STAD80: Homework #2

Yulun Wu

Due: 2023-02-16

Contents

Question 1 (30 Points) A Simple Linear Regression

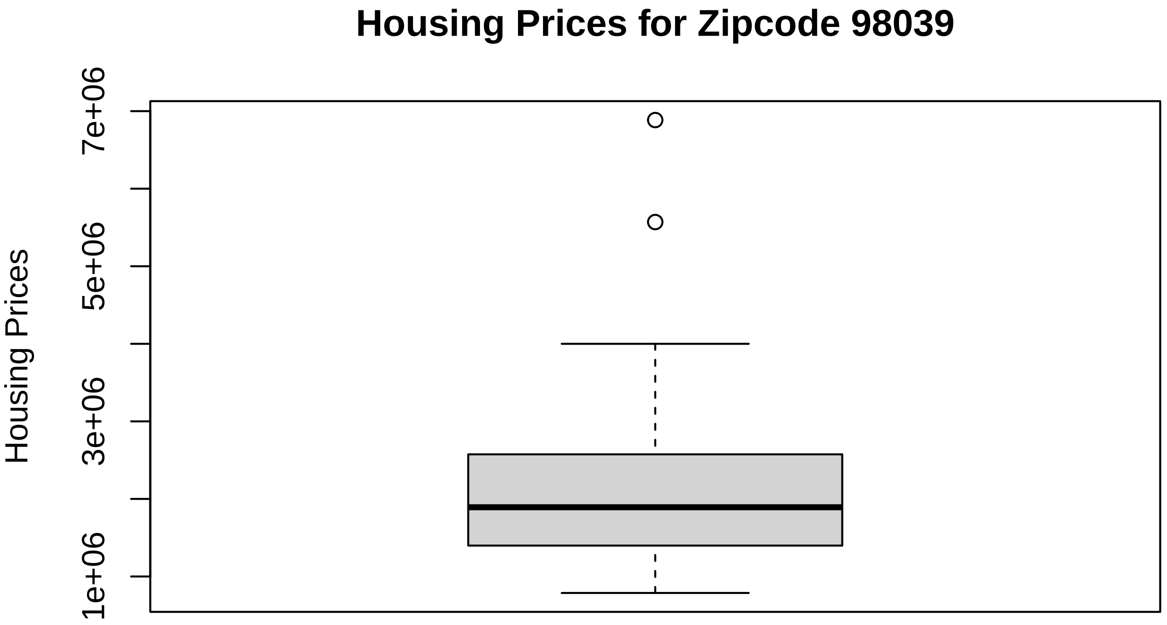
Answer:

(a)

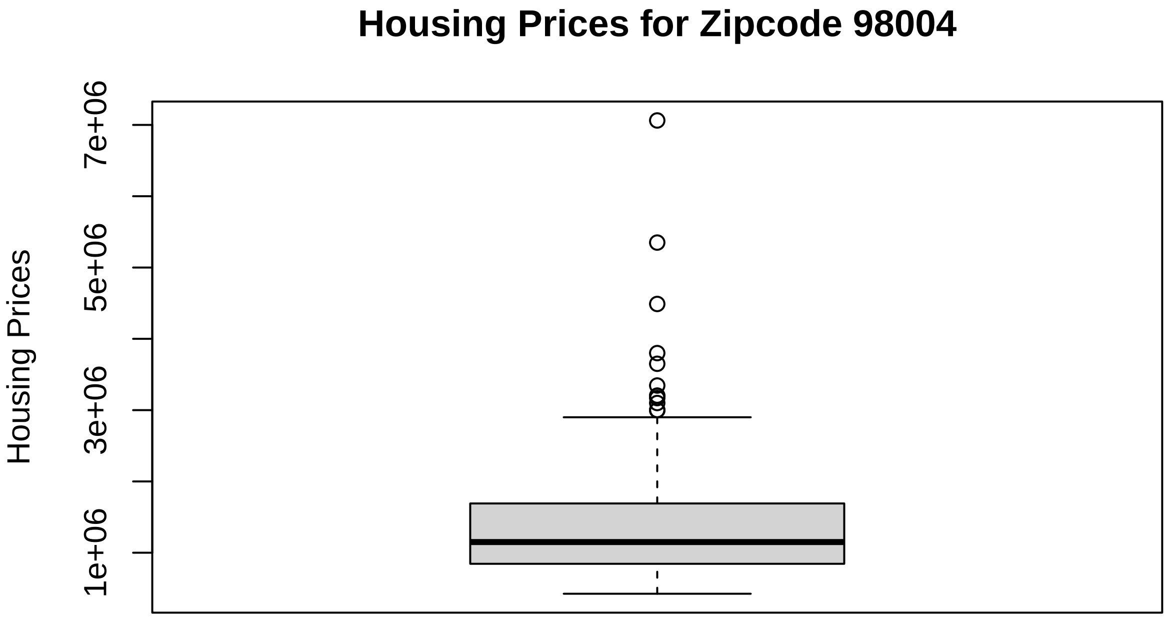
```r
hp = read.csv("housingprice.csv",header = T)
hp$zipcode = factor(hp$zipcode) # convert zipcode into factor
average_price = tapply(hp$price, hp$zipcode, mean) # average price of each zipcode
sort_zip_byavgprice = sort(average_price,decreasing=T) # sort average price of each zipcode in decreasi
names(sort_zip_byavgprice[1:3]) # he top 3 zipcodes whose average housing prices are most expensive
```

```
## [1] "98039" "98004" "98040"
```

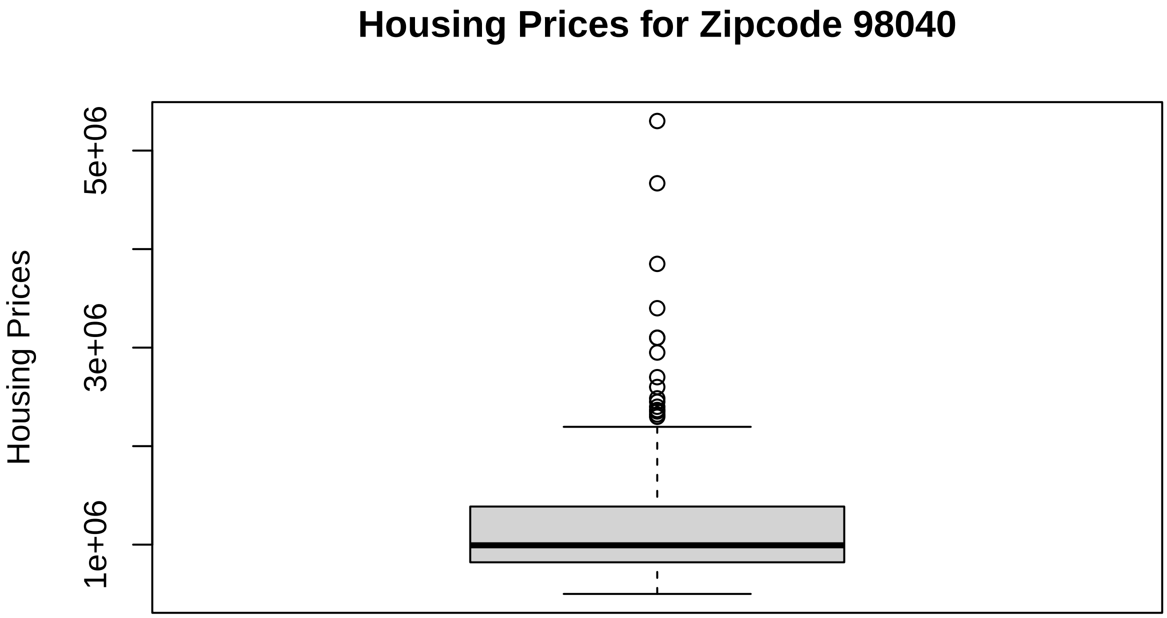
```r
boxplot(hp[which(hp["zipcode"]==names(sort_zip_byavgprice[1])),"price"],main="Housing Prices for Zipcod
```



```
boxplot(hp[which(hp["zipcode"]==names(sort_zip_byavgprice[2])),"price"],main="Housing Prices for Zipcode
```



