Project_1_Econ_104

James Freedman, Jack Posner, Michael Tam, Byeonghyeok Yoo

2022-10-11

Contents

Question 1

Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

Loading packages

Loading our dataset of wage data with several potential explanatory variables

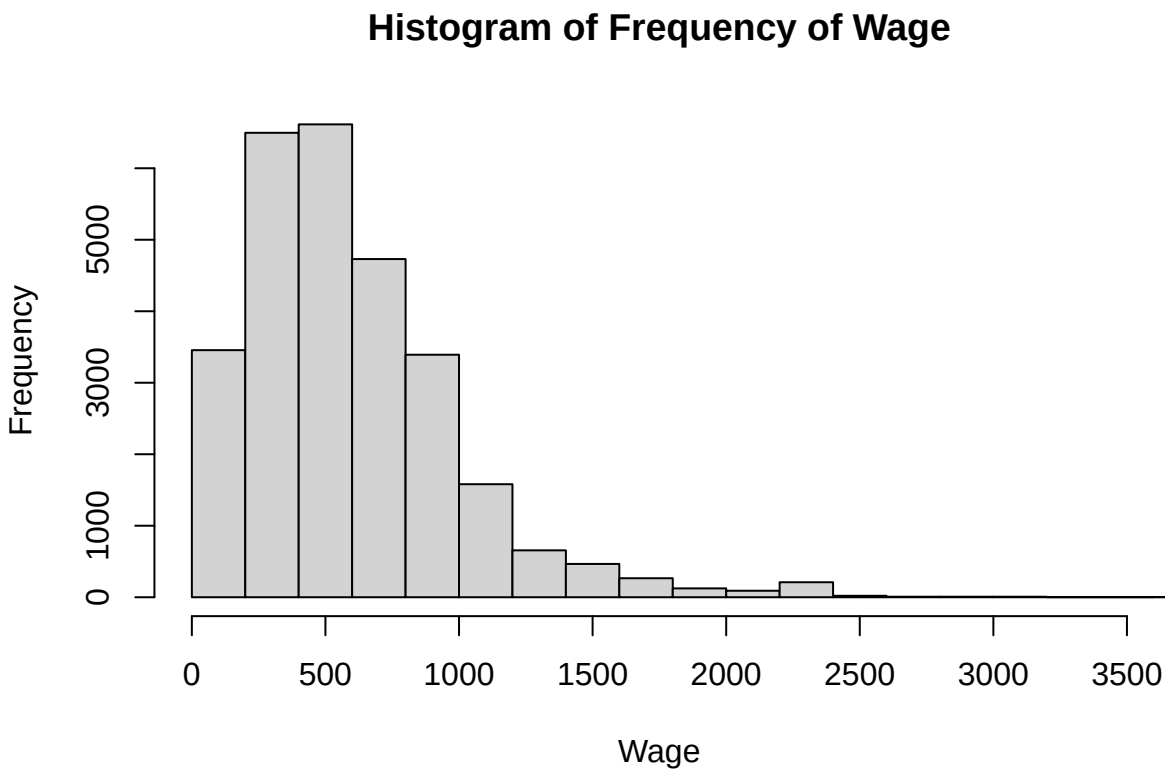
Here are the first 6 rows present in the in the CPS1988 dataset

```
CPS1988_wage_data <- CPS1988
head(CPS1988)
```

```
##     wage education experience ethnicity smsa    region parttime
## 1 354.94         7         45      cauc  yes northeast       no
## 2 123.46        12          1      cauc  yes northeast      yes
## 3 370.37         9          9      cauc  yes northeast       no
## 4 754.94        11         46      cauc  yes northeast       no
## 5 593.54        12         36      cauc  yes northeast       no
## 6 377.23        16         22      cauc  yes northeast       no
```

Histograms and Bar Plots

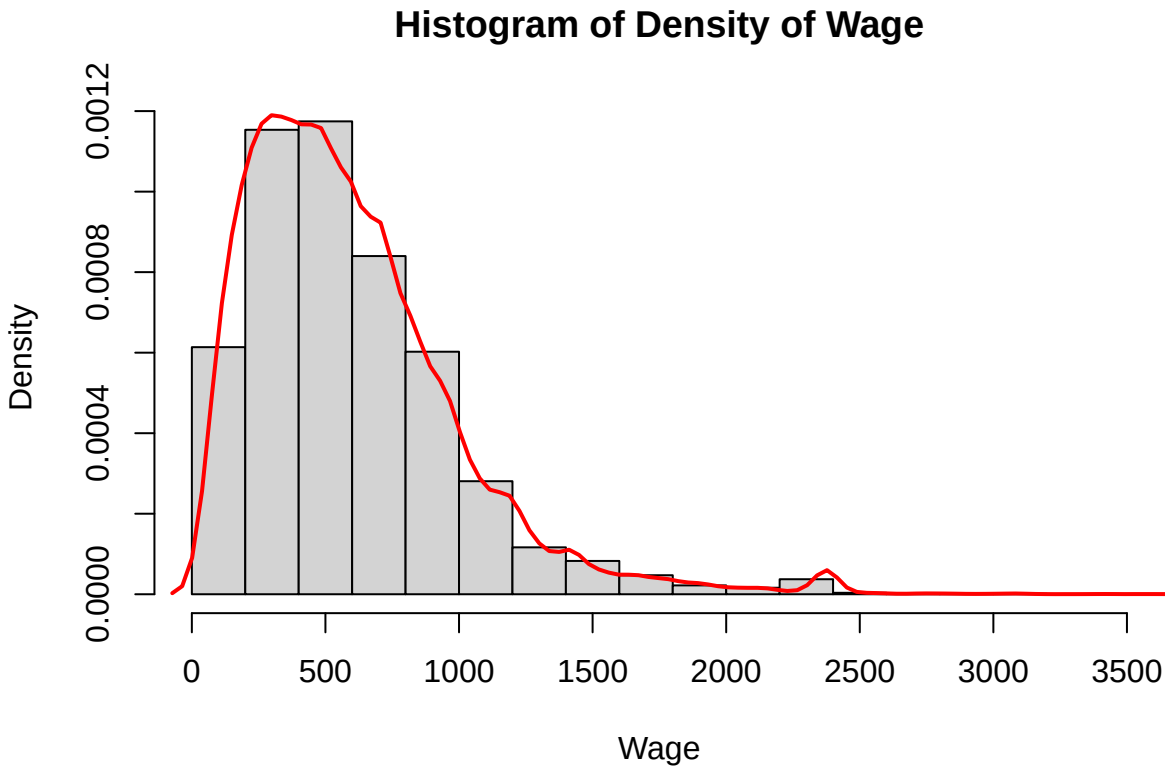
```
# Frequency histogram distribution
hist(CPS1988$wage, xlab = "Wage", ylab = "Frequency", main = "Histogram of Frequency of Wage",
    xlim = c(0, 3500), breaks = 80)
```




```
# Density histogram distribution
hist(CPS1988$wage, xlab = "Wage", ylab = "Density", main = "Histogram of Density of Wage",
    xlim = c(0, 3500), breaks = 80, probability = TRUE)
summary(CPS1988$wage)
```

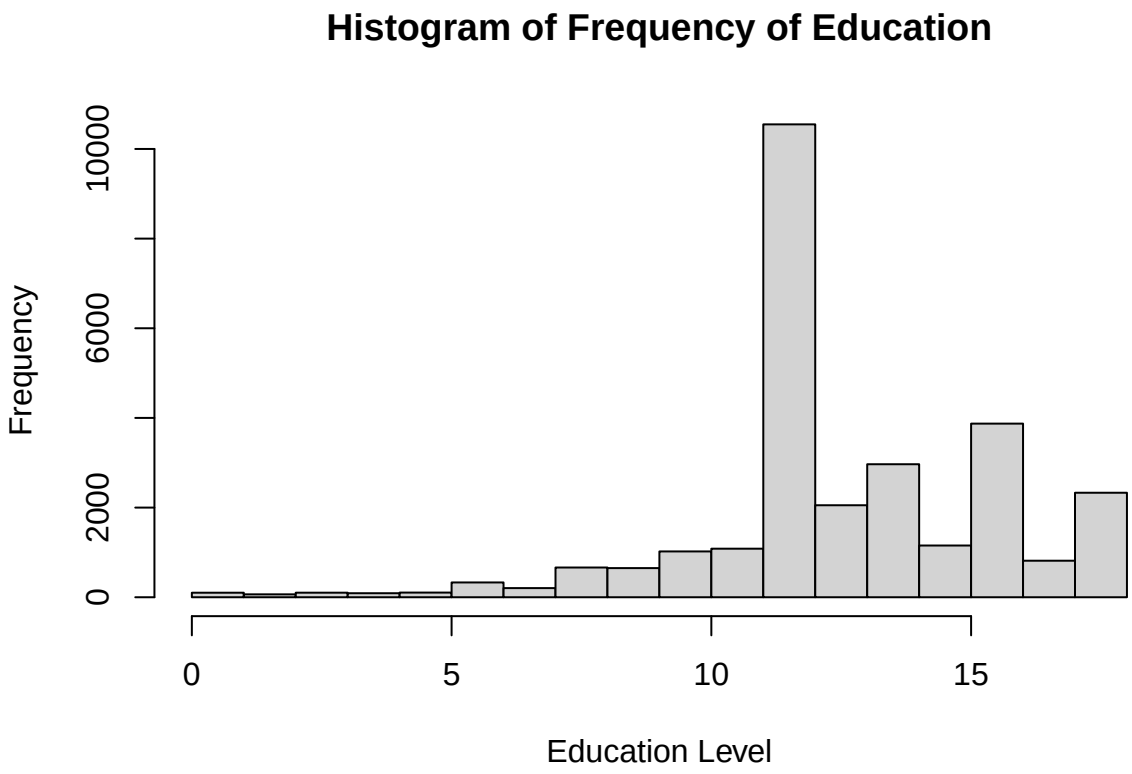
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    50.05   308.64   522.32   603.73   783.48 18777.20
```

