Study for Airline Passenger Satisfaction with the open data analysis software - KNIME and Microsoft Power BI

Fotakelis Chronis

July 2022

Synopsis

Technological developments in the field of information technology have provided useful data analysis tools, such as the KNIME and Microsoft Power BI for the approach and study of problems in many sciences and everyday life. This paper aims to present the analysis and modeling of a data set from the Maven Airline Passenger Satisfaction Challenge which took place from the 18th of May 2022 to 24th of June 2022. Customer satisfaction scores from 129,000+ airline passengers, including additional information about each passenger about their flight, and type of travel, as well as their evaluation upon different factors like cleanliness, comfort, service, and overall experience.

1. General Terms

Modelling, Simulation, Data Analysis, Neural Networks, Linear Regression, Decision Trees, Linear Regression Trees.

Keywords

KNIME, Naïve Bayes, Decision Trees, Backpropagation

2. Contribution

With this research paper the user of KNIME and Microsoft Power BI software comes into contact with the study of modeling and simulation and pulling out useful insights using the data file used in the research work of John D and can be accessible at Kaggle's repository. The data file is free for the public and can be found in the following url: https://cutt.ly/UZoK7wO

3. Introduction

Airline industry faced tremendous challenges the past two years in terms of financial terms because of covid-19 and not only. Beside of course economic factors airline businesses must also take under account the satisfaction of their passengers and the factors which influence the pleasure of travelling with their airplane company. Utilizing tools such as Microsoft Power BI and KNIME we have the opportunity for deeper knowledge and analysis of data and drawing critical conclusions. In particular, in this study, correlation techniques between data file variables were used, linear regression techniques, prediction algorithms for satisfied or not satisfied passengers were also applied. (Naive Bayes Algorithm, Decision Trees, Artificial Neural Networks)

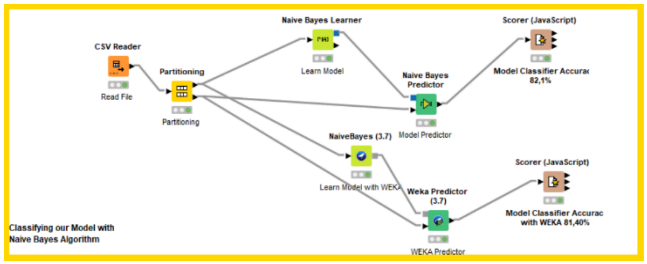


Figure 1. Sample Stage of studying the model

4. Model Presentation

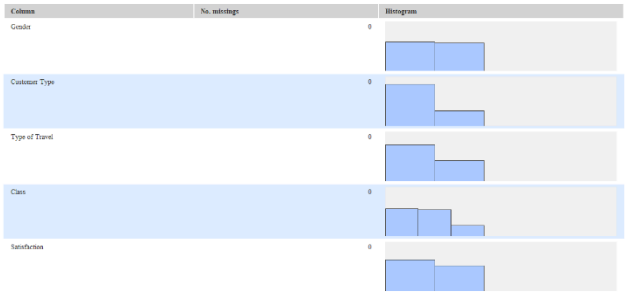
4.1 Variables

The data set consists of 18 numeric and 5 categorical variables. The "Satisfaction" attribute can be used as a class tag. "Age" represents the actual age of the passengers, "Gender" represents the gender of the passengers (Female, Male), "Type of Travel" represents the purpose of the flight of the passengers (Personal Travel, Business Travel), "Class" represents the travel class in the plane of the passengers (Business, Eco, Eco Plus), "Customer Type" represents the customer type (Loyal customer, disloyal customer), "Flight distance" represents the flight distance of this journey, "Inflight wi-fi service" showing satisfaction level of the inflight wi-fi service (0:Not Applicable;1-5), "Ease of Online booking" indicating satisfaction level of online booking, "Inflight service" indicating satisfaction level of inflight service, "Online boarding" representing satisfaction level of online boarding, "Inflight entertainment" indicating satisfaction level of inflight entertainment, "Food and drink" appearing for satisfaction level of Food and drink, "Seat comfort" symbolizing satisfaction level of Seat comfort, "On-board service" representing satisfaction level of On-board service, "Leg room service" appearing for satisfaction level of Leg room service, "Departure/Arrival time convenient" indicating satisfaction level of Departure/Arrival time convenient, "Baggage handling" representing satisfaction level of baggage handling, "Gate location" indicating satisfaction level of Gate location, "Cleanliness" appearing for satisfaction level of Cleanliness, "Check-in service" representing satisfaction level of Check-in service, "Departure Delay in Minutes" representing minutes delayed when departure, "Arrival Delay in Minutes" indicating minutes delayed when "Arrival Flight cancelled" whether the Flight cancelled or not (Yes, No), "Flight time in minutes" representing the flight duration in minutes.