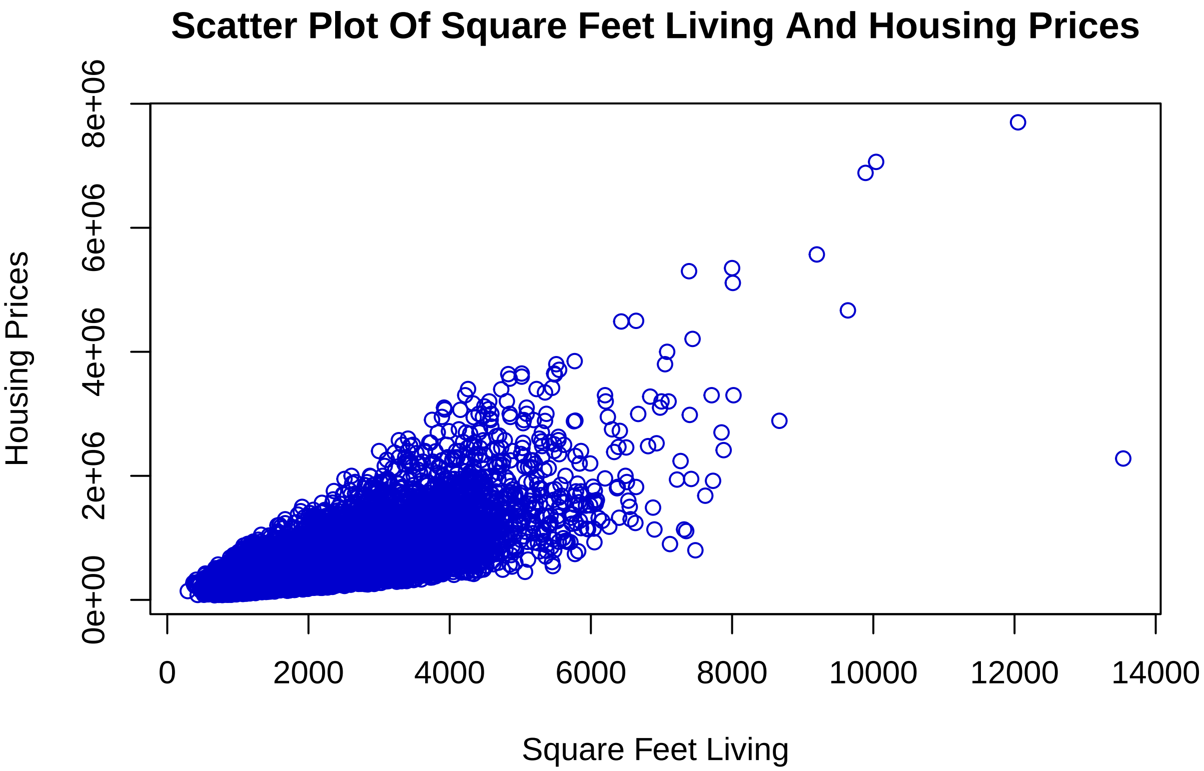
```
boxplot(hp[which(hp["zipcode"]==names(sort_zip_byavgprice[3])),"price"],main="Housing Prices for Zipcode
```



The top 3 zipcodes whose average housing prices are most expensive are: 98039, 98004, 98040 (listed in decreasing order of average housing prices).

(b)

```
plot(hp$sqft_living,hp$price,col="blue3",ylab="Housing Prices",xlab="Square Feet Living",main="Scatter I
```



(c)

```r
# Read file of training and testing data
train_data = read.csv("train.data.csv",header = T)
test_data = read.csv("test.data.csv",header = T)
# Fit linear model
model1 = lm(price~bedrooms+bathrooms+sqft_living+sqft_lot,data = train_data)
Ybar=mean(test_data[,"price"])
SST=sum((test_data[,"price"]-Ybar)^2)

# Calculate R^2 for testing data
Yhat=predict(model1,test_data) # Predicted Y
SSE=sum((test_data[,"price"]-Yhat)^2)

cat("R^2 on training data is",summary(model1)$r.squared,"\n")
```

```
## R^2 on training data is 0.5101139
```

```r
cat("R^2 on testing data is",1-(SSE/SST),"\n")
```

```
## R^2 on testing data is 0.5049945
```

R^2 of the model on training data is 0.5101139. R^2 on testing data is 0.5049945.

(d)

```r
model2 = update(model1, .~. + zipcode)
# Calculate R^2 for testing data with model after adding zipcode
Ybar=mean(test_data[,"price"])
SST=sum((test_data[,"price"]-Ybar)^2)
Yhat=predict(model2,test_data) # Predicted Y
SSE=sum((test_data[,"price"]-Yhat)^2)
```

```
cat("R^2 on training data is",summary(model2)$r.squared,"\n")
```

```
## R^2 on training data is 0.5162971
```

```
cat("R^2 on testing data is",1-(SSE/SST),"\n")
```

```
## R^2 on testing data is 0.5120097
```

R^2 of the model on training data after adding zipcode to linear model is 0.5162971. R^2 on testing data after adding zipcode to linear model is 0.5120097.

(e)

```
# Read file of training and testing data
fancyhouse = read.csv("fancyhouse.csv",header = T)
fancyhouse
```

```
##   X bedrooms bathrooms sqft_living sqft_lot floors zipcode condition grade
## 1 1        8        25       50000   225000      4   98039        10    10
##   waterfront view sqft_above sqft_basement yr_built yr_renovated      lat
## 1          1    4      37500         12500     1994         2010 47.62761
##        long sqft_living15 sqft_lot15
## 1 -122.2421          5000      40000
```

```
price_fancyhouse=predict(model2,fancyhouse)
price_fancyhouse
```

```
##        1
## 15642273
```

The predicted price for Bill Gates' house is 15642273. I don't think the predicted price is reasonable, the real price probably is 10 times of that, based on what I saw from luxury house tour on YouTube, the house that has 5 bedrooms already worth more than this predicted price.

(f)

Since $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (Y_i - X\widehat{\beta})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. SST remains unchanged no matter how $X\widehat{\beta}$ changes, so the only term that can change is $SSE = \sum_{i=1}^n (Y_i - X\widehat{\beta})^2$, and since $SSE = \sum_{i=1}^n (Y_i - X\widehat{\beta})^2 \propto ||Y - X\widehat{\beta}||_2^2$, so we can compare R^2 of model with d covariates and d+1 covariates by comparing OLS. And model with smaller OLS will have bigger R^2. Further more, since Y remains unchanged, basically we are comparing $X\widehat{\beta}$ and $X_1\widehat{\beta}_1$, the one closer to Y will gives smaller OLS therefore bigger R^2.

$\because \widehat{\beta} = X(X'X)^{-1}X'Y$,$\widehat{\beta}_1 = X_1(X_1'X_1)^{-1}X_1'Y$ and $n > d+1$

$\therefore \widehat{\beta}$ is a d by 1 vector and $\widehat{\beta}_1$ is d+1 by 1 vector

$\therefore$ We have chance to get a better approximation of Y by $X_1\widehat{\beta}_1$ compare to $X\widehat{\beta}$ since we have additional flexibility in $\widehat{\beta}_1$ (ie: df of SSE with d+1 covariates is 1 greater than df of SSE with d covariates). Also R^2 can stay unchanged if $\widehat{\beta}_{d+1} = 0$ because basically $X\widehat{\beta} = X_1\widehat{\beta}_1$ in this case, so $SSE_{d+1} = SSE_d$ implies $R^2_{d+1} = R^2_d$.

$\therefore |Y - X_1\widehat{\beta}_1| \leq |Y - X\widehat{\beta}|$

$\therefore SSE_{d+1} \leq SSE_d$

$\therefore R^2_{d+1} \geq R^2_d$

Thus, adding another covariate in the model never hurts R^2 over the training data.

Question 2 (20 Points) Feature Engineering

Answer:

(a)

```
model3 = update(model2, .~. + bedrooms*bathrooms)
# Calculate R^2 for testing data with model after adding bedrooms*bathrooms
Ybar=mean(test_data[,"price"])
SST=sum((test_data[,"price"]-Ybar)^2)
Yhat=predict(model3,test_data) # Predicted Y
SSE=sum((test_data[,"price"]-Yhat)^2)

cat("R^2 on training data is",summary(model3)$r.squared,"\n")
```

```
## R^2 on training data is 0.5223738
```

```
cat("R^2 on testing data is",1-(SSE/SST),"\n")
```

```
## R^2 on testing data is 0.5165114
```

R^2 of the model with interaction of bedrooms and bathrooms added on training data is 0.5223738. R^2 of the model with interaction of bedrooms and bathrooms added on testing data is 0.5165114.

