**Airline Passenger Satisfaction**

**Business Data Analysis using Python and Tableau**

**Sarita Namia**

14 June 2020

### Abstract

This report studies airline customer satisfaction using data from a US airline survey. Attempting to give answers to a problem that is common among businesses, more than 100K respondent data was collected. This report aims to deep dive into which factors influence airline customer satisfaction and provide decision-makers some useful insights on how the airline can offer a better service. Python was used to segment and analyze the data and Tableau to visualize the results.

The results of the data analysis show that, overall, business class airline customers are the most satisfied. The statistical test that was chosen for this analysis showed how each variable contributes to the overall satisfaction level. This report underlines that the distance of the flight plays an important role in how customers will rate their flight experience. Moreover, the departure and arrival delays as well as the online boarding service have a strong relationship with customer satisfaction. Gender and gate location, on the other hand, do not affect the overall satisfaction results.

Contents

[Abstract 1](#_Toc43186636)

[Introduction 2](#_Toc43186637)

[Data Structure 2](#_Toc43186638)

[Initial Cleaning 2](#_Toc43186639)

[Notes & Assumptions 3](#_Toc43186640)

[Business Questions 3](#_Toc43186641)

[Exploratory Data Analysis 3](#_Toc43186642)

[Passengers 3](#_Toc43186643)

[Flights 4](#_Toc43186644)

[Quality service ratings 5](#_Toc43186645)

[Arrival and Departure Delay 5](#_Toc43186646)

[Results 6](#_Toc43186647)

[Correlation heatmaps 6](#_Toc43186648)

[Satisfaction across all the variables 6](#_Toc43186649)

[Conclusions 7](#_Toc43186650)

### Introduction

The dataset which was used is based on a US Airline passenger satisfaction survey. The data were collected from Kaggle website. This dataset had already been split into train and test csv files. 80% of the total dataset is in train file and 20% is in the test file. In this report only the train dataset will be used as the purpose of this report is not to build a predictive model.

Exploratory data analysis, initial cleaning, correlation, and clustering analysis were conducted in Python, using NumPy, Pandas, and Scikit-learn (known as Sklearn) libraries. Tableau was used for visualizing the results together with Seaborn library in which correlation heatmaps were created.

### Data Structure

The dataset contains 103,904 unique rows and 25 columns. The columns can be divided into five sections based on their content: (the type of the variable is written in *italics*)

1. Passenger’s information:
   * Unnamed:0 Numerical-doesn’t provide any information *integer*
   * Id: Customer’s id *integer*
   * Gender: (Female, Male) *string*
   * Customer Type: (Loyal customer, disloyal customer) *string*
   * Age: Actual age of the passenger *integer*
2. Details about the flight:
   * Type of Travel: (Personal Travel, Business Travel) *string*
   * Class: (Business, Eco, Eco Plus) *string*
   * Flight distance *integer*
3. Rating on different attributes (0:not applicable, 1-5) This section contains only *integers*:
   * Inflight wifi service
   * Departure/Arrival time convenient
   * Ease of Online booking
   * Gate location
   * Food and drink
   * Online boarding
   * Seat comfort
   * Inflight entertainment
   * On-board service
   * Leg room service
   * Baggage handling
   * Check-in service
   * Inflight service
   * Cleanliness
4. Flight delays:
   * Departure Delay in Minutes *integer*
   * Arrival Delay in Minutes *float*
5. Satisfaction level:

* Satisfaction: (neutral or dissatisfied, satisfied) *string*

### Initial Cleaning

The dataset was already fairly clean. Columns ‘Unnamed:0’ and ‘Id’ were removed as they did not provide any useful information. In the ‘Arrival Delay in minutes’ column, there were 310 nulls which later were replaced with zeros - assumption that these trips had no arrival delay.

### Notes & Assumptions

1. The satisfaction level is divided into two groups: neutral or dissatisfied and satisfied. The fact that satisfaction was not divided into three groups can be explained by the fact that **dissatisfied or disengaged** (neutral) **customers** have similar behavior- they tend **not to choose again the same product/service**. This means fewer passengers and less revenue. Customers should have a **positive experience** every time they travel as this contributes to more **loyal customers.**
2. On the ‘Flight Distance’ column the measurement unit is not defined. Based on the fact that the dataset comes from a US airline and by looking at the data it is obvious that the flight distance is calculated in **miles**.
3. In the service industry, **service quality** and **price** have an important role in influencing **customer satisfaction**. It would add extra value to the study if the prices of the tickets were recorded.

