DS6030 Module 1 Exercises

Alanna Hazlett

2024-05-19

Problem 1

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a)
We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
CEO salary is a continuous numerical variable, so we would perform a regression on this data. To understand the factors that affect CEO salary we would be interested in inference. n = 500 and p = 3

(b)
We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
This is a classification problem with two categories, success or failure. This is a prediction problem. n = 20 and p = 13

(c)
We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
Percent change is a continuous numerical variable, so we would perform a regression on this data. This is a prediciton problem. n = 52 (as there are 52 weeks in a year) and p = 3

Problem 2

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?
A parametric method makes an assumption on the shape of f, while non-parametric do not make these explicit assumptions.
Advantages of parametric: Does not require as many observations as non-parametric approach. It reduces the problem of estimating f down to estimating a set of parameters.
Disadvantages of parametric: The chosen model we use to estimate f will almost never match the true form of f. By assuming the functional form of f there is a smaller range of possible shapes for f.

Problem 3

Explore the dataset ISLR2::Boston

(a)
Create histograms and/or densityplots of each feature using ggplot2. Use patchwork to combine multiple graphs into a figure. Look for interesting patterns in the distributions. For example, are distributions highly skewed? Do you notice any outliers? Document your findings. Are there any variables that should be transformed?

```
Data<-ISLR2::Boston
?ISLR2::Boston
```

The response variable in this dataset is medv, which is the median value of owner-occupied homes in the thousands of dollars.

```
print(summary(Data))
```

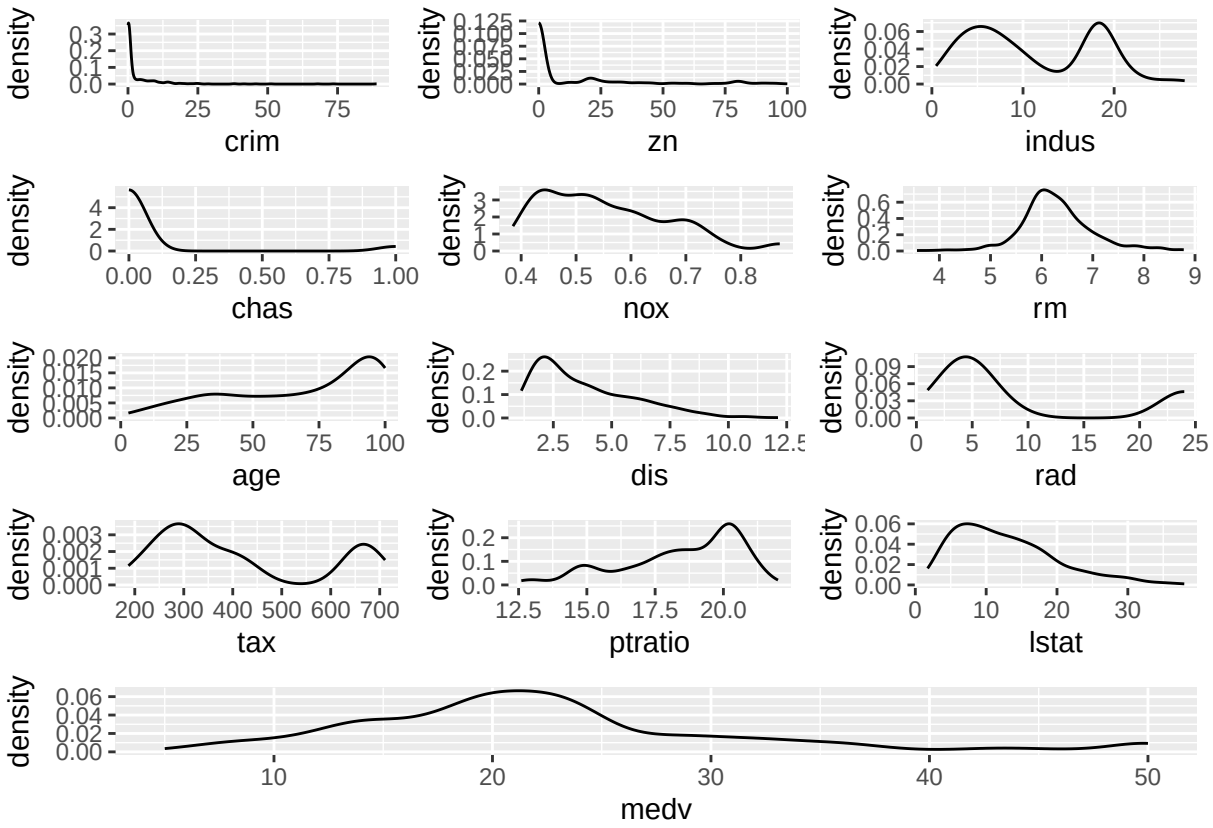
```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

Based on our summary information it seems there are outliers in crim, zn, indus, dis, rad, lstat, and medv, as there is a significant increase from the 3rd quartile value and the max value. Here I have filtered to show values of the data where for each of these variables the values are greater than the 3rd quartile.

For dis it appears that while there is a significant increase from the 3rd quartile to the max value there is a fairly even distribution between these values.For rad there is a significant increase from the second highest value of 8 to the max value of 24. This seems to show a large difference in the accessibility to highways, with most houses having a lower index of accessibility.

