Main

2022-12-11

```
salaries= read.csv("AI_MLsalaries.csv")
my_data <- salaries
```

```
#summarizing the data
str(my_data)
```

```
## 'data.frame':    1332 obs. of  11 variables:
##  $ work_year         : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ experience_level  : chr  "MI" "MI" "MI" "MI" ...
##  $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
##  $ job_title         : chr  "Machine Learning Engineer" "Machine Learning Engineer" "Data Scientist"
##  $ salary            : int  130000 90000 120000 100000 85000 78000 161000 110000 136000 104000 ...
##  $ salary_currency   : chr  "USD" "USD" "USD" "USD" ...
##  $ salary_in_usd     : int  130000 90000 120000 100000 85000 78000 161000 110000 136000 104000 ...
##  $ employee_residence: chr  "US" "US" "US" "US" ...
##  $ remote_ratio      : int  0 0 100 100 100 100 100 100 100 100 ...
##  $ company_location  : chr  "US" "US" "US" "US" ...
##  $ company_size      : chr  "M" "M" "M" "M" ...
```

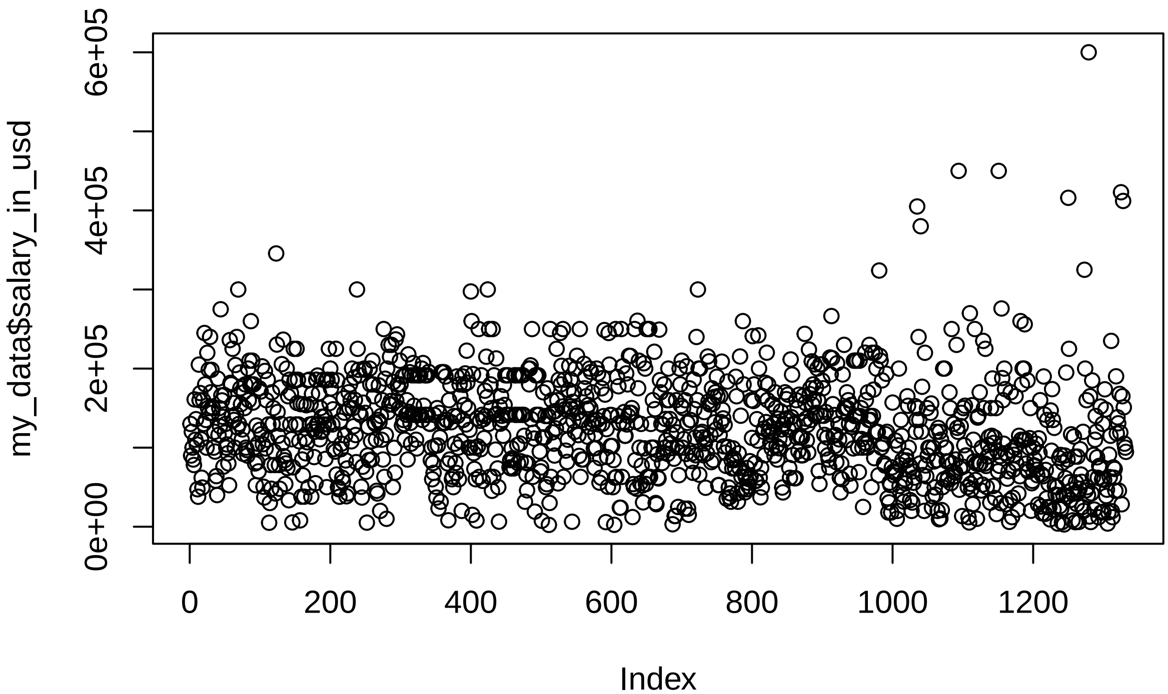
```
summary(my_data)
```

```
##    work_year    experience_level   employment_type     job_title
##  Min.   :2020   Length:1332        Length:1332        Length:1332
##  1st Qu.:2022   Class :character   Class :character   Class :character
##  Median :2022   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2022
##  3rd Qu.:2022
##  Max.   :2022
##      salary         salary_currency    salary_in_usd    employee_residence
##  Min.   :    2324   Length:1332        Min.   :  2324   Length:1332
##  1st Qu.:   80000   Class :character   1st Qu.: 75593   Class :character
##  Median :  130000   Mode  :character   Median :120000   Mode  :character
##  Mean   :  237712                      Mean   :123375
##  3rd Qu.:  175100                      3rd Qu.:164997
##  Max.   :30400000                      Max.   :600000
##   remote_ratio    company_location   company_size
##  Min.   :  0.00   Length:1332        Length:1332
##  1st Qu.:  0.00   Class :character   Class :character
##  Median :100.00   Mode  :character   Mode  :character
##  Mean   : 63.85
##  3rd Qu.:100.00
##  Max.   :100.00
```

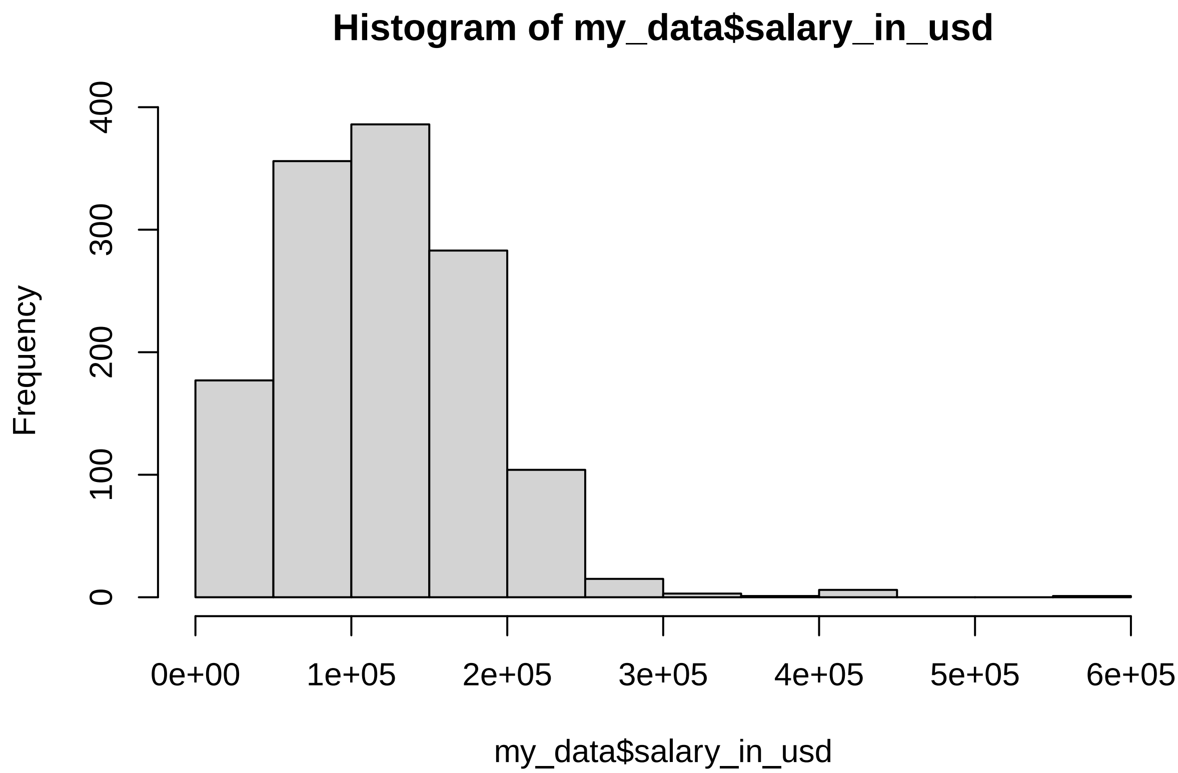
```r
auxiliar <- my_data
my_data[sapply(my_data, is.character)] <- data.matrix(my_data[sapply(my_data, is.character)])
summary(my_data)
```

```
##    work_year    experience_level employment_type   job_title
##  Min.   :2020   Min.   :1.000    Min.   :1.000   Min.   : 1.00
##  1st Qu.:2022   1st Qu.:3.000    1st Qu.:3.000   1st Qu.:21.00
##  Median :2022   Median :4.000    Median :3.000   Median :22.00
##  Mean   :2022   Mean   :3.348    Mean   :2.994   Mean   :26.83
##  3rd Qu.:2022   3rd Qu.:4.000    3rd Qu.:3.000   3rd Qu.:31.00
##  Max.   :2022   Max.   :4.000    Max.   :4.000   Max.   :64.00
##      salary         salary_currency salary_in_usd    employee_residence
##  Min.   :    2324   Min.   : 1.00   Min.   :  2324   Min.   : 1.00
##  1st Qu.:   80000   1st Qu.:18.00   1st Qu.: 75593   1st Qu.:34.00
##  Median :  130000   Median :18.00   Median :120000   Median :63.00
##  Mean   :  237712   Mean   :15.92   Mean   :123375   Mean   :51.61
##  3rd Qu.:  175100   3rd Qu.:18.00   3rd Qu.:164997   3rd Qu.:63.00
##  Max.   :30400000   Max.   :18.00   Max.   :600000   Max.   :64.00
##   remote_ratio    company_location  company_size
##  Min.   :  0.00   Min.   : 1.00    Min.   :1.000
##  1st Qu.:  0.00   1st Qu.:37.75    1st Qu.:2.000
##  Median :100.00   Median :58.00    Median :2.000
##  Mean   : 63.85   Mean   :48.05    Mean   :1.842
##  3rd Qu.:100.00   3rd Qu.:58.00    3rd Qu.:2.000
##  Max.   :100.00   Max.   :59.00    Max.   :3.000
```

