**Group 3: Airline Prices**

*Kyle Cielencki, Luis Cerrilla, Ahasan Hossain, Artie Edwards*

Majority of people determine between 54- or 60-days searching departure flights , which in turn they can find a reasonable cost for their flight destination. Airline prices overtime fluctuate during peak season and off season for the cities that are desirable throughout the year. While gathering all the data information from Transportation Statistics Airline Origin and Destinations Survey, we based examples of visualizations over the last three years.

Overview we created an interactive app Plane Price Prediction determine airline arrival and departure flight throughout the United States with dynamic content. The website is deployed on Heroku then laid out our analysis and prediction model on five pages; which include (Home, Price Predictor, Visualizations, Methodology, About). The website utilize bootswatch, CSS and D3 result listeners, this website is a creative demonstration for end- users.

**Flight Price Trend and Geographical Breakout Q1**

**A picture containing screenshot

Description automatically generated**

**Flight Price trend and Geographical Breakout 2017 – Q1 OF 2020**

**A close up of a map

Description automatically generated**

**1. Data Gathering**

The data used in this app is from the Bureau of Transportation Statistics Airline Origin and Destination Survey ("DB1BCoupon" and "DB1BTicket" tables).

To make more meaningful data visualizations, we included the airport names, city names and coordinates (latitude, longitude). This required finding an appropriate dataset that contained these fields. We then joined this data to the BTS data using Python Pandas library.

We based our visualizations on a random 5% sample of the last three years of data. This data was combined with an airport city, latitude and longitude for better visualizations. These can be found on the "Visualizations" page.

Our flight price prediction model used a random 25% sample of Q1 2020 data from above. Our model is based on the most recent quarter as this most accurately reflects the current travel environment. Our model is based on Origin, Destination, Distance, Operating Carrier, Quarter, and Passenger Numbers. We used a "linear regression" model using these inputs as we noticed a higher correlation between these and flight ticket price.

For data cleanup, there were fortunately not any N/As to begin with. However, we did notice some itineraries with well over 100 tickets (a Boeing 747 has 276-467 seats). We narrowed our dataset to exclude any itineraries with less than 100 tickets. Additionally, there were some itineraries that cost as much as $20,000. We excluded any itineraries that cost more than $2,000.

**2. Data Exploration**

To understand how the data is distributed, we used Seaborn library's distplot( ) function on a some of the data features:

**Number of tickets ordered:**

A screenshot of a cell phone

Description automatically generated

Most customers purchased only one ticket per itinerary. There were some outliers in which there were 100+ tickets ordered. We limited our dataset to tickets with passengers less than 100 (instances with 50+ were left in as they did not bring down the "Price per ticket").

**Miles Flown:**

A screen shot of a computer

Description automatically generated

The median flight distance flown is ~874 miles (roughly the distance between Minneapolis and Denver). The vast majority of values fall under 2,000 miles; however, there are some outliers in the 3,000-5,000+ miles traveled range.

**Ticket Prices:**

A picture containing knife

Description automatically generated

The median price per passenger is ~$451. The vast majority of prices fall under $1,000 with a number of outliers.

**3. Visualizations**

*Ahason’s portion*

**4. Machine Learning**

*Luis’ portion*A picture containing knife

Description automatically generated

**A picture containing people

Description automatically generatedMachine Learning Model**

**Overview**

The purpose of the machine learning model is to predict the fare for domestic flights in the USA territory given only two inputs, origin airport code and destination airport code. To accomplish this purpose the dataset had to be fitted, trained and tested.

**Data**

Figure Data

The dataset used for the model is from the [*Bureau of Transportation Statistics*](https://www.bts.gov/) which according to their website “is one of more than a dozen principal federal statistical agencies”. The sample used for the model is a 5% of all the available data from the first quarter of 2020 After cleaning the data, this ended up in 32 columns and 3,267 rows

**Model**

To build this model, first we needed a correlation between our target, the Fare column, and the rest of the columns. With that result we decided to use the columns that had more than 5.6% correlation as futures in our model, the other columns were ignored because they were not truly relevant.