(b)

```
model4 = update(model3, .~. + bathrooms*sqft_living)
# Calculate R^2 for testing data with model after adding bathrooms*sqft_living
Ybar=mean(test_data[,"price"])
SST=sum((test_data[,"price"]-Ybar)^2)
Yhat=predict(model4,test_data) # Predicted Y
SSE=sum((test_data[,"price"]-Yhat)^2)

cat("R^2 on training data is",summary(model4)$r.squared,"\n")
```

```
## R^2 on training data is 0.5490765
```

```
cat("R^2 on testing data is",1-(SSE/SST),"\n")
```

```
## R^2 on testing data is 0.5451303
```

R^2 of the model with interaction of sqft_living and bathrooms added on testing data is 0.5451303.

(c)

```
model5 = update(model2, .~. + poly(bedrooms, 3)+ poly(bathrooms, 3))
# Calculate R^2 for testing data with model after adding bedrooms*bathrooms
Ybar=mean(test_data[,"price"])
SST=sum((test_data[,"price"]-Ybar)^2)
Yhat=predict(model5,test_data) # Predicted Y
SSE=sum((test_data[,"price"]-Yhat)^2)

cat("R^2 on training data is",summary(model5)$r.squared,"\n")
```

```
## R^2 on training data is 0.5424973
```

```
cat("R^2 on testing data is",1-(SSE/SST),"\n")
```

```
## R^2 on testing data is 0.5285074
```

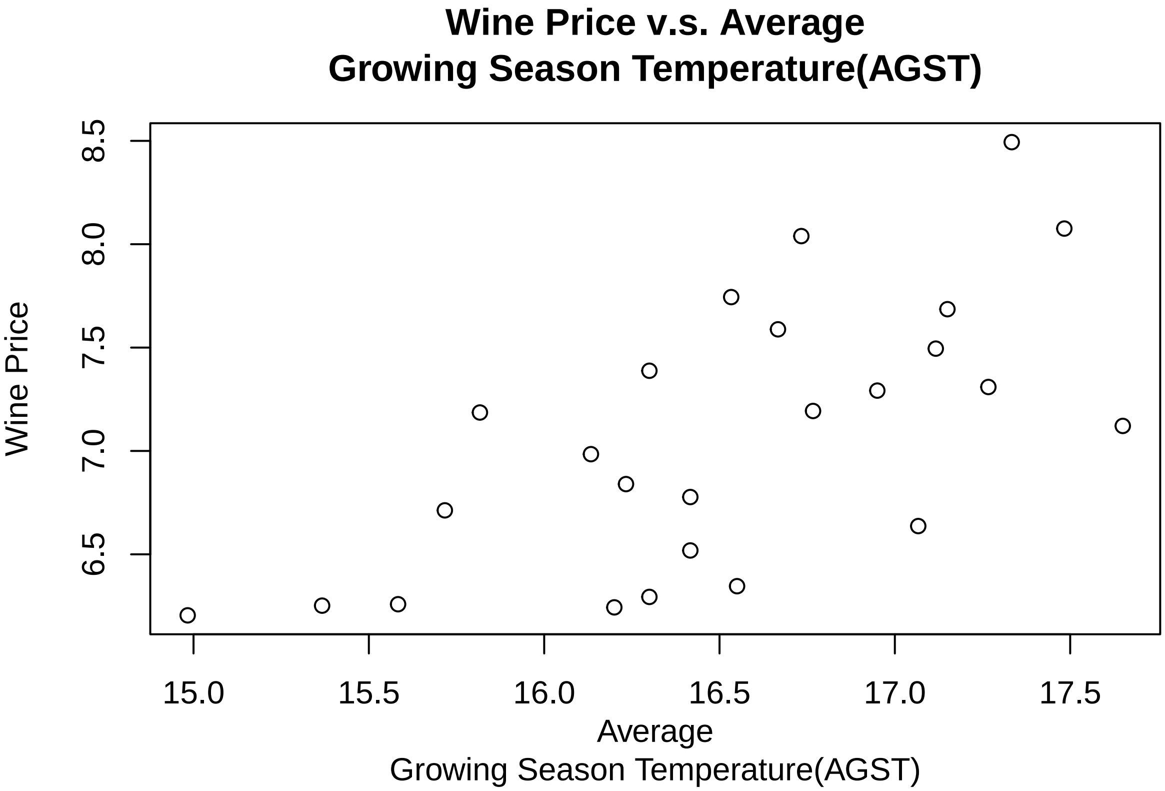
R^2 of the model with polynomial term with degree 2 and 3 of bedrooms and bathrooms added on training data is 0.5424973. R^2 of the model with polynomial term with degree 2 and 3 of bedrooms and bathrooms added on testing data is 0.5285074.

Question 3 (20 Points) Wine Pricing

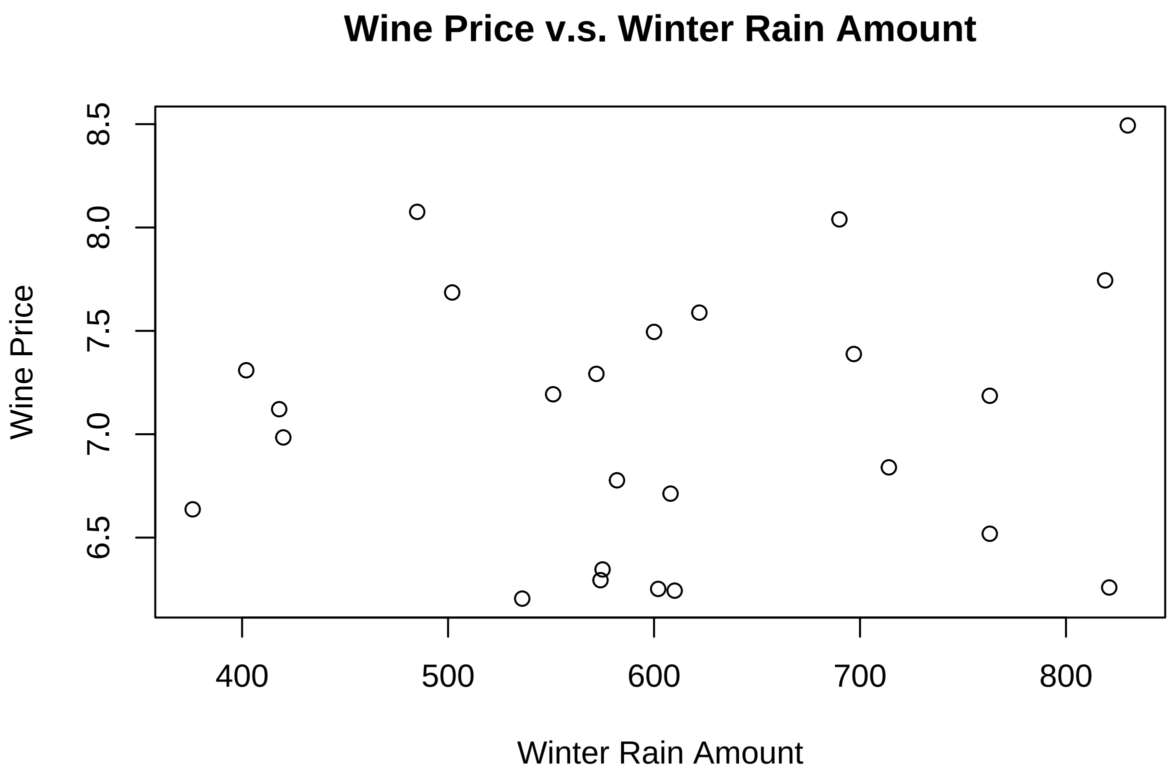
Answer:

Part I

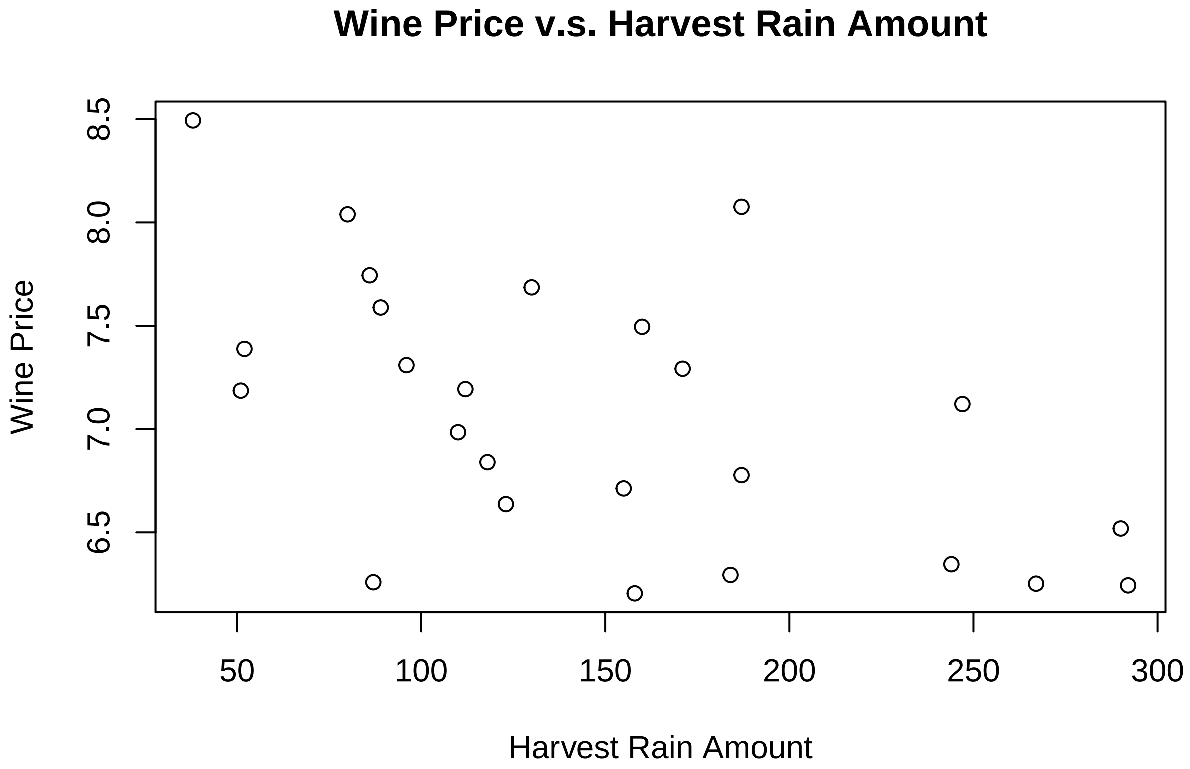
```
# Read file of wine data
wine = read.csv("wine.csv",header = T)
# 4 Scatter plots
plot(wine$AGST,wine$Price,main="Wine Price v.s. Average
Growing Season Temperature(AGST)",ylab="Wine Price",xlab="Average
Growing Season Temperature(AGST)")
```



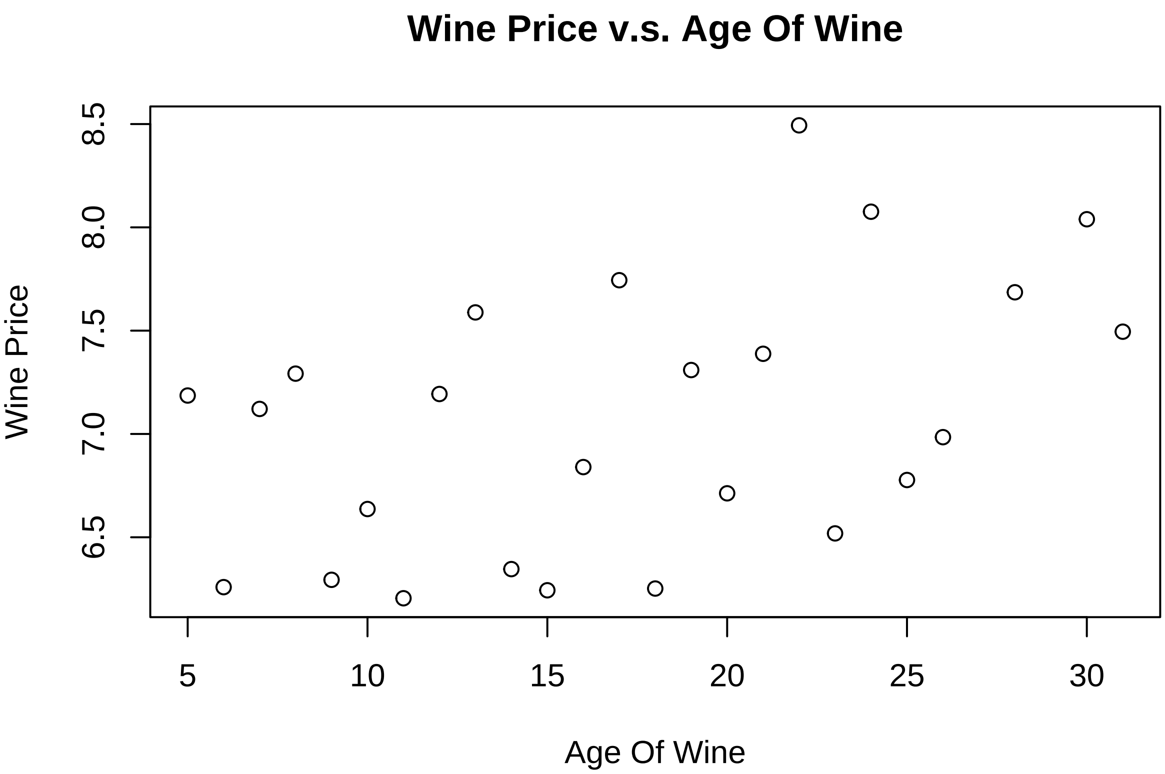
```
plot(wine$WinterRain,wine$Price,main="Wine Price v.s. Winter Rain Amount",ylab="Wine Price",xlab="Winte
```



```
plot(wine$HarvestRain,wine$Price,main="Wine Price v.s. Harvest Rain Amount",ylab="Wine Price",xlab="Har
```



```
plot(wine$Age,wine$Price,main="Wine Price v.s. Age Of Wine",ylab="Wine Price",xlab="Age Of Wine")
```



```
mean_price = mean(wine[,"Price"]) # ybar
mean_AGST = mean(wine[,"AGST"]) # mean of AGST
mean_WinterRain = mean(wine[,"WinterRain"]) # mean of WinterRain
mean_HarvestRain = mean(wine[,"HarvestRain"]) # mean of HarvestRain
mean_Age = mean(wine[,"Age"]) # mean of Age

rAGST = sum((wine[,"AGST"]-mean_AGST)*(wine[,"Price"]-mean_price))/sqrt(sum((wine[,"AGST"]-mean_AGST)^2
rAGST
```

```
## [1] 0.6595629
```

```
rWinterRain = sum((wine[,"WinterRain"]-mean_WinterRain)*(wine[,"Price"]-mean_price))/sqrt(sum((wine[,"Wi
rWinterRain
```

```
## [1] 0.1366505
```

```
rHarvestRain = sum((wine[,"HarvestRain"]-mean_HarvestRain)*(wine[,"Price"]-mean_price))/sqrt(sum((wine[
rHarvestRain
```

```
## [1] -0.5633219
```

```
rAge = sum((wine[,"Age"]-mean_Age)*(wine[,"Price"]-mean_price))/sqrt(sum((wine[,"Age"]-mean_Age)^2)*sum
rAge
```

```
## [1] 0.4477679
```

Based on the scatter plot and Pearson's correlation calculated, average growing season temperature (AGST) is most correlated with Price, their Pearson's correlation is 0.6595629.

Part II

```
model1 = lm(Price~AGST,data=wine)
cat("Fitted coefficient of model Price~AGST is",summary(model1)$coef[1],",", summary(model1)$coef[2],"\n
## Fitted coefficient of model Price~AGST is -3.417761 , 0.6350943
```

```
cat("R^2 of marginal model of Price~AGST is",summary(model1)$r.squared,"\n")
```

```
## R^2 of marginal model of Price~AGST is 0.4350232
```

The fitted coefficient values for Price~AGST is

$$\beta_0 = -3.4177613$$

and

$$\beta_1 = 0.6350943$$

. R^2 is 0.4350232.

