STAT 605 Project

Group 5: Abby Terzis , Anna Hayes, and Marwa Eltgani

Link to GitHub

Introduction

The data provides ticket prices for one-way flights found on Expedia. We sought to find if seatsRemaining and totalTravelDistance affect the price of a plane ticket and used Multiple Linear Regression to do this. The original data is a 31GB file of flight tickets found on Expedia between 04/16/2022 and 10/05/2022.

Statistical Analysis

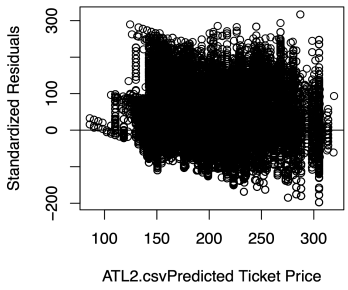
We split the large data set into 16 smaller files based on departing airport and randomly sampled 10% of the data. We used seatsRemaining and totalTravelDistance as predictors in all our 16 models for basic economy tickets. The summary table below shows the results we achieved from fitting the regression model for ATL airport.

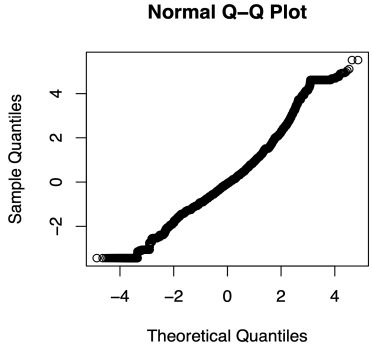
```
Call:
lm(formula = totalFare ~ seatsRemaining + totalTravelDistance,
    data = sampledat)

Residuals:
     Min       1Q   Median       3Q      Max
-196.820  -39.086   -3.466   32.400  284.010

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.135e+01  2.545e+00   28.04   <2e-16 ***
seatsRemaining      3.684e+00  2.948e-01   12.50   <2e-16 ***
totalTravelDistance 6.007e-02  4.138e-04  145.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.99 on 43083 degrees of freedom
  (177 observations deleted due to missingness)
Multiple R-squared:  0.3287,    Adjusted R-squared:  0.3287
F-statistic: 1.055e+04 on 2 and 43083 DF,  p-value: < 2.2e-16
```

The linear regression model for Atlanta is as follows:

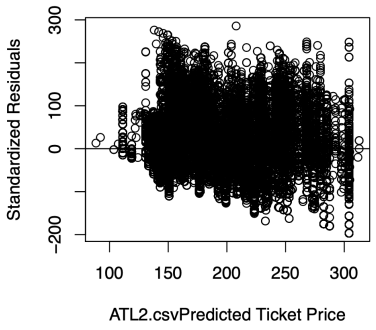
$$totalFare = 71.35 + 3.68seatsRemaining + 0.06totalTravelDistance$$

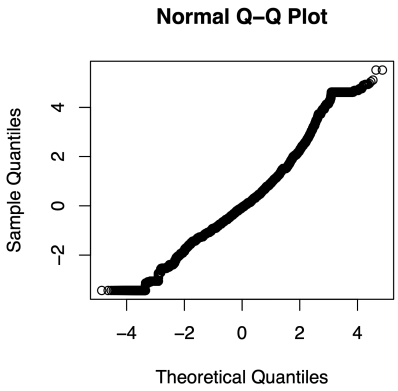
For each seat remaining, holding distance constant, the ticket price will increase by about $3.68. For each 100 miles traveled, holding seats remaining constant, the ticket price will increase by about $6. Since the p-value for both coefficients is below $\alpha = 0.05$, we can conclude that both seatsRemaining and totalTravelDistance are significant in predicting ticket prices.





The above plots are a couple diagnostic plots before taking the 10% random sample. As we can see, there are too many observations to get a good look at the underlying pattern. The below plots are after taking the 10% sample and unfortunately there are still so many observations but we did not want to lose the nature of the original data so we decided not to reduce the random sample further.





A weakness to our analysis is that we aren't taking into consideration a temporal effect of ticket prices so holidays like Memorial Day, July 4th and Labor Day are treated the same as other days.

Condor Details

We created a shell script to split the large 31GB dataset, filtered by basic economy tickets, and removed unnecessary columns to reduce memory, leaving us with 16 20 to 30 MB files. Afterwards, we tested our R script by running a regression model on only one csv (ATL) to ensure our scripts were working before running them on all 16 files. We initially used 1 GB of memory to run our first job but got an error message stating that more memory was required to run the job. We increased our memory to 5 GB but that also didn't work. We realized that lm() requires a lot of memory to run in CHTC so we ended up increasing our memory to 8 GB to get just one job to run. Once our first job was successful, we ran three more jobs and then finally, ran all 16 jobs which took about 6 minutes to run in total and required 20 GB of memory along with 2 GB of disk space.

Since linear regression requires a lot of memory to be able to run in CHTC, it may not be the best option in terms of memory when there are a hundred jobs to run. It was also hard to view visualizations on CHTC. The only way to view plots was by transferring the files to our local machine which can be time consuming when there are multiple jobs to run.

Conclusion

We wanted to see what factors affected the price of a plane ticket. Based on our analysis, seatsRemaining and totalTravelDistance significantly impact the price of a plane ticket regardless of the 16 departing airports we analyzed. For future analysis we could add the temporal effect of the ticket since we did not look at that aspect for this analysis.

Contributions

Abby, Anna, and Marwa all contributed to: writing the proposal, writing code, making the presentation, and writing the report.