# Report Airline Passenger Satisfaction

Kilian Sennrich

26 3 2021

This is the report of Kilian Sennrich for the first assignment of the class "10616 Machine Learning" held by Prof. Dr. Dietmar Maringer (University of Basel). The task is to perform a complete analysis for one of the 10 datasets provided in the lecture. I chose the "Airline Passenger Satisfaction" dataset from kaggle: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction Time to complete task: 10 Days (Hand-in on April, 6., 2021)

From the Website:

## What factors lead to customer satisfaction for an Airline?

This dataset contains an airline passenger satisfaction survey. What factors are highly correlated to a satisfied (or dissatisfied) passenger? Can you predict passenger satisfaction?

Note that this dataset was modified from another dataset by John D on Kaggle. It has been cleaned up for the purposes of classification.

1. Gender: Gender of the passengers (Female, Male)
2. Customer Type: The customer type (Loyal customer, disloyal customer)
3. Age: The actual age of the passengers
4. Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
5. Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6. Flight distance: The flight distance of this journey
7. Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
8. Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
9. Ease of Online booking: Satisfaction level of online booking
10. Gate location: Satisfaction level of Gate location
11. Food and drink: Satisfaction level of Food and drink
12. Online boarding: Satisfaction level of online boarding
13. Seat comfort: Satisfaction level of Seat comfort
14. Inflight entertainment: Satisfaction level of inflight entertainment
15. On-board service: Satisfaction level of On-board service
16. Leg room service: Satisfaction level of Leg room service
17. Baggage handling: Satisfaction level of baggage handling
18. Check-in service: Satisfaction level of Check-in service
19. Inflight service: Satisfaction level of inflight service
20. Cleanliness: Satisfaction level of Cleanliness
21. Departure Delay in Minutes: Minutes delayed when departure
22. Arrival Delay in Minutes: Minutes delayed when Arrival
23. Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

# Introduction

This Report explains the analysis of the dataset. It is organized according to the main.R file, that contains the analysis. This report explains the thought process behind the analysis. Assumptions in the analysis are clearly documented here!!

# Table of contents



## Inspection, Balance, Missing Values

A first look into the data revealed, that the data is pretty clean. There were 2 columns to be deleted (V1, id) that contain unique identifiers to single observations and that don't add value to the analysis, since we do not try to make statements about single customers. The data is generally well balanced, with approx. 50 percent of the criterion (satisfaction) in both groups. Therefore, in this regard, no further steps are necessary. Missing values (NAs and sometimes NULL objects) can only be upserved in the "Arrival Delay in Minutes" column. There are 310 missing values in total, that distributed approximately equally between the two attributes of the criterion. I had the hypothesis, that the 310 values might all originate from the same plane, since an average airplane could transport about that amount of passengers. To check this i made the following assumption **If the NA values originated from passengers of the same flight, the flight distance for all these passengers would be equivalent** . The hypothesis is wrong, since the max number of passengers that had the same flight distance was 5. In order to get good data for the classifier, I wanted to get rid of the NAs. For imputation I had three ideas in mind: replacement by mean, delete rows or MICE. Mice (Multiple Imputation by Chained Equantions) seemed an over-the-top approach for only 310 missing values. Both of the other approaches would work, but i chose to omit the NA rows, since the proportion of data that falls away is very small (0.2 %) and having to many identical values (mean imputation) in the data, sometimes affects the accuracy of the model.

## Visualisation

*Nominal variables*: Note: Some variables could also be considered ordinal (E.g. Class). The genders are equally distributed in both the "neutral or dissatisfied" and in the "satisfied" box. This is a good sign, since the data is likely to be a sample of a much larger database itself. Of interest is the interaction between "Type of Travel" and "Satisfaction". Where most of the "Personal Travel" customers are dissatisfied, this cannot be said about the "Business Travel" customers. Comparing it with the plot for the class, this is a clear hint, that the environment in Business Class leads to more satisfaction. For interventions, this would be a good starting-point. *Ordinal variables*: (0 means "no answer" given). For ordinal variables, to see whether a variable has let to more satisfaction, the skewedness can be interpreted: When both "satisfied" and "neutral or dissatisfied" plots are left skewed, the variable is likely to not have a great impact on the satisfaction rating (E.g. Departure time convenient). Further, the height of the barplots can reveal information on the factors that led to dissatisfaction. The heigher the "neutral or dissatisfied" plot in oppose to the "satisfied" plot, the more an item contributes to the dissatisfaction of the customers.
*Interval variables*: Many of the plots are heavily skewed, which must be considered when implementing the classifier. What is interesting is, that especially younger people are dissatisfied. But my initial hypothesis that younger people are more likely to travel in the "Eco" class is not really valid, since the correlation is -.13 (Age ~ Class, Spearmans Rho; See Figure 4). So there must be another reason.

