Report

HEC LIÈGE
Management School - Liège Université

Flight Prices Analysis

Data Management

Project in Data Management with a specialisation in Data collection, Data cleaning, Data visualisation and Modelling.

Student 1: Adelia Rahmawati Safitri
Student 2: Céline Lam
Student code 1: s175873
Student code 2: s191578

Course: ECON2306-1
Lecturer: Malka Guillot
June 2023

Abstract

The goal of this project is to develop a model that can accurately predict the prices of round-trip airline tickets from Lufthansa and Swiss airlines to São Paulo and New York City. By determining the optimal time to purchase tickets for each destination, it is hoped that significant savings can be achieved, allowing for the maximisation of the travel budget. To accomplish this, historical data on ticket prices and other relevant factors such as holidays will be analysed in order to identify patterns and trends that can be used to forecast future ticket prices. This project is expected to provide valuable experience in working with time series data and developing predictive models, while also uncovering practical ways to save money on air travel during peak travel times.

Keywords: Destinations, Historical data, Ticket prices, Patterns, Trends, Forecasting, Time series data, Predictive models

1. Introduction

As students, we are constantly seeking opportunities to maximise our limited budgets. One area where we identified potential for cost savings is air travel, particularly during peak travel times such as holidays. During these periods, ticket prices can increase significantly, making it challenging for us to afford travel for visiting family or taking vacations.

With this in mind, we decided to undertake a project to develop a model capable of predicting the prices of round-trip tickets from Lufthansa and Swiss airlines to São Paulo and New York City. By analysing historical data on ticket prices and other relevant factors such as holidays, we aim to identify patterns and trends that can assist us in forecasting future ticket prices. Utilising this information, our goal is to determine the optimal time to purchase a ticket for each destination, allowing us to save money and make the most of our travel budget.

Through this project, we hope to gain valuable experience in working with time series data and developing predictive models while also discovering practical ways to save money on air travel during peak travel times. By focusing on two specific airlines and destinations, we aim to develop a model that is both accurate and applicable to our travel needs.

1.1 Research questions/hypotheses

The objective of this study is to create a model that can predict the prices of round-trip tickets for Lufthansa and Swiss flights to Sao Paulo and NYC. The first hypothesis posits that historical ticket prices for these flights exhibit identifiable patterns and trends. By examining ticket price data over time, it is anticipated that these patterns can be discerned and used to forecast future prices. To evaluate this hypothesis, historical ticket price data and other relevant factors will be analyzed to identify patterns and trends. This analysis will employ statistical methods to reveal relationships between ticket prices and other variables such as the time at which flight ticket data is collected.

The second hypothesis proposes that the time at which flight ticket data is collected significantly impacts ticket prices. It is anticipated that collecting data at certain times during the week will result in lower ticket prices.

The third hypothesis asserts that a model trained on historical data can accurately predict future ticket prices. To test these hypotheses, a predictive model will be created using machine learning techniques to forecast future ticket prices based on historical data. This model will consider a variety of factors known to affect ticket prices, including the time at which flight ticket data is collected and airline-specific factors.

The research questions are: What are the primary factors that influence ticket prices for Lufthansa and Swiss flights to Sao Paulo and NYC? How does the time at which flight ticket data is collected impact ticket prices? When is the optimal time to collect flight ticket data to minimize cost? And how accurately can future ticket prices be predicted using a model trained on historical data? In order to answer these questions, the study will involve collecting and analyzing large amounts of data on ticket prices and other relevant factors. This data will be used to train and evaluate the predictive model, allowing us to determine its accuracy and effectiveness in forecasting future ticket prices. By identifying the key factors that influence ticket prices and determining the optimal time to collect flight ticket data, we hope to provide valuable insights for travelers looking to save money on air travel.

2. Data

In this project, data was collected from the online travel booking platform booking.com. This platform was chosen due to its extensive coverage of flights from Lufthansa and Swiss airlines to Sao Paulo and NYC. Data collection was conducted on a daily basis, with observations recorded every 2 hours between 8 am and 10 pm. This schedule was chosen to capture variations in ticket prices throughout the day. In total, 11619 data points were collected over the course of the project.

The data collected includes a range of variables relevant to the research questions. These include the price for round-trip tickets for the specified destinations using Lufthansa and Swiss airlines, as well as the departure and arrival times/dates for both the outbound and inbound flights. Additionally, data on the duration of the outbound and inbound flights and the number of stopovers for both flights were recorded. Since the weight allowance for luggage is consistent across both airlines, this variable was not included in the data collection.

