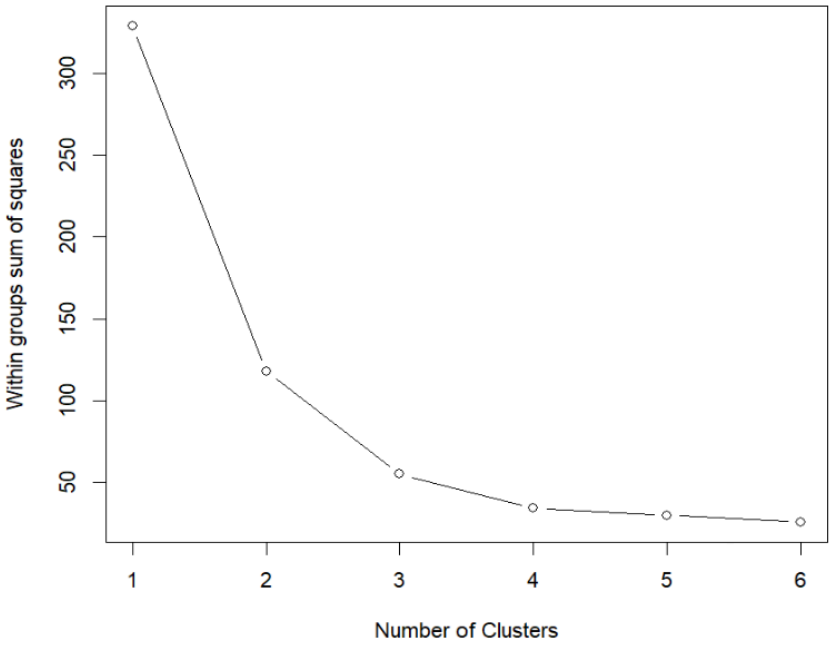
1. We choose g initial "poles" (also called seeds, or origin points), i.e. g points in the p-dimensional space that constitute the centroids of the clusters in the initial partition. The poles can be identified through different criteria, generally such that they are quite distant from each other. The initial partition is then constructed, consisting of g groups, allocating each unit to the cluster whose pole is the closest.
2. For each element, the distance from the centroids of the g groups is calculated: if the minimum distance is not obtained in correspondence with the centroid of the group it belongs to, the unit is reassigned to the cluster corresponding to the closest centroid. In case of reallocation of a unit, the centroid of both the new and the old group to which it belongs is recalculated.
3. Step 2 is repeated until the convergence of the algorithm is reached, that is, until no modification of the poles - and therefore of the groups - occurs compared to the previous iteration. Alternatively, if you want to reduce calculation times, the stopping condition provided for in point 3 of the procedure can be replaced by a less restrictive rule, which provides for the interruption of the procedure in each of the following cases:
    a. convergence of the algorithm;
    b. distance between each centroid calculated in the current iteration and the corresponding centroid in the previous iteration not exceeding a predetermined threshold;
    c. achievement of the chosen maximum number of iterations.

To obtain a partition with a different number of groups, for example g* it is necessary to repeat all the steps of the procedure, starting from phase 1 and replacing g with g*. The subsequent phases of the illustrated methodology require the repeated calculation of the distance between each point and the centroids of the g groups: this procedure therefore belongs to the class of classification algorithms that adopt the technique called "nearest centroid sorting". The metric used in calculating these distances is usually the Euclidean one, as it guarantees the convergence of the iterative procedure. The k-means method - with the use of the Euclidean distance - has as its implicit objective the search for the partition (with g clusters) that satisfies an internal cohesion criterion based on the deviance in the groups, i.e. on the minimization of the internal variance of the group.

We therefore proceed with our analyzes by choosing the initial number of groups to create. This aspect can be analyzed with two distinct techniques, the first relating to the wss plot (within sum of squares plot) and the second relating to the silhouette plot. Within-Cluster Sum of Squares (WSS) is a measure of how far away each centroid is from their respective instances or points. The larger the WSS, the more dispersed the cluster values are from the centroid. The objective of this metric is to find the “elbow” of the WSS curve in order to determine the smallest number of clusters that captures the most amount of signal in our data. As we can see, the optimal

number of clusters to choose is 3, in which there is a clear decrease in the within variance and we can notice the so-called elbow of the wss plot.

Fig. 2.2: wws plot



In any case we can rely on the second method, called the silhouette method. The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation) in a range of [-1, 1]. Silhouette coefficients near 1 indicate that the sample is far away from the neighboring clusters, so a score of 1 denotes that the data point is very compact within the cluster to which it belongs and far away from the other clusters. A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters and we could have two or more clusters overlapping. Finally, negative values indicate that those samples might have been assigned to the wrong cluster. The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \; if \left|C_i\right| > 1$$

s(i) is defined to be equal to zero if i is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters. Here, a(i) is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.

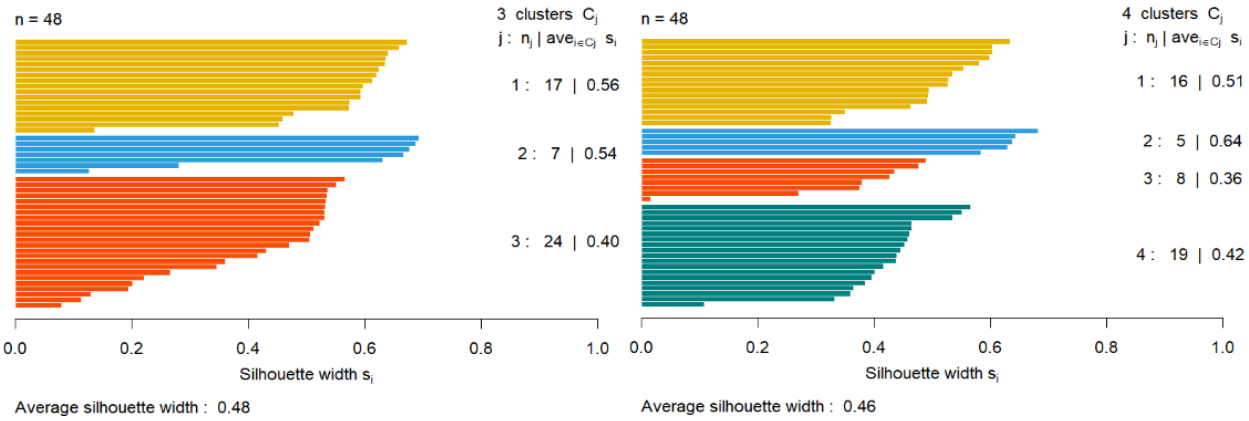
$$a(i) = \frac{1}{\left|C_i\right| - 1} \sum_{j \in C_i, \, i \neq j} d(i, j)$$

Similarly, b(i) is the measure of dissimilarity of i from points in other clusters. b(i) is the minimum average distance from i to all clusters to which i does not belong. d(i, j) is the distance between points i and j. Generally, Euclidean Distance is used as the distance metric.

$$b(i) = min_{i \neq j} \frac{1}{|C_i|} \sum_{j \in C_{i'}} d(i, j)$$

Therefore, looking at the graph of the silhouette that considers 3 clusters and comparing it with the graph of the silhouette that considers 4 clusters we notice how the silhouette value decreases. This explains why we can consider 3 clusters as the best option for our analyses

Fig. 2.3: comparison between silhouette plots with 3 and 4 clusters

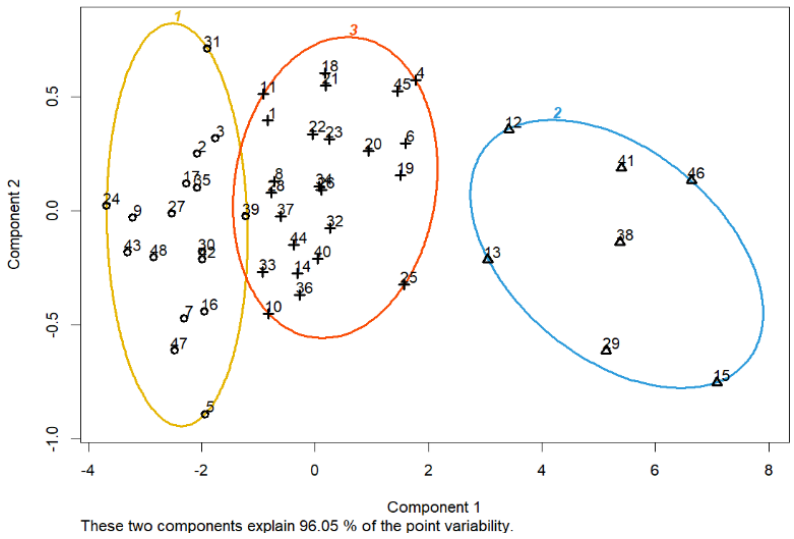


We therefore have three clusters in which the majority of observations fall into the first and second groups, with 17 and 24 observations respectively. Only 7 airlines fall into the second cluster. In order to have a clearer idea of what separates one cluster from another, we can carry out summary statistics to identify what separates one cluster from another. In this case we note that the first cluster includes the airlines rated negatively by users, or in general the airlines among those considered which have the lowest ratings. On the contrary, in the second cluster, composed of only 7 airlines, the values of the variables are on average higher than in the remaining two clusters, indicating how the companies most appreciated by passengers are represented. Finally, in the last cluster we find the averagely appreciated companies, i.e. with average values intermediate to the first and second clusters. This results can be seen in the figure 2.5 displayed down below

Fig. 2.4: means of every variables in the cluster

Groups	Seat	Cabin	Food	Ground	Inflight	Wifi	Value
1	1.706549	1.893014	1.544732	1.450995	1.454501	1.250741	1.431598
2	3.853852	4.141050	3.755335	3.851894	3.854037	3.443604	3.775636
3	2.452862	2.676612	2.290893	2.063925	2.383343	1.997863	2.115294

Fig. 2.5: graphical representation of the airlines companies



1	Aer Lingus
2	Aeromexico
3	Air Canada
4	Air France
5	Air India
6	Air New Zealand
7	Air Serbia
8	Alaska Airlines
9	Avianca
10	Breeze Airways
11	British Airways
12	Cathay Pacific Airways
13	China Airlines
14	China Eastern Airlines
15	China Southern Airlines
16	Condor Airlines
17	Copa Airlines
18	Delta Air Lines
19	El Al Israel Airlines
20	Emirates
21	Etihad Airways
22	Finnair
23	French Bee
24	Go First

24	Go First
25	Gulf Air
26	Hawaiian Airlines
27	ITA Airways
28	Icelandair
29	Japan Airlines
30	Jet Airways
31	Jetblue Airways
32	KLM Royal Dutch Airlines
33	Kenya Airways
34	Kuwait Airways
35	LOT Polish Airlines
36	Philippine Airlines
37	Qantas Airways
38	Qatar Airways
39	Royal Jordanian Airlines
40	Saudi Arabian Airlines
41	Singapore Airlines
42	SpiceJet
43	Spirit Airlines
44	SriLankan Airlines
45	Turkish Airlines
46	Vistara
47	WOW air
48	flydubai

We can therefore proceed with the graphic representation of the 3 clusters obtained. The blue cluster represents the airlines evaluated very positively while the yellow cluster represents the companies evaluated very negatively. The intermediate cluster is represented by the color orange, which includes companies rated quite positively. From the graph we can see how cluster 1 and cluster 3 overlap slightly and this explains why cluster 3 presents the lowest silhouette value among all three clusters. Among the airlines rated positively by passengers we can see that the majority are oriental companies unlike Vistara which is based in India and Qatar Airways which is based in Qatar. These are very large airlines that operate on national and above all international routes for which the comfort of the seat, the cabin, the food served etc. are fundamental variables for a passenger, which is why they could be evaluated more positively than low cost companies that operate only on national or local routes. Among these, we note that China Southern Airlines is the most positively rated, with average values very close to 5, i.e. the maximum rating, especially in the case of the 'Ground' variable, relating precisely to the services provided on the ground by the company. On the contrary, the companies 'Cathay Pacific Airways' and 'China Airlines', numbers 12 and 13 respectively in figure 2.6, present the lowest values and can be found in the middle of cluster 3 and cluster 2, almost as if they were outliers within the latter cluster. In cluster 1 as well as in cluster 3 we find companies such as Air France, British Airways and Emirates and in general national airlines that operate mostly within national territories.

In this case we could compare the results obtained from our clustering analyzes with the results obtained from the previous dataset in which the single airline was analyzed. The two datasets are not identical but they use similar variables and with the same scale of values(1-5) so some

variables coincide and can be compared. For example, in the first dataset that was used for the PCA and relating to the single airline we can take into consideration the variables 'seat', 'food and drink', 'inflight entertainment', 'inflight wifi service' which are considered to be practically the same as in the second dataset relating to clustering.

Fig. 2.6: airlines companies in the second cluster

Airline Name	Seat	Cabin	Food	Ground	Inflight	Wifi	Value	Cluster
Cathay Pacific Airways	3.333333	3.363636	3.272727	3.393939	3.757576	2.939394	2.939394	2
China Airlines	3.266667	3.533333	3.100000	3.000000	3.100000	2.833333	3.333333	2
China Southern Airlines	4.520000	4.680000	4.280000	4.920000	4.360000	3.560000	4.480000	2
Japan Airlines	3.944444	4.138889	3.750000	4.000000	3.555556	3.277778	3.944444	2
Qatar Airways	3.784314	4.450980	4.000000	3.764706	4.000000	3.392157	3.745098	2
Singapore Airlines	3.820513	4.358974	3.615385	3.846154	3.974359	3.871795	3.717949	2
Vistara	4.307692	4.461538	4.269231	4.038462	4.230769	4.230769	4.269231	2

Some variables, however, can be assimilated to each other, for example the variable 'on-board service', which was evaluated in the first dataset, can be assimilated to the variable 'cabin staff service', which was instead considered in the second dataset. Similarly the 'baggage handling' variable can be compared to the 'Ground Service' variable of the second dataset. We can then proceed with our analyzes and compare the averages obtained from the first dataset with the averages obtained from the second dataset and note how our airline company would most likely fall into the first dataset. This allows us to more clearly analyze the strengths and weaknesses of our airline and have a benchmark that allows us to compare our airline with the market. We could therefore develop a business strategy that reinforces the most negative aspects, such as the comfort of the seat.

