# CSCI E-106:Assignment 1

Shreya Bajpai

2024-09-06

## Problem 1: teengamb

The dataset teengamb concerns a study of teenage gambling in Britain. You can download this data set by installing faraway library. To get the data set, copy and paste the r command: install.packages("faraway");library(faraway); data(teengamb, package="faraway"). (40 points)

The list variables are described below:

sex:0=male, 1=female

status: Socioeconomic status score based on parents' occupation

income: in pounds per week

verbal: verbal score in words out of 12 correctly defined

gamble: expenditure on gambling in pounds per year

As per the instructions, we first import the library, faraway, and extract the data from the dataset: teengamb.

```
library(faraway)
library(ggplot2)
data(teengamb, package="faraway")
```

### Problem 1A

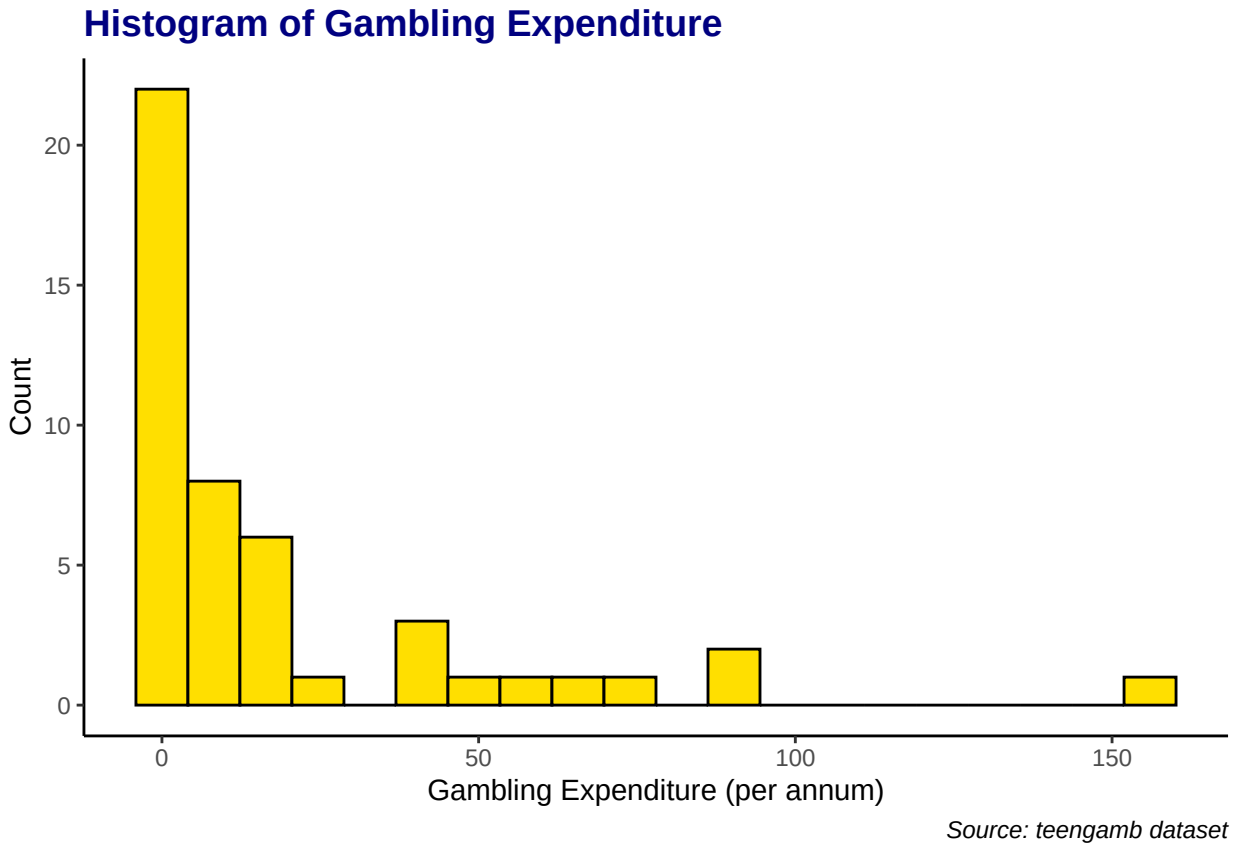
**We are interested in predicting the expenditure on gambling. What is the dependent variable? and What are the independent variables? (10 points)**

If we intend to predict the expenditure on gambling, then the **dependent variable** is `gamble`, which is the focus of the study and the variable we expect will change based on the variations in the **independent variable(s)**: `sex`, `socioeconomic status`, `income` and `verbal score`.

Viewing the dependent variable's, `gamble`, histogram shows the spread (variability) and central tendency (mean, median) of the dependent variable which gives insight into how values are distributed across the range for this sample. The histogram below depicts that our dependent variable is **right-skewed** indicating that the linear regression model, which assumes normality of distribution in residuals (errors), might struggle to capture the pattern, leading to higher residuals (errors) and lower predictive accuracy if a linear regression model is applied. However, as part of this analysis, we do not intend to implement linear regression, but this is interesting to capture before we dive in.

```
ggplot(teengamb, aes(gamble)) +  geom_histogram(color = "#000000", fill = "#FFDF00", bins=20) + labs(
    title = "Histogram of Gambling Expenditure",
    caption = "Source: teengamb dataset",
    x = "Gambling Expenditure (per annum)",
    y = "Count"
  ) + theme_classic() +
  theme(
```

```
    plot.title = element_text(color = "#000080", size = 14, face = "bold"),
    plot.subtitle = element_text(size = 10, face = "bold"),
    plot.caption = element_text(face = "italic")
  )
```




**Problem 1B**

**Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data. (30 points)**

To understand the data of 47 observations (rows) of 5 variables (columns) (`sex`, `status`, `income`, `verbal` and `gamble`), we run a summary on the data set. The numerical summary shows the dataset's mean, median, minimumns, maximums, and more for each given variable.

```
str(teengamb)
```

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```
summary(teengamb)
```

```
##       sex             status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
```

```
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

At first glance, we see a few discrepancies with the data:

1. `income` is in pounds (£) per week, but `gamble` is in pounds (£) per year, which prevents us from comparing the two variables on the same cadence as the timeline is different. As a result, when we consider the proportion of weekly income to gambling expendtiures, we can consider multiplying it by 52 (weeks in a year) to gain perspective on the proportion of total annual income spent on annual gambling expenditure.
2. `sex` is not a quantitative variable, so the numeric summary details in the raw form for this variable are not informative. To turn a binary variable, sex, from a quantitative to a categorical variable, we can use the factor() function.