Furthermore, the time at which each observation was recorded was also noted in order to facilitate analysis of trends in ticket prices relative to the day/time of the week. This information is expected to provide valuable insights into how ticket prices vary over time.

By analyzing this comprehensive dataset, it is anticipated that patterns and trends can be identified that will aid in predicting future ticket prices for these specific airlines and destinations. The data collected will be used to train a predictive model that can forecast future ticket prices with a high degree of accuracy. This model will be a valuable tool for travelers seeking to minimize costs and maximize their travel budgets.

2.1 Data preparation

The initial phase of the methodology involves collecting data on ticket prices for Lufthansa and Swiss flights to Sao Paulo and New York City. This objective is achieved by utilizing the Selenium web automation tool to extract ticket price data from the airlines' websites. Selenium is a robust tool that facilitates the automation of web browser interactions, enabling the visiting of websites, inputting flight information, and retrieving ticket price data in an automated fashion.

Data collection occurs at regular intervals, with ticket price data being collected every 2 hours from 8 am to 10 pm daily. This frequent data collection captures changes in ticket prices over time, providing a rich dataset for analysis.

Once the data has been collected, it is converted into a structured format using CSV (Comma-Separated Values) files. The data for each destination is stored in separate folders: one for flights from Brussels to Sao Paulo (*bxl_to_sao*) and one for flights from Brussels to New York City (*bxl_to_nyc*). Each CSV file contains information on ticket prices, flight dates, and other relevant factors.

Furthermore, the collected data is cleaned using the Pandas library in Python. This process includes checking the data for missing or incorrect values and ensuring that it has been correctly collected for the chosen destinations and airlines. The cleaned data is then stored in data frames, which are two-dimensional data structures that enable easy manipulation and analysis of the data.

After cleaning the data, it is preprocessed to prepare it for analysis. This step involves transforming the data into a format that can be easily analyzed. For example, categorical variables may be converted into numerical values or the data may be normalized to ensure that all variables are on a similar scale. By cleaning and preprocessing the data, its accuracy is ensured and it is ready for analysis. This enables identification of patterns and trends in the data, which form the basis of a predictive model. By using tools such as Pandas and Python, large amounts of data can be efficiently manipulated and analyzed.

2.2 Data visualizations

In our analysis of flight prices, we used the Seaborn, plotnine and Matplotlib libraries in Python to create visualizations of our data. These powerful tools allowed us to create bar charts, scatter plots, line plots, boxplots and heatmaps to gain insights into trends and factors influencing ticket prices over time. Our goal was to identify potential features for use in a predictive model.

One of the visualizations we created was a simple bar chart showing the number and frequency of flights to New York City and Sao Paulo operated by Lufthansa and Swiss Airline. As shown in Figure 1, we observed that Lufthansa operates more flights to Sao Paulo (53.16%), while Swiss Airline has a higher number of flights to New York City (51.46%). However, the differences were not significant enough to suggest a higher popularity of one destination over another or a dominance of one airline company over another. This made the two destinations and airlines comparable in terms of price.

We also created additional stacked bar charts, shown in Figure 2, that display the minimum ticket price depending on the day of the week on which the flight is booked. These charts allowed for easy visual comparison between the two companies for each destination. We observed a clear variation in prices, mainly due to Swiss Airline, which tends to increase or decrease its prices more than Lufthansa. For flights to New York City, shown in Figure 2(a), booking on Monday, Tuesday, Saturday or Sunday would be more beneficial. For flights to Sao Paulo, shown in Figure 2(b), the price with Swiss Airline was significantly lower on Tuesday compared to other days of the week.

The second type of plot allows us to see the distribution of flight prices, shown by Figure 3. These boxplots especially draw attention to outliers, representing prices