Part III

```
# Read file of winetest data
winetest = read.csv("winetest.csv",header = T)

Ybar = mean(winetest$Price)
SST = sum((winetest$Price-Ybar)^2)

model2 = update(model1, .~. + HarvestRain) # Add HarvestRain
cat("R^2 for training data after adding HarvestRain:",summary(model2)$r.squared,"\n")
```

```
## R^2 for training data after adding HarvestRain: 0.7073708
```

```
# Calculate R^2 for testing data with model after adding HarvestRain
Yhat2=predict(model2,winetest) # Predicted Y
SSE = sum((winetest$Price-Yhat2)^2)
cat("R^2 for testing data with model after adding HarvestRain:",1-(SSE/SST),"\n")
```

```
## R^2 for testing data with model after adding HarvestRain: -2.503339
```

```
model3 = update(model2, .~. + Age) # Add Age
summary(model3)$r.squared # R^2 for training data after adding Age
```

```
## [1] 0.7900362
```

```
cat("R^2 for training data after adding Age:",summary(model3)$r.squared,"\n")
```

```
## R^2 for training data after adding Age: 0.7900362
```

```
# Calculate R^2 for testing data with model after adding Age
Yhat3=predict(model3,winetest) # Predicted Y
SSE = sum((winetest$Price-Yhat3)^2)
cat("R^2 for testing data with model after adding Age:",1-(SSE/SST),"\n")
```

```
## R^2 for testing data with model after adding Age: -0.5080824
```

```
model4 = update(model3, .~. + WinterRain) # Add WinterRain
summary(model4)$r.squared # R^2 for training data after adding WinterRain
```

```
## [1] 0.8285662
```

```
cat("R^2 for training data after adding WinterRain:",summary(model4)$r.squared,"\n")
```

```
## R^2 for training data after adding WinterRain: 0.8285662
```

```
# Calculate R^2 for testing data with model after adding WinterRain
Yhat4=predict(model4,winetest) # Predicted Y
```

```
SSE = sum((winetest$Price-Yhat4)^2)
cat("R^2 for testing data with model after adding WinterRain:",1-(SSE/SST),"\n")
```

```
## R^2 for testing data with model after adding WinterRain: 0.3343905
```

```
model5 = update(model4, .~. + FrancePop) # Add FrancePop
summary(model5)$r.squared # R^2 for training data after adding FrancePop
```

```
## [1] 0.8293592
```

```
cat("R^2 for training data after adding FrancePop:",summary(model5)$r.squared,"\n")
```

```
## R^2 for training data after adding FrancePop: 0.8293592
```

```
# Calculate R^2 for testing data with model after adding FrancePop
Yhat5=predict(model5,winetest) # Predicted Y
SSE = sum((winetest$Price-Yhat5)^2)
cat("R^2 for testing data with model after adding FrancePop:",1-(SSE/SST),"\n")
```

```
## R^2 for testing data with model after adding FrancePop: 0.2120672
```

```
summary(model4)$coef
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -3.429980187 1.7658975180 -1.942344 6.631093e-02
## AGST         0.607209348 0.0987022158  6.151932 5.197012e-06
## HarvestRain -0.003971534 0.0008537981 -4.651608 1.537556e-04
## Age          0.023930832 0.0080968750  2.955564 7.818874e-03
## WinterRain   0.001075505 0.0005072784  2.120148 4.669359e-02
```

Based on testing R^2, we should choose model of Price~AGST+HarvestRain+Age+WinterRain. Since $\widehat{\beta}_{HarvestRain}$ is negative, so more rain in harvest season will reduce the wine price. Since $\widehat{\beta}_{AGST}$, $\widehat{\beta}_{Age}$, $\widehat{\beta}_{WinterRain}$ are positive, so higher average growing season temperature (AGST) and/or older age and/or more rain in winter season will increase the wine price. The interpretation of my model agree with Prof. Ashenfelter's finding.

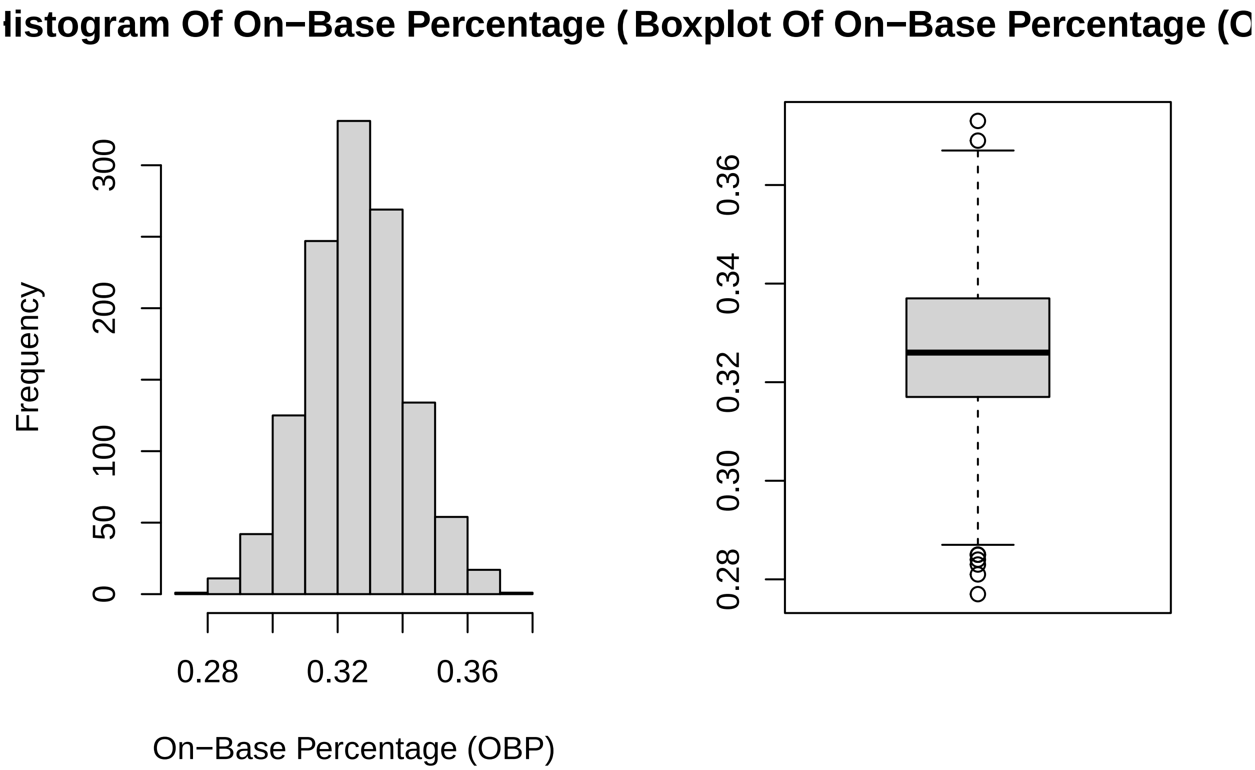
Question 4 (30 Points) Moneyball: The Analytics Edge in Sports

Answer:

Part I

```
# Read file of baseball data
baseball = read.csv("baseball.csv",header = T)

# Histogram and Boxplot of OBP
par(mfrow=c(1,2))
hist(baseball$OBP,main="Histogram Of On-Base Percentage (OBP)",xlab="On-Base Percentage (OBP)")
boxplot(baseball$OBP,main="Boxplot Of On-Base Percentage (OBP)")
```



