Group Homework 2

STAT011-F23

Introduction and Purpose

The data set you will analyse in this homework records information from different flights for an airline carrier named Envoy Air in the year 2013. The purpose of this assignment is to practice using software to investigate and describe a rather large data set consisting of 19 variables and $25,037$ observations.

Required Tech

Excel

The skills necessary to complete this assignment in Excel are covered in the following seven videos:

- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak (1:52)
- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak: Common Procedures (3:08)
- Excel 2016 with Data Analysis Toolpak Descriptive Statistics and Confidence Intervals for a Mean (2:57)
- Excel 2016 with Data Analysis Toolpak Histogram (3:08)

RStudio

The skills necessary to complete this assignment in RStudio are covered in the following seven videos:

- R Studio Video Introduction to R and RStudio (1:52)
- R Studio Video Getting Started (3:51)
- R Studio Video Working with Data Objects 1 (3:29)
- R Studio Video Working with Data Objects 2 (5:15)
- R Studio Video Importing Data (4:36)
- R Studio Video Descriptive Statistics (3:33)
- R Studio Video Plotting - Histograms, Bar Charts, Boxplots, Scatterplots (4:08)

Instructions

If you are analyzing this data in Excel you first need to download the data set for HW 1 from our Stat 11 Github Data page. Do this by right clicking on the link "View Raw" and save the link with the name `EnvoyAir_flights.csv`. This may take a few moments as it is a large data set.

If you are analyzing this data in RStudio, you will import the data with the following command

```
EnvoyAir_flights <- read.delim(
        "https://raw.githubusercontent.com/ProfSuzy/Stat11/main/Data/EnvoyAir_flights.txt",
        sep=",")
```

This may take a few moments as it is a large data set. The data object is called `EnvoyAir_flights`.

Once you have access to the data set, complete all parts of the five problems in this assignment. You are encouraged to work with your classmates on this assignment but you must hand in your own, unique write up of the solutions. In a Word document, clearly label each problem's solution. Most solutions will include graphics which can be copied from Excel or RStudio and pasted into your solution document. All solutions require a written component. When you are ready to submit your assignment, save the Word document as a PDF and upload it to the Moodle link for Group Homework #2.

Problem 1

Answer the following questions about the data set called `EnvoyAir_flights`:

1. What constitutes an observational unit?
2. What are the different variables being collected?
3. Which of the variables are quantitative and which are categorical? Are there any that could be both?
4. What kind of relationship would you expect (if any) between the variables `dep_delay` (departure delay) and `arr_delay` (arrival delay)?

Solution 1

1. Individual flights conducted by this airline are the observational units.
2. and 3. There. are 19 different variables. Carrier, flight, tailnum, origin, and dest are all categorical variables. dep_delay, arr_delay, air_time, distance, hour, minute, and time_hour are most likely to be used as numerical variables. The other variables (year, month, day, dep_time, sched_dep_time, arr_time, and sched_arr_time) could reasonably be treated as either categorical or numerical.
3. One might anticipate a linear relationship between any delays in departure and any delays in arrival since one typically (though not always) corresponds to the other. Sometimes pilots may be able to make up for lost time by flying faster or along a different route, for example.

For graders: All questions must be answered correctly for full credit.

Potentially useful R Code:

```
names(EnvoyAir_flights)
```

```
##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"
```

Problem 2

Perform a one variable exploratory data analysis by providing of the variable `distance` by creating:

1. A labeled histogram (with a reasonable number of bins);
2. A five-number summary of the minimum, maximum, lower 25% quantile, lower 75% quantile, and the mean.

What story do these analyses tell about the variable `distance`?

Solution 2

The distance traveled by these Envoy Air flights does not seem to follow a unimodal distribution. There are at least two (possibly four) prominent modes around which the data is clustered. The data is roughly symmetric however which we can tell from the histogram and by the fact that the mean and median are so close (also we can see this from the fact that the difference between the minimum and first quartile approximately the same as the distance between the maximum and the third quartile, the latter distance being a bit longer indicates a slight right skew but not a severe one). The majority of the flights are between 400 and 600 miles long. There are almost no flights between 800 and 1000 miles long. In summary, very long flights would be those over 1000 miles and very short flights are under 400 miles. These very long and very short flights don't occur as often as the mid-range flights but there is a noticeable dip in the amount of flights around the mean. From this we can gather that this is a regional airline and not one that provides cross continental travel.