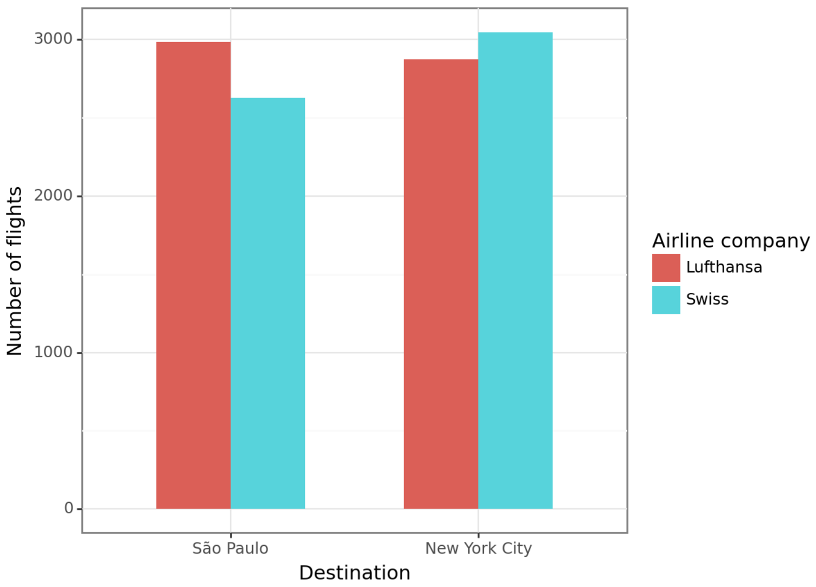
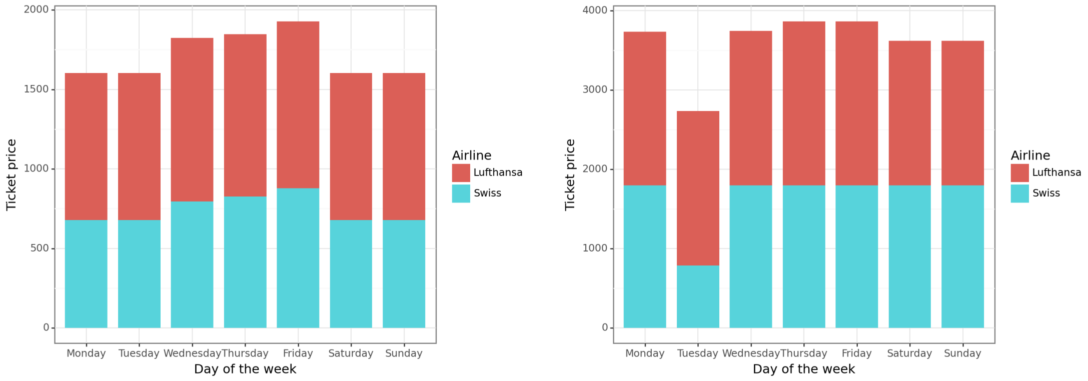


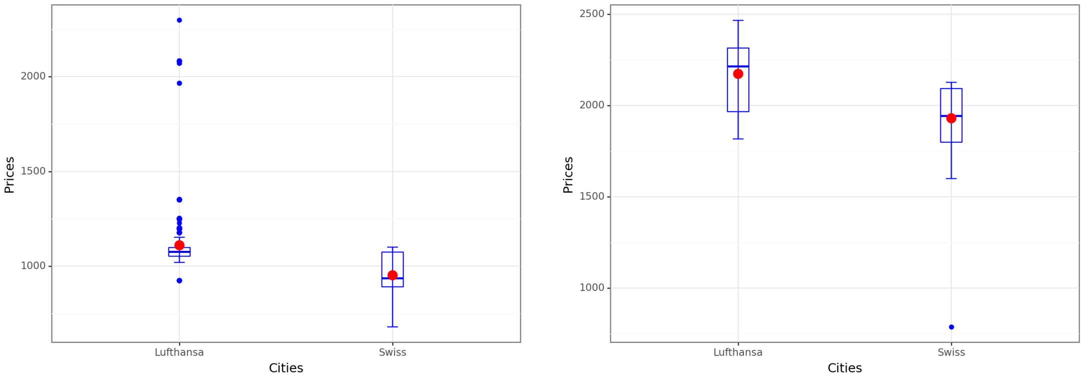
Figure 1: Bar chart showing the number and frequency of flights to New York City and Sao Paulo operated by Lufthansa and Swiss Airline.



(a) Stacked bar chart for flights to New York

(b) Stacked bar chart for flights to Sao Paulo

Figure 2: Stacked bar charts showing the minimum ticket price for flights to New York City and Sao Paulo depending on the day of the week, operated by Lufthansa and Swiss Airline.