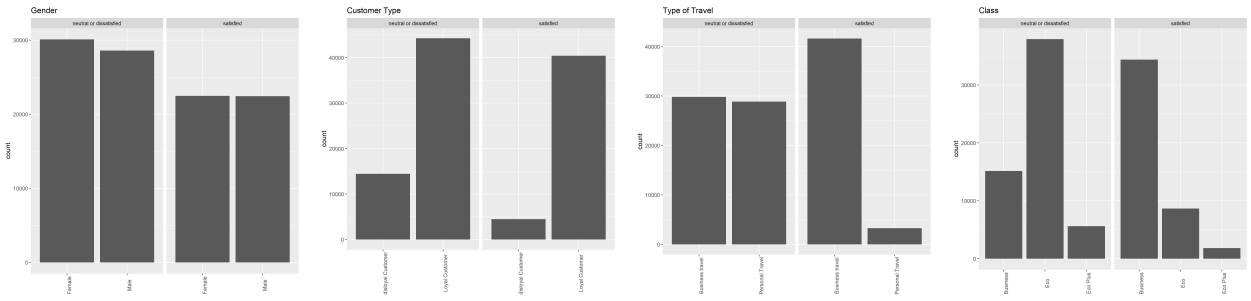



Figure 1: Visualization of all nominal variables by criterion

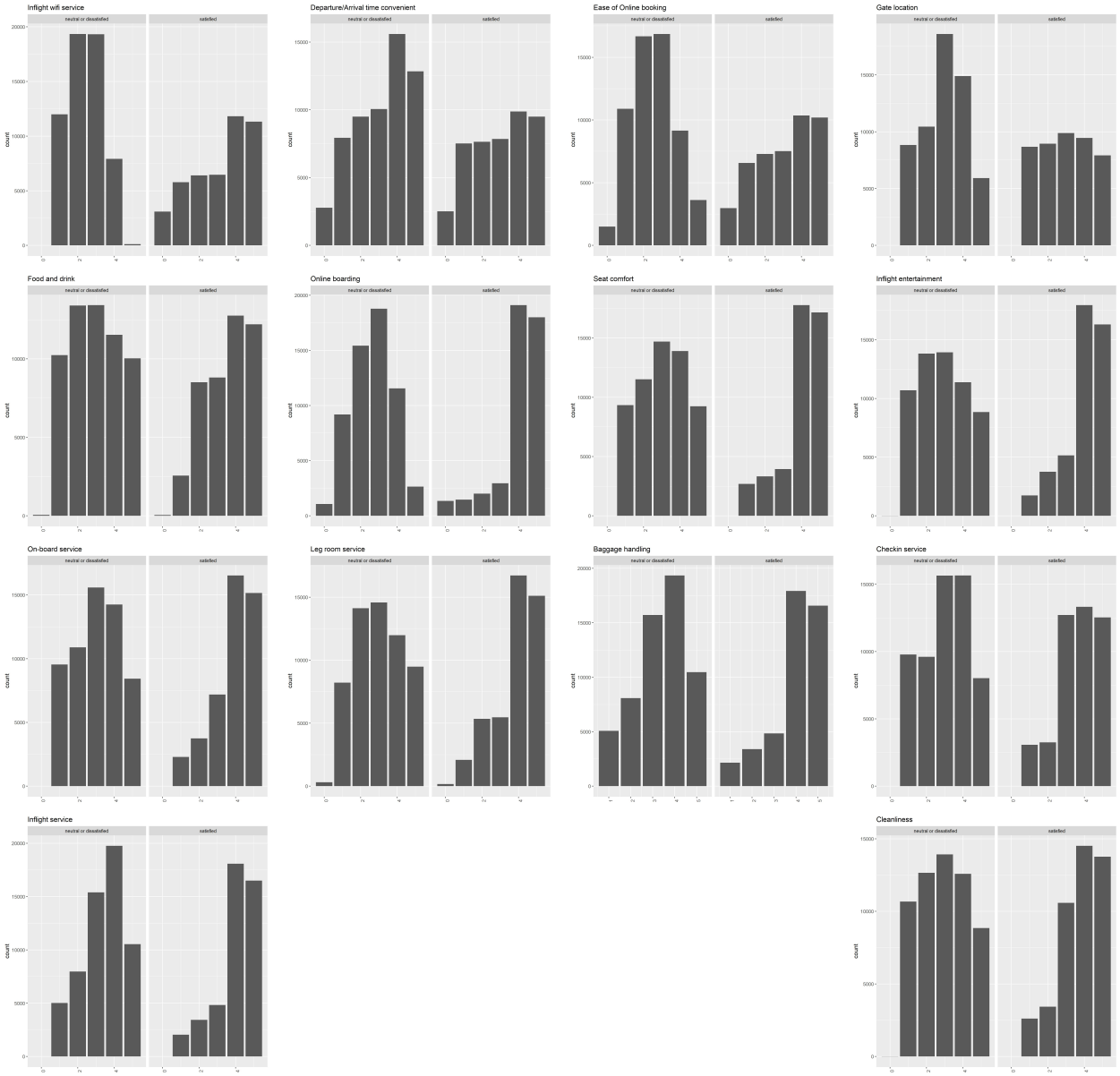



Figure 2: Visualization of all ordinal variables by criterion

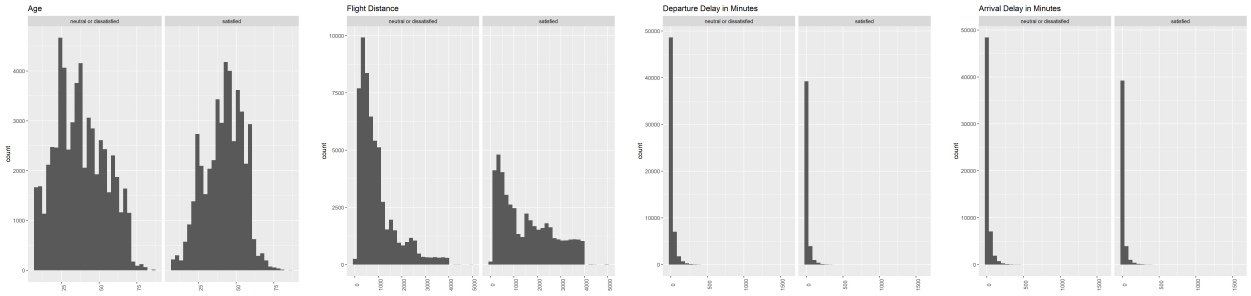



Figure 3: Visualization of all intervall variables by criterion

*Note: The plots are held really small to save space. But since the quality is really good, one can simply zoom in!*

## Correlation Analysis

To extract variables that are likely to be associated with the criterion (satisfaction), a correlation analysis is conducted. The correlation coefficient is spearmans rho, which is made for correlation of ordinal variables. Spearmans Rho, in oppose to Pearson is a rank correlation measure and is therefore, to a certain degree, also able to detect non-linear relationships. Spearmans Rho was preferred over Kendals Tau, because the dataset is rather large. Some say that Kendals Tau does better on smaller datasets. There is quite a problem with a correlation analysis since some variables are nominal. This must be considered when interpreting the results.

Another problem with the correlation analysis at hand is, that there are no significance measures included. Since there are 24 comparisons, using p-values would likely lead to alpha inflation. Also, since the dataset is rather large, even small correlation coefficinets are likely to get significant. Bonferroni correcting the p-values would not be a good idea, since there are 22 (!) single comparisons meaning alpha would be 0.05/22 = 0.002272727. A slight different approach is