```
lines(density(CPS1988$wage), lwd = 2, col = "red")
```

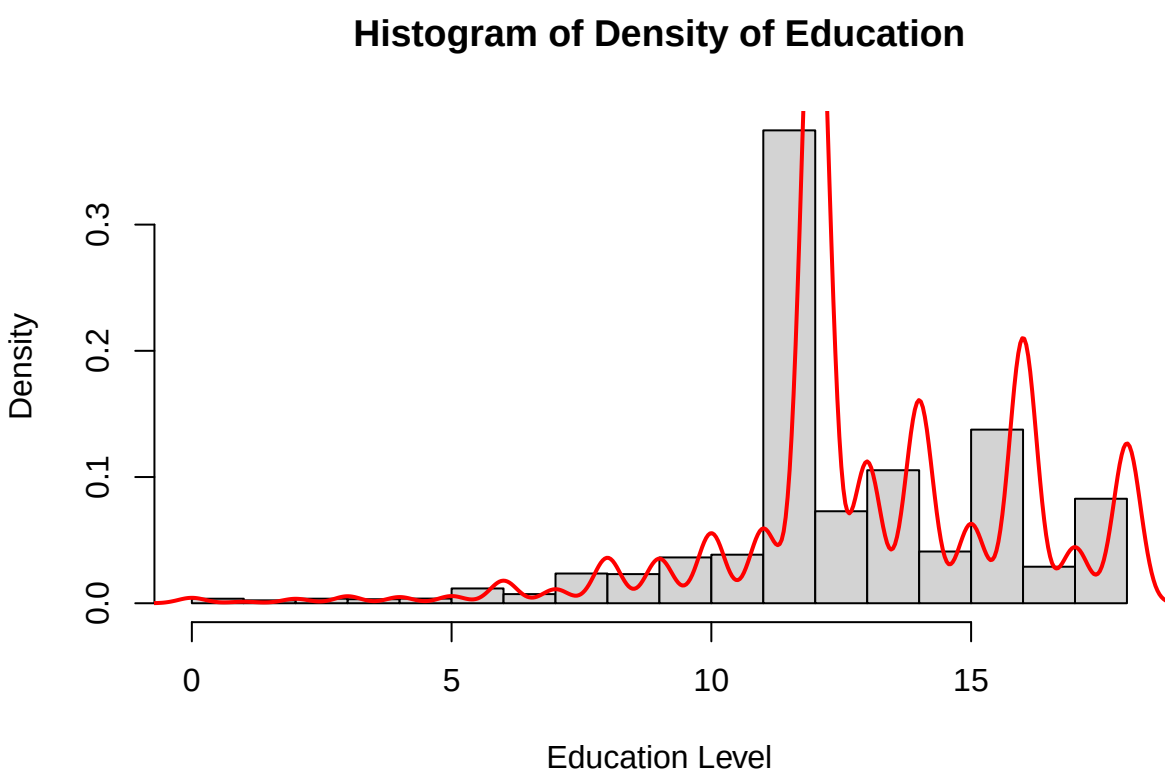


These histograms shows that more than 20,000 of the individuals' wages fall between 0 and 1,000. The density is highest just before a wage level of 500. Less than 5,000 people our of the 28,155 entries earn more than 1,000 in wages over the given time period. **One interesting observation is that this appears to follow a lognormal distribution, with a right-tailed skew.** This will potentially be explored as a source of misspecification/respecification. The median wage is 522.32 per time unit, and the mean is 603.73 (most likely time unit is weeks). This skew follows an expected wealth/wage distribution, with several high earners and a bulk mass of median/low income earners.

```
# Frequency histogram
hist(CPS1988$education, xlab = "Education Level", ylab = "Frequency",
    main = "Histogram of Frequency of Education")
```




```
# Density Histogram
hist(CPS1988$education, xlab = "Education Level", ylab = "Density",
    main = "Histogram of Density of Education", probability = TRUE)
lines(density(CPS1988$education), lwd = 2, col = "red")
```

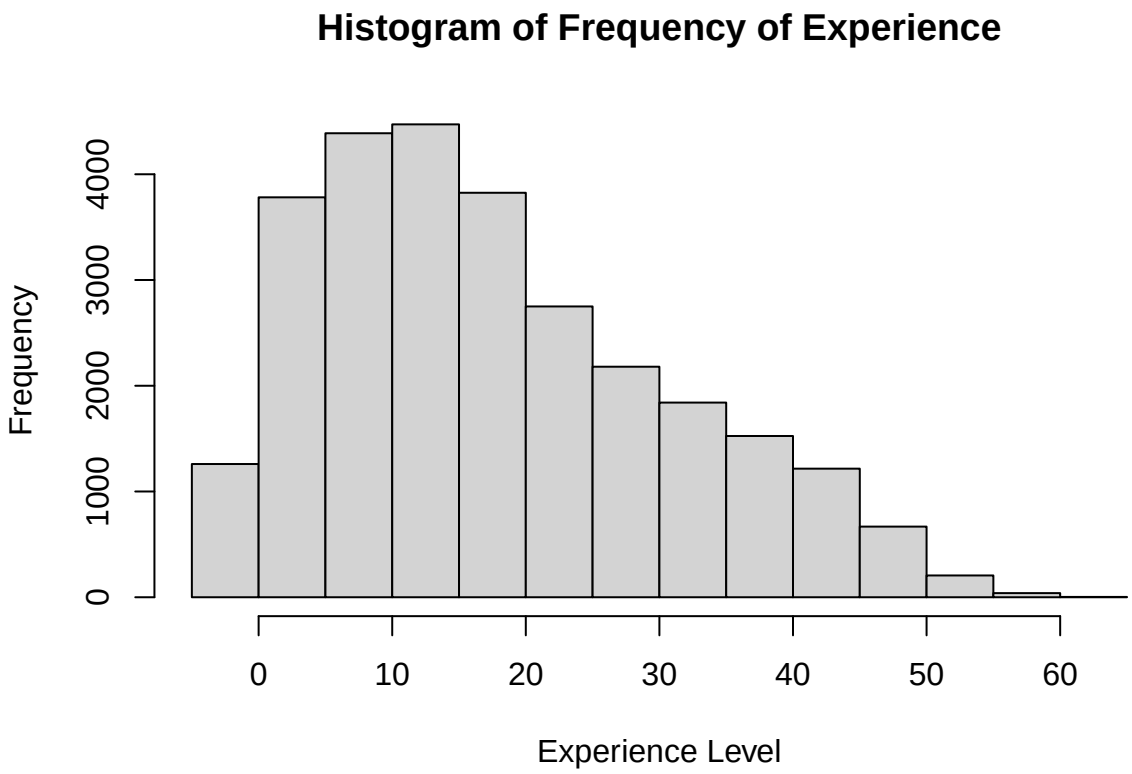



```
summary(CPS1988$education)
```

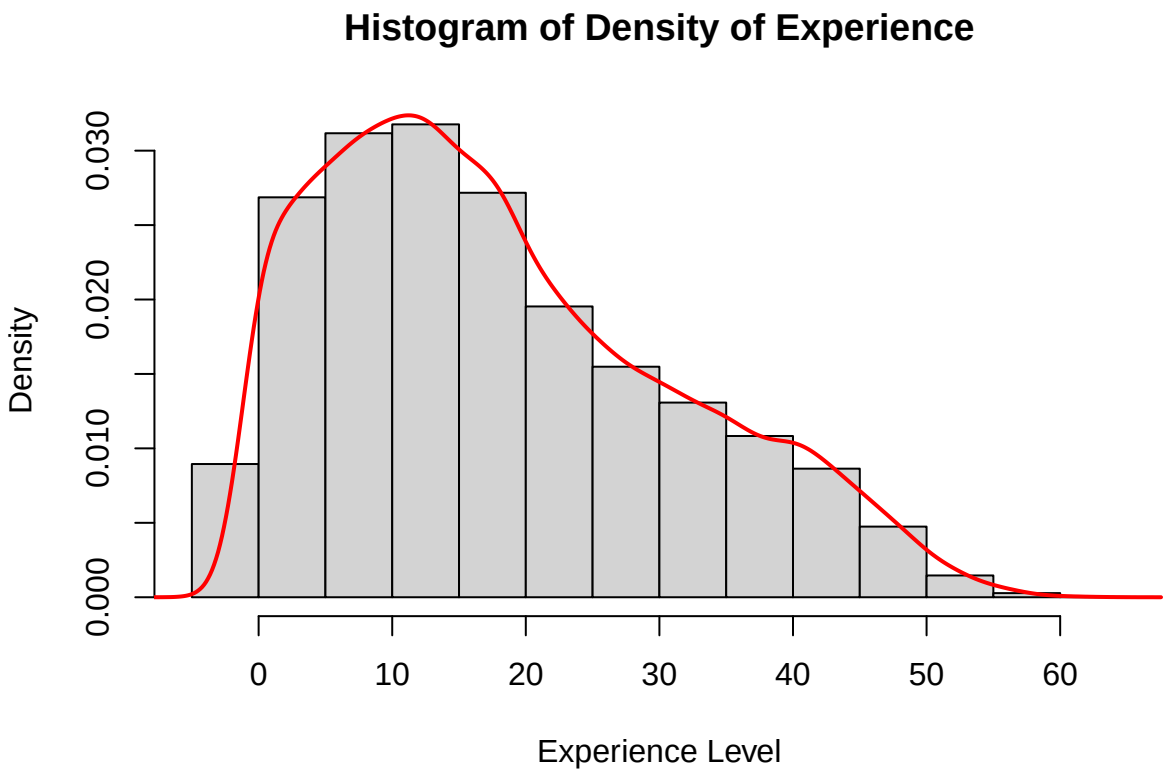
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   13.07   15.00   18.00
```

Above are the histograms for education level, in years in the data set, as well as the 5 number summary. Based on the spike at 12 years of education - where the highest density is as well - it appears that this data set has a significant portion of high school graduates, and then a spread of individuals with varying higher education. This could be useful in running filtered statistical analyses based on education status (i.e. high school dropouts vs. grads vs. higher ed.).

```
# Frequency histogram
hist(CPS1988$experience, xlab = "Experience Level", ylab = "Frequency",
    main = "Histogram of Frequency of Experience")
```




```
# Density histogram
hist(CPS1988$experience, xlab = "Experience Level", ylab = "Density",
    main = "Histogram of Density of Experience", probability = TRUE)
lines(density(CPS1988$experience), col = "red", lwd = 2)
```

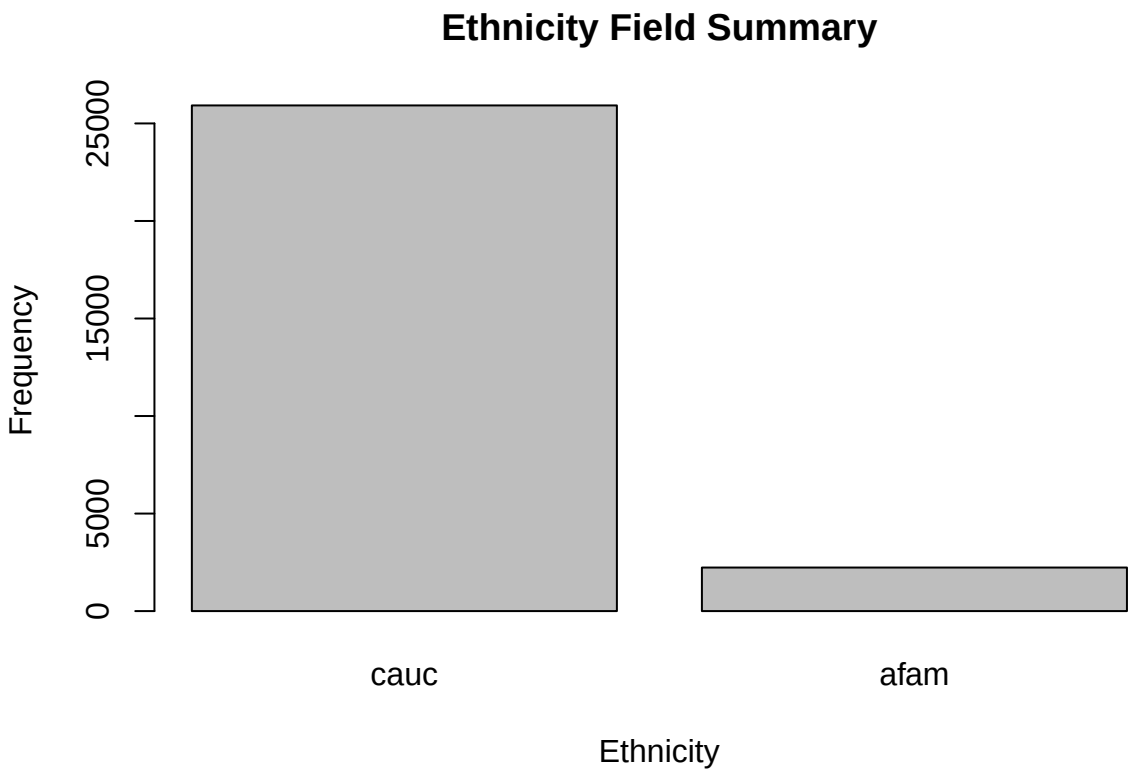



```
summary(CPS1988$experience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -4.0     8.0    16.0    18.2    27.0    63.0
```