(a) Boxplot for flights to New York

(b) Boxplot for flights to Sao paulo

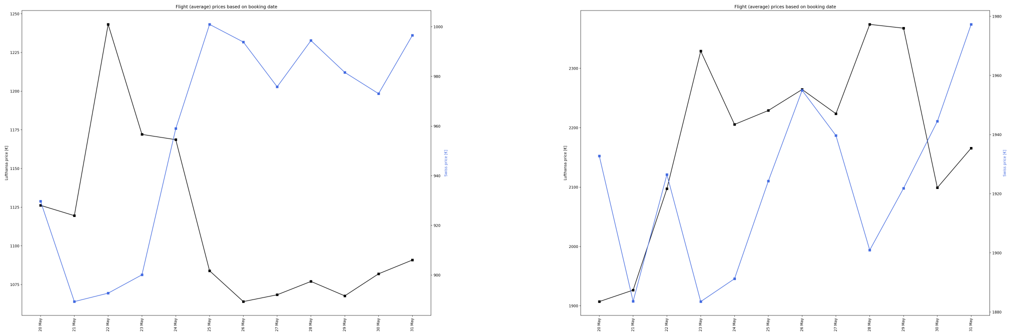
Figure 3: Flight price distribution for different airlines to New York City and Sao Paulo.

that deviated from the typical price range of an airline. The Figure 3(a) reveals that, in the case of flights to New York City, there is a noticeable and higher variability in prices associated with Swiss Airline. The opposite can be observed in the case of Sao Paulo, shown by Figure 3(b). Overall, prices are around 1000 Euros for New York City, but prices can go until 2000 Euros for Sao Paulo. The mean shows that Swiss Airline offers lower prices specifically for each destination, in average.

Line plots are essential for visualizing the fluctuations in prices over time, which allows for a better understanding of any patterns that may emerge. As the departure date was approaching, as shown by Figure 4, the average price of tickets for both destinations did begin to increase and varied more. This highlights the importance of considering the timing of booking to secure more favorable prices. Swiss Airline exhibits a rather consistent increase in prices for both New York City and Sao Paulo as the departure date approaches. This suggests that booking earlier with Swiss Airline could result in more favorable prices. On the other hand, Lufthansa's price trend is less obvious and does not show a clear pattern of increasing prices as the departure date approaches. This could mean that booking timing may not be as crucial when flying with Lufthansa.

Scatter plots, Figure 5, are an effective way to visualize the relationship between flight cost and duration. By visually assessing the correlation between these two variables, we were able to gain insights into their relationship. For example, the Figure 5(a) for New York City revealed a clear negative correlation between flight price and duration, indicating that longer flights tend to be less expensive. This could be a useful feature for predicting flight prices. In contrast, the relationship between these variables was less clear for Sao Paulo, but still tended towards a negative correlation, as shown by Figure 5(b). Additionally, we observed that Lufthansa flights tended to have higher costs for the same flight duration.

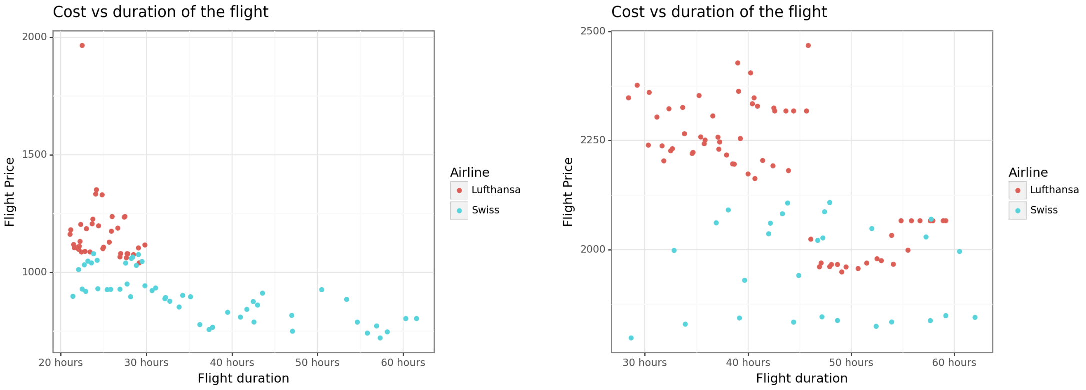
The final step in our data visualization process was to create a heatmap of all numerical variables to better identify features for use in data modeling. Our analysis of the New York City destination revealed a strong negative correlation between ticket price and flight duration, as previously indicated by the scatter plot. Additionally,