I intend to extract insights on basis of `sex`, a categorical variable, so I distribute the dataset into a male and female dataset and note whether there are differences or biases in the relationship between `sex` and the other independent variables.

```
teengamb$sex <- factor(teengamb$sex)
levels(teengamb$sex) <- c("male", "female")
gamb_male <- subset(teengamb, sex == "male")
gamb_female <- subset(teengamb, sex == "female")
```

1. First, we explore the *socioeconomic status* for males in contrast to females and we see that the median and mean socioeconomic status score based on parents' occupation is higher for males (Median: 51.00, μ: 52.00) than females (Median: 30.00 μ: 35.26). We need to be mindful that there is a greater number of males than females in this sample, but this contrast is stark.
2. Second, we explore *income*, which we see is not as stark for the medians of the two genders (Male Median: £3.38/week (~ £175.76 annual income per year), Female Median: £3.00/week (~ £156.00 annual income per year)) despite the sample containing more males than females and the assumption that males generally earn more than females.
3. Third, we examine the *verbal score in words out of 12 correctly defined* which showcases that males and females perform similarly when considering the median of both datasets (Male Median: 7.000, Female Median: 6.000).
4. Finally, we examine how the *sex* of an individual in this sample impacts the dependent variable, gambling, and the observations are worth examining. There is a notable difference between all metrics when comparing Males and Females, but the primary being the effect of the gender of the sample on the median and mean of gambling expenditures—in males, £14.25 (Median) and £29.78 (μ) per year and in females, £1.70 (Median) and £3.87 (μ) per year.

```
summary(gamb_female)
```

```
##      sex         status          income          verbal          gamble
##  male  : 0   Min.   :18.00   Min.   : 1.500   Min.   :4.000   Min.   : 0.000
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.:6.000   1st Qu.: 0.100
##              Median :30.00   Median : 3.000   Median :6.000   Median : 1.700
```

```
##              Mean   :35.26   Mean   : 4.149   Mean   :6.421   Mean   : 3.866
##              3rd Qu.:43.00   3rd Qu.: 5.750   3rd Qu.:8.000   3rd Qu.: 6.000
##              Max.   :65.00   Max.   :10.000   Max.   :8.000   Max.   :19.600
```

```
summary(gamb_male)
```

```
##      sex         status          income          verbal
##  male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.000
##  female: 0   1st Qu.:38.00   1st Qu.: 2.000   1st Qu.: 6.000
##              Median :51.00   Median : 3.375   Median : 7.000
##              Mean   :52.00   Mean   : 4.976   Mean   : 6.821
##              3rd Qu.:65.25   3rd Qu.: 6.625   3rd Qu.: 8.250
##              Max.   :75.00   Max.   :15.000   Max.   :10.000
##      gamble
##  Min.   :  0.000
##  1st Qu.:  2.775
##  Median : 14.250
##  Mean   : 29.775
##  3rd Qu.: 42.175
##  Max.   :156.000
```

```
summary(teengamb)
```

```
##      sex         status          income          verbal          gamble
##  male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
##              Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
##              Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
##              3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
##              Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

This is a telling analysis as it is, but I propose using linear regression to identify the most significant variables to explore. We will use the `lm()` function to fit a simple linear regression model, with `gamble` as the response and the four independent variables, `sex`, `status`, `income` and `verbal` as the predictors. The basic syntax is `lm(y ~ x, data)`, where `y` is the response, `x` is the predictor, and `data` is the data set in which these two variables are kept.