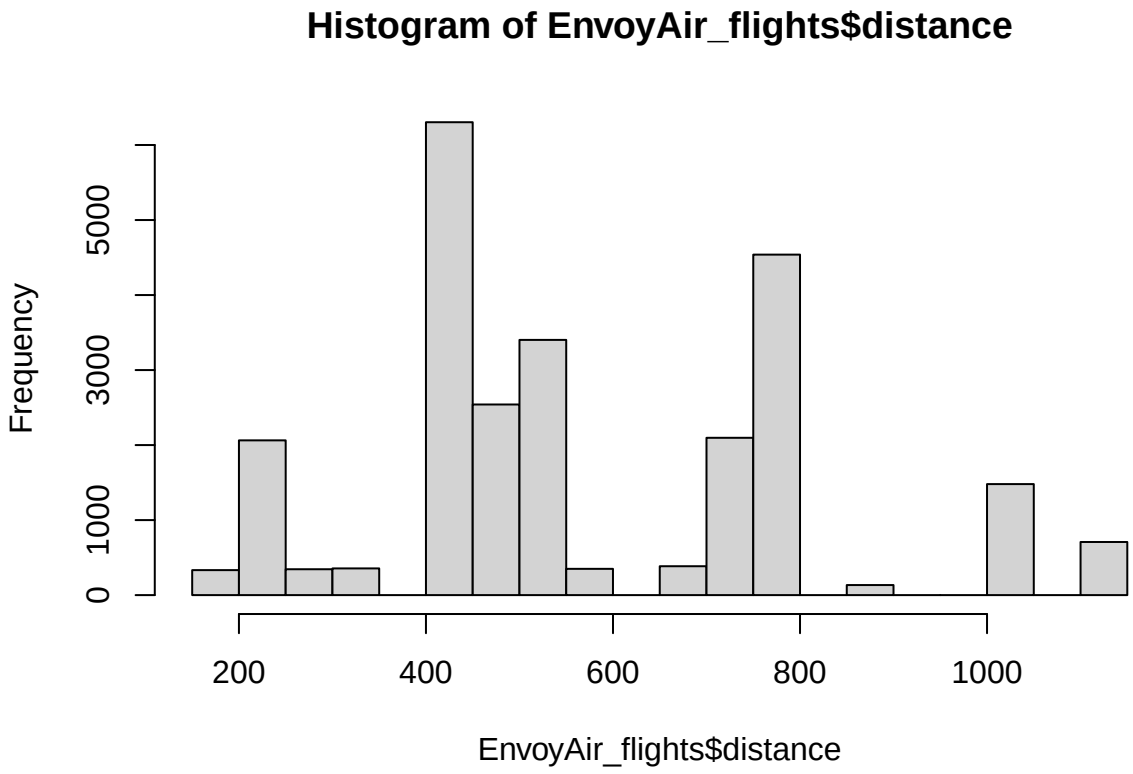
For graders: Must include the five number summary, a histogram, and an interpretation of these to receive full credit.

Potentially useful R Code:

```
summary(EnvoyAir_flights$distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   184.0   431.0   502.0   570.4   762.0  1147.0
```

```
hist(EnvoyAir_flights$distance, breaks=20)
```



Problem 3

Create a labeled bar/frequency chart for the variable `origin`. What story does this tell about the variable `origin`?

Solution 3

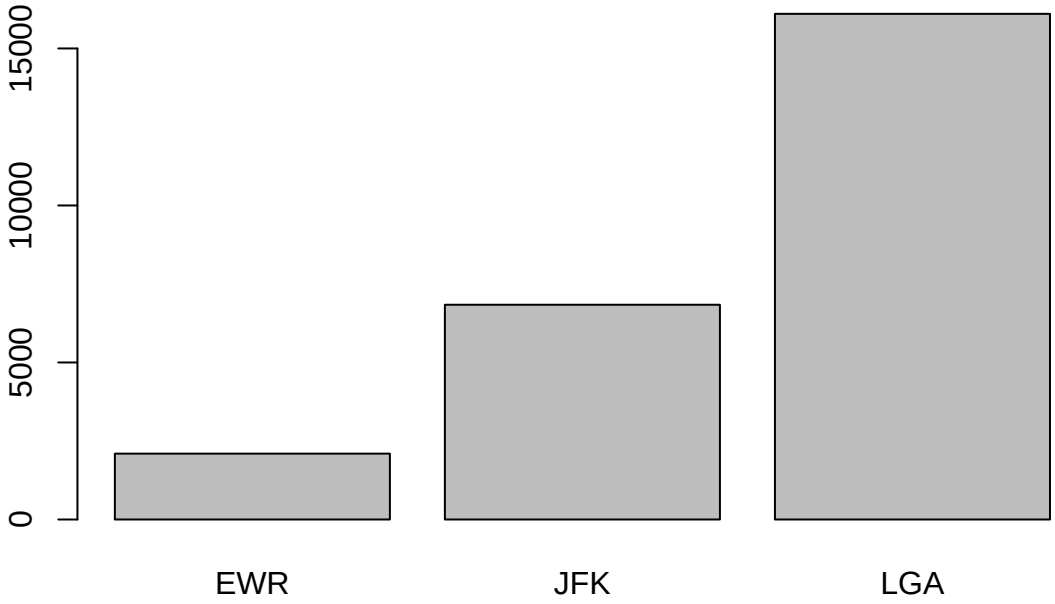
Most Envoy Air flights start in LGA, about half as many originate at JFK and even fewer from EWR.

