7002 A1
CHENGYANG ZHOU
3036167854

Q1:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$$Y_2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \varepsilon$$

$$\text{when } X_1 = 1 \quad \begin{cases} Y_1 = \beta_0 + \beta_1 + (\beta_2 + \beta_3) X_2 + \varepsilon \\ Y_2 = \alpha_0 + \alpha_1 + (\alpha_2 + \alpha_3) X_2 + \varepsilon \end{cases}$$

$$\text{when } X_1 = \begin{cases} 0 \\ -1 \end{cases} \quad \begin{cases} Y_1 = \beta_0 + \beta_2 X_2 + \varepsilon \\ Y_2 = \alpha_0 - \alpha_1 + (\alpha_2 - \alpha_3) X_2 + \varepsilon \end{cases}$$

$$\begin{cases} \beta_0 + \beta_1 = \alpha_0 + \alpha_1 \\ \beta_2 + \beta_3 = \alpha_2 + \alpha_3 \\ \beta_0 = \alpha_0 - \alpha_1 \\ \beta_2 = \alpha_2 - \alpha_3 \end{cases} \Longrightarrow \begin{cases} \beta_0 = \alpha_0 - \alpha_1 \\ \beta_1 = 2\alpha_1 \\ \beta_2 = \alpha_2 - \alpha_3 \\ \beta_3 = 2\alpha_3 \end{cases}$$

Q2:

```
## Q2 - Production Time Run

ProdTime.dat contains information about 20  production  runs  supervised  by  each  of  three managers.
Each  observation gives the time  (in minutes)  to  complete  the task, Time  for Run,
as  well  as  the number  of  units produced, Run Size, and the manager involved, Manager.

Which manager performs  the best?
```

```{r}
q2 = read.csv("ProdTime.dat")
q2
```

Description: df [61 × 3]

Time.for.Run <int>	Manager <chr>	Run.Size <int>
252	a	204
215	a	103
238	a	143
261	a	210
297	a	334
236	a	102
282	a	261
264	a	118
254	a	198
223	a	87

1-10 of 61 rows   Previous 1 2 3 4 5 6 7 Next

```{r}
dim(q2)
names(q2)
str(q2)
```

```
[1] 61  3
[1] "Time.for.Run" "Manager"      "Run.Size"
'data.frame':	61 obs. of  3 variables:
 $ Time.for.Run: int  252 215 238 261 297 236 282 264 254 223 ...
 $ Manager     : chr  "a" "a" "a" "a" ...
 $ Run.Size    : int  204 103 143 210 334 102 261 118 198 87 ...
```

```{r}
q2$Efficiency <- q2$Time.for.Run / q2$Run.Size
q2
```

Description: df [61 × 4]

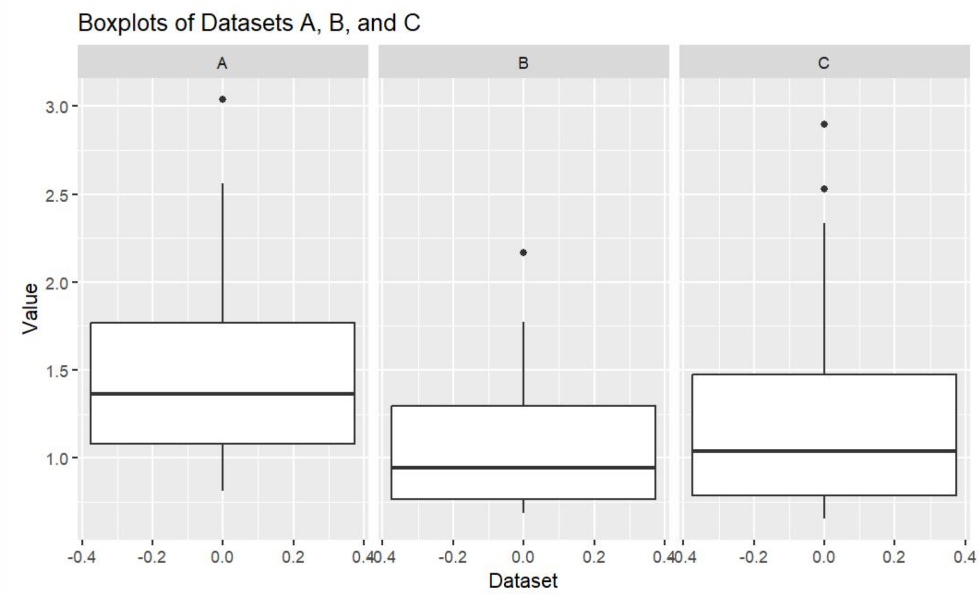
Time.for.Run <int>	Manager <chr>	Run.Size <int>	Efficiency <dbl>
252	a	204	1.2352941
215	a	103	2.0873786
238	a	143	1.6643357
261	a	210	1.2428571
297	a	334	0.8892216
236	a	102	2.3137255
282	a	261	1.0804598
264	a	118	2.2372881
254	a	198	1.2828283
223	a	87	2.5632184