```
fit_gamb <- lm(gamble ~ sex+status+income+verbal, data=teengamb)
summary(fit_gamb)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sexfemale   -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```
fit_gamb$coefficients
```

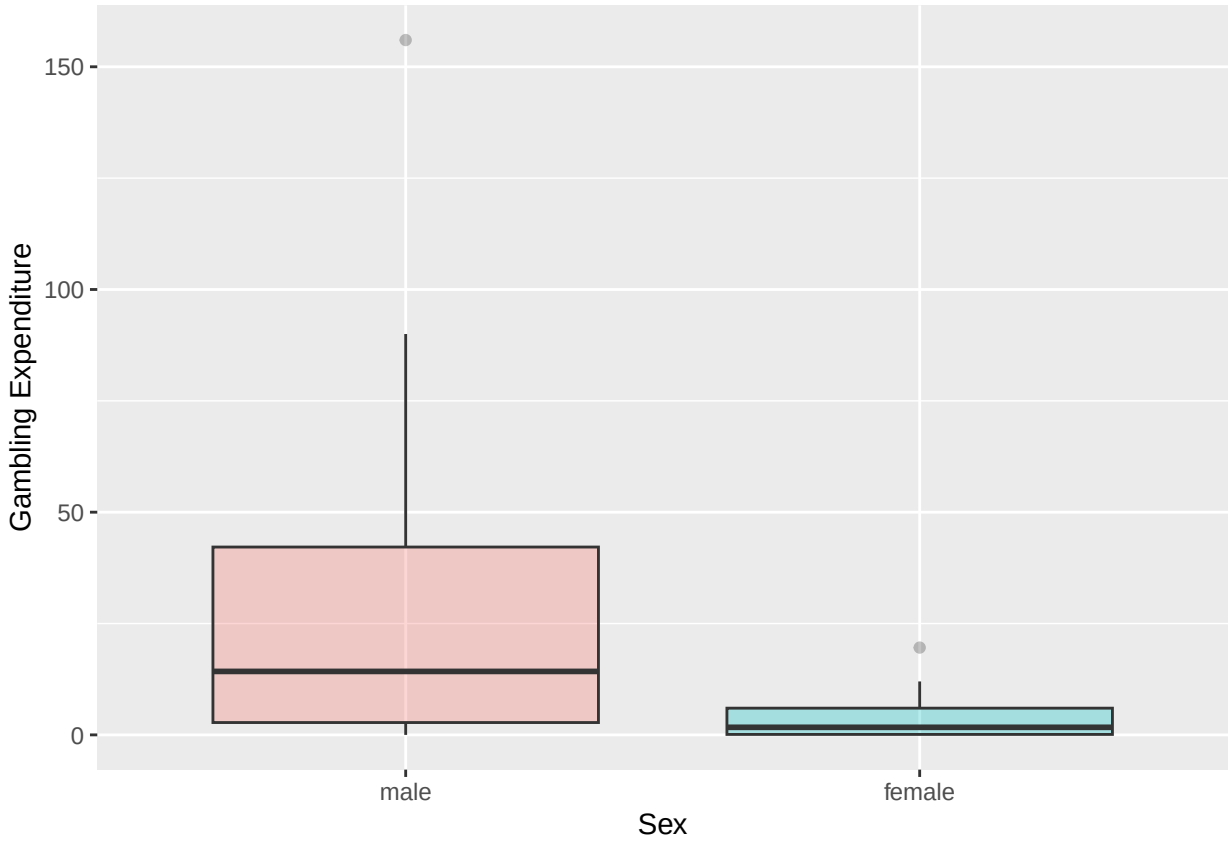
```
##  (Intercept)    sexfemale       status       income       verbal
##  22.55565063 -22.11833009   0.05223384   4.96197922  -2.95949350
```

The results above showcase that **income** and **sex** are the two most significant independent variables as they have the least P values (*the probability of observing a greater absolute value of t under the null hypothesis*), with the former being the more significant independent variable to predict gambling expenditure. Now, I will examine both of these independent variable's effect visually on gambling expenditures.

The first variable I want to explore is: **sex** via the basic function ggplot().

- The box plot shows the median level of expenditure on gambling of male is much higher than a female (*represented by the thick black line in each box*). Recall, the summary values told the same story.
- Further, the variability of expenditure on gambling of a male is much higher than the variability of the one of female (*represented by the length of the box*), with the outliers also being vastly far apart in the two genders of the sample.

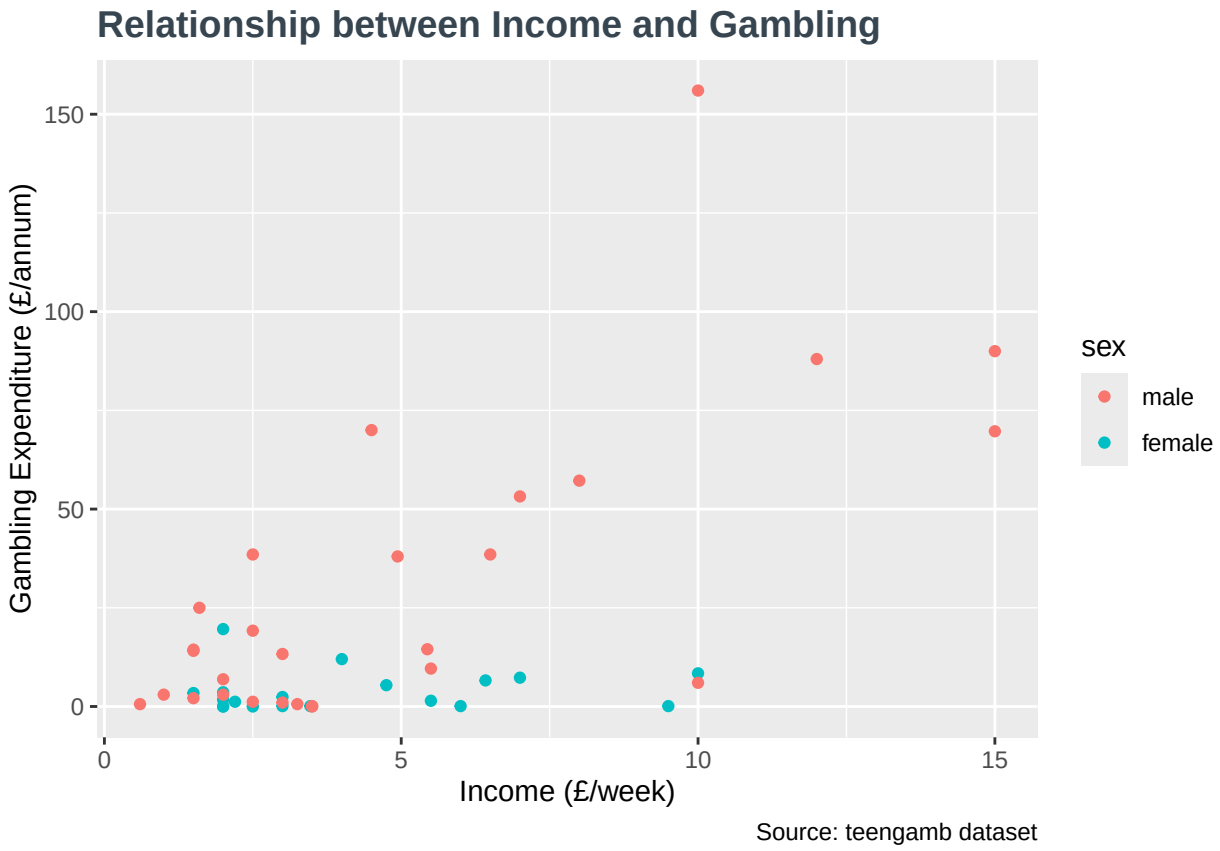
```
ggplot(data=teengamb, mapping = aes(x = sex, y = gamble, fill=sex)) +
   geom_boxplot(alpha=0.3) +
    theme(legend.position="none") + xlab("Sex") + ylab("Gambling Expenditure")
```




Next, we examine the weekly income of the sample and its impact on annual gambling expenditure. We show the distinction of the sex in the data points plotted to visualize what we observed above about gender through the summary() of the malegamb and femalegamb datasets.

Based on our sample and this plot, we can see that females overall have lower annual gambling expenditures than males. Many of the extreme values or outliers are from males, with women on average, regardless of income, spending less than men on gambling expenditures. As per the observations from the histogram, we can see that the cluster of the sample data points are concentrated to the bottom left of the graph, showing the right skew when the dependent variable, `gamble`, is plotted. Visually, we can conclude that plotting a linear regression fit line will not accurately predict the outliers or the right-skewed cluster that we observe in this dataset.