```
p_crim<-ggplot(Data,aes(x=crim))+
  geom_density()
```

```
p_zn<-ggplot(Data,aes(x=zn))+
  geom_density()
p_indus<-ggplot(Data,aes(x=indus))+
  geom_density()
p_chas<-ggplot(Data,aes(x=chas))+
  geom_density()
p_nox<-ggplot(Data,aes(x=nox))+
  geom_density()
p_rm<-ggplot(Data,aes(x=rm))+
  geom_density()
p_age<-ggplot(Data,aes(x=age))+
  geom_density()
p_dis<-ggplot(Data,aes(x=dis))+
  geom_density()
p_rad<-ggplot(Data,aes(x=rad))+
  geom_density()
p_tax<-ggplot(Data,aes(x=tax))+
  geom_density()
p_ptratio<-ggplot(Data,aes(x=ptratio))+
  geom_density()
p_lstat<-ggplot(Data,aes(x=lstat))+
  geom_density()
p_medv<-ggplot(Data,aes(x=medv))+
  geom_density()
(p_crim+p_zn+p_indus)/(p_chas+p_nox+p_rm)/(p_age+p_dis+p_rad)/(p_tax+p_ptratio+p_lstat)/p_medv
```

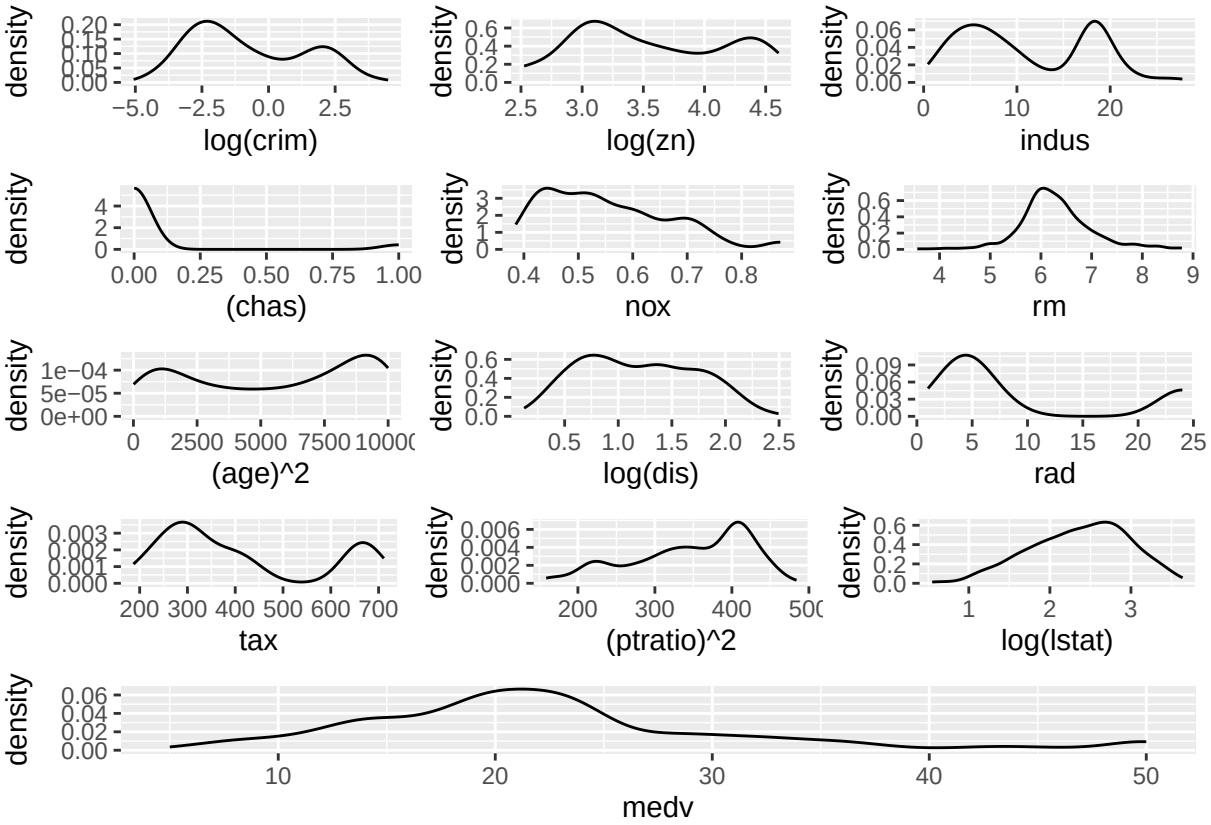


Highly skewed: crim, zn, chas, dis, and lstat are all skewed towards zero. Rad has is also skewed toward zero, but does has some significant density at about 22.5. Age is skewed towards the right at about 90. ptratio is

skewed towards the right with majority of the density occurring at 20. We will try log transforming crim, zn, dis, and lstat and try squaring age and ptratio to see if the distribution improves.

