Airline Passenger Satisfaction
Exploratory Analysis

Carson Brower, cbrower@bellarmine.edu
Caleb Bush, cbush6@bellarmine.edu

I. INTRODUCTION

The data set we are using contains an airline survey that displays satisfaction ratings based 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)

II. 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."

Extra Table which provides a sample of our data

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	...	5	4	3	4	4	5	5	25	18.0	neutral or dissatisfied
1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	...	1	1	5	3	1	4	1	1	6.0	neutral or dissatisfied
2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	...	5	4	3	4	4	4	5	0	0.0	satisfied
3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	...	2	2	5	3	1	4	2	11	9.0	neutral or dissatisfied
4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	...	3	3	4	4	3	3	3	0	0.0	satisfied
5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180	3	4	2	...	1	3	4	4	4	4	1	0	0.0	neutral or dissatisfied
6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2	4	2	...	2	3	3	4	3	5	2	9	23.0	neutral or dissatisfied
7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4	3	4	...	5	5	5	5	4	5	4	4	0.0	satisfied
8	79485	Female	Loyal Customer	41	Business travel	Business	853	1	2	2	...	1	1	2	1	4	1	2	0	0.0	neutral or dissatisfied
9	65725	Male	disloyal Customer	20	Business travel	Eco	1061	3	3	3	...	2	2	3	4	4	3	2	0	0.0	neutral or dissatisfied

10 rows × 24 columns

Table 1: Data Types and Missing Data

```
In [24]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 24 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   id                                103904 non-null  int64
 1   Gender                            103904 non-null  object
 2   Customer Type                     103904 non-null  object
 3   Age                               103904 non-null  int64
 4   Type of Travel                    103904 non-null  object
 5   Class                             103904 non-null  object
 6   Flight Distance                   103904 non-null  int64
 7   Inflight wifi service             103904 non-null  int64
 8   Departure/Arrival time convenient 103904 non-null  int64
 9   Ease of Online booking            103904 non-null  int64
 10  Gate location                     103904 non-null  int64
 11  Food and drink                    103904 non-null  int64
 12  Online boarding                   103904 non-null  int64
 13  Seat comfort                      103904 non-null  int64
 14  Inflight entertainment            103904 non-null  int64
 15  On-board service                  103904 non-null  int64
 16  Leg room service                  103904 non-null  int64
 17  Baggage handling                  103904 non-null  int64
 18  Checkin service                   103904 non-null  int64
 19  Inflight service                  103904 non-null  int64
 20  Cleanliness                       103904 non-null  int64
 21  Departure Delay in Minutes        103904 non-null  int64
 22  Arrival Delay in Minutes          103904 non-null  object
 23  satisfaction                      103904 non-null  object
dtypes: int64(18), object(6)
```

III. Data Set Summary Statistics

In this section of our findings from the data you will find four tables and a heat map. The first two tables represent statistical information regarding each categorical variable and the statistics based on our data set with information like proportion percentage, mean, max, min, and standard deviation. The third table is a correlation table this provides correlations between each categorical variable within the dataset and how each variable correlate to the other. The heat map is in connection with the correlation table and explains based on colors where the lighter the color the stronger the positive correlation there is and the darker the color the more negative correlation is there based upon variable correlations.