```
ggplot(data=teengamb, aes(x = income, y = gamble, color=sex)) +
    geom_point() +
    labs(
    title = "Relationship between Income and Gambling",
    caption = "Source: teengamb dataset",
    x = "Income (£/week)",
    y = "Gambling Expenditure (£/annum)") +
    theme(plot.title = element_text(color = "#36454F", size = 14, face = "bold"))
```




Our endeavor to explain this data using income and sex as the two most significant variables proved to provide insights worth examining in this sample. The discernment gathered from separating the sample based on its one factor variable, `sex`, provided a closer look at other independent variables that may not have otherwise been observed if we overlooked the significance of this independent variable in conjunction to the others (independent variables) when observing gambling expenditures.

## Problem 2: uswages

The dataset uswages is drawn as a sample from the Current Population Survey in 1988. You can download this data set by installing faraway library. To get the data set, copy and paste the r command: install.packages("faraway"); data(uswages, package="faraway"). (60 points, 10 points each)

The wage is the response variable. Please see below for the full list of variables.

wage: Real weekly wages in dollars (deflated by personal consumption expenditures - 1992 base year)

educ:Years of education

exper:Years of experience

race:1 if Black, 0 if White (other races not in sample)

smsa:1 if living in Standard Metropolitan Statistical Area, 0 if not

ne:1 if living in the North East

mw:1 if living in the Midwest

we:1 if living in the West

so:1 if living in the South

pt:1 if working part time, 0 if not

### Problem 2A

**How many observations are in the data set?**

There are 2000 observations of 10 variables (`wage`, `educ`, `expr`, `race`, `smsa`, `ne`, `mw`, `so`, `pt`).

```
data(uswages, package="faraway")
str(uswages)
```

```
## 'data.frame':    2000 obs. of  10 variables:
##  $ wage : num  772 617 958 617 902 ...
##  $ educ : int  18 15 16 12 14 12 16 16 12 12 ...
##  $ exper: int  18 20 9 24 12 33 42 0 36 37 ...
##  $ race : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smsa : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ ne   : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ mw   : int  0 0 0 0 1 0 0 1 0 1 ...
##  $ so   : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ we   : int  0 1 0 0 0 1 0 0 1 0 ...
##  $ pt   : int  0 0 0 0 0 0 1 1 1 0 ...
```

### Problem 2B

**Calculate the mean and median of each variable. Are there any outliers in the data set?**

When we calculate the mean and median of each of the data elements, we extract the numerical summary below:

```
summary(uswages)
```

```
##       wage              educ           exper            race       
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000  
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000  
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000  
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078  
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000  
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000  
##       smsa             ne              mw              so        
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
```

```
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we              pt
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.00   Median :0.0000
##  Mean   :0.21   Mean   :0.0925
##  3rd Qu.:0.00   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000
```

What is odd is that variable `exper` has a negative minimum value of -2.00. Executing a pull of the values of this variable results in the following. While it is permissible for a sample individual to have 0 years of work experience, we do not expect negative values of work experience to be applicable. This may be due to incorrect sampling, but we pursue the analysis knowing this **data quality constraint** when building the regression model.

```
head(sort(uswages$exper,decreasing=FALSE), n = 50)
```

```
##  [1] -2 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
## [26] -1 -1 -1 -1 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

In addition, there seem to be a handful of categorical variables (`race`, `smsa`, `pt`, `ne`, `mw`, `so`, `we`, `so`, `pt`) which have binary values ('Yes', 'No'; 'Black', White') for which a numerical summary is not ideal to depict understanding of these elements as the "mean" of a binary/factor is difficult to decipher, so I proceed to categorize them as factors to make them easier to decipher in our analysis.

```
# Factor the categorical variables
uswages$race <- factor(uswages$race)
levels(uswages$race) <- c("White","Black")
uswages$smsa <- factor(uswages$smsa)
levels(uswages$smsa) <- c("No","Yes")
uswages$pt <- factor(uswages$pt)
levels(uswages$pt) <- c("No","Yes")
uswages$ne <- factor(uswages$ne)
levels(uswages$ne) <- c("No","Yes")
uswages$mw <- factor(uswages$mw)
levels(uswages$mw) <- c("No","Yes")
uswages$we <- factor(uswages$we)
levels(uswages$we) <- c("No","Yes")
uswages$so <- factor(uswages$so)
levels(uswages$so) <- c("No","Yes")
uswages$pt <- factor(uswages$pt)
levels(uswages$pt) <- c("No","Yes")
```

### Problem 2C

**Calculate the correlation among wage,education and experience. Plot each of the predictors against the response variable. Identify the variables that are strongly correlated with the response variable.**

To identify the *correlation among the three variables*: wage (*independent variable*), years of education and years of experience (*dependent variables*), we can use the `cor()` function.

We see from the results below that there is a higher correlation between wages and years of education (0.248) than there is between wages and years of experience (0.183). There is also a negative correlation between

years of education and years of experience (rround(cor(uswages*educ*, *uswages*exper), 3)‘).

```
# Plot the correlation between the three variables mentioned
round(cor(uswages$wage, uswages$educ), 3)
```

```
## [1] 0.248
```

```
round(cor(uswages$wage, uswages$exper), 3)
```

```
## [1] 0.183
```