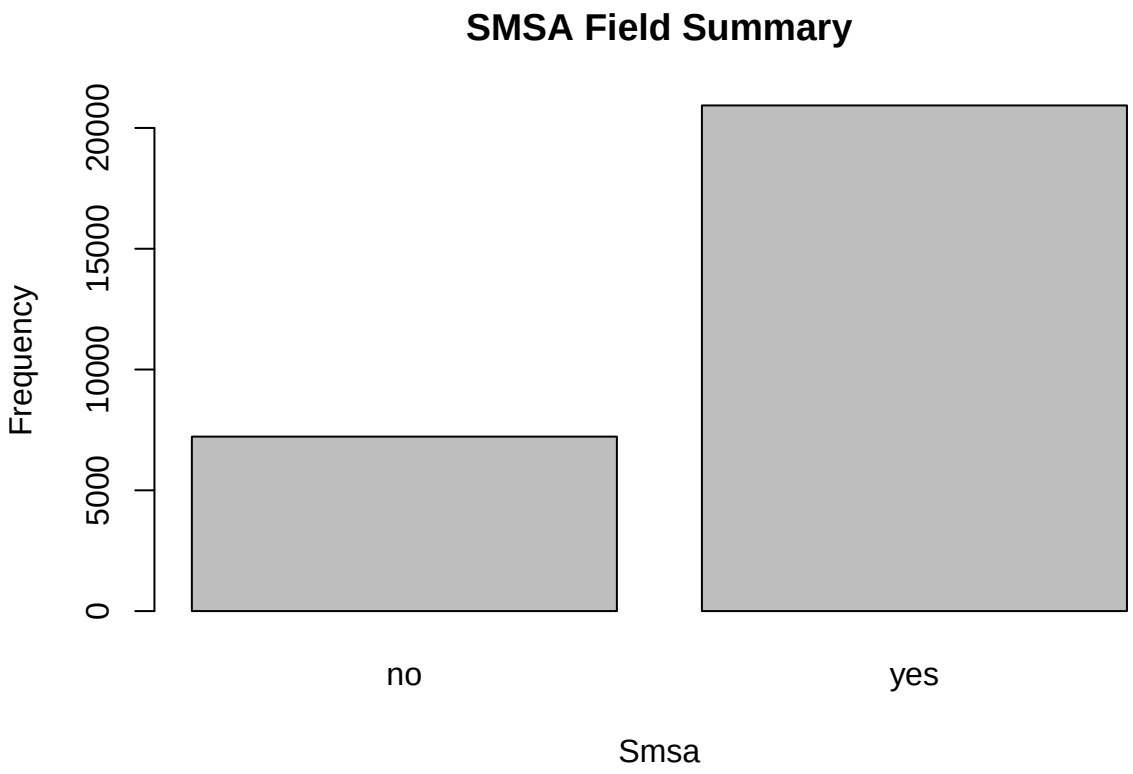
Above are the histograms for experience level, in years in the data set as well as the 5 number summary. The histogram has a right-skewed distribution, although not as prominently as to say it is a lognormal distribution like that for wage. The median experience level is 16 years, and the mean is 18.2 years. The highest density of experience occurs just after the 10 year mark.

```
barplot(table(CPS1988$ethnicity), xlab = "Ethnicity", ylab = "Frequency",
    main = "Ethnicity Field Summary")
```



Above is the barplot for the ethnicities of the observed population, either caucasian or african-american. This dataset is primarily comprised of caucasian individuals.

```
barplot(table(CPS1988$smsa), xlab = "Smsa", ylab = "Frequency",
    main = "SMSA Field Summary")
```

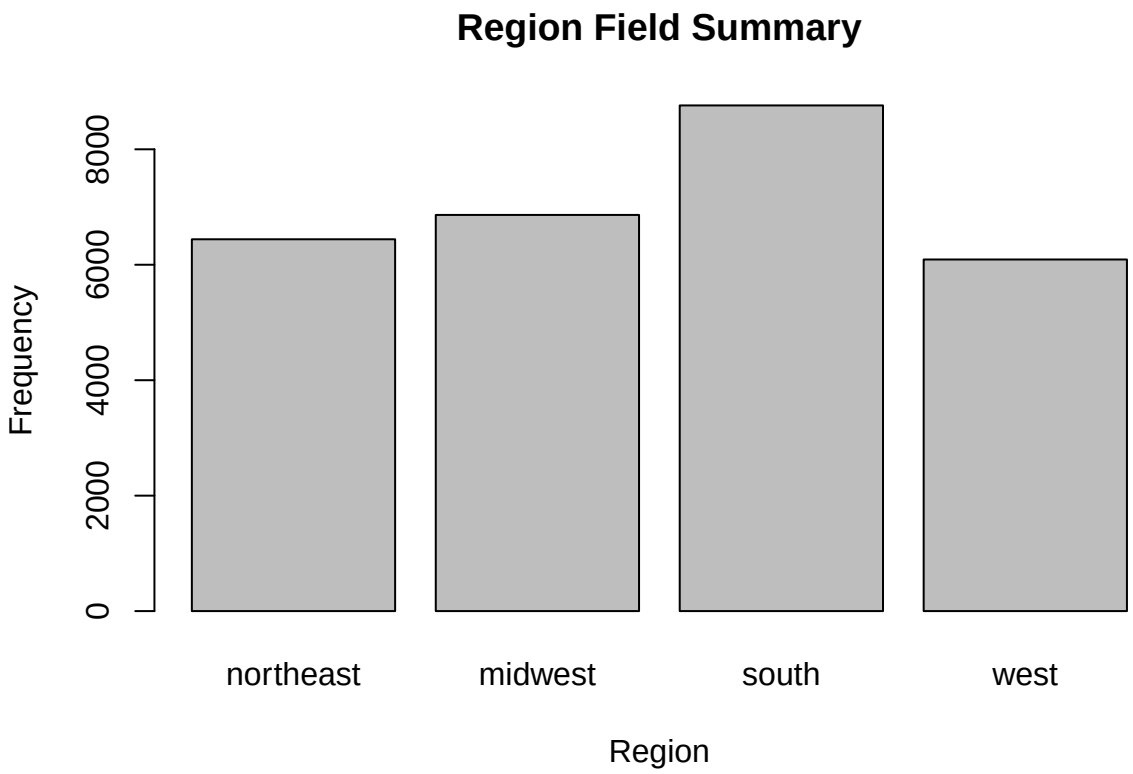



```
summary(CPS1988$smsa)
```

```
##    no   yes
##  7223 20932
```

Above is the barplot indicating whether the observed lives in a Standard Metropolitan Statistical Area (SMSA). The majority of individuals surveyed live in a SMSA.

```
barplot(table(CPS1988$region), xlab = "Region", ylab = "Frequency",
    main = "Region Field Summary")
```

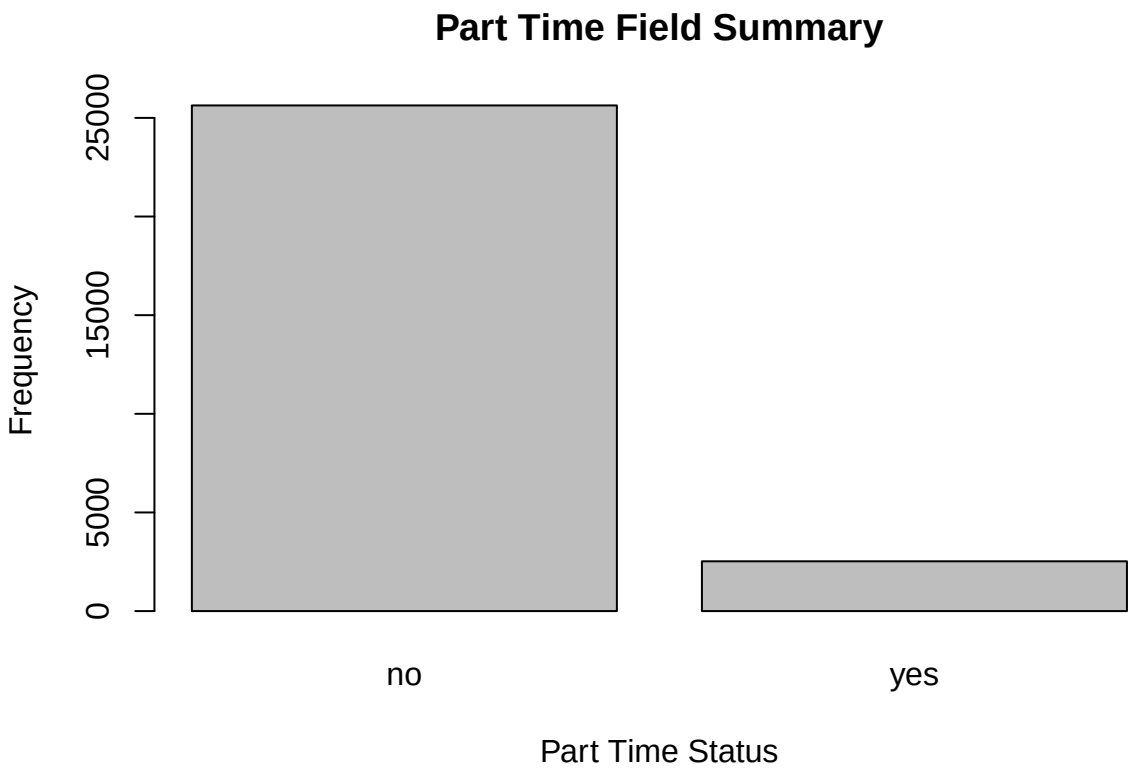



```
summary(CPS1988$region)
```

```
## northeast   midwest     south      west
##      6441      6863      8760      6091
```

Above is the barplot for region of the observed population. There is a relatively mixed distribution of regions in this data set, with the most individuals coming from the South, and the least (by a small margin) coming from the West.

```
barplot(table(CPS1988$parttime), xlab = "Part Time Status", ylab = "Frequency",
    main = "Part Time Field Summary")
```