Finally, the data set consists of features belonging to 129,880 passengers. The data contains 389 missing values, especially from the column of "Arrival Delay" and there has been an effort to handle those values. The features of our data set in Figures 2, 3 show the statistical observation tables of the numeric and nominal variables.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Age	7	39,428	?	85	15,1194	-0,0036	-0,7191	0	0	0	
Flight Distance	31	1.190,3164	?	4.983	997,4525	1,1081	0,2655	0	0	0	
Departure Delay	0.0	14,7137	?	1.592	38,0711	6,822	100,6445	0	0	0	
Arrival Delay	0.0	15,0455	?	1.584	38,4164	6,679	95,3551	0	0	0	
Departure and Arrival Time Convenience	0.0	3,0576	?	5	1,5267	-0,3325	-1,0409	0	0	0	
Ease of Online Booking	0.0	2,7569	?	5	1,4017	-0,0188	-0,9135	0	0	0	
Check-in Service	0.0	3,3063	?	5	1,2662	-0,3666	-0,8299	0	0	0	
Online Boarding	0.0	3,2526	?	5	1,3507	-0,4569	-0,6987	0	0	0	
Gate Location	0.0	2,9769	?	5	1,2785	-0,0383	-1,0316	0	0	0	
On-board Service	0.0	3,383	?	5	1,2871	-0,4213	-0,8888	0	0	0	
Seat Comfort	0.0	3,4414	?	5	1,3193	-0,4858	-0,9228	0	0	0	
Leg Room Service	0.0	3,3509	?	5	1,3163	-0,3484	-0,983	0	0	0	
Cleanliness	0.0	3,2863	?	5	1,3137	-0,3009	-1,0148	0	0	0	
Food and Drink	0.0	3,2048	?	5	1,3299	-0,1551	-1,1454	0	0	0	
In-flight Service	0.0	3,6422	?	5	1,1767	-0,6916	-0,3583	0	0	0	
In-flight Wifi Service	0.0	2,7287	?	5	1,3293	0,0405	-0,8486	0	0	0	
In-flight Entertainment	0.0	3,3581	?	5	1,334	-0,3664	-1,0611	0	0	0	
Baggage Handling	1	3,6321	?	5	1,18	-0,6774	-0,3838	0	0	0	

Figures 2. Statistical data analysis for Numeric Variables



Column	No. missings	Histogram
Gender	0	
Customer Type	0	
Type of Travel	0	
Class	0	
Satisfaction	0	

Figure 3. Statistical analysis of Nominal Variables

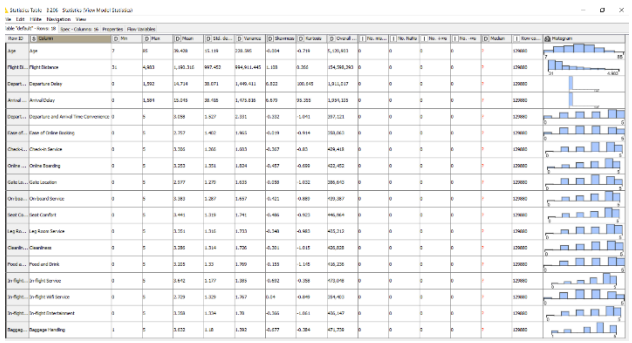


Figure 4. Statistical analysis of model variables

4.2 Results – Variable Correlations

The study began by investigating the correlations between the variables in relation to the Satisfaction variable. After examination in Figure 5 a correlation matrix table has been created to show all possible correlations. In addition, we observe that the Satisfaction variable is correlated with the most of the variables.