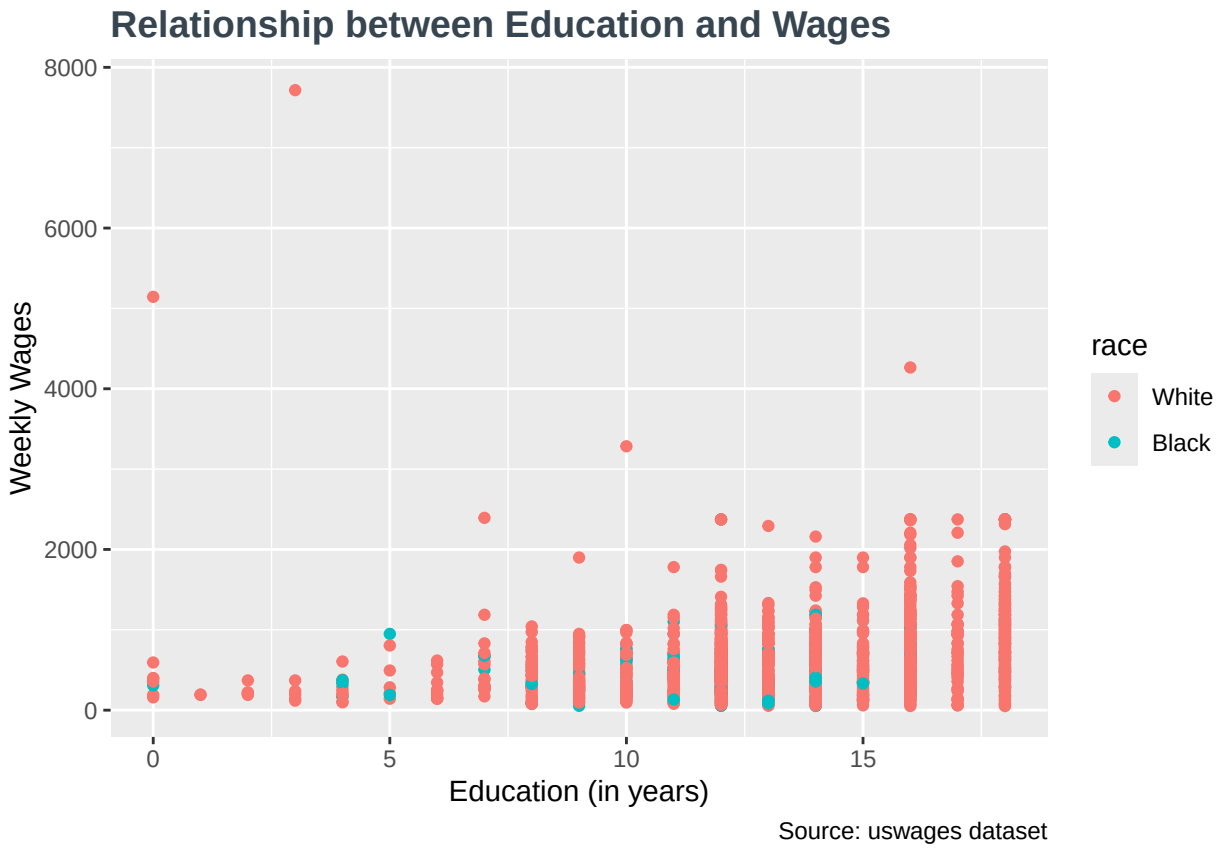
```
round(cor(uswages$educ, uswages$exper), 3)
```

```
## [1] -0.302
```

We plot the first relationship between **wage** as the response variable and **education (in years)** as the predictor variable.

The data shows with more years of education there is a gradual increase in wages, indicating a positive relationship that is depicted by the correlation coefficient of 0.248.

```
ggplot(data=uswages, aes(x = educ, y = wage, color=race)) +
    geom_point() +
    labs(
    title = "Relationship between Education and Wages",
    caption = "Source: uswages dataset",
    x = "Education (in years)",
    y = "Weekly Wages") +
    theme(plot.title = element_text(color = "#36454F", size = 14, face = "bold"))
```

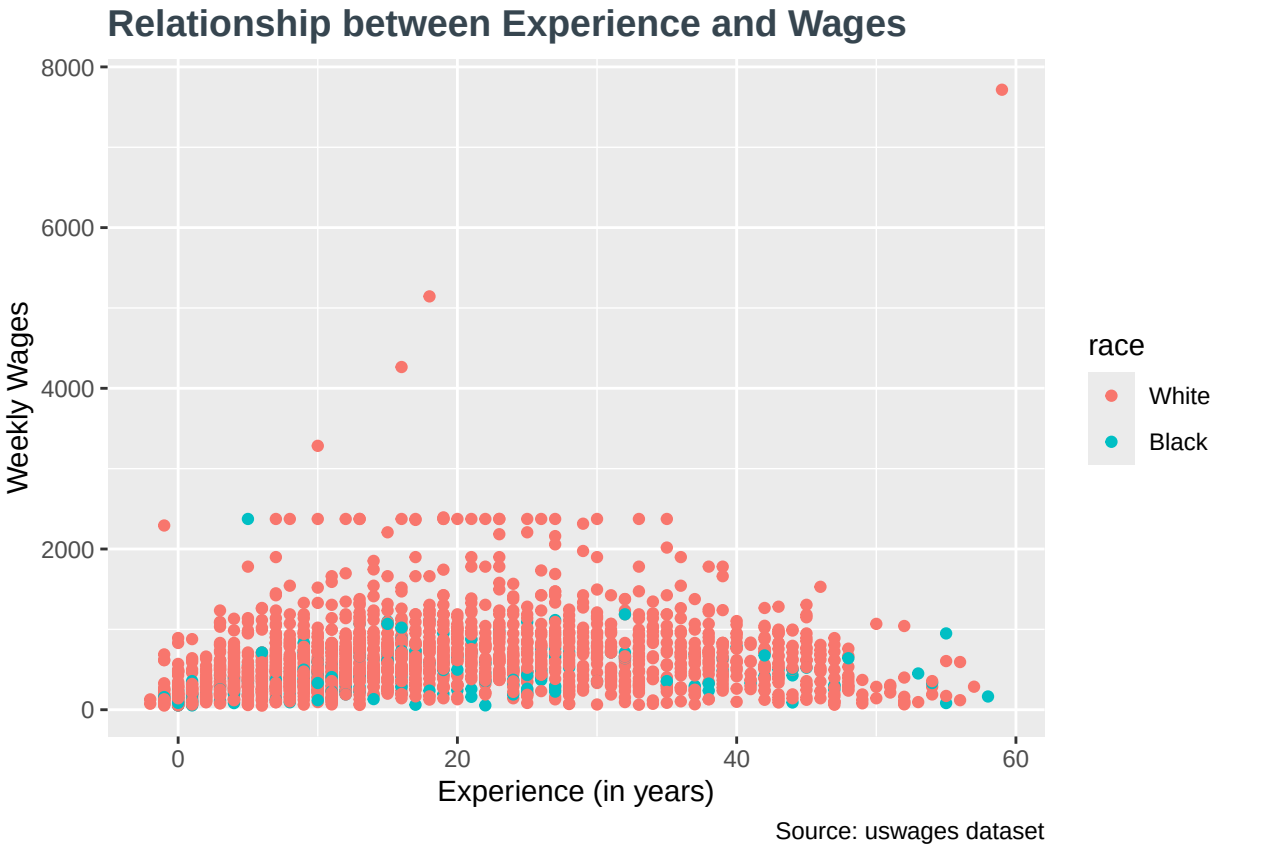



We plot the second relationship between **wage** as the response variable and **experience (in years)** as the

predictor variable.

