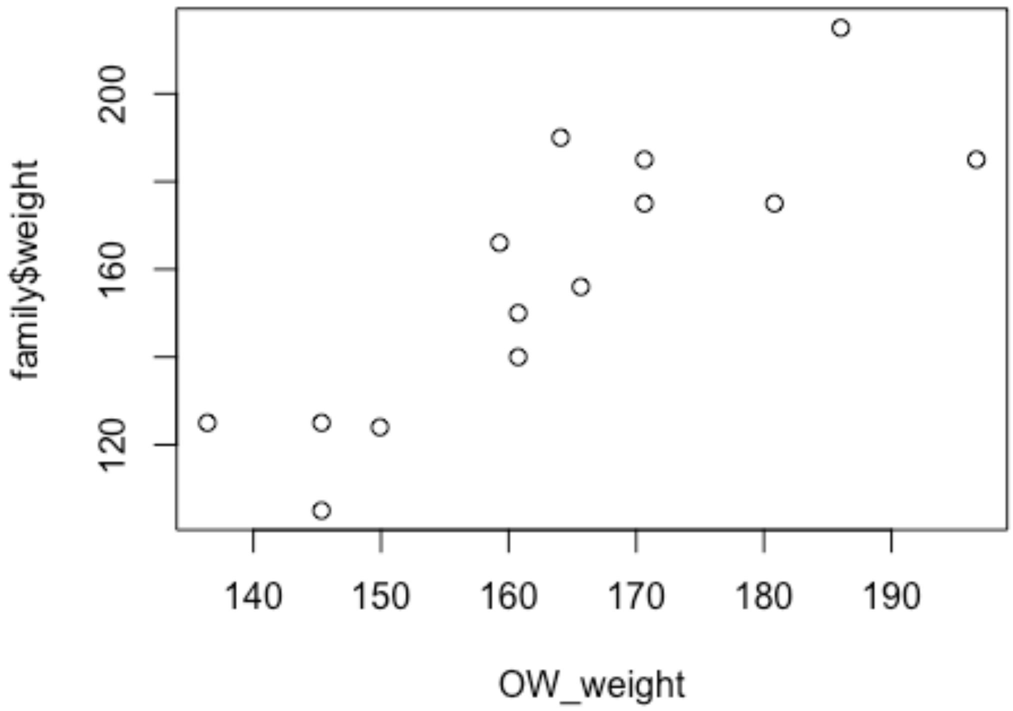
```
# HW 2 Due Tuesday Sept 13, 2016. Upload R file to Moodle with name:
HW2_490IDS_YOUR_NetID.R
# Do Not remove any of the comments. These are marked by #

###Name: Zixin Ouyang

# In this assignment you will manipulate a data frame, by taking subsets and creating new
variables,
# with the goal of creating a plot.

# You will work with a dataset called Baseball in the R library. The Baseball dataset describes
# baseball players' stats from the '86 and '87 season, as well as career stats.

# Before beginning with the housing data however, you will do some warm up
# exercises with the small family data set that we have used in class.

#PART 1.  Family Data
# Load the data from the Web into your R session with the following command:
load(url("http://courseweb.lis.illinois.edu/~jguo24/family.rda"))

# In the following exercises try to write your code to be as general as possible
# so that it could still work if the family had 27 members in it or if the
# variables were in a different order in the data frame.

# Q1. (2 pts.)
# The NHANES survey (the source of the family data) used different cut-off values for
# men and women when classifying them as over weight. Suppose that a man is classified
# as obese if his bmi exceeds 26 and a woman is classified as obese if her bmi exceeds 25.

# Write a logical expression to create a logical vector, called OW_NHANES, that is TRUE if
# a member of family is obese and FALSE otherwise
OW_NHANES=(family$gender=="m"&family$bmi>26)|(family$gender=="f"&family$bmi>25)

# Q2. (4 pts.)
# Here is an alternative way to create the same vector that introduces
# some useful functions and ideas

# We will begin by creating a numeric vector called OW_limit that is 26 for each male in
# the family and 25 for each female in the family.

# To do this, first create a vector of length 2 called OWval whose first element
# is 26 and second element is 25.
OWval=c(26,25)

# Create the OW_limit vector by subsetting OWval by position, where the
# positions are the numeric values in the gender variable
# (i.e. use as.numeric to coerce the factor vector to a numeric vector)
OW_limit=OWval[as.numeric(family$gender)]

# Finally, us OW_limit and bmi to create the desired logical vector, and
# call it OW_NHANES2.
OW_NHANES2=family$bmi>OW_limit

# Q3. (2 pts.)
# Use the vector OW_limit and each person's height to find the weight
# that they would have if their bmi was right at the limit (26 for men and
# 25 for women). Call this weight OW_weight

# To do this, start with the formula:
# bmi = (weight/2.2) / (2.54/100 * height)^2
# and find re-express it in terms of weight.
OW_weight=OW_limit*2.2*(2.54/100*family$height)^2

# Make a plot of the weight at wihich they would
# be over weight aginst actual weight
plot(OW_weight, family$weight)
```



```
#PART 2.  Baseball data
#Load the data into R.
#In order to access this data set we will install the relevant package and use the following code to do so:
install.packages("vcd")
library(vcd)
attach(Baseball)

#This means that the dataset Baseball was in the vcd package.

# Q4.  (4 pts.)
# How many variables are in the dataset Baseball?
### Your code below
objects(Baseball)

### Your answer here
"assist86" "atbat"    "atbat86"  "div86"    "error86"  "hits"     "hits86"   "homer86"
"homeruns" "league86" "league87" "name1"
"name2"    "outs86"   "posit86"  "rbi"      "rbi86"    "runs"     "runs86"   "sal87"
"team86"   "team87"   "walks"    "walks86"  "years"
There are 25 variables in the dataset Baseball.

# How many observations are in Baseball?
### Your code below
length(name1)
### Your answer here
322

# For a more DETAILED description of ALL of the variables is this data set, visit:
# https://vincentarelbundock.github.io/Rdatasets/doc/vcd/Baseball.html

# Run the summary function and anwser the following questions:
# For the variable team87, which state had the most baseball players in the dataset?
### Your code below
summary(team87)
### Your answer here
Atl Bal Bos Cal Chi Cin Cle Det Hou  KC  LA Mil Min Mon  NY Oak Phi Pit  SD  SF Sea StL Tex Tor
 12  15  11  12  22  11  13  13  12  14  13  14  14  13  24  13  13  16  10  15  11   9  12  10
The state has the most baseball players is NY.

# Make an observation about the variable, sal87, which is the yearly salary of the selected
# baseball players in the dataset.
# Who is the highest paid player in the data set?
### Your code below
max(Baseball$sal87, na.rm=TRUE)
2460
Baseball[Baseball$sal87==2460, c("name1","name2")]
```

```
### Your answer here
```