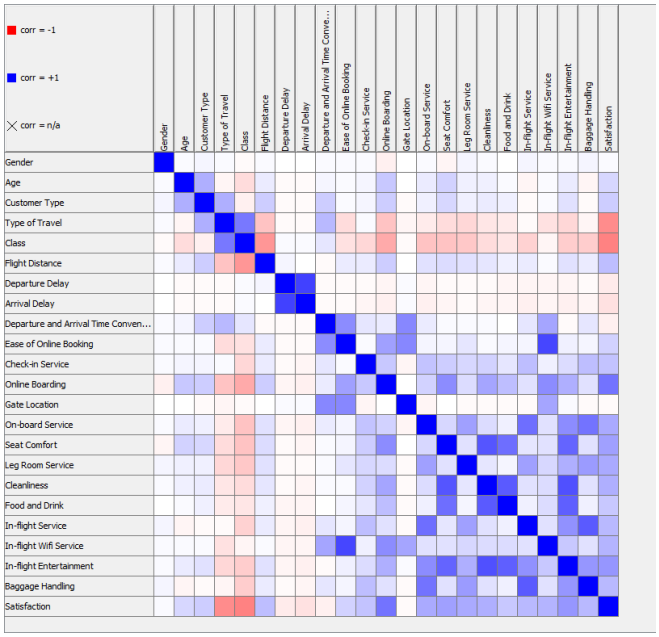


Figure 5. Correlation Matrix

5. Useful Insights for our data set with KNIME Software

After the findings from the correlation matrix table for all variables there has been an effort to explore certain metrics for our data set. These metrics are shown to Figure 6.

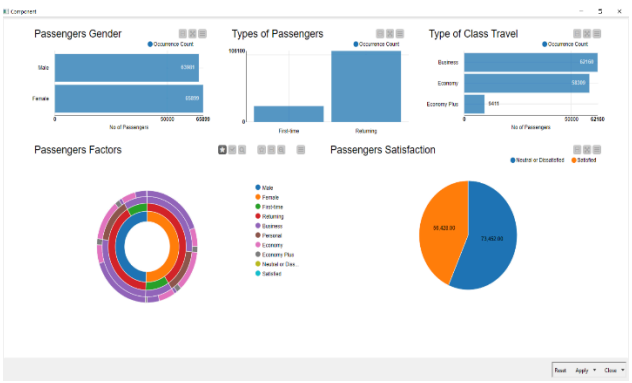


Figure 6. KNIME Dashboard Analytics

5.1 Linear Regression

The correlation among the variables was also confirmed by the linear regression study in parallel with the WEKA linear regression tool. The results were "moderate" for the linear correlation. The coefficient R^2 after examination had values 0.529 (Figure 9) and 0.545 (Figure 8) respectively. Linear regression was then applied in combination with Linear Regression Tree, where no improvement was observed with the coefficient R^2 value in 0.214 (Figure 10). Moreover, the coefficient R^2 in the linear regression tree was improved after all with a value of R^2 0.779 (Figure 11), which also proves linear correlation between the variables.

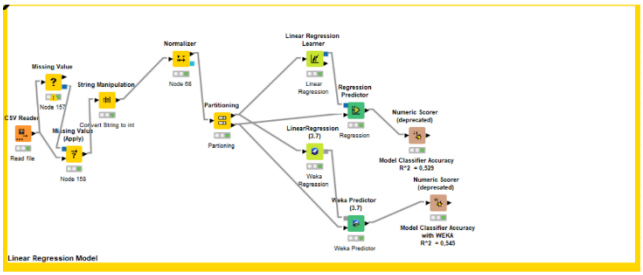


Figure 7. Application of Linear Regression and Weka Linear Regression tool

Statistics - 3:63 - Numeric Scorer (deprecated) (Model Classifier Accuracy)

File Edit Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Prediction (Satisfa...
R^2	0.545
mean absolute error	0.25
mean squared error	0.112
root mean squared deviation	0.335
mean signed difference	-0.003

Figure 8. Accuracy of Weka Linear Regression Model tool

Statistics - 3:62 - Numeric Scorer (deprecated) (Model Classifier Accuracy)

File Edit Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Prediction (Satisfaction)
R^2	0.529
mean absolute error	0.257
mean squared error	0.116
root mean squared deviation	0.34
mean signed difference	0.007

Figure 9. Accuracy of Linear Regression Model

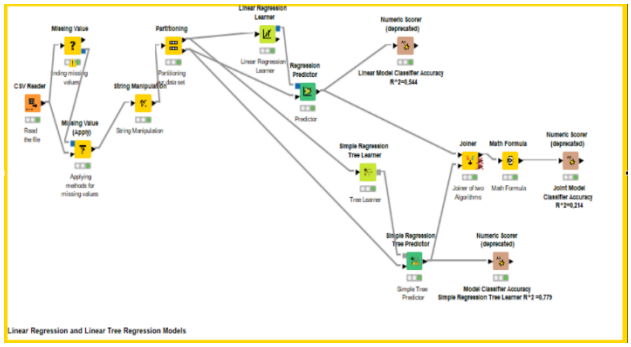


Figure 10. Application of Linear Regression Model and Linear Regression Tree Model

