STAT346: Statistical Data Science I

HW#2 – Due: Saturday, October 10, 2020 by 6 p.m.

September 26, 2020

Instruction: Answer to the following questions and write your report using R Markdown. You may use the R Markdown template, accompanied with this homework assignment. You should submit two files, through KU Black Board system (https://kulms.korea.ac.kr), which should have the following naming format:

- `stat346_hw2_your_id.rmd`
- `stat346_hw2_your_id.pdf` or `stat346_hw2_your_id.docx`

---

1. Download the `gapminder` data by clicking *here* or by calling `library(gapminder)`:

```
library(gapminder)
gapminder
```

   Use `dplyr` functions to address the following questions:

   a. How many unique countries are represented per `continent`? (Hint: `group_by` then `summarize` with a call to `n_distinct(...)`).

   b. Which European nation had the lowest `GDP per capita` in 1997 and 2007? (Hint: `filter`, `arrange`, `head(n=1)`).

   c. According to the data available, what was the average life expectancy across each continent in the 1980's? (Hint: `filter`, `group_by`, `summarize`).

   d. What 5 countries have the highest total GDP over all years combined? (Hint: GDP per capita is simply GDP divided by the total population size. To get GDP back, you'd mutate to calculate GDP as the product of GDP per capita times the population size. `mutate`, `group_by`, `summarize`, `arrange`, `head(n=5)`).

   e. What countries and years had life expectancies of at least 80 years? Provide the columns of interest only: `country`, `life expectancy` and `year` (in that order).

   f. What 10 countries have the strongest correlation (in either direction) between `life expectancy` and `per capita GDP`?

   g. Which combinations of continent (besides Asia) and year have the highest average population across all countries? Your output should include all results sorted by highest average population. (Hint: `filter` where `continent != Asia`, `group_by` two variables, `summarize`, then `arrange`).

h. Which three countries have had the most consistent population estimates (i.e. lowest standard deviation) across the years of available data?

---

2. Use the `nycflights13` package and the `flights`, `planes`, and `weather` data to answer the following questions:

```
library(nycflights13)
flights
planes
weather
```

   a. What month had the highest proportion of cancelled flights? What month had the lowest? Interpret any seasonal patterns.

   b. What plane (specified by the `tailnum` variable) traveled the most times from NYC airports in 2013? Plot the number of trips per week over the year.

   c. What is the oldest plane (specified by the `tailnum` variable) that flew from NYC airports in 2013? How many airplanes that flew from NYC are included in `planes` table?

   d. How many planes have a missing data of manufacture? What are the five most common manufacturers? Has the distribution of manufacturer changed over time as reflected by the airplanes flying from NYC in 2013? (Hint: you may need to recode the manufacturer name and collapse rare venders into a category called `Other`).

   e. What is the distribution of temperature in July 2013? Identify any important outliers in terms of the `wind_speed` variable. What is the relationship between `dewp` and `humid`? What is the relationship between `precip` and `visib`?

   f. On how many days was there precipitation in the NY area in 2013? Where there differences in the mean visibility (`visib`) based on the day of the week and/or month of the year?

---

3. For this problem, we'll use the `diamonds` dataset from the `ggplot2` package.

```
library(ggplot2)
diamonds
```

   a. Use the `hist` function to create a histogram of carat with bars colored `steelblue`.

   b. Use the `qplot` function from the `ggplot2` package to create a histogram of `depth`.

   c. Use the `qplot` function from the `ggplot2` library to create violin plots showing how `price` varies across diamond `cut`. Specify `fill = cut` to get all the boxplots to be coloured differently. Hint: For this exercise, it will be useful to know that `violin` is a geometry (`geom`) built into `ggplot2`, and that `qplot` can be called with the arguments:

```
qplot(x, y, data, geom, fill)
```

---

4. For this exercise, we'll use the `Cars93` data set in the `MASS` library.

```
library(MASS)
library(tidyverse)
as_tibble(Cars93)
```

a. Define a `ggplot` object using the `Cars93` dataset that you can use to view `Price` on the y-axis, `MPG.highway` on the x-axis, and set the size mapping to be based on `Horsepower`. Use `geom_point()` to create a scatterplot from your `ggplot` object.

b. Repeat part (a), this time also setting the `color` mapping to be based on `Origin`.

c. Repeat part (b), this time using `stat_smooth()` to add a layer showing the smoothed curve representing how `Price` varies with `MPG.highway`.

d. Use your `ggplot` object from part (b) along with the `geom_point()` and `facet_grid()` layers to create scatterplots of `Price` against `MPG.highway`, broken down by (conditioned on) `Origin`.

e. Modify your solution to part (d) to also display regression lines for each scatterplot.