Week 9 Graded Lab exercise

490241543

05/05/2020

```
data <- read.csv('NSW_Crime_data_clean-1.csv')

#head(data)
```

```
fraud <- data$Fraud
log.fraud <- log(fraud + 1)

mortgage <- data$Median_mortgage_repay_monthly
rent <- data$Median_rent_weekly
highschool <- data$X..of.People.who.completed.Year.12
age.65.74 <- data$X..Population.65.74
```

I have selected fraud(Number of fraud offences in each LGA) as a dependent variable. Due to its skewness, logarithm transformation was conducted. The graph that illustrates how the distribution was changed after transformation will be shown below.

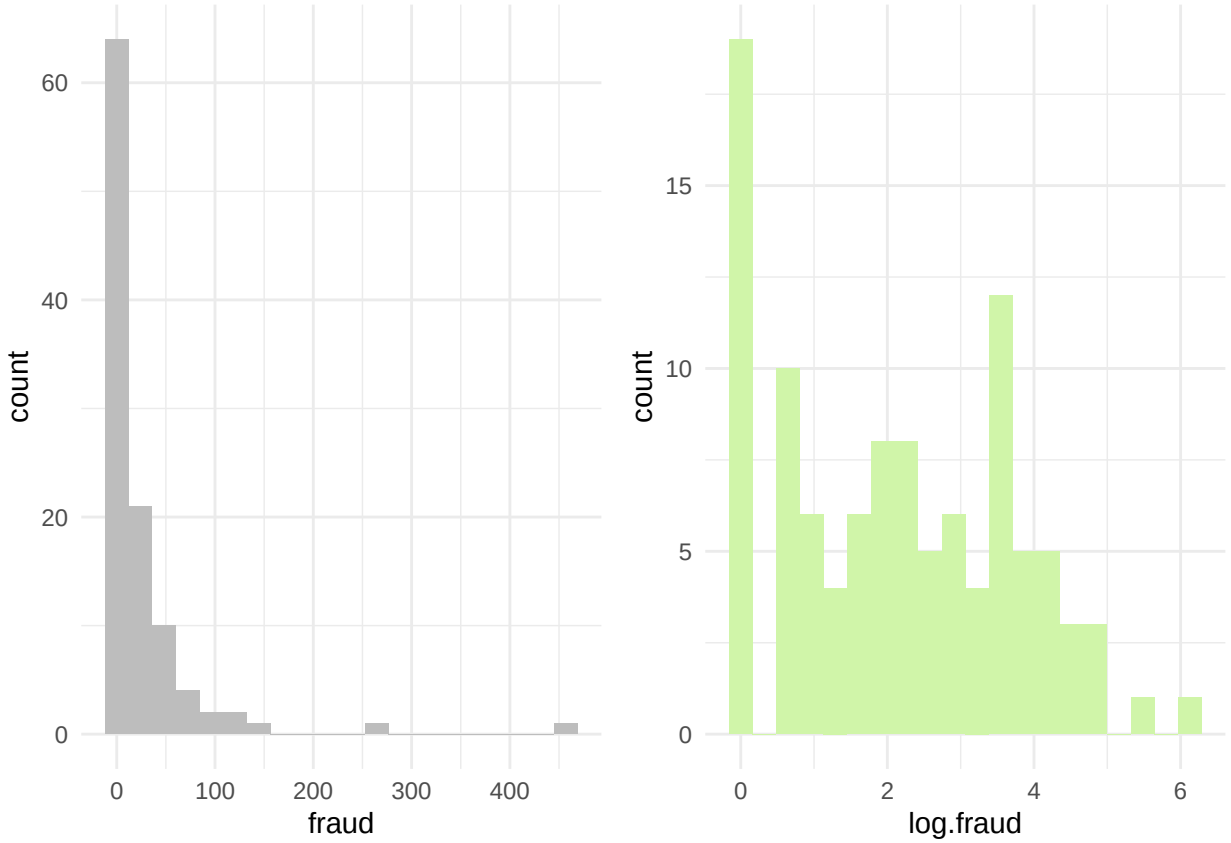
For independent variables, I have selected Median_mortgage_repay_monthly, Median_rent_weekly, Completed.Year.12, and Population.65.74. Firstly, I assume that mortgage repayments and rent affect on the number of fraud offences such as tax fraud because people living at the houses that are relatively pricey are more likely to commit this kind of fraud. Next, in terms of the percentage of people who completed high school, I had two different views. On the one hand, people who finished high school might have a deeper insight into the world, and where the money goes, therefore they tend to be involved in fraud. On the other hand, those people might be aware of how they are supposed to be punished when they commit fraud, and they are conscious to avoid it. I thought this variable would be related to fraud in either way, so I decided to choose as one of the independent variables. Lastly, the percentage of people within the ages of 65-74 has also been chosen since I thought where a higher number of elderly people live tend to show the lower number of fraud.