Statistics - 3:10 - Numeric Scorer (deprecated) (Model Classifier Accuracy)

File Edit Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Predictio...
R^2	0.779
mean absol...	0.054
mean squa...	0.054
root mean ...	0.233
mean signe...	0.005

Figure 11. Accuracy of Linear Regression Tree Model

6. Neural Networks

In addition to the above techniques for machine learning techniques the Backpropagation algorithm was applied (Figure 12). In order to train the data model and then to be classified, the Satisfaction classification class (column) of our data file was used. Once the data was normalized, 80% of it was used for training and the remaining 20% to check if it was properly classified according to the Satisfaction class.

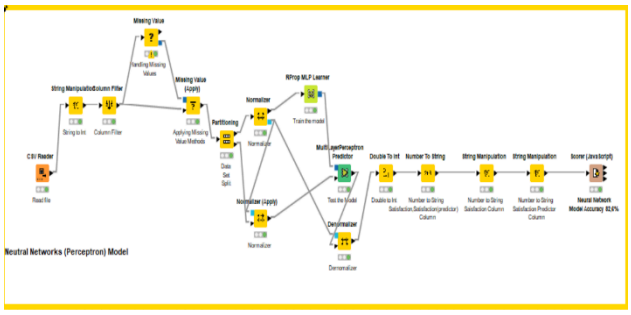


Figure 12. Back Propagation Train Model

The results of our trained model showed that it was trained with a high accuracy close to 82.6% (Figure 13).

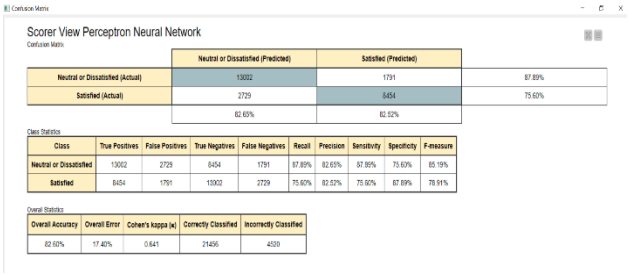


Figure 13. Accuracy and performance Matrix Table of trained Model with Back Propagation

7. Classifying and Predictive Algorithms

7.1 Naïve Bayes Algorithm

In this study of the particular data set, an attempt was made to study its performance also in other classification algorithms. So, we decided to use the Naïve Bayes classification algorithm and the same algorithm from the add-on tool from WEKA tool using the classification class Satisfaction (column). For this reason, we kept the 80% of the total data for model training and the remaining 20% to verify if our model has been properly trained. The model that was built is shown in Figure 14.

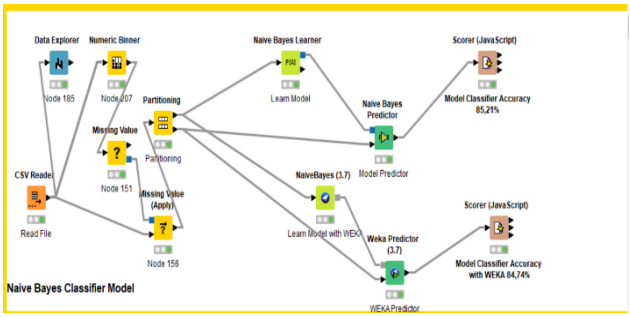


Figure 14. Naïve Bayes trained Model

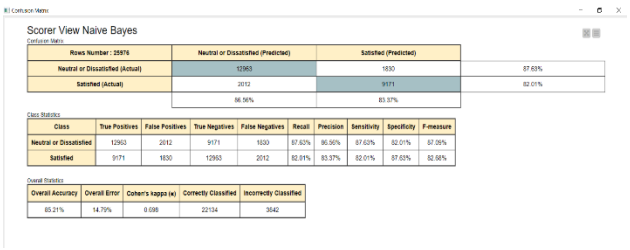


Figure 15. Accuracy and Performance Matrix of our Model with Naïve Bayes Algorithm