being chosen. The correlation coefficients are evaluated against a well known/accepted criterion of Cohen, 1988. According to this criterion, correlations above +-0.5 are considered strong, correlations between +-0.3 and +- 0.5 are considered of medium strength and correlations between +-0.1 and +-0.3 are considered of low strength. Correlations between -0.1 and +0.1 are considered as non existing. (https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/ReferencesPapers.aspx?ReferenceID=2041144) Interpretation with cohens‘ coefficients are of course not perfect and do also come with limitations, but I think that they depict the reality better than significance levels from correlation-tests.

Since spearmans rho does at least require ordinal ranks, correlating the items that code a 0 as “no answer” would lead to biased correlation coefficients (ordinal would mean, that 0 is worse than 1, which does not hold true for our data). I controlled for this problem by recoding 0 as NA and not using NAs for correlation (use = “complete.obs”). Considering that 0 does have a different meaning for “Flight Distance”, “Departure Delay in Minutes” and “Arrival Delay in Minutes”, these variables were not recoded with NA. Assumption for not considering NAs: **Choosing “no answer” is compeletely at random and does not have any relationship with satisfaction**

The results of the correlation analysis can be found in Tables 1 to 4.

According to the Cohen, 1988, online Boarding is most severely associated with satisfaction rating. A good intervention would be to further expand online checkin service.