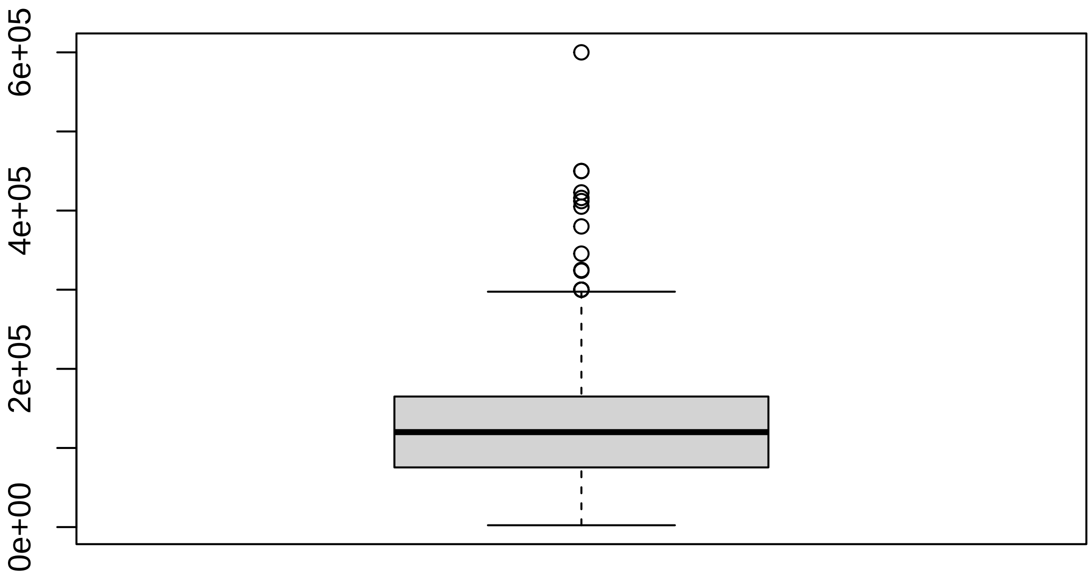
```r
plot(my_data$salary_in_usd)
```



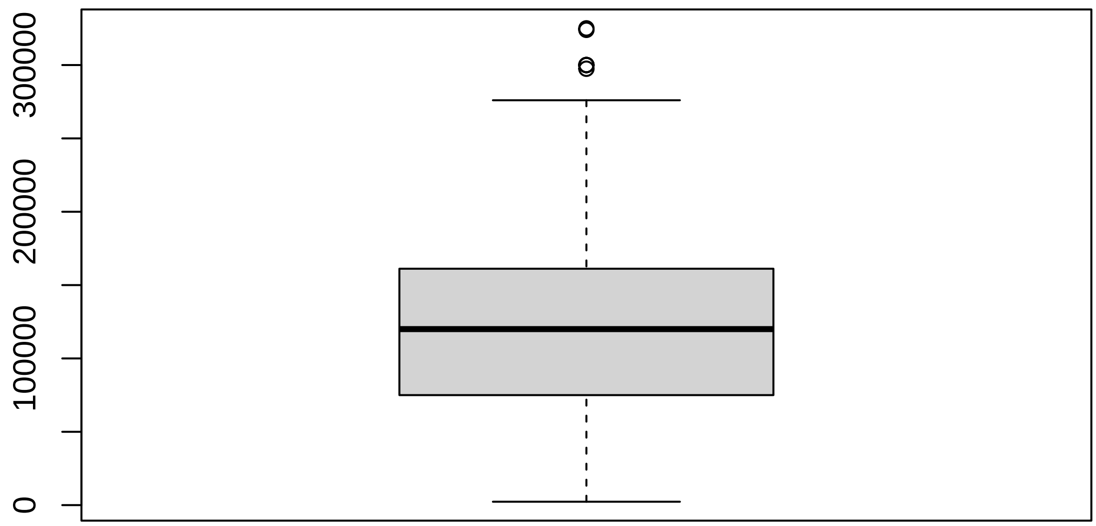

```
hist(my_data$salary_in_usd)
```



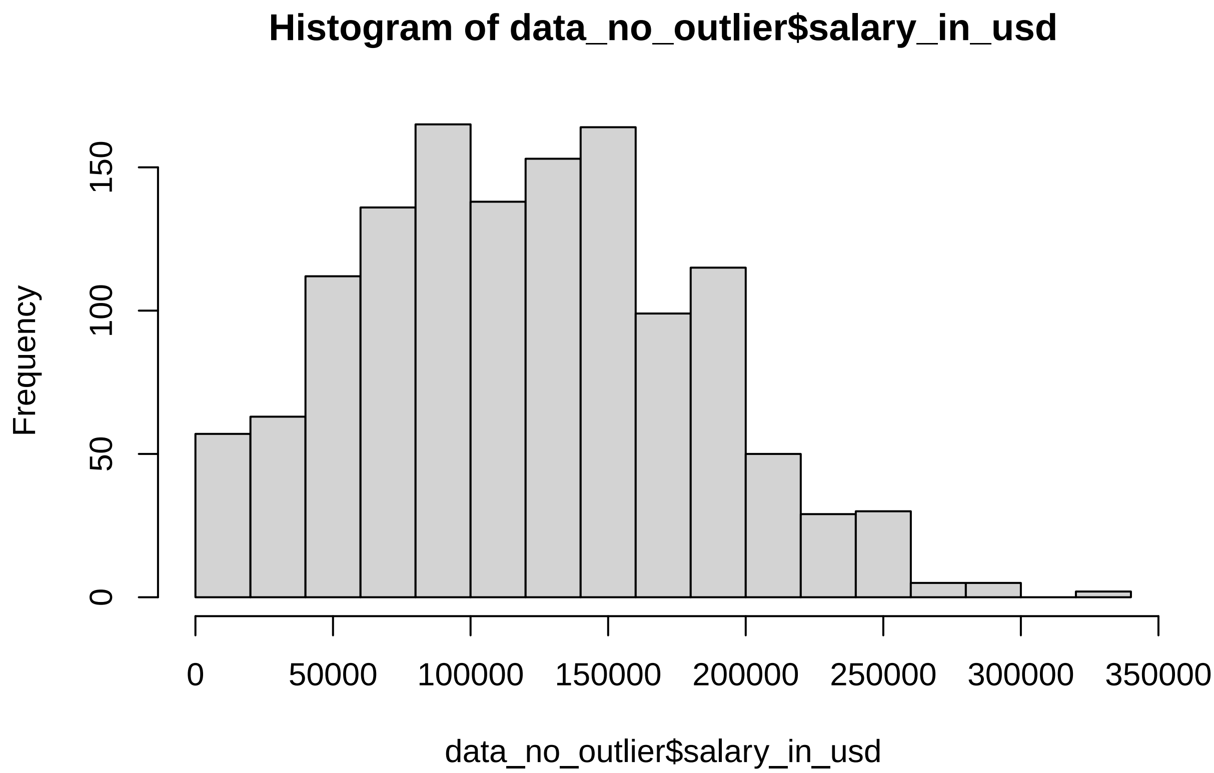

```
#removing outliers
boxplot(my_data$salary_in_usd)
```



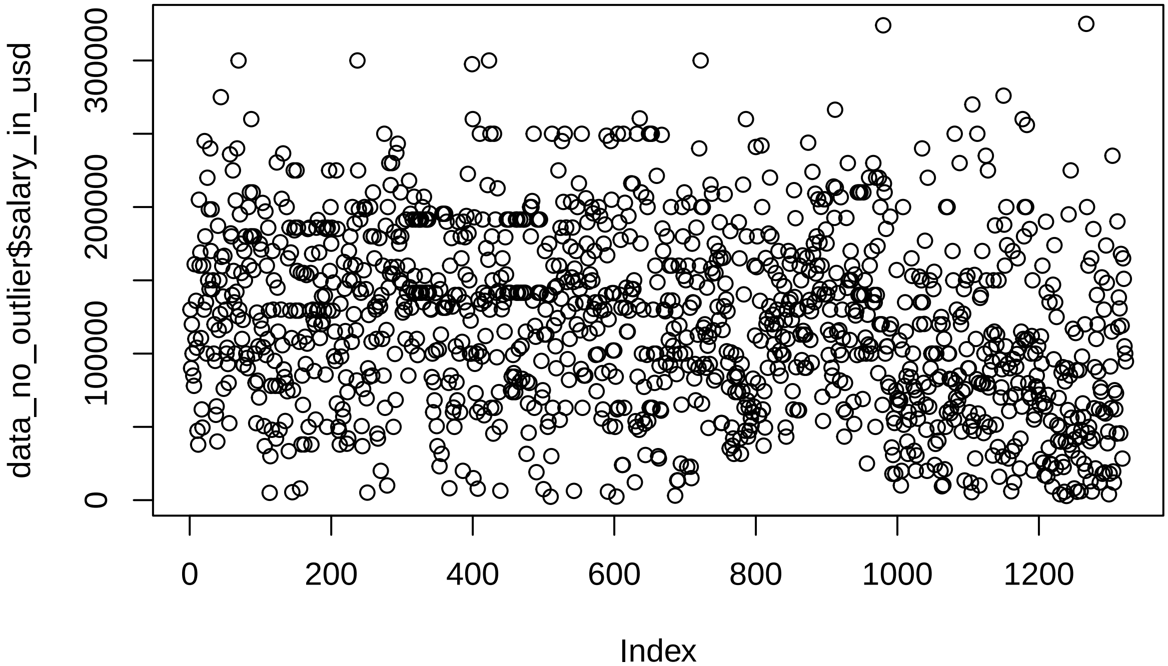

```
quartiles1 <- quantile(my_data$salary_in_usd, probs=c(.01, .90), na.rm = FALSE)
IQR <- IQR(my_data$salary_in_usd)

Lower1 <- quartiles1[1] - 1.5*IQR
Upper1 <- quartiles1[2] + 1.5*IQR

data_no_outlier <- subset(my_data,my_data$salary_in_usd > Lower1 & my_data$salary_in_usd < Upper1)
boxplot(data_no_outlier$salary_in_usd)
```




```
hist(data_no_outlier$salary_in_usd)
```



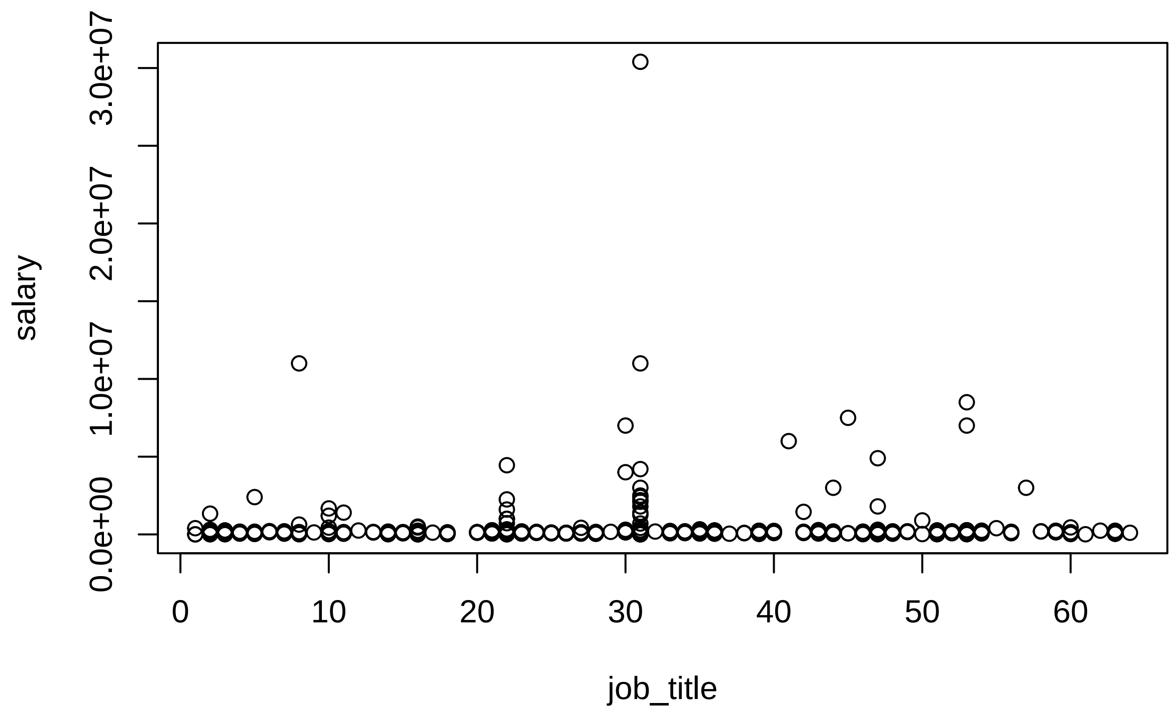

```
plot(data_no_outlier$salary_in_usd)
```




```
salaTitle <-  data_no_outlier[, c(4, 5)]
summary(salaTitle)
```

```
##    job_title         salary
##  Min.   : 1.00   Min.   :    2324
##  1st Qu.:21.00   1st Qu.:   80000
##  Median :22.00   Median :  130000
##  Mean   :26.79   Mean   :  236396
##  3rd Qu.:31.00   3rd Qu.:  175000
##  Max.   :64.00   Max.   :30400000
```

```
plot(salaTitle)
```




```
summary(data_no_outlier$experience_level)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.349   4.000   4.000
```

```
summary(data_no_outlier$employee_residence)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.00   63.00   51.53   63.00   64.00
```

```
summary(data_no_outlier$employment_type)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   2.995   3.000   4.000
```

```
summary(data_no_outlier$company_size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.845   2.000   3.000
```

```
summary(data_no_outlier$remote_ratio)
```

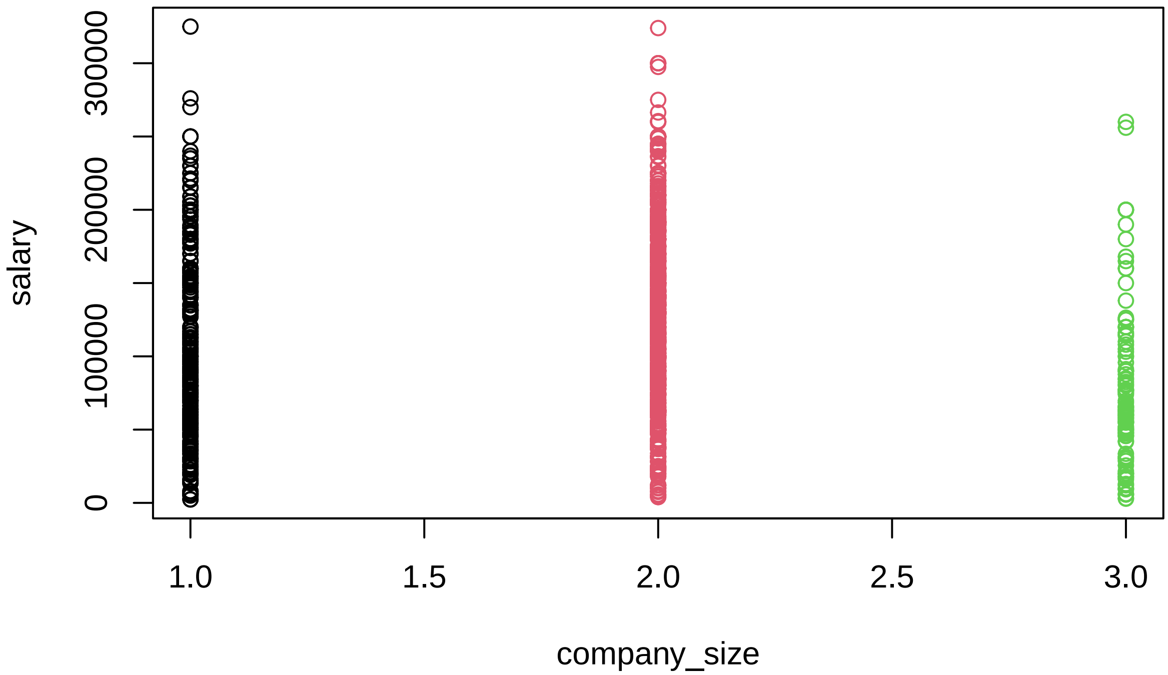
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00  100.00   63.79  100.00  100.00
