**Lab 2 – 60 points**

**Name: Christian M. Bridge**

**Note:** The word document of this assignment can be downloaded from CANVAS course page.

Please first read instructions and/or examples below and then **answer the numbered questions below**. For all Labs, you many use any software tool you want (Excel, R, Python, SAS, SPSS, etc.).

For Download instructions, see the Statistical Software page in Canvas: <https://unt.instructure.com/courses/67141/pages/the-statistical-software>

In this lab, we are going to look at both numerical and categorical data using descriptive measures and graphical techniques. Open the nycflights.csv, a random sample of flights departing New York City airports in 2013. If something is wrong with the data set in Canvas, a copy is available at <https://www.openintro.org/book/statdata/index.php?data=nycflights>.

.

**TASKS:**

1. **Using the “distance” variable, answer the following questions:**
   1. **What is the mean? (4 points) \_\_\_\_\_\_\_1,046.24\_\_\_\_\_\_\_**
   2. **What is the median? (4 points) \_\_\_\_\_\_\_888.00\_\_\_\_\_\_\_\_\_**
   3. **What is the variance? (4 points) \_\_\_\_\_\_538,289.14\_\_\_\_\_\_\_\_\_\_**
   4. **What is the standard deviation? (4 points) \_\_\_\_\_\_\_\_733.68\_\_\_\_\_\_**
   5. **What is the Inter Quartile Range? (4 points) \_\_\_\_\_\_\_889.0\_\_\_\_\_\_**
2. **Plot the “distance” variable in a histogram? Which of the following responses best describes the skewness of the histogram? (4 points) \_\_\_\_\_\_\_\_C\_\_\_\_\_\_**
   1. **Symmetric**
   2. **Left skewed**
   3. **Right skewed**
3. **Plot a scatterplot of the two variables “dep\_delay” and “arr\_delay”. Which response best describes their relationship. (4 points) \_\_\_\_A\_\_\_\_\_**
   1. **The variables have a positive linear correlation**
   2. **The variables have a negative linear correlation**
   3. **The variables have a non-linear association**
4. **Since negative times represent early departures/arrivals, a flight was “Delayed” if the dep\_delay value is greater than zero and a flight is “Late” if the arr\_delay value is greater than zero. Given those definitions, fill out this contingency table (12 points, 1.5 points each):**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Delayed Departure | Not Delayed Departure | Total |
| Late Arrival | 9291 | 4171 | 13462 |
| Not Late Arrival | 3508 | 15765 | 19273 |
| Total | 12799 | 19936 | 32735 |

1. **Convert the values in the table you just filled out in Question 4 to proportions and fill out this frequency table (12 points, 1.5 points each):**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Delayed Departure | Not Delayed Departure | Total |
| Late Arrival | 28.38 | 12.74 | 41.12 |
| Not Late Arrival | 10.72 | 48.16 | 58.88 |
| Total | 39.10 | 60.90 | 100% |

1. **Transform the “distance” variable to log(distance), using base 10 log. What is the mean of this new variable? (4 points) \_\_\_\_\_\_\_2.90\_\_\_\_\_\_\_\_**

1. **Plot a histogram of the “minute” variable? Which modality best describes this variable? (4 points) \_\_\_\_\_\_C\_\_\_\_\_\_\_\_\_**
   1. **Unimodal**
   2. **Bimodal**
   3. **Multimodal**