Exploratory Data Analysis

```
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr
```

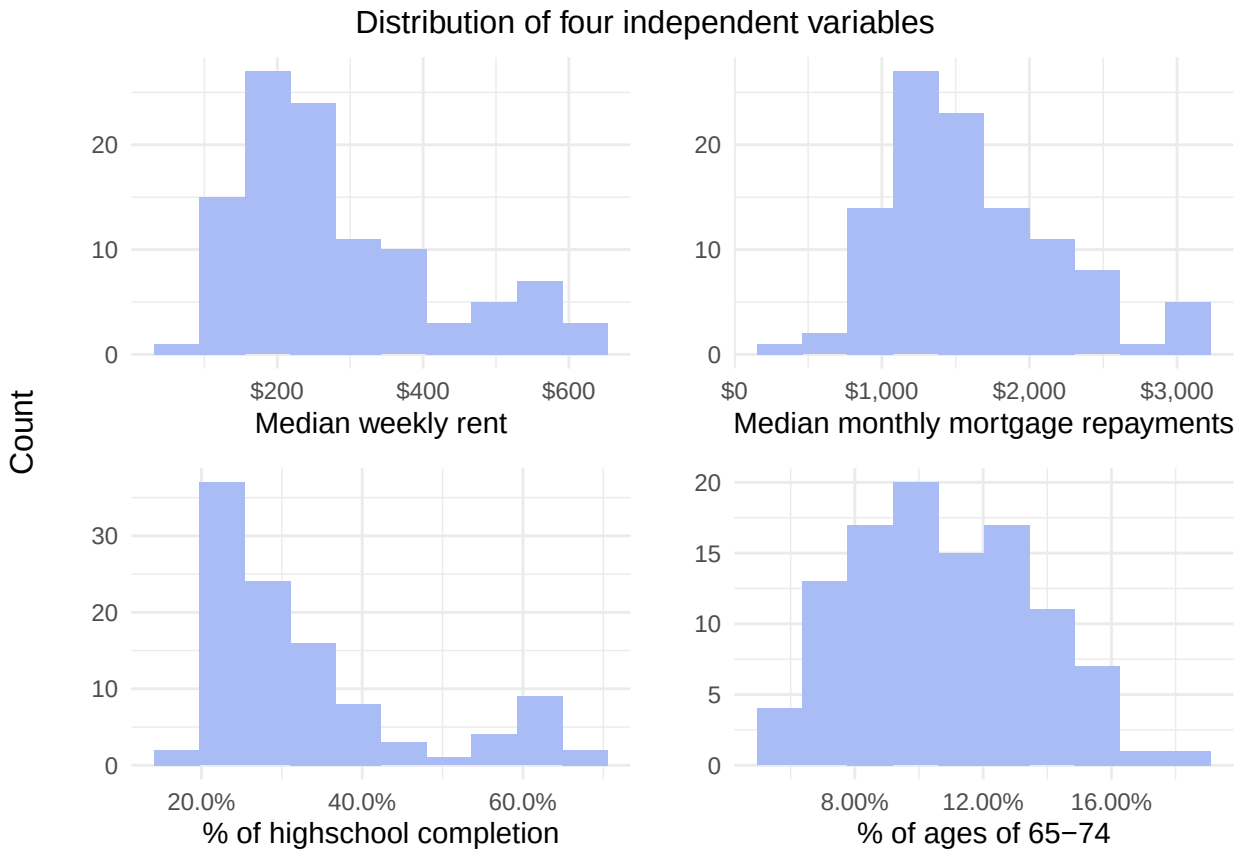
```
fraud_plot <- ggplot(data, aes(x = fraud)) +
              geom_histogram(bins = 20, fill='#BDBDBD') + theme_minimal()

log_fraud_plot <- ggplot(data, aes(x = log.fraud)) +
                  geom_histogram(bins = 20, fill='#D0F5A9') + theme_minimal()

ggarrange(fraud_plot, log_fraud_plot, ncol = 2, nrow = 1)
```