Potentially useful R Code:

```
table(EnvoyAir_flights$origin)
```

```
##
##   EWR   JFK   LGA
##  2097  6838 16102
```

```
barplot(table(EnvoyAir_flights$origin))
```



Problem 4

To compare the flight distance across all airports of origin we could either create three boxplots or create three histograms, one for each airport of origin. Create a boxplot for the distance of Envoy Air flights by each of the three airports of origin. Then, answer the following two questions. What story does this plot tell you about any apparent relationship between these variables? Hypothetically, is there something else you could see if you plotted three separate histograms for distance (one for each origin) instead of three boxplots?

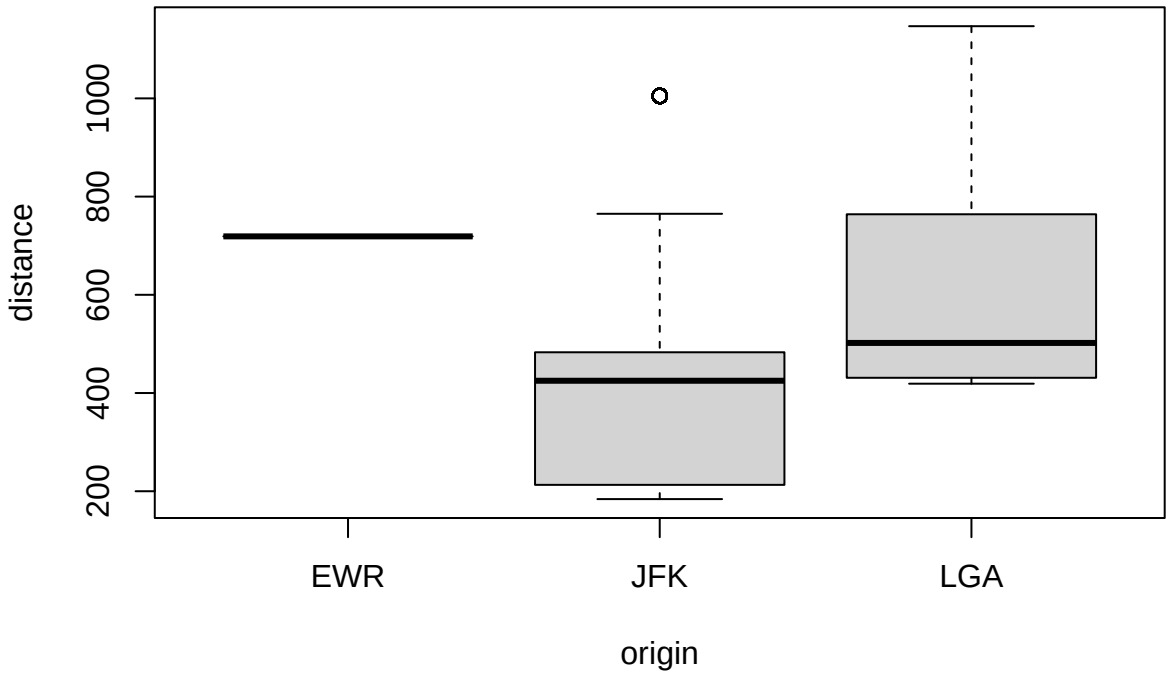
Solution 4

Right away, we can see that flights out of EWR only go to a single destination. Flights originating in LGA tend to fly longer distances than those from JFK, and, flights out of LGA cover a broader range of distances ($\approx 1700 - 400 = 1300$) than those from JFK ($\approx 1000 - 200$). There is one flight out of JFK that is much longer than the others at around 1000 miles (assuming the units of distance are miles). If we were to compare three histograms for distance as opposed to three boxplots, we would see a strong right skew for LGA and a single, large bin for EWR. The histogram for JFK may appear somewhat symmetric because of the large outlier and because the median is also quite large.

