HW 2 Due Monday Sept 25, 2017. Upload R file to Moodle with name: HW2_490ID_41.R
You can upload plots to Moodle in separate files and refer to them in your R file, or use an
R notebook http://rmarkdown.rstudio.com/r_notebooks.html
Do not remove any of the comments. These are marked by

Class ID: 41

In this assignment you will practice how to manipulate a dataframe,
such as taking subsets and creating new variables, with the goal of creating a plot.

You will work with the mtcars data in R library and a dataset called SFHousing.

Before beginning with the housing data however, you will do some warm up
exercises with the small mtcars data set.

PART 1. mtcars Data
Q1.(2 pts.)
Use R to generate descriptions of the mtcars data which is already included in R base.
The description could be a summary of each column and the dimensions of the dataset (hint:
you may find the summary() command useful). Write up your descriptive findings and observations
of the R output.

Your code below
```
head(mtcars)
summary(mtcars)
dim(mtcars)
```

I used the head function to display a subset of the data contained in mtcars. We see that it displays the variables mpg, cyl, disp, hp, drat, qsec, vs, am, gear, carb. The dataset has dimensions 32x11. By using the function head() we see that "vs" and "am" are just 0's and 1's. By using the function summary we see that the maximum value of "cyl" and "carb" is 8 while for "gear" is 5 and by using head() we see that they are integers. The mean "mpg" and "hp" are 20.09 and146.7 respectively.

Q2.(2 pts)
Create a vector mpg_cl based on mpg in the dataset.
For automatic cars, the vector should have value TRUE when mpg > 16 and value FALSE when mpg <= 16.
For manual cars, the vector should have value TRUE when mpg > 20 and value FALSE when mpg <= 20.

Your code below
```
mpg_cl = (mtcars$am==1 & mtcars$mpg > 20) | (mtcars$am==0 & mtcars$mpg > 16)
```

Q3.(2 pts)

Here is an alternative way to create the same vector in Q2.

First, we create a numeric vector mpg_index that is 16 for each automatic cars
and 20 for each manual cars. To do this, first create a vector of length 2 called
id_val whose first element is 16 and second element is 20.

Your code below
```
id_val = c(16,20)
```

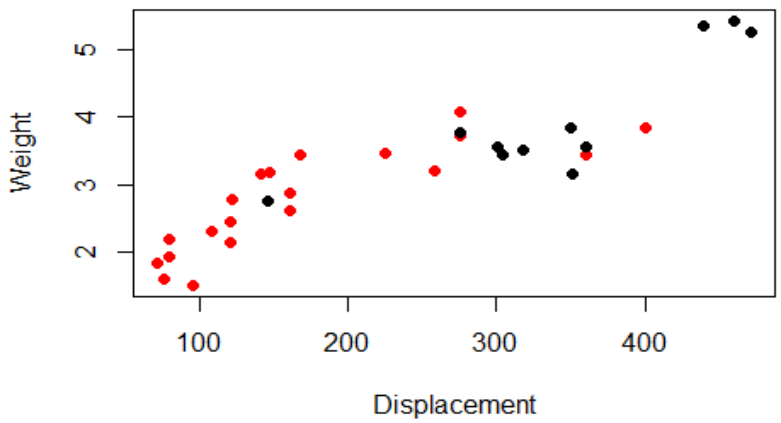
Create the mpg_index vector by subsetting id_val by position, where the
positions could be represented based on am column in mtcars.

Your code below
```
mpg_index = id_val[mtcars$am+1]
```

Finally, us mpg_index and mpg column to create the desired vector, and
call it mpg_cl2.

Your code below
```
mpg_cl2 = mtcars$mpg>mpg_index
```

Q4.(2 pts)
Make a plot of the variable disp against the variable weight. Color cars with different
mpg_cl value differently, and also format your plots with appropriate labels. Describe any
notable observations you have of the plot.
Your code below
```
plot(mtcars$disp,mtcars$wt, col = as.factor(mpg_cl),
xlab="Displacement",ylab="Weight",pch=19)
```



The variable "am" describes the tranmission (0=automatic, 1=manual). In the plot, the red points, represent automatic cars and the black points represent the manual cars. We see that

automatic cars mostly have a lower displacement and weight, compared to manual cars. In general, it seems that displacement and weight behave almost linear relation ship regardless of the type of transmission.

#PART 2. San Francisco Housing Data
#
# Load the data into R.
load(url("https://www.stanford.edu/~vcs/StatData/SFHousing.rda"))

Q5. (2 pts.)
What objects are in SFHousing.rda? Give the name and class of each.
Your code below
ls()
class(cities)
class(housing)

The function ls() shows use what objects where loaded to the global environment and by using class(), we can see both of them are data frames.

Give a summary of each object, including a summary of each variable and the dimension of the object.