As you can see the graphs above, after 'fraud' variable has been converted into log, skewness of distribution seems to be corrected.

```
library(scales)

rent_plot <- ggplot(data, aes(x = rent)) +
              geom_histogram(bins = 10, fill = '#A9BCF5') + theme_minimal() +
              xlab("Median weekly rent") + ylab(" ") +
              scale_x_continuous(labels = dollar)

mortgage_plot <- ggplot(data, aes(x = mortgage)) +
                geom_histogram(bins = 10, fill = '#A9BCF5') + theme_minimal() +
                xlab("Median monthly mortgage repayments") + ylab(" ") +
                scale_x_continuous(labels = dollar)

highschool_plot <- ggplot(data, aes(x = highschool)) +
                    geom_histogram(bins = 10, fill='#A9BCF5') + theme_minimal() +
                    xlab("% of highschool completion") + ylab(" ") +
                    scale_x_continuous(labels = percent)

age.65.74_plot <- ggplot(data, aes(x = age.65.74)) +
                    geom_histogram(bins = 10, fill='#A9BCF5') + theme_minimal() +
                    xlab("% of ages of 65-74") + ylab(" ") +
                    scale_x_continuous(labels = percent)

dep.plot.fin <- ggarrange(rent_plot, mortgage_plot, highschool_plot, age.65.74_plot,
```

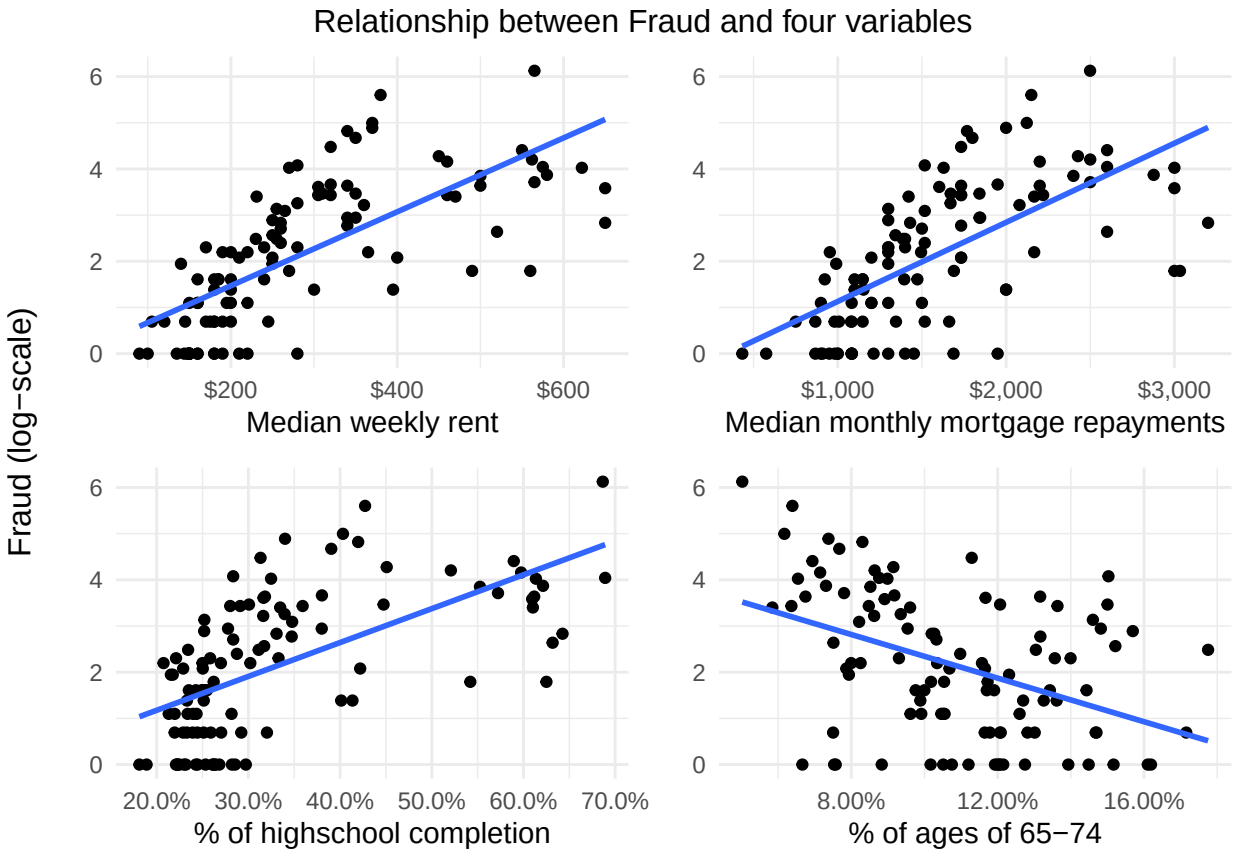
```
                        ncol = 2, nrow = 2)

annotate_figure(dep.plot.fin,
                top = text_grob("Distribution of four independent variables"),
                left = text_grob("Count", rot = 90))
```



The graphs above show the distribution of four different independent variables that have been selected earlier. They are normally distributed, aside from the data of the percentage of people who completed high school.