For graders: For full credit, answers must include a boxplot for each airport of origin in addition to answering the two questions above.

Potentially useful R Code:

```r
boxplot(distance ~ origin, EnvoyAir_flights)
```



Problem 5

Generate a scatter plot of the variables `dep_delay` (departure delay) and `arr_delay` (arrival delay). Then, answer the following questions. What story does this scatter plot tell about any apparent relationship between these variables? Is this consistent with your answer from Problem 1?

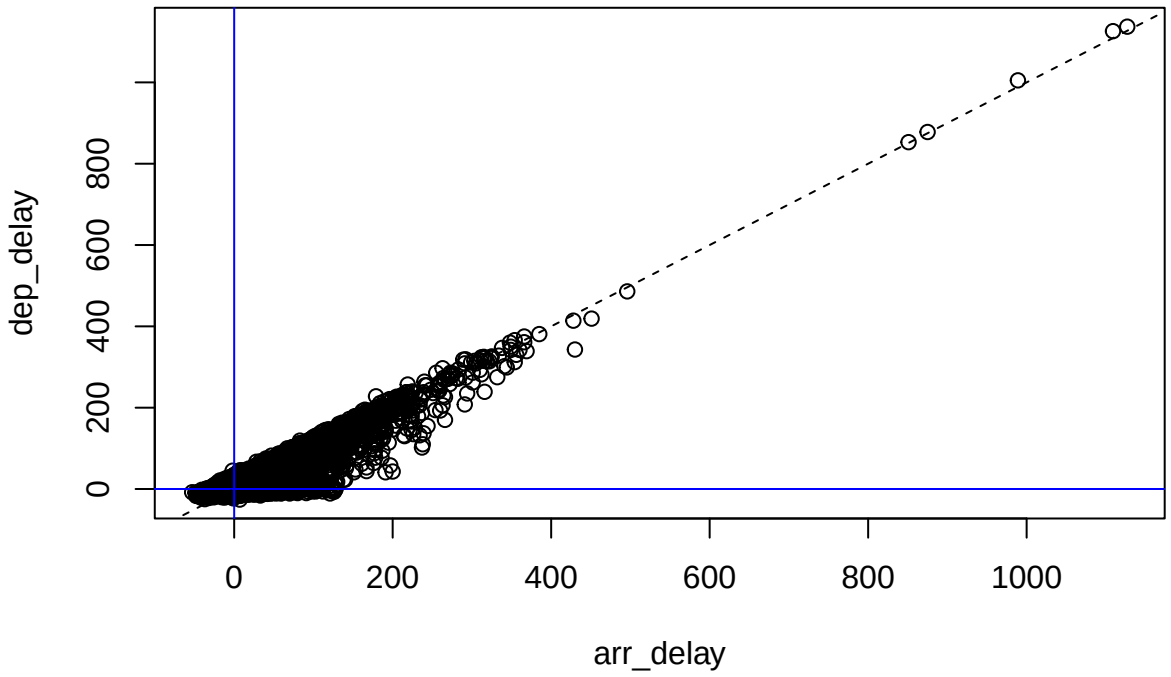
Solution 5

From the scatterplot we see a strong, positive linear trend between these two variables. More specifically, the data falls really close to the diagonal line $y = x$ (dotted line below), which is what we might expect since it seems likely that all other things equal, the amount of time a flight is delayed departing will also be the amount of time it's delayed arriving. In the plot below, we see that the most deviation from this linear trend occurs in the negative direction. This means that when the arrival and departure delays are not the same, the arrival delays tend to be longer than the departure delays. (As a student pointed out in office hours, this makes sense as there is a limit to how many flights can fit at the airline's gates.) Finally, although the five data points in the upper right hand corner are further away from the rest of the data, they are not likely to be very influential as removing them doesn't look like it would change the line of best fit through all the data points.

For graders: (The lines aren't necessary) For full credit answers must display this scatterplot and contain answers to both questions above.

Potentially useful R Code:

```r
plot(dep_delay ~ arr_delay, EnvoyAir_flights)
abline(h=0, col="blue")
abline(v=0, col="blue")
abline(a=0, b=1, lty=2)
```


dep_delay
0
200
400
600
800
0
200
400
600
800
1000
arr_delay