Assignment 1-Office Hours

test

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

install.packages("datasets")

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

dat ### Problem 2

Use this command to make the state names into a new variable called State. It is important to rename data sets in order to practice renaming files, as well as to stay organized and keep track of the datasets.

dat$state <- tolower(rownames(USArrests))

head(dat)

names(dat) summary (dat)

Variables in the data set USArrests: state, murder, assault, urbanpop, rape

Problem 3 What type of variable (from the DVB chapter) is `Murder`?

Answer: categorical

What R Type of variable is it?

Answer: character

Problem 4

What information is contained in this dataset, in general? What do the numbers mean? This dataset contains information on the type and frequency of crime (murder, assautl, urbanpop, and rape) committed in different states. The numbers are the frequency at which each crime is committed in each state.

hist(dat$murder, main="Histogram of Murder", xlab=State, ylab="Frequency of Murder")

Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.? The mean is 7.788. The median is 7.3. Mean is the average of all the values, while median is the middle point of a number set. A quartile is a type of quantile which divides the number of data points into four parts (quarters). I think R gives us the 1st and 3rd quartile because the 1st and 3rd quartile tell us how spread out the middle 50% of the data set i.

Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

hist(dat$Assault, main = "Histogram of Assault", xlab = State, ylab = "Frequency of Assault")$hist(dat$Rape, main="Histogram of Rape", xlab=State, ylab="Frequency of Rape")

The mean is 170.76 for assault and 21.232 for rape. The median is 159 for assault and 20.1 for rape. Mean is the average of all the values, while median is the middle point of a number set. A quartile is a type of quantile which divides the number of data points into four parts (quarters). I think R gives us the 1st and 3rd quartile because the 1st and 3rd quartile tell us how spread out the middle 50% of the data set i.

What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: The command par combines several plots onto one graph.

What can you learn from plotting the histograms together?

Answer: You can compare the frequencies as well as the spread of data.

Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
library('maps')
library('ggplot2')

ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```

What does this code do? Explain what each line is doing.

Answer: The first two lines lets you plot the murder data on a map. The last line allows you to expand the parameters of the map to fit the parameters of the dataset.