### Business Questions

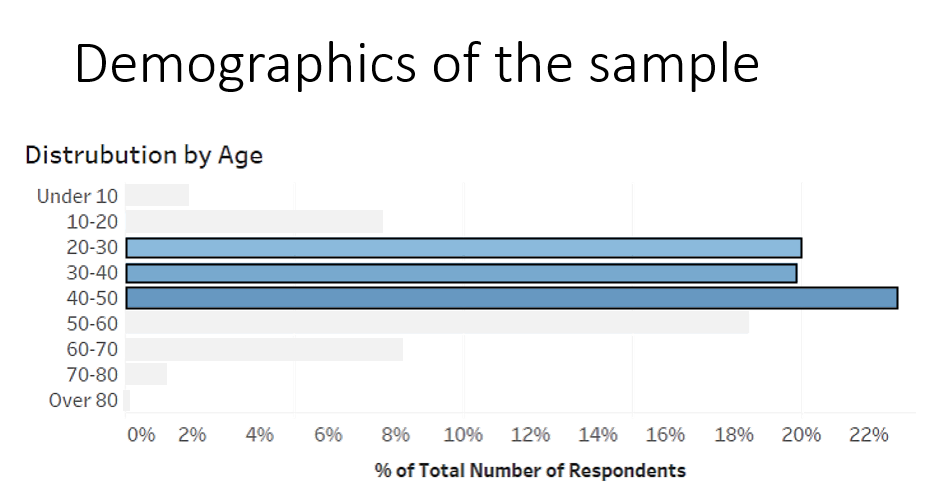
Based on the structure and the information provided by the dataset, this report aims to answer the following questions:

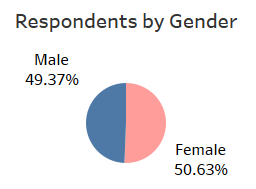
* Which is the age group that travels the most, according to the sample?
* The business class trip experience is better, satisfaction wise, comparing the other classes. Is this assumption true?
* As customer satisfaction usually leads to customer loyalty, loyal dissatisfied/neutral customers are fewer than loyal satisfied. Is this true?
* Which are the variables that have the strongest and weakest relationship with satisfaction?
* Does the customer’s gender affect satisfaction results?

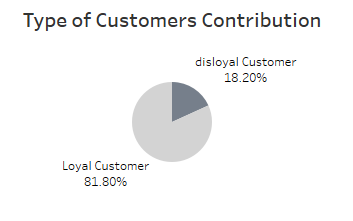
### Exploratory Data Analysis

# Passengers

Out of the 103,904 passengers, 50.7% were female. The average age is 39 (in both genders) (σ=15), with the youngest age to be 7 and the oldest 85. The vast majority of the passengers were loyal customers (81.7%). Passengers in the **decades of 30-50** are those who **travel the most**, according to the sample. This explains the fact that most of the passengers travel for **business** (next graphs) purposes since this decade is one of the most productive in a person’s professional life. More than **80%** of passengers are **loyal** customers.