```
Data2<-Data %>%
  dplyr::mutate(
    log(crim),
    log(zn),
    log(dis),
    log(lstat),
    (age)**2,
    (ptratio)**2)
p_crim2<-ggplot(Data2,aes(x=log(crim)))+
  geom_density()
p_zn2<-ggplot(Data2,aes(x=log(zn)))+
  geom_density()
p_indus2<-ggplot(Data2,aes(x=indus))+
  geom_density()
p_chas2<-ggplot(Data2,aes(x=(chas)))+
  geom_density()
p_nox2<-ggplot(Data2,aes(x=nox))+
  geom_density()
p_rm2<-ggplot(Data2,aes(x=rm))+
  geom_density()
p_age2<-ggplot(Data2,aes(x=(age)^2))+
  geom_density()
p_dis2<-ggplot(Data2,aes(x=log(dis)))+
  geom_density()
p_rad2<-ggplot(Data2,aes(x=rad))+
  geom_density()
p_tax2<-ggplot(Data2,aes(x=tax))+
  geom_density()
p_ptratio2<-ggplot(Data2,aes(x=(ptratio)^2))+
  geom_density()
p_lstat2<-ggplot(Data2,aes(x=log(lstat)))+
  geom_density()
p_medv2<-ggplot(Data2,aes(x=medv))+
  geom_density()
(p_crim2+p_zn2+p_indus2)/(p_chas2+p_nox2+p_rm2)/(p_age2+p_dis2+p_rad2)/(p_tax2+p_ptratio2+p_lstat2)/p_m
```

