DSCI 6607 – Fall 2024

Assignment 3[*]

Question 1

Using the `mtcars` dataset and create a scatter plot of miles per gallon (mpg) vs horsepower (hp). Facet the plot by the number of cylinders (cyl) in the dataset. [**20 points**]

a. What does each panel in the faceted plot represent?

b. How can you adjust the appearance of points in the scatter plot (e.g., color or shape)?

c. Which variable is used for faceting in this plot?

d. What insights can you gain from comparing mpg and hp across different values of cyl?

```r
head(mtcars)
```

Question 2

Generate the correct format string to parse each of the following dates and times: [**20 points**]

```r
a1 <- "12/30/14" # Dec 30, 2014
a2 <- "07-Jan-2017"
a3 <- c("August 19 (2015) - 3:04PM", "July 1 (2015) - 4:04PM")
a4 <- "January 1, 2010"
a5 <- "2015-Mar-07"
```

Question 3

Consider the `cities` data set. [**20 points**]

a. create a new feature named `city_density` by dividing the city population `city_pop` by the city area `city_area`.

b. Use the `select` function to select the city name (name), population, area and density.

c. The numbers in (b) are very small. Modify the units in `city_density` by multiplying the city density by 1000.

d. Now report the average city density by continent. *Hint*: You should notice that the results include some missing values:

e. Create a plot with city density on the x-axis and metro density on the y-axis. Use a log scale for the axes and include points and text repel labels with the city names.

---

[*] This content is protected and may not be shared, uploaded, or distributed without written permission from Dr. Armin Hatefi.

Question 4

You will try to recreate a **plot** from an Economist article showing the relationship between well-being and financial inclusion.

You can find the accompanying article at this **link**

Load the data set `EconomistData.csv`.

```
head(EconomistData)
```

a. Create a scatter plot similar to the one in the article, where the `x` axis corresponds to percent of people over the age of 15 with a bank account (the `Percent.of.15plus.with.bank.account` column) and the `y` axis corresponds to the current SEDA score `SEDA.Current.level`.
b. Color all points `blue`.
c. Color points according to the `Region` variable.
d. Overlay a fitted smoothing trend on top of the scatter plot. Try to change the span argument in `geom_smooth` to a low value and see what happens.
e. Overlay a regression line on top of the scatter plot Hint: use `geom_smooth` with an appropriate method argument.
f. Facet the previous plot by `Region`. [**20 points**]

Question 5

Consider again Questions 1 and `mtcars` dataset. Sometimes continuous variables can be used for faceting by converting them into factors. [**20 points**]

a. Convert the hp (horsepower) variable in mtcars into a factor with three levels: "Low," "Medium," and "High".
b. Create a scatter plot of mpg vs weight (wt), faceted by this new hp factor.
c. How does converting hp into categorical groups enhance the interpretability of the plot?
d. Describe the differences observed in mpg for different hp levels.
e. What function is used to create categorical levels from continuous variables?
f. Can faceting by grouped levels provide more insight than using hp as a continuous variable on the x-axis?

Question 6

Load the dataset `movies.csv` used in the lecture:

https://raw.githubusercontent.com/Juanets/movie-stats/master/movies.csv

a. Find a subset of the movies produced after 2005. Save the subset in `movies.sub` variable.
b. Keep columns `name`, `director`, `year`, `country`, `genre`, `budget`, `gross`, `score` in the `movies.sub`.
c. Find the profit for each movie in `movies.sub` as a fraction of its budget. Convert `budget` and `gross` columns million dollar units rounded to the first decimal point. Use `round()` to round numbers
d. Count the number of movies in `movies.sub` produced by each genre, and order them in the descending count order.

e. Now group movies in `movies.sub` by countries and genre. Then, count the number of movies in each group and the corresponding median fractional profit, the mean and variance of the movie score for each group. [**20 points**]

Question 7

Load in the dataset `movies.csv` used in the lecture:

https://raw.githubusercontent.com/Juanets/movie-stats/master/movies.csv

Using pipes, for each genre find the two directors the top mean movie scores received for the movies produced after 2001, after filtering out the directors with fewer than 4 movies in total.

Hint: Use `top_n()` function to select top n from each group. [**20 points**]

Question 8

The continuous random variable X has the following probability density function (pdf), for some positive constant c,

$$f(x) = \frac{3}{(1+x)^3}, \quad 0 \leq x \leq c.$$

a. Find c which makes f a legitimate pdf?

b. Use R and plot the pdf curve of the random variable.

c. What is $E(X)$?

d. Use R and simulate 1000 observations from this statistical population?

e. Use the generated data from part (d), estimate the mean and variance of the distribution? [**20 points**]

Due on Friday, November 15, by 5 pm

Have fun!