Consult the heatmap in Figure 4 to get insights to intercorrelations. Handle with care, since not all variables in there meet the requirements of Spearmans Rho. The darker the color, the stronger the relationship.

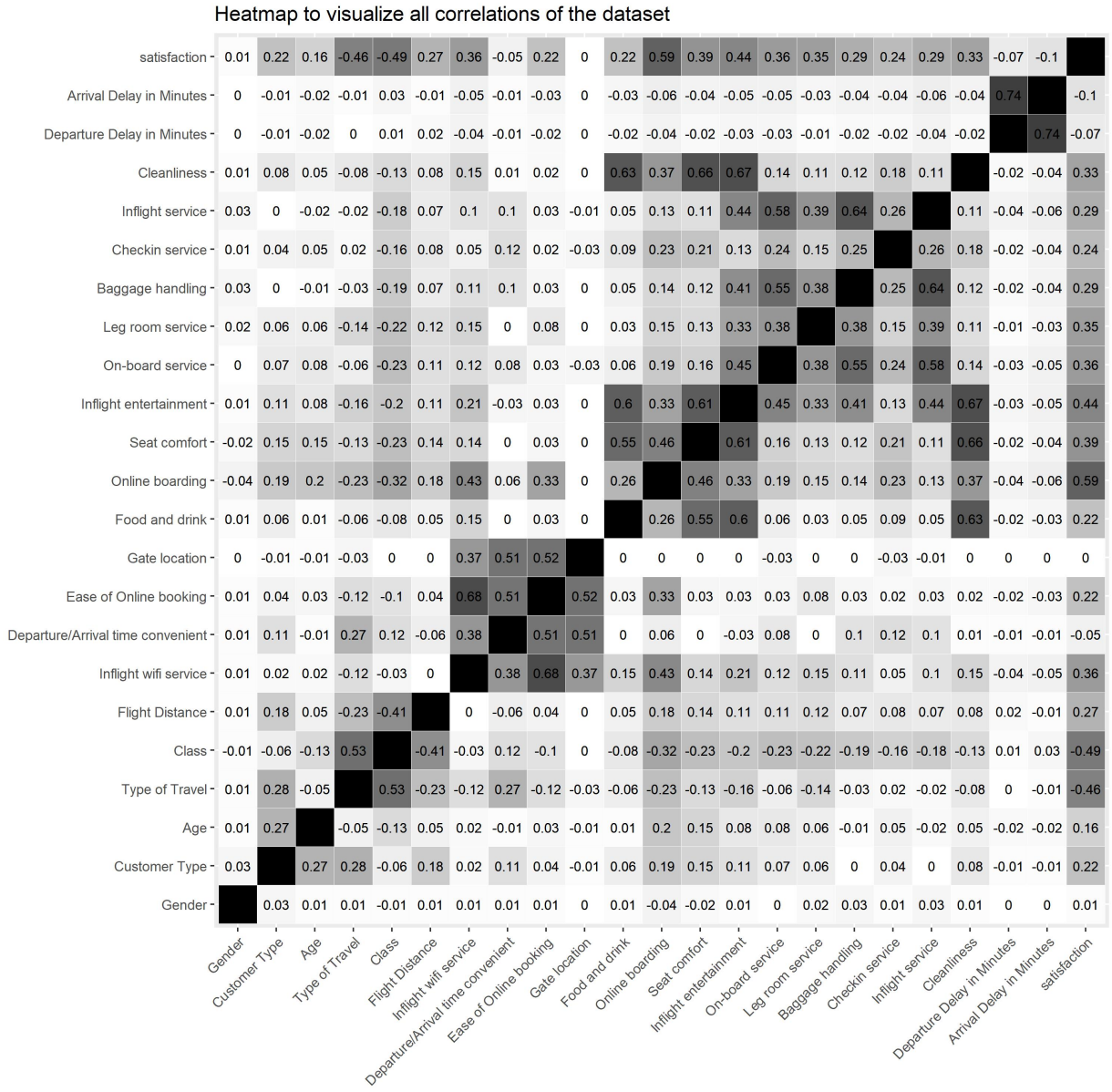



Figure 4: The darker the color, the more higher the correlation. The correlation-coefficient is Spearmans Rho. Be careful with interpretation since there were nominal variables in there!