```
## Warning: Removed 372 rows containing non-finite outside the scale range
## (`stat_density()`).
```

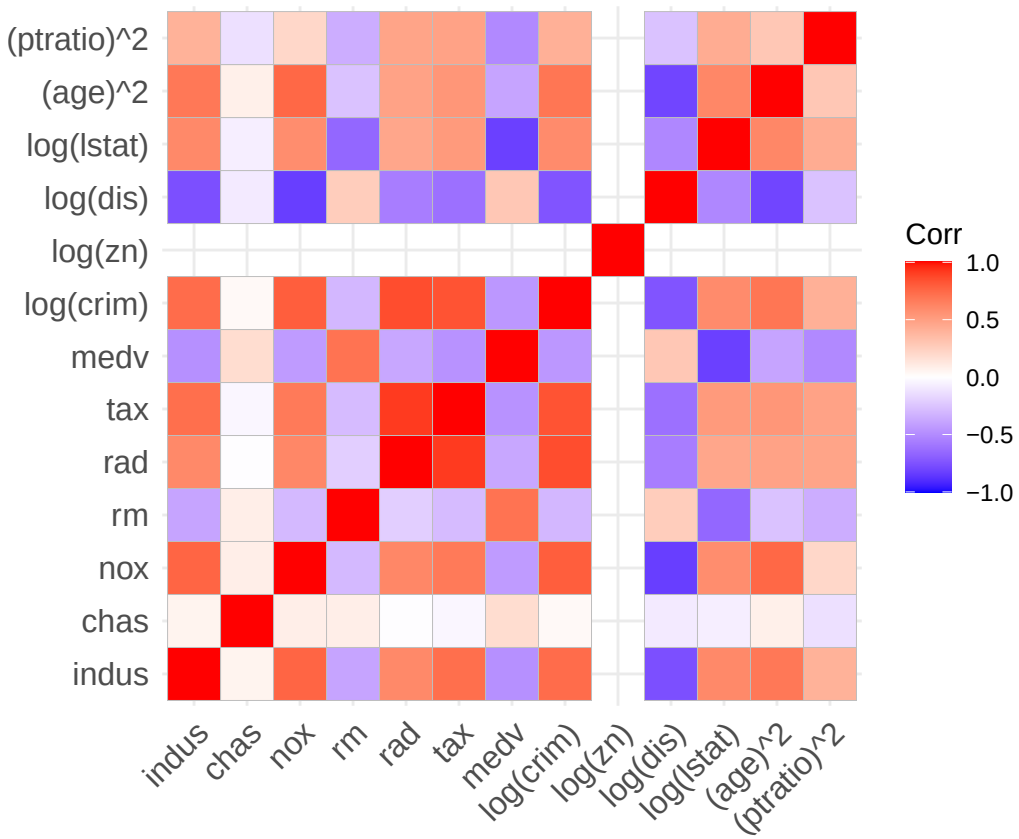


It appears that log transforming crim, zn, dis, lstat, and rad does help with the distribution and squaring age and ptratio also helped the distribution.

(b)

With the processed dataset, create a number of plots using ggplot2 to explore the relationship between the variables. Document your findings. Do you see strong correlation between some of the variables? What type of correlation do you see? What would be the consequence of these if you want to train a regression model?

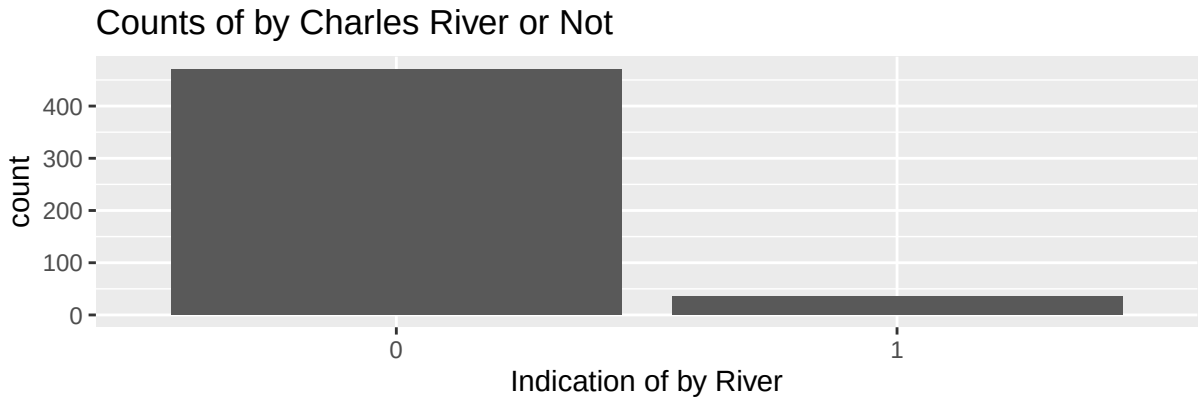
```
corr<-Data2 %>%
  dplyr::select(-c(crim,zn,age,dis,ptratio,lstat)) %>%
  dplyr::select(where(is.numeric)) %>%
  cor()
ggcorrplot(corr)
```

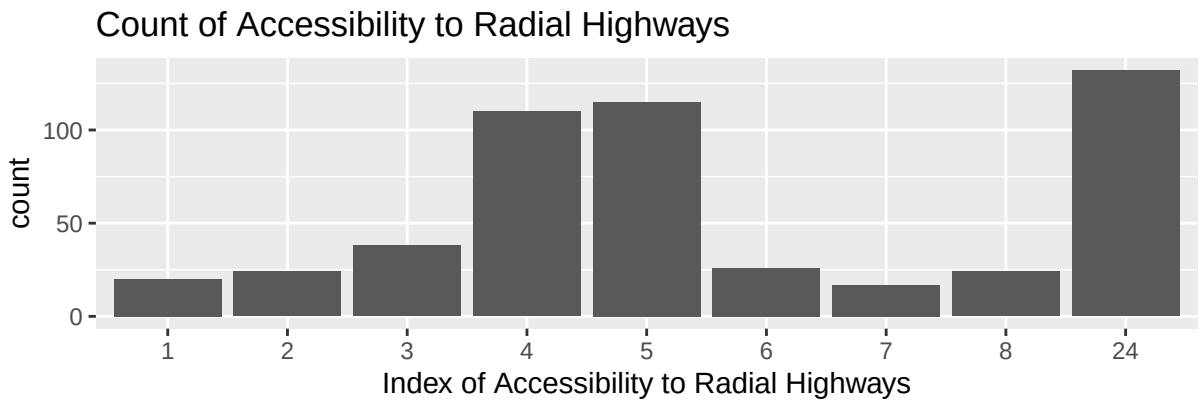


Based on our heatmap of the corrleation values, we can see that tax and rad, tax and log(crim), and rad and log(crim) are highly positively correlated. We see highly negative correlations between indus and log(dis), nox and log(dis), medv and log(stat), and log(dis) and age squared. The consequence of these high correlations in regards to regression modeling is that several of these variables would be excluded from the model, as we can model the relationship of the predictors to the response variable utilizing only one of the variables to represent the relationship. medv and log(lstat) have a high correlation, it is likely that log(lstat) may have a larger impact on our response variable than the other variables do.

Proportion Visualizations for Categorical and Discrete Variables

