# MA334-Coursework

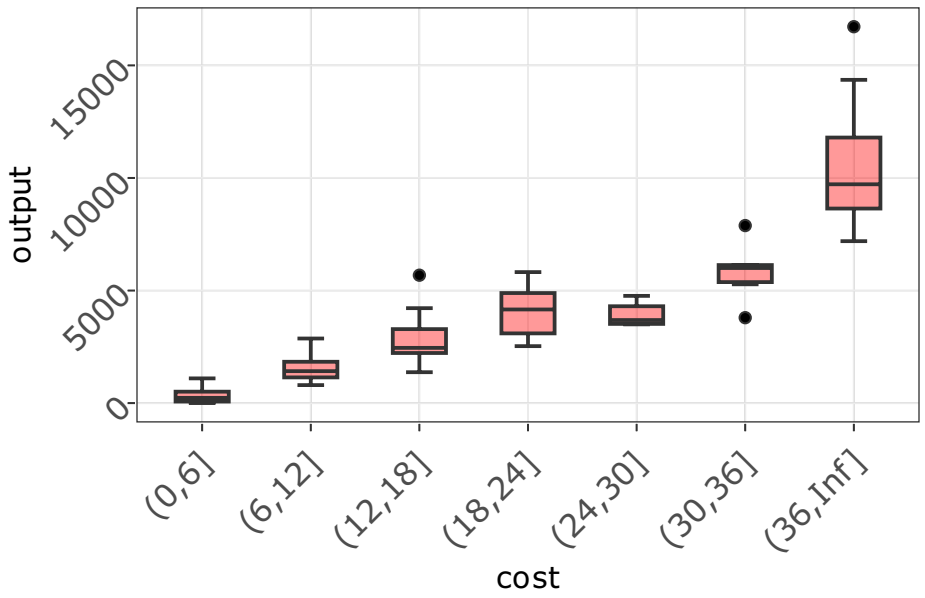
Analysis of a electricity producers dataset from US in 1955

2111159-Md Hazzaz Rahman-Antu

## Introduction

The name of the dataset chosen for this project is "Cost Function of Electricity Producers (1955, Nerlove Data). It is a cost function data of 159 US electricity producers in 1955. The data frame contains 159 observations on 8 variables. The 8 variables are: 1. cost = Total cost for the producer to generate certain amount of electricity. 2. output = Total output or total amount of electricity produced. 3. labor = Labors's wage rate. 4. laborshare = Cost share for labor. 5. capital = Capital price index. 6. capitalshare = Cost share for capital. 7. fuel = Fuel price. 8. fuelshare = Cost share for fuel.

From the summary, all 8 variables in this dataset is continuous data. Among them the variable chosen for target variable is "output" and predictor variable is "cost". Also, in the summary we can notice the minimum, maximum, and mean cost for the predictor variable is 0.082, 139.422, and 12.976. For target variable "output", it is 2, 2133, and 16719.



