DATA ANALYSIS- College-Majors

San Jose State University

Professor - Dr. Shilpa Gupta

Date - 05-11-2022

I care about the college-Majors data set, reason being - as a graduate student, hunting job is pivotal.

It is equally important for others to look at this data because everyone needs job and to know what majors will help them in securing the job.

The context of the data includes the major code, major, major category, total number of students, Employed full time year, unemployed, unemployed rate, median and the percentile salary.

“A college degree is no guarantee of economic success. But through their choice of major, they can take at least some steps toward boosting their odds.”

College-major data

College-major all ages data

Link to write up and analysis

All data is from American Community Survey 2010-2012 Public Use Microdata Series.

cases incluede - variation in the number of employed and unemployed with respect to the major and major category

will be studying the categorical and numerical variables.

This is an observational study

```
getwd()

## [1] "/Users/Kavya/Desktop"

setwd("~/Desktop")

edu <- read.csv("all-ages-data.csv", header = TRUE, sep = ",")
dim(edu)

## [1] 173  11

mis <- is.na(edu)
dim(mis)

## [1] 173  11
```

```
dupli <- unique(edu)
dim(dupli)

## [1] 173  11

sum(is.na(edu))

## [1] 0

head(edu)

##   Major_code                                 Major
## 1       1100                   GENERAL AGRICULTURE
## 2       1101 AGRICULTURE PRODUCTION AND MANAGEMENT
## 3       1102                AGRICULTURAL ECONOMICS
## 4       1103                       ANIMAL SCIENCES
## 5       1104                          FOOD SCIENCE
## 6       1105            PLANT SCIENCE AND AGRONOMY
##                      Major_category  Total Employed
Employed_full_time_year_round
## 1 Agriculture & Natural Resources 128148    90245
74078
## 2 Agriculture & Natural Resources  95326    76865
64240
## 3 Agriculture & Natural Resources  33955    26321
22810
## 4 Agriculture & Natural Resources 103549    81177
64937
## 5 Agriculture & Natural Resources  24280    17281
12722
## 6 Agriculture & Natural Resources  79409    63043
51077
##   Unemployed Unemployment_rate Median P25th P75th
## 1       2423        0.02614711  50000 34000 80000
## 2       2266        0.02863606  54000 36000 80000
## 3        821        0.03024832  63000 40000 98000
## 4       3619        0.04267890  46000 30000 72000
## 5        894        0.04918845  62000 38500 90000
## 6       2070        0.03179089  50000 35000 75000
```

- The dimention of data is 173 rows and 11 colums
- there is no missing data
- there is no duplicate data

```
summary(edu)

##    Major_code      Major           Major_category         Total
##  Min.   :1100   Length:173          Length:173          Min.   :    2396
##  1st Qu.:2403   Class :character   Class :character   1st Qu.:  24280
##  Median :3608   Mode  :character   Mode  :character   Median :  75791
##  Mean   :3880                                         Mean   : 230257
##  3rd Qu.:5503                                         3rd Qu.: 205763
##  Max.   :6403                                         Max.   :3123510
##     Employed       Employed_full_time_year_round   Unemployed
##  Min.   :   1492   Min.   :   1093               Min.   :     0
##  1st Qu.:  17281   1st Qu.:  12722               1st Qu.:  1101
##  Median :  56564   Median :  39613               Median :  3619
##  Mean   : 166162   Mean   : 126308               Mean   :  9725
##  3rd Qu.: 142879   3rd Qu.: 111025               3rd Qu.:  8862
##  Max.   :2354398   Max.   :1939384               Max.   :147261
##  Unemployment_rate     Median           P25th           P75th
##  Min.   :0.00000   Min.   : 35000   Min.   :24900   Min.   : 45800
##  1st Qu.:0.04626   1st Qu.: 46000   1st Qu.:32000   1st Qu.: 70000
##  Median :0.05472   Median : 53000   Median :36000   Median : 80000
##  Mean   :0.05736   Mean   : 56816   Mean   :38697   Mean   : 82506
##  3rd Qu.:0.06904   3rd Qu.: 65000   3rd Qu.:42000   3rd Qu.: 95000
##  Max.   :0.15615   Max.   :125000   Max.   :78000   Max.   :210000
```

My research question on the data that I have selected are

Q1. How is the median salary distributed

Q2. Which Major has the highest salary earning and lowest salary earning

Q3. What were the most common majors (will not be showing all 173, as it will be huge)

Q4. Which Major categor is making

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.8
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard
## The following object is masked from 'package:readr':
## 
##     col_factor
```