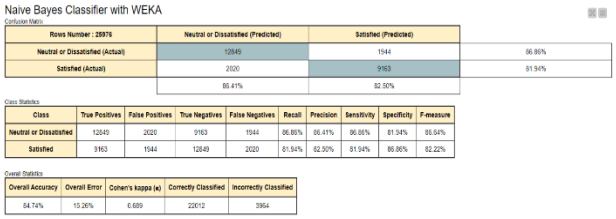


Figure 16. Accuracy and Performance Matrix of our Model with Naïve Bayes Algorithm from WEKA add-on tool

After the training process followed the test accuracy of our model. As can be seen from the table in Figure 15, the model correctly classified 12,963 cases in the class **No (Neutral or Dissatisfied)** and 9,171 cases in the class **Yes (Satisfied)**, resulting in a forecast accuracy of 85.21%. Similar results were found when the add-on was used by the WEKA tool for the Naïve Bayes algorithm. In this version of our research the created model correctly classified 12,849 cases in the class **No (Neutral or Dissatisfied)** and 9,163 cases in the class **Yes (Satisfied)** having this result of a forecast accuracy of 84.74% as shown in the table in Figure 16.

7.2 Classification with Decision Trees

Next step to our research, was an attempt to study our data further with a classification model based on decision trees and the C4.5 algorithm. In this particular implementation, first our data were normalized and then with a classification class (column) Satisfaction the 80% of the total data was used for model training and the remaining 20% to determine if the model was properly trained. As shown in Figure 17, the model created correctly classified 14,108 cases in the class **No (Neutral or Dissatisfied)** and 10,530 cases in the class **Yes (Satisfied)**, resulting in a forecast accuracy of 94.849% very high beyond the performance index of our model when the Backpropagation algorithm was used earlier (yield 85.21%).

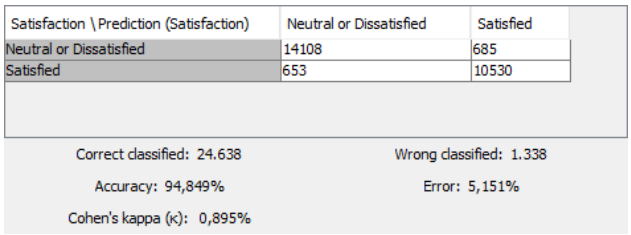


Figure 17. Accuracy and performance matrix of our model with Decision Trees (C4.5 algorithm)

8. Studying the data set with Microsoft Power BI

After analyzing our data file with the KNIME tool, we decided to step into further examination of our data in order to gain useful insights with Microsoft Power BI.

Firstly, we found that were 129,880 passengers examined (Figure 17). After examination of our data set, we found 389 missing values in 'Arrival Delay' column and we treated in a way for better results. In addition, there has been also a change in 'Age' column of airline passengers in order to present four age-segments. At this point we have also tried to create a correlation matrix table in order to examine the correlation of the class Satisfaction with the other variables (Figure 18). Our findings were similar with the KNIME analysis which preceded above.

The second phase of this research was to present useful insights of our model. Firstly, about the Satisfaction factor it can been seen that Neutral or Dissatisfied passengers are nearly up to 73,450 and Satisfied passengers are 56,430 (Figure 17). Moreover, our sample consists of 50,74% female passengers and 49,26% male

passengers (Figure 17). Another issue about our passengers were the type of travel with the certain airline. It can be seen that 69.06% of our passengers are travelling for business purposes and for personal reasons the 30.94%. About the 'Age' column after the segmentation process of this particular column we found that the majority of our passengers are between 25 to 64 years old (76.41%) followed by the age group of 15-24 with 14,01%. As far as with the type of the profile of our passengers we found that 81.69% belong s to "Returning" passengers and 18.31% are "First-time" passengers (Figure 18).

On the other hand, we must take under consideration the following thoughts also. The dissatisfied customers are close to 73,450. The 39.07% of them are from age group 25 - 64, they are travelling for personal reasons in 28.75% and 28.97% percent of them are male passengers. Moreover, the company must take under consideration the low performance according to passengers opinion about easy online-booking, online boarding, average gate location, in-flight services and as well as the in-flight Wi-Fi services. Also, when passengers class is Economy is more likely for the passengers to be dissatisfied. In addition, from dissatisfied customers 42.63% percent are returned passengers and 27.81% are travelled for personal reasons (Figure 18).