```

```
summary(data_no_outlier$job_title)
```

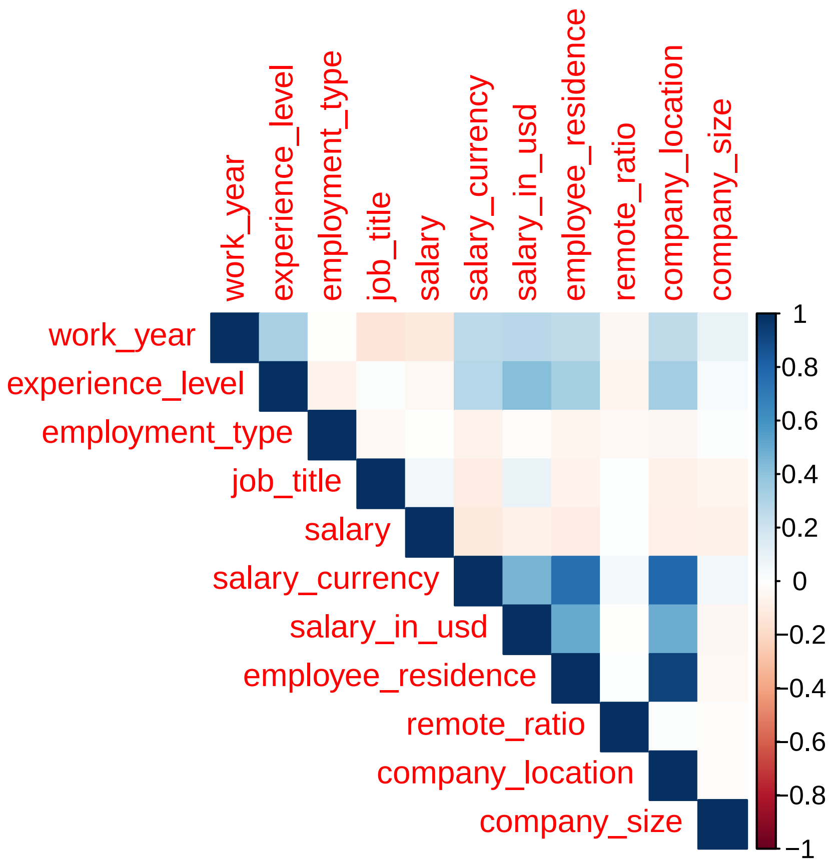
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   21.00   22.00   26.79   31.00   64.00
```

plot de los salarios por experiencia diferenciado en años

```
plot(data_no_outlier$company_size, data_no_outlier$salary_in_usd , col=data_no_outlier$company_size, xla
```




```
corrplot(cor(data_no_outlier), method = "color", type = "upper")
```




```r
#reiniciar los datos
data_no_outlier <- read.csv("AI_MLsalaries.csv",stringsAsFactors = TRUE)

#removing outliers
quartiles1 <- quantile(data_no_outlier$salary_in_usd, probs=c(.01, .90), na.rm = FALSE)
IQR <- IQR(data_no_outlier$salary_in_usd)

Lower1 <- quartiles1[1] - 1.5*IQR
Upper1 <- quartiles1[2] + 1.5*IQR

data_no_outlier <- subset(data_no_outlier,data_no_outlier$salary_in_usd > Lower1 & data_no_outlier$salar

#eliminating columns
data_no_outlier["salary"] <- NULL
data_no_outlier["employee_residence"] <- NULL
data_no_outlier["salary_currency"] <- NULL

#eliminating not FT
head(data_no_outlier)
```

```
##   work_year experience_level employment_type                job_title
## 1      2022               MI              FT Machine Learning Engineer
## 2      2022               MI              FT Machine Learning Engineer
## 3      2022               MI              FT            Data Scientist
## 4      2022               MI              FT            Data Scientist
## 5      2022               MI              FT            Data Scientist
```

```
## 6      2022               MI              FT            Data Scientist
##   salary_in_usd remote_ratio company_location company_size
## 1        130000            0               US            M
## 2         90000            0               US            M
## 3        120000          100               US            M
## 4        100000          100               US            M
## 5         85000          100               US            M
## 6         78000          100               US            M
```