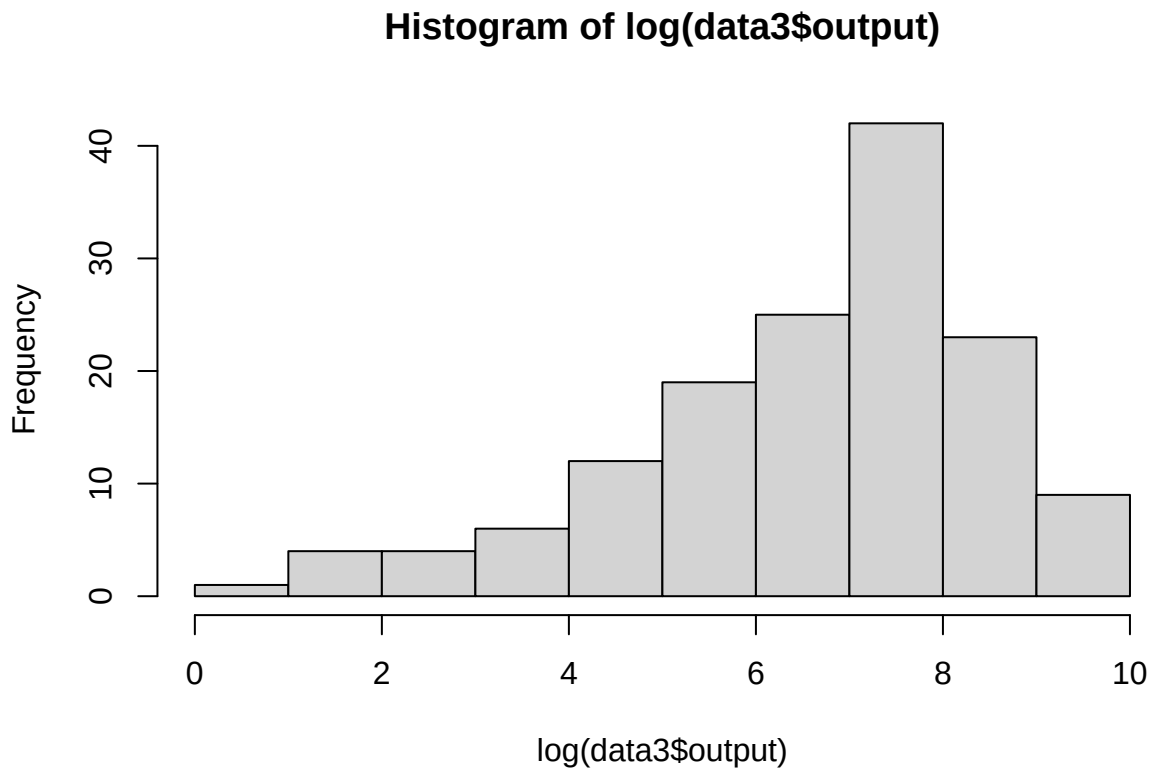

This is a boxplot between two continuous variable cost and output. Cost has been divided into 7 intervals. For 7 different intervals of cost output we can see, the output is gradually increasing as the cost goes higher.

The highest output is produced when cost is highest (36-infinity). *Source:* Online complements to Greene (2003). Table F14.2. http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm *Reference:* Greene, W.H. (2003). Econometric Analysis, 5th edition. Upper Saddle River, NJ: Prentice Hall. Nerlove, M. (1963) "Returns to Scale in Electricity Supply." In C. Christ (ed.), Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld. Stanford University Press, 1963.

## Data pre-processing

For further analysis, dataset needed some cleaning. The dataset contains several extra observations that are aggregates of commonly owned firms. Only the first 145 observations should be used for this analysis. So, last 14 rows have been removed from the dataset. The first column of the dataset which is 'x' also been removed because it is just label or counting of the rows.

## Results




*Histogram:* In the histogram we can clearly see that the data is normally distributed because it forms a bell-shaped structure when it is plotted. It is indeed a typical example of data displaying normal distribution. So as a result, most of the data in terms of frequency is around the mean and as it is move away from the mean, the frequency start to decrease.

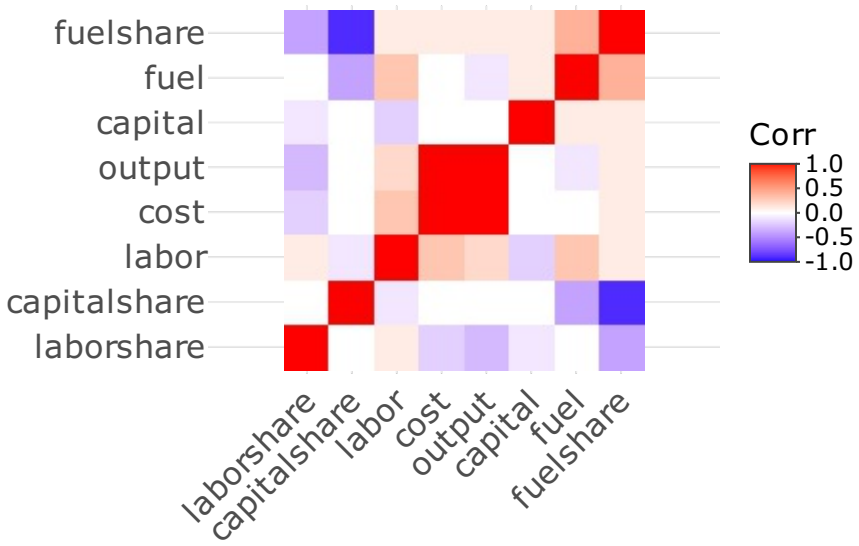
```
## [1] "numeric"
```

```
## [1] 2133.083
```