9. References

[1] Gorzałczany, Marian B., Filip Rudziński, and Jakub Piekoszewski. (2021). Business Intelligence in Airline Passenger Satisfaction Study—A Fuzzy-Genetic Approach with Optimized Interpretability-Accuracy Trade-Off. *Applied Sciences* 11, no. 11: 5098. https://doi.org/10.3390/

[2] Suki, N. (2014). Passenger satisfaction with airline service quality in Malaysia: A structural equation modeling approach. *Research In Transportation Business & Management*, *10*, 26-32. https://doi.org/10.1016/j.rtbm.2014.04.001

10. Microsoft Power Bi Model Figures

The following figures (Figure 17-18) represent our dashboard analytics for our model.

Figure 18. Airlines Dashboard Analytics

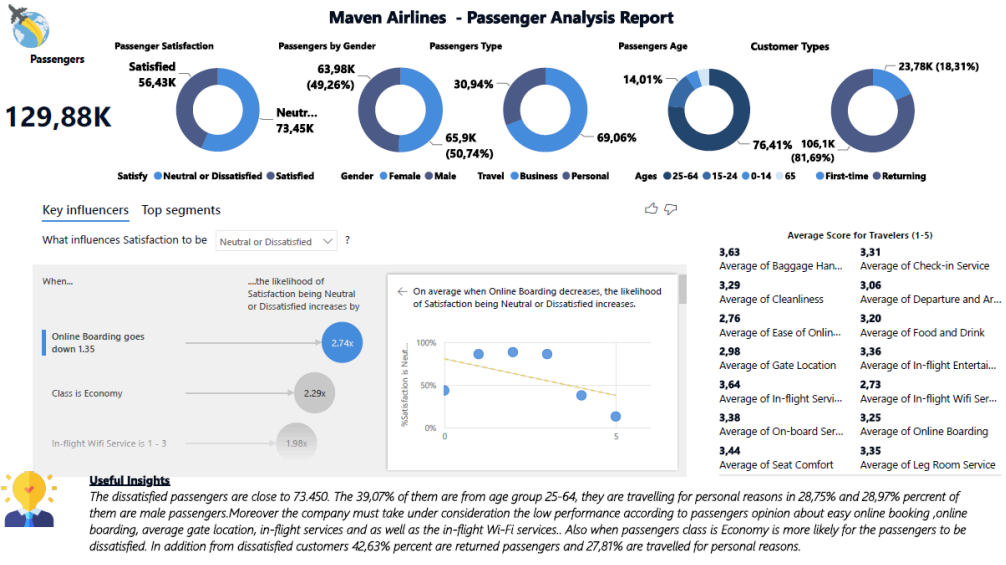
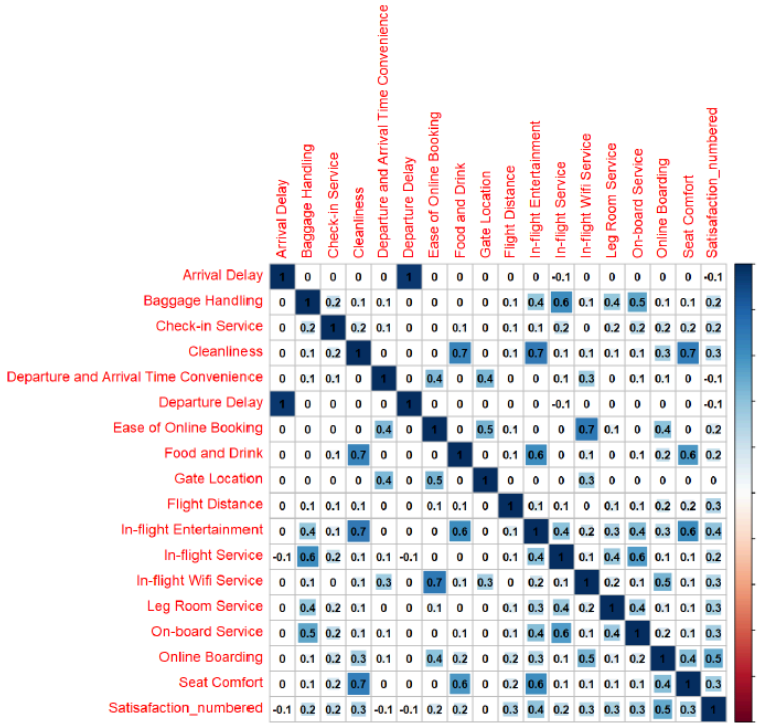



Figure 19. Correlation Matrix with Power BI