```
summary(edu)
```

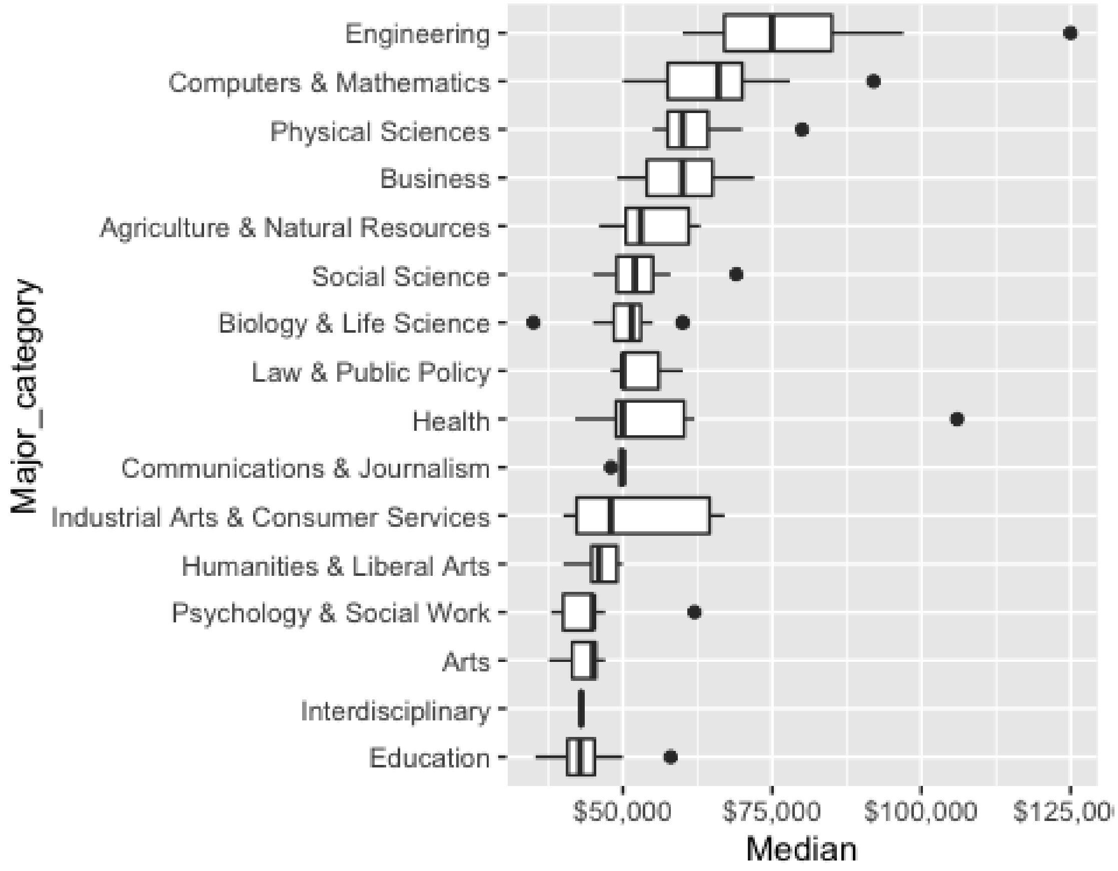
```
##    Major_code      Major           Major_category         Total        
##  Min.   :1100   Length:173         Length:173         Min.   :   2396  
##  1st Qu.:2403   Class :character   Class :character   1st Qu.:  24280  
##  Median :3608   Mode  :character   Mode  :character   Median :  75791  
##  Mean   :3880                                         Mean   : 230257  
##  3rd Qu.:5503                                         3rd Qu.: 205763  
##  Max.   :6403                                         Max.   :3123510  
##     Employed       Employed_full_time_year_round   Unemployed    
##  Min.   :   1492   Min.   :   1093              Min.   :     0  
##  1st Qu.:  17281   1st Qu.:  12722              1st Qu.:  1101  
##  Median :  56564   Median :  39613              Median :  3619  
##  Mean   : 166162   Mean   : 126308              Mean   :  9725  
##  3rd Qu.: 142879   3rd Qu.: 111025              3rd Qu.:  8862  
##  Max.   :2354398   Max.   :1939384              Max.   :147261  
##  Unemployment_rate     Median           P25th           P75th       
##  Min.   :0.00000   Min.   : 35000   Min.   :24900   Min.   : 45800  
##  1st Qu.:0.04626   1st Qu.: 46000   1st Qu.:32000   1st Qu.: 70000  
##  Median :0.05472   Median : 53000   Median :36000   Median : 80000  
##  Mean   :0.05736   Mean   : 56816   Mean   :38697   Mean   : 82506  
##  3rd Qu.:0.06904   3rd Qu.: 65000   3rd Qu.:42000   3rd Qu.: 95000  
##  Max.   :0.15615   Max.   :125000   Max.   :78000   Max.   :210000
```