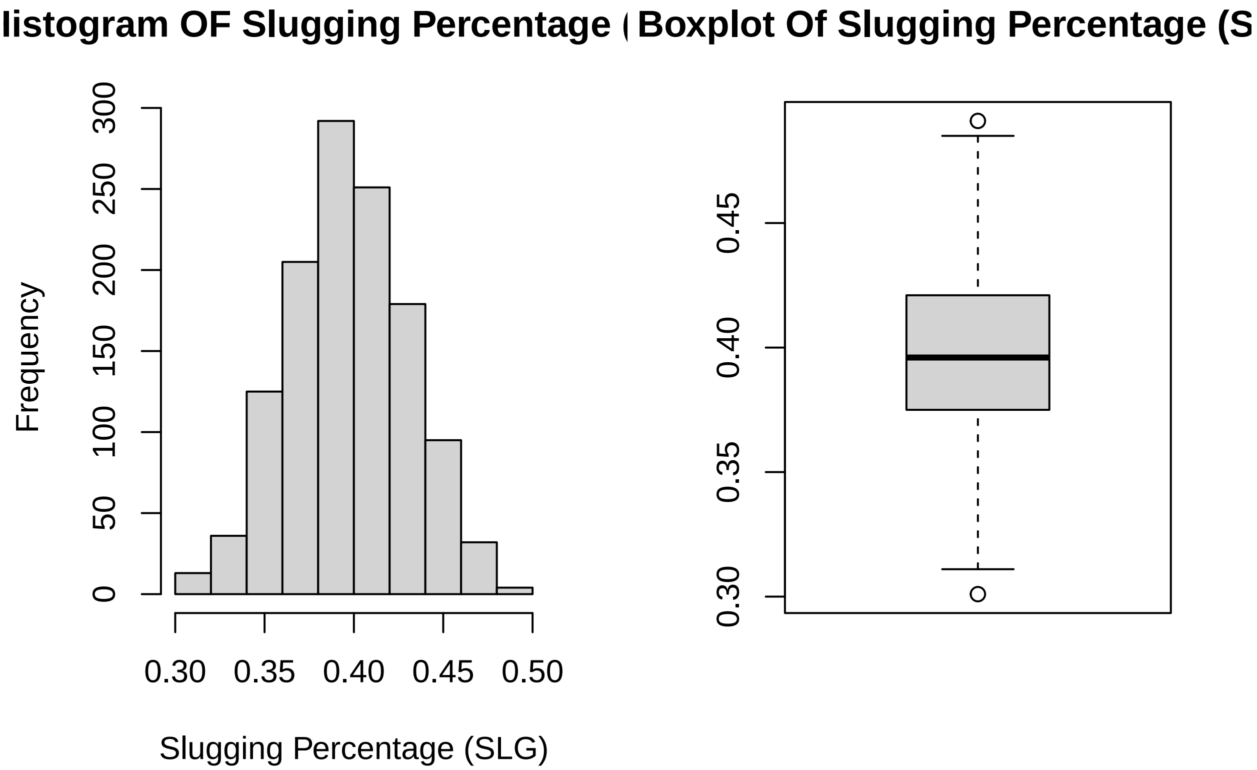
```
mean_OBP = mean(baseball$OBP) # Mean of OBP
cat("The mean of OBP:",mean_OBP,"\n")
```

```
## The mean of OBP: 0.3263312
```

```
median_OBP = median(baseball$OBP) # Median of OBP
cat("The median of OBP:",median_OBP,"\n")
```

```
## The median of OBP: 0.326
```

```
# Histogram and Boxplot of SLG
par(mfrow=c(1,2))
hist(baseball$SLG,main="Histogram OF Slugging Percentage (SLG)",xlab="Slugging Percentage (SLG)")
boxplot(baseball$SLG,main="Boxplot Of Slugging Percentage (SLG)")
```



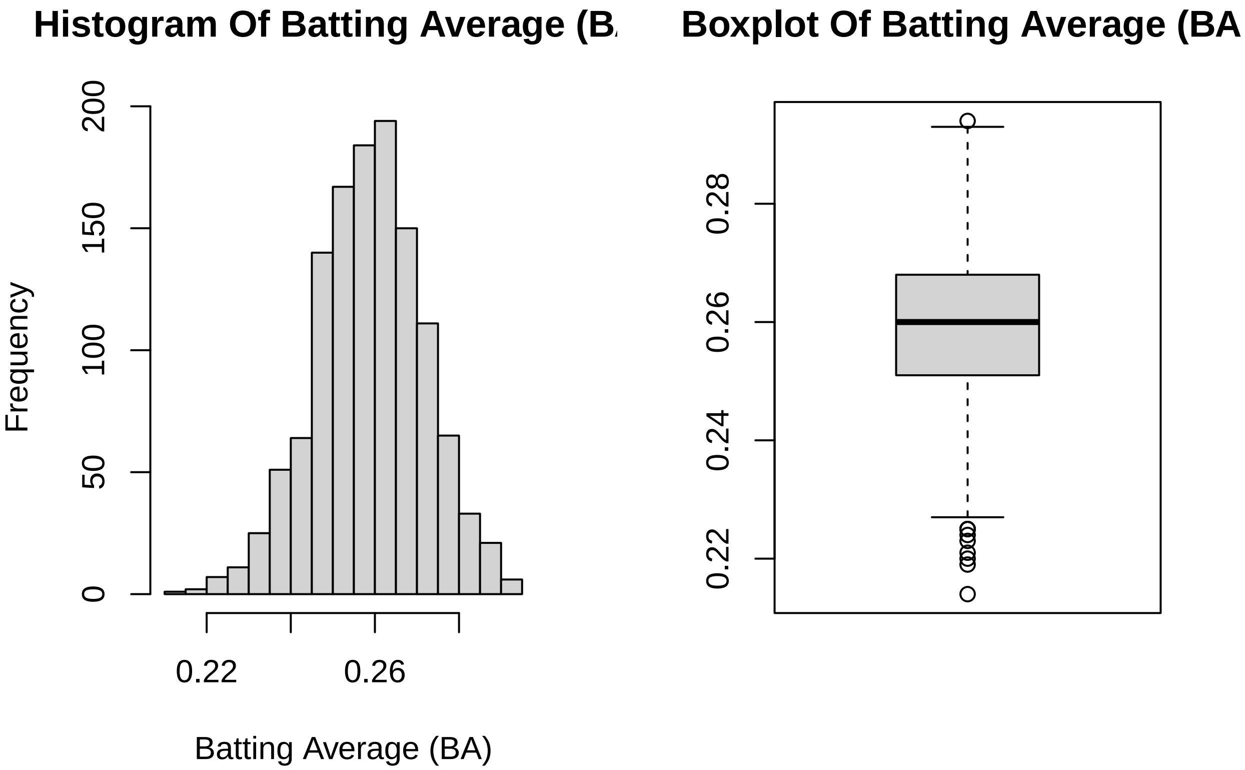
```
mean_SLG = mean(baseball$SLG) # Mean of SLG
cat("The mean of SLG:",mean_SLG,"\n")
```

```
## The mean of SLG: 0.3973417
```

```
median_SLG = median(baseball$SLG) # Median of SLG
cat("The median of SLG:",median_SLG,"\n")
```

```
## The median of SLG: 0.396
```

```
# Histogram and Boxplot of BA
par(mfrow=c(1,2))
hist(baseball$BA,main="Histogram Of Batting Average (BA)",xlab="Batting Average (BA)")
boxplot(baseball$BA,main="Boxplot Of Batting Average (BA)")
```



```
mean_BA = mean(baseball$BA) # Mean of BA
cat("The mean of BA:",mean_BA,"\n")
```

```
## The mean of BA: 0.2592727
```

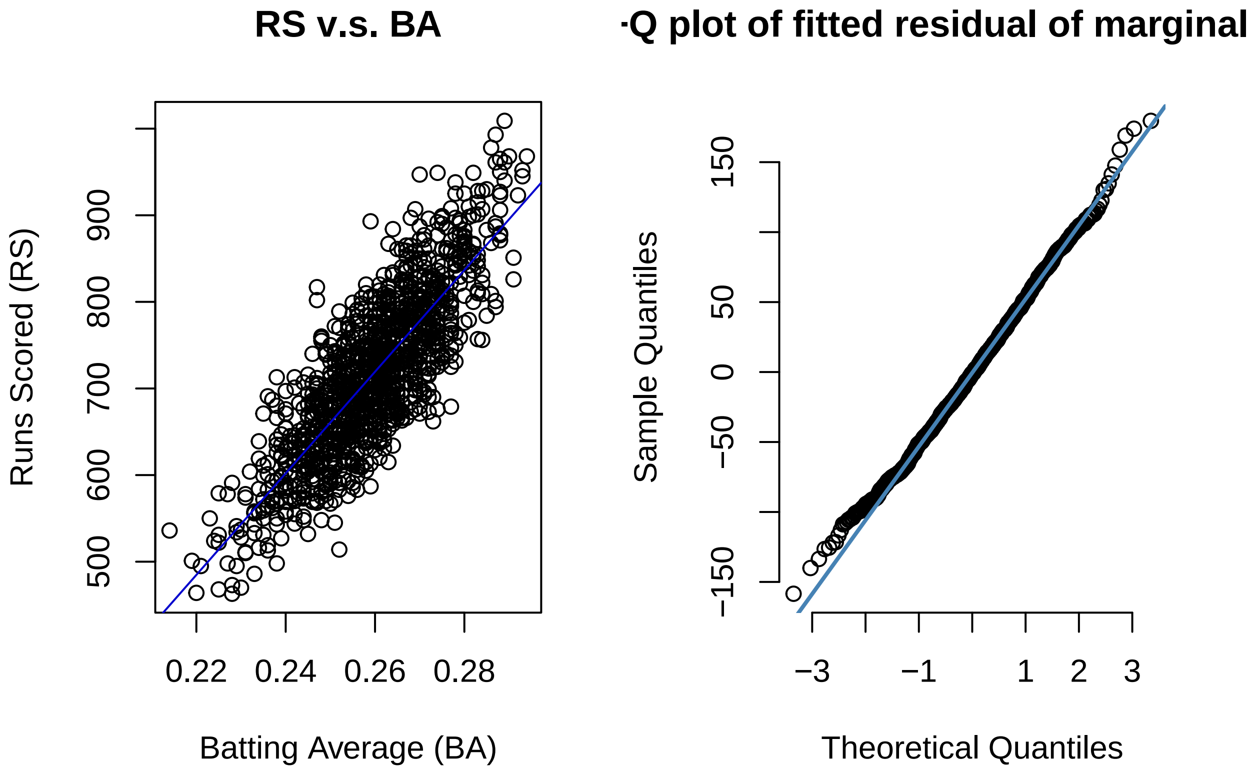
```
median_BA = median(baseball$BA) # Median of BA
cat("The median of BA:",median_BA,"\n")
```

```
## The median of BA: 0.26
```

The mean of OBP is 0.3263312 and the median of OBP is 0.326, meaning that the distribution of OBP is not skewed at all. The mean of SLG is 0.3973417 and the median of SLG is 0.396, meaning that the distribution of SLG a little bit skew.