The data here is a bit hard to decode but reflects the understanding that most individuals earn the average salary for a given job for their entire career, with fewer individuals earning at the top range of the projected salary range for a given job. Years of experience appears **not** to have significant impact on weekly wages given the limited variance of data points as experience in years increases.

```
ggplot(data=uswages, aes(x = exper, y = wage, color=race)) +
    geom_point() +
    labs(
    title = "Relationship between Experience and Wages",
    caption = "Source: uswages dataset",
    x = "Experience (in years)",
    y = "Weekly Wages") +
    theme(plot.title = element_text(color = "#36454F", size = 14, face = "bold"))
```

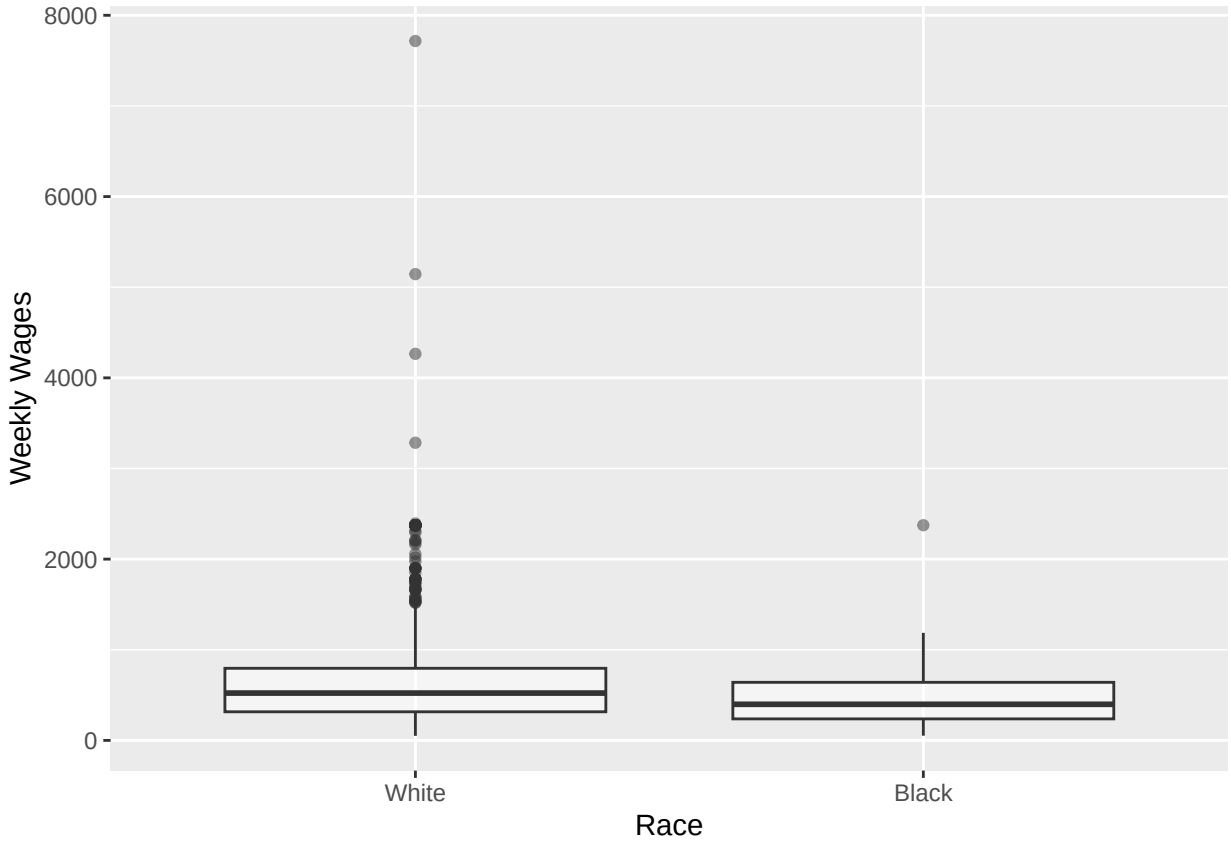



### Problem 2D

### Is there difference in wages based on race?

There is a bias in the dataset with the sample consisting of 1,812 White race observations and 155 Black race observations (with no other races captured). The box plot below shows that there is a higher variability in the wages earned weekly by an individual of White race as opposed to an individual of Black race in this sample, as noted by the many outliers between the upper quartile and upper extreme of the 'White' sample box in comparison to the one outlier of the 'Black' sample box.

```
ggplot(data=uswages, mapping = aes(x = race, y = wage)) +
   geom_boxplot(alpha=0.5) +
    theme(legend.position="none") + xlab("Race") + ylab("Weekly Wages")
```




However, when we disperse the sample into two distinct subsets on basis of `race`, we see the vast differences clearly that are seen above in the box plot. What is interesting is that between the two subsets of the population, the μ of `educ` and `exper` is similar, however the range (min to max) of weekly wages exhibits a substantial disparity, with wages for 'White' race showing a much wider range compared to the other.

