CPTS 575 Data Science

Assignment 4

Jinyang Ruan

011696096

Libraries prepare:

Text

Description automatically generated

Data prepare:

Text

Description automatically generated

Problem 1.

1.a Filter the dataset (using a left join) to display the tail number, year, month, day, hour, origin, and humidity for all flights heading to Tampa International Airport (TPA) after 12pm on November 1, 2013.

Graphical user interface, text

Description automatically generated

Table

Description automatically generated

1.b What is the difference between the following two joins?

* anti\_join(flights, airports, by = c("dest" = "faa")): this operation will drop from table flights all observations that have a match with the condition ("dest" = "faa") in table airports. the result is a subset of table flights.
* anti\_join(airports, flights, by = c("faa" = "dest")): this operation will drop from table airports all observations that have a match with the condition ("dest" = "faa") in table flights the result is a subset of table airports.

1.c Filter the table flights to only show flights with planes that have flown at least 100 flights. Hint: tailnum is used to identify planes.

I am not sure whether the “year” in the planes table represents the same thing as the “year” in the table flights. If not, semi-join the planes table with flights by “tailnum”.

Text

Description automatically generated

Totally there are 1118 flights with planes that have flown at least 100 flights.

Table

Description automatically generated

If we semi-join the planes table with flights by “tailnum” and “year”:

Text

Description automatically generated

Totally there are 11 flights with planes that have flown at least 100 flights.

Table

Description automatically generated

1.d What weather conditions make it more likely to see a delay? Briefly discuss any relations/patterns you found.

Text, letter

Description automatically generated

Generally, I left-join the weather table with flights, compute the total delay time (|dep\_delay|+|arr\_delay|), sort the table by decreasing delay time, pick first 100 rows which has more significant delay time and analyze them.

The correlation score matrix of the first 100 rows is shown as below.

Table

Description automatically generated

As the correlation matrix shows, temperature and dewpoint temperature make it more likely to see a delay. As the temperature and dewp goes down, delay times go up. From my perspective, temperature usually can lead other weather conditions. With the low temperature it is, the weather conditions become worse for flights.

1.e Produce a map that sizes each destination airport by the number of incoming flights. You may use a continuous scale for the size. Here is a code snippet to draw a map of all flight destinations, which you can use as a starting point. You may need to install the maps packages if you have not already. Adjust the title, axis labels and aesthetics to make this visualization as clear as possible.

Graphical user interface, text, application

Description automatically generated

Map

Description automatically generated

Problem 2

I failed to get the geocode for each place through geocode function from google, so I manually get latitude and longitude for each place, I also post the sources where I got geocode, there might be several little mistakes. (I found there is one geocode is weird)

Text

Description automatically generated

Draw the first map which shows the total number of Janssen doses administered.

Graphical user interface, text

Description automatically generated

Map

Description automatically generated

Draw the second map which shows the total number of Moderna doses administered.

Text

Description automatically generated

Map

Description automatically generated

Draw the third map which shows the total number of Pfizer doses administered.

Text

Description automatically generated

Map

Description automatically generated

Problem 3

I chose an argumentative essay I wrote when I was learning English last year.

I also generate the word frequency data frame for the figure.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Jinyang Ruan’s argumentative essay on AI future, written in December 2020.