**Assignment #6: Introduction to Working with R/RStudio** (Due Sunday, **April 9**, 2017 at **9:00 am**)

**Before you start**

For this assignment, you’ll run simple analyses by modifying the R script you used in the ICA #10.2 (**Descriptives.r**). You will also need a new data set – **OnTimeAirport-Jan14.csv**, which contains actual data regarding on-time flight statistics for 84,656 flights, by airline and airport, for January 2014.

**IMPORTANT!** When downloading the .csv file, please make sure that the name doesn’t change, and that it is in the same folder as the *Descriptives.r file that you are modifying*.

The metadata for the **OnTimeAirport-Jan14.csv** spreadsheet is below:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| **FlightDate** | The date of the flight (mm/dd/yyyy) |
| **UniqueCarrier** | The unique carrier code |
| **AirlineFullName** | The full name of the airline |
| **AirlineID** | The numeric ID of the airline |
| **Origin** | The origin airport of the flight |
| **OriginCityName** | The origin city of the flight |
| **DestCityName** | The destination city of the flight |
| **DepDelayMinutes** | The delay in departing from the origin gate |
| **TaxiOut** | The minutes spent taxiing out to the runway at origin |
| **TaxiIn** | The minutes spent taxiing in from the runway at destination |
| **ArrDelayMinutes** | The delay in arrive to the destination gate |
| **Cancelled** | Whether the flight was cancelled (0 = no, 1 = yes) |
| **Distance** | The total distance of the flight |

**What to Submit**

Submit the following four files through Blackboard:

1. The completed, working **R script** that produced the analysis in Steps 1 through 8.
2. The output file – **descriptivesOutput.txt**
3. Another output file – **histogram.pdf**.
4. The completed **answer sheet** provided on the last page and also as a separate word file.

*If you do not follow the above instructions, your assignment will be counted late.*

**Guidelines**

To complete the assignment, modify the Descriptives.r script to perform an analysis of departure delays by origin airport, following the instructions below, and complete the answer sheet on the last page.

1. Use OnTimeAirport-Jan14.csv as the input file.

|  |
| --- |
| HINT: In line 21 of the Descriptives.r script, it says:  INPUT\_FILENAME <- "NBA14Salaries.csv"  Change that line to:  INPUT\_FILENAME <- "OnTimeAirport-Jan14.csv" |

1. Present the number of flights, grouped by origin airport.

|  |
| --- |
| HINT: In line 61, change the line to read:  summary(dataSet$Origin)  This presents the number of observations/rows (flights) by Origin airport. You will need the output from this command to answer the first question in the answersheet on the last page.  In addition, add a “#” at the beginning of line 66, so that this line will be treated as a comment:  #describe(dataSet$Salary) |

1. Present summary statistics for departure delay (using DepDelayMinutes).

|  |
| --- |
| HINT: In line 66, change the line by replacing Salary with DepDelayMinutes:  describe(dataSet$DepDelayMinutes) |

1. Present summary statistics for departure delay (using DepDelayMinutes), grouped by origin airport.

|  |
| --- |
| HINT: Check line 73 in the script:  describeBy(dataSet$Salary,dataSet$Position)  This presents summary statistics for salary by position (for the NBA salary data).  Now that we are using a different data set, you should be able to figure out how to change line 73 to present summary statistics for departure delay (DepDelayMinutes), grouped by origin airport (Origin).  If you get that, you will now be able to answer questions 2 through 4 on the answer sheet! |

1. Compare, using a t-test, the departure delays from Philadelphia (PHL) versus Pittsburgh (PIT). (PHL and PIT are both origin airports.)

|  |
| --- |
| HINT: Now please change line 87 and line 93 on your own. Hopefully the first few steps will get you started!  Check line 87:  subset <- dataSet[ which(dataSet$Position=='PG' | dataSet$Position=='SF'), ]  This create a subset with only two positions: PG and SF (for the NBA salary data). Now that we are using a different data set, you should be able to figure out how to change this line to create a subset with only two origin airports: PHL and PIT.  Check line 93:  t.test(subset$Salary~subset$Position)  This runs a t-test by using Salary as your dependent variable and Position as your grouping variable (for the NBA salary data). Now with the airport data, you should be able to change this line by using DepDelayMinutes as the dependent variable, and Origin as the grouping variable. |

1. Create a histogram, properly labeled, of the overall distribution of departure delays for all flights.

|  |
| --- |
| HINT: You will need to change the hist() function in both line 106 and line 112. You also need to change line 25 & line 27 for the label and title of the histogram. |

Once you’ve completed this part, add several new lines to the script that does the following (see next page):

|  |
| --- |
| NOTE: Make sure you add these lines right before the sink() function (line 96) so that the results are included in your text file output. |

1. Use **describeBy()** to compare the flight distance (Distance) across airlines (using AirlineFullName).
2. Use **describeBy()** to compare the taxiing out time (TaxiOut) across origin airports (Origin).
3. Answer this question using a t-test: Do planes spend more time taxiing out to the runway in Chicago (ORD) or Los Angeles (LAX) as the origin airport? (using TaxiOut as the taxiing out time);

**Once you’ve completed all the 9 steps, you can set the working directory and run the script. Based on your script output, answer the 11 questions listed on the answersheet on the next page.**

**Answer Sheet on the Next Page……**

Answer Sheet for Assignment: Introduction to Working with R/RStudio

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Answer the questions below based on your script output*

|  |  |  |
| --- | --- | --- |
| Question | | Answer |
| 1 | How many total flights (including cancelled flights) came out of Philadelphia (PHL) during January 2014? |  |
| 2 | What was the average departure delay across all flights during January 2014? |  |
| 3 | What was the average departure delay for Philadelphia (PHL) during January 2014? |  |
| 4 | What was the longest departure delay for Los Angeles (LAX) during January 2014? |  |
| 5 | On average, which city experienced greater departure delays: Philadelphia (PHL) or Pittsburgh (PIT)? |  |
| 6 | For question #5, was this difference statistically significant? What is the p-value? (answer both questions in the blank to the right) |  |
| 7 | Which airline(s) had longest average flight distance? (you can list more than one if it’s a tie) |  |
| 8 | Which airline (s) had shortest average flight distance?  (you can list more than one if it’s a tie) |  |
| 9 | On average, which city experienced greater taxi out times: Chicago (ORD) or Los Angeles (LAX)? |  |
| 10 | For question #9, was this difference statistically significant? What is the p-value? (answer both questions in the blank to the right) |  |
| 11 | Looking at the histogram. Is the distribution symmetric? Are most flights delayed less than 30 minutes or more than 30 minutes? |  |