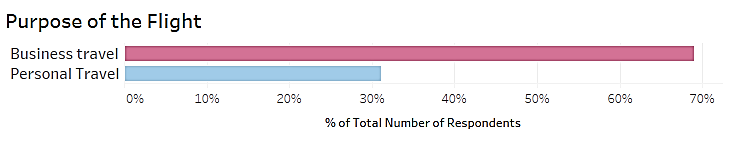


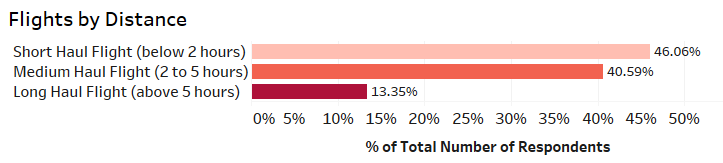




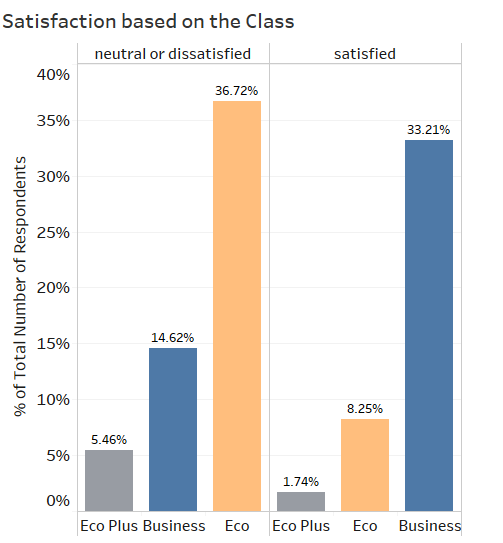
# Flights

Most of the customers travel for **business** and also more than 86% of the flights were **domestic** (less than 5 hours duration).

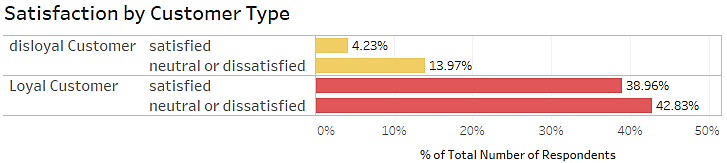




Examing the classes and the satisfaction results, as assumed, most of the **satisfied customers** traveled with **Bussiness class**, not surprisingly as business class trips provide more comfort for the passenger. On the other hand, **Economy class** ‘created‘ the most **neutral/dissatisfied** customers.



The assumption that loyal **neutral/dissatisfied** customers are fewer than **loyal** satisfied ones is not verified as it can be noticed in the graph below.



# Quality service ratings

As can be viewed in the table below, the winners of quality service ratings are **inflight service**, **baggage handling**, and **seat comfort**. At the bottom of the list are **ease of online booking** and **inflight wifi service.**



# Arrival and Departure Delay

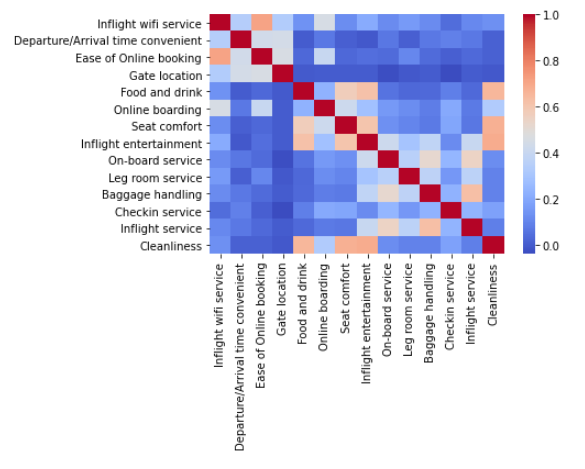
The average absolute value of the subtraction of arrival and departure delay is 5 minutes thus arrival delay can be representative for both of the delays.

**15 minutes** is the average **arrival delay** (σ=38), with a maximum delay of 26.5 hours. **64.1%** of all the flights **arrived on time**.

### Results

# Correlation heatmaps

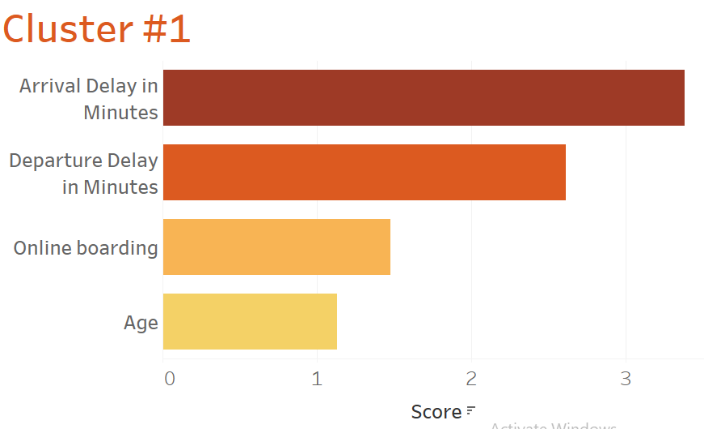
Interesting insights from correlation heatmap between all the numerical variables is that **ease of online booking** is correlated to **inflight wife** **service** and **seat comfort** is correlated to **inflight entertainment**. Arrival and departure delays are highly correlated which was expected.