```
str(edu)
```

```
## 'data.frame':    173 obs. of  11 variables:
##  $ Major_code                   : int  1100 1101 1102 1103 1104 1105 1106
1199 1301 1302 ...
##  $ Major                        : chr  "GENERAL AGRICULTURE" "AGRICULTURE
PRODUCTION AND MANAGEMENT" "AGRICULTURAL ECONOMICS" "ANIMAL SCIENCES" ...
##  $ Major_category               : chr  "Agriculture & Natural Resources"
"Agriculture & Natural Resources" "Agriculture & Natural Resources"
"Agriculture & Natural Resources" ...
##  $ Total                        : int  128148 95326 33955 103549 24280
79409 6586 8549 106106 69447 ...
##  $ Employed                     : int  90245 76865 26321 81177 17281 63043
4926 6392 87602 48228 ...
##  $ Employed_full_time_year_round: int  74078 64240 22810 64937 12722 51077
4042 5074 65238 39613 ...
##  $ Unemployed                   : int  2423 2266 821 3619 894 2070 264 261
4736 2144 ...
##  $ Unemployment_rate            : num  0.0261 0.0286 0.0302 0.0427 0.0492
...
##  $ Median                       : int  50000 54000 63000 46000 62000 50000
63000 52000 52000 58000 ...
##  $ P25th                        : int  34000 36000 40000 30000 38500 35000
```

```
39400 35000 38000 40500 ...
##  $ P75th                           : num  80000 80000 98000 72000 90000 75000
88000 75000 75000 80000 ...
```

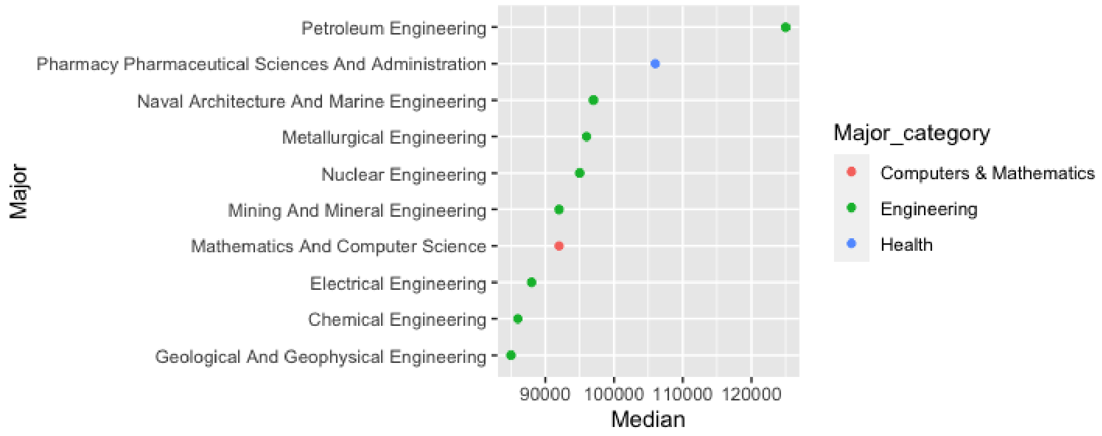
```
edu %>% mutate(Major_category = fct_reorder(Major_category, Median)) %>%
ggplot(aes(Major_category, Median)) + geom_boxplot() +
scale_y_continuous(labels = label_dollar ()) + coord_flip()
```



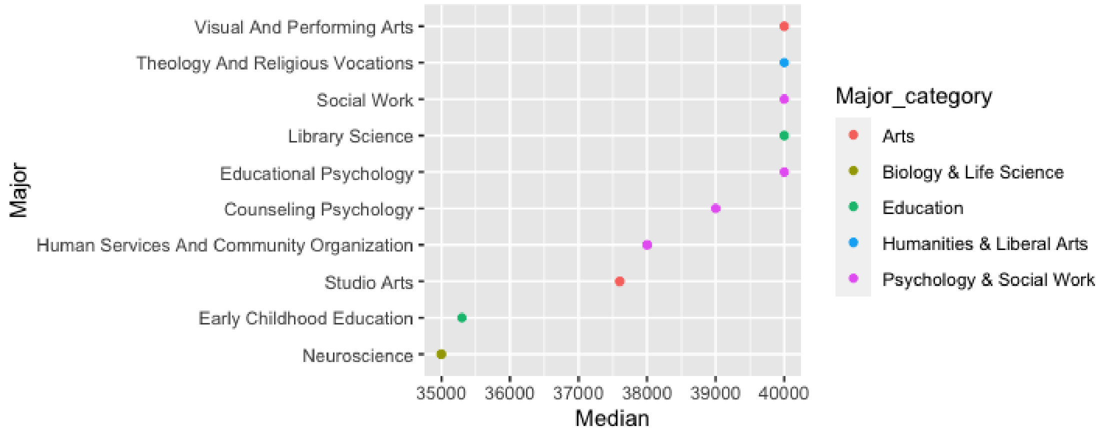
Ans 1.

From the distribution it is clear that the Engineering Major_category has highest median earning of $75000 and Education category has the lowest median salary of around $35000

