Project Homework 2

STAT011-S23

Due: Friday, Sept 29 by 11:59pm

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 four 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)

R

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

- R Video - Introduction to R (0:59)
- R Video - Getting Started (4:01)
- R Video - Working with Data Objects 1 (3:08)
- R Video - Working with Data Objects 2 (4:28)
- R Video - Importing Data (03:20)
- R Video - Descriptive Statistics (3:34)
- R Video - Plotting - Histograms, Bar Charts, Boxplots, Scatterplots (3:56)

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 R, you will import the data with the following command

```
EnvoyAir_flights <- read.delim(
        "https://raw.githubusercontent.com/dr-suz/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 R 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)?

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`?

Problem 3

Create a labeled bar/frequency chart for the variable `origin`. What story does this tell about the variable `origin`? (Hint for R users: First create a table from this data and then put that table inside the `barplot` function.)

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?

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?