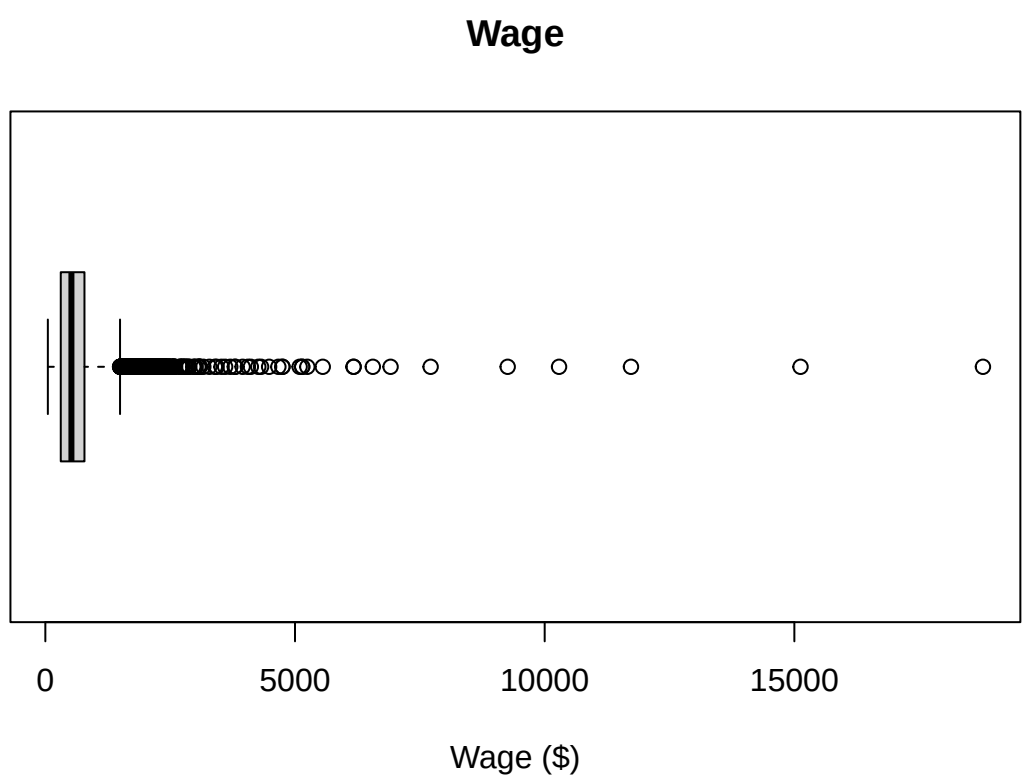



```
summary(CPS1988$parttime)
```

```
##    no   yes
## 25631  2524
```

Above is the barplot for the full-time vs. part-time status of the observed population. The overwhelming majority of individuals in this data set are not part-time (i.e. are full-time employees).

```
boxplot(CPS1988$wage, main = "Wage", xlab = "Wage ($)", horizontal = TRUE)
```

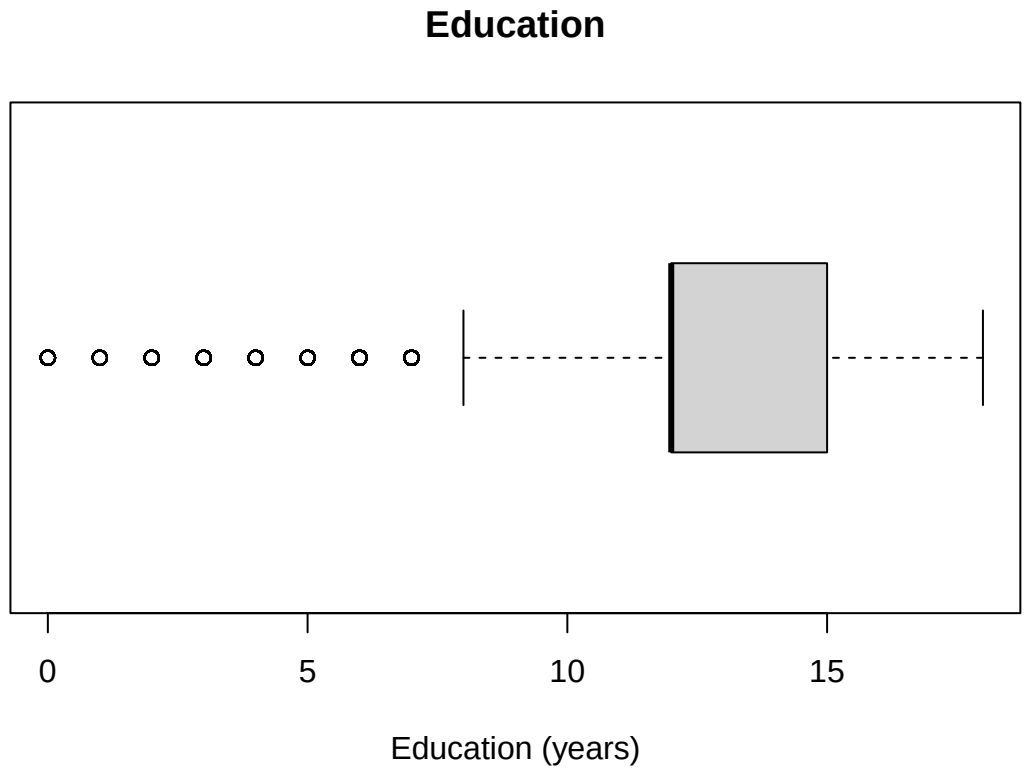



```
summary(CPS1988$wage)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    50.05   308.64   522.32   603.73   783.48 18777.20
```

Above is the boxplot for wage. Notice that the data exhibits a right-tailed skew, and potentially shows signs of a lognormal distribution. The majority of the individuals in this data set make very little money. As wage increases, there are less and less individuals who make more money.

```
boxplot(CPS1988$education, main = "Education", xlab = "Education (years)",
    horizontal = TRUE)
```

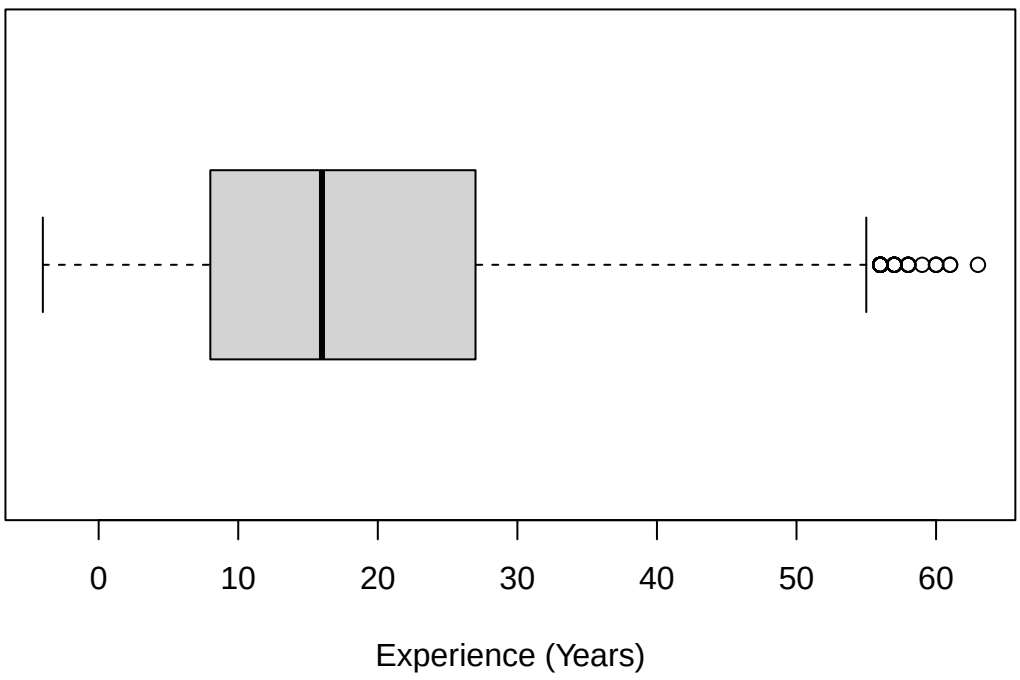



```
summary(CPS1988$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   13.07   15.00   18.00
```

Above is the boxplot for education. Notice that most of the data is concentrated between 12 and 15 years - individuals who have earned a high school degree and attended some college. Very few people have less than 10 years of education and are considered outliers.

```
boxplot(CPS1988$experience, main = "Experience", xlab = "Experience (Years)",
    horizontal = TRUE)
```




```
summary(CPS1988$experience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -4.0     8.0    16.0    18.2    27.0    63.0
```

Above is the boxplot for experience from 0 to 70 years. Notice that most people have between 8 years and 27 years of experience before they retire. Also notice the erroneous entry for -4 years of experience that we omit from our data as an outlier in the coming questions.

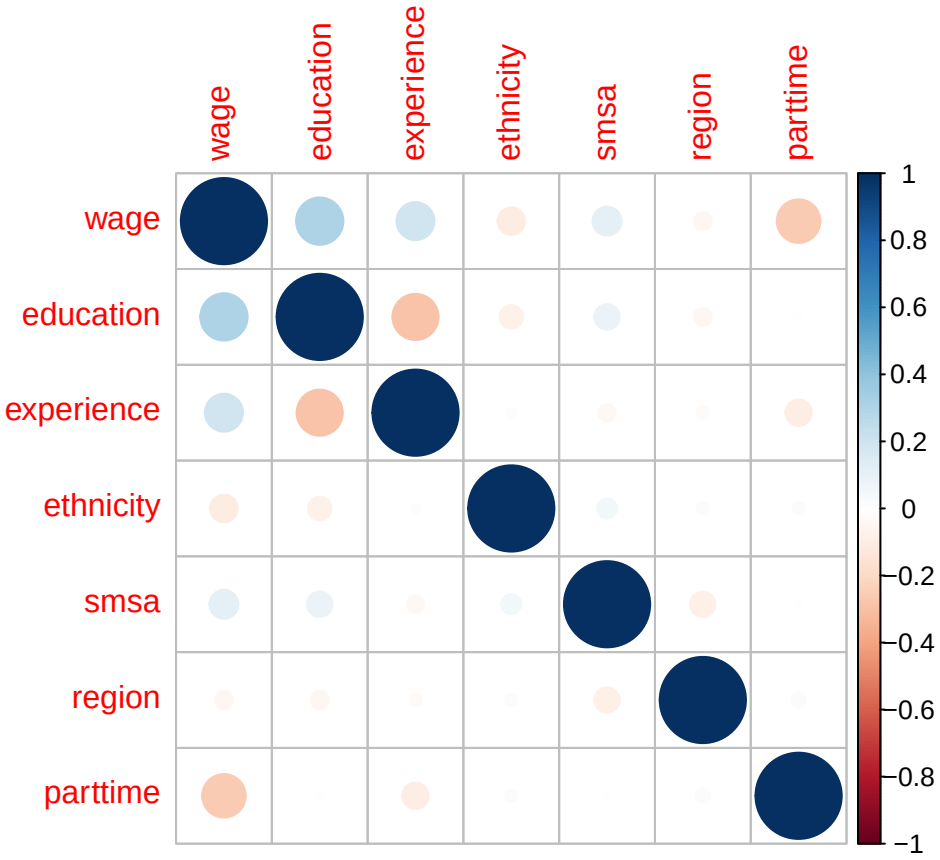
Correlation Charts and Scatterplots

Below is the correlation plot for the numeric columns for our data set.

```
library(corrplot)
```