Fig. 2.7: means of the airline company

Seat	Cabin	Food	Ground	Inflight	Wifi
2.950451	3.476751	2.987455	3.708893	3.467385	3.260836

2.3 - Hierarchical clustering

Now that we have finished our analyzes with the k-means method we can carry out a second analysis with the hierarchical clustering method. Cluster analysis consists of searching for groups of similar units in the n p-dimensional observations, without knowing a priori whether such homogeneous groups actually exist in the data set. Given n units, to which the p-dimensional vectors correspond, many cluster analysis methods require the calculation of the distance matrix (i.e. the similarity indices), which contains the "proximity" measures between all pairs of units. It is therefore necessary to choose the distance or the distance index in the case of quantitative variables.

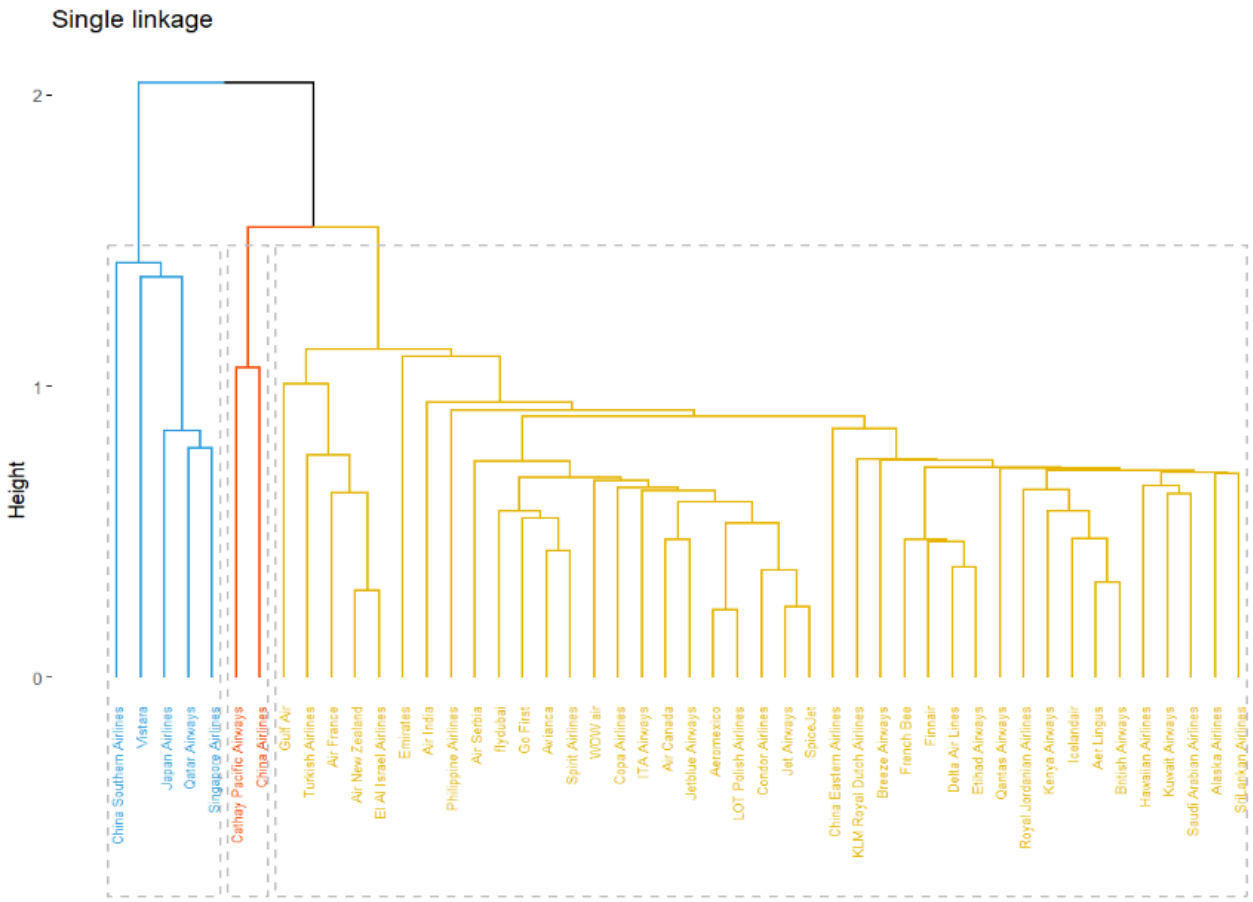
This choice conditions the results of the classification, since by varying the type of distance the ordering of the pairs of units does not generally remain unchanged (from those most similar to each other to those most different) and therefore even the groups of "homogeneous" units can differ. The objective is to classify these units into groups with the characteristics of internal cohesion (the units assigned to the same group must be similar to each other) and external separation (the groups must be as distinct as possible).

A partition is satisfactory when the variance or variability within the identified groups is small (the units of each group present modest differences between them) and furthermore the groups are well distinct from each other. It must be taken into account that there is a trade-off between the number of groups and the homogeneity within them: by reducing the number of groups a more concise classification is obtained, and therefore generally more useful for operational purposes, but a cost must be paid price in terms of greater variability in the groups, since more different units are aggregated. The partition with the optimal number of groups will therefore be the one that best manages to reconcile these opposing needs of synthesis of the units in classes and internal cohesion of the groups. Our analysis will be carried out using 4 different methods, namely the single linkage method, the complete linkage method, the average linkage method and the Ward method. Starting with the single linkage or nearest neighbor method we can define the distance between two groups as the minimum of the $n_1 n_2$ distances between each of the units of one group and each of the units of the other group:

$$d(C_{1,}\, C_2) = min(d_{rs}),\ per\ r \in C_1,\ s \in C_2$$

The partition family obtained with this method can be visualized in the dendrogram below, which divides the airlines into 3 clusters. In the graph below as well as in the next ones, the blue cluster represents the airlines evaluated very positively while the yellow cluster represents the companies evaluated very negatively. The intermediate cluster is represented by the color orange, which includes companies rated quite positively. We can see how, using this method, the observations are not optimally divided between the clusters and are more concentrated in the yellow one, which represents companies rated negatively. On the contrary, only two airlines are included in the orange cluster, which in this case represents airlines rated on average well. We therefore do not have a completely meaningful representation of the results

Fig. 2.8: dendrogram obtained with the single linkage method

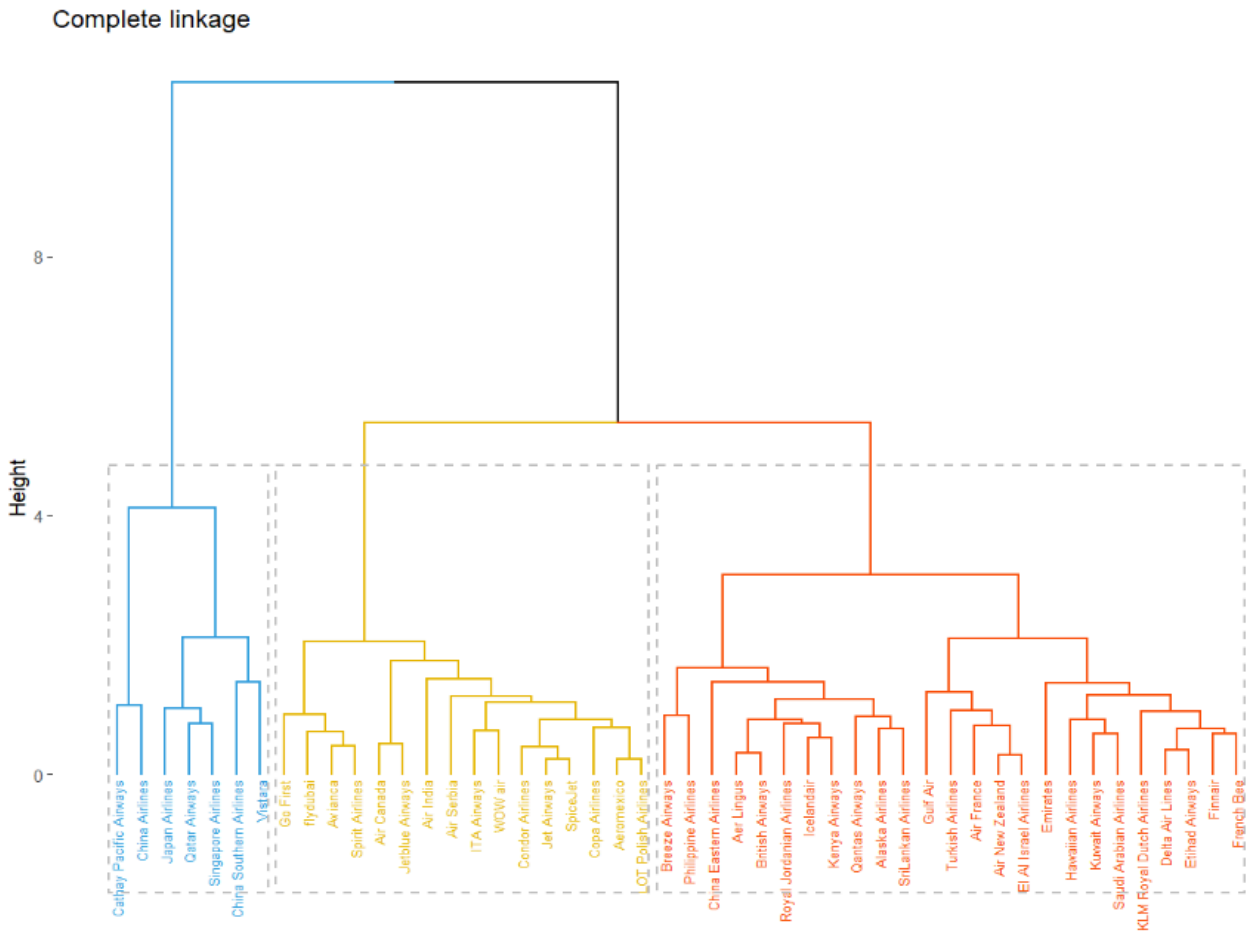


We continue by considering the complete linkage or furthest neighbor method. The distance between two groups is defined as the maximum of $n_1 n_2$ distances between each of the units of one group and each of the units of the other group:

$$d(C_{1,}\, C_2) = max(d_{rs}),\ per\ r \in C_1,\ s \in C_2$$

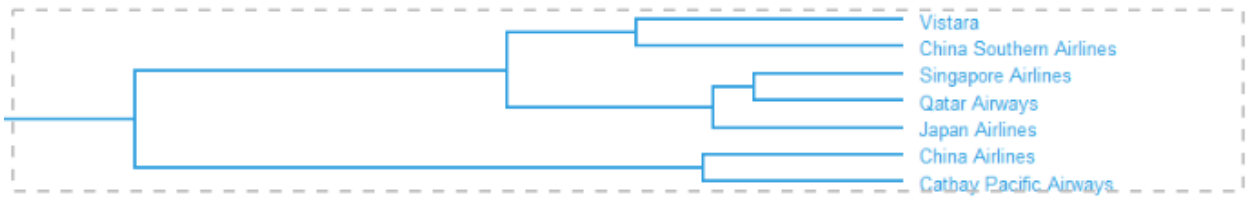
Looking at the dendrogram obtained with the complete link method we can see how the partition of the airlines is more balanced, i.e. there is a more homogeneous partition of the units. The result obtained is very similar to what we previously obtained with the k-means analysis, as the airlines appear to be divided in the same way into the clusters except for the 'Royal Jordanian Airlines' which is classified in the orange cluster and not in the yellow one, i.e. it is classified in the cluster rated on average well and is not placed in the cluster which includes the worst companies, so the yellow one. At the moment, therefore, the dendrogram obtained with the complete linkage method is better than the dendrogram obtained with the single linkage method

Fig. 2.9: dendrogram obtained with the complete linkage method



In more detail we can see how Qatar Airways and Singapore Airlines, belonging to the blue cluster which indicates the most positively evaluated airlines, are very similar to each other. At a slightly greater distance we find Japan Airlines joining that group. At a greater distance these three airlines join a second group formed by the companies China Southern Airlines and Vistara. At an even greater level of distance these 5 units join the remaining 2, formed by Cathay Pacific Airways and China Airlines and at this point all the units are united in a single cluster. These last two airlines are therefore the most different within the group and could almost be considered outliers as they join the remaining variables at very high levels of distance. In fact, if we consider the dendrogram that was obtained using the single link method, we notice how these two airlines were classified in a separate cluster

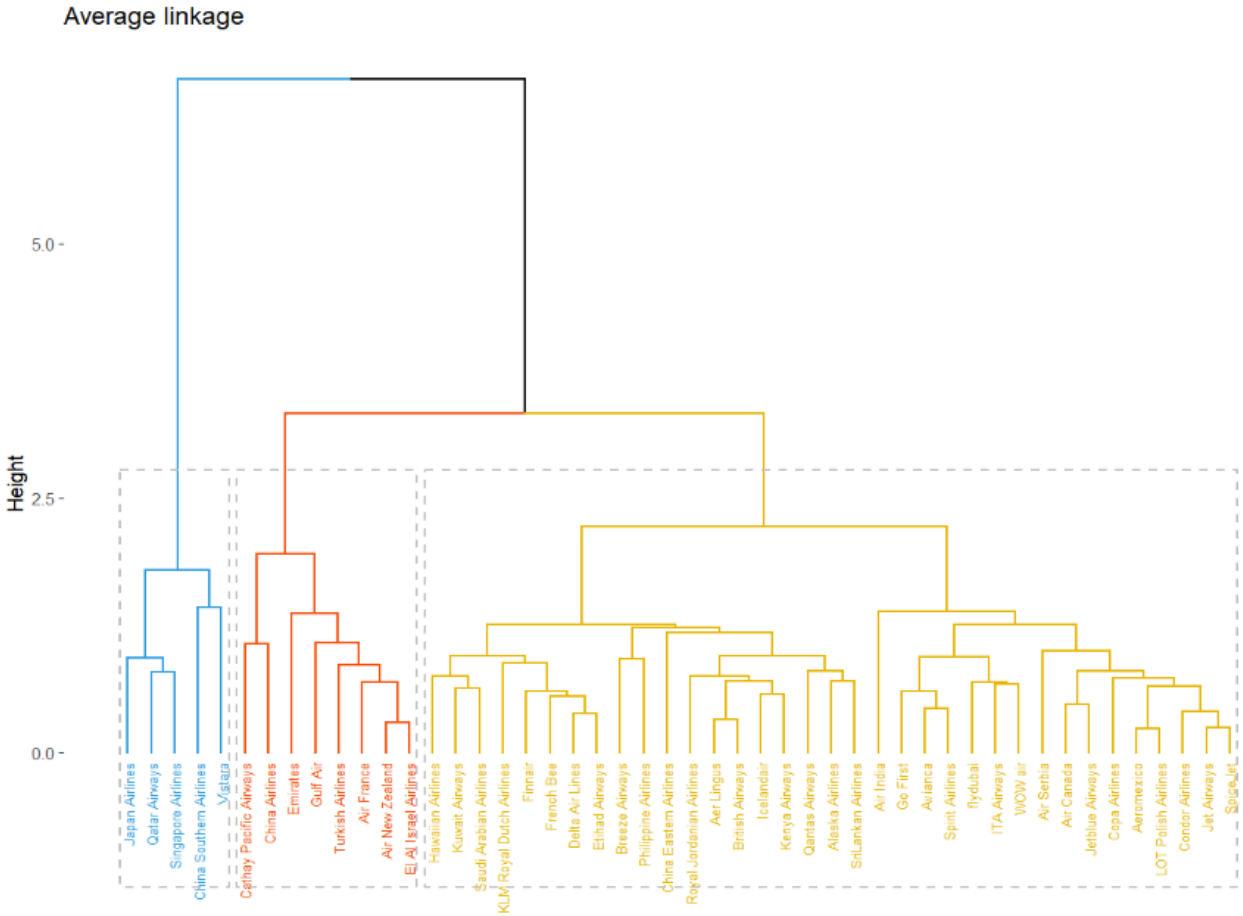
Fig. 2.10: close up of the best airlines companies