```
Data2$chas<-factor(Data2$chas)
Data2$rad<-factor(Data2$rad)
p1<-Data2 %>%
  ggplot2::ggplot(aes(x=chas))+
  geom_bar()+
  labs(x="Indication of by River",title="Counts of by Charles River or Not")
p2<-ggplot2::ggplot(Data2,aes(x=rad))+
  geom_bar()+
  labs(x="Index of Accessibility to Radial Highways",title="Count of Accessibility to Radial Highways")
p1 / p2
```



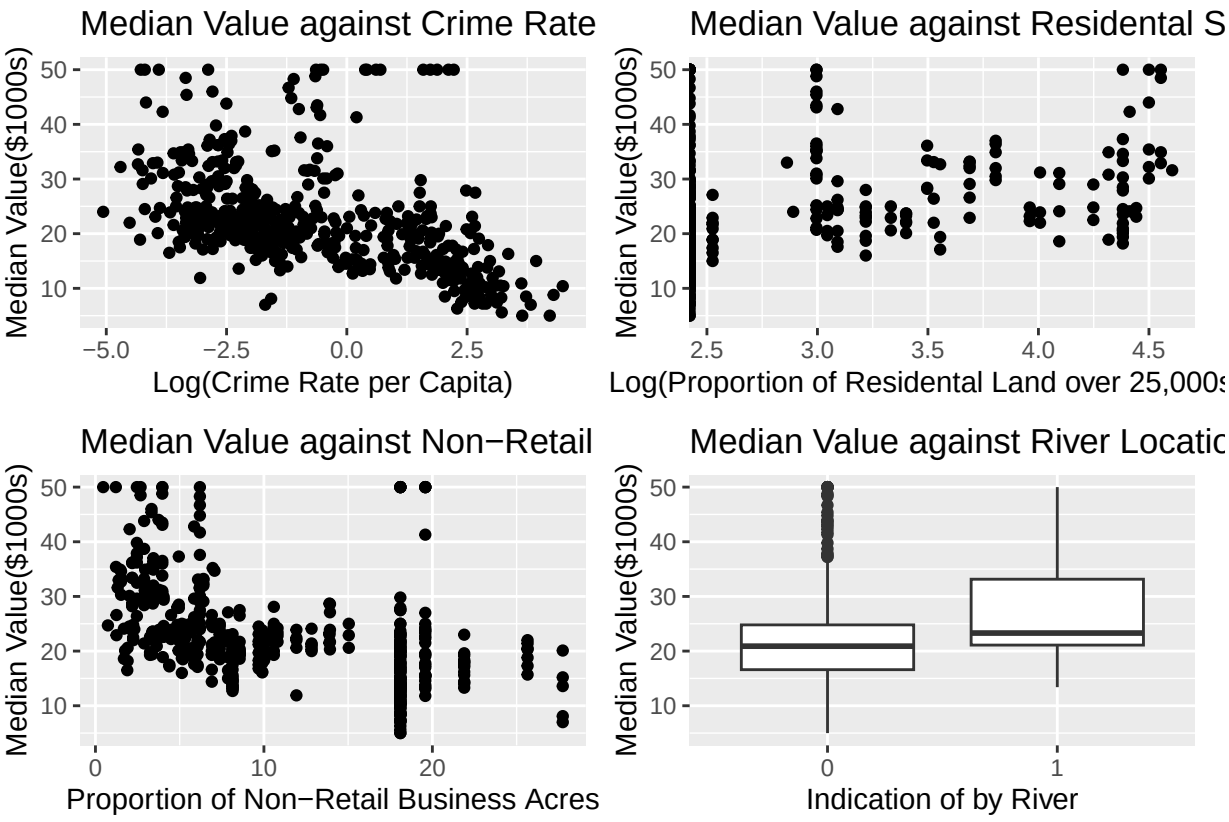


We can see that the majority of the suburbs of Boston are not by the Charles River, as indicated by a zero value for the variable chas. In the second plot we can see that many suburbs of Boston seem to have an index of accessiblity to radial highways of 4 or 5, this may be a relatively middle value for most suburbs. The index of 24 has the highest proportion of accessibility, perhaps this is the downtown area that has a higher density of population and is closer to the highways.

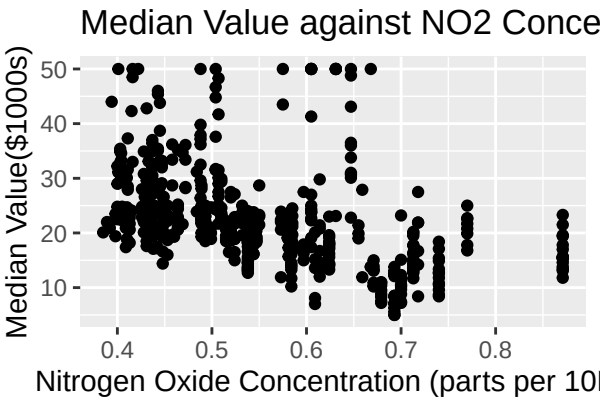
Predictors VS Response for Linear Regression

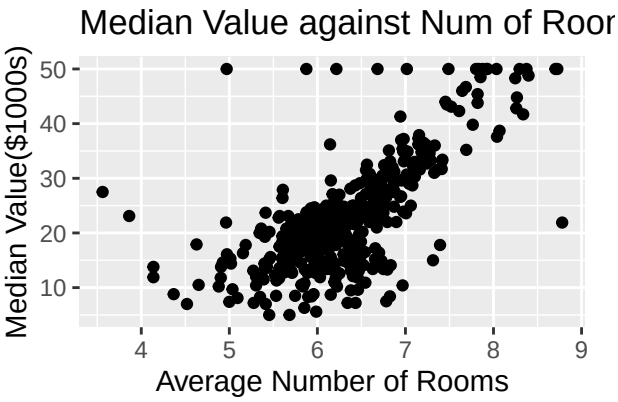
```
p3<-ggplot(Data2,aes(x=log(crim),y=medv))+
  geom_point()+
  labs(x="Log(Crime Rate per Capita)",y="Median Value($1000s)",title="Median Value against Crime Rate")
p4<-ggplot(Data2, aes(x=log(zn), y=medv))+
  geom_point()+
  labs(x="Log(Proportion of Residental Land over 25,000sqft)", y ="Median Value($1000s)",title="Median
p5<-ggplot(Data2, aes(x=indus, y=medv))+
  geom_point()+
  labs(x="Proportion of Non-Retail Business Acres", y="Median Value($1000s)", title="Median Value again