Table 1: correlations higher than +-0.5

| y | x | cor |
|---|---|---|
| satisfaction | Online boarding | 0.5899783 |

Table 2: correlations between +-0.3 and +-0.5

| y | x | cor |
|---|---|---|
| satisfaction | Type of Travel | -0.4489949 |
| satisfaction | Class | -0.4801065 |
| satisfaction | Inflight wifi service | 0.3722062 |
| satisfaction | Seat comfort | 0.3618708 |
| satisfaction | Inflight entertainment | 0.3995540 |
| satisfaction | On-board service | 0.3283335 |
| satisfaction | Leg room service | 0.3200487 |
| satisfaction | Cleanliness | 0.3031766 |

Table 3: correlations between +-0.1 and +-0.3

| y | x | cor |
|---|---|---|
| satisfaction | Customer Type | 0.1875584 |
| satisfaction | Age | 0.1470610 |
| satisfaction | Flight Distance | 0.2573588 |
| satisfaction | Ease of Online booking | 0.2335138 |
| satisfaction | Food and drink | 0.2075063 |
| satisfaction | Baggage handling | 0.2693956 |
| satisfaction | Checkin service | 0.2323059 |
| satisfaction | Inflight service | 0.2654399 |

Table 4: items not associated with criterion (-0.1 and 0.1)

| y | x | cor |
|---|---|---|
| satisfaction | Gender | 0.0123560 |
| satisfaction | Departure/Arrival time convenient | -0.0468449 |
| satisfaction | Gate location | -0.0003662 |
| satisfaction | Departure Delay in Minutes | -0.0737981 |
| satisfaction | Arrival Delay in Minutes | -0.0735977 |

## Exploratory Factor Analysis

Since the task at hand asks for the factors that lead to customer satisfaction, it would be interesting to see, whether some variables/predictors are driven by "higher" factors. The appropriate approach here would be to conduct an exploratory factor analysis, but this is not possible since some variables are qualitative and various other assumptions are not met. The approach chosen here is to conduct a Factor Analysis for Mixed Data (FAMD), since this approach is capable of working with qualitative and quantitative variables. The FAMD was carried out using the FactoMineR and the factoextra package. A tutorial can be found here: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-

guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/ FAMD is not really a Factor Analysis but rather a principal component analysis. But due to the high similarity of the two, the interpretation is rather similar. What has to be said though is, that in pca, the Dimensions are, unlike the factors in Factor Analysis, not uncorrelated. This must be considered when interpreting the results.

Table 5: Explained portion of variance by the principal components

| | eigenvalue | variance.percent | cumulative.variance.percent |
|---|---|---|---|
| Dim.1 | 4.095675 | 17.807284 | 17.80728 |
| Dim.2 | 2.374696 | 10.324765 | 28.13205 |
| Dim.3 | 2.204845 | 9.586283 | 37.71833 |
| Dim.4 | 1.971187 | 8.570377 | 46.28871 |
| Dim.5 | 1.836173 | 7.983361 | 54.27207 |

The results of the FAMD can be seen in Table 5 and in Figure 5. The results are not really great. Capturing a total of 17.8 percent of the total variance most likely leads to the interpretation that there are not really overarching factors. But still, a look at the variables that load primarely onto the first 4 dimensions cannot be a bad idea (Figure 6).