Then we fit the data and divide it into 80% for training purposes and 20% for testing purposes, these parameters will apply to all the tested models. Also, the data of all the tested model was fitted and unfitted to see the results in both ways.

A screenshot of a cell phone

Description automatically generated

Figure Model

First, the data was tested with linear models which included a Linear Regression, Logistic Regression model, K Neighbors Classifier, Support Vector Machine (SVM) and also Tree Models like: Random Forest Regressor, Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier and XTREME Gradient Boosting Classifier. That is a total of 22 models.

Out of these 22 models the two with best results were the Linear Regression and Random Forest Regressor. Here is a comparison for both models:

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Although the Random Forest Regressor had a negative r2 score on the test, it had a better result on the Train and that is why it was tested on production with better results.

A screenshot of a cell phone

Description automatically generated

Figure Selected Model

Because no one knows exactly how many miles their flight is going to take given that it is one of our main features and that the app is only asking for the origin airport code and destination airport code, it was necessary to pull the latitude and longitude from each one of the airports and then calculate the distance between both coordinates to be able to get the distance in miles and include it into our futures.

**Tableau**

This Dashboard provides historical data on travel prices throughout the last 3 years to help gauge how much does usually cost to travel to a specific destination. The time range is also segmented in Quarters, which provides more visibility to the user as to when tickets more expensive from prior years.

A close up of a map

Description automatically generated

The map provides an overall view of how much it cost to travel to a specific States. Prices will change when utilizing the filters.

A close up of a map

Description automatically generated

The vertical bar graphs show prices per year and quarter. The horizontal bar graph shows the cost to travel to a city sorted from least to most expensive. Adjusting the filters will change the two graphs.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The optional filters allow the user to select:

* Their Origin City
* Origin Airport
* Destination State
* Destination City
* Destination City
* Option to include specific years/quarters to their price results.

It is advised to utilize all filters to reach an accurate price.

Here is an example of how a user may want to use the dashboard. The dashboard shows how much it will usually cost for someone who travels from LA to NY cities during Q4. This could help the user gauge prices to travel during the holidays.