Part II

```
model1 = lm(RS~BA,data=baseball) # Marginal model RS~BA
par(mfrow=c(1,2))
plot(baseball$BA, baseball$RS,main="RS v.s. BA",xlab="Batting Average (BA)",ylab="Runs Scored (RS)") # 
abline(summary(model1)$coef[1],summary(model1)$coef[2],col="blue3") # Fitted line
qqnorm(summary(model1)$residuals, pch = 1, frame = FALSE,main="Normal Q-Q plot of fitted residual of ma
qqline(summary(model1)$residuals, col = "steelblue", lwd = 2)
```



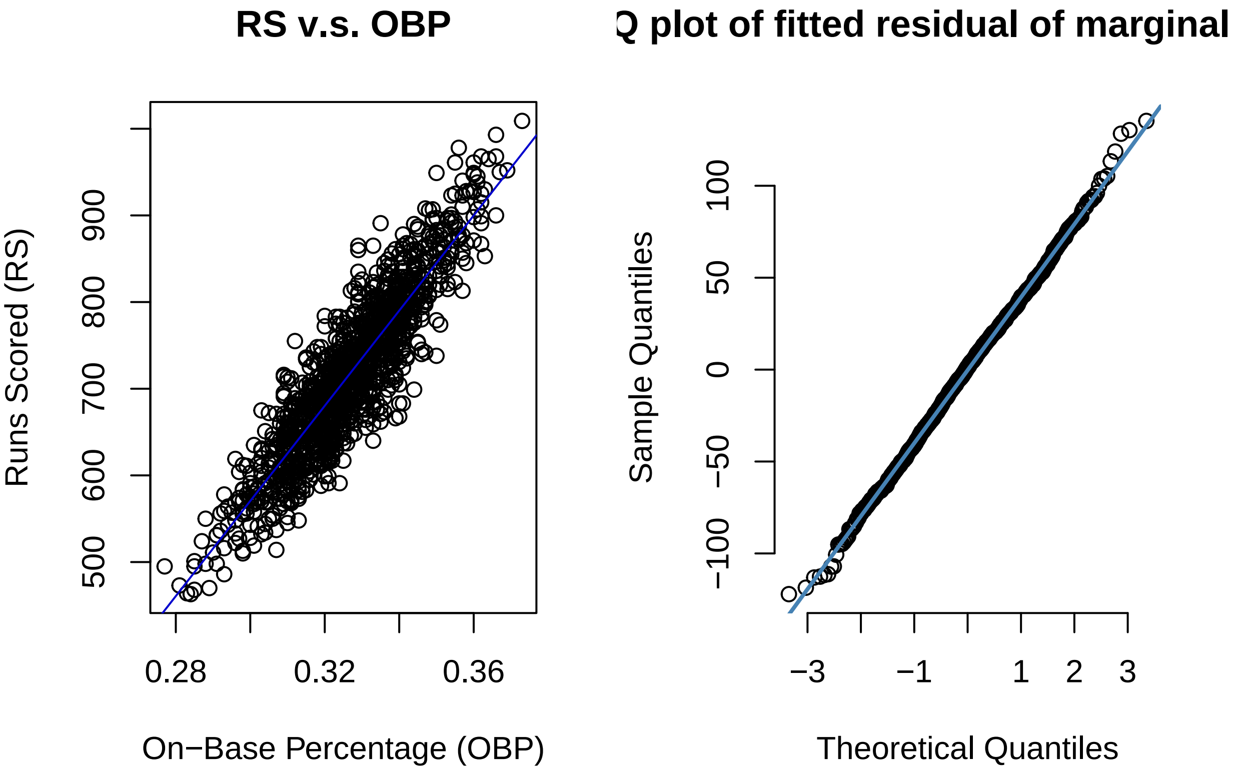
```
cat("Fitted coefficient of model RS~BA is",summary(model1)$coef[1],",", summary(model1)$coef[2],"\n")
```

```
## Fitted coefficient of model RS~BA is -805.511 , 5864.84
```

```
cat("R^2 of marginal model of RS~BA is",summary(model1)$r.squared,"\n")
```

```
## R^2 of marginal model of RS~BA is 0.6839284
```

```
model2 = lm(RS~OBP,data=baseball) # Marginal model RS~OBP
par(mfrow=c(1,2))
plot(baseball$OBP, baseball$RS,main="RS v.s. OBP",xlab="On-Base Percentage (OBP)",ylab="Runs Scored (RS)
abline(summary(model2)$coef[1],summary(model2)$coef[2],col="blue3") # Fitted line
qqnorm(summary(model2)$residuals, pch = 1, frame = FALSE,main="Normal Q-Q plot of fitted residual of mar
qqline(summary(model2)$residuals, col = "steelblue", lwd = 2)
```



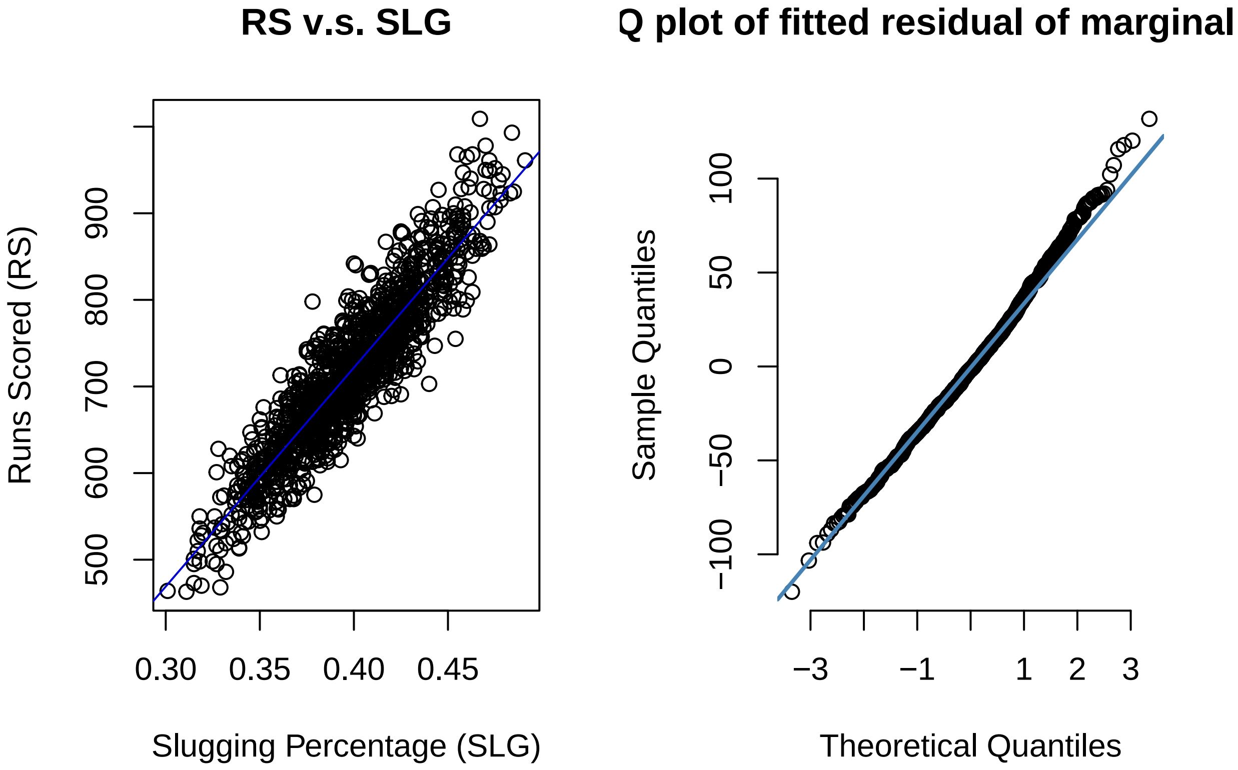
```
cat("Fitted coefficient of model RS~OBP is",summary(model2)$coef[1],",", summary(model2)$coef[2],"\n")
```

```
## Fitted coefficient of model RS~OBP is -1076.602 , 5490.386
```

```
cat("R^2 of marginal model of RS~OBP is",summary(model2)$r.squared,"\n")
```

```
## R^2 of marginal model of RS~OBP is 0.8108862
```

```
model3 = lm(RS~SLG,data=baseball) # Marginal model RS~SLG
par(mfrow=c(1,2))
plot(baseball$SLG, baseball$RS,main="RS v.s. SLG",xlab="Slugging Percentage (SLG)",ylab="Runs Scored (R
abline(summary(model3)$coef[1],summary(model3)$coef[2],col="blue3") # Fitted line
qqnorm(summary(model3)$residuals, pch = 1, frame = FALSE,main="Normal Q-Q plot of fitted residual of ma
qqline(summary(model3)$residuals, col = "steelblue", lwd = 2)
```



```
cat("Fitted coefficient of model RS~SLG is",summary(model3)$coef[1],",", summary(model3)$coef[2],"\n")
```

```
## Fitted coefficient of model RS~SLG is -289.368 , 2527.925
```

```
cat("R^2 of marginal model of RS~SLG is",summary(model3)$r.squared,"\n")
```

```
## R^2 of marginal model of RS~SLG is 0.8440831
```