1-10 of 61 rows   Previous 1 2 3 4 5 6 7 Next

```{r}
A <- q2[q2$Manager == 'a', "Efficiency"]
B <- q2[q2$Manager == 'b', "Efficiency"]
C <- q2[q2$Manager == 'c', "Efficiency"]
summary(A)
summary(B)
summary(C)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.8145  1.0800  1.3618  1.5545  1.7701  3.0400 
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.6869  0.7638  0.9413  1.0630  1.2931  2.1667 
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.6528  0.7861  1.0369  1.2799  1.4742  2.8966 
```

```{r}
df_A <- data.frame(Value = A)
df_B <- data.frame(Value = B)
df_C <- data.frame(Value = C)

combined_df <- rbind(df_A, df_B, df_C)
combined_df$Dataset <- rep(c("A", "B", "C"), each = nrow(df_A))

ggplot(combined_df, aes(y = Value)) +
  geom_boxplot() +
  facet_wrap(~ Dataset, nrow = 1) +
  xlab("Dataset") + ylab("Value") +
  ggtitle("Boxplots of Datasets A, B, and C")
```




```{r}
t.test(B, A , alternative = "less")
t.test(B, C , alternative = "less")
```

```
        Welch Two Sample t-test

data:  B and A
t = -3.0511, df = 32.651, p-value = 0.002252
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.218792
sample estimates:
mean of x mean of y
 1.062995  1.554498


        Welch Two Sample t-test

data:  B and C
t = -1.2594, df = 30.907, p-value = 0.1087
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.07511878
sample estimates:
mean of x mean of y
 1.062995  1.279865
```

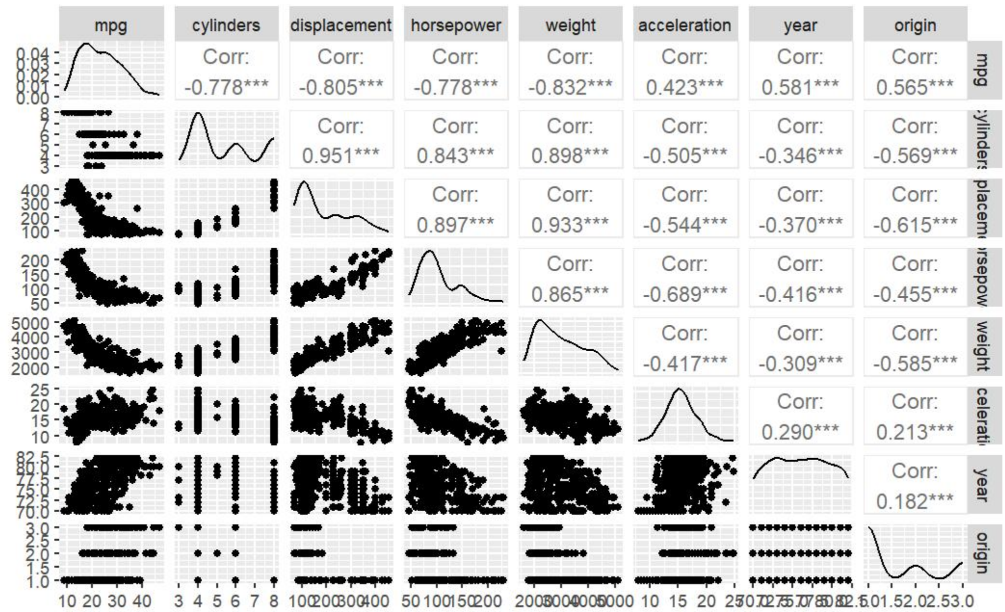
As a result:
According to the above figure and t-test, B Manager performs better than A Manger.
But B Manager and C Manager do not have significant difference.