While we have a larger sample population in the 'White' subset, it is notable that this does not affect the proportion of the sample of each race's statistics on working part-time, with the majority of samples working full-time regardless of their race (White: Yes: 167, No: 1677; Black: Yes: 18, No: 138).

```
uswages_blk <- subset(uswages, race == "Black")
uswages_wte <- subset(uswages, race == "White")
summary (uswages_blk)
```

```
##       wage             educ           exper         race      smsa
##  Min.   :  52.23   Min.   : 0.00   Min.   :-1.00   White:  0   No : 29
##  1st Qu.: 237.42   1st Qu.:11.75   1st Qu.: 9.75   Black:156   Yes:127
##  Median : 398.46   Median :12.00   Median :17.50
##  Mean   : 456.04   Mean   :12.11   Mean   :19.83
##  3rd Qu.: 641.03   3rd Qu.:14.00   3rd Qu.:27.00
##  Max.   :2374.15   Max.   :18.00   Max.   :58.00
##    ne        mw        so        we        pt
##  No :134   No :130   No :61   No :143   No :138
##  Yes: 22   Yes: 26   Yes:95   Yes: 13   Yes: 18
##
##
##
##
```

```
summary (uswages_wte)
```

```
##       wage              educ          exper          race       smsa
##  Min.   :  50.39   Min.   : 0.0   Min.   :-2.00   White:1844   No : 459
##  1st Qu.: 315.81   1st Qu.:12.0   1st Qu.: 8.00   Black:   0   Yes:1385
##  Median : 522.32   Median :12.0   Median :15.00
##  Mean   : 620.98   Mean   :13.2   Mean   :18.29
##  3rd Qu.: 795.59   3rd Qu.:16.0   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.0   Max.   :59.00
##    ne         mw         so         we         pt
##  No :1408   No :1373   No :1314   No :1437   No :1677
##  Yes: 436   Yes: 471   Yes: 530   Yes: 407   Yes: 167
##
##
##
##
```

We examine whether living in a Standard Metropolitan Statistical Area further cements the differences we have seen in race. We see that the majority of the sample resides in the `smsa` area and this is where we see the outlier of wage(s) present. This corroborates the assumption most have that `smsa` areas being more desirable to find work in as there may be more opportunities in these area that pay more as opposed to other areas across the state.

```
uswages_smsa <- subset(uswages, smsa == "Yes")
uswages_nonsmsa <- subset(uswages, smsa == "No")
summary (uswages_smsa)
```

```
##       wage              educ          exper          race       smsa
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   White:1385   No :   0
##  1st Qu.: 333.27   1st Qu.:12.00   1st Qu.: 8.00   Black: 127   Yes:1512
##  Median : 547.47   Median :13.00   Median :15.00
##  Mean   : 643.72   Mean   :13.24   Mean   :18.48
##  3rd Qu.: 830.96   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00
##    ne         mw         so         we         pt
##  No :1121   No :1157   No :1060   No :1198   No :1377
##  Yes: 391   Yes: 355   Yes: 452   Yes: 314   Yes: 135
##
##
##
##
```

```
summary (uswages_nonsmsa)
```

```
##       wage              educ          exper          race      smsa
##  Min.   :  54.61   Min.   : 0.0   Min.   :-1.0   White:459   No :488
##  1st Qu.: 260.80   1st Qu.:12.0   1st Qu.: 8.0   Black: 29   Yes:  0
##  Median : 427.35   Median :12.0   Median :16.0
##  Mean   : 497.80   Mean   :12.7   Mean   :18.2
##  3rd Qu.: 664.77   3rd Qu.:14.0   3rd Qu.:27.0
##  Max.   :2374.15   Max.   :18.0   Max.   :56.0
##    ne         mw         so         we         pt
##  No :421   No :346   No :315   No :382   No :438
##  Yes: 67   Yes:142   Yes:173   Yes:106   Yes: 50
##
##
```

##
##

### Problem 2E

**Build a regression model by using only education to predict the response variable. State the regression model.**

Our linear model using *only* education as the predictor accounts for 6.167% of the variation in the response of wages. From the model summary, we see that the fitted regression equation is:

$$Wages = 109.754 + 38.011 \times (X\ i\ )$$

where $X_i$ corresponds to the independent variable, *education*.

- This informs that an additional year (or unit) of education is associated with an average increase in wages of $38.01 per week. The y-intercept of 109.75 gives us the average expected weekly income for an individual who has 0 years of education.
- The value of $P$, the probability of finding the given t statistic if the null hypothesis of no relationship were true, for education ($<$ 2e-16) is significantly less than .05 so we can conclude that there is a statistically significant association between wages and education.
- The *Residual Standard Error*, the average distance that the observed values fall from the regression line, is quite high (445.5 on 1998 df) indicating the regression line is *not* able to match the observed data. In this case, the average observed wage falls $445.50 away from the wage predicted by the regression line.
- While $\mathbf{R^2}$ measures the strength of the relationship between our model and the dependent variable, wages, it is not a formal test for the relationship between wages and education. The *F-statistic* is 131.3 which is a measure of how well the regression model fits the data compared to a model with no predictors (the null model).
  - The p-value $<$ 2.2e-16 indicates the probability of observing an F-statistic as large as 131.3 (or larger) under the null hypothesis (i.e., no relationship between the independent and dependent variables) is extremely small. A p-value this small suggests strong evidence to reject the null hypothesis, meaning that the independent variable, education, in our model significantly explains the variation in the dependent variable, wages.

```
fit_wage_edu <- lm(wage ~ educ, data=uswages)
summary(fit_wage_edu)
```