p6<-ggplot(Data2,aes(x=chas,y=medv))+
  geom_boxplot()+
  labs(x="Indication of by River",y="Median Value($1000s)",title="Median Value against River Location")
p7<-ggplot(Data2,aes(x=nox,y=medv))+
  geom_point()+
  labs(x="Nitrogen Oxide Concentration (parts per 10M)",y="Median Value($1000s)",title="Median Value ag
p8<-ggplot(Data2,aes(x=rm,y=medv))+
  geom_point()+
  labs(x="Average Number of Rooms",y="Median Value($1000s)",title="Median Value against Num of Rooms")
p9<-ggplot(Data2,aes(x=age^2,y=medv))+
  geom_point()+
  labs(x="(Homes Built Prior to 1940)^2",y="Median Value($1000s)",title="Median Value against Age")
```

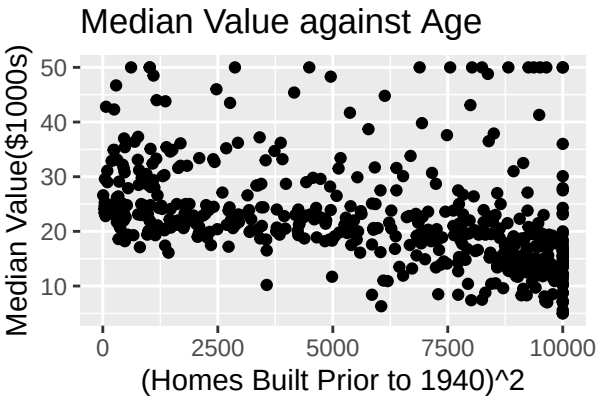
```
p10<-ggplot(Data2,aes(x=log(dis),y=medv))+
  geom_point()+
  labs(x="Log(Distance to Employment Centers)",y="Median Value($1000s)",title="Median Value against Dis
p11<-ggplot(Data2,aes(x=rad,y=medv))+
  geom_point()+
  labs(x="Accessibility to Radial Highways",y="Median Value ($1000s)",title="Median Value against Acces
p12<-ggplot(Data2,aes(x=tax,y=medv))+
  geom_point()+
  labs(x="Tax Rate per $10,000",y="Median Value($1000s)",title="Median Value against Tax Rate")
p13<-ggplot(Data2,aes(x=ptratio,y=medv))+
  geom_point()+
  labs(x="Pupil:Teacher Ratio",y="Median Value($1000s)",title="Median Value against Pupil:Teacher Ratio
p14<-ggplot(Data2,aes(x=log(lstat),y=medv))+
  geom_point()+
  labs(x="Log(Lower Status of Population)",y="Median Value($1000s)",title="Median Value against Lower S
(p3 + p4)/(p5 + p6)
```

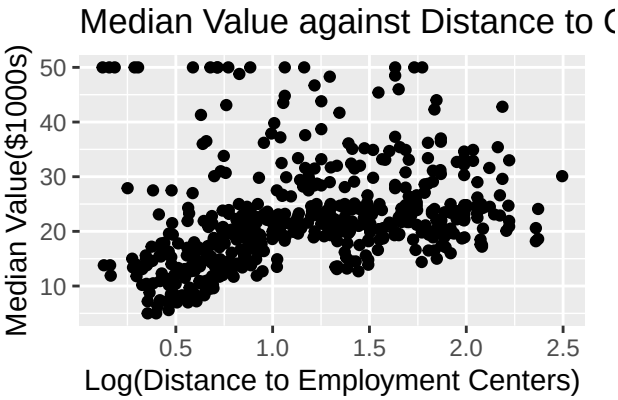