Q3:

Q3.1:

```
```{r, eval = F}
# check if you have ISLR package, if not, install it
if(!requireNamespace('ISLR')) install.packages('ISLR')
auto_data <- ISLR::Auto
```
```

```
Q3.1
Explore the data, with particular focus on pairwise plots and summary statistics.
Briefly summarize your findings and any peculiarities in the data.
```

```
```{r}
dim(auto_data)
names(auto_data)
str(auto_data)
sum(is.na(auto_data))
summary(auto_data)
```
```

```
[1] 392   9
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"       "acceleration" "year"
[8] "origin"       "name"
'data.frame':	392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : num  3504 3693 3436 3433 3449 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2 ...
[1] 0
      mpg          cylinders      displacement     horsepower        weight      acceleration        year
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50   Median :76.00
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00

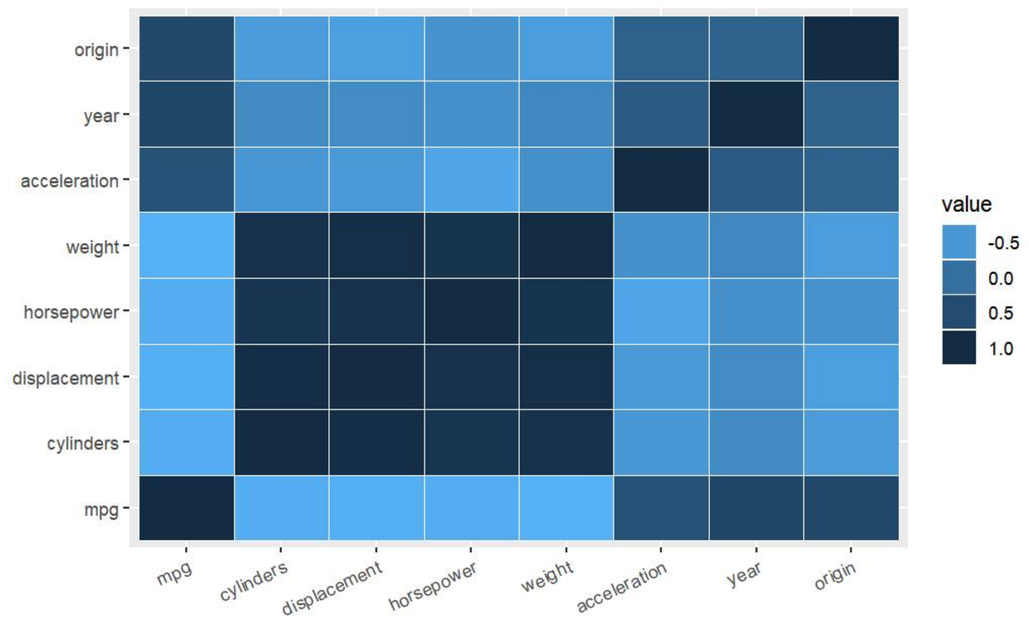
     origin                      name
 Min.   :1.000   amc matador       :  5
 1st Qu.:1.000   ford pinto        :  5
 Median :1.000   toyota corolla    :  5
 Mean   :1.577   amc gremlin       :  4
 3rd Qu.:2.000   amc hornet        :  4
 Max.   :3.000   chevrolet chevette:  4
                 (Other)           :365
```

```
```{r}
auto_data %>%
  select_if(is.numeric) %>%
  ggpairs(progress =FALSE)
```
```



```{r}
auto_data %>%
  select_if(is.numeric)  %>%
  cor()
```

```
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

```{r}
ggplot(data = auto_data %>% select_if(is.numeric) %>% cor() %>% reshape2::melt(),
aes(x = Var1 ,y = Var2, fill = value)) +
geom_tile(color="white",size=0.1) +
xlab("") +
ylab("") +
guides(fill = guide_legend(data.crdle = "Correlation")) +
scale_fill_gradient( low = "#56B1F7", high = "#132B43") +
theme(axis.text.x = element_text(angle = 25, hjust = 1))
```



Q3.2:

i:

```{r}
Q3_2_1 <- lm(mpg~year,auto_data)
summary(Q3_2_1)
```

```
Call:
lm(formula = mpg ~ year, data = auto_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0212  -5.4411  -0.4412   4.9739  18.2088 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -70.01167    6.64516  -10.54   <2e-16 ***
year          1.23004    0.08736   14.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.363 on 390 degrees of freedom
Multiple R-squared:  0.337,	Adjusted R-squared:  0.3353 
F-statistic: 198.3 on 1 and 390 DF,  p-value: < 2.2e-16
```

p-value < 0.05, year is a significant variable at the .05 level.
When year increases 1 unit, mpg will increase 1.230 units.

ii:

```{r}
Q3_2_2 <- lm(mpg~year+horsepower,auto_data)
summary(Q3_2_2)
```

```
Call:
lm(formula = mpg ~ year + horsepower, data = auto_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0768  -3.0783  -0.4308   2.5884  15.3153 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.739166   5.349027  -2.382   0.0177 *  
year          0.657268   0.066262   9.919   <2e-16 ***
horsepower   -0.131654   0.006341 -20.761   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.388 on 389 degrees of freedom
Multiple R-squared:  0.6855,	Adjusted R-squared:  0.6839 
F-statistic: 423.9 on 2 and 389 DF,  p-value: < 2.2e-16
```

p-value < 0.05, year is a significant variable at the .05 level.
When year increases 1 unit, mpg will increase 0.657 units.

iii:

When adding another parameter to the original model, it will affect the significance of the other original parameters on the results.

iv:

```{r}
Q3_2_4 <- lm(mpg~year*horsepower,auto_data)
summary(Q3_2_4)
```

```
Call:
lm(formula = mpg ~ year * horsepower, data = auto_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.3492  -2.4509  -0.4557   2.4056  14.4437 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.266e+02  1.212e+01 -10.449   <2e-16 ***
year             2.192e+00  1.613e-01  13.585   <2e-16 ***
horsepower       1.046e+00  1.154e-01   9.063   <2e-16 ***
year:horsepower -1.596e-02  1.562e-03 -10.217   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.901 on 388 degrees of freedom
Multiple R-squared:  0.7522,    Adjusted R-squared:  0.7503 
F-statistic: 392.5 on 3 and 388 DF,  p-value: < 2.2e-16
```

p-value < 0.05, the interaction effect is a significant variable at the .05 level.

Q3.3:

i:

```{r}
Q3_3_1 <- lm(mpg ~ horsepower + cylinders, ISLR::Auto)
summary(Q3_3_1)
```

```
Call:
lm(formula = mpg ~ horsepower + cylinders, data = ISLR::Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.4378  -3.2422  -0.3721   2.3532  16.9289 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.94842    0.77880  55.147  < 2e-16 ***
horsepower  -0.08612    0.01119  -7.693 1.19e-13 ***
cylinders   -1.91982    0.25261  -7.600 2.24e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.584 on 389 degrees of freedom
Multiple R-squared:  0.6569,	Adjusted R-squared:  0.6551 
F-statistic: 372.4 on 2 and 389 DF,  p-value: < 2.2e-16
```

```
p-value < 0.01, the cylinders is a significant variable at the 0.01 level.
When cylinders increases 1 unit, mpg will decrease 1.920 units.
But we know that cylinders and mpg are positively related, so the model has some problems.
```

ii:

```{r}
Q3_3_2 <-lm(mpg ~ horsepower + as.factor(cylinders), ISLR::Auto)
summary(Q3_3_2)
```

```
Call:
lm(formula = mpg ~ horsepower + as.factor(cylinders), data = ISLR::Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5917 -2.7067 -0.6102  1.9001 16.3258 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           30.77614    2.41283  12.755  < 2e-16 ***
horsepower            -0.10303    0.01133  -9.095  < 2e-16 ***
as.factor(cylinders)4  6.57344    2.16921   3.030  0.00261 ** 
as.factor(cylinders)5  5.07367    3.26661   1.553  0.12120    
as.factor(cylinders)6 -0.34406    2.18580  -0.157  0.87501    
as.factor(cylinders)8  0.49738    2.27639   0.218  0.82716    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.27 on 386 degrees of freedom
Multiple R-squared:  0.7046,	Adjusted R-squared:  0.7008 
F-statistic: 184.1 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
Only for category 4, p-value < 0.01, the cylinders is a significant variable at the 0.01 level.
For category 5, 6, 8, the cylinders is not a significant variable at the 0.01 level.
Cylinder 4 cars have average 6.573 units higher mpg than cylinder 3 cars do.
```

iii:

```{r}
anova(Q3_3_1, Q3_3_2)
```

```
Analysis of Variance Table

Model 1: mpg ~ horsepower + cylinders
Model 2: mpg ~ horsepower + as.factor(cylinders)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    389 8172.5                                  
2    386 7036.7  3    1135.8 20.769 1.705e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on this ANOVA table, it seems that treating cylinders as a factor variable provides a significantly better model fit compared to treating it as a numeric variable.

The fundamental difference between treating "cylinders" as a numeric or a factor variable lies in the interpretation and representation of the variable.

The choice between treating "cylinders" as numeric or factor depends on the nature of the variable and the specific requirements of your analysis.
If the number of cylinders has inherent numerical meaning, treating it as numeric is appropriate.
However, if the cylinder counts represent distinct categories without an ordering pattern, treating it as a factor variable is more appropriate.

Q4:

```{r}
crime <- read.csv("CrimeData_sub.csv", stringsAsFactors = F, na.strings = c("?"))
crime <- na.omit(crime)
crime
```

Description: df [368 × 103]

	state <chr>	fold <int>	population <dbl>	household.size <dbl>	race.pctblack <dbl>	race.pctwhite <dbl>	race.pctasian <dbl>	race.pcthisp <dbl>
1	CA	1	10.132971	2.89	21.34	49.42	17.21	26.78
2	CA	1	12.100929	2.62	1.30	74.02	14.14	20.96
3	CA	1	10.182520	3.34	18.97	53.60	20.84	10.73
4	FL	1	9.681843	2.63	13.79	83.94	1.42	2.40
5	CA	1	11.392948	2.70	2.92	87.36	2.82	13.97
6	CA	1	9.389490	2.71	0.24	88.57	1.11	17.04
7	CA	1	10.056337	4.17	7.95	82.42	1.36	19.55
8	CA	1	10.172293	2.37	0.94	91.30	6.07	6.46
9	CA	1	10.050657	2.09	0.70	95.77	1.74	6.86
10	CA	1	10.372616	2.20	1.70	91.28	5.46	5.40