```
data_no_outlier <- data_no_outlier[data_no_outlier$employment_type == "FT",]

#remote_ratio as factor
data_no_outlier["remote_ratio"] <- as.factor(data_no_outlier$remote_ratio)
data_no_outlier["work_year"] <- as.factor(data_no_outlier$work_year)
#NR -> no remote work, PR -> partially remote, FR -> fully remtote
levels(data_no_outlier$remote_ratio) <- list(NR  = "0", PR = "50", FR = "100")

#grouping by continents
data_no_outlier$continent <- countrycode(sourcevar = data_no_outlier[, "company_location"],
                            origin = "iso2c",
                            destination = "continent")
data_no_outlier$continent[data_no_outlier$company_location == "CA" | data_no_outlier$company_location ==
data_no_outlier$continent = as.factor(data_no_outlier$continent)

#gruouping jobs (for clustering approach)
data_no_outlier$job_title_grouped <- data_no_outlier$job_title
data_no_outlier[grepl("BI", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <- "BI An
data_no_outlier[grepl("Data Analy", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <
data_no_outlier[grepl("Sci", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <- "Data
data_no_outlier[grepl("Machine Learning", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_gro
data_no_outlier[grepl("Data Engi", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <-
data_no_outlier[grepl("NLP", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <- "Data
data_no_outlier[grepl("Analytics", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <-