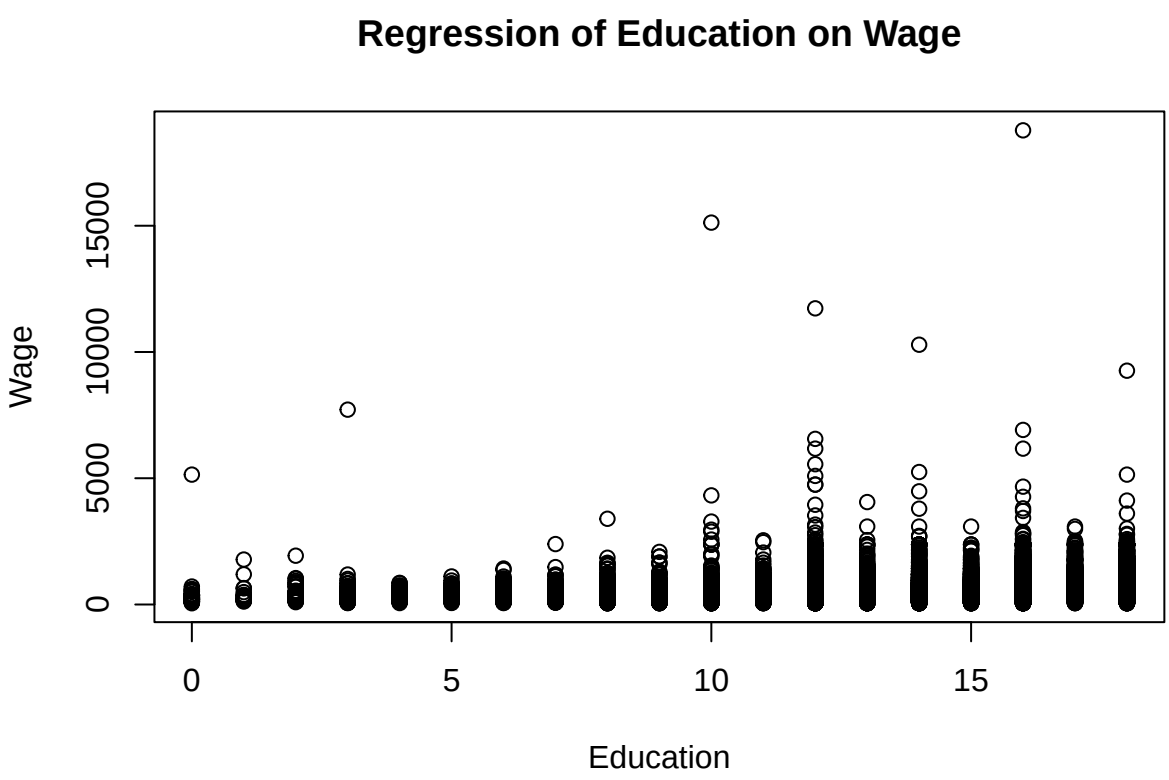
```
## corrplot 0.92 loaded
```

```
CPS1988$region <- as.numeric(CPS1988$region)
CPS1988$ethnicity <- as.numeric(CPS1988$ethnicity)
CPS1988$smsa <- as.numeric(CPS1988$smsa)
CPS1988$parttime <- as.numeric(CPS1988$parttime)
X <- cor(CPS1988)
corrplot(X, method = "circle")
```



As you can see in the correlation plot above, education is positively correlated with wage at a level of approximately 0.5. Individuals who have more years of schooling tend to earn more wages. Wage is also positively correlated with experience, but less so than education. This means that individuals who have more experience tend to earn higher wages. A more sublte positive correlation can be found for smsa and wage, meaning that those who live in metropolitan areas tend to earn higher wages. On the other hand, wage is negatively correlated with the parrtime variable at approximately -0.3. This makes intuitive sense because individuals who work parttime earn less money than those who work full time. Moreover, education is negatively correlated with experience and vice versa. This means that the more experience an individual has, the less likely they are to have more years of education.

```
educ_wage_lm <- lm(CPS1988$wage ~ CPS1988$education)
plot(CPS1988$wage ~ CPS1988$education, xlab = "Education", ylab = "Wage",
    main = "Regression of Education on Wage")
```




```
summary(educ_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$education)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -786.0  -264.9   -54.8   174.5 18035.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12.8281    11.8970  -1.078    0.281
## CPS1988$education  47.1810     0.8888  53.085   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.4 on 28153 degrees of freedom
## Multiple R-squared:  0.09099,    Adjusted R-squared:  0.09096
## F-statistic:  2818 on 1 and 28153 DF,  p-value: < 2.2e-16
```

This regression shows that increases in total years of education are visually slightly associated with an increase in wage, although any correlation appears to be quite weak. It is notable that the vast majority of high-earning outliers appear on the high side of Education (implying that increased education provides, at least, an increased chance of becoming an *extremely* high earner), even if the overall coefficient w.r.t. wage is

small. The summary output of the correlation model indicates a positive coefficient of 47$ of wage resulting from an additional year of education, with a 0.091 R^2 value, implying very little, but not totally negligible predictive power.

```
exp_wage_lm <- lm(CPS1988$wage ~ CPS1988$experience)
plot(CPS1988$wage ~ CPS1988$experience, xlab = "Experience",
    ylab = "Wage", main = "Regression of Experience on Wage")
```


```
summary(exp_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$experience)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -823.5  -271.2   -77.1   174.0 18275.8
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         481.1510     4.5437  105.89   <2e-16 ***
## CPS1988$experience    6.7350     0.2027   33.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 444.9 on 28153 degrees of freedom
## Multiple R-squared:  0.03772,    Adjusted R-squared:  0.03769
## F-statistic:  1104 on 1 and 28153 DF,  p-value: < 2.2e-16
```

This regression shows that there is no clear association between an increase in wage due to more years of experience. It could be argued that the same occurence with high-earning outliers observed with the education scatterplot is evident here, but it is less notable as the high earning outliers appear more distributed in the middle of the graph then on the right side. The summary of the regression model confirms this, with an R^2 <0.038, implying extremely low predictive power.

The remainder of the potential explanatory variables are "binary" or Y/N variables and so cannot as elegantly be displayed with a scatterplot. Therefore, we will provide the linear model summaries of wage correlated to each variable.

```
eth_wage_lm <- lm(CPS1988$wage ~ CPS1988$ethnicity)
summary(eth_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -567.2  -284.9   -76.5   190.0 18160.0
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         787.615     11.074   71.12   <2e-16 ***
## CPS1988$ethnicity -170.381      9.953  -17.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 451.2 on 28153 degrees of freedom
## Multiple R-squared:  0.0103, Adjusted R-squared:  0.01027
## F-statistic:   293 on 1 and 28153 DF,  p-value: < 2.2e-16
```

```
smsa_wage_lm <- lm(CPS1988$wage ~ CPS1988$smsa)
summary(smsa_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$smsa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -583.6  -285.3   -73.0   195.3 18260.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    400.18      11.05   36.20   <2e-16 ***
## CPS1988$smsa   116.75       6.15   18.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 450.7 on 28153 degrees of freedom
## Multiple R-squared:  0.01264,    Adjusted R-squared:  0.0126
## F-statistic: 360.4 on 1 and 28153 DF,  p-value: < 2.2e-16
```

```
reg_wage_lm <- lm(CPS1988$wage ~ CPS1988$region)
summary(reg_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -580.6  -291.7   -83.9   183.2 18182.3
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     649.754      6.911  94.021  < 2e-16 ***
## CPS1988$region  -18.301      2.529  -7.236 4.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 453.1 on 28153 degrees of freedom
## Multiple R-squared:  0.001856,   Adjusted R-squared:  0.001821
## F-statistic: 52.35 on 1 and 28153 DF,  p-value: 4.756e-13
```

```
pt_wage_lm <- lm(CPS1988$wage ~ CPS1988$parttime)
summary(pt_wage_lm)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$parttime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -589.8  -262.2   -78.3   162.3 18137.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1046.599     10.303  101.58   <2e-16 ***
## CPS1988$parttime -406.436      9.147  -44.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 438.4 on 28153 degrees of freedom
## Multiple R-squared:  0.06554,    Adjusted R-squared:  0.06551
## F-statistic:  1975 on 1 and 28153 DF,  p-value: < 2.2e-16
```

All of the categorical variables have very low R^2, most below 0.02, implying middling to no meaningful correlation with wage. It is useful to note that indicating part-time status has a R^2 of 0.066, but a *very* negative correlation with wage - with a coefficient of -406.4$

Question 2

Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates

```
multiple_regression <- lm(CPS1988$wage ~ CPS1988$education +
    CPS1988$experience + CPS1988$ethnicity + CPS1988$smsa + CPS1988$region +
    CPS1988$parttime)
summary(multiple_regression)
```

```
##
## Call:
## lm(formula = CPS1988$wage ~ CPS1988$education + CPS1988$experience +
##     CPS1988$ethnicity + CPS1988$smsa + CPS1988$region + CPS1988$parttime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1034.7  -206.8   -48.8   134.5 18187.8
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           37.881     21.594   1.754   0.0794 .
## CPS1988$education     57.270      0.854  67.062   <2e-16 ***
## CPS1988$experience     9.803      0.189  51.880   <2e-16 ***
## CPS1988$ethnicity   -130.893      8.757 -14.948   <2e-16 ***
## CPS1988$smsa         100.227      5.440  18.425   <2e-16 ***
## CPS1988$region        -2.273      2.217  -1.025   0.3052
## CPS1988$parttime    -356.741      8.286 -43.055   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.9 on 28148 degrees of freedom
## Multiple R-squared:  0.2419, Adjusted R-squared:  0.2417
## F-statistic:  1497 on 6 and 28148 DF,  p-value: < 2.2e-16
```

Above, we generate an initial multiple regression model based on all the possible explanatory variables in our data set. The adjusted R-squared is 0.2429, which *could* imply low predictive power among other things, and indicates that under 1/4 of variance in the data is explained by the model.

Question 3

Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.

We now will idenfity outliers and high leverage observations using a IQR *1.5 threshold for outliers of the response and continuous predictor variables, and Cooks Distances for leverage observations.

