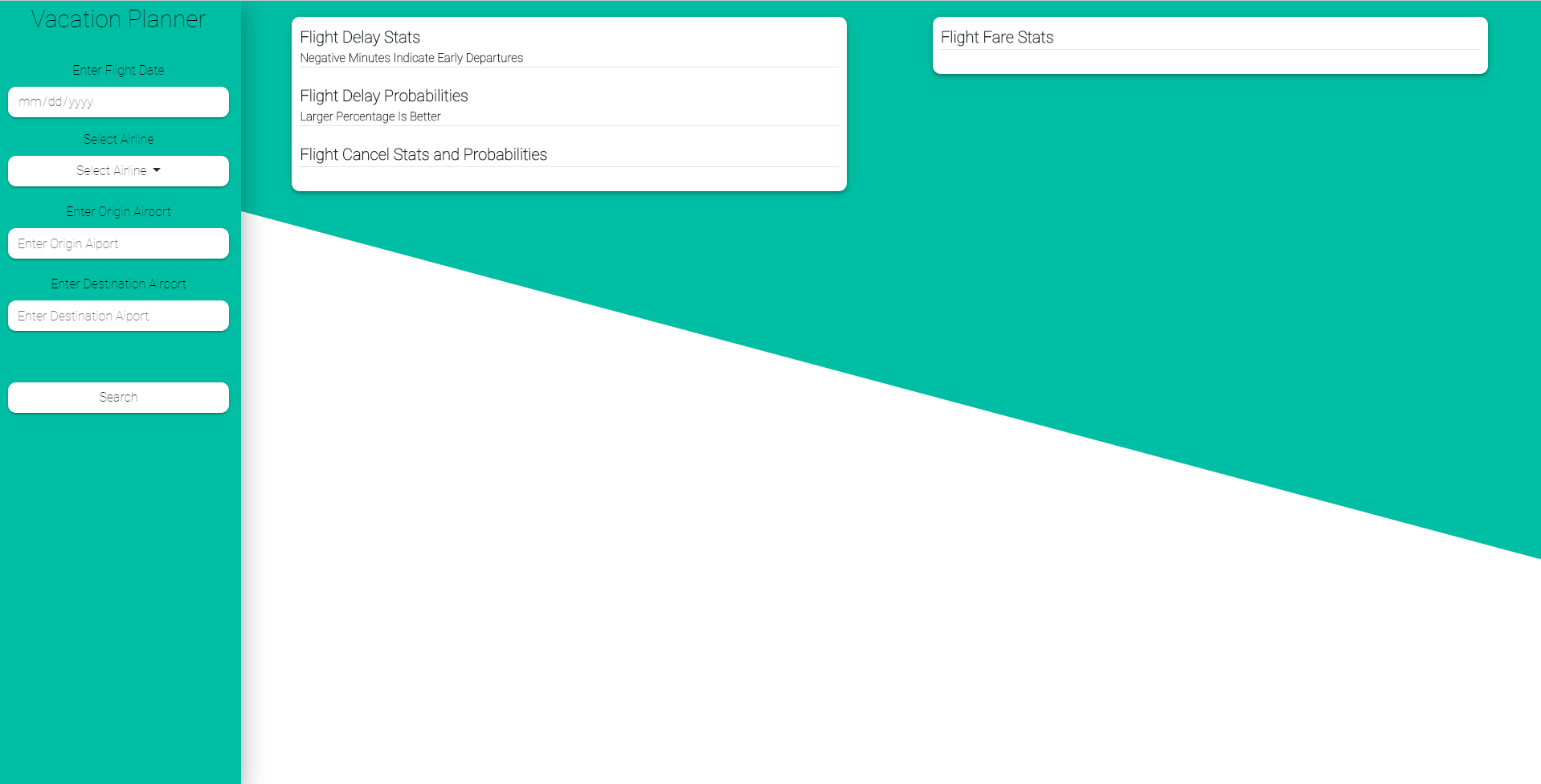
**Vacation Planner**

A Statistical Tool for Assisting in Planning Your Vacation

Cameron Stark

MA-412 Probability & Statistics Final Project

04/21/19

**Introduction**

The purpose of this report and the tool itself is to analyze the rate and probability at which flights get delayed between two different airports and the likely that the flight will get delayed based on previous years of information. The tool itself is built as a website, with Angular JavaScript as the core with an advanced array manipulation library called LoDash, and two math and statistics libraries called Simple Statistics and JStat.

The tool takes in 4 parameters from the user, being the date of the flight, the airline/carrier (which is consisted of all U.S based airlines from 2018, only one year because the data was already at over 1 million items), the airport of origin and the destination airport. With this the tool outputs the stats for all of the previous flights and also provides probabilities that the flight will get delay on the chosen date, aswell as giving the average air fare price and the probability that it will increase or decrease.

