**Assignment #5: Introduction to Working with R/RStudio**

|  |
| --- |
| **Submission Instructions****Due:** **Tuesday, April 2nd, 2019 at 9:30 AM.*** Submit the following four files through Canvas>Assignments> Assignments:
1. The completed, working **R script** that produced the analysis in Steps 1 through 9
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 instructions, your assignment will be counted late.
	+ **Late Assignment policy**: All assignments will be assessed a **50% penalty** (subtracted from that assignment's score) for the first day (i.e. 24 hours) they are late.
* Your submission will be graded based on the correctness of the completed answer sheet, with other files as supporting documents.
 |

**Before you start**

For this assignment, you’ll run simple analyses by modifying the R script you used in the ICA #10 (**Descriptives.r**). You will also need a new data set – **OnTimeAirport2017Dec.csv**, which contains actual data regarding on-time flight statistics for 83,915 flights, by airline and airport, for December 2017, collected from Bureau of Transportation Statistics.[[1]](#footnote-1)

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 – **OnTimeAirport2017Dec.csv** spreadsheet is below:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| FlightDate | The date of the flight (mm/dd/yyyy) |
| UniqueCarrier | The unique carrier code |
| CarrierlName | The name of the carrier |
| FlightNum | Flight Number |
| Origin | The origin airport of the flight |
| OriginCity | The origin city of the flight |
| Dest | The destination airport of the flight |
| DestCity | The destination city of the flight |
| DepDelay | The delay in departing from the origin gate (in minutes) |
| TaxiOut | The minutes spent taxiing out to the runway at origin |
| TaxiIn | The minutes spent taxiing in from the runway at destination |
| ArrDelay | The delay in arrive to the destination gate (in minutes) |
| Cancelled | Whether the flight was cancelled (0 = no, 1 = yes) |
| AirTime | Flight Time (in minutes) |
| Distance | The total distance of the flight (in miles) |

**Modifying the Descriptives.r script**

To complete the assignment, modify the **Descriptives.r** script (used in ICA #10) 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 OnTimeAirport2017Dec.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 <- "OnTimeAirport2017Dec.csv" |

1. Present the number of flights, grouped by destination airport (using Dest).

|  |
| --- |
| HINT: In line 61, change the line to read:summary(dataSet$Dest)This presents the number of observations/rows (flights) by destination airport. You will need the output from this command to answer the first question in the answersheet on the last page. |

1. Present summary statistics for arrival delay (using ArrDelay).

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

1. Present summary statistics for arrival delay (using ArrDelay), grouped by airline carriers (using UniqueCarrier).

|  |
| --- |
| 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 arrival delay (ArrDelay), grouped by airline carrier (UniqueCarrier).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 arrival delays for two airline carriers (using UniqueCarrier), American Airlines (AA) and United Airlines (UA).

|  |
| --- |
| 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 airline carriers: AA and UA.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 ArrDelay as the dependent variable, and UniqueCarrier as the grouping variable. |

1. Create a histogram, properly labeled, of the overall distribution of arrival delays (using ArrDelay) 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. In addition, in line 24, change the number of breaks (NUM\_BREAKS) to 50 so you will see more vertical bars in the histogram. |

Once you’ve completed this part, ***add several additional lines*** to the script that does the following 7), 8), and 9):

|  |
| --- |
| 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 UniqueCarrier).
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 Newark (EWR) or Philadelphia (PHL) as the origin airport? (using TaxiOut as the taxiing out time, and Origin as the origin airport);

**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) have Philadelphia (PHL) as the destination airport during December 2017? |  |
| 2 | What was the average arrival delay (in minutes) across all flights during December 2017? |  |
| 3 | What was the average arrival delay (in minutes) for American Airlines (with UniqueCarrier code of AA) during December 2017? |  |
| 4 | What was the longest arrival delay for United Airlines (with UniqueCarrier code of UA) during December 2017? |  |
| 5 | On average, which airline (using UniqueCarrier) experienced greater arrival delays: American Airline (AA) or United Airlines (UA)? |  |
| 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 origin airport (using Origin) experienced greater taxi out times: Newark (EWR) or Philadelphia (PHL)? |  |
| 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 50 minutes or more than 50 minutes? |  |

1. https://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID=236 [↑](#footnote-ref-1)