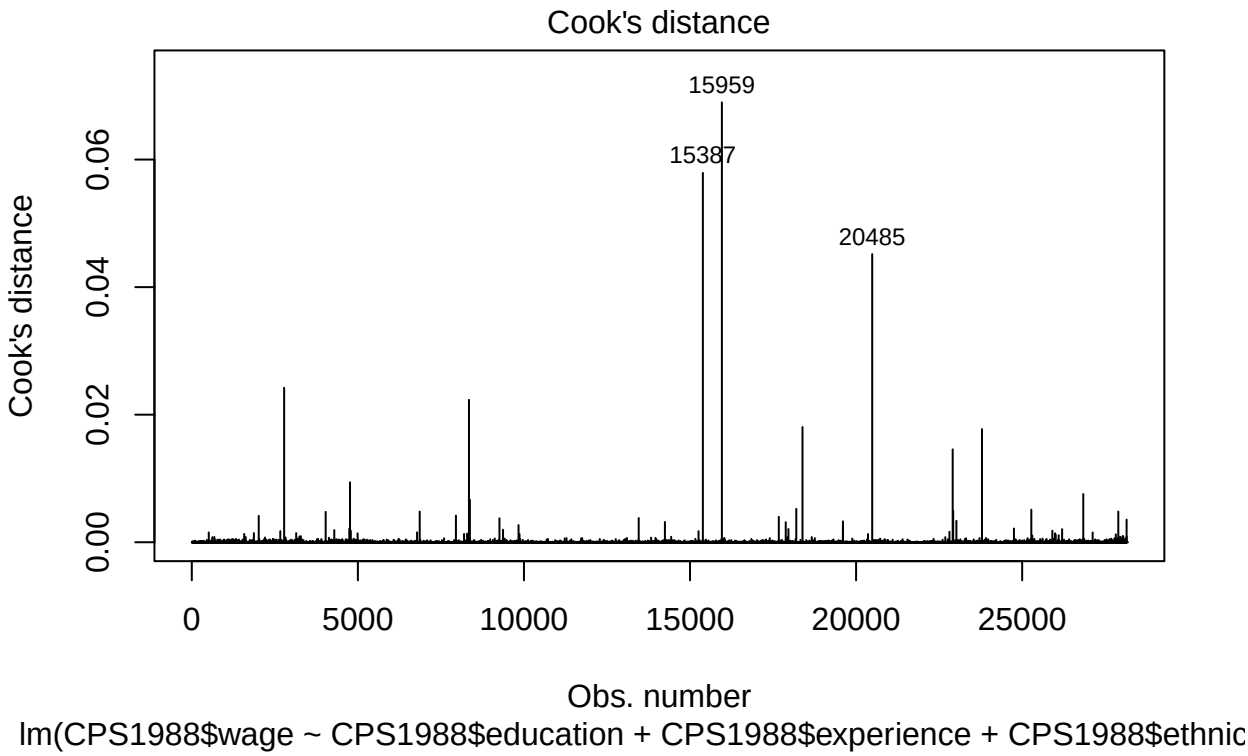
IQR Outliers

```
# Investigate outliers for all variables Wage
quartiles <- quantile(CPS1988$wage, probs = c(0.25, 0.75), na.rm = FALSE)
IQR <- IQR(CPS1988$wage)
UQ <- quartiles[2] + 1.5 * IQR
LQ <- quartiles[1] - 1.5 * IQR
out_rem <- subset(CPS1988, CPS1988$wage > LQ & CPS1988$wage <
    UQ)

# Experience
quartiles <- quantile(out_rem$experience, probs = c(0.25, 0.75),
    na.rm = FALSE)
IQR <- IQR(out_rem$experience)
UQ <- quartiles[2] + 1.5 * IQR
LQ <- quartiles[1] - 1.5 * IQR
out_remv <- subset(out_rem, out_rem$experience > LQ & out_rem$experience <
    UQ)

# Education
quartiles <- quantile(out_remv$education, probs = c(0.25, 0.75),
    na.rm = FALSE)
IQR <- IQR(out_remv$education)
UQ <- quartiles[2] + 1.5 * IQR
LQ <- quartiles[1] - 1.5 * IQR
outliers_removed <- subset(out_remv, out_remv$education > LQ &
    out_remv$education < UQ)
```

We identified outliers using a range threshold of 1.5 * the Inter-Quartile Range for the continuous, non-categorical variables in our data set, and removed the outliers for each observation. Now we will proceed to analyze high leverage and influential observations using Cooks Distance formulas.

Cooks Distance

```
p = 6
n = 27198
cooksD <- cooks.distance(multiple_regression)
plot(multiple_regression, which = 4)
```




```
influential_obs <- as.numeric(names(cooksD)[(cooksD > (2 * (p +
    2)/n))])
influential_obs2 <- as.numeric(names(cooksD)[(cooksD > (2 * mean(cooksD)))])
influential_obs3 <- as.numeric(names(cooksD)[(cooksD > 1)])

CPS1988_clean <- outliers_removed[-influential_obs, -influential_obs2]
```

Out rationale behind excluding the observations with high Cooks values can be seen through the nature of the data set as well as intuitively: specifically for wage, the presence of several high earners, who may have a completely disparate set of explanatory parameters due to confounding variables or sheer chance makes the extremely influential over the model parameters. Therefore, we decided to exclude wage outliers and high leverage observations as identified by the Cooks Distance method. Due to the Cooks Distance method of identifying influential observations, we now have a cleaned dataset from several iterations of outlier removal, and a round of Cooks Distance analysis. Let us now use this data to create an updated regression model.

Cleaned Data Multiple Regression

```
# Create new regression model from cleaned data set
multiple_regression_clean <- lm(CPS1988_clean$wage ~ CPS1988_clean$education +
    CPS1988_clean$experience + CPS1988_clean$ethnicity + CPS1988_clean$smsa +
    CPS1988_clean$region + CPS1988_clean$parttime)
summary(multiple_regression_clean)
```

```
##
```

```
## Call:
## lm(formula = CPS1988_clean$wage ~ CPS1988_clean$education + CPS1988_clean$experience +
##     CPS1988_clean$ethnicity + CPS1988_clean$smsa + CPS1988_clean$region +
##     CPS1988_clean$parttime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -900.59 -175.37  -24.97  148.77 1213.50
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                169.8828    15.1372  11.223  < 2e-16 ***
## CPS1988_clean$education     47.4942     0.6848  69.359  < 2e-16 ***
## CPS1988_clean$experience    7.8984     0.1284  61.508  < 2e-16 ***
## CPS1988_clean$ethnicity  -103.3124     5.8865 -17.551  < 2e-16 ***
## CPS1988_clean$smsa         74.4613     3.6508  20.396  < 2e-16 ***
## CPS1988_clean$region       -4.7886     1.5019  -3.188  0.00143 **
## CPS1988_clean$parttime   -351.2003     5.5337 -63.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 256.8 on 26157 degrees of freedom
## Multiple R-squared:  0.3364, Adjusted R-squared:  0.3362
## F-statistic:  2210 on 6 and 26157 DF,  p-value: < 2.2e-16
```

First, we cleaned our data from outliers with a 1.5*IQR threshold for wage, as well as using the Cooks Distances of our observations to remove high leverage and overly influential observations, and also excluding instances where an explanatory variable with a floor of zero had a listed value of less than 0. Then, we run a linear regression model on the data set without the cleaned observations.

The new adjusted R^2 is 0.3368, which is still not very high, implying that the model only accounts for 33.68% of the variance observed in the data. This *could* imply poor predictive power of the explanatory variables, functional form misspecification, or simply result from a particularly random data spread.

Question 4

Use Mallows Cp for identifying which terms you will keep in the model (based on part 3) and also use the Boruta algorithm for variable selection. Based on the two results, determine which subset of predictors you will keep.

```
set.seed(123)
boruta <- Boruta(CPS1988_clean$wage ~ ., data = CPS1988_clean,
    doTrace = 1)
```

```
## After 10 iterations, +2.1 mins:
```

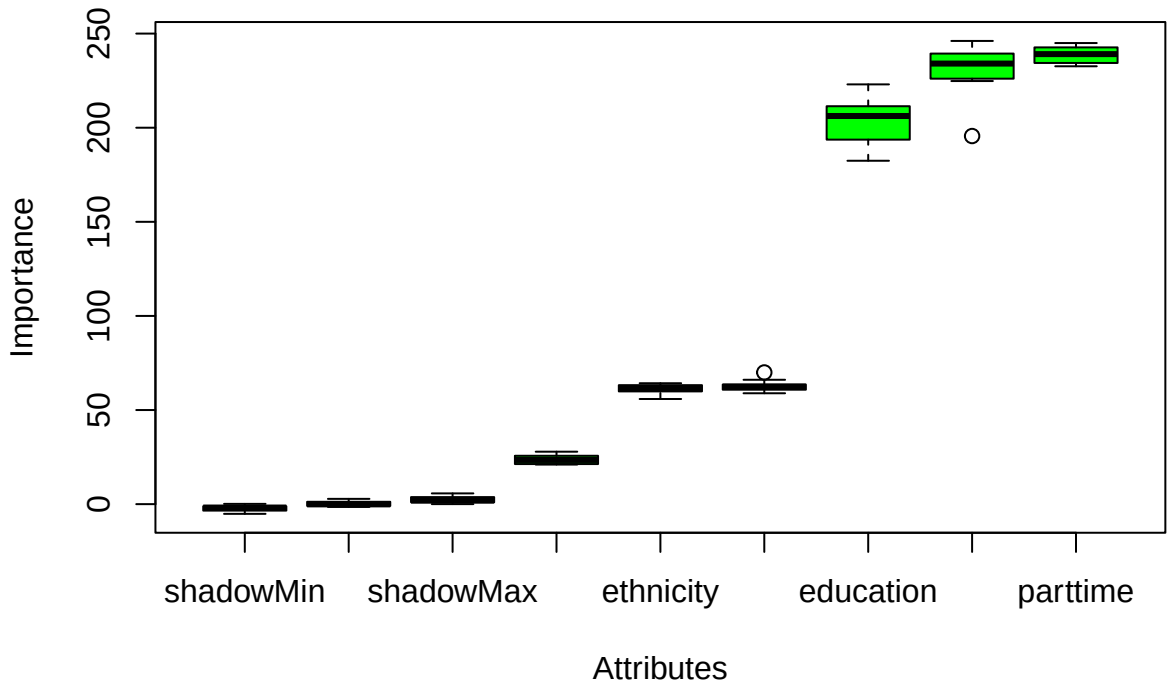
```
##  confirmed 6 attributes: education, ethnicity, experience, parttime, region and 1 more;
```

```
##  no more attributes left.
```

```
print(boruta)
```

```
## Boruta performed 10 iterations in 2.133248 mins.
##  6 attributes confirmed important: education, ethnicity, experience,
## parttime, region and 1 more;
##  No attributes deemed unimportant.
```

```
plot(boruta)
```



As seen in the output and in the graph, the boruta algorithm deemed *all 6 explanatory variables as important.* Therefore, this suggests we should keep all of our potential explanatory variables in the regression as is.

Below, we will assess using the Mallows CP method which selects the best model out of a set of possible models with differing explanatory variables

Mallows CP

```
cp_model_1 <- lm(CPS1988_clean$wage ~ CPS1988_clean$experience +
    CPS1988_clean$education + CPS1988_clean$parttime)
cp_model_2 <- lm(CPS1988_clean$wage ~ CPS1988_clean$experience +
    CPS1988_clean$education + CPS1988_clean$parttime + CPS1988_clean$ethnicity)
cp_modeL_3 <- lm(CPS1988_clean$wage ~ CPS1988_clean$experience +
    CPS1988_clean$education + CPS1988_clean$parttime + CPS1988_clean$smsa)
```

```
cp_model_4 <- lm(CPS1988_clean$wage ~ CPS1988_clean$experience +
    CPS1988_clean$education + CPS1988_clean$parttime + CPS1988_clean$region)
```

```
ols_mallows_cp(cp_model_1, multiple_regression_clean)
```

```
## [1] 707.6726
```

```
ols_mallows_cp(cp_model_2, multiple_regression_clean)
```

```
## [1] 444.9728
```

```
ols_mallows_cp(cp_modeL_3, multiple_regression_clean)
```