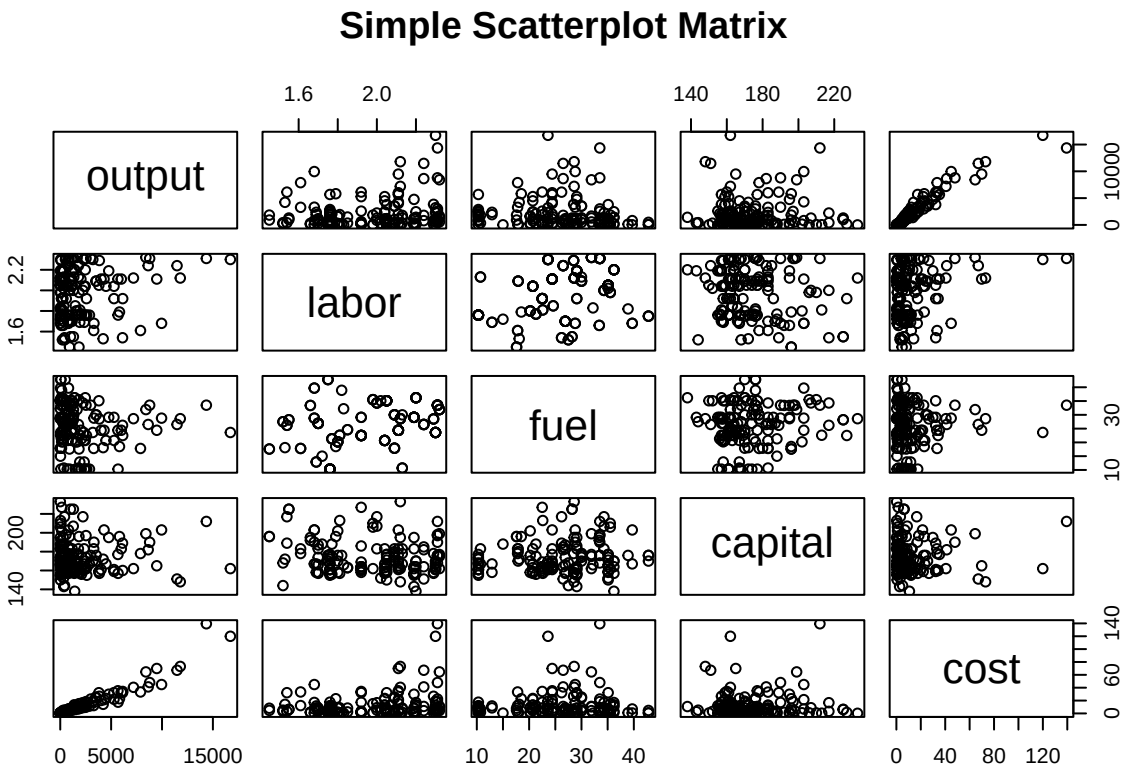
```
## [1] 8596285
```

```
## [1] 2931.942
```

Here, variance is 8596285 and staandard deviation is 2931.942. So, most of the data is plus or minus 2931.942 from the mean 2133.083.




*Correlation:* From the correlation matrix above we can see that, 1. Target variable "output" and predictor variable "cost" have high positive correlation because the correaltion coefficient value is 1. 2. Labor and fuelshare is also postively correlated with the variable output but correlation is much weaker. 3. Laborshare and fuel have a very low negative correaltion with output variable. 4. Other features do not have any impact on




output.