Finally we consider the method of the average link between groups. The distance between two groups is defined as the arithmetic mean of the $n_1 n_2$ distances between each of the units of one group and each of the units of the other group:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_r \sum_s d_{rs}, \; per\; r \in C_1, \; s \in C_2$$

Fig. 2.11: dendrogram obtained with the average linkage method

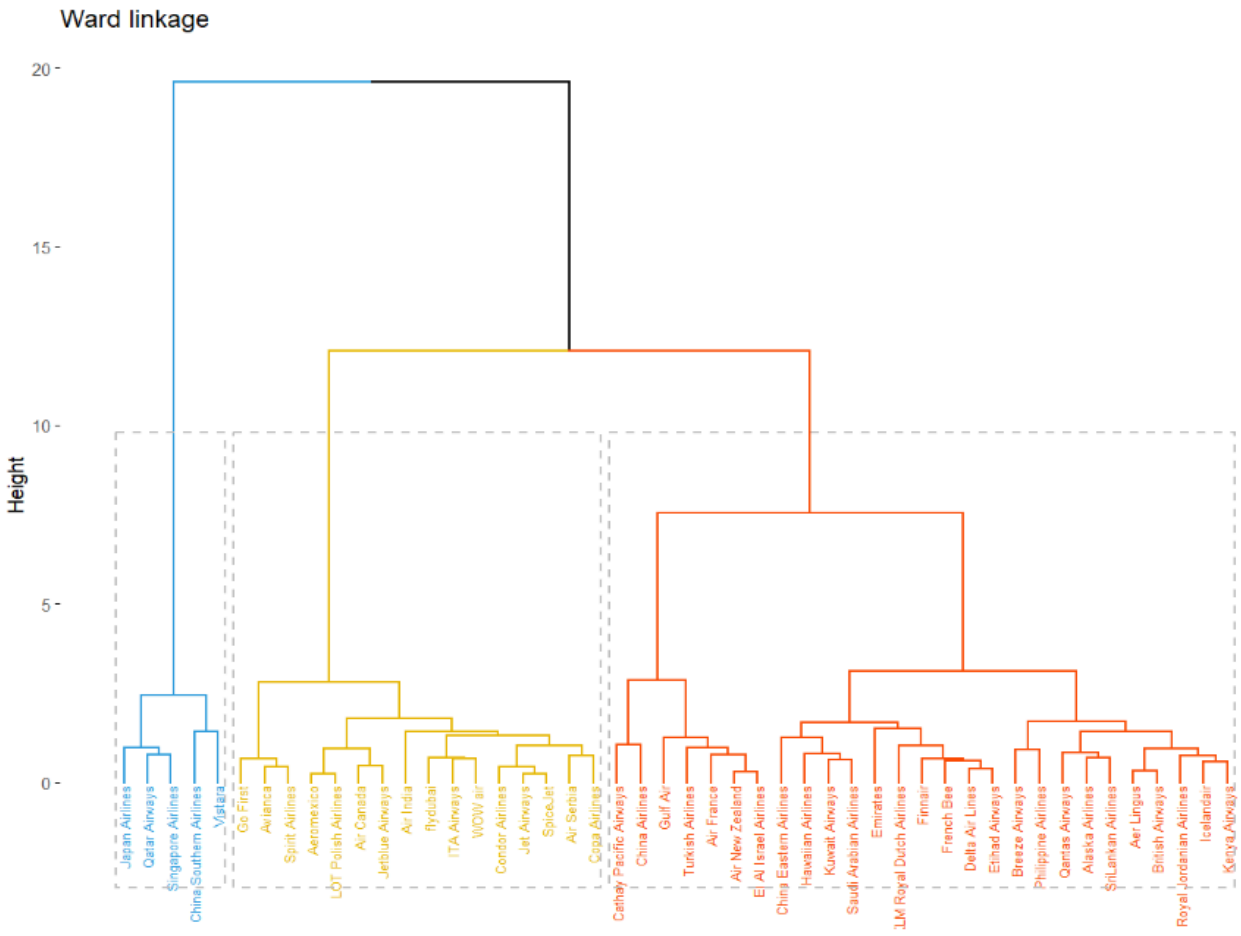


From the graph we can see how the representation of the observations is a middle ground between the dendrogram obtained with the complete linkage method and the dendrogram obtained with the single linkage method. In fact, we can see how most of the observations are concentrated in the third cluster, so the yellow one, which brings this dendrogram to the single linkage, however the first two clusters are more numerous than the first two clusters obtained with the single linkage method and this is more common to this dendrogram to that obtained with the complete linkage method.

Let's now move on to a final methodology for calculating clusters, namely the Ward method. This method differs from the first three because it uses the starting data matrix and not the

distance matrix. Since the aim of the classification is to obtain groups with the greatest internal cohesion, the decomposition of the total variance of the p variables into variance within the groups and variance between the groups is considered. At each step of the hierarchical procedure, the groups (possibly made up of a single unit) are aggregated together which cause the least increase in variance in the groups, i.e. which ensure the greatest possible internal cohesion. Therefore, we look for a partition that minimizes the within variance and maximizes the between variance

Fig. 2.12: dendrogram obtained with the Ward method



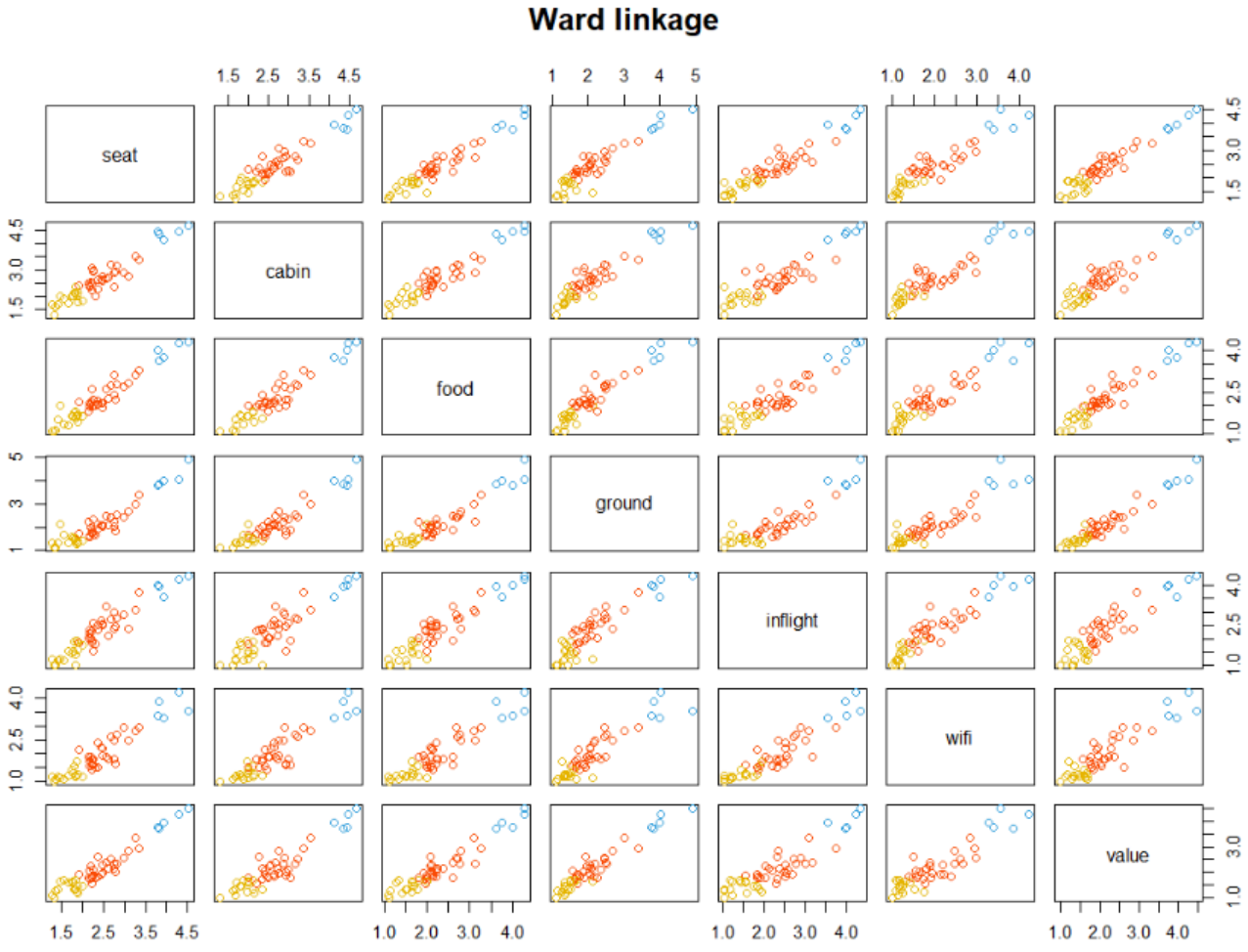
At the moment the subdivision of the clusters obtained with the Ward method is the best as the observations are homogeneously divided between the various clusters and there are no evident outliers in each cluster. If desired, one could cut the dendrogram at height 5 to obtain a further cluster which would make the partition of the observations more elaborate. In this case the orange cluster, which represents the airlines rated fairly, would be divided into two subgroups, one that is closer to the companies rated very negatively and one that is closer to the companies rated very positively.

Comparing Ward's method with the complete linkage method we note how the Ward method and the complete linkage method present very similar results except for the third cluster, i.e. the

	ward		
complete	1	2	3
1	25	0	0
2	0	16	0
3	2	0	5

one relating to airlines evaluated very positively. In fact, 2 companies, Cathay Pacific Airway and China Airlines, are classified in the first cluster by the Ward method, i.e. the one relating to airlines rated on quite well. These two companies, in fact, had previously been classified as outliers during the analyzes obtained with the complete linkage method and during the k-means analysis. Therefore, the result obtained with the Ward method is not only better than the methods that use the distance matrix but also provides better results than the k-means method

Fig. 2.13: scatterplot with Ward linkage method



Subsequently we can create a scatter plot matrix which summarizes the correlations between the variables used. In particular, each observation, i.e. each airline, is differentiated based on the cluster it belongs to. First of all we can notice a positive correlation between the variables. Furthermore, a clear division of the variables is observed on the basis of the cluster they belong to, therefore the negatively evaluated airlines will be found in the lower left part of each scatterplot and the positively evaluated airlines will occupy the upper right area

To conclude our analysis we can compare the means of the variables divided into clusters according to the Ward method and we obtain mainly the same result obtained with the k-means method. A particular difference can be seen in the cluster in which the companies rated most positively by users are present, in which it is possible to notice a remarkable increase in the average after the two airlines, Cathay Pacific Airways and China Airlines, classified as outliers, were removed from the cluster

Fig. 2.14: means of every variables in the ward cluster

Groups	Seat	Cabin	Food	Ground	Inflight	Wifi	Value
1	2.510335	2.708728	2.346450	2.132715	2.440199	2.04460	2.178996
2	1.668811	1.886327	1.516278	1.438234	1.431183	1.23624	1.409004
3	4.075393	4.418076	3.982923	4.113864	4.024137	3.66650	4.031344

Fig. 2.16: means of every standardized variables in the ward cluster

Groups	Seat	Cabin	Food	Ground	Inflight	Wifi	Value
1	0.1532446	0.1209229	0.1340465	0.02958007	0.1971512	0.1256566	0.07712717
2	-0.9444560	-0.9141593	-0.9132306	-0.78807779	-0.9638301	-0.8849946	-0.85500935
3	2.1947384	2.2723261	2.1984869	2.36211652	2.0196400	2.1534369	2.31954322

3 - Supervised analysis

Once our clustering analyzes have been completed, we can move on to supervised learning techniques. In particular we will focus on a logistic regression that we will obtain using the stepwise regression method and we will evaluate the result with the Linear Discriminant Analysis method. At the end of the chapter we will find the decision trees and the k-fold cross validation that will conclude our supervised analyses. We can therefore begin by explaining the theory behind the logistic regression model. By regression model we mean a mathematical model that links one or more variables, called independent, with a dependent variable. The purpose of the model is to estimate the conditional expected value of the dependent variable Y as the regressors X vary. In this case, the response variable Y has limited support and can only take two values, i.e. 0 when the opinion expressed about the flight is negative and 1 when the opinion is positive. Therefore, the response variable Y is distributed as a Bernoulli random variable with parameter p: $Yi \sim Be(p)$. Since we are interested in the conditional distribution of Yi with respect to Xi we will have: $Yi|Xi = x \sim Be(p(x))$ (where $p(x)$ is some function of x) with conditional expected value:

$$E[Yi\ |Xi = x] = Pr(Yi = 1|Xi = x) = p(x)$$

Therefore, the regression model that best estimates the relationship between independent variables and the dependent one is the logistics model. These model is based on a latent variable Y*, an unobserved variable whose support is the entire real axis and which comes used as a proxy for Y, the observed variable of interest. The higher the value of Y* and the higher the probability of observing $Y = 1$. Conversely, the lower the value of the latent variable, the higher the probability of observing $Y = 0$. The following function therefore results:

$$Yi = 1\ se\ Yi* > 0\ e\ Yi = 0\ se\ Yi* \leq 0 \text{ con } Y* = \beta 0 + \beta 1 Xi + \varepsilon i$$

