This data set consists of domestic flight details from 2015. It includes features such as airlines, departure and destination airport, different types of delays and reasons, and trip information. The original data set in the flights.csv file consists of roughly 5.8 million observations. For this assignment,

please use the flight.csv file flights.csv . This assignment has two parts: 1) exploratory

analysis and 2) building a linear regression model.

Part I: Exploratory Analysis

The following questions are just starters, you are more than welcomed to explore more.

1. How many observations are there? How many features are there? (2pts)
2. How many flights arrived at SFO? How many airlines fly to SFO? (2pts)
3. How many missing values are there in the departure delays? How about arrival delays? Do they match? Why or why not? Remove these observations afterwards. (5pts)
4. What is the average and median departure and arrival delay? What do you observe? (2pts)
5. Display graphically the departure delays and arrival delays for each airline. What do you notice? Explain. (2pts)
6. Now calculate the 5 number summary (min, Q1, median, Q3, max) of departure delay for each airline. Arrange it by median delay (descending order). Do the same for arrival delay. (2pts)
7. Which airline has the most averaged departure delay? Give me the top 10 airlines. (2pts)
8. Do you expect the departure delay has anything to do with distance of trip? What about arrival delay and distance? Prove your claims. (2pts)
9. What about day of week vs departure delay? (2pts)
10. If there is a departure delay (i.e. positive values for departure delay), does distance have anything to do with arrival delay? Explain. (My experience has been that longer distance flights can make up more time.)

(2pts)

(2pts)

.1. Are there any seasonal (monthly) patterns in departure delays for all flights? (2PTS)

Part II: Regression Analysis

Now we want to build a model to analyze the arrival delay. We will use linear regression

here.

Subpart I

1. Your response is ARRIVAL DELAY. First, remove all the missing data in the WEATHER DELAY column. Once you do this, there shouldn't be any more missing values in the data set (except for the cancellation reason feature). Check that. (2pts)

2. Build a regression model using all the observations, and the following predictors:

[LATE AIRCRAFT DELAY, AIR SYSTEM DELAY, DEPARTURE DELAY , WEATHER DELAY, SECURITY DELAY, DAY OF WEEK, DISTANCE, AIRLINE] a total of 8 predictors. (5pts)

3. Perform model diagnostics. What do you observe? Explain. (5pts)

4. Provide interpretations for a few of the coefficients, and comment on whether they make sense. (3pts)

Subpart II

If you have done the above steps correctly, you will notice a lot of things "doesn't seem right". We will try to fix a couple of these things here.

1. Removing outliers: first is to remove outliers. Using the boxplot method, remove the outliers in the ARRIVAL DELAY variable. (2pts)

2. Refit the linear regression model, but now with log(ARRIVAL DELAY) as your response. Also, remove the non-significant predictors from the previous model (with p-values larger than 0.05). (Remember that when removing non-significant predictors one can only eliminate one variable per step, but for now we will ignore this rule and remove everything in one step.)

Also take the log transform of a DELAY variable and the square of another DELAY variable of your choice.

(5pts)

3. Perform model diagnostics. Did anything improve? (5pts)

4. Provide interpretations to a few of the coefficients. Do you think they make sense? (3pts)

5. Obviously there's still a lot that needs to be done. Provide a few suggestions on how we can further improve the model fit (you don't need to implement them). (5pts)