Table 2: Summary Statistics for our dataset

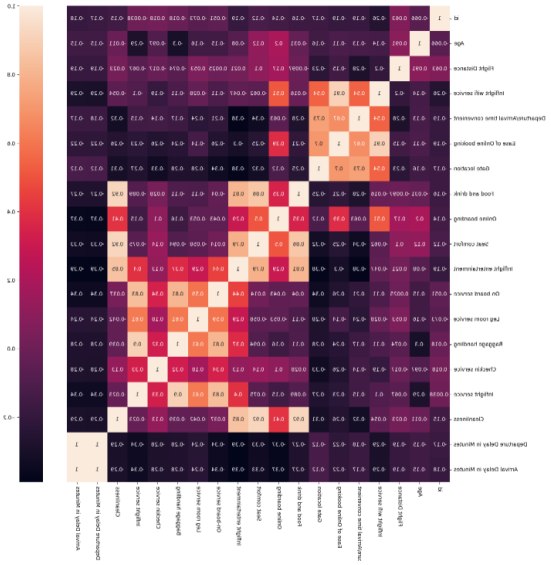
	count	mean	std	min	25%	50%	75%	max
id	103904.0	64924.210502	37463.812252	1.0	32533.75	64856.5	97368.25	129880.0
Age	103904.0	39.379706	15.114964	7.0	27.00	40.0	51.00	85.0
Flight Distance	103904.0	1189.448375	997.147281	31.0	414.00	843.0	1743.00	4983.0
Inflight wifi service	103904.0	2.729683	1.327829	0.0	2.00	3.0	4.00	5.0
Departure/Arrival time convenient	103904.0	3.060296	1.525075	0.0	2.00	3.0	4.00	5.0
Ease of Online booking	103904.0	2.756901	1.398929	0.0	2.00	3.0	4.00	5.0
Gate location	103904.0	2.976883	1.277621	0.0	2.00	3.0	4.00	5.0
Food and drink	103904.0	3.202129	1.329533	0.0	2.00	3.0	4.00	5.0
Online boarding	103904.0	3.250375	1.349509	0.0	2.00	3.0	4.00	5.0
Seat comfort	103904.0	3.439396	1.319088	0.0	2.00	4.0	5.00	5.0
Inflight entertainment	103904.0	3.358158	1.332991	0.0	2.00	4.0	4.00	5.0
On-board service	103904.0	3.382363	1.288354	0.0	2.00	4.0	4.00	5.0
Leg room service	103904.0	3.351055	1.315605	0.0	2.00	4.0	4.00	5.0
Baggage handling	103904.0	3.631833	1.180903	1.0	3.00	4.0	5.00	5.0
Checkin service	103904.0	3.304290	1.265396	0.0	3.00	3.0	4.00	5.0
Inflight service	103904.0	3.640428	1.175663	0.0	3.00	4.0	5.00	5.0
Cleanliness	103904.0	3.286351	1.312273	0.0	2.00	3.0	4.00	5.0
Departure Delay in Minutes	103904.0	14.815618	38.230901	0.0	0.00	0.0	12.00	1592.0
Arrival Delay in Minutes	103904.0	15.178678	38.640909	0.0	0.00	0.0	13.00	1584.0

Table 3: Proportions from the dataset

	Category Type	Category	Frequency	Proportion (%)
0	Gender	Female	52727	50.745881
1	Gender	Male	51177	49.254119
2	Customer Type	Loyal Customer	84923	81.732176
3	Customer Type	Disloyal Customer	18981	18.267824
4	Type of Travel	Business Travel	71655	68.962696
5	Type of Travel	Personal Travel	32249	31.037304
6	Class	Business	49665	47.798930
7	Class	Eco	46745	44.988643
8	Class	Eco Plus	7494	7.212427
9	Satisfaction	Neutral or Dissatisfied	58879	56.666731
10	Satisfaction	Satisfied	45025	43.333269