**Data Collection**

The source of the data was off the Bureau of Transportation Statistics website, specifically under the Airlines and Airports data section. For the Delay and Cancel data I combined the data for every airline for the past 4 years, which totaled out at over 700,000 items (in which each item counts as a flight) for each month. The Delay and Cancel has 6 columns of data being the flight date, the ICAO airline code, origin airport code, destination airport code, delay minutes (with positive being the delay minutes and negative being an early departure) and cancel column with 0 being no cancel and 1 being a cancelled flight. The flight fare data is a combination of the average fare of tickets for each quarter for each airport from the years of 2012-2018.

**Data Summary**

Delay and Cancel Data

The data itself appears to be normally or evenly distributed around the mean delay minutes for each month, however there is a slight right skew to the delay side because the average delay for each month is positive and the majority of flights are in fact a delay rather than an on time or even an early departure. The means and standard deviations for each of the twelve data tables (which are separated by which month they are associated with and contain around 600,000 items each) are displayed in the table below (these values are not based on airport or airline, they are the total values for all options).

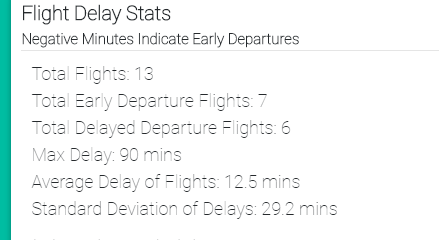
|  |  |  |
| --- | --- | --- |
| Month | Mean | Standard Deviation |
| January | 9.48 | 46.31 |
| February | 9.387 | 44.51 |
| March | 7.644 | 38.85 |
| April | 8.44 | 41.0 |
| May | 10.671 | 45.62 |
| June | 13.05 | 49.12 |
| July | 13.62 | 48.94 |
| August | 13.74 | 50.83 |
| September | 7.37 | 41.27 |
| October | 6.99 | 40.07 |
| November | 9.46 | 45.14 |
| December | 8.99 | 42.96 |

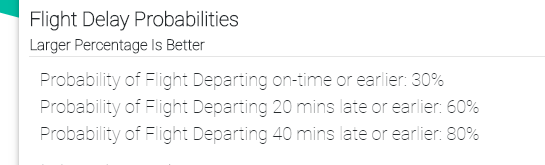
Airfare Ticket Prices

The airfare ticket prices are a basic mean and standard deviation of ticket prices based on past ticket prices and it independent of airline, the only narrowing factors are the airport and the date of the flight. The distribution of the data is represented below.

|  |  |  |
| --- | --- | --- |
| Quarter | Mean | Standard Deviation |
| Q1 | 443.53 | 174.043 |
| Q2 | 446.65 | 167.842 |
| Q3 | 443.55 | 166.958 |
| Q4 | 443.50 | 165.2621 |

**Statistical Analysis** The tool does its statistical analysis based on four inputs for each search, the date, the airline, the origin airport and the destination airport. With this information the tool parses through the over 1 million flight delay and cancels data rows, to get the rows that match all four criteria. The first test and walk through of the tools statistical analysis, is using the filters of May 16th, American Airlines, DFW airport and ORD airport. From this the tool sorts through the json files which contain the rows of data and then performs four categories of math on the resulting items, a basic overview of Delay Statistics, then Delay Probabilities, then a Delay Analysis and lastly a Cancel Stats and Probabilities.

 The Flight Delay Stats contains the total returned flights from the filters, the total flights that had a negative delay time which indicate an early departure, the total flights that were delayed, the max delay time, the mean of the total flights and then the standard deviation of the flights.

 The Flight Delay Probabilities section contains the probability the flight will either leave on-time or earlier (which is discovered by finding the cdf of the normal distribution of flights where the “X” is 0 mins), the probability the flight will leave at most 20 mins late or earlier (which is the same process but the “X” is 20 mins), then lastly with the “X” at 40 mins.

 The Flight Delay Analysis section contains the Confidence Intervals at 90%, 95% and 99%. These are obtained by using the confidence interval, the mean of filtered flights and the data set itself. The returned result is the range of values for the mean delay minutes.