1-10 of 368 rows | 1-9 of 103 columns Previous 1 2 3 4 5 6 ... 37 Next

```{r}
dim(crime)
names(crime)
str(crime)
sum(is.na(crime))
summary(crime)
```

```
[1] 368 103
  [1] "state"                         "fold"                         "population"
  [4] "household.size"                "race.pctblack"                "race.pctwhite"
  [7] "race.pctasian"                 "race.pcthisp"                 "age.pct12to21"
 [10] "age.pct12to29"                 "age.pct16to24"                "age.pct65up"
 [13] "num.urban"                     "pct.urban"                    "med.income"
 [16] "pct.wage.inc"                  "pct.farmself.inc"             "pct.inv.inc"
 [19] "pct.socsec.inc"                "pct.pubasst.inc"              "pct.retire"
 [22] "med.family.inc"                "percap.inc"                   "white.percap"
 [25] "black.percap"                  "indian.percap"                "asian.percap"
 [28] "other.percap"                  "hisp.percap"                  "num.underpov"
 [31] "pct.pop.underpov"              "pct.less9thgrade"             "pct.not.hsgrad"
 [34] "pct.bs.ormore"                 "pct.unemployed"               "pct.employed"
 [37] "pct.employed.manuf"            "pct.employed.profserv"        "pct.occup.manuf"
 [40] "pct.occup.mgmtprof"            "male.pct.divorce"             "male.pct.nvrmarried"
 [43] "female.pct.divorce"            "total.pct.divorce"            "ave.people.per.fam"
 [46] "pct.fam2parents"               "pct.kids2parents"             "pct.youngkids2parents"
 [49] "pct.teens2parents"             "pct.workmom.youngkids"        "pct.workmom"
 [52] "num.kids.nvrmarried"           "pct.kids.nvrmarried"          "num.immig"
 [55] "pct.immig.recent"              "pct.immig.recent5"            "pct.immig.recent8"
 [58] "pct.immig.recent10"            "pct.pop.immig"                "pct.pop.immig5"
 [61] "pct.pop.immig8"                "pct.pop.immig10"              "pct.english.only"
 [64] "pct.no.english.well"           "pct.fam.hh.large"             "pct.occup.hh.large"
 [67] "ave.people.per.hh"             "ave.people.per.ownoccup.hh"   "ave.people.per.rented.hh"
 [70] "pct.people.ownoccup.hh"        "pct.people.dense.hh"          "pct.hh.less3br"
 [73] "med.num.br"                    "num.vacant.house"             "pct.house.occup"
 [76] "pct.house.ownoccup"            "pct.house.vacant"             "pct.house.vacant.6moplus"
 [79] "med.yr.house.built"            "pct.house.nophone"            "pct.house.no.plumb"
 [82] "value.ownoccup.house.lowquart" "value.ownoccup.med"           "value.ownoccup.highquart"
 [85] "rent.lowquart"                 "rent.med"                     "rent.highquart"
 [88] "med.rent"                      "med.rent.aspct.hhinc"         "med.owncost.aspct.hhinc.wmort"
 [91] "med.owncost.as.pct.hhinc.womort" "num.in.shelters"            "num.homeless"
 [94] "pct.foreignborn"               "pct.born.samestate"           "pct.samehouse1985"
 [97] "pct.samecity1985"              "pct.samestate1985"            "land.area"
[100] "pop.density"                   "pct.use.publictransit"        "pct.police.drugunits"
[103] "violentcrimes.perpop"
'data.frame':	368 obs. of  103 variables:
 $ state                          : chr  "CA" "CA" "CA" "FL" ...
 $ fold                           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ population                     : num  10.13 12.1 10.18 9.68 11.39 ...
 $ household.size                 : num  2.89 2.62 3.34 2.63 2.7 2.71 4.17 2.37 2.09 2.2 ...
 $ race.pctblack                  : num  21.34 1.3 18.97 13.79 2.92 ...
```

```{r}
crime[crime==0] <- NA
sapply(crime, function(x) sum(is.na(x)))
```

```
         state            fold      population
             0               0               0
household.size   race.pctblack   race.pctwhite
             0               0               0
 race.pctasian    race.pcthisp   age.pct12to21
             0               0               0
 age.pct12to29   age.pct16to24     age.pct65up
             0               0               0
     num.urban       pct.urban      med.income
            51              51               0
  pct.wage.inc pct.farmself.inc    pct.inv.inc
             0               3               0
pct.socsec.inc pct.pubasst.inc      pct.retire
             0               0               0
med.family.inc      percap.inc    white.percap
             0               0               0
  black.percap   indian.percap    asian.percap
```

```
                          0                            0                              0
med.owncost.as.pct.hhinc.womort               num.in.shelters                 num.homeless
                          0                          183                            184
             pct.foreignborn             pct.born.samestate             pct.samehouse1985
                          0                            0                              0
            pct.samecity1985              pct.samestate1985                     land.area
                          0                            0                              0
                 pop.density           pct.use.publictransit            pct.police.drugunits
                          0                           20                            294
        violentcrimes.perpop
                          0
```

```{r}
crime$pct.police.drugunits <- NULL
crime$num.homeless <- NULL
crime$num.in.shelters <- NULL
crime$pct.urban <- NULL
crime$num.urban <- NULL
crime[is.na(crime)] <- 0
```