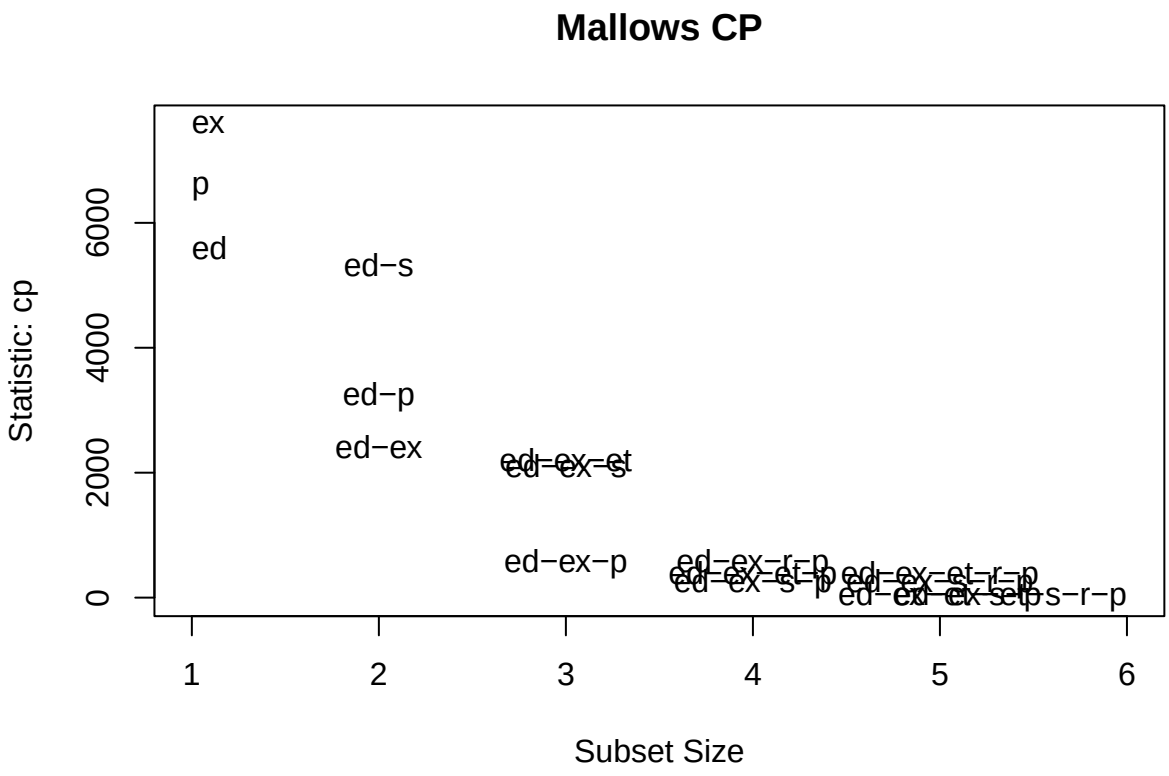
```
## [1] 325.0136
```

```
ols_mallows_cp(cp_model_4, multiple_regression_clean)
```

```
## [1] 679.4287
```

Let us now plot the Mallows CP values for the exhaustive list of potential predictive models, using subset size as the x axis and CP statistic as the Y axis.

```
ss <- regsubsets(wage ~ education + experience + ethnicity +
    smsa + region + parttime, method = c("exhaustive"), nbest = 3,
    data = CPS1988)
subsets(ss, statistic = "cp", legend = F, main = "Mallows CP")
```




```
##            Abbreviation
## education            ed
## experience           ex
## ethnicity            et
## smsa                  s
## region                r
## parttime              p
```

```
mallowsum <- summary(ss)
mallowsum[order(mallowsum$cp)]
```

```
## $<NA>
## NULL
##
## $<NA>
## NULL
##
## $<NA>
## NULL
##
## $<NA>
## NULL
##
## $<NA>
## NULL
```

```
##
## $<NA>
## NULL
##
## $<NA>
## NULL
##
## $outmat
##          education experience ethnicity smsa region parttime
## 1  ( 1 ) "*"       " "        " "       " "  " "    " "
## 1  ( 2 ) " "       " "        " "       " "  " "    "*"
## 1  ( 3 ) " "       "*"        " "       " "  " "    " "
## 2  ( 1 ) "*"       "*"        " "       " "  " "    " "
## 2  ( 2 ) "*"       " "        " "       " "  " "    "*"
## 2  ( 3 ) "*"       " "        " "       "*"  " "    " "
## 3  ( 1 ) "*"       "*"        " "       " "  " "    "*"
## 3  ( 2 ) "*"       "*"        " "       "*"  " "    " "
## 3  ( 3 ) "*"       "*"        "*"       " "  " "    " "
## 4  ( 1 ) "*"       "*"        " "       "*"  " "    "*"
## 4  ( 2 ) "*"       "*"        "*"       " "  " "    "*"
## 4  ( 3 ) "*"       "*"        " "       " "  "*"    "*"
## 5  ( 1 ) "*"       "*"        "*"       "*"  " "    "*"
## 5  ( 2 ) "*"       "*"        " "       "*"  "*"    "*"
## 5  ( 3 ) "*"       "*"        "*"       " "  "*"    "*"
## 6  ( 1 ) "*"       "*"        "*"       "*"  "*"    "*"
##
## $obj
## Subset selection object
## Call: regsubsets.formula(wage ~ education + experience + ethnicity +
##     smsa + region + parttime, method = c("exhaustive"), nbest = 3,
##     data = CPS1988)
## 6 Variables  (and intercept)
##            Forced in Forced out
## education      FALSE      FALSE
## experience     FALSE      FALSE
## ethnicity      FALSE      FALSE
## smsa           FALSE      FALSE
## region         FALSE      FALSE
## parttime       FALSE      FALSE
## 3 subsets of each size up to 6
## Selection Algorithm: exhaustive
##
## $<NA>
## NULL
##
## $adjr2
##  [1] 0.09095681 0.06550565 0.03768737 0.17678220 0.15505528 0.09841255
##  [7] 0.22727605 0.18504710 0.18256320 0.23569524 0.23243851 0.22746147
## [13] 0.24170762 0.23571687 0.23259128 0.24170901
##
## $cp
##  [1] 5598.96296 6543.88589 7576.69247 2413.44584 3220.07081 5322.96563
##  [7]  539.81303 2107.53216 2199.74504  228.24900  349.14812  533.91064
## [13]    6.05158  228.43776  344.46498    7.00000
```

```
##
## $bic
##  [1] -2665.445 -1888.002 -1062.109 -5448.382 -4714.933 -2888.072 -7221.318
##  [8] -5723.234 -5637.551 -7520.518 -7400.804 -7218.829 -7733.629 -7512.070
## [15] -7397.162 -7724.435
##
## $which
##   (Intercept) education experience ethnicity  smsa region parttime
## 1        TRUE      TRUE      FALSE     FALSE FALSE  FALSE    FALSE
## 1        TRUE     FALSE      FALSE     FALSE FALSE  FALSE     TRUE
## 1        TRUE     FALSE       TRUE     FALSE FALSE  FALSE    FALSE
## 2        TRUE      TRUE       TRUE     FALSE FALSE  FALSE    FALSE
## 2        TRUE      TRUE      FALSE     FALSE FALSE  FALSE     TRUE
## 2        TRUE      TRUE      FALSE     FALSE  TRUE  FALSE    FALSE
## 3        TRUE      TRUE       TRUE     FALSE FALSE  FALSE     TRUE
## 3        TRUE      TRUE       TRUE     FALSE  TRUE  FALSE    FALSE
## 3        TRUE      TRUE       TRUE      TRUE FALSE  FALSE    FALSE
## 4        TRUE      TRUE       TRUE     FALSE  TRUE  FALSE     TRUE
## 4        TRUE      TRUE       TRUE      TRUE FALSE  FALSE     TRUE
## 4        TRUE      TRUE       TRUE     FALSE FALSE   TRUE     TRUE
## 5        TRUE      TRUE       TRUE      TRUE  TRUE  FALSE     TRUE
## 5        TRUE      TRUE       TRUE     FALSE  TRUE   TRUE     TRUE
## 5        TRUE      TRUE       TRUE      TRUE FALSE   TRUE     TRUE
## 6        TRUE      TRUE       TRUE      TRUE  TRUE   TRUE     TRUE
##
## $rsq
##  [1] 0.09098910 0.06553884 0.03772155 0.17684068 0.15511530 0.09847659
##  [7] 0.22735839 0.18513393 0.18265030 0.23580382 0.23254757 0.22757123
## [13] 0.24184229 0.23585261 0.23272757 0.24187061
##
## $rss
##  [1] 5264467695 5411860930 5572962645 4767264752 4893085664 5221104448
##  [7] 4474695296 4719235022 4733618792 4425784199 4444642575 4473462646
## [13] 4390812899 4425501674 4443600110 4390648869
```

The Mallows CP tests indicate that the most optimal subset of predictor variables for predicting our response variable is the inclusive set, i.e. we will retain all of our predictor variables from the base model. This is perhaps because of the low predictive power of any individual variable resulting in a lack of push towards a reduced subset with a few predictors, and instead a larger set of predictors as shown above.

Question 5

Test for multicollinearity using VIF on the model from (4) . Based on the test, remove any appropriate variables, and estimate a new regression model based on these findings.

We can see that these all have very high Mallows CP values, indicating that they are poor replacements for the multiple_regression_clean model.

Now let us run the VIF test on the chosen model (multiple_regression_clean), in order to test for multicollinearity.

VIF Test

```
vif(multiple_regression_clean)
```

```
##  CPS1988_clean$education CPS1988_clean$experience  CPS1988_clean$ethnicity
##                 1.067605                 1.064451                 1.011929
##       CPS1988_clean$smsa     CPS1988_clean$region   CPS1988_clean$parttime
##                 1.024924                 1.011145                 1.012768
```

We observe VIF return values near 1 for all of the explanatory variables, implying that there is no significant multicollinearity between variables, and therefore that we do not need to exclude any for that reason.

Question 6

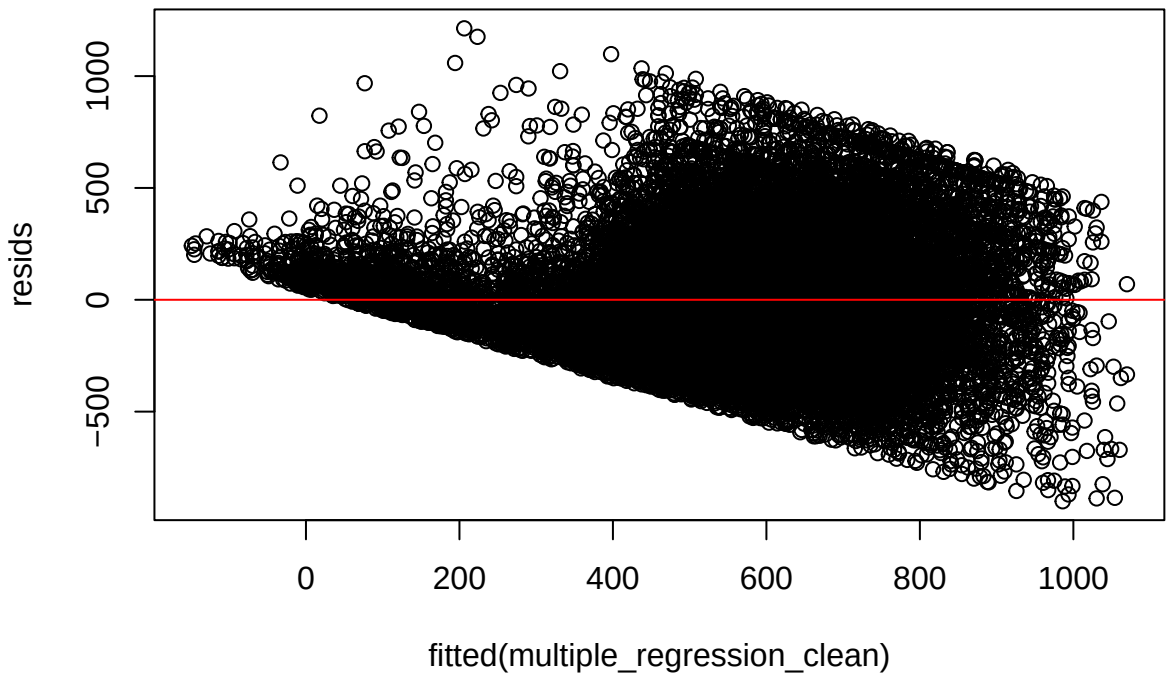
For your model in part (5) plot the respective residuals vs. predicted y and comment on your results.