```
edu_data <- edu %>% arrange(desc(Median)) %>%
  select(Major, Major_category, Median, P25th, P75th) %>%
  head(10) %>%
  mutate(Major= str_to_title(Major), Major = fct_reorder(Major, Median)) %>%
  ggplot(aes(Major, Median, color = Major_category)) +
  geom_point() +
  coord_flip()
  edu_data
```



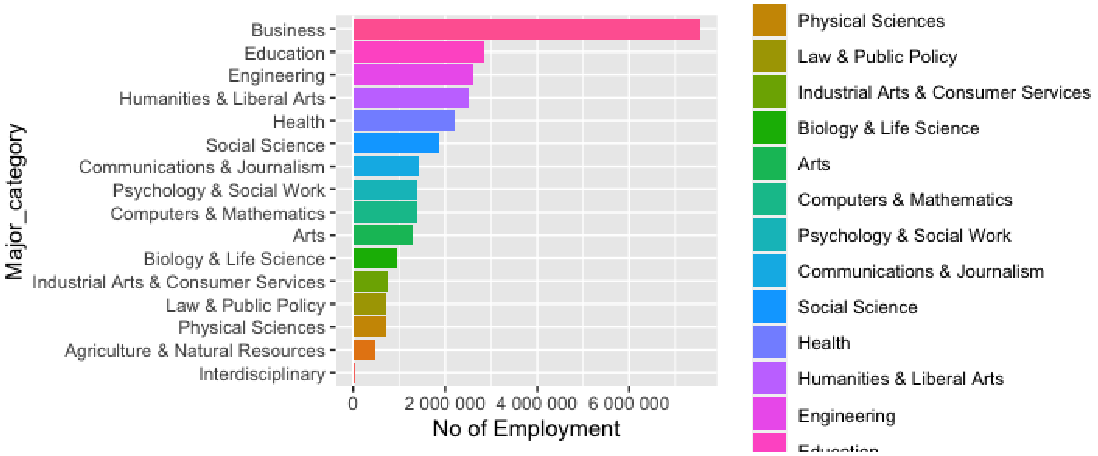
```
edu_data <-edu %>% arrange(desc(Median)) %>%
 select(Major, Major_category, Median, P25th, P75th) %>%
 tail(10) %>%
 mutate(Major= str_to_title(Major), Major = fct_reorder(Major, Median)) %>%
 ggplot(aes(Major, Median, color = Major_category)) +
 geom_point() +
 coord_flip()
edu_data
```



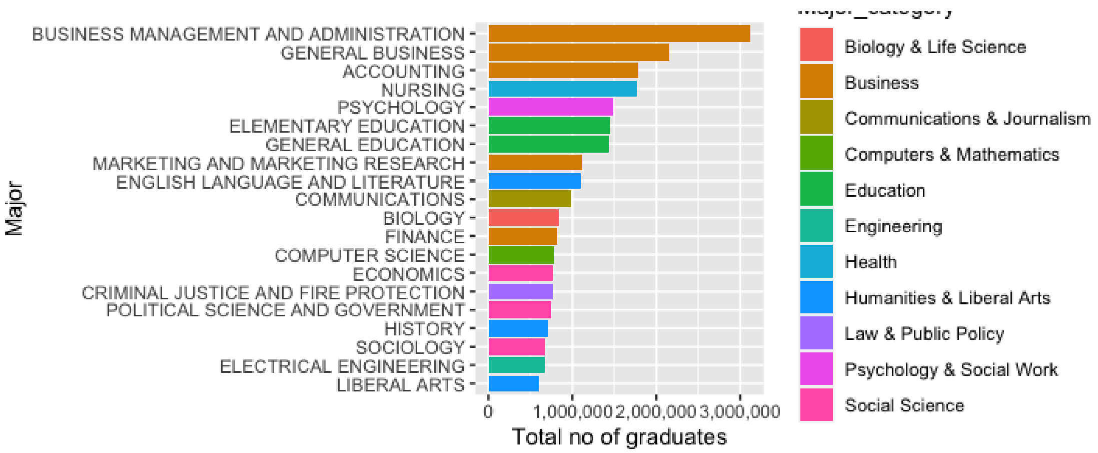
Ans 2.

Petroleum Engineering Major has the highest median paying of around 120000 and Neuroscience has the lowest salary earning of 35000

```
edu_analysing <- edu %>% count(Major_category, wt = Employed, sort = TRUE)
%>%
   mutate(Major_category = fct_reorder(Major_category, n)) %>%
 ggplot(aes(Major_category, n, fill = Major_category )) +
 geom_col() +
 coord_flip() +
 labs( y = "No of Employment") +
 scale_y_continuous(labels = label_number())

edu_analysing
```



```
edu_analysing_1 <- edu %>%
    mutate(Major = fct_reorder(Major, Total)) %>%
  arrange(desc(Total)) %>%
  head(20) %>%
  ggplot(aes(Major, Total, fill = Major_category )) +
  geom_col() +
  coord_flip() +
  labs( y = "Total no of graduates") +
  scale_y_continuous(labels = comma_format())
edu_analysing_1
```



Ans 3.

From the graph it can be inferred that Majors has the common categories (with same color) e.g > Business Management and administration > general Business > accounting > Marketing research >Finance have the common Major_category of Business

---

Hypothesis