```
(p7+p8)/(p9+p10)
```

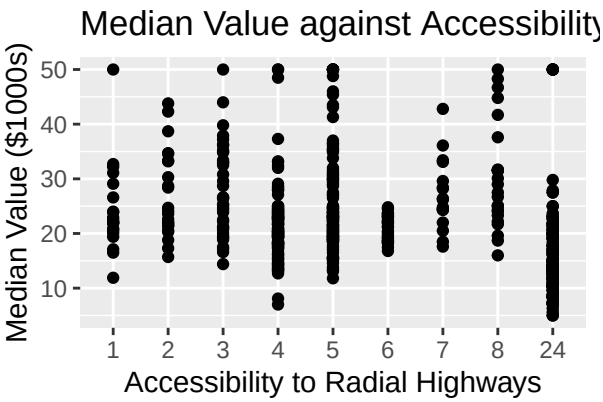


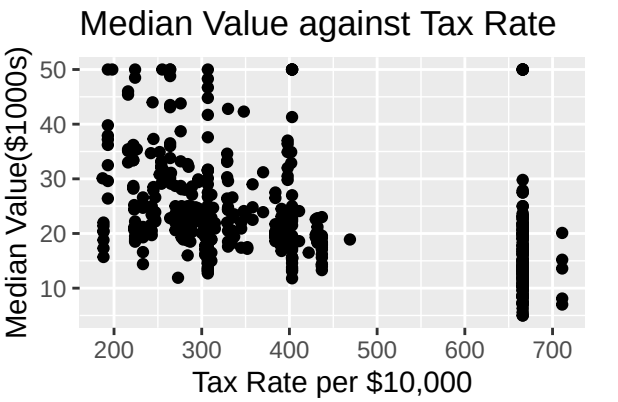


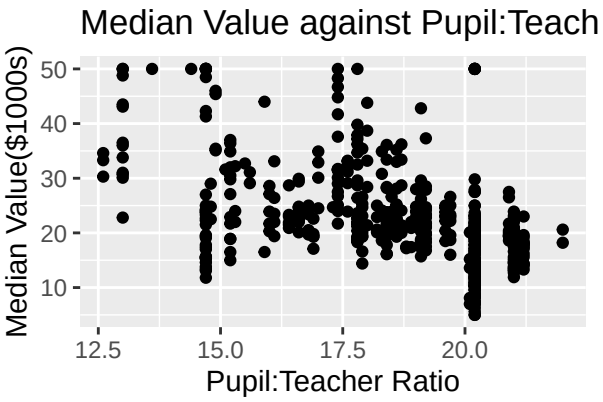


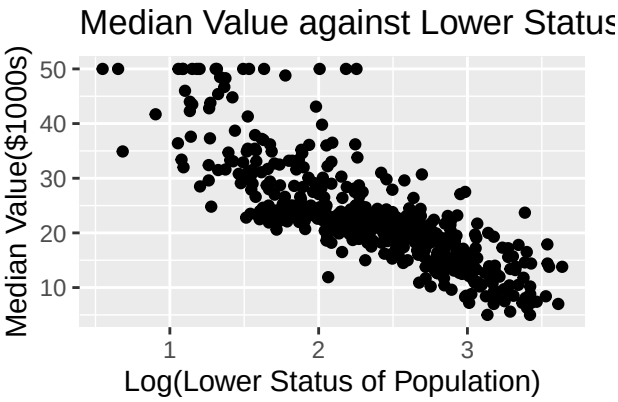