The choice of the threshold value 0 is purely arbitrary since Yi* it is not observable and has no well-defined scale. In the above function $\beta 0$ and $\beta 1$ are our parameters of interest to estimate and εi is the error term associated with the latent variable. Starting from this function it is easy to demonstrate that:

$$E[Yi\ |Xi = x] = 1*\Pr(Yi = 1|Xi = x) + 0*\Pr(Yi = 0|Xi = x) = \Pr(Yi = 1|Xi = x) = Pr(Yi* > 0|Xi = x) = = Pr(\beta 0 + \beta 1 Xi + \varepsilon i > 0|Xi = x)$$

Therefore, the correct specification of the conditional mean of Yi depends on the error distribution in the latent model. Let us assume that the error εi is an independent variable and identically distributed(i.i.d.) with $F(\cdot)$ distribution function and $f(\cdot)$ density function symmetric around 0. Therefore:

$$E[Yi\ |Xi = x] = Pr(\varepsilon i < \beta 0 + \beta 1 x) = F(\beta 0 + \beta 1 x)$$

where $F(\beta 0 + \beta 1 x)$ represents the distribution function of the random variable (not observed) εi. The linearity of $E[Yi*|Xi]$ therefore implies the non-linearity of $E[Yi|Xi]$, since that $F(\cdot)$ is a

nonlinear function that maps values that belong to the entire real axis on values inside the interval [0, 1]. If we assume that εi is distributed according to a logistics distribution function, that is, $F(t) = \Psi(t)$, where $\Psi(\cdot)$ is the distribution function given by $\Psi(t) = \frac{e^t}{1+e^t}$ we will have the logistics model. At this point, the maximum likelihood estimator allows us to obtain non-estimations distorted for $\beta 0$ and $\beta 1$. It is important to underline how the marginal effects are not constant but vary as xi varies as the parameters β represent the marginal effects of the latent theoretical model. Therefore, the marginal effect is specific for each individual. We can exploit two strategies, that is, calculate the average of the marginal effects in the sample (AME) or report marginal effects for a representative individual (MEM). The AME (Average Marginal Effect) is given by the average on the observed distribution of Xi of the individual marginal effects and is easily calculable from the sample data. In our subsequent analyses, reference will be made to the AME whenever we talk about effects marginals of the logit models

$$\widehat{AME} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta 1} * f(\widehat{\beta 0} + \widehat{\beta 1} * X_i)$$

3.1 - Logistic regression assumption

Firstly, logistic regression does not need a linear relationship between the dependent and independent variables. Logistic regression can handle all sorts of relationships, because it applies a non-linear log transformation to the predicted odds ratio. Secondly, the independent variables do not need to be multivariate normal – although multivariate normality yields a more stable solution. Also the error terms (the residuals) do not need to be multivariate normally distributed. Thirdly, homoscedasticity is not needed. Logistic regression does not need variances to be heteroscedastic for each level of the independent variables. However some other assumptions still apply.

1. The error terms need to be independent. Logistic regression requires each observation to be independent. Also the model should have little or no multicollinearity. That is that the independent variables should be independent from each other. If multicollinearity is present centering the variables might resolve the issue, i.e. deducting the mean of each variable.
2. Logistic regression assumes linearity of independent variables and log-odds. While it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log-odds. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and absence of multicollinearity are met.
3. It requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression); whilst OLS needs 5 cases per independent variable in the analysis, ML

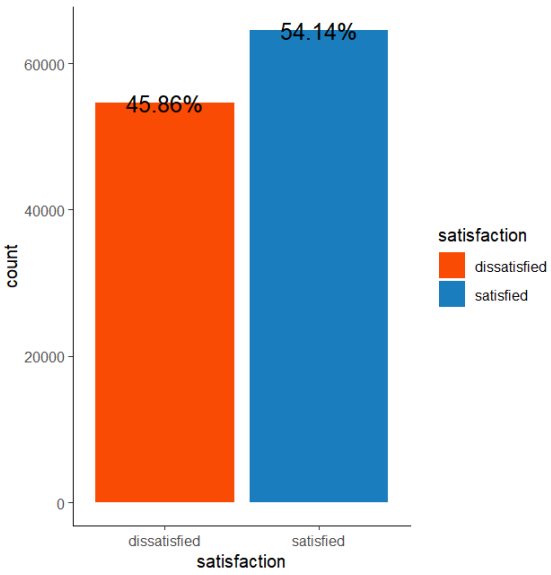
needs at least 10 cases per independent variable, some statisticians recommend at least 30 cases for each parameter to be estimated.

4. Outliers in logistic regression can skew parameter estimates, distort the relationship between variables, compromise model fit and lead to misleading interpretations. Handling outliers appropriately is crucial to ensure the validity and reliability of logistic regression analyses, which may involve removal or transformation

All these assumptions will be verified in the following paragraphs, but we can already see how the second and third have already been verified as most of the variables are categorical. Furthermore, the sample is made up of 119,255 observations, so the third hypothesis is also respected.

3.2 - Data preparation - outliers

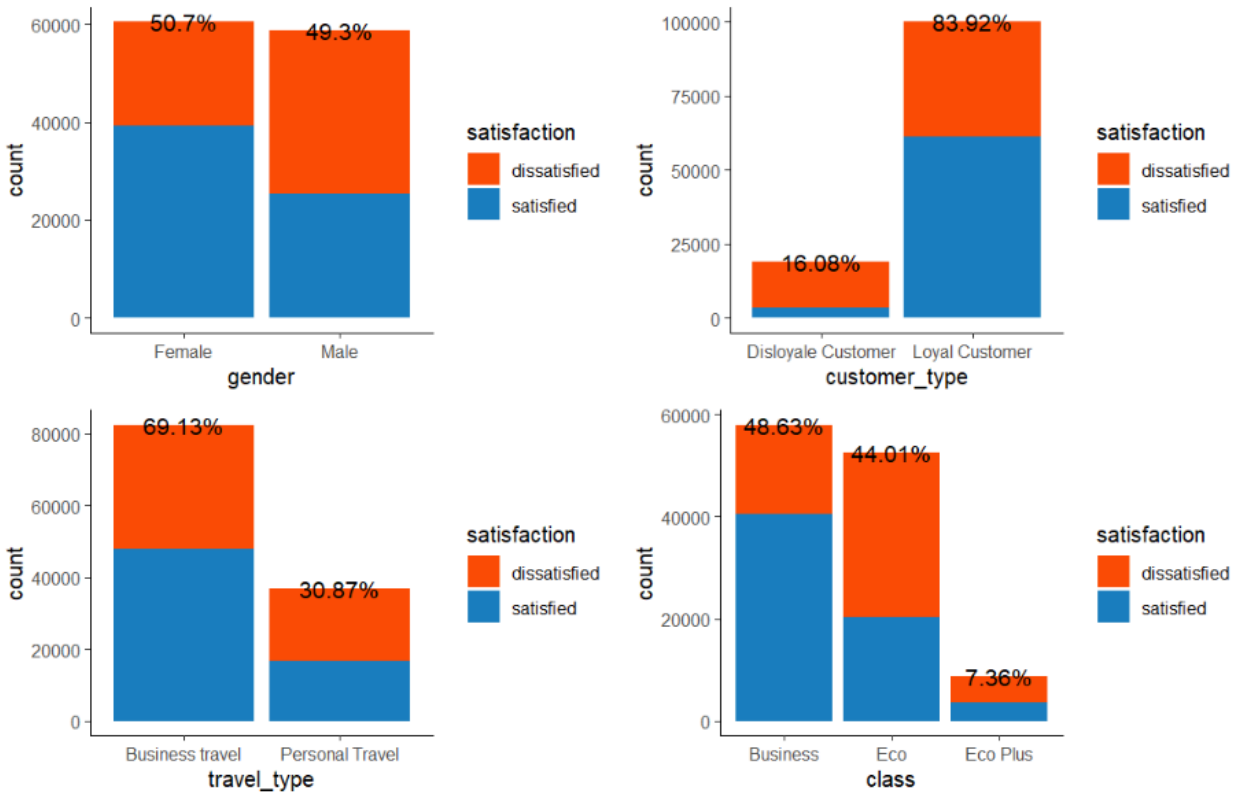
Fig. 3.1: barplot of the dependent variable satisfaction



First of all, the dataset we will use is the one used for the PCA, only that the variables taken into consideration will be different. In fact, as dependent variable of our model, i.e. Y, the dichotomous variable relating to passenger satisfaction with the airline will be considered. This variable in fact takes two values: dissatisfied and satisfied, which in the logistic model will be interpreted as 0 and 1 respectively. Regarding the independent variables, i.e. our x's, we will have a total of 6 variables of which 2 are numeric and 4 are categorical or binary. Among these we find the sex of the passenger, the type of customer, i.e. whether loyal or disloyal, the age, the type of trip made, i.e. whether personal or business, the class in which the traveler was traveled, divided into eco, eco plus and business and finally the distance traveled by flight. These three variables will determine our logistics model and we will be able to understand how these variables will influence our Y, or satisfaction at the end of the flight relating to the airline. We therefore start with descriptive statistics of the variables to get an initial idea of the sample. Starting from our response variable we can see how the majority of passengers are satisfied with the airline as the percentage of positive responses is slightly higher than negative responses, therefore there is a fairly homogeneous distribution and this is fundamental in our analyzes to have a predictive model that is not too biased towards a single response. Moving now to the remaining categorical independent variables, we find that the distribution of sex in the sample is homogeneous. It is also possible to notice how women tend to be more satisfied than men. Let's proceed with the variable relating to the type of customer. In this case the distribution is not homogeneous as the majority of passengers who took part in the questionnaire define themselves as loyal and this variable

could lead to bias problems. However, it is important to note that customers who define themselves as loyal have not always expressed a favorable opinion of the company, so this aspect could compensate for the lack of balance between loyal and non-loyal customers. Moving on to the type of trip taken, we note how the majority of the sample took a trip for work and not for personal purposes. Also in this case it is possible to notice a fairly homogeneous division between satisfied and dissatisfied customers in both categories. To conclude, let's consider the variable relating to the class in which the passengers traveled. We can see how the majority of customers traveled in business and eco and only 7.36% of those interviewed declared they had traveled in eco plus. Therefore, in order to improve our analyzes and make the model more significant we could combine the eco plus category with the eco category, so as to have an even more homogeneous distribution of the variables

Fig. 3.2: barplot of the 4 categorical independent variables before outliers removal



Moving on to the numerical variables instead, we can see how the age variable has a fairly normal distribution with an average value of 40 years, whether we consider satisfied or unsatisfied customers. Regarding the distance however, we can notice the presence of outliers in the right part of the graph which makes the distribution far from normal. This can also be verified using a boxplot which gives information on the range of attributes and helps to identify any outliers. Therefore, we proceed with the removal of these outliers by exploiting the first and third quantiles. From the boxplots it is possible to see how the outliers have been eliminated from the distribution and the dataset has gone from 119.255 observations to 116.962 observations

Fig. 3.3: distribution of the numerical variables before outliers removal: age and distance

