```
# Homework
# Use the following code to read dataset "WestRoxbury.csv"
# The dataset is on Canvas.

d= read.csv("WestRoxbury.csv")
View(d)
getwd()

# Use ggplot2 package to answer the following questions.

library(ggplot2)

# Questions


# 1) YR.BUILT column has a typo in it. Imputate it with median YR.BUILT.
median_yr_built <- median(d$YR.BUILT, na.rm = TRUE)
d %>%
  mutate(YR.BUILT = IFELSE(YR.BUILT == 'typo.value', median_yr_built, YR.BUILT))


# 2) Create a histogram for the YR.BUILT variable. Use 40 bins.
#    Comment on the shape of the distribution.
d%>%
  ggplot()+
  geom_histogram(aes(x=YR.BUILT), bins=40)


# 3) We want to investigate the relationship between TOTAL.VALUE and YR.BUILT.
#    Create a scatter plot using the two variables. Color is "navy" and transparency
is 20%.
#    Discuss any noticeable patterns in the plot.
d%>%
  ggplot()+
  geom_point(aes(x=TOTAL.VALUE, y=YR.BUILT), color="navy", alpha= 0.2)


# 4) Create a scatter plot using total value, lot sqft and, remodel as color
parameter.
#    Alpha is 0.40, color is navy Discuss any noticeable patterns in the plot.
d%>%
  ggplot()+
  geom_point(aes(x=LOT.SQFT, y=TOTAL.VALUE), color="navy", alpha = 0.40)



# 5) By modifying your code in Question 3, introduce rooms variable into the plot in
the size parameter.
#
d%>%
  ggplot()
geom_point(aes(x=TOTAL.VALUE, y=YR.BUILT), color="navy", alpha= 0.2, size= "ROOMS")


# 6) Create a bar plot in which x-axis shows ROOMS and y-axis shows average
TOTAL.VALUE.
#    Use aggregate function, or dplyr group_by & summarize approach
library(dplyr)

d %>%
  group_by(ROOMS) %>%
  summarise(
    MEANTOTALVALUE= mean(TOTAL.VALUE)
```

```
  ) %>%
  ggplot()+
  geom_bar(aes(x= ROOMS, y= MEANTOTALVALUE), stat = "identity")



# 7) Create a 4-panel chart in which the following plots show up;
# a) scatter plot of total value vs tax
# b) scatter plot of total value vs lot sqft
# c) box plot of total value and remodel
# d) box plot of lot sqr foot and remodel

PLOTa = d%>%
  ggplot()+
  geom_point(aes(x=TAX, y=TOTAL.VALUE))

PLOTb = d%>%
  ggplot()+
  geom_point(aes(x= LOT.SQFT, y= TOTAL.VALUE))

PLOTc = d%>%
  ggplot()+
  geom_boxplot(aes(x= REMODEL, y= TOTAL.VALUE))

PLOTd = d%>%
  ggplot()+
  geom_boxplot(aes(x= REMODEL, y=LOT.SQFT))

install.packages("gridExtra")
library(gridExtra)
grid.arrange(PLOTa, PLOTb, PLOTc, PLOTd)
```