Assignment3

Jinyang Ruan

011696096

1.0 Load the data into R and print the first few values of the columns with a header containing the string “time”.

Graphical user interface

Description automatically generated

In this case, I printed first 10 rows with a header containing the string “time”.

1.a (10 pts) Count the number of flights that departed NYC in the first week (first 7 days) of January and February combined.

Graphical user interface, text, application, email

Description automatically generated

1.b (10 pts) Print the year, month, day, carrier and air\_time of the flights with the 6 longest air times, in descending order of air\_time.

Graphical user interface, text, application

Description automatically generated

1.c (10 pts) Add a new column to the dataframe; speed (in miles per hour) is the ratio of distance to air\_time. Note that the unit of speed should be miles per hour. If you think they might be useful, feel free to extract more features than these, and describe what they are.

Table

Description automatically generated

A picture containing text

Description automatically generated

1.d (14 pts) Display the average, min and max air\_time times for each month.

I excluded NAs first then computed average, min and max air\_times for each month.

Graphical user interface, text

Description automatically generated

1.e (16 pts) Impute the missing air\_times as the distance divided by the average speed of flights for that destination (dest). Make a second copy of your dataframe, but this time impute missing air\_time with the average air\_time for that destination. What assumptions do these data filling methods make? Which is the best way to impute the data, or do you see a better way, and why? You may impute or remove other variables as you find appropriate. Briefly explain your decisions.

Firstly, I imputed the missing air\_times as the distance divided by the average speed of flights for that destination.

Graphical user interface, text, application, email

Description automatically generated

Then I imputed missing air\_time with the average air\_time for that destination.

Text

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedWhen we impute the missing air\_times as the distance divided by the average speed of flights for that destination, we assume that for the same destination, the average speeds of the flights are similar; while choosing the average air\_time, the air\_time of the samples are similar. Here imputation by distance divided by average speed is a better way for the air\_time which usually associates with the distance.

However, usually this "Mean Imputation" is not a good choice in practice. there are many other better solutions, such as "Hot deck imputation", "Regression imputation", and "Stochastic regression imputation". In this case, to impute missing air\_time. The first way which is the distance divided by the average speed of flights for that destination would be good.

2.0 Lode the dataset into R and tidy the dataset

A picture containing scatter chart

Description automatically generated

Table

Description automatically generated

2.a

It replaces the header with name "newrel" to "new\_rel", which can make the format of all headers consistent.

If it is skipped, we cannot extract detailed features by simply separate the header name.

For example, given a header name of "new\_sp\_f5564" we can apply separate function on it and

extract three new features:new, sp, f5564, while in case of "newrel\_m014" we cannot.

2.b (5 pts) How many entries are removed from the dataset when you set values\_drop\_na to true in the pivot\_longer command (in this dataset)?

Text

Description automatically generated

There is the data frame for who2.Table

Description automatically generated

The removed entries can be computed as following:

Graphical user interface, text, application

Description automatically generated

2.c(5 pts) Explain the difference between an explicit and implicit missing value, in general. Can you find any implicit missing values in this dataset, if so, where?

According to the "R for Data Science, An explicit missing value is the presence of an absence; an implicit missing value is the absence of a presence", That means for explicit missing, there will be a specific representation to indicate the missing of the value (e.g., NA). While for implicit missing, there will be no specific representation for the value.

In this dataset, we can consider the “case == 0” as the implicit missing value, we can get the sub-samples with following command.

Table

Description automatically generated

2.d (5 pts) Looking at the features (country, year, var, sex, age, cases) in the tidied data, are they all appropriately typed? Are there any features you think would be better suited as a different type? Why or why not?

Table

Description automatically generated

As shown above, all features are typed appropriately except the age which is typed as character. Usually, it is better to type this feature as integer, when we try to analyze this feature (e.g., distribution of the age, average age), we must compute the results with integer.

2.e (10 pts) Generate an informative visualization, which shows something about the data. Give a brief description of what it shows, and why you thought it would be interesting to investigate.

I. Look at the data grouped by country.

Graphical user interface, text, application

Description automatically generatedChart, bar chart

Description automatically generated

In this case, top TB cases distribution by countries are shown as following graph, from which we can know which countries are worse off.

II. look at the data grouped by sex

Text

Description automatically generated

Chart, bar chart

Description automatically generated

the above graph shows TB case distribution among female and male, it indicates that male are more easily affected.

III. Look at the data grouped by age.

Text

Description automatically generated

Chart, bar chart

Description automatically generated

the above graph shows TB case distribution among different ages, it indicates that people with young and the middle-aged are more possible to be affected.

2.f

Firstly, I built the data frame and input all data.

A picture containing text

Description automatically generated

Then I tidied the data as required.

Text

Description automatically generated with medium confidence

Table

Description automatically generated