R^2 of marginal model of RS~BA is 0.6839284 which is lower than R^2 of RS~OBP and R^2 of RS~SLG, so this contradict to the intuition that BA is thought to be most responsible for RS.

Part III

```
model4 = lm(RS~BA + SLG + OBP,data=baseball)
summary(model4)$coef
```

```
##              Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) -806.0845   17.39190 -46.348260 5.904672e-272
## BA          -134.9050  113.73431  -1.186141  2.357959e-01
## SLG         1533.8848   37.75868  40.623372 2.187242e-229
## OBP         2900.9403   97.87168  29.640243 4.386860e-146
```

```
cat("Since p-value of fitted coefficient of BA > 0.05, coefficient of BA is not significant.\n")
```

```
## Since p-value of fitted coefficient of BA > 0.05, coefficient of BA is not significant.
```

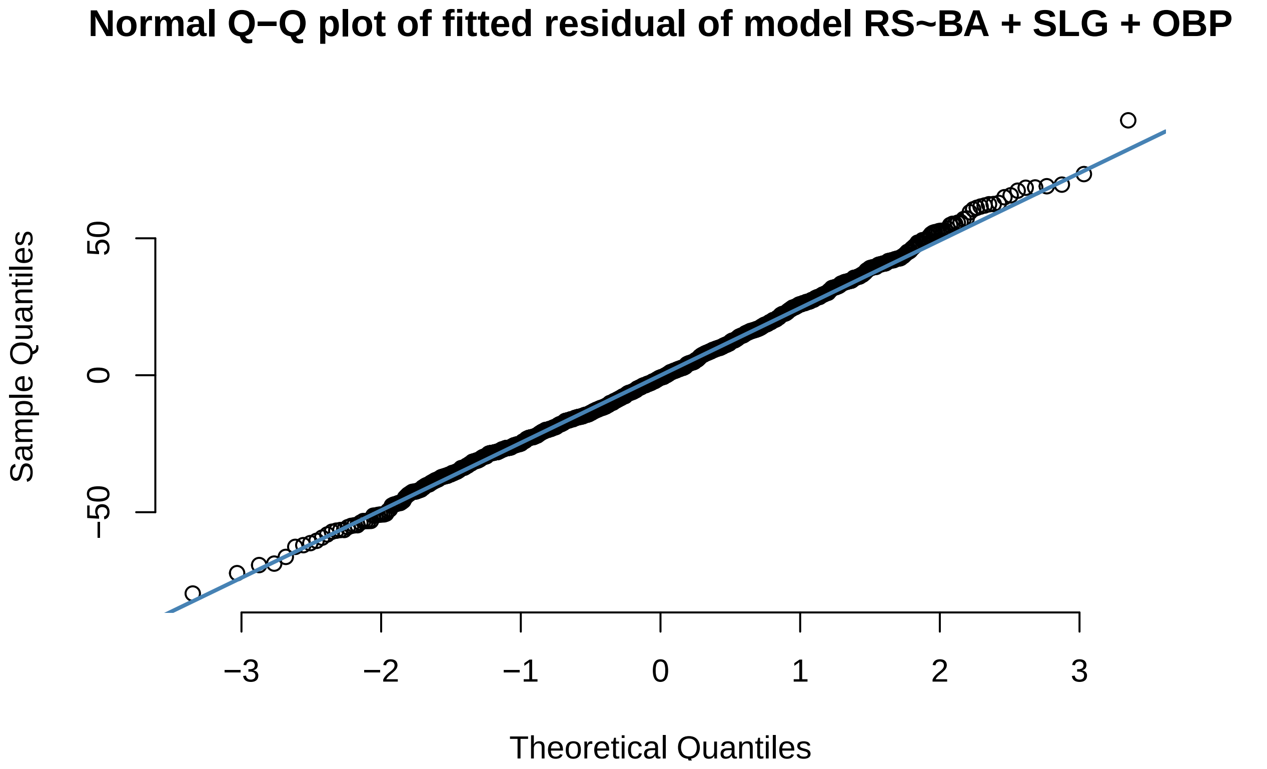
```
cat("All other fitted coefficients are significant.\n")
```

```
## All other fitted coefficients are significant.
```

```
cat("R^2 of model of RS~BA + SLG + OBP is",summary(model4)$r.squared,"\n")
```

```
## R^2 of model of RS~BA + SLG + OBP is 0.9248834
```

```
qqnorm(summary(model4)$residuals, pch = 1, frame = FALSE,main="Normal Q-Q plot of fitted residual of mo
qqline(summary(model4)$residuals, col = "steelblue", lwd = 2)
```



```
model5 = lm(RS~BA + SLG,data=baseball)
cat("R^2 of model of RS~BA + SLG is",summary(model5)$r.squared,"\n")
```

```
## R^2 of model of RS~BA + SLG is 0.871143
```

BA is not significant because its p-value = 2.357959e-01 < 0.05, this consist with the low R^2 in marginal model RS~BA in Part II. Also SLG and OBP are significant consist with the high R^2 in marginal model RS~SLG and marginal model RS~OBP in Part II. Model RS~BA + SLG + OBP is clearly better because it has a much higher R^2 than model of RS~BA + SLG.

Part IV

```
training_baseball = baseball[which(baseball$Year<2002),]
training_baseball["RD"] = training_baseball$RS-training_baseball$RA # Add column RD=RS-RA
model6 = lm(W ~ RD,data=training_baseball) # Model of W ~ RD
model7 = lm(RS~OBP + SLG,data=training_baseball) # Model of RS~OBP + SLG
model8 = lm(RA~OOBP + OSLG,data=training_baseball) # Model of RA~OOBP + OSLG

# Create dataframe for Xnew
OBP = 0.349
SLG = 0.430
OOBP = 0.307
OSLG = 0.373
Xnew = data.frame(OBP,SLG,OOBP,OSLG)
Xnew
```

```
##     OBP  SLG  OOBP  OSLG
## 1 0.349 0.43 0.307 0.373
```

```
RA_pred = predict(model8,Xnew) # Predicted RA
RA_pred
```

```
##        1
## 621.9258
```

```
RS_pred = predict(model7,Xnew) # Predicted RS
RS_pred
```

```
##        1
## 832.3647
```

```
RD = RS_pred-RA_pred # Predicted RD
W_pred = predict(model6,data.frame(RD)) # Predicted W
cat("Oakland Athletics is predicted to have",W_pred,"win games in 2002.")
```

```
## Oakland Athletics is predicted to have 103.1386 win games in 2002.
```

```
cat("The true number of win games for Oakland Athletics in 2002 is",baseball[which(baseball$Year==2002&
```

```
## The true number of win games for Oakland Athletics in 2002 is 103
```

```
cat(", which is the same as our prediction after round our prediction to integer.","\n")
```

```
## , which is the same as our prediction after round our prediction to integer.
```