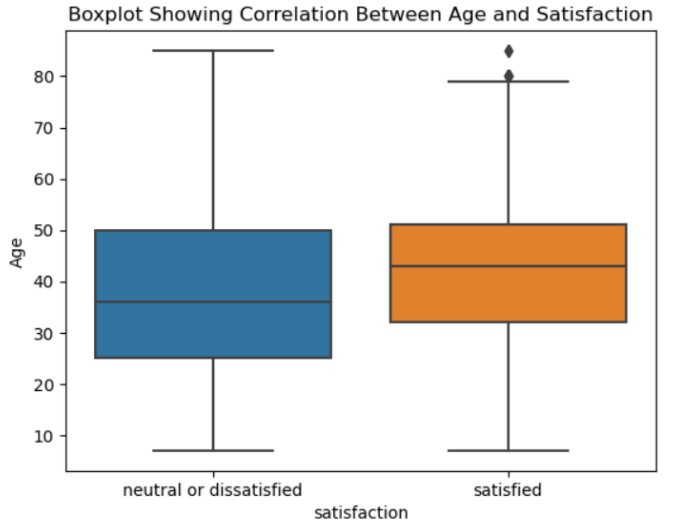
Table 4: Correlation Table/Tables

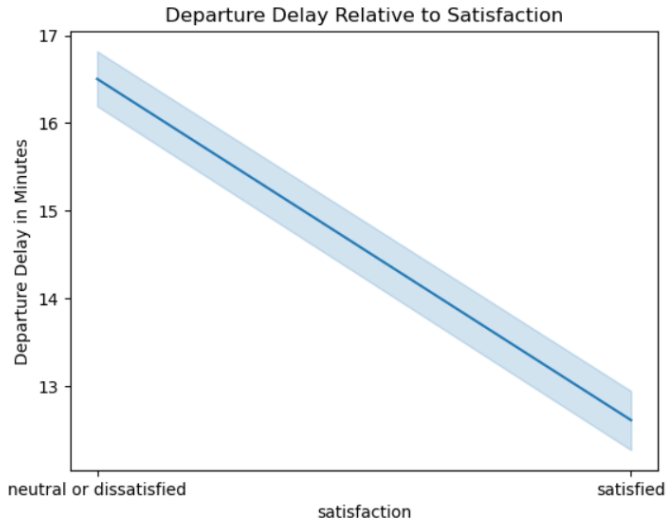
	id	Age	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
id	1.000000	0.022857	0.095544	-0.021276	-0.002110	0.014163	-0.000606	0.001063	0.055477	0.052903	0.002300	0.055241	0.044634	0.074940	0.079273	0.079346	0.024965	-0.019546	-0.037196
Age	0.022857	1.000000	0.099461	0.017859	0.038125	0.024842	-0.001330	0.023000	0.208939	0.160277	0.076444	0.057594	0.040583	-0.047529	0.035482	-0.049427	0.053611	-0.010152	-0.012127
Flight Distance	0.095544	0.099461	1.000000	0.007131	-0.020043	0.065717	0.004793	0.056994	0.214869	0.157333	0.128740	0.109526	0.133916	0.063184	0.073072	0.057540	0.093149	0.002158	-0.002423
Inflight wifi service	-0.021276	0.017859	0.007131	1.000000	0.343845	0.715856	0.336248	0.134718	0.456970	0.122658	0.209321	0.121500	0.160473	0.120923	0.043193	0.110441	0.132698	-0.017402	-0.019067
Departure/Arrival time convenient	-0.002110	0.038125	-0.020043	0.343845	1.000000	0.436961	0.444757	0.004906	0.070119	0.011344	-0.004861	0.068882	0.012441	0.072126	0.093333	0.073318	0.014292	0.001005	-0.000863
Ease of Online booking	0.014163	0.024842	0.065717	0.715856	0.436961	1.000000	0.458655	0.031873	0.404074	0.030014	0.047032	0.038833	0.107601	0.038762	0.011081	0.035272	0.016179	-0.006371	-0.007972
Gate location	-0.000606	-0.001330	0.004793	0.336248	0.444757	0.458655	1.000000	-0.001159	0.001688	0.003669	0.003517	-0.028373	-0.005873	0.002313	-0.035427	0.001681	-0.003830	0.005467	0.005136
Food and drink	0.001063	0.023000	0.056994	0.134718	0.004906	0.031873	-0.001159	1.000000	0.234468	0.574556	0.622512	0.059073	0.032498	0.034746	0.087299	0.033993	0.657760	-0.029926	-0.032472
Online boarding	0.055477	0.208939	0.214869	0.456970	0.070119	0.404074	0.001688	0.234468	1.000000	0.420211	0.285066	0.155443	0.123950	0.083280	0.204462	0.074573	0.331517	-0.018982	-0.021915
Seat comfort	0.052903	0.160277	0.157333	0.122658	0.011344	0.030014	0.003669	0.574556	0.420211	1.000000	0.610590	0.131971	0.105559	0.074542	0.191854	0.069218	0.678534	-0.027898	-0.029851
Inflight entertainment	0.002300	0.076444	0.128740	0.209321	-0.004861	0.047032	0.003517	0.622512	0.285066	0.610590	1.000000	0.420153	0.299692	0.378210	0.120867	0.404855	0.691815	-0.027489	-0.030658
On-board service	0.055241	0.057594	0.109526	0.121500	0.068882	0.038833	-0.028373	0.059073	0.155443	0.131971	0.420153	1.000000	0.355495	0.519134	0.243914	0.550782	0.123220	-0.031569	-0.035172
Leg room service	0.044634	0.040583	0.133916	0.160473	0.012441	0.107601	-0.005873	0.032498	0.123950	0.105559	0.299692	0.355495	1.000000	0.369544	0.153137	0.368656	0.096370	0.014363	0.011824
Baggage handling	0.074940	-0.047529	0.063184	0.120923	0.072126	0.038762	0.002313	0.034746	0.083280	0.074542	0.378210	0.519134	0.369544	1.000000	0.233122	0.628561	0.095793	-0.005573	-0.008530
Checkin service	0.079273	0.035482	0.073072	0.043193	0.093333	0.011081	-0.035427	0.087299	0.204462	0.191854	0.120867	0.243914	0.153137	0.233122	1.000000	0.237197	0.179583	-0.018453	-0.020339
Inflight service	0.079346	-0.049427	0.057540	0.110441	0.073318	0.035272	0.001681	0.033993	0.074573	0.069218	0.404855	0.550782	0.368656	0.628561	0.237197	1.000000	0.088779	-0.054813	-0.059104
Cleanliness	0.024965	0.053611	0.093149	0.132698	0.014292	0.016179	-0.003830	0.657760	0.331517	0.678534	0.691815	0.123220	0.096370	0.095793	0.179583	0.088779	1.000000	-0.014093	-0.015749
Departure Delay in Minutes	-0.019546	-0.010152	0.002158	-0.017402	0.001005	-0.006371	0.005467	-0.029926	-0.018982	-0.027898	-0.027489	-0.031569	0.014363	-0.005573	-0.018453	-0.054813	-0.014093	1.000000	0.961161
Arrival Delay in Minutes	-0.037196	-0.012127	-0.002423	-0.019067	-0.000863	-0.007972	0.005136	-0.032472	-0.021915	-0.029851	-0.030658	-0.035172	0.011824	-0.008530	-0.020339	-0.059104	-0.015749	0.961161	1.000000