```
(p11+p12)/(p13+p14)
```









- We see as crime rate increases the median value of houses decreases, this appears to be a relatively negative linear relationship. There are some outlying data points between Log(0-2) that maintain higher house values despite the higher crime rates.

- There is a slightly positive linear relationship for most of the data for zn, the proportion of land zoned for lots over 25,000 square feet. We have a large number of suburbs of Log(2.25) that span the entire range of values of homes.

- There is a negative relationship between indus, the proportion of non-retail business acres per town and median values of homes. As the proportion gets higher the values become more discrete.

- We see that suburbs near the river have a higher interquartile range for median values of homes, however the median is only slightly higher for the suburbs by the river than the ones not by the river.

- There is a negative linear relationship between nitrogen oxide concentration and median home values, with some outliers at about 0.6-0.7 maintaining a higher than average median home value.

- There is a positive linear relationship between the average number of rooms in a house for each suburb and the median home value.

- There is a negative linear association between age, the proportion of owner-occupied units built prior to 1940, and median home values. There are a handful of observations that maintain higher median home values than the average across all ages of homes.

- There is a positive linear relationship between distance to employment centers and median home values. This could indicate that areas of lower home values and likely lower income need more readily available access to employment centers and more expensive homes are further away from these resources as they are not as needed. There are a handful of observations that maintain a higher median home value across all distances.

- Generally across most values of accessibility index for highways the values span most of the range. The most significant is that the accessibility index of 24 has far more median home values that are lower in the range.

- For tax rate it is difficult to determine a relationship. There are observations for tax rate between 200-500 that span most of the range of median home values. Then there is a more discrete pattern for tax rates of 650-750 and they tend to be in the lower part of the range of median home values.

- There is a negative linear relationship between ptratio, pupil to teacher ratio, and median home value. This indicates that areas of lower home values have more students to a teacher than areas of higher home values.

- There is a negative linear relationship between the log(lstat), the percent of lower status of the population and median home values. This means that as lower status percent increases the home values are decreasing on average.