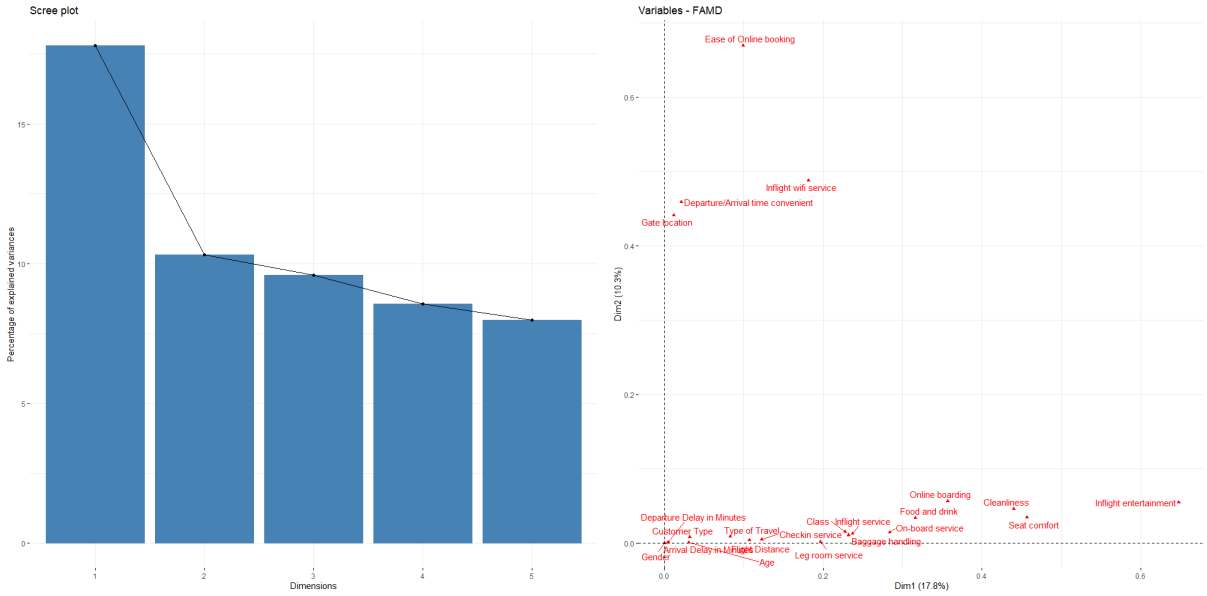



Figure 5: Left: Screeplot of FAMD: Traditionally one would only accept dimension 1, but variance on this dimension is pretty small; Right: The variables plotted on the first two dimensions

As stated above the validity of an interpretation of the dimensions is vastly in doubt. However, here are possible factors that the first two dimensions might cover (Note again that the dimensions are NOT uncorrelated and some items load on more than one dimension):

- **Dimension 1**: Inflight entertainment, Seat comfort, Cleanliness, Online boarding, Food and drink, On-board service, Inflight service, Baggage handling, Class, Leg room service. **Interpretation**: These are mostly variables that concern inflight services and inflight personal comfort.
- **Dimension 2**: Ease of online booking, Inflight wifi service, Arrival/Departure time convenient, Gate location. **Interpretation**: These are mostly variables that concern factors that surround the flight.

Conclusion: FAMD is not really useful here and no conclusion can be drawn from the results.

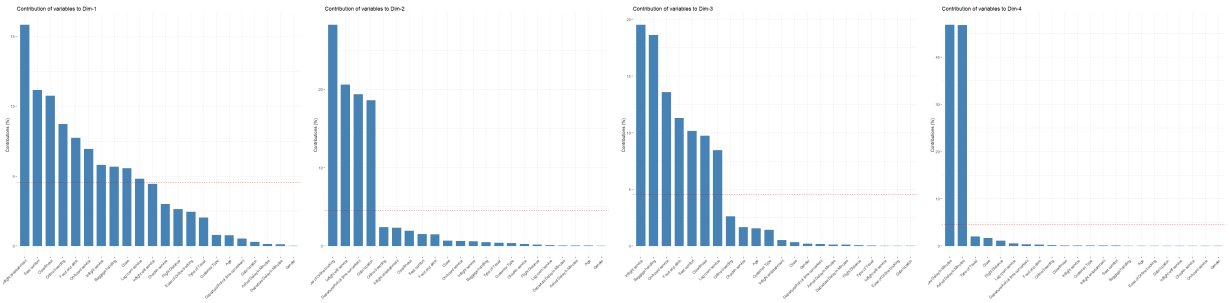



Figure 6: Zoom in to see more. The variables that load to the first 4 dimensions. The red dashed line on the graph above indicates the expected average value, if the contributions were uniform.

## Machine Learning for Deeper Insights.

With the dataset and the task at hand, it is actually not ideal to use a model that maximizes accuracy but is not interpretable. However, the assignment, this report is for, requires to maximize accuracy, so the modeling here is split into two parts: First there will be a discussion of decision trees, which also includes in interpretation for the task. Second there will be an accuracy optimizing part, where ensemble methods (all treebased) were implemented to maximize predictive validity. These models cannot directly be used for interpretation, since they are blackbox models (at least the boosting method).

### Cross Validation

The data was first split into a training dataset and a testing dataset (80% vs. 20%). The training dataset was further split into ten folds. Therefore, to improve model robustness **10-fold cross validation** was used. After the model was fit to the data, it got benchmarked on the training dataset (in-sample) and the testing dataset (out-of-sample). The trained model has therefore never seen the data from the testing dataset. The 10-fold cross validation technique could have been further extended with a repeated cross validation. However, since I wanted to use the same validation technique for all my models and because the ensemble methods were already computationally expensive themselves, I did decide against it. (10 folds repeated 10 times would lead tho 100 models that had to be fit for every single hyperparameter combination)

All models were fitted in parallel. All models were implemented using the caret package.