data_no_outlier[grepl("Research", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <-
data_no_outlier[grepl("ETL", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <- "Data
data_no_outlier[grepl("Data Operations", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grou
data_no_outlier[grepl("Computer Vision", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grou
data_no_outlier[grepl("Data Architect", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_group
data_no_outlier[grepl("Head of", data_no_outlier$job_title_grouped, fixed=TRUE),]$job_title_grouped <-
data_no_outlier$job_title_grouped <- factor(data_no_outlier$job_title_grouped)
summary(data_no_outlier$job_title_grouped)
```

```
## 3D Computer Vision Researcher                    BI Analyst
##                             7                            17
##                  Data Analyst                Data Architect
##                           220                            39
##                 Data Engineer                  Data Manager
##                           413                            12
##                Data Scientist               Data Specialist
##                           465                             6
##                   ML Engineer
##                           119
```

```r
levels(data_no_outlier$job_title_grouped)
```

```
## [1] "3D Computer Vision Researcher" "BI Analyst"
## [3] "Data Analyst"                  "Data Architect"
## [5] "Data Engineer"                 "Data Manager"
## [7] "Data Scientist"                "Data Specialist"
## [9] "ML Engineer"
```

```r
#agrupar los precios por rangos
data_no_outlier$quartile <- ntile(data_no_outlier$salary_in_usd, 4)
data_no_outlier["quartile"] <- as.factor(data_no_outlier$quartile)
levels(data_no_outlier$quartile) <- list(Low  = "1", Medium_low = "2", Medium_high = "3", High = "4")