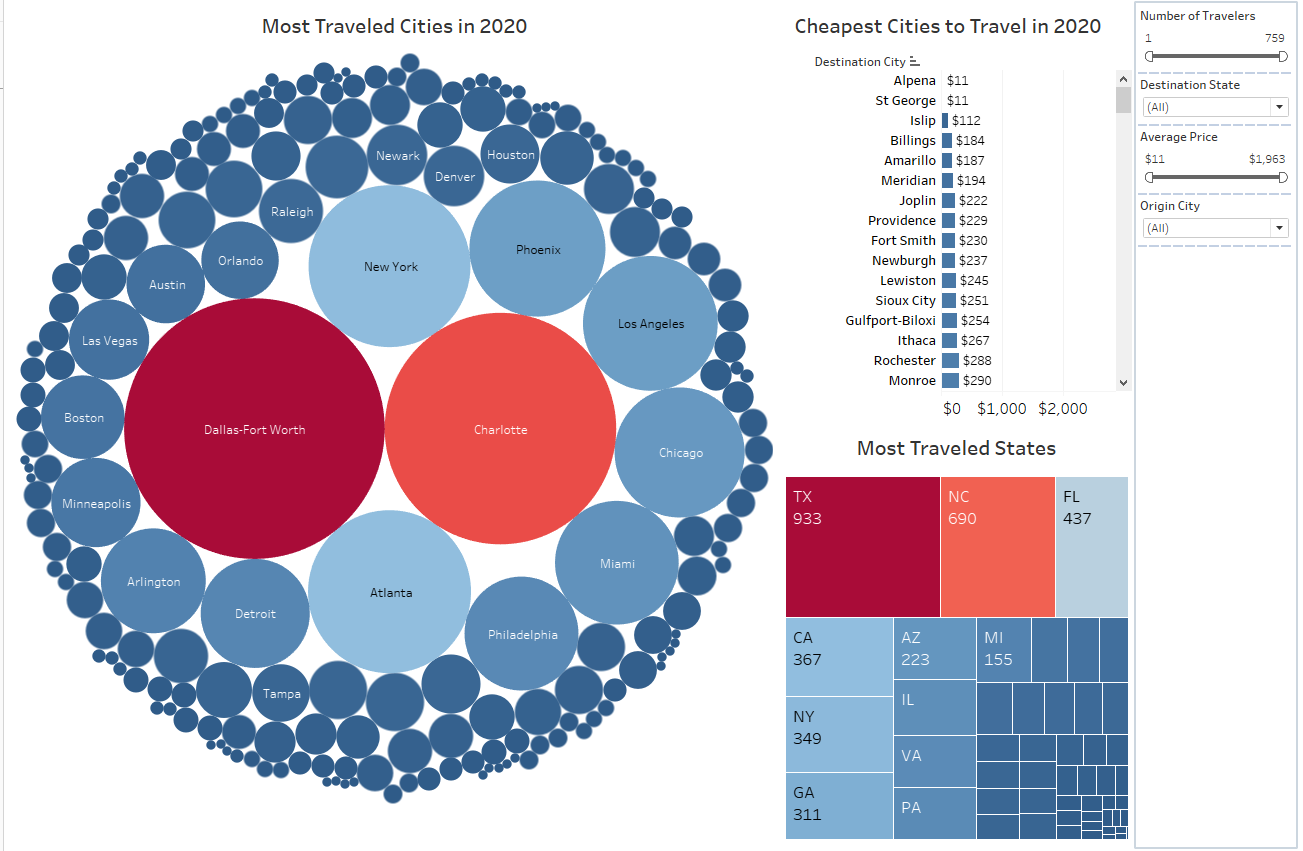
A screenshot of a cell phone

Description automatically generated

The Second Dashboard shows most traveled cities and states across the U.S. in Q1 2020. If a user wants to travel to a place with a limited range of travelers due to COVID, as well as select a destination within a price range they could utilize this dashboard to narrow their search. The optional filters allow the user to:

* Limit the number of travelers per destination
* Choose a desired state that they would like to travel to
* Set a specific price range
* And select an origin city for more accurate prices

This dashboard only contains data from Q1 2020. Q2 data has not been released yet. Prices and number of travelers may seem skewed when looking at it, but that is only because a lot of this was uploaded right before COVID started to really effect the airline industry.



The bubble chart and the tree-map highlight which city and state contains the most travelers recorded in 2020 Q1. Understanding the number of travelers could affect decision making when booking travel tickets. Users may want to know the number of travelers to specific areas to understand where a popular place is to travel, or to know which areas to stay away from to avoid COVID.

The Tree map displays how many people have traveled to different states in 2020. This would allow the user to understand number of travelers for each State. It will also provide the average cost to travel to that specific state.

A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The bar-chart displays cheapest cities to travel to in the U.S. the chart is sorted from least to greatest, so the cheapest city will always be the first one listed. The cities will change when the filters are utilized

A screenshot of a cell phone

Description automatically generated

**Responsibilities:**

* + Ahasan – ML, Data Exploration, HTML, Tableau
  + Artie – Tableau, Write up, Data Exploration
  + Kyle – HTML, Data Exploration, ML
  + Luis – ML, Data Exploration, HTML

**Inspiration:**

The inspiration for this project came from Ryan Zernach website on Airline Price Predictor.

<https://ryan.zernach.com/portfolio/airline-price-predictor-how-are-flight-prices-calculated/>

His work is good; however, we’re creating more interactive visualizations and a prediction model for ticket prices per airline.

A screenshot of a cell phone

Description automatically generated

**Footnotes**:

1. Research & Policy (https://www.transtats.bts.gov/)
2. Ryan Zernach (inspiration for project): (https://ryan.zernach.com/portfolio/airline-price-predictor-how-are-flight-prices-calculated/
3. Bureau of Transportation Statistics (https://www.bts.dot.gov/topics/airlines-and-airports-0)