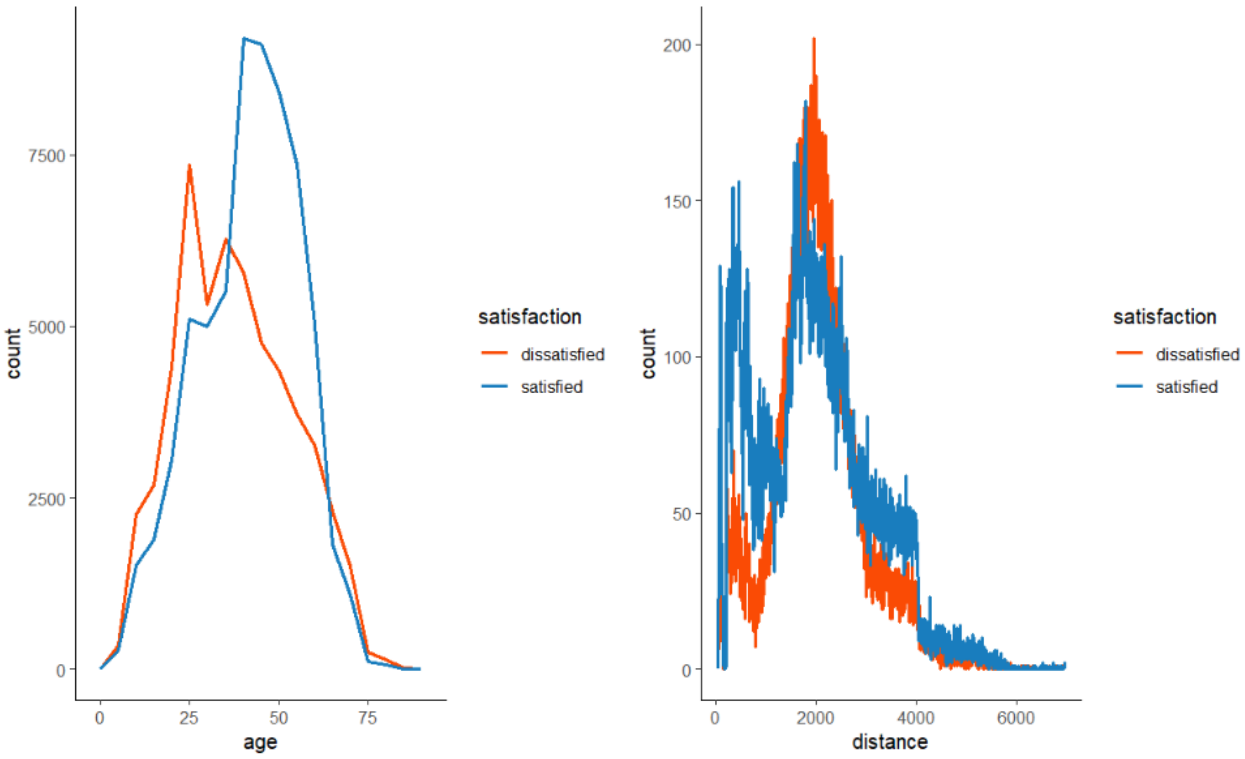
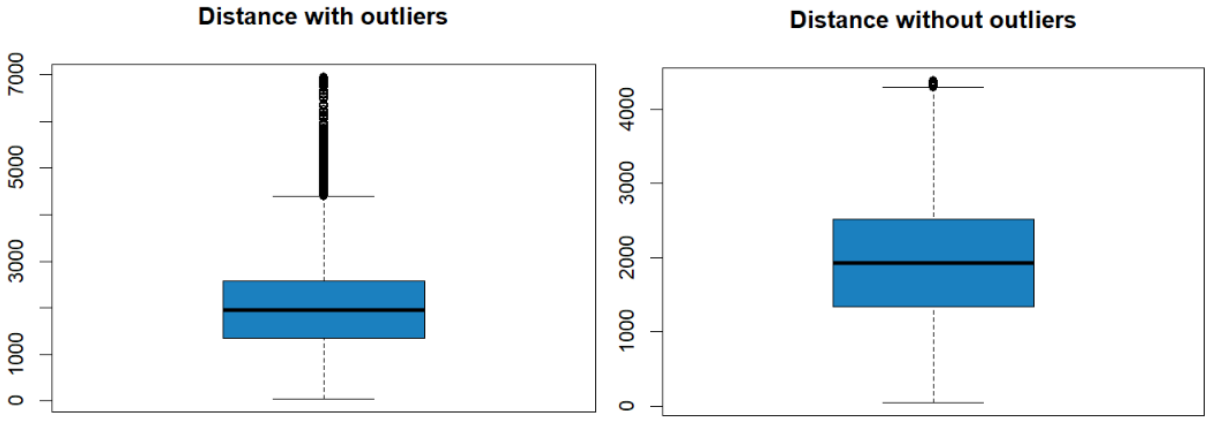
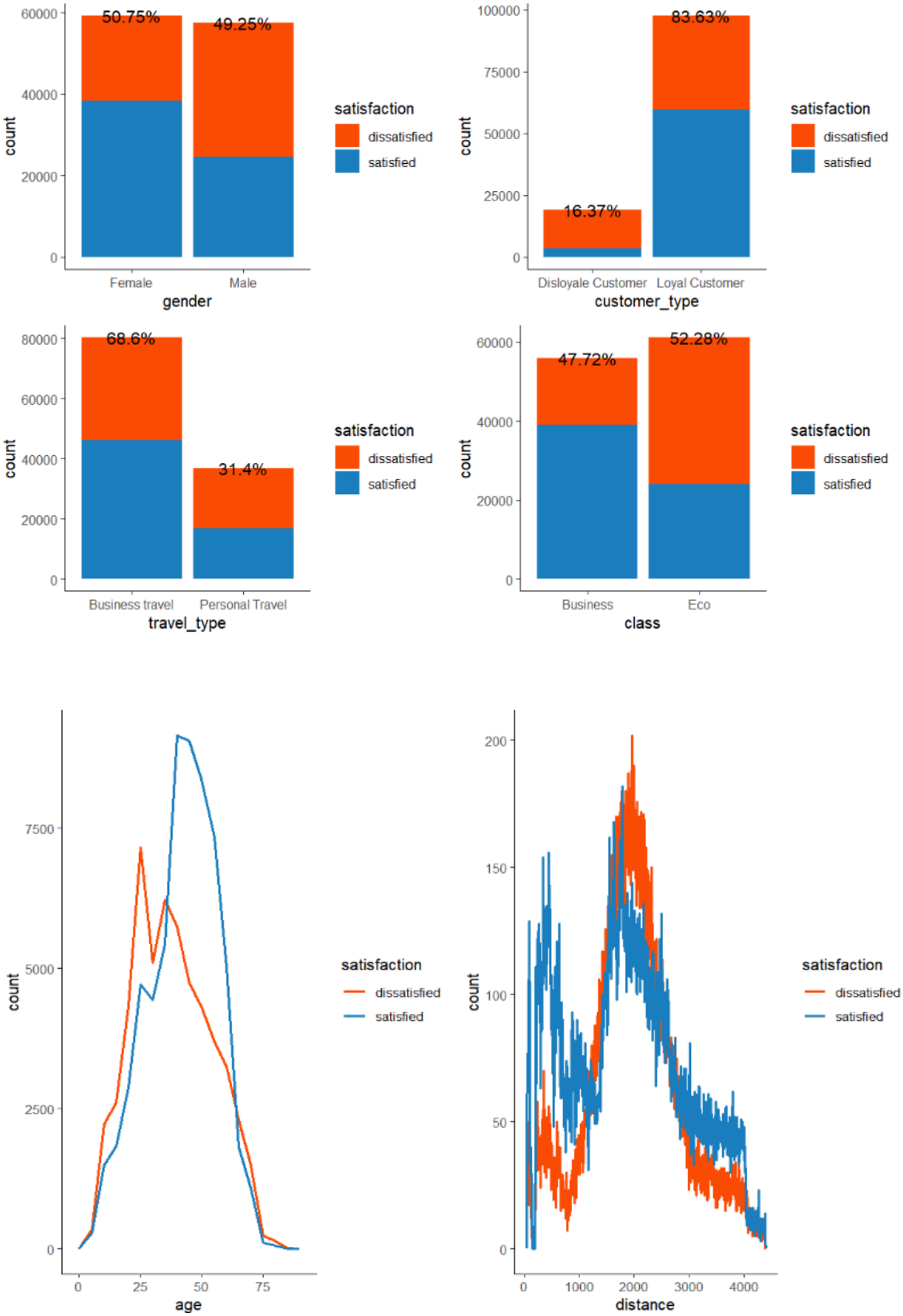


Fig. 3.4: comparison between the distribution of the variable distance before and after the removal of outliers



Therefore, once the outliers have been removed and the Eco and Eco plus categories have been grouped, our variables are finally ready for the logistical analysis which will be carried out in the next chapter. The number of satisfied and unsatisfied passengers remains practically identical, going from 45.86% to 46.09% of dissatisfied people and from 54.14% to 53.91% of satisfied people. The new distributions of the variables will be exposed below

Fig. 3.5: barplot and distribution of the independent variables after outliers removal



3.3 - Data preparation - dependence and correlation

Now we will study the dependence between variables, starting with the relationship that exists between the dependent variable and the independent variables using the chi-square test with a 95% confidence interval. The test consists in verifying whether the frequencies observed in one or more categories correspond to the expected frequencies. The null hypothesis for any pair of variables is that they are independent of each other. A high chi-square value and a low p-value will allow us to reject the null hypothesis and therefore conclude that the variables are dependent on each other. However, to measure the extent of the association between two variables we will use the Cramer's V method. This function calculates Cramer's V, a measure of association between two categorical variables. It ranges from 0 to 1 where:

- 0 indicates no association between the two variables.
- 1 indicates a strong association between the two variables.

It is calculated as:

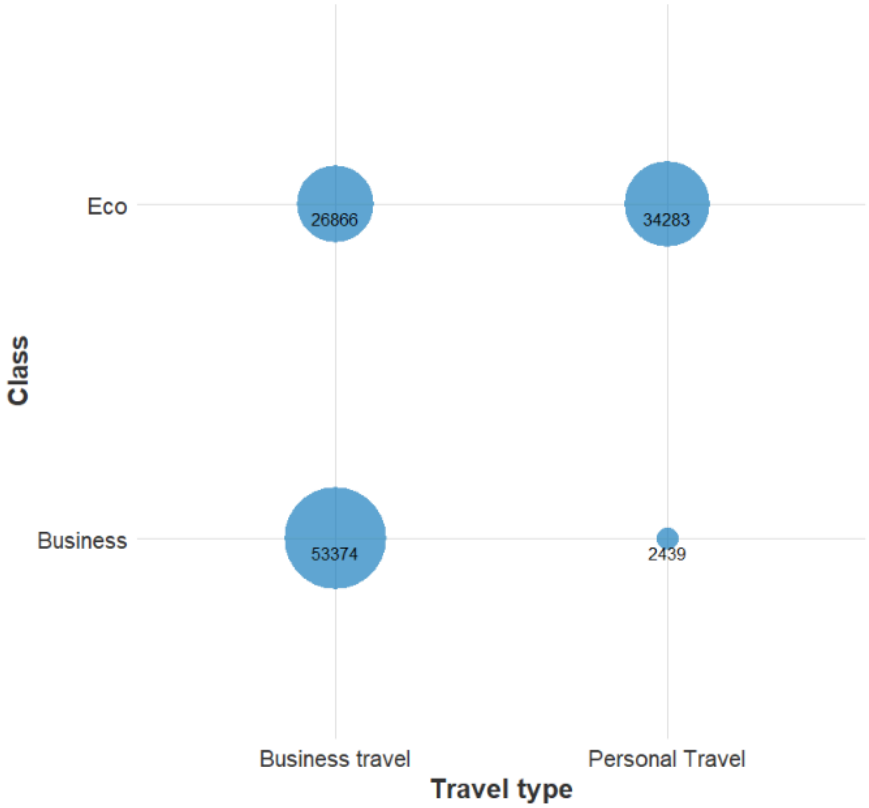
$$Cramer's\ V\ = \sqrt{\frac{(X^2/n)}{min(c-1, r-1)}}$$

where: X^2 is the Chi-square statistic, n is the total sample size, r is the number of rows and c is the number of columns.

Tab. 3.6: level of dependence between variables

Variable 1	Variable 2	Chi-square test	P-values	Cramer's V test
Satisfaction	Gender	5635.8, df = 1	< 2.2e-16	0.2195
Satisfaction	Customer Type	11715, df = 1	< 2.2e-16	0.3165
Satisfaction	Travel Type	1477.9, df = 1	< 2.2e-16	0.1124
Satisfaction	Class	11175, df = 1	< 2.2e-16	0.3091
Gender	Customer Type	108.02, df = 1	< 2.2e-16	0.0304
Gender	Travel Type	7.6759, df = 1	0.005596	0.0081
Gender	Class	7.9303, df = 1	0.004861	0.0082
Customer Type	Travel Type	9862.3, df = 1	< 2.2e-16	0.2904
Customer Type	Class	724.69, df = 1	< 2.2e-16	0.0787
Travel Type	Class	36201, df = 1	< 2.2e-16	0.5564

Fig. 3.7: balloon plot between class and travel type



We can therefore see from Table 3.6 that the categorical variables are poorly correlated with each other except for the 'Travel Type' and 'Class' variables with a Cramer's V test result of 0.56. This can be explained by the fact that business flights will be associated with the Business class while for personal flights less expensive classes will be preferred. It is possible to note what has been said in the balloon plot in figure 3.7, in which this dependence between business flights and business trips is evident as the flights are paid by the company and not by the individual worker. Looking instead at the two numerical variables, the correlation is -0.24, therefore a low negative correlation linking the age of the passenger with the distance traveled. We can therefore conclude that our variables are not highly correlated and can be safely inserted into our logistic model. We also remember that the outliers have been removed so as to fully respect the assumptions of a logistic regression.

3.4 - Training and test

We now move on to our analyzes by dividing the dataset of 116.962 observations into a training set and a test set. The training set will be formed by taking 80% of the observations from our initial dataset for a total of 93.570 observations and the remaining 23.392 will be included in the test set which forms the remaining 20% of the initial dataset. We then proceed to carry out our regression using the train dataset, so as to have a first idea of the statistically significant variables. Looking at the complete logistic model which includes all the variables, we can see that they are all statistically significant at 0.01%.

The age variable, on the contrary, is statistically significant at 10% and even considering the logarithm does not improve the significance, on the contrary it worsens because the p-value increases. The pseudo R square is the same in both cases, but considering the AIC (Akaike Information Criterion) which is based on the likelihood function, the preferred model should be the first as it has a lower value

Fig. 3.8: full logistic model

	Full model	
(Intercept)	-0.38 ***	(0.00)
genderMale	-1.13 ***	(0.00)
customer_typeLoyal Customer	2.24 ***	(0.00)
age	-0.01	(0.06)
travel_typePersonal Travel	-0.26 ***	(0.00)
classEco	-1.32 ***	(0.00)
distance	-0.12 ***	(0.00)
N	93570	
AIC	104131.24	
BIC	104197.36	
Pseudo R2	0.31	

All continuous predictors are mean-centered and scaled by 1 standard deviation. The outcome variable is in its original units. Standard errors are heteroskedasticity robust.
*** p < 0.001; ** p < 0.01; * p < 0.05.

Fig. 3.9: full logistic model with log(age)

	Full model with log(age)	
(Intercept)	-0.37 ***	(0.00)
genderMale	-1.14 ***	(0.00)
customer_typeLoyal Customer	2.22 ***	(0.00)
`log(age)`	0.01	(0.24)
travel_typePersonal Travel	-0.25 ***	(0.00)
classEco	-1.32 ***	(0.00)
distance	-0.12 ***	(0.00)
N	93570	
AIC	104133.37	
BIC	104199.49	
Pseudo R2	0.31	

All continuous predictors are mean-centered and scaled by 1 standard deviation. The outcome variable is in its original units. Standard errors are heteroskedasticity robust.
*** p < 0.001; ** p < 0.01; * p < 0.05.

Again based on the AIC we will now build a forward and backward stepwise regression model. Both methods aim to select a subset of predictors that maximizes the model's predictive ability, but they differ in their starting points and progression.

Forward stepwise selection builds the model incrementally from no predictors, while backward stepwise selection starts with all predictors and gradually eliminates the least influential ones.The algorithm will add one regressor at a time into the model until it reaches the complete model (stepwise) and will do the same backwards by selecting the model that minimizes the AIC:

$$AIC = 2k - 2ln(L)$$

where K is the number of model parameters and ln(L) is the log-likelihood of the model.

Fig. 3.10: model obtained with forward and backward stepwise regression

	Forward and backward stepwise regression model	
(Intercept)	-0.38 ***	(0.00)
genderMale	-1.13 ***	(0.00)
customer_typeLoyal Customer	2.24 ***	(0.00)
age	-0.01	(0.06)
travel_typePersonal Travel	-0.26 ***	(0.00)
classEco	-1.32 ***	(0.00)
distance	-0.12 ***	(0.00)
N	93570	
AIC	104131.24	
BIC	104197.36	
Pseudo R2	0.31	

All continuous predictors are mean-centered and scaled by 1 standard deviation. The outcome variable is in its original units. Standard errors are heteroskedasticity robust.
*** p < 0.001; ** p < 0.01; * p < 0.05.