```
mean_edu <- mean(edu$Median)
max(sapply(edu$Median, max))

## [1] 125000
```

```
min(sapply(edu$Median, min))

## [1] 35000

sd(edu$Median)

## [1] 14706.23
```

The typical recent college graduate with a full-time job earns about $36,000 a year, according to the American Community Survey.

But graduates with a degree in petroleum engineering is earning $125,000 and Neuroscience has the lowest earning of $35,000

The mean median salary is 56816.18 from the data.

For the graduates, is the mean median salary less than the typical salary of recent graduate obtained from the American Community Survey?

1. H0: $\mu = 36{,}000$ H1: $\mu > 36{,}000$

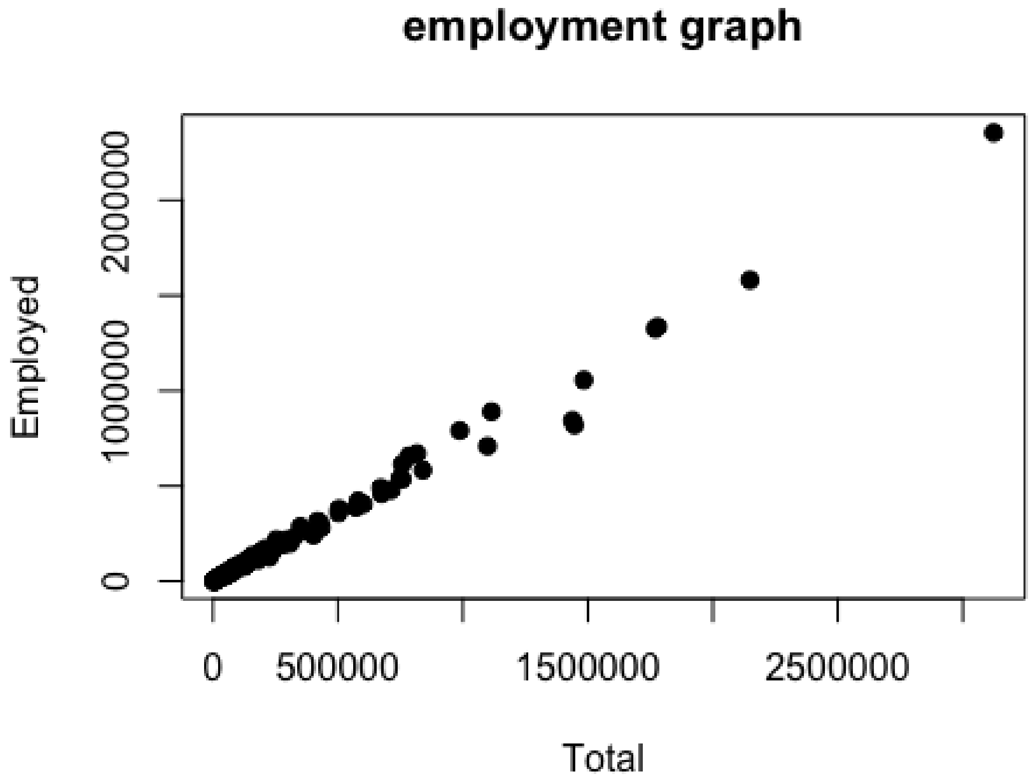
The Mean median salary is more than the salary obtained in the survey, in the alternate hypothesis seems to be true in case of 1st hypothesis formulation.

But even in more closely related fields, there are clear differences in earnings between majors. Actuarial science majors earn more than accounting majors; public policy majors out-earn history majors; and court reporting is better earnings bet than criminology.

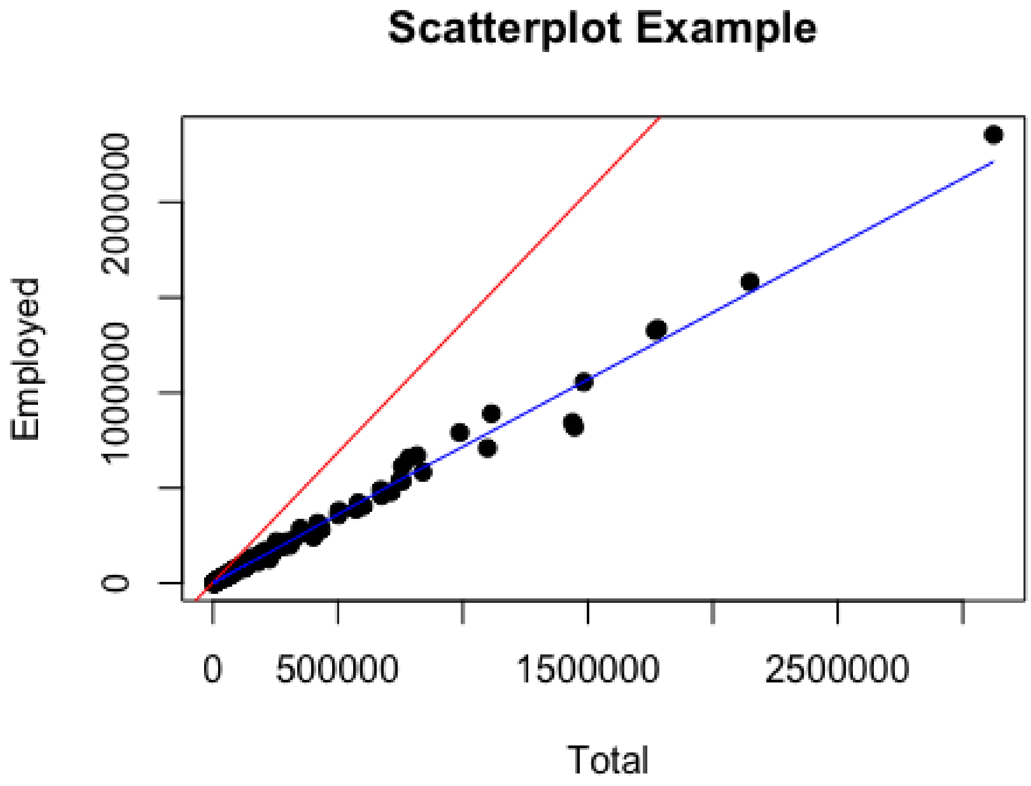
```
x <- edu$Median
t.test(x, mu = 36000)