str(data_no_outlier)
```

```
## 'data.frame':    1298 obs. of  11 variables:
##  $ work_year        : Factor w/ 3 levels "2020","2021",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ experience_level : Factor w/ 4 levels "EN","EX","MI",..: 3 3 3 3 3 3 4 4 4 4 ...
##  $ employment_type  : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title        : Factor w/ 64 levels "3D Computer Vision Researcher",..: 47 47 31 31 31 31 22 2
##  $ salary_in_usd    : int  130000 90000 120000 100000 85000 78000 161000 110000 136000 104000 ...
##  $ remote_ratio     : Factor w/ 3 levels "NR","PR","FR": 1 1 3 3 3 3 3 3 3 3 ...
##  $ company_location : Factor w/ 59 levels "AE","AL","AR",..: 58 58 58 58 58 58 58 58 58 58 ...
##  $ company_size     : Factor w/ 3 levels "L","M","S": 2 2 2 2 2 2 2 2 2 2 ...
##  $ continent        : Factor w/ 6 levels "Africa","Americas",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ job_title_grouped: Factor w/ 9 levels "3D Computer Vision Researcher",..: 9 9 7 7 7 7 5 5 7 7 ...
##  $ quartile         : Factor w/ 4 levels "Low","Medium_low",..: 3 2 2 2 2 2 3 2 3 2 ...
```

```r
head(data_no_outlier)
```

```
##   work_year experience_level employment_type                 job_title
## 1      2022               MI              FT Machine Learning Engineer
## 2      2022               MI              FT Machine Learning Engineer
## 3      2022               MI              FT            Data Scientist
## 4      2022               MI              FT            Data Scientist
## 5      2022               MI              FT            Data Scientist
## 6      2022               MI              FT            Data Scientist
##   salary_in_usd remote_ratio company_location company_size     continent
## 1        130000           