Eddie Murray

```
# Q5. (2 pts.)
# Now, we only want to use the baseball players in the National League.
# This information is found through the variable, league86. The letter N indicates that the
# player is in the National League. The letter A indicates that the player is in the American
League.
# Subset the new data frame so that all of the baseball players are in the National League,
# and only keep the following variables: name1, name2, years, hits86, homer86, homeruns,rbi,
and sal87.
# To clarify, the variable, homer86 are the homeruns in that the player hit in '86, and the
# variable homeruns are career homeruns for each player.
# Call the new data Baseball1 (your code below)
Baseball1=Baseball[Baseball$league86=="N", c("name1", "name2", "years", "hits86",
"homer86", "homeruns","rbi", "sal87")]
Baseball1
```

```
# Q6. (2 pts.)
# We want to remove unusually large values in order to further subset the data.
# Use the quantile function to determine the 99% of variable sal87 (the salaries of the players in
'87).
# Then remove those baseball players that are above the 99th percentile.
# Call this new dataset Baseball1 as well.
quantile(Baseball1$sal87,0.99,na.rm = TRUE)
```

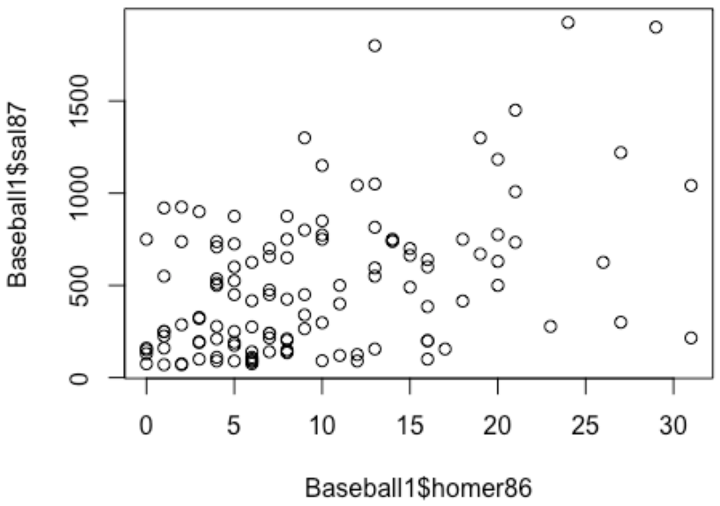
```
  99%
1936.681
```

The 99% of variable sal87 is 1936.681

```
Baseball1=Baseball1[Baseball1$sal87<=1936.681, ]
Baseball1=na.omit(Baseball1)
Baseball1
```

```
# Q7. (2 pts.)
# Create a new vector called hitsperhome.
# Divide hits86 by homer86, and this will create our new vector.
# Now add this new variable to the data frame.
hitsperhome=Baseball1$hits86/Baseball1$homer86
Baseball1=data.frame(name1=Baseball1$name1, name2=Baseball1$name2,
years=Baseball1$years, hits86=Baseball1$hits86, homer86=Baseball1$homer86,
homeruns=Baseball1$homeruns, rbi=Baseball1$rbi, sal87=Baseball1$sal87, hitsperhome)
```

```
# Q8. (2 pts.)
# Create a vector called hr15, this will be the number of homeruns hit in the year 1986
# (NOT total) so use the variable, homer86, if this number is greater than 15, it is set to 15.
# So if a player has 15 or more homeruns in that year, then hr15 will be 15, otherwise
# it will be the actual number of homeruns.
hr15=sapply(Baseball1$homer86, function(x) ifelse(x>15, 15, x))

# Q9. (2 pts.)
# Find out if there is a significant association between homeruns hit in 1986, variable homer86,
#and the salary of the players on opening day in 1987, variable sal87 (which is USD 1000).
# Answer this using several functions, including the plot function.
# Make 3 observations below.
plot(Baseball1$homer86, Baseball1$sal87)
```



```
cor(Baseball1$homer86, Baseball1$sal87)
[1] 0.4473575
```

The variable homer86 and the variable sal87 have a moderately strong positive linear relationship: when the value of homer86 increases, the value of sal87 also intend to increase.

Most players' homerun hits are below 20, and their salaries are below 1000.

There are some outliers that do not fit the pattern of the rest of the data.