##
##  One Sample t-test
##
## data:  x
## t = 18.618, df = 172, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 36000
## 95 percent confidence interval:
##  54609.23 59023.14
## sample estimates:
## mean of x
##  56816.18
```

Here we reject the alternate hypothesis as the mean value is not equal to 36000 and the P-value is less

```
attach(edu)
plot(Total, Employed, main="employment graph",
   xlab="Total", ylab="Employed", pch=19)
```



```
plot(Total,  Employed, main="Scatterplot Example",
   xlab=" Total", ylab="Employed", pch=19)
abline(lm(Total~Employed), col="red") # regression line (y~x)
lines(lowess(Total, Employed), col="blue") # lowness line (x,y)
```



There is a linear relationship between the total number of students and the number of Employment

If the number of students enrolled is more then the no of employees will also be more

The regression line is added in the second graph

```
linear_reg <- lm(formula = Employed~Total, data = edu)
summary(linear_reg)

##
## Call:
## lm(formula = Employed ~ Total, data = edu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228060   -2024     809    3926   92130
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.545e+02  2.674e+03  -0.245    0.807
## Total        7.245e-01  5.574e-03 129.967   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 30860 on 171 degrees of freedom
## Multiple R-squared:   0.99,  Adjusted R-squared:  0.9899
## F-statistic: 1.689e+04 on 1 and 171 DF,  p-value: < 2.2e-16
```

From the coefficient Estimate it is seen that there is a positive relationship between "Total" number of students and the "Employment"

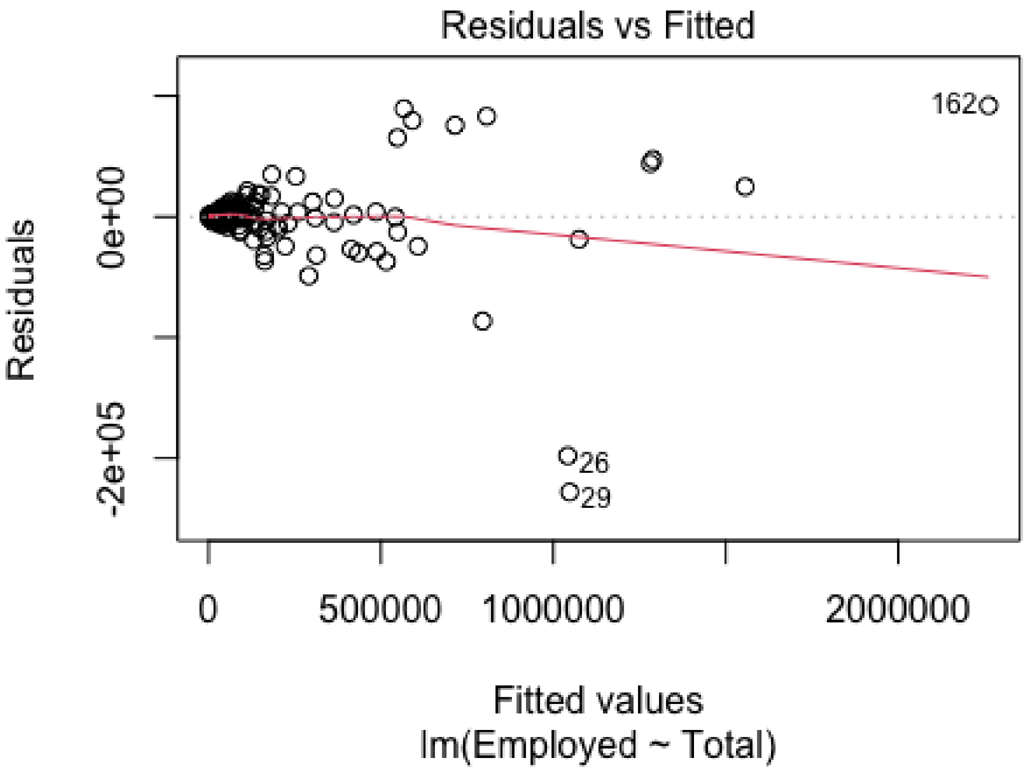
R-squared value here is 0.99, i.e the 'monthly energy' usage explains 99% of the variability in 'peak-hour' demand

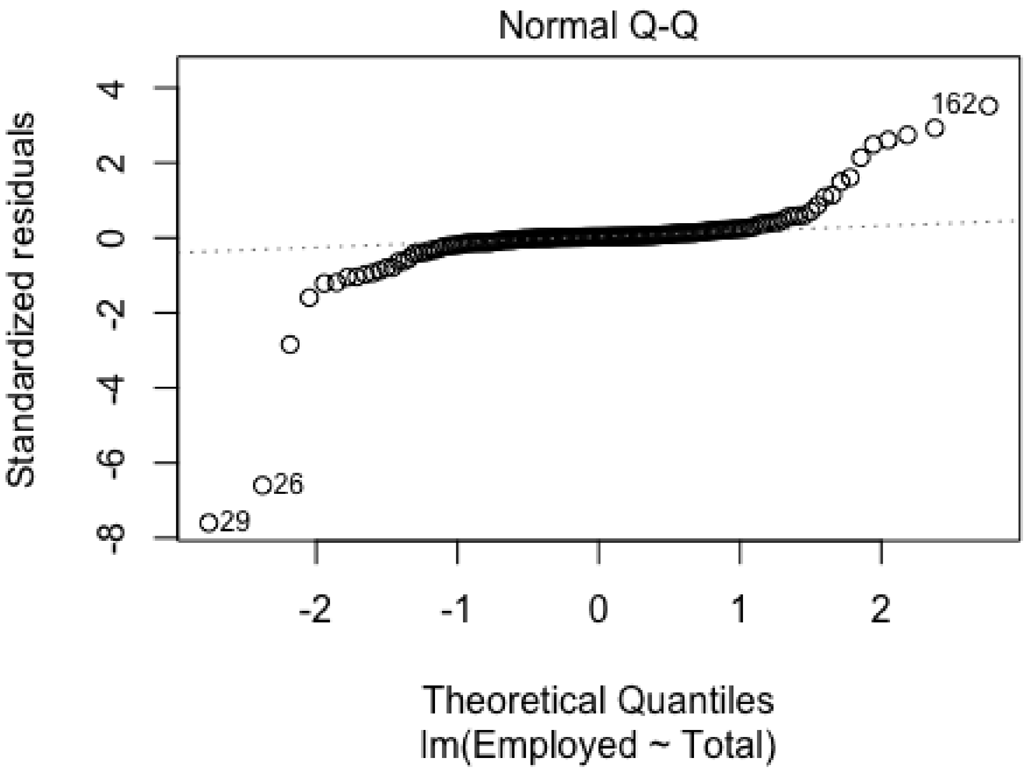
correlation coefficient = sqrt(R-square) For alpha = 0.05, data frame = 173-2 = 173-2 = 171

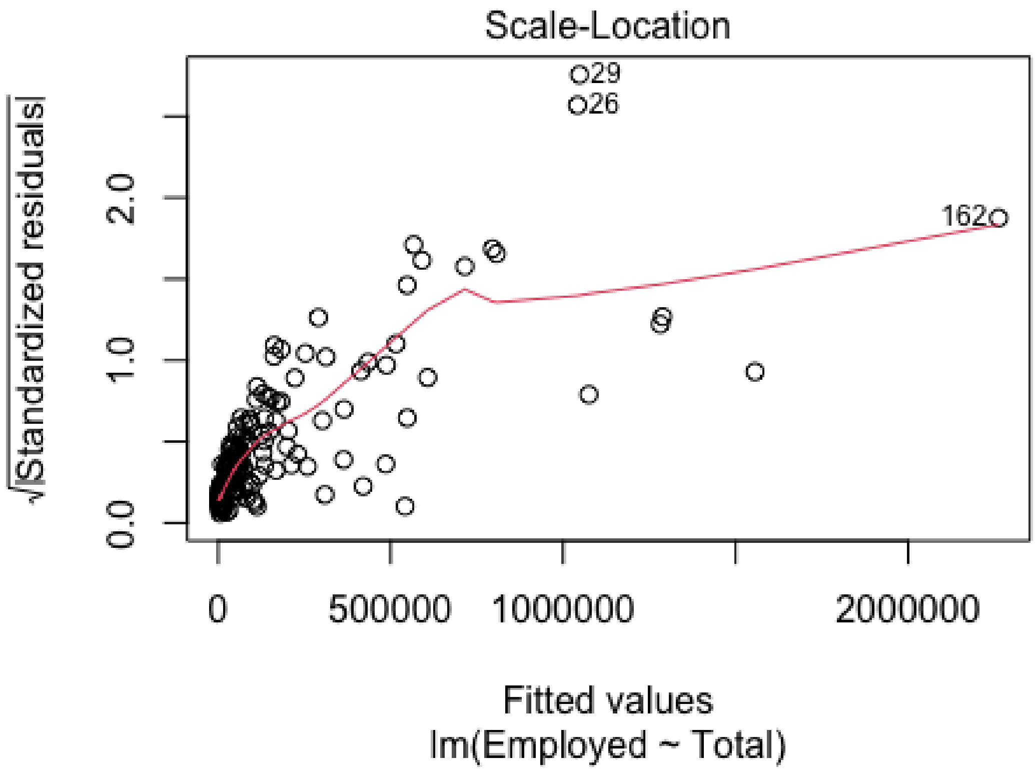
```
ct_edu <- cor.test(edu$Total, edu$Employed)
ct_edu