*Scatter Plot:* In the first two plots which are output vs labor and output vs fuel, we hardly see any pattern because they have very low correlation with output but there is no pattern in the output vs capital plot. We might have expected that the cost variety in different section would have effect output but there is no obvious pattern. But as a matter of fact if we combine these cost and plot a scatter diagram with output vs total cost we can clearly see that, as the cost increases, the output increase as well.

```
## [1] 2133.083
```

```
## [1] 2369.5
```

```
##   Type.of.Data     Mean
## 1   Population 2133.083
## 2       Sample 2369.500
```

```
##
##  One Sample t-test
##
## data:  sampledata$output
## t = 0.41829, df = 19, p-value = 0.6804
## alternative hypothesis: true mean is not equal to 2133.083
## 95 percent confidence interval:
##  1186.513 3552.487
## sample estimates:
## mean of x
##    2369.5
```

*One sample t-test* In here, randomly 20 output of electricity has been taken to check the distribution. The random sample mean is 1402.200. So, after conducting the t-test one can say that there is no statistical difference between the sample mean and the population mean for the output variable. Null Hypothesis, H0: There is no statistical difference between the sample mean and population mean Alternative Hypothesis, H1: There is a statistical difference between the sample mean and population mean Significance Level: 0.05 (Equivalent to 95% confidence interval) p-value = 0.1571; in here, p-value is more than 0.05. So, there is no clear evidence that contradicts the null hypothesis. As a result, fail to reject the null hypothesis.

```
##
## Call:
## lm(formula = output ~ labor + capital + fuel, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3083.9 -1777.2  -729.5   328.0 13616.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3807.87    3387.81  -1.124   0.2629
## labor        2883.49    1098.15   2.626   0.0096 **
## capital        10.30      13.67   0.753   0.4526
## fuel          -58.89      32.75  -1.798   0.0743 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2884 on 141 degrees of freedom
## Multiple R-squared:  0.05248,    Adjusted R-squared:  0.03232
## F-statistic: 2.603 on 3 and 141 DF,  p-value: 0.05438
```

*Regression Analysis:* output = -3807.87+2883.49labor + 10.30capital + fuel

From the regression we can see that, labor has the more significance. The 3 variables labor, capital, and fuel In combination will contribute 52% to manipulate the output.

## Conclusion

In the analysis we can see that, even if the total cost of electricity production is highly correlated with the output of the electricity but the individual costs are barely correlated with the output variable.

##Appendix

webshot::install_phantomjs() library(dsEssex) library(tidyverse) library(tidytext) library(ggrepel) tinytex::install_tinytex() library(dplyr) library("janitor") library(ggplot2) library("ggcorrplot") library("plotly") library(htmlwidgets) library(reshape2) library(gridExtra) library(ggrepel) library(caTools) tinytex::install_tinytex()

#load the data data1 = read.csv("C:\Users\User\Desktop\ma334 project\Electricity1955.csv", header= TRUE) data3 = data1[-c(146:159),-1]

summary(data3) #box-plot data3$cat_cost = cut(data3$cost, c(0,6,12,18,24,30,36,Inf)) g <- ggplot(data3, aes(x=factor(cat_cost), y=output)) + xlab("cost") + ylab("output") g1 <- g + geom_boxplot(fill="red", alpha=0.4)+ theme_bw() + theme(axis.text=element_text(face='bold', size = 12, angle = 45, hjust = 1)) ggplotly(g1) #histogram hist(log(data3$output))

mode(data3$output)mean(data3$output) var(data3$output)sd(data3$output)

#Correlation corr.mat <- round(cor(data3), 1) pval.cor <- cor_pmat(data3) G <- ggcorrplot(corr.mat, hc.order = TRUE) (Fig = ggplotly(G)) #Scatterplot pairs(output~labor+fuel+capital+cost,data=data3, main="Simple Scatterplot Matrix") #t-test sampledata <- data3 %>% sample_n(20) population_mean <- mean(data3$output) $population_mean sample_mean < -mean(sampledata$output) sample_mean mean_comparison <- data.frame("Type of Data" = c("Population", "Sample"), "Mean" = c(population_mean, sample_mean)) mean_comparison t.test(sampledata$output, mu = mean(data3$output)) #Regression analysis reg<- lm(output~ labor+capital+fuel, data=data3) summary(reg)