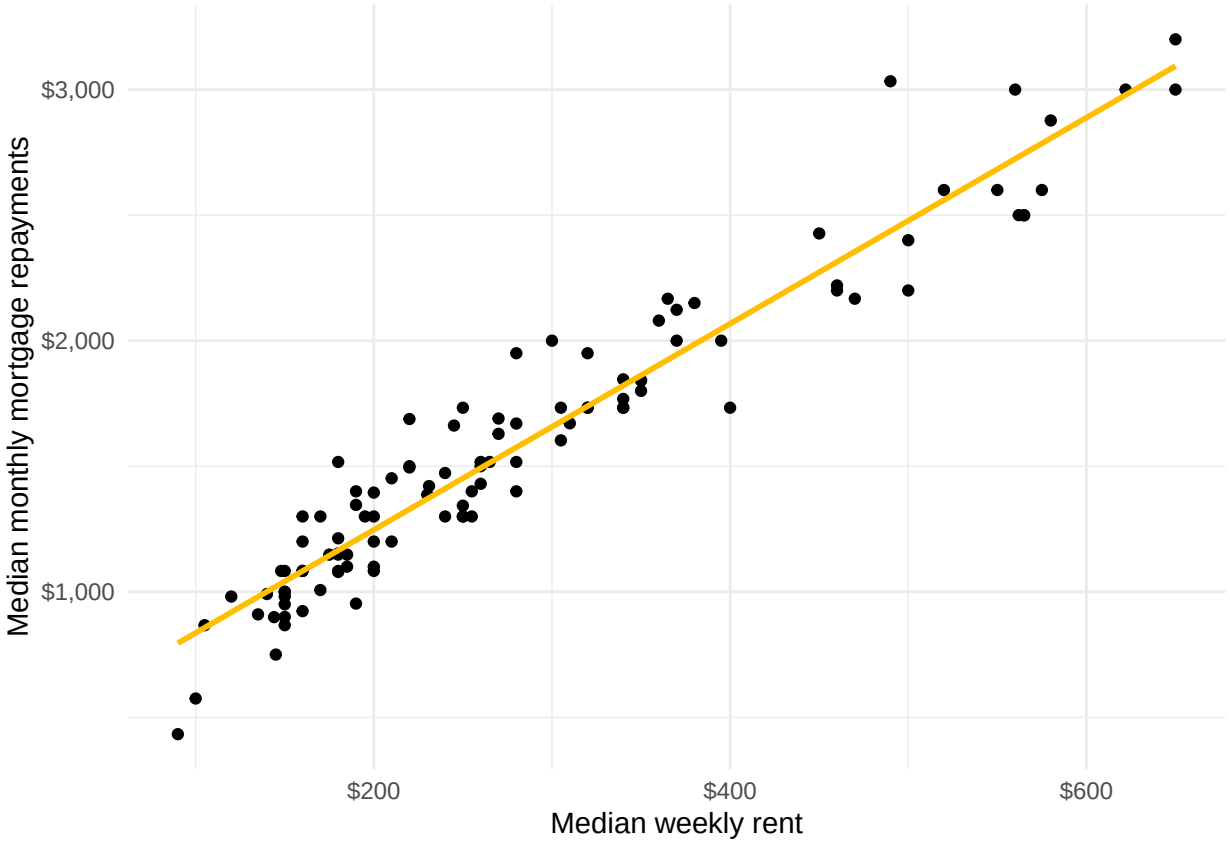
```
corplot.rent <- ggplot(data, aes(x = rent, y = log.fraud)) + geom_point() +
                geom_smooth(method = lm, se = FALSE) + theme_minimal() +
                xlab("Median weekly rent") + ylab(" ") +
                scale_x_continuous(labels = dollar)

corplot.mortgage <- ggplot(data, aes(x = mortgage, y = log.fraud)) + geom_point() +
                    geom_smooth(method = lm, se = FALSE) + theme_minimal() +
                    xlab("Median monthly mortgage repayments") + ylab(" ") +
                    scale_x_continuous(labels = dollar)

corplot.age <- ggplot(data, aes(x = age.65.74, y = log.fraud)) + geom_point() +
               geom_smooth(method = lm, se = FALSE) + theme_minimal() +
               xlab("% of ages of 65-74") + ylab(" ") +
               scale_x_continuous(labels = percent)

corplot.hischool <- ggplot(data, aes(x = highschool, y = log.fraud)) + geom_point() +
                    geom_smooth(method = lm, se = FALSE) + theme_minimal() +
```

```
                    xlab("% of highschool completion") + ylab(" ") +
                    scale_x_continuous(labels = percent)

corplot.fin <- ggarrange(corplot.rent, corplot.mortgage, corplot.hischool, corplot.age,
                         ncol = 2, nrow = 2)

annotate_figure(corplot.fin,
                top = text_grob("Relationship between Fraud and four variables"),
                left = text_grob("Fraud (log-scale)", rot = 90))
```