In this case the model proposed by the algorithm is identical to the full model indicated in figure 3.8. We can therefore proceed to calculate the average marginal values(AME), which allow us to interpret the coefficients of the logistic model and calculate the predictive power of the model by exploiting the test set

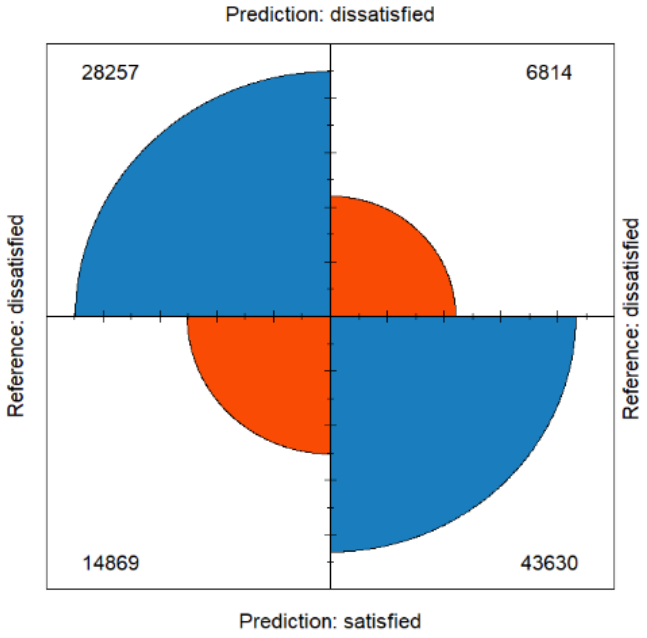
Fig. 3.11: Average Marginal Effects(AME) for every variable, with standard errors, T-statistic, p-values a confidence interval

factor	AME	SE	z	p	lower	upper
age	-0.0002	0.0001	-1.8506	0.0642	-0.0004	0.0000
classEco	-0.2650	0.0037	-71.3439	0.0000	-0.2723	-0.2577
customer_typeLoyal Customer	0.4205	0.0037	114.8439	0.0000	0.4133	0.4277
distance	0.0000	0.0000	-15.3930	0.0000	0.0000	0.0000
genderMale	-0.2196	0.0029	-76.0220	0.0000	-0.2253	-0.2139
travel_typePersonal Travel	-0.0499	0.0039	-12.6958	0.0000	-0.0576	-0.0422

As we can see from table 3.11, it can be stated that, all things being equal, an additional year of age leads on average to a reduction in the probability of being satisfied with the flight by 0.02 percentage points. However, remember that this coefficient is only 10% significant compared to the other coefficients which are much more significant. Continuing we find that on average those traveling in the Eco class, and therefore also Eco Plus, have a probability of expressing a favorable opinion of the flight by 26.5 percentage points lower than those traveling in Business, all things being equal. Loyal customers are 42.05 percentage points more likely to express a positive opinion on the flight compared to non-loyal customers, holding the remaining variables constant. Looking at the distance, we note that it does not influence the probability of expressing a favorable opinion of the flight as the coefficient is zero. Continuing from this, we note that men have a lower probability of 21.96 percentage points than women of expressing a positive

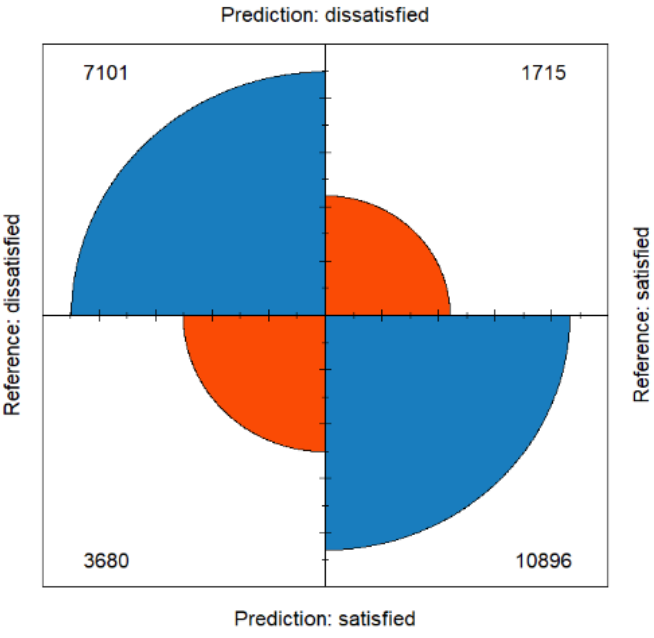
opinion of the flight, always under equal conditions and on average. Finally, we conclude with the variable relating to the type of flight. Those who travel for personal purposes have on average a lower probability of 4.99 percentage points of expressing a favorable opinion compared to those who travel for work and business purposes, all other things being equal. From an economic point of view, therefore, we can generally state that older women, who travel in business, who are loyal to the company and who travel for work purposes have a higher probability of being satisfied with the flight compared to those who do not respect these requirements. Let us now move on to the evaluation of the model and the quality of the predictions. To do this we will exploit the confusion matrix and the ROC curve. The confusion matrix is a tool for analyzing the errors made by a machine learning model and evaluating its performance. It is a table that shows the number of correct and incorrect predictions made by the model on a test dataset and provides a summary of the model's performance.

Fig. 3.12: training set confusion matrix




```
               Accuracy : 0.7683
                 95% CI : (0.7656, 0.771)
    No Information Rate : 0.5391
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5273

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8649
            Specificity : 0.6552
         Pos Pred Value : 0.7458
         Neg Pred Value : 0.8057
             Prevalence : 0.5391
         Detection Rate : 0.4663
   Detection Prevalence : 0.6252
      Balanced Accuracy : 0.7601

       'Positive' Class : satisfied
```

Fig. 3.13: test set confusion matrix




```
               Accuracy : 0.7694
                 95% CI : (0.7639, 0.7748)
    No Information Rate : 0.5391
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5297

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8640
            Specificity : 0.6587
         Pos Pred Value : 0.7475
         Neg Pred Value : 0.8055
             Prevalence : 0.5391
         Detection Rate : 0.4658
   Detection Prevalence : 0.6231
      Balanced Accuracy : 0.7613

       'Positive' Class : satisfied
```

The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives, as well as the overall accuracy and other performance metrics of the model.The elements of a confusion matrix are:

- True Positives (TP): the number of examples that were predicted as positive (label satisfied) and are actually positive (true label satisfied).
- True Negatives (TN): the number of examples that were predicted as negative (label dissatisfied) and are actually negative (true label dissatisfied).
- False Positives (FP): the number of examples predicted as positive (label satisfied) but actually negative (true label dissatisfied).
- False Negatives (FN): the number of examples predicted as negative (label dissatisfied) but actually positive (true label satisfied).

The confusion matrix can be used to calculate various evaluation metrics, such as error rate, accuracy, precision and recall or sensitivity:

- Accuracy measures the percentage of exact predictions out of the total instances. Ranges from 0 (worst) to 1 (best).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

  Looking at the test set we can see an accuracy of the estimates of 0.77, i.e. 77% of the estimates were correct. We can therefore be quite satisfied with the result obtained from our model.

- Recall or sensitivity is the percentage of correct positive predictions (TP) out of the total positive instances. Ranges from 0 (worst) to 1 (best)

$$Sensitivity = \frac{TP}{TP + FN}$$

  In this case our sensitivity is very high, indicating how the model correctly estimates the people who were satisfied with the flight. However, it must be said that in our sample there were more satisfied passengers than unsatisfied passengers, with percentages of 53.91% versus 46.09%. So even if the difference is minimal, the presence of a greater number of people categorized as satisfied may have made our model slightly more unbalanced towards positive results

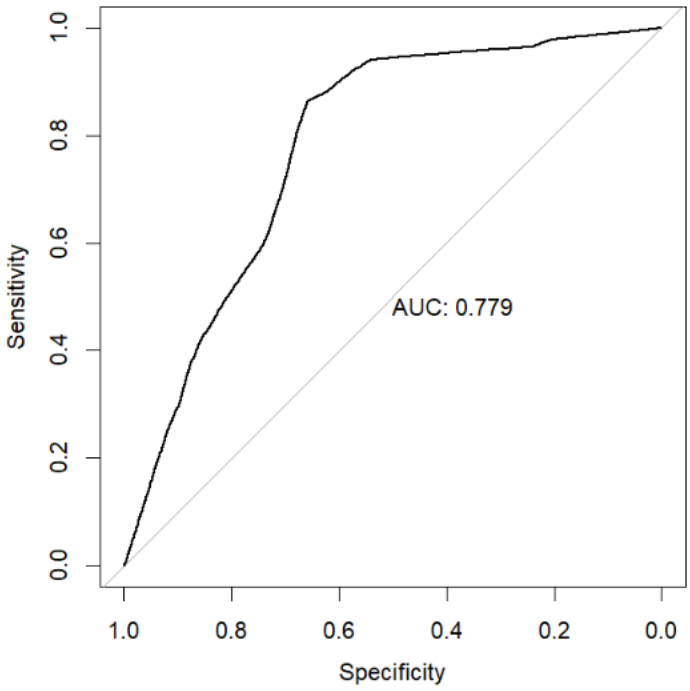
- Specificity is the percentage of correct negative predictions (TN) out of the total negative instances. Ranges from 0 (worst) to 1 (best)

$$Specificity = \frac{TN}{TN + FP}$$

  As mentioned previously for sensitivity, in the case of specificity the opposite situation arises as we find a value of 0.66 indicating that only 66% of cases manage to correctly

classify true negatives. Therefore, although it is not a terrible value because it is closer to 1 than to 0, the fact that we obtained a not very high value indicates a sort of prevalence in classifying the units positively

Fig. 3.14: ROC curve



Let's now move on to the ROC curve (Receiver Operating Characteristics). The ROC curve is a graph that relates the sensitivity and specificity of a diagnostic test to changes in the threshold value. The analysis of the ROC curve of a diagnostic test allows us to evaluate its accuracy, determine the most appropriate threshold value and compare the performance of two, or more, different tests. By increasing the threshold value, the number of false negatives increases, while the number of false positives decreases. As a result, the test is highly specific but not very sensitive. Conversely, by lowering the threshold value, the number of false positives increases, while the number of false negatives decreases, therefore the test is highly sensitive but not very specific. On the ordinate axis (y axis) we find the sensitivity values of the test, or true positive rate (TPR – True Positive Rate) and on the x-axis we find the specificity of the test, or false positive rate (FPR – False Positive Rate). For each cut-off we have a certain value for sensitivity and a certain value for specificity, and we report the corresponding point on the graph, obtaining the ROC curve

In general, a test is more accurate the closer its ROC curve is to the upper left corner of the graph. Furthermore, the point closest to this angle represents the cut-off value that simultaneously maximizes the sensitivity and specificity of the test. That said, you can now understand why the subgraph area (AUC) is a measure of how accurate a test is. If the AUC is 0.5 the test is not informative (the curve would correspond to the bisector), if instead the AUC has a value between 0.5 and 1 then the test turns out to be increasingly more accurate. Therefore, looking at graph 3.14 we can initially state that an AUC value of 0.779 is positive for our analyses, indicating that the test is very accurate. It means there is a 77.9% chance that the model will be able to distinguish between positive class and negative class. Furthermore we can see how the optimal threshold value that maximizes the area under the curve is 0.5, as can be seen from table 3.15, which means the data is perfectly balanced. In this case our model is better able to identify false negatives, which decrease while the number of false positives increases and the accuracy of the model is maximized.

Fig. 3.15: threshold, specificity and sensitivity of the ROC curve

threshold	specificity	sensitivity
0.4940712	0.658195	0.8645627

Therefore, if the probability is greater than the threshold value, i.e. greater than 0.5, then we can conclude that the passenger will be satisfied with the flight and will express a positive opinion. On the contrary, a value below the threshold value will lead to a negative opinion and the customer will therefore be dissatisfied with the service offered by the airline.

3.5 - Linear Discriminant Analysis

Linear Discriminant Analysis or LDA is a supervised machine learning approach employed primarily for multi-class classification tasks. Its fundamental aim is to reduce data dimensionality while preserving crucial class discriminatory information. This framework involves modeling the data distribution for each class and leveraging Bayes' theorem to classify new data points based on conditional probabilities. In more details, the approach is to model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain Pr(Y|X):

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}$$

where:

- $f_k(x) = Pr(X = x|Y = k)$ is the density for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = Pr(Y = k)$ is the prior probability for class k.