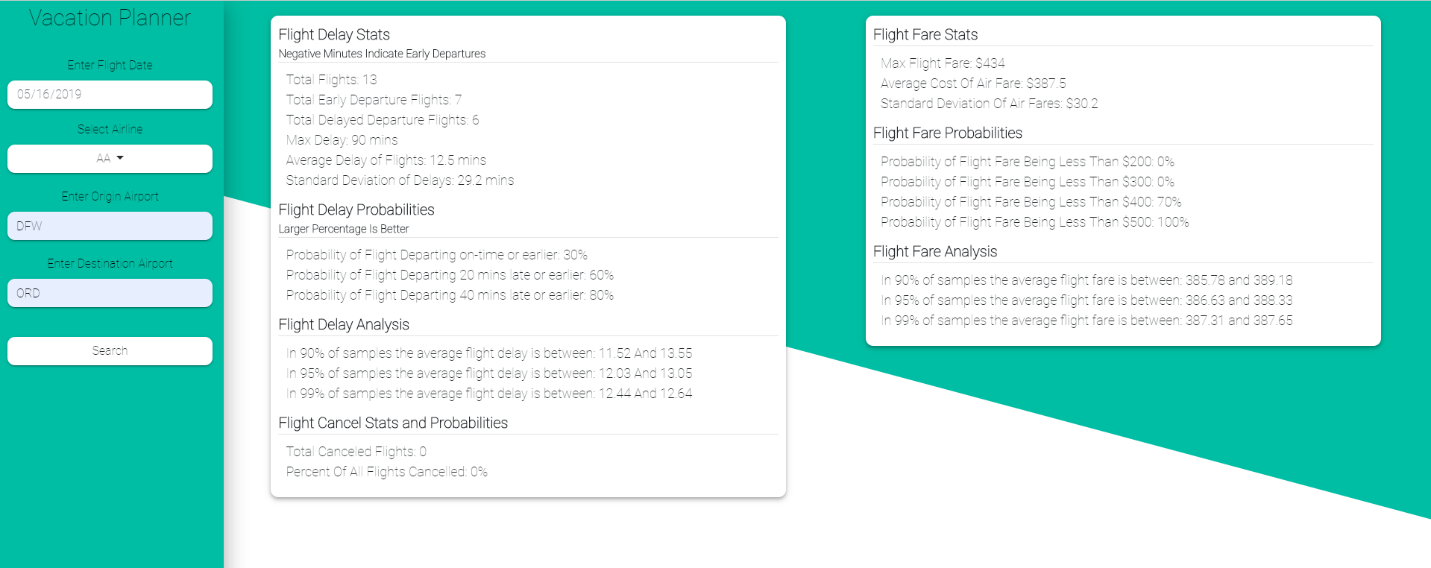
 The last section for the Delay and Cancel information is the Flight Cancel Stats and Probabilities, which doesn’t contain much since the information for cancels is just a binary 0 for no cancel or a 1 for a cancel, therefore the tool outputs the total amount of cancels for the filtered flights and the percent of the total flights that were cancelled.

The Flight Fare Section is broken down into sections the basic stats, the probability of the tickets being below certain prices and then confidence interval based on the filtered fares. The data for these sections is very simplistic therefore the amount of computations that can be done on them is minimal

 The Flight Fare Stats sections contains the max flight fare, the average flight fare and then the standard deviation of the flight fares. These are calculated by using the filtered results and performing average, max and standard deviation computations on them.

 The Flight Fare Probabilities section contains the probability that a flight fare is under a certain price of $200, $300, $400, and $500. This is calculated my taking the mean and standard deviation of the filtered fares and performing a normal cdf calculation on it with X value of each of the 4 prices above.

 The Flight Fare Analysis section contains the confidence interval for where the filtered fares mean likely lies within under a 90% confidence, 95% confidence and a 99% confidence. This is calculated by taking the filtered fares mean, the confidence interval and the filtered fares list, and performing the normal confidence interval calculations on them, which returns the confidence interval for each of the confidence levels.

 The Tool overall should be able to prepare anybody for their trip to the airport or help plan the family get away with a little less stress.

**Conclusion**

To conclude the tool acts as an almost complete statistical analysis and prediction tool. The tool covers basic statistical values such as mean and standard deviation of the data set. As well as advanced statistics with normal confidence interval and probabilities with normal cumulative distribution functions. The normal confidence interval uses the means, standard deviations and confidence interval values to determine the interval of confidence for the average delay of a flight or the average cost of an air fare. The Normal cumulative distribution function uses the expected value (which is the flight delay mins or the cost of a ticket), the mean of delay mins and the standard deviation of mins to determine the probability of a flight delay at various amounts.

**Steps to Run and Use Tool**

Extract the whole emailed folder, go into the folder named VacationPlanner and open the index.html file in any web browser of choice, by double clicking the file. From there wait approximately 1 or 2 mins for the data to load properly, which will be indicated by the white circle disappearing. From there then select a Date, Airline, Origin Airport and Destination Airport, then the search button will appear, click to run search.

**Sources**

https://www.bts.gov/topics/airlines-and-airports-0

https://simplestatistics.org/docs/#zscore

https://lodash.com/

http://jstat.github.io/