Q4.1:

```{r}
set.seed(123)
split<- sample(c(rep(0, 0.8 * nrow(crime)), rep(1, 0.2 * nrow(crime))))
train_data <- crime[split == 0, ]
test_data <- crime[split == 1, ]
```

```{r}
model_train_q4_1 <- lm(train_data$violentcrimes.perpop~., data= train_data)
summary(model_train_q4_1)
fitrain <- summary(model_train_q4_1)
```

```
Call:
lm(formula = train_data$violentcrimes.perpop ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-892.66 -156.81  -11.97  146.13 1023.96

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.685e+03  1.515e+04   0.441  0.65945
stateFL                    2.123e+02  2.552e+02   0.832  0.40638
fold                      -7.305e+00  8.056e+00  -0.907  0.36561
population                -5.805e+02  3.487e+02  -1.665  0.09760 .
household.size            -2.354e+02  4.531e+02  -0.519  0.60404
race.pctblack              2.960e+01  9.642e+00   3.070  0.00244 **
race.pctwhite              3.052e+00  7.453e+00   0.409  0.68266
race.pctasian             -6.340e+00  1.253e+01  -0.506  0.61344
race.pcthisp              -1.133e+01  9.620e+00  -1.178  0.24033
age.pct12to21              5.403e+01  5.209e+01   1.037  0.30090
age.pct12to29             -3.943e+01  3.910e+01  -1.008  0.31451
age.pct16to24              3.681e+00  6.153e+01   0.060  0.95236
age.pct65up                5.933e+01  2.881e+01   2.059  0.04078 *
med.income                -1.480e-02  2.012e-02  -0.736  0.46290
pct.wage.inc               1.943e+01  2.000e+01   0.972  0.33235
```

```{r}
model_test_q4_1 = predict(model_train_q4_1,test_data)
summary(model_test_q4_1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -424.9   441.1   824.6   888.7  1128.4  2843.8 
```

```{r}
cat("Training RMSE:", mean(fitrain$residuals^2), "\n")
cat("Training R-squared:", mean(fitrain$residuals^2), "\n")
cat("Testing RMSE:", RMSE(model_test_q4_1, test_data$violentcrimes.perpop), "\n")
cat("Testing R-squared:", R2(model_test_q4_1, test_data$violentcrimes.perpop), "\n")
```

```
Training RMSE: 70752.12 
Training R-squared: 70752.12 
Testing RMSE: 463.0018 
Testing R-squared: 0.6056186 
```

Q4.2:

i:

```{r}
X.train <- model.matrix(violentcrimes.perpop~., data=train_data)
Y.train <- train_data$violentcrimes.perpop
```

```{r}
matrix.crimes <- data.frame(CRIMES = Y.train, X.train)
```

```{r}
set.seed(123)
fit.lasso.cv <- cv.glmnet(X.train, Y.train, alpha=1, nfolds=5,
                    lambda = 10^seq(-3,0,length=100))
names(fit.lasso.cv)
```

```
 [1] "lambda"     "cvm"        "cvsd"       "cvup"       "cvlo"       "nzero"      "call"       "name"       "glmnet.fit"
[10] "lambda.min" "lambda.1se" "index"
```

```{r}
coef.min <- coef(fit.lasso.cv, s="lambda.min")
coef.min <- coef.min[which(coef.min !=0),]
var.min <- rownames(as.matrix(coef.min))
var.min[2]='state'
lm.input <- as.formula(paste("violentcrimes.perpop", "~", paste(var.min[-1], collapse = "+")))
lm.input
```

```
violentcrimes.perpop ~ state + fold + population + household.size + 
    race.pctblack + race.pctwhite + race.pcthisp + age.pct12to21 + 
    age.pct12to29 + age.pct65up + med.income + pct.wage.inc + 
    pct.inv.inc + pct.pubasst.inc + pct.retire + white.percap + 
    black.percap + indian.percap + asian.percap + other.percap + 
    hisp.percap + pct.pop.underpov + pct.less9thgrade + pct.not.hsgrad + 
    pct.bs.ormore + pct.unemployed + pct.employed + pct.employed.manuf + 
    pct.occup.manuf + male.pct.divorce + male.pct.nvrmarried + 
    female.pct.divorce + ave.people.per.fam + pct.fam2parents + 
    pct.kids2parents + pct.teens2parents + pct.workmom + pct.kids.nvrmarried + 
    num.immig + pct.immig.recent + pct.immig.recent8 + pct.pop.immig8 + 
    pct.english.only + pct.no.english.well + ave.people.per.rented.hh + 
    pct.people.ownoccup.hh + pct.people.dense.hh + pct.hh.less3br + 
    med.num.br + num.vacant.house + pct.house.occup + pct.house.vacant + 
    pct.house.vacant.6moplus + med.yr.house.built + pct.house.nophone + 
    pct.house.no.plumb + value.ownoccup.house.lowquart + value.ownoccup.med + 
    value.ownoccup.highquart + rent.lowquart + rent.highquart + 
    med.rent + med.rent.aspct.hhinc + med.owncost.aspct.hhinc.wmort + 
    med.owncost.as.pct.hhinc.womort + pct.foreignborn + pct.born.samestate + 
    pct.samehouse1985 + pct.samecity1985 + pct.samestate1985 + 
    land.area + pop.density + pct.use.publictransit
```