##
##  Pearson's product-moment correlation
##
## data:  edu$Total and edu$Employed
## t = 129.97, df = 171, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9932204 0.9962783
## sample estimates:
##       cor
## 0.9949763
```

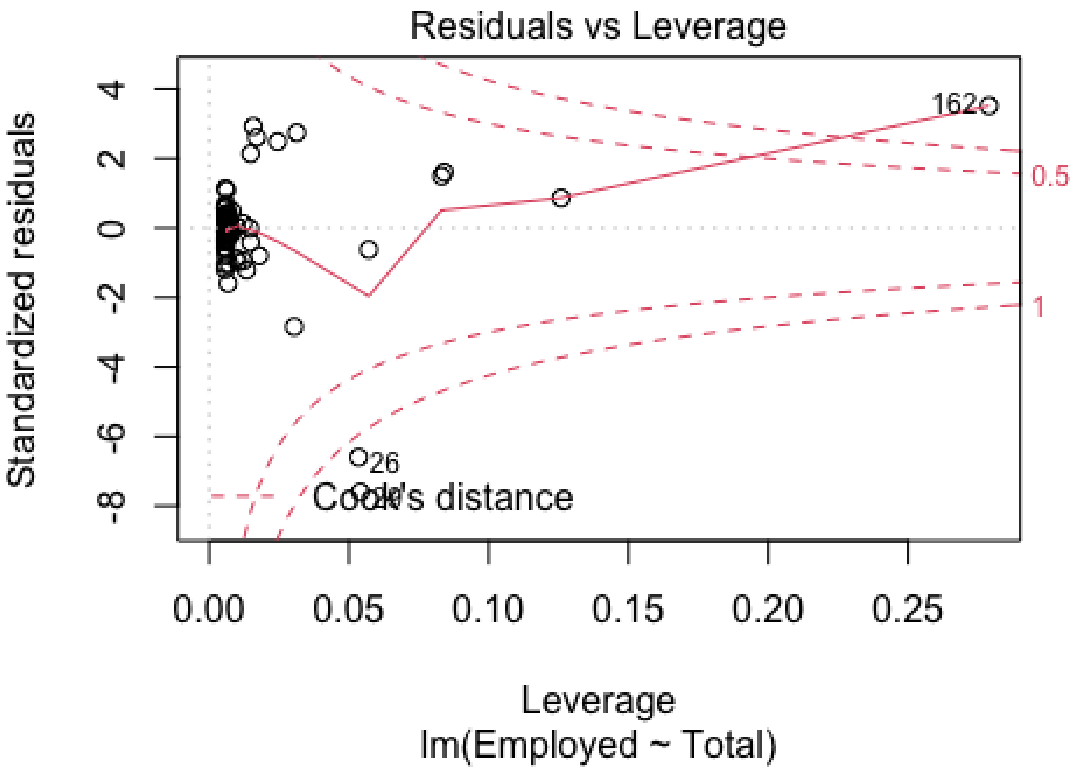
The correlation coefficient with respect to 95% of confidence interval is found to be 0.9932 and 0.9962

```
plot(linear_reg)
```


Residuals vs Fitted
Residuals
0e+00
-2e+05
0
500000
1000000
2000000
Fitted values
lm(Employed ~ Total)
162
26
29


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-6
-8
-2
-1
0
1
2
162
26
29
Theoretical Quantiles
lm(Employed ~ Total)


Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
0
500000
1000000
2000000
29
26
162
Fitted values
lm(Employed ~ Total)



Graph 1 - The linearity assumption is not met in the 1st plot, hence there is a pattern and also the variation is not constant

Graph 2 - From the second plot the error are not normally distributed as there is no linearity in the distribution(points are not falling roughly on a diagonal line)

Graph 3, 4 - from these graph it is seen that there is non linearity, the variance is not constant