(a) Flight price trend to NYC

(b) Flight price trend to NYC

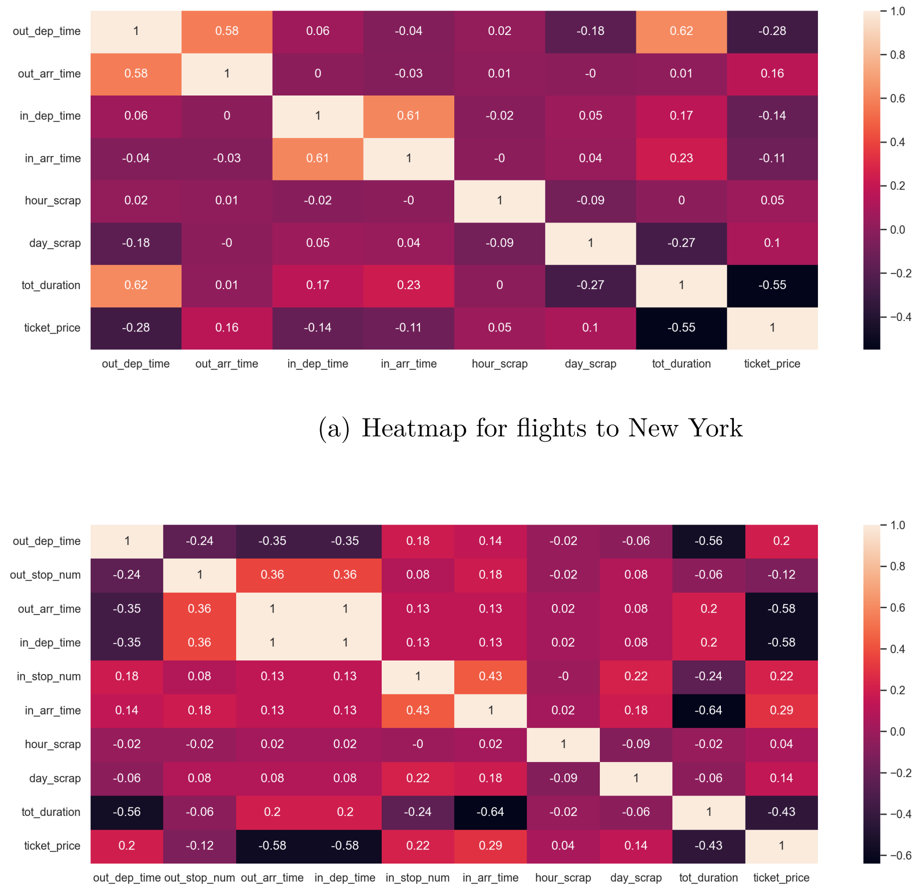
Figure 4: Line plots showing the fluctuations in flight prices over time for Lufthansa and Swiss Airline to New York City and Sao Paulo.



(a) Scatter plot for flights New York

(b) Scatter plot for flights Sao Paulo

Figure 5: Scatter plots showing the relationship between flight cost and duration for New York City and Sao Paulo.



(a) Heatmap for flights to New York

(b) Heatmap for flights for Sao paulo

Figure 6: Heatmaps showing correlations between numerical variables for New York City and Sao Paulo.

departure time also appeared to be negatively correlated with our prediction target. For Sao Paulo, prices showed a stronger correlation with flight duration and time-related variables. These observations and anlysis concerning the heatmaps is illustrated by Figure 6.

2.3 Data modelling

The data visualization process has yielded several key insights that could inform data modeling efforts. For instance, line plots have shown that ticket prices for both New York City and Sao Paulo tend to increase and exhibit greater variation as the departure date approaches. In particular, Swiss Airline has displayed a consistent increase in prices. Scatter plots have also revealed a distinct negative correlation between flight price and duration for New York City. While the relationship for Sao Paulo is less clear, it still tends towards a negative correlation. Heatmaps have further shown strong negative correlations between ticket price and flight duration for New York City, as well as stronger correlations between prices and flight duration/time-related variables for Sao Paulo. These insights highlight the importance of considering factors such as booking timing, flight duration, and time-related variables when modeling data.

In this project, several libraries and functions from the scikit-learn and statsmodels packages were utilized to perform data analysis and modeling. The data was split into training and testing sets using the train_test_split function, while cross-validation was performed using the cross_val_score function. The performance of the models was evaluated using various metrics, including accuracy_score, mean_absolute_error, mean_squared_error, and r2_score. Although several regression models were employed to fit the data and make predictions, including LinearRegression, LogisticRegression and LogisticRegressionCV, only the results from the LinearRegression model will be taken into account. Additionally, statistical analysis was performed and multicollinearity was assessed using the statsmodels library and its api module as well as the variance_inflation_factor function.