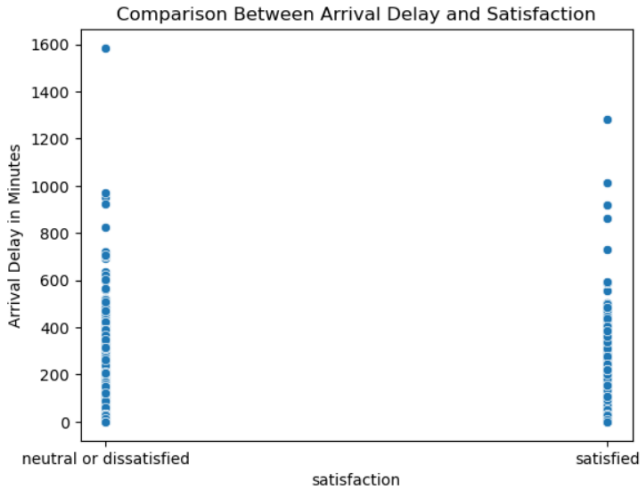


IV. DATA SET GRAPHICAL EXPLORATION

Within the data set graphical exploration section of the analysis we used three graphs and charts to display our data and to answer the major question of what we were trying to find which was what variables directly correlated to satisfaction. The first graph we did was a boxplot displaying age and their satisfaction. The second graph was a line graph comparing departure delay time to satisfaction. While the last graph was scatterplot comparing arrival time delay and satisfaction. These were only a few of the many graphs we did that will be shared within our presentation report.







V. SUMMARY OF FINDINGS

After going through our dataset and cleaning it up finding statistical information, using multiple graphs, and discussing the data together. Through this entire process the major questions we are trying to answer are, what variables directly correlate with satisfaction? As well as can we predict whether a customer will

be satisfied or not? Through our exploration of the dataset, we concluded that there are certain categorical variables that are related and correlate to satisfaction as well as satisfaction can be predicted. Quite a few of the variables that are calculated from passengers such as inflight Wi-Fi services, gate location, food and drink, seat comfort, on-board service and many more variables like these were related to satisfaction. When looking at those variables when the passengers gave those a higher rating on a scale from 1 to 5, they were more likely satisfied than the passengers who gave it a lower rating. Other calculated variables such as age, departure and arrival delay, class, gender, and customer type, still correlated to whether the customer was satisfied or not but were much closer in terms of a 50/50 percentage to whether the customer was satisfied or not unlike the clearer definitive results from the other variables. To answer the second question, passenger satisfaction can be predicted based upon all the variables that can be calculated. Variables such as age, gender, and flight distance would not help predict satisfaction for future customers but all the other variables that can be calculated and rated upon by passengers is a direct indicator of satisfaction and can be predicted.