The graphs above represent the relationship between the fraud and four dependent variables. Firstly, as I mentioned earlier, median rent and median mortgage repayment seem to be linearly related to the number of fraud committed. The percentage of people who completed high school is less likely impacted with fraud, but I could see some correlation even though there are quite many of data points that are unfitted with the blue line on the graph. Unlike other variables, the percentage of people within the ages of 65-74 shows negative correlations with fraud.

```
rent.mortgage.plot <- ggplot(data, aes(x = rent, y = mortgage)) + geom_point() +
                      geom_smooth(method = lm, se = FALSE, colour = '#FFBF00') + theme_minimal() +
                      ylab("Median monthly mortgage repayments") +
                      xlab("Median weekly rent") +
                      scale_x_continuous(labels = dollar) +
                      scale_y_continuous(labels = dollar)

rent.mortgage.plot
```



I came up with one extra graph above, it represents that there is very strong relationship between mortgage repayments and weekly rent.

Model Building and Analysis

The new data frame called 'data2' was made in order to be used for building a model, and it contains the dependent and independent variables. As I saw a linear relationship between dependent variables and fraud, I built linear model as below.

```
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /Users/yeseul/Desktop/USYD/2020-1/Data Analysis in the Social Science/Tutorial/W
```

```
##
## Attaching package: 'arm'
```

```
## The following object is masked from 'package:scales':
##
##     rescale
```

```
data2 <- data.frame(Fraud = log.fraud, rent = rent, mortgage = mortgage,
                    highschool = highschool, age.65.74 = age.65.74)

model <- lm(Fraud ~ ., data = data2)

summary(model)
```

```
##
## Call:
## lm(formula = Fraud ~ ., data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16185 -0.86712 -0.04431  0.69678  2.65582
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.0893140  0.7583474   2.755  0.00696 **
## rent         0.0170648  0.0033666   5.069 1.82e-06 ***
## mortgage    -0.0014163  0.0006347  -2.232  0.02785 *
## highschool  -4.6908016  2.1309450  -2.201  0.02999 *
## age.65.74   -9.0489507  4.2378170  -2.135  0.03516 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 101 degrees of freedom
## Multiple R-squared:  0.5582, Adjusted R-squared:  0.5407
## F-statistic: 31.91 on 4 and 101 DF,  p-value: < 2.2e-16
```

Looking at the summary of the model above, standard errors of rent and mortgage are relatively lower than the other two variables, which means that these two variables are considered stastifically significant. Because residual standare error is 1.054, I would say this model may be fairy accurate. Moreover, R-squared value is approximaely 0.55, which is not perfectly high.

Discussion

I thought the LGA where more elderly (or retired) people live than younger people would be involved in a lower number of fraud rate. According to my analysis, my assumptions proved, and I feel this result quite accurately describes reality.

Moreover, I think job availability (or the percentage of people who have a proper job) and drug/alcohol abuse rate would be other great factors to predict crime. Especially in terms of robbery and housebreaking, these sorts of crime tend to happen when people are suffering from poverty due to unemployment.

Ethical Consideration

The data collected for this crime rates prediction project are considered quite sensitive, such as information on education level and the data that could be used to assume respondents' races. Some people possibly would

not want to respond the questions about the race, hence data collectors need to make sure respondents aware that participation of the survey is absolutely voluntary and know some questions of the survey might be emotionally harmful prior to the progress. These procedures are necessary in order to avoid any unethical data collection.