In the context described, when classifying a new point, we consider which class's density is highest at that point. However, if the prior probabilities of the classes are different, we take them into account as well. This means we compare $\pi_k f_k(x)$ for each class k, where $f_k(x)$ represents the probability density function of class k evaluated at the new point x. The decision rule can be expressed as follows:

- If $\pi_1 * f_1(x) > \pi_2 * f_2(x)$, we classify x into class 1(dissatisfied in our case)
- If $\pi_1 * f_1(x) < \pi_2 * f_2(x)$, we classify x into class 2(satisfied in our case)

This decision rule accounts for both the likelihood of observing x in each class (represented by the PDF) and the prior belief or expectation of encountering each class (represented by π_k). It allows for a principled classification approach, considering the relative importance of each class and their distributions. When the priors are different, the decision boundary may shift accordingly. For instance, if one class has a higher prior probability, the decision boundary might

favor that class. This ensures that the classification decision takes into account not only the data but also the prior information about the classes. By taking logarithms and discarding terms that are independent of k, this task is equivalent to assigning x to the class with the largest discriminant score $\delta_k(x)$:

$$\delta_k(x) = x * \frac{u_k}{\sigma^2} - \frac{u_k^2}{2\sigma^2} + log(\pi_k)$$

where:

- $\delta_k(x)$ is a linear function of x
- u_k is the mean of the predictor variable in class k
- σ^2 is the variance of the predictor variable
- π_k is the prior probability of class k

In a scenario with K=2 classes, as in our case and equal prior probabilities $\pi_1 = \pi_2 = 0.5$ the decision boundary can be found by setting $\delta_1(x) = \delta_2(x)$, leading to:

$$x = \frac{u_1 + u_2}{2}$$

This equation represents the value of x at which the decision boundary is located between the two classes. Any x value less than this boundary is classified into one class, while values greater than or equal to the boundary are classified into the other class.
Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{k} e^{\hat{\delta}_l(x)}}$$

This formula calculates the probability of belonging to class k given the observed value x. It uses the exponentiated discriminant scores to convert them into probabilities, ensuring that the sum of probabilities across all classes equals 1. In the case where there are only two classes (K=2), the decision rule for classification simplifies as follows:

- If $Pr(Y = 2|X = x) \geq 0.5$, we classify x into class 2(satisfied in our case)
- Otherwise, we classify x into class 1(dissatisfied in our case)

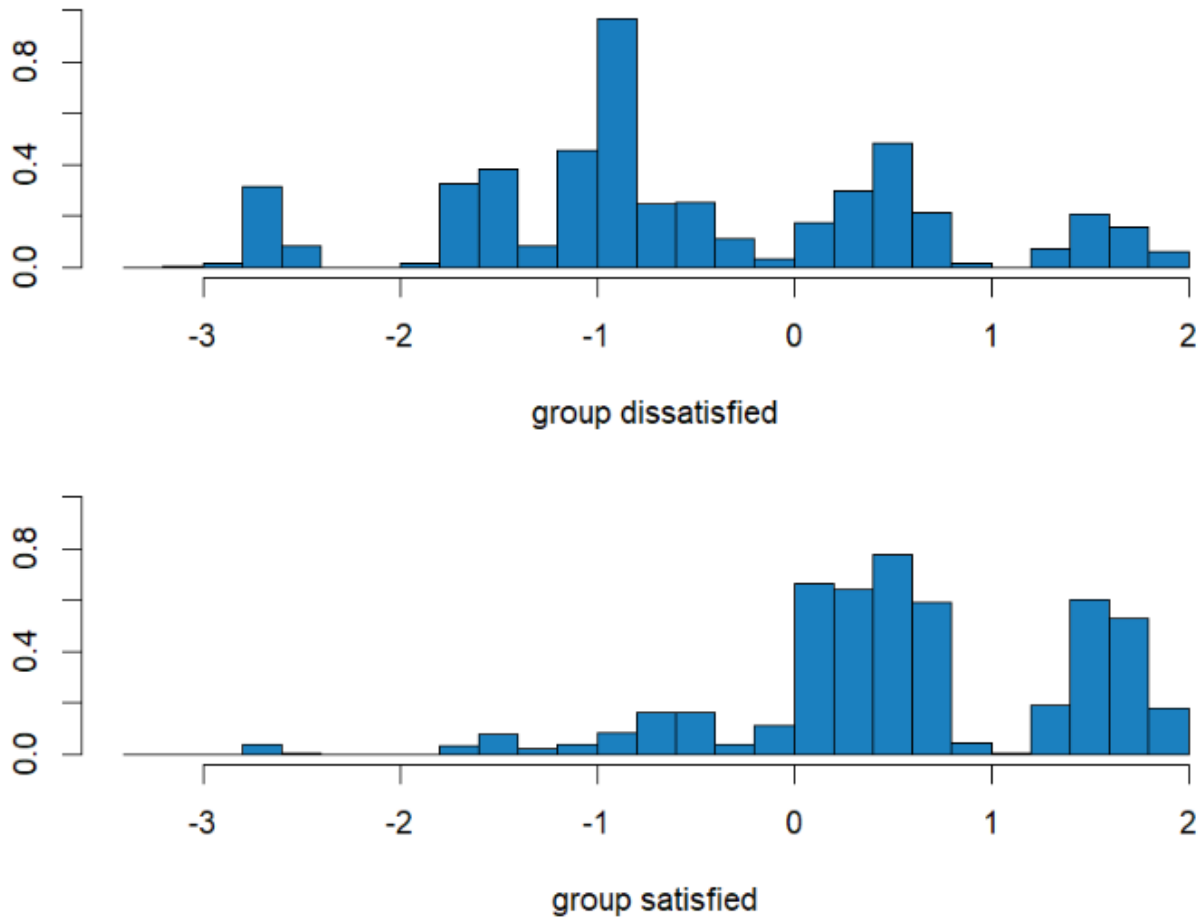
Linear Discriminant Analysis seeks to maximize the separation between class means while minimizing within-class variance. The technique involves projecting high-dimensional data onto a lower-dimensional space, often referred to as dimensionality reduction.

Fig. 3.16: coefficients of Linear Discriminant Analysis

	Coefficients of LDA
genderMale	-1.03660641
customer_typeLoyal Customer	2.03738330
scale(age)	-0.02034136
travel_typePersonal Travel	-0.33408796
classEco	-1.21062623
scale(distance)	-0.12525314

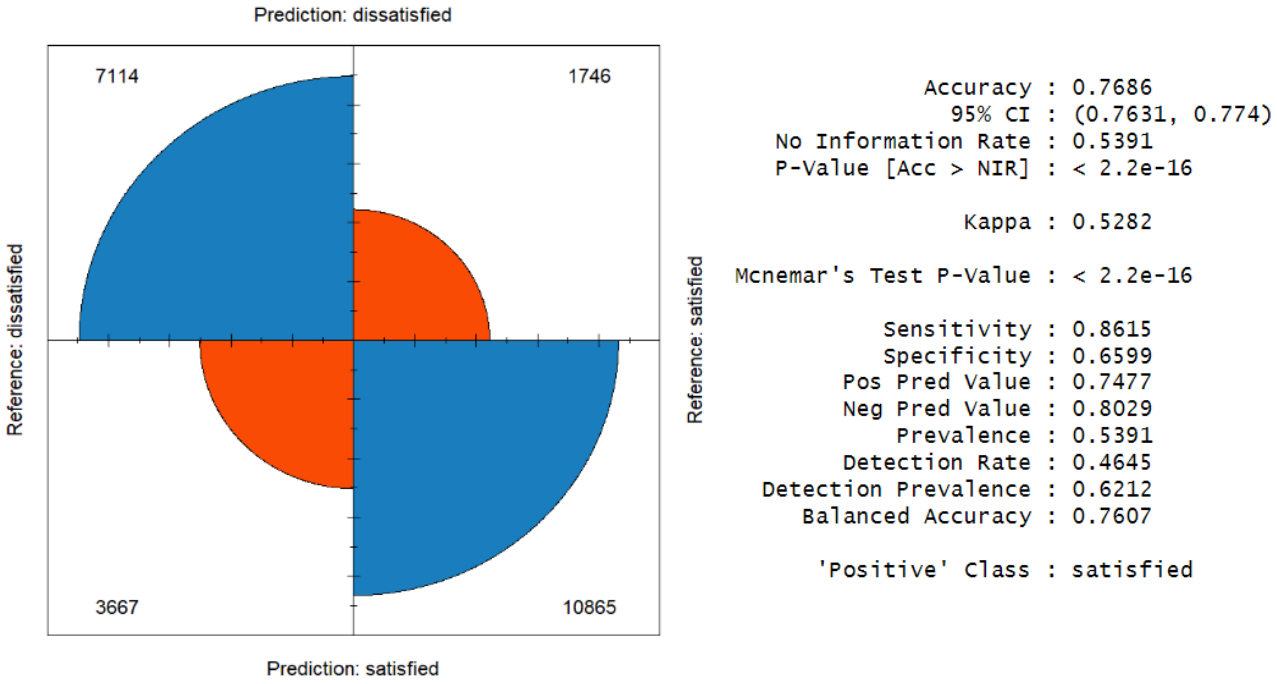
Therefore, exploiting this model we obtain the coefficients of linear discriminants for each predictor variable in the model. These coefficients indicate the contribution of each predictor variable to the linear combination used for classification. From the graph above in fig. 3.17 we have a histogram of the linear component and we can see that the separation between the group of satisfied and dissatisfied people is in some points overlapped, mainly in the right part of the graph. This can create some problems for the model as a fairly clear separation between classes is required

Fig. 3.17: separation between the two classes



We now proceed to evaluate the model predictions by looking at the confusion matrix obtained in Fig. 3.18. As we can see we obtain a good level of accuracy, very similar to that obtained from the stepwise logistic model, in which the level of accuracy was 0.7694. Similar results are therefore also obtained in relation to sensitivity and specificity, indicating how the LDA model and the logistic model have led to the same results and performances.

Fig. 3.18: confusion matrix of the Linear Discriminant Model



3.6 - Decision Trees

This methodology has the objective of obtaining a hierarchical segmentation of a set of units, sometimes even very large ones, through the identification of "rules" that exploit the relationship existing between the class to which they belong and the variables detected for each unit. The graphical output of the procedure consists of a tree structure, with nodes, branches and leaves. The application of decision trees requires a priori knowledge of the class to which each unit belongs: the purpose of the technique is to better predict the class to which the individual units belong. Decision trees do not place all available variables on the same logical level. Contrary to what happens in cluster analysis, one variable here takes on the role of dependent variable, while the others (also called attributes) are considered explanatory. Decision trees are therefore an asymmetric segmentation technique and the homogeneity of the groups refers only to the modalities of the dependent variable. The examination of the individual effect of each character allows us to select only the most relevant variables for the purposes of classifying the units and therefore to arrive at decision rules which are usually easy to interpret and immediately use. Tree procedures are in fact much more flexible and do not require the satisfaction of stringent hypotheses, neither on the type of relationship nor on the form of distribution of the dependent

variable. From a visual point of view, a tree represents a finite set of elements called nodes. The node from which the subsequent ones branch off is called the root and is sometimes indicated with the letter R or with the term "node 0". The set of nodes, with the exception of the root node, can be divided into h distinct sets S1, S2, ..., S, which are referred to as subtrees of the node R. The set of nodes descending from a given intermediate node is called a branch. A node is called a parent with respect to the nodes it generates, while it is called a child with respect to the node from which it descends. Threshold values of a variable that divide the units of a given node are called splits. The terminal nodes are called leaves.

In summary, the hierarchical segmentation obtained through a decision tree can therefore be defined as a stepwise procedure, through which the set of n statistical units is progressively divided, according to an optimization criterion, into a series of disjoint subgroups which present within them a greater degree of homogeneity than the initial set. Segmentation therefore provides a hierarchical succession of partitions of the set of n units obtained with a splitting or top down criterion. At each step of the process the heterogeneity in the groups is reduced compared to the previous step. At the end, the leaves of the tree, used to graphically describe the procedure, present such a degree of homogeneity that they can be attributed to one of the starting classes. The decision rule defined by the leaves of the tree can then be used to classify new cases whose class they belong to is not known. But how does it work ?

- We divide the predictor space — that is, the set of possible values for X1, X2, . . . , Xp — into J distinct and non-overlapping regions, R1, R2, …, Rj .
- For every observation that falls into the region Rj , we make the same prediction, which is simply the mean of the response values for the training observations in Rj .

The goal is to find boxes R1, …, Rj that minimize the RSS, given by:

$$RSS = \sum_{J=1}^{J} \sum_{I \in R_j} (y_i - \widehat{y_{R_j}})^2$$

where $\widehat{y_{R_j}}$ is the mean response for the training observations within the jth box. In practical scenarios, considering every possible partition of the feature space into j boxes is computationally infeasible due to the exponentially growing number of combinations. To address this challenge, decision tree algorithms employ a top-down, greedy approach known as recursive binary splitting. This approach starts at the root node, representing the entire feature space, and recursively splits the predictor space into two child nodes at each step. Each split divides the feature space into two subsets based on the value of a chosen predictor variable. At each step of the tree-building process, the algorithm selects the best split among all possible splits available at that particular step. It evaluates each predictor variable and its possible values to determine the split that optimally separates the data into two subsets, maximizing the homogeneity or purity of the response variable within each subset. The algorithm uses a splitting criterion, such as Gini impurity or information gain for classification tasks, or mean squared error reduction for regression tasks, to evaluate the effectiveness of each split.

It chooses the split that maximizes the improvement in homogeneity or minimizes the impurity within the resulting child nodes. The splitting process continues recursively until a stopping criterion is met, such as reaching a maximum tree depth, minimum number of samples in a node, or when further splits no longer provide significant improvement in homogeneity or purity. By employing recursive binary splitting, decision tree algorithms efficiently construct a hierarchical tree structure by making locally optimal decisions at each step. While this greedy approach may not always lead to the globally optimal tree, it often results in a satisfactory tree that effectively captures the underlying patterns in the data.

In the process of building a regression tree, we begin by selecting a predictor variable Xj and a cut point s that divides the predictor space into two regions: {X|Xj < s} and {X|Xj ≥ s}. This division is chosen to maximize the reduction in the residual sum of squares (RSS). We then repeat this process iteratively, seeking the best predictor and cutpoint to further partition the data and minimize the RSS within each resulting region. However, instead of splitting the entire predictor space each time, we focus on one of the two previously identified regions, creating three regions in total. We continue this process, further partitioning the regions to minimize the RSS. This iterative approach continues until a stopping criterion is met, such as when no region contains more than five observations. Through this process, a regression tree is constructed, effectively capturing the underlying patterns in the data.