Your code below
summary(cities)
dim(cities)
summary(housing)
dim(housing)

Your answer here

The object cities has data of the location (longitude and latitude), county, median price, size bedrooms, and number of houses. Also, it has dimensions 163x7. The object housing has also informaton about the county,,city and logitude and latitude, but also includes zip code, street, number of bedrooms, bedroom square feet, year and date. This data frame has dimensions 281,506x15.

After exploring the data (maybe using the summary() function), describe in words the connection
between the two objects.

While they have some overlapping information (city,county,price, bedroom, lon,lat), housing contains the actual prices and bedrooms in cities is the median of those variables so they differ among data frames.

Describe in words two problems that you see with the data.

After observing the data we see that for some of the columns have missing values (different amount of them too). The variable names can be hard to interpret what they mean and there seems to be a large spread among some of the values. This could skew the observations we can make about the data.

Q6. (2 pts.)
We will work the houses in Oakland, San Francisco, Campbell, and Sunnyvale only.
Subset the housing data frame so that we have only houses in these cities
and keep only the variables county, city, zip, price, br, bsqft, and year.
Call this new data frame SelectArea. This data frame should have 28843 observations
and 7 variables. (Note you may need to reformat any factor variables so that they
do not contain incorrect levels)

Your code below
```
SelectArea = housing[c(housing$city=="Oakland"|housing$city=="San
Francisco"|housing$city=="Campbell"|housing$city=="Sunnyvale"),
c('county','city','zip','price','br','bsqft','year')]
```

Q7. (3 pts.)
We are interested in making plots of price and size of house, but before we do this
we will further subset the housing dataframe to remove the unusually large values.
Use the quantile function to determine the 95th percentile of price and bsqft
and eliminate all of those houses that are above either of these 95th percentiles
Call this new data frame SelectArea (replacing the old one) as well. It should
have 26418 observations.

Your code below
```
rem.NA = SelectArea[!is.na(SelectArea$bsqft),]
SelectArea=
rem.NA[c(rem.NA$price<quantile(rem.NA$price,0.95)&(rem.NA$bsqft<quantile(rem.NA$bsqf
t,0.95))),]
```

Q8 (2 pts.)
Create a new vector that is called price_per_sqft by dividing the sale price by the square footage
Add this new variable to the data frame.

Your code below
```
price_per_sqft = SelectArea$price/SelectArea$bsqft
SelectArea = cbind(SelectArea,price_per_sqft)
```

Q9 (2 pts.)
Create a vector called br_new that is the number of bedrooms in the house, except
if this number is greater than 5, it is set to 5. That is, if a house has 5 or more
bedrooms then br5 will be 5. Otherwise it will be the number of bedrooms.

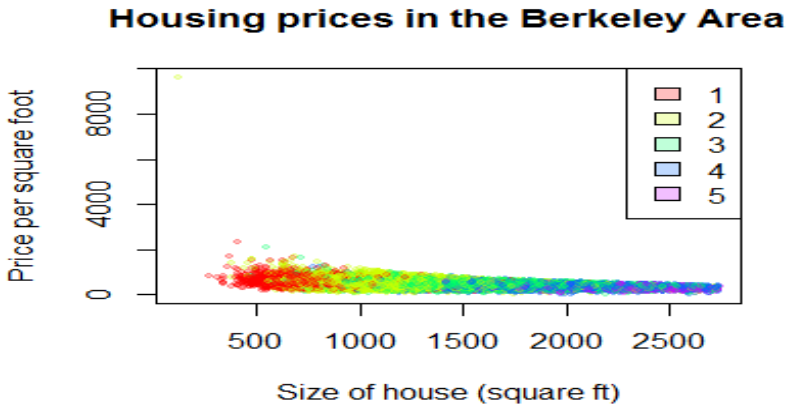
Your code below
```
br_new = SelectArea$br # Make a copy of br vector
br_new[br_new >= 5] = 5  # Change those no. br greater than 5 to 5.
```

```
# Q10. (4 pts. 2 + 2 - see below)
# Use the rainbow function to create a vector of 5 colors, call this vector rCols.
# When you call this function, set the alpha argument to 0.25.
# Create a vector called brCols where each element's value corresponds to the color in rCols
# indexed by the number of bedrooms in the br_new.
# For example, if the element in br_new is 3 then the color will be the third color in rCols.
# (2 pts.)
```

Your code below
```
rCols = rainbow(5,alpha=0.25)
brCols = rCols[br_new]
```

```
# We are now ready to make a plot!
# Try out the following code (check R documentation to make sure you understand it),
plot(price_per_sqft ~ bsqft, data = SelectArea,
    main = "Housing prices in the Berkeley Area",
    xlab = "Size of house (square ft)",
    ylab = "Price per square foot",
    col = brCols, pch = 19, cex = 0.5)
legend(legend = 1:5, fill = rCols, "topright")
```

What interesting feature do you see that you didn't know before making this plot?
(2 pts.)



On the plot, we see that the vertical spread of the observation gets smaller as the size of of the house increases. This could indicated that while houses with more bedrooms have a higher price, when you distribute it according to the number of bedrooms it is not that high of a difference.