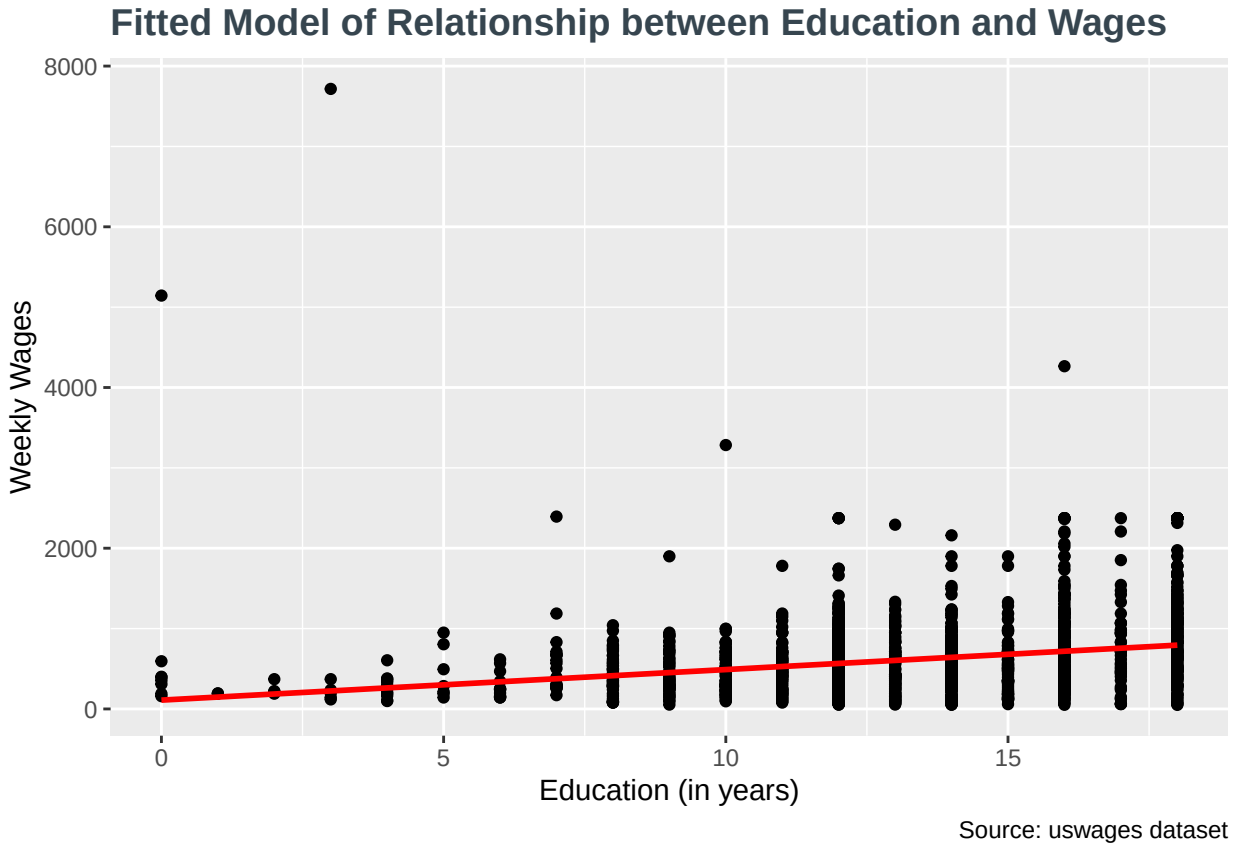
```
##
## Call:
## lm(formula = wage ~ educ, data = uswages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -743.6 -269.5  -67.7  173.0 7492.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  109.754     44.616    2.46    0.014 *
## educ          38.011      3.317   11.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 445.5 on 1998 degrees of freedom
## Multiple R-squared:  0.06167,    Adjusted R-squared:  0.0612
## F-statistic: 131.3 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```
fit_wage_edu$coefficients
```

```
## (Intercept)        educ
##   109.75385    38.01114
```

```
ggplot(data=uswages, aes(x = educ, y = wage)) +
    geom_point() +
    labs(title = "Fitted Model of Relationship between Education and Wages",
    caption = "Source: uswages dataset",
    x = "Education (in years)",
    y = "Weekly Wages") +
    theme(plot.title = element_text(color = "#36454F", size = 14, face = "bold")) + geom_smooth(color='
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




**Problem 2F**

**Build a regression model by using only experience to predict the response variable. State the regression model.**

Our linear model using *only* experience as the predictor accounts for 3.356% of the variation in the response of wages. From the model summary, we see that the fitted regression equation is:

$$Wages = 492.1669 + 6.2981 \times (X\ i\ )$$

where $X_i$ corresponds to the independent variable, *experience*.

- This informs that an additional year (or unit) of experience is associated with an increase in wages of \$6.30 per week. The y-intercept of 492.17 gives us the average expected weekly income for an individual

who has 0 years of experience.

- The value of $P$, the probability of finding the given t statistic if the null hypothesis of no relationship were true, for experience (<2e-16) is significantly less than .05 so we can conclude that there is a statistically significant positive association between wages and experience.
- The *Residual Standard Error*, the average distance that the observed values fall from the regression line, is high (452.2 on 1998 df) indicating the regression line is *not* able to match the observed data well. In this case, the average observed wage falls $452.20 dollars away from the wage predicted by the regression line.
- While $\mathbf{R^2}$ measures the strength of the relationship between our model and the dependent variable, wages, it is not a formal test for the relationship between wages and experience. The *F-statistic* is 69.39 which is a measure of how well the regression model fits the data compared to a model with no predictors (the null model).
  - The p-value < 2.2e-16 indicates the probability of observing an F-statistic as large as 69.39 (or larger) under the null hypothesis (i.e., no relationship between the independent and dependent variables) is extremely small. A p-value this small suggests strong evidence to reject the null hypothesis, meaning that the independent variable, experience, in our model significantly explains the variation in the dependent variable, wages.

```
fit_wage_exper <- lm(wage ~ exper, data=uswages)
summary(fit_wage_exper)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = uswages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -755.5 -271.0  -77.5  165.7 6852.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 492.1669    17.2043   28.61   <2e-16 ***
## exper         6.2981     0.7561    8.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 452.2 on 1998 degrees of freedom
## Multiple R-squared:  0.03356,    Adjusted R-squared:  0.03308
## F-statistic: 69.39 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```
fit_wage_exper$coefficients
```

```
## (Intercept)       exper
##  492.166861    6.298091
```

```
ggplot(data=uswages, aes(x = exper, y = wage)) +
    geom_point() +
    labs(title = "Fitted Model of Relationship between Experience and Wages",
    caption = "Source: uswages dataset",
    x = "Experience (in years)",
    y = "Weekly Wages") +
    theme(plot.title = element_text(color = "#36454F", size = 14, face = "bold")) + geom_smooth(color='
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Fitted Model of Relationship between Experience and Wages
8000
6000
4000
2000
0
Weekly Wages
0
20
40
60
Experience (in years)
Source: uswages dataset