LAB 3 and Assignment 3

Jillian Morrison

9/26/2019

General rules

- For some questions, the needed methods may not have been covered in class. For them, please do some research to solve them.
- You must show your work in order to get points. Providing correct answers without supporting codes or intermediate steps does not receive full credit.
- You must submit both the R file as a .R file and the Assignment file as a PDF. For the Assignment file include the code, the output and explanations (if necesssary).
- LAB 3 - Questions 1-7
- Assignment 3 - Questions 8-15
- Needed files: `Alldata.csv`, `Population.csv`, `Documentation_WHO.doc`

Questions

Loading dataset into R. Rememer to save all files in a folder which you will set as your working directory for this Lab and Assignment.

```
> library(readr)
> Table4 <- read_csv("Alldata.csv")
```

The dataset used in this assignment comes from WHO (World Health Organization) and it lists the number of deaths for different causes by year, country, gender, age ranges, etc. The source is: https://www.who.int/healthinfo/statistics/mortality_rawdata/en/

You can look at the website above for documentaton and more information about the dataset. I merged three datasets from this site to create the dataset used in this assignment.

1. Remove rows where `Cause.Name` is `"All causes"`
2. Group the dataset from (1) by `Year`, `Cause.Name` (Cause of death) and `Name` (Country name) and Sum `Deaths1` (number of deaths for all ages). Do this because the same cause of death can be listed under different codes, for different genders and so on. By doing this, you sum over all the other variables that we will not consider in this assignment to get the total number of deaths for each combination of Year, cause of death and country.

Be sure to name this dataset. You will be using this dataset for most of the rest of the assignment.

3. a. How many different causes of death are listed in the dataset?

```
b. List  the top 3 causes of death in the dataset.
```

4. Find the total number of deaths from each country (Country is the `Name` variable in dataset) for people who died from the top 3 causes of death. Note that not every country has recorded deaths for each cause of death.

5. a. Construct a barplot of total number of deaths by country for the top 3 causes of death. You should have 3 barplots - one for each cause of death. Remember to add titles (plot and axis) and change legend name and labels (if necessary).

```
b. What is the problem with these plots? How would you fix it? Write at least 3 sentences.
```

6 a. For each of the top 3 causes of death, create a barplot of the total number of deaths for the top 5 countries (5 countries with the most deaths for each cause). Remember to add titles (plot and axis) and change legend name and labels (if necessary)

b. Say something about the plots by comparing them to each other. Write at least 3 sentences.

7. a. Construct a histogram of the total number of Deaths for the top 3 causes of death (i.e. sum over Year - so each country should have one number for the total number of deaths for each cause). You will construct 3 histograms on the same graph - one for each of the top 3 causes of death. Remember to add titles (plot and axis) and change legend name and labels (if necessary)

```
b.   Compare the histograms. Write a couple sentences.
```

8. List the overall top 3 causes of death for every year.
9. List the top cause of death for `2003` to `2013`
10. List the top cause of death for each year in the dataset
11. Create a list like you did in question 10 for the country you are from and for a country in another continent which is different from your home country. If your country is not listed in the dataset, use any country and compare your result to a country in another continent.
12. Choose 6 countries (for which you have data for at least 10 years). Look for a cause of death that is common to all these 6 countries. a. Create a boxplot for each country (with all 6 graphed on the same plot), of the number of deaths over the years from that chosen disease (so you should have at least 10 datapoints (number of deaths) for each country). Remember to add titles (plot and axis) and change legend name and labels (if necessary)

    b. Say something about your boxplots in part a. Write at least 3 sentences.

13. Create a graph or table of your own based on the dataset. Justify the reasoning behind why you created what you created (i.e. you will want to create a hypothesis about the variables in the dataset and create a table/graph to help you look at your data in terms of that hypothesis). Describe what you see (i.e. does the data show the pattern of your hypothesis?). Remember to add titles (plot and axis) and change legend name and labels (if necessary)
14. You are interested in the Caribbean Region and two of the top 10 causes of deaths in the region. Use the code below to generate a table that lists the top 10 causes of death for each country of interest (given in code).

Notes on code:

- `Carib` creates a vector with the lists of countries you are interested in

- filter(Name %in% Carib) filters dataset to choose only observations whose name is in the vector Carib
- group_by(Name,Cause.Name) since you are interested in looking at the number of deaths by name of country (Name) and cause of death (Cause.Name).
- summarize:
  - n=n() counts the number of times each cause of death per country appears in the dataset (this essentially counts the number of years each cause of death per country appears)
  - Deaths=sum(Deaths1) sums the total deaths by Name and Cause.Name
- arrange(desc(n)) sorts the dataset by n in descending order.
- slice(1:10) takes the first (1st) to tenth (10th) observations for each country (Name) (essentially takes the top 10 observations since the dataset is sorted)

```
> Carib=c("Cayman Islands", "Belize", "Jamaica", "Antigua and Barbuda",
+         "Barbados", "Trinidad and Tobago","Dominican Republic", "Grenada")
> Caribbean=Tabl%>%filter(Name %in% Carib)%>%group_by(Name, Cause.Name)%>%
+   summarize(num=n(), Deaths=sum(Deaths1))%>%arrange(desc(Deaths))%>%slice(1:10)
> head(Caribbean)
# A tibble: 6 x 4
# Groups:   Name [1]
  Name          Cause.Name                              num Deaths
  <fct>         <fct>                                 <int>  <dbl>
1 Antigua and~ Cerebrovascular disease                   12    688
2 Antigua and~ Signs, symptoms and ill-defined co~    12    485
3 Antigua and~ Diseases of pulmonary circulation ~    12    482
4 Antigua and~ Endocrine and metabolic diseases, ~    12    398
5 Antigua and~ Ischaemic heart disease                   12    340
6 Antigua and~ Hypertensive disease                      12    331
```

Another note on results: looks like most of the causes of death for these countries were reported for all years where data was submitted to the WHO.

a. Job for you: Create a scatter plot of total yearly deaths for x= Cerebrovascular disease versus y=Ischaemic heart disease and color points by name of country. (i.e. find the total deaths by Cause.Name(i.e x and y in this case), year and Name (name of country) in order to plot) Remember to add titles (plot and axis) and change legend name and labels (if necessary).

b. Make observations based on the plot in part a. Write at least 3 sentences.

15. I have included a dataset called Population.csv which has the population in 2018 for each of the countries in 14. Merge this dataset to your existing dataset to create a scatterplot of yearly Deaths as a percent of population in 2018 (i.e. calculate a new variable as Deaths1/Population2018 * 100 and redo plot in 14 with this percent instead of number of deaths) for x= Cerebrovascular disease versus y=Ischaemic heart disease and color points by name of country. Remember to add titles (plot and axis) and change legend name and labels (if necessary)