ii:

````
```{r}
fit.min.lm <- lm(lm.input, data=train_data)
lm.output <- coef(fit.min.lm)
model_q4_2_2 = summary(fit.min.lm)
```
````

````
```{r}
test_q4_2_2 = predict(fit.min.lm,test_data)
```
````

````
```{r}
cat("Training RMSE:", mean(model_q4_2_2$residuals^2), "\n")
cat("Training R-squared:", mean(model_q4_2_2$residuals^2), "\n")
cat("Testing RMSE:", RMSE(test_q4_2_2, test_data$violentcrimes.perpop), "\n")
cat("Testing R-squared:", R2(test_q4_2_2, test_data$violentcrimes.perpop), "\n")
```
````

```
Training RMSE: 74413.41
Training R-squared: 74413.41
Testing RMSE: 442.6142
Testing R-squared: 0.6268232
```

iii:

````
```{r}
my_model = as.data.frame(summary(fit.min.lm)$coefficients)
var = rownames(my_model[my_model$'Pr(>|t|)'<0.05,])
lm.input <- as.formula(paste("violentcrimes.perpop", "~", paste(var, collapse = "+")))
lm.input
```
````

```
violentcrimes.perpop ~ race.pctblack + age.pct12to21 + pct.pubasst.inc +
    indian.percap + pct.less9thgrade + pct.not.hsgrad + male.pct.nvrmarried +
    pct.kids2parents + pct.pop.immig8 + pct.no.english.well +
    pct.people.ownoccup.hh + pct.house.nophone + med.owncost.as.pct.hhinc.womort +
    land.area
```

iv:

````
```{r}
set.seed(123)
fit.ridge.cv <- cv.glmnet(X.train, Y.train, alpha = 0, nfolds = 5,
                        lambda = 10^seq(-3,0,length=100))
names(fit.ridge.cv)
```
````

```
 [1] "lambda"     "cvm"        "cvsd"       "cvup"       "cvlo"       "nzero"      "call"       "name"       "glmnet.fit"
[10] "lambda.min" "lambda.1se" "index"
```

````
```{r}
coef.min <- coef(fit.ridge.cv, s="lambda.min")
coef.min <- coef.min[which(coef.min !=0),]
var.min <- rownames(as.matrix(coef.min))
var.min[2]='state'
lm.input <- as.formula(paste("violentcrimes.perpop", "~", paste(var.min[-1], collapse = "+")))
lm.input
```
````

```
violentcrimes.perpop ~ state + fold + population + household.size +
    race.pctblack + race.pctwhite + race.pctasian + race.pcthisp +
    age.pct12to21 + age.pct12to29 + age.pct16to24 + age.pct65up +
    med.income + pct.wage.inc + pct.farmself.inc + pct.inv.inc +
    pct.socsec.inc + pct.pubasst.inc + pct.retire + med.family.inc +
    percap.inc + white.percap + black.percap + indian.percap +
    asian.percap + other.percap + hisp.percap + num.underpov +
    pct.pop.underpov + pct.less9thgrade + pct.not.hsgrad + pct.bs.ormore +
    pct.unemployed + pct.employed + pct.employed.manuf + pct.employed.profserv +
    pct.occup.manuf + pct.occup.mgmtprof + male.pct.divorce +
```

```{r}
fit.min.lm <- lm(lm.input, data=train_data)
lm.output <- coef(fit.min.lm)
model_q4_2_4 = summary(fit.min.lm)
```

```{r}
test_q4_2_4 = predict(fit.min.lm, test_data)
```

```{r}
cat("Training RMSE:", mean(model_q4_2_4$residuals^2), "\n")
cat("Training R-squared:", mean(model_q4_2_4$r.squared), "\n")
cat("Testing RMSE:", RMSE(test_q4_2_4, test_data$violentcrimes.perpop), "\n")
cat("Testing R-squared:", R2(test_q4_2_4, test_data$violentcrimes.perpop), "\n")
```

```
Training RMSE: 70752.12
Training R-squared: 0.8340532
Testing RMSE: 463.0018
Testing R-squared: 0.6056186
```