### Decision Tree Analysis

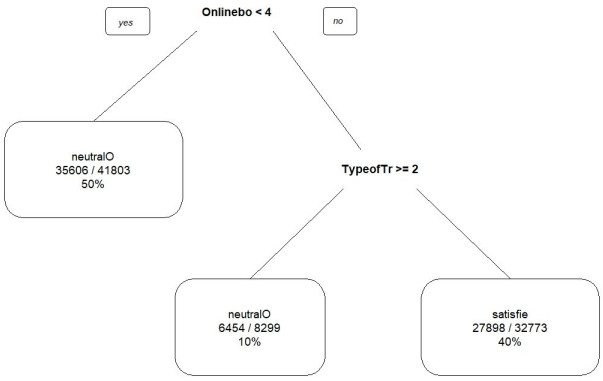
A decision tree seems appropriate for the task. Decision trees have several benefits: They are insensitive to outliers (so no outlier detection was carried out), they are insensitive to skewed data (so no transformations were used) and they work well with mixed data (nominal, ordinal and interval). Further, decision trees are easily to interpret, which is important for finding interventions for the airline company. It is also good to mention, that decision trees exploit variables by their predictive validity and do automatically exclude variables, that are not of further use. Therefore, no variables were excluded as predictors beforehand. The final tree provided an accuracy of 0.8441 in-sample and 0.8407 out-of-sample, which is acceptable considering the a priori probability of 0.56. The mean accuracy of the resampling in the cross validation is 0.85 (this is the value that can be generalized/that has the most validity).

Since decision trees exploit variables as much as possible, they contain information on the importance of the variables. The variables were ranked as follows:

Table 6: Variables with the most importance

| Rank | Variable |
|---|---|
| 1 | Online Boarding |
| 2 | Seat Comfort |
| 3 | Inflight Wifi Service |
| 4 | Type of Travel |
| 5 | Class |

Online Boarding (neutral or dissatisfied when Online Boarding $< 4$) and Type of Travel (Personal Travel (2) leads to neutral or dissatisfied) represent the two primary nodes. This can be used to make the decision about what the intervention will look like: Improving Online Boarding would probably lead to great improvement of the satisfaction rating. Type of Travel might mean, that the conditions for personal travel are not really good, interventions in this regard would also improve the satisfaction rating. The output (Figure 7) further shows, that when customers travel in the Eco class, they state that they are dissatisfied. Improving conditions in this class, would therefore lead to better satisfaction ratings.




```
Node number 1: 82875 observations,    complexity param=0.5123539
  predicted class=neutralOrDissatisfied  expected loss=0.4336652  P(node) =1
    class counts: 46935 35940
   probabilities: 0.566 0.434
  left son=2 (41803 obs) right son=3 (41072 obs)
  Primary splits:
      Onlineboarding        < 3.5  to the left,  improve=13743.290, (0 missing)
      Class                 < 1.5  to the right, improve=10334.700, (0 missing)
      TypeofTravel          < 1.5  to the right, improve= 8224.773, (0 missing)
      Inflightentertainment < 3.5  to the left,  improve= 7020.021, (0 missing)
      Inflightwifiservice   < 3.5  to the left,  improve= 6812.699, (0 missing)
  Surrogate splits:
      Seatcomfort           < 3.5  to the left,  agree=0.731, adj=0.458, (0 split)
      Inflightwifiservice   < 3.5  to the left,  agree=0.715, adj=0.425, (0 split)
      EaseofOnlinebooking   < 3.5  to the left,  agree=0.675, adj=0.345, (0 split)
      Class                 < 1.5  to the right, agree=0.671, adj=0.336, (0 split)
      Inflightentertainment < 3.5  to the left,  agree=0.660, adj=0.313, (0 split)

Node number 2: 41803 observations
  predicted class=neutralOrDissatisfied  expected loss=0.1482429  P(node) =0.5044103
    class counts: 35606  6197
   probabilities: 0.852 0.148

Node number 3: 41072 observations,    complexity param=0.1282415
  predicted class=satisfied               expected loss=0.2758327  P(node) =0.4955897
    class counts: 11329 29743
   probabilities: 0.276 0.724
  left son=6 (8299 obs) right son=7 (32773 obs)
  Primary splits:
      TypeofTravel          < 1.5  to the right, improve=5238.846, (0 missing)
      Class                 < 1.5  to the right, improve=3788.172, (0 missing)
      Inflightentertainment < 3.5  to the left,  improve=2418.106, (0 missing)
      Legroomservice        < 3.5  to the left,  improve=2397.701, (0 missing)
      Onboardservice        < 3.5  to the left,  improve=1610.287, (0 missing)
  Surrogate splits:
      Age            < 60.5 to the right, agree=0.815, adj=0.084, (0 split)
      Class          < 1.5  to the right, agree=0.813, adj=0.073, (0 split)
      Legroomservice < 0.5  to the left,  agree=0.803, adj=0.023, (0 split)
```

Figure 7: Left: Decision tree: For the split with TypeofTravel, the Class-variable could also be used; Right: Extensive overview on the decision rules

**Improving Accuracy With Ensemble Methods.**

For improving accuracy, I still wanted to stick to tree based methods. Therefore, I used Random Forest as a bagging approach and Gradient Boosting Sequential Trees as a boosting approach.