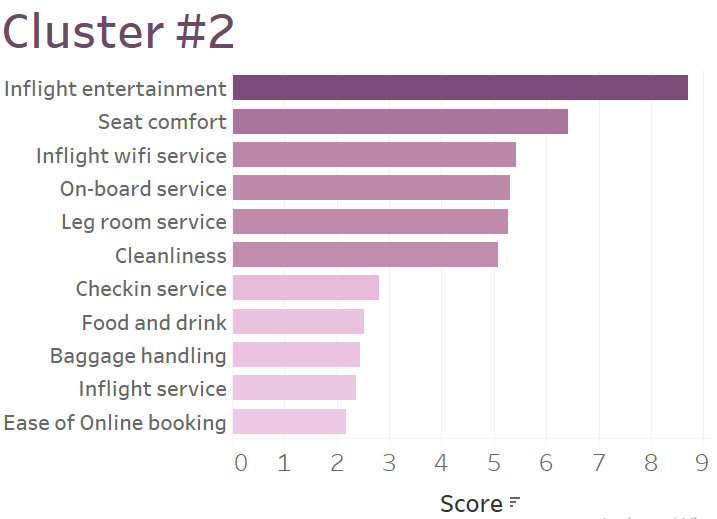


The correlation heatmap between the variables that were rated by the passengers is displayed on the left. It seems like easy of **online booking** is correlated to **inflight wifi service** and also **food and drink**, **online boarding**, and **seat comfort** are correlated to **cleanliness**.

# Satisfaction across all the variables

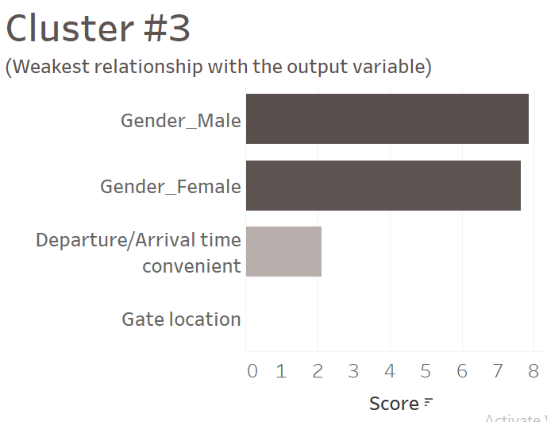
Based on the chi2 statistical method provided by **scikit-learn** library, **flight distance** has the strongest relationship with the **satisfaction** level, significantly ahead from the second one. An explanation on that could be that the longest the trip, the longest is the **duration onboard** which leads to different needs and expectations by the customers.

Based on the different order of magnitude of satisfaction relationship, clusters were created so that the different variables to be compared. Following flight distance, **delays** have a significant effect on the satisfaction outcome result. Moreover, **online boarding** service seems to be very closely connected to satisfaction. The **age** of the passenger also plays its role and should not be ignored.



The second cluster contains a bigger number of variables, the top of which are **inflight entertainment service** and **seat comfort**. Inflight wifi service, on-board service, leg room, and cleanliness are following.

Finally, gender has a very weak relationship with satisfaction and gate location doesn’t affect at all the satisfactory outcome.



### Conclusions

* **40** years oldis the average age of the passengers. **Most of them** (70%) travel for **business purposes** half of which travel with **business class.**
* The **business class** trip experience creates a higher amount of **satisfied** customers compared to the other classes.
* **Less than half** (47.6%) of **loyal** customers were **satisfied** with their trip experience, this s something that should be investigated further in future analysis.
* Priority for the airline company decision-makers should be on how to avoid **departure/arrival delays** and improve the **online boarding** service.
* Second on their agenda should be the development of **inflight entertainment** service.