**Airline Passenger Satisfaction**

**Exploratory Analysis**

Carson Brower, cbrower@bellarmine.edu

Caleb Bush, cbush6@bellarmine.edu

1. **INTRODUCTION**

The data set we are using contains an airline survey that displays satisfaction ratings based of of factors that are correlated. The data set uses multiple factors to see whether satisfaction can be predicted and what factors may or not correlate with each other and satisfaction ratings. We chose this data set because we thought it would be easier to work with based on the variables within the dataset. As well as it would be interesting to see why they would be satisfied or dissatisfied with their airline travel. Our data set can be found using this link: [Airline Passenger Satisfaction (kaggle.com)](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)

1. **DATA SET DESCRIPTION**

This data set contains 103,903 samples with 25 columns with various data types. We dropped the first column which was called, “Unknown:0” which was our row id but that was already provided. We changed the missing data for the column, “Arrival delay in minutes” and converted it using the mean of the numerical data within that column. After filling in the missing data as seen from the category in the table “Non-Null” all the numbers are the same therefore we can see there is no more missing data. A sample of the data is shown in the table below. The data types for our columns are in the image below in the category “Dtype.”

A screenshot of a computer

Description automatically generated

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| V1 |  | 0% |
| V2 |  | 0% |
| V3 |  | 0% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

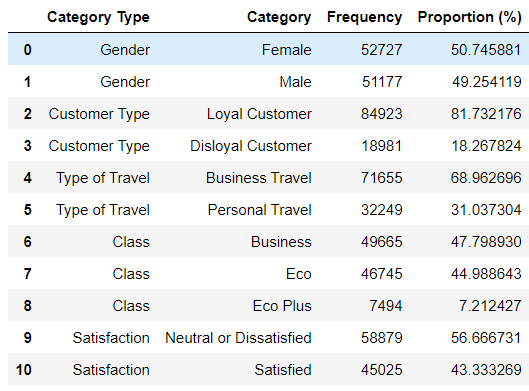
**Table 2: Summary Statistics for XXX (name of dataset)**

*A table with numbers and text

Description automatically generated*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

**

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables

After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

A diagram of a box plot

Description automatically generated with medium confidence

*A graph showing a line

Description automatically generated with medium confidence*

*A graph with blue dots

Description automatically generated*

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.