Residual Plot

```
resids <- resid(multiple_regression_clean)
summary(resids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -900.59 -175.37  -24.97    0.00  148.77 1213.50
```

```
plot(fitted(multiple_regression_clean), resids)
abline(0, 0, col = "Red")
```



We can observe that the residual variance appears to increase as a function of the predicted wage. However, one reason for this is a skew factor - the negative residuals cannot be larger than the predicted wage, as there cannot be a wage below zero. Even accounting for this, the residuals appear to follow either a linear or an exponential heteroskedastic function, i.e. variance as a function of an explanatory variable X or X squared. Several explanations could exist for the potential heteroskedasticity of our data, including the cross-sectional nature of wage analysis, the reality that several possible explanatory variables are not addressed in this study, as well as other factors.

Question 7

For your model in part (5) perform a RESET test and comment on your results.

RESET Test

```
resettest(multiple_regression_clean, power = 2, type = "fitted",
    data = CPS1988_clean)
```

```
##
##  RESET test
##
## data:  multiple_regression_clean
## RESET = 190.09, df1 = 1, df2 = 26156, p-value < 2.2e-16
```

```r
# The test indicates that our model is misspecified:
# p-value ~0. Let us try with a lognormal model, taking the
# log of wage.


# Let us also add columns for the log of wage into the data
# set in case of incorrect model specification.
CPS1988_clean$wage_log <- log(CPS1988_clean$wage)
lognormal_regression <- lm(CPS1988_clean$wage_log ~ CPS1988_clean$education +
    CPS1988_clean$experience + CPS1988_clean$ethnicity + CPS1988_clean$smsa +
    CPS1988_clean$region + CPS1988_clean$parttime)
resettest(lognormal_regression, power = 2, type = "fitted", data = CPS1988_clean)
```

```
##
##  RESET test
##
## data:  lognormal_regression
## RESET = 3.649, df1 = 1, df2 = 26156, p-value = 0.05611
```

```r
# The test indicates that our lognormal model is slightly
# less misspecified, but still the value is effectively
# zero, indicating misspecification.
```

The RESET test indicates model misspecification, however it is unclear how to rectify this given the data. The response variable, wage, appears to follow a lognormal distribution, but the lognormal model does not pass the RESET test as its p-value is still nearly zero. Therefore, something else must be causing the misspecification of this model.

Question 8

For your model in part (5) test for heteroskedasticity and comment on your results. If you identify heteroskedasticy, make sure to account for it before moving on to (9).

```r
gq_output <- gqtest(multiple_regression_clean, point = 0.5, fraction = 0,
    alternative = "two.sided", order.by = NULL)
gq_output
```

```
##
##  Goldfeld-Quandt test
##
## data:  multiple_regression_clean
## GQ = 1.0362, df1 = 13075, df2 = 13075, p-value = 0.04205
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
# The GQ test tests the null hypothesis of homoskedasticity
# with a given alternative hypothesis, in this case a
# two-tailed not equal to - hypothesis.
```

We can see from the p-value of 0.3576 that the null hypothesis is **not** rejected at a significance level of 0.05. Therefore we fail to reject the null and cannot claim that the residuals are heteroskedastic.

This is interesting when compared to our visual observations from the residual plot, where it appeared that heteroskedasticity was possible. However, the visual appearance of the plot affected by the floor of wage at zero clearly is somewhat misleading, considering that the Goldfield-Quant test for heteroskedasticity failed to reject the null hypothesis for homoskedasticity.

Question 9

Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison.

Let us create dummy variables for all of our Y/N explanatory variables

Indicator (Dummy) Variables

dummy for region

```
CPS1988_clean$region_ne <- ifelse(CPS1988_clean$region == "northeast",
    1, 0)
CPS1988_clean$region_west <- ifelse(CPS1988_clean$region == "west",
    1, 0)
CPS1988_clean$region_midwest <- ifelse(CPS1988_clean$region ==
    "midwest", 1, 0)
CPS1988_clean$region_south <- ifelse(CPS1988_clean$region ==
    "south", 1, 0)
```

dummy for ethnicity

```
CPS1988_clean$ethnicity_cauc <- ifelse(CPS1988_clean$ethnicity ==
    "cauc", 1, 0)
CPS1988_clean$ethnicity_afam <- ifelse(CPS1988_clean$ethnicity ==
    "afam", 1, 0)
```

dummy smsa

```
CPS1988_clean$smsa_y <- ifelse(CPS1988_clean$smsa == "yes", 1,
    0)
CPS1988_clean$smsa_n <- ifelse(CPS1988_clean$smsa == "no", 1,
    0)
```

dummy parttime

```
CPS1988_clean$parttime_y <- ifelse(CPS1988_clean$parttime ==
    "yes", 1, 0)
CPS1988_clean$parttime_n <- ifelse(CPS1988_clean$parttime ==
    "no", 1, 0)
```

We created dummy variables for the various possible outcomes of region, ethnicity, smsa, and partime - adding additional columns to classify every individual in the dataset

```
indicator_regression <- lm(CPS1988_clean$wage ~ CPS1988_clean$education +
    CPS1988_clean$experience + CPS1988_clean$parttime_y + CPS1988_clean$parttime_y +
    CPS1988_clean$smsa_y + CPS1988_clean$smsa_n + CPS1988_clean$ethnicity_cauc +
    CPS1988_clean$ethnicity_afam + CPS1988_clean$region_south +
    CPS1988_clean$region_midwest + CPS1988_clean$region_ne +
    CPS1988_clean$region_west)
summary(indicator_regression)
```

```
##
## Call:
## lm(formula = CPS1988_clean$wage ~ CPS1988_clean$education + CPS1988_clean$experience +
##     CPS1988_clean$parttime_y + CPS1988_clean$parttime_y + CPS1988_clean$smsa_y +
##     CPS1988_clean$smsa_n + CPS1988_clean$ethnicity_cauc + CPS1988_clean$ethnicity_afam +
##     CPS1988_clean$region_south + CPS1988_clean$region_midwest +
##     CPS1988_clean$region_ne + CPS1988_clean$region_west)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -990.03 -191.29  -24.11  171.48 1184.54
##
## Coefficients: (9 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -267.1483    10.7461  -24.86   <2e-16 ***
## CPS1988_clean$education         50.8169     0.7383   68.83   <2e-16 ***
## CPS1988_clean$experience        8.7711     0.1387   63.26   <2e-16 ***
## CPS1988_clean$parttime_y            NA         NA      NA       NA
## CPS1988_clean$smsa_y                NA         NA      NA       NA
## CPS1988_clean$smsa_n                NA         NA      NA       NA
## CPS1988_clean$ethnicity_cauc        NA         NA      NA       NA
## CPS1988_clean$ethnicity_afam        NA         NA      NA       NA
## CPS1988_clean$region_south          NA         NA      NA       NA
## CPS1988_clean$region_midwest        NA         NA      NA       NA
## CPS1988_clean$region_ne             NA         NA      NA       NA
## CPS1988_clean$region_west           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279.2 on 26161 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2153
## F-statistic:  3591 on 2 and 26161 DF,  p-value: < 2.2e-16
```

Here we run the AIC test using the indicator_regression, original multiple_regression_clean, and the "best" of the mallows CP models, cp_model_3. The AIC test identifies the best fitting model out of a series of models and provides descriptive statistics of the relative predictive power of the input models.

AIC Test

```
library(AICcmodavg)
models <- list(indicator_regression, multiple_regression_clean,
    cp_modeL_3)
mod.names <- c("edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west",
```

```
    "edu, exp, eth, smsa, reg, pt", "exp, edu, pt, smsa")
a <- aictab(cand.set = models, modnames = mod.names)
print(a)
```

```
##
## Model selection based on AICc:
##
##                                                                                  K
## edu, exp, eth, smsa, reg, pt                                                     8
## exp, edu, pt, smsa                                                               6
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west 4
##                                                                                      AICc
## edu, exp, eth, smsa, reg, pt                                                     364586.2
## exp, edu, pt, smsa                                                               364902.3
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west 368959.3
##                                                                                  Delta_AICc
## edu, exp, eth, smsa, reg, pt                                                           0.00
## exp, edu, pt, smsa                                                                   316.13
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west  4373.06
##                                                                                  AICcWt
## edu, exp, eth, smsa, reg, pt                                                          1
## exp, edu, pt, smsa                                                                    0
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west    0
##                                                                                  Cum.Wt
## edu, exp, eth, smsa, reg, pt                                                          1
## exp, edu, pt, smsa                                                                    1
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west    1
##                                                                                      LL
## edu, exp, eth, smsa, reg, pt                                                     -182285.1
## exp, edu, pt, smsa                                                               -182445.2
## edu, exp, pt_y, pt_n, smsa_y, smsa_n, eth_cauc, eth_afam, south, midwest, ne, west -184475.6
```

The indicator_regression model has the lowest AIC value and therefore is implied as the "best fitting" model of the three. However, we can clearly see that AIC values are very high in this case, and that this can be related to the indication of model misspecification by the RESET test. Interestingly, it appears for all of the tests that the indicator_regression model and the multiple_regression_clean model are returning the same values, and as well we get a warning about model redundancy. This may indicate that adding the dummy variables has no affect on the model fit, precision, or other similar factors.

BIC Test

```
library(flexmix)
BIC(indicator_regression)
```

```
## [1] 368992
```

```
BIC(multiple_regression_clean)
```

```
## [1] 364651.6
```

```
BIC(cp_modeL_3)
```

```
## [1] 364951.4
```

Again, indicator_regression and multiple_regression_clean both have the lowest (and very high) BIC values. Therefore, of our options, we will choose indicator_regression as our model of choice.

Question 11

1 paragraph summary

Our project aimed to analyze a dataset and generate a model predicting wage based on a set of continuous and discrete possible explanatory variables. Ultimately, our model did not have significant predictive power as evidenced by, among other things: Low R^2 statistics, ~0 p-values from the RESET score, implying model misspecification, and high AIC and BIC return values. While our initial guess at the source of the misspecification errors was the fact that the distribution of wage in the data set appears to follow a right-skewed lognormal distribution, running the RESET test on the log-adjusted model still returned a p-value ~0 (*although slightly better than the general linear model*) , which indicates that there are other causes of this issue. Some ideas we brainstormed that could cause the above indicators of poor predictive power are: General randomness and/or high variance in the data set (this could be related to our findings of homoskedasticity with a GQ test p value of .3276, which could indicate some random variance not correlated with any predictive variable), lognormal explanatory variable distributions that could provide better fit with a linear (or log-adjusted) response variable, poor predictive power of any of the variables on wage as an exogenous observation, and others.

Ultimately, going through this process was a very interesting task - we learned many new statistical methods for model evaluation and understanding the rigorous, step-by-step process of filtering through model possibilities was rewarding. It would be good to further explore the CPS1988 dataset and attempt all possible model combinations and functional form specifications to understand where the source of variance and the source of misspecification in the linear model lies.