Several models were tested on different dataframes. Ultimately, a linear regression model was selected with the following function:

$$Ticket\_price = \beta_0 + \beta_1 \times tot\_duration\_seconds + \beta_2 \times hour\_scrap + \beta_3 \times airline\_company\_dummy + \beta_4 \times destination\_dummy + \beta_5 \times out\_dep\_time\_dummy + \beta_6 \times in\_arr\_time\_dummy + \beta_7 \times day\_of\_week\_.$$

The linear regression model incorporates several independent variables and a dependent variable. The dependent variable is ticket_price, which represents the ticket price for a flight and can take any non-negative numerical value. The first independent variable is tot_duration_seconds, which denotes the total duration of the flight in seconds. The second variable is hour_scrap, which specifies the hour at which the data was collected. The third and fourth variables, airline_company_dummy and destination_dummy, are dummy variables that represent the airline company and destination, respectively. The fifth and sixth variables, out_dep_time_dummy and in_arr_time_dummy, are dummy variables that indicate the departure time from the origin and arrival time at the destination, respectively. Lastly, day_of_week_* represents dummy variables that indicate the day of the week on which the data was collected.

	Values
Intercept	2097.4909712064773
Training RMSE	522.364425
Test RMSE	524.069444
Cross-validation RMSE	516.31323
Mean Absolute Error	171.879278
Mean Squared Error	274648.782087
Root Mean Squared Error	524.069444
R-squared	0.552468

Table 1: Different metrics to assess the model and their results

The linear regression model was assessed and several values were obtained, as shown by Table 1. The intercept value of 2097.49 represents the expected value of the dependent variable *ticket_price* when all independent variables are equal to zero. The training and test RMSE values of 522.36 and 524.07, respectively, represent the root mean squared error of the model on the training and test data. The cross-validation RMSE value of 516.31 represents the root mean squared error of the model when evaluated using cross-validation. The mean absolute error value of 171.88 represents the average absolute difference between the predicted and actual values of the dependent variable. The mean squared error value of 274648.78 represents the average squared difference between the predicted and actual values of the dependent variable. The root mean squared error value of 524.07 represents the square root of the mean squared error. The R-squared value of 0.55 represents the proportion of variance in the dependent variable that is explained by the independent variables in the model.

	coef	**p-values**
const	2103.6217	1.613956e-219
tot_duration_seconds	-0.0011	5.714598e-08
hour_scrap	-2.8844	5.003298e-02
day_scrap	16.5914	2e-24
airline_company_dummy	-291	2e-128
destination_dummy	-1207	0
out_dep_time_dummy	-159	1e-33
in_arr_time_dummy	89	2e-09

Table 2: Coefficients of the model and their p-values

The coefficients and p-values for each independent variable represent the estimated effect of that variable on the dependent variable and the statistical significance of that effect, respectively. For example, the coefficient for tot_duration_seconds is -0.0012, indicating that for every one-unit increase in tot_duration_seconds, ticket_price is expected to decrease by 0.0012 units on average, holding all other variables constant. The p-value for tot_duration_seconds is 5.71e-08, indicating that this effect is statistically significant at a conventional level $e.g., p < 0.05$. Similarly, the coefficient for hour_scrap is -2.31, indicating that for every one-unit increase in hour_scrap, ticket_price is expected to decrease by 2.31 units on average, holding all other variables constant. The p-value for hour_scrap is 0.050, indicating that this effect is marginally statistically significant at a conventional level. They are summarize by the Table 2.

3. Discussion

What do the results mean (their significance and to whom) and how do they answer your research questions/hypotheses? From the evaluation and validation of the tool, what can you conclude?

3.1 Limitations and Challenges

What could have been investigated if given more time? What have been difficult when solving the problem and getting answers for your research questions/hypotheses?

4. Conclusion

Write the conclusion. What did you gain from the project assignment? Briefly explain your questions/hypotheses, findings, and meaningful discussion points in relation to the data collection, data management, data analysis, visualization and interaction concepts, and evaluation of your tool. What additional investigations need to be performed (or what is the limitation) in order to say that your solution is a good one for the problem?

5. Reflections on own work

- Describe how you decided to scope (and/or re-scope) your problem formulation during the work, and given the data you had access to.
- Describe how you searched for knowledge on how to scope the DS question you wanted to answer
- understand how to implement, test, and validate your results.
- Which sources helped you to get progress and how?
- What would you have done differently if you were to start over again, for understanding the problem faster and for understanding what could be done with the data you had access to?
- What else is there that you would have changed about this assignment?