**Random Forest** The random forest was estimated using the Rborist package in caret which holds a very efficient approach to random forests (Breiman, L. (2001). Random forests. Mach Learn, 45:5-32. https://doi.org/10.1023/A:1010933404324.). To further improve the speed of the estimation, I did not use the formula specification but rather entered the predictors and the criterion directly. This is recommended in various forums. The Rborist function holds two parameters that were automatically tuned. Therefore, no further specification was needed. The best values are: predFixed = 22 and minNode = 2. The random

forest was able to get an accuracy of 0.9734 in sample and an accuracy of 0.9745 out of sample. The mean accuracy of the resampling in the cross validation is 0.91 (this is the value that can be generalized/that has the most validity).

**Gradient Boosting Sequential Trees.** Gradient Boosting Sequential Trees were implemented using the gbm package in caret. The model has 4 parameters, two of them are being hold constant by default. To vary them as well, I specified custom values for all parameters. The best values estimated by the model for the 4 parameters are: n.trees: 250, interaction.depth = 7, shrinkage = 0.15 and n.minobsinnode = 15. For all parameters the maximum value specified in advance (in the gradientboostgrid) was chosen. I think that accuracy could be further improved by allowing the model higher values in all parameters. However, higher values, for example, in the n.trees parameter will increase computation times by quite a bit, which is why I have not re-estimated the model. The Gradient Boosting Sequential Trees were able to achieve an accuracy of 0.9624 in sample and an accuracy of 0.9615 out of sample. Therefore, they benchmark slightly worse than the random forest (not considering the resampling mean). I assume that with better parameters, they could beat the random forest. However, the mean accuracy of the resampling in the cross validation is 0.96 (this is the value that can be generalized/that has the most validity), which is better than the random forest.

**Model Selection.** For selecting the best model, boxplots of the 10 folds (from cross validation) of all tree models were compared. For both Kappa and Accuracy, the Gradient Boosting Sequential Trees seem to be slightly better than the random forest. **So this means that the Gradient Boosting Sequential Tree Model is my best model and my final choice**

```
Call:
summary.resamples(object = comparison)

Models: Decision_Tree, Random_Forest, Gradient_Boost
Number of resamples: 10

Accuracy
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Decision_Tree  0.8345801 0.8415433 0.8458521 0.8503412 0.8601424 0.8673826    0
Random_Forest  0.9125483 0.9159970 0.9175153 0.9172635 0.9187200 0.9201737    0
Gradient_Boost 0.9582046 0.9594344 0.9600830 0.9604031 0.9608099 0.9631274    0

Kappa
                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Decision_Tree  0.6592622 0.6738165 0.6828167 0.6926136 0.7138414 0.7282017    0
Random_Forest  0.8208867 0.8280138 0.8311041 0.8305886 0.8335940 0.8367199    0
Gradient_Boost 0.9145930 0.9171089 0.9183780 0.9190702 0.9199371 0.9246029    0
```

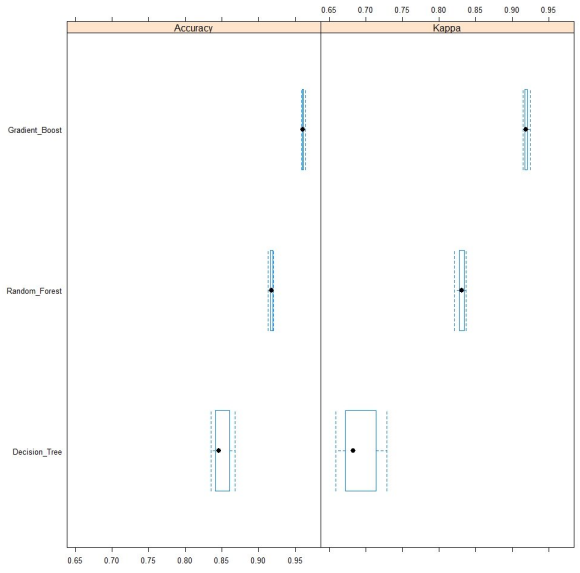



Figure 8: Left: Summary of the comparison, Right: Visualisation of the comparison