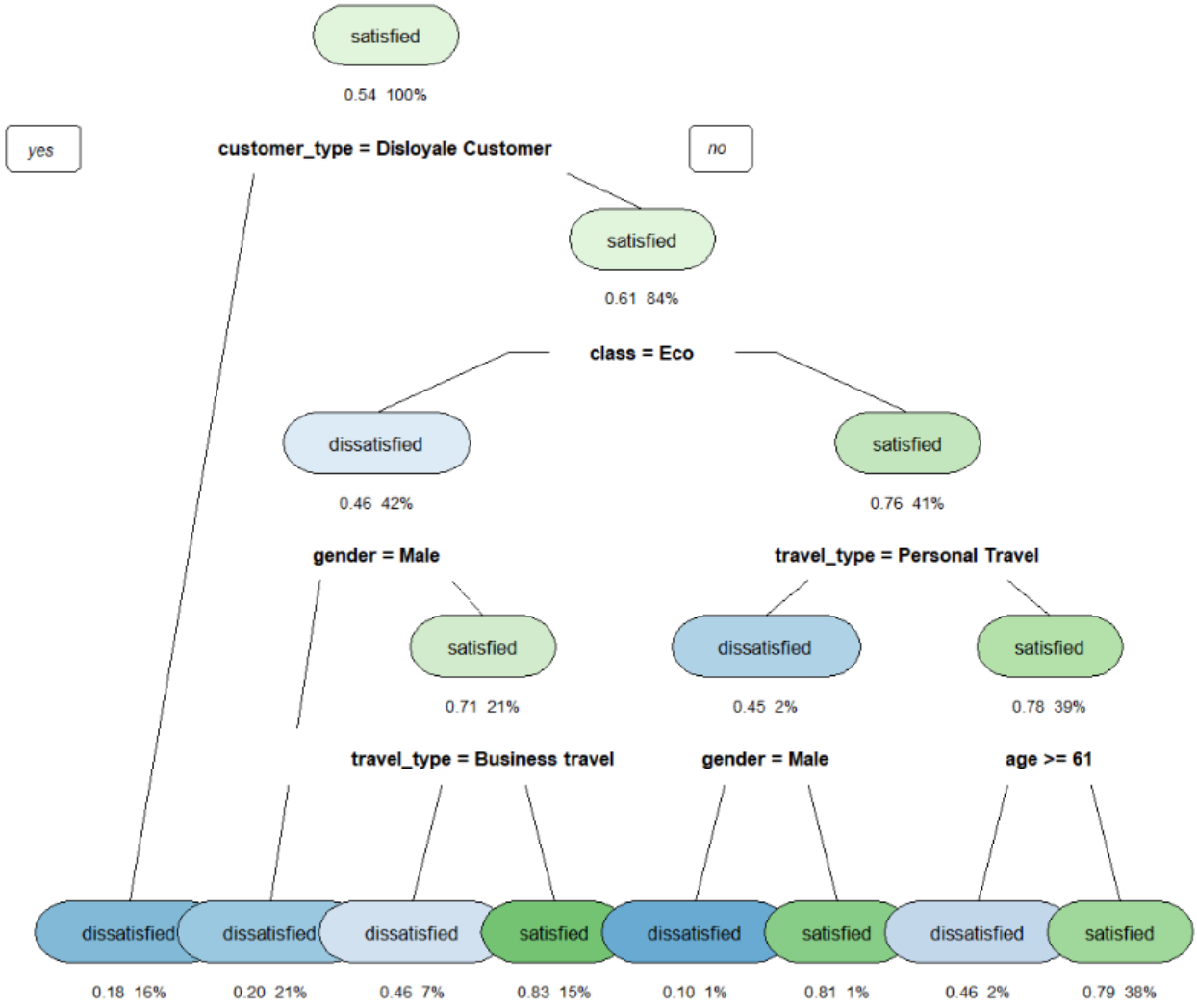
Fig. 3.19: results obtained from the decision tree algorithm

CP	nsplit	rel error	xerror	xstd
0.226197653	0	1.0000000	1.0000000	0.003535630
0.135753374	1	0.7738023	0.7738023	0.003397596
0.011849001	3	0.5022956	0.5022956	0.002991785
0.009205584	4	0.4904466	0.4904466	0.002966772
0.003014423	6	0.4720354	0.4720354	0.002926466
0.003000000	7	0.4690210	0.4708760	0.002923868

Based on the provided data table, we can observe distinct trends associated with different values of the complexity parameter (CP). As the CP value decreases, indicating a higher level of model complexity, there is a noticeable increase in the number of splits represented by the "nsplit" column. This escalation in splits suggests a more intricate decision tree structure, capable of capturing finer details in the data. Simultaneously, we witness a decline in the "rel error" column as CP decreases, indicating a reduction in error relative to the initial error. This implies that the model becomes more adept at fitting the training data with lower CP values, potentially leading to improved performance. Moreover, the cross-validation error ("xerror") shows a consistent decrease with diminishing CP values. A lower cross-validation error suggests enhanced generalization ability of the model, implying that it can better handle unseen data. Accompanying the decline in cross-validation error, we observe a reduction in the cross-validation standard deviation ("xstd"). This decrease signifies a decrease in the variability of model performance across different cross-validation folds, indicating more stable and reliable predictions. The provided data underscores the trade-off between model complexity and generalization performance. While lower CP values result in more complex

models with increased splits, they also tend to offer superior generalization performance, as evidenced by lower cross-validation error and standard deviation. However, selecting the optimal CP value necessitates careful consideration of these trade-offs, alongside the specific requirements and constraints of the problem being addressed. So by looking at table 3.19 we can say that 0.003 is the best choice since the relative error and the error of the cross validation stop decreasing significantly.

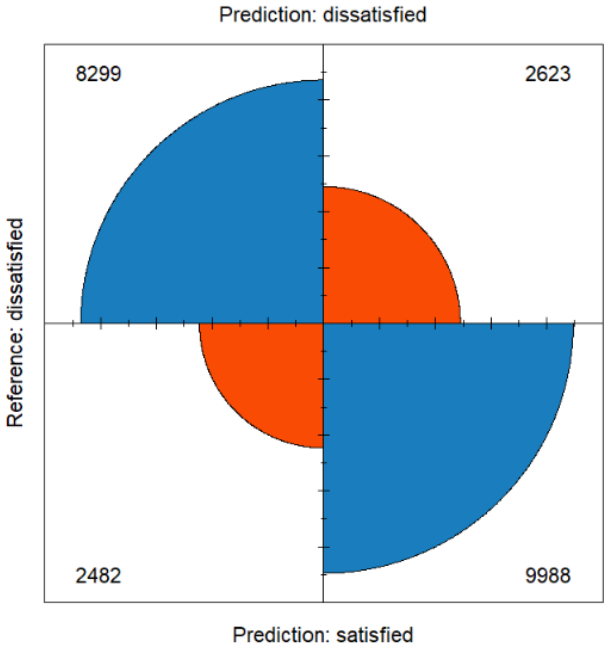
Fig. 3.20: decision tree

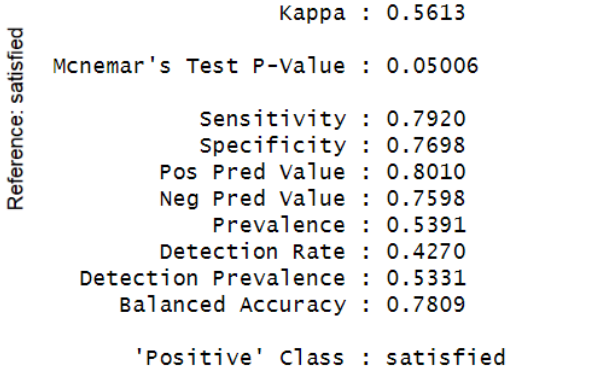


Looking at the decision tree in figure 3.20 we therefore obtain a graphical representation of the logistic regression obtained in the previous paragraphs. It can be seen that the model used all variables except the distance variable, which actually had a coefficient of zero in the logistic regression despite being statistically significant. It is also noted that the independent variable of greatest importance among all is that relating to the type of customer, as it is the variable from which the tree begins. We also note that the individuals who will most certainly be satisfied with the services offered by the airline during flights are female loyal customers, who travel in

Economy and who travel for personal reasons. In this case, 15% of the units in the sample have an 83% probability of remaining satisfied, despite traveling in economy and for personal reasons decreasing this probability. Relative to the group of most satisfied people we find male loyal customers, those who travel for business and for personal reasons. There is therefore a 10% probability of expressing a positive opinion, therefore these units are rightly classified as dissatisfied. However, this is 1% of the units, therefore a very low proportion.

Fig. 3.21: confusion matrix of the decision tree





As we can see from figure 3.21, which presents the confusion matrix of the decision tree model, we have a slightly higher level of accuracy than that obtained with previous models. In fact, let us remember that the logistic model obtained with the stepwise method had an accuracy value of 0.7694 and the LDA model a value of 0.7686. In this case the value is 0.7818 because as can be seen by comparing the confusion matrices of the three models the sensitivity has slightly decreased but the specificity, lacking in the remaining models, has increased in this case and has reached the same level as the sensitivity. Therefore we can summarize that the tree model proposes more balanced estimates between false negatives and false positives and is less distorted, as the number of the latter decreases.

3.7 - K-fold cross validation

Previously we divided our dataset into two parts, namely the training set and the test set, obtaining a fairly significant model with a good percentage of accuracy. However, when employing the validation approach for model evaluation, it's crucial to recognize that the estimation of test error can exhibit high variability. This variability arises from the specific

composition of the training set, meaning that slight variations in which observations are included in each set can lead to different performance estimates. This is particularly significant because, in the validation approach, only a subset of the data, namely those in the training set, is used to fit the model. Consequently, the model may not fully capture the complexities present in the entire dataset. As a result, the validation set error, which is based on this limited subset, may tend to overestimate the true test error for the model fitted on the entire dataset. This phenomenon occurs because the model might generalize better or worse when exposed to unseen data that differs from the training set. Therefore, it's essential to interpret validation set error estimates with caution, understanding that they may not accurately reflect the model's performance on completely unseen data. To mitigate this issue, techniques like cross-validation, which repeatedly partition the data into training and validation sets, can provide more robust estimates of model performance by averaging out the variability inherent in single validation set approaches.

A widely used approach for estimating test error is k-fold cross-validation. The concept behind k-fold cross-validation is to randomly divide the dataset into K equal-sized parts or folds. Then, iteratively, one of the K parts is left out as the validation set while the model is trained on the remaining K-1 parts combined. Predictions are then made for the left-out k-th part. This process repeats for each part k = 1, 2, ..., K, ensuring that each part serves as the validation set exactly once. Finally, the results from each fold are combined, often by averaging, to obtain an overall estimate of model performance. This approach provides a robust estimate of the model's generalization ability by leveraging multiple train-test splits of the data, thus offering a more reliable assessment compared to a single train-test split. By utilizing k-fold cross-validation, practitioners can make informed decisions about model selection and have a better understanding of the expected performance of the chosen model on unseen data.

Let the K parts be $C_1, C_2, \ldots C_k$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/k$

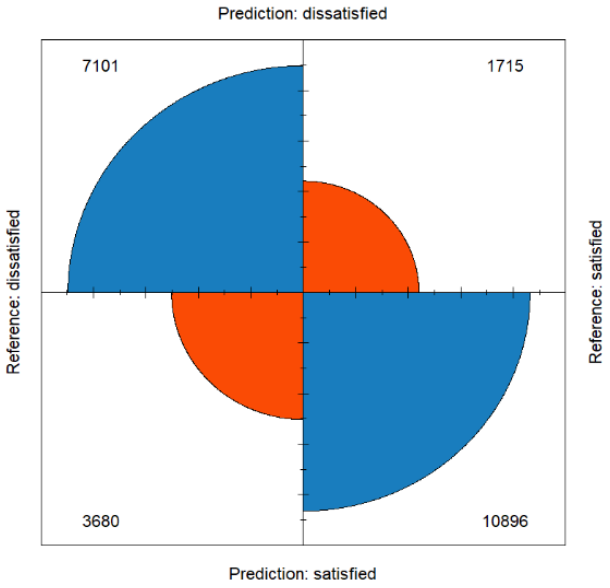
$$CV_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

where $MSE_k = \sum_{i \in C_k} (y_i - \widehat{y_i})^2 / n_k$ and $\widehat{y_i}$ is the fit for the observation i, obtained from the data with part k removed. The upward bias in k-fold cross-validation occurs because each training set is only a fraction of the original dataset's size, leading to an underrepresentation of the data's complexity. This bias is minimized in leave-one-out cross-validation (LOOCV), where each fold consists of a single observation, but LOOCV often suffers from high variance. To strike a balance between bias and variance, practitioners commonly use k values like 5 or 10, offering a reasonable compromise. These values provide a sufficient reduction in bias while maintaining a manageable level of variance, making them suitable for reliable model evaluation. As we can see, cross validation with 10 folds was applied. The training set composed of 93.570 observations was divided into 10 fairly equal parts. In total, the model was cross-validated 10

times across 10 folds, where each fold served as the validation sample once, while the remaining samples were used for model estimation. The estimated classification error was averaged across folds to measure overall predictive performance. The final model was trained on the entire training dataset to enhance accuracy, particularly for logistic regression's correct classification rate regarding the dichotomous outcome variable. The output provided average accuracy and Cohen's kappa values, indicating model performance across folds. Accuracy denotes the proportion of correctly classified cases relative to all instances. Cohen's kappa, on the other hand, considers baseline probabilities from a null model lacking predictor variables, accommodating imbalanced outcome levels.

Fig. 3.22: model obtained with k-fold cross validation

```
Generalized Linear Model

93570 samples
    6 predictor
    2 classes: 'dissatisfied', 'satisfied'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 84213, 84212, 84214, 84213, 84214, 84213, ...
Resampling results:

  Accuracy   Kappa
  0.7682804  0.5273038
```




```
               Accuracy : 0.7694
                 95% CI : (0.7639, 0.7748)
    No Information Rate : 0.5391
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5297

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8640
            Specificity : 0.6587
         Pos Pred Value : 0.7475
         Neg Pred Value : 0.8055
             Prevalence : 0.5391
         Detection Rate : 0.4658
   Detection Prevalence : 0.6231
      Balanced Accuracy : 0.7613

       'Positive' Class : satisfied
```

The model thus obtained is identical to the one previously obtained with the stepwise and backward regression, with the same regressors and the same levels of significance. Furthermore, considering the accuracy of the model equal to 0.77 we can be satisfied with the result obtained.

3.8 - Key points and conclusions

We conclude our analyzes with a brief summary that includes the main points of our work:

1. With the PCA we have verified that there are mainly two categories of passengers: those who focus on the speed and ease of online booking and for whom the online services offered by the airline company are fundamental and those who favor more luxurious aspects of a flight. The former, therefore, may prefer shorter and last-minute flights, while the latter may prefer flights with more distant destinations and for which flight comfort is important.

2. with the k-means and hierarchical clustering method we studied the composition of the airline market by analyzing 48 of the most important airlines in the world. In this way it was possible to compare the results obtained individually by our airline company and compare them with those of the other airlines. In fact, by dividing the airlines into three clusters we discovered that our company is part of the first cluster, which is the smallest but most important as it contains the flight companies judged most positively by passengers. In this way it was possible to better study the vision that passengers have of us and have a reference benchmark.

3. Having concluded the unsupervised analysis, we moved on to the supervised ones. In this section, various statistical learning methods were applied, such as stepwise regression, Linear Discriminant Analysis, decision trees and k-fold cross validation. During these analyses, we obtained a highly predictive logistics model that allows us to identify which individuals were satisfied with the flight, and with our airline in general, and which were not. What we discovered is that women are generally more satisfied than men at the end of the flight. Furthermore, regular and loyal customers consider themselves happier than non-regular customers at the end of the flight. Continuing we discovered that young people on average are more satisfied than older people and that those who travel for work reasons are happier with their trip than those who travel for personal reasons. Furthermore, those traveling in Economy are more likely to be dissatisfied with their trip than those flying in Business. Finally, distance is inversely correlated with flight satisfaction, which means that those who fly short flights are on average more satisfied than those who make long journeys.