NR               US            M North America
## 2         90000           NR               US            M North America
## 3        120000           FR               US            M North America
## 4        100000           FR               US            M North America
## 5         85000           FR               US            M North America
## 6         78000           FR               US            M North America
##   job_title_grouped    quartile
## 1       ML Engineer Medium_high
## 2       ML Engineer  Medium_low
## 3    Data Scientist  Medium_low
## 4    Data Scientist  Medium_low
## 5    Data Scientist  Medium_low
## 6    Data Scientist  Medium_low
```

```
summary(data_no_outlier)
```

```
##  work_year   experience_level employment_type                     job_title
##  2020:  69   EN:135           CT:   0         Data Scientist           :339
##  2021: 213   EX: 43           FL:   0         Data Engineer            :317
##  2022:1016   MI:324           FT:1298         Data Analyst             :187
##              SE:796           PT:   0         Machine Learning Engineer: 86
##                                               Analytics Engineer       : 42
##                                               Data Architect           : 36
##                                               (Other)                  :291
##  salary_in_usd    remote_ratio company_location company_size
##  Min.   :  2324   NR:409       US     :919      L:310
##  1st Qu.: 77301   PR:128       GB     : 87      M:885
##  Median :120191   FR:761       CA     : 39      S:103
##  Mean   :122484                IN     : 34
##  3rd Qu.:164996                DE     : 33
##  Max.   :325000                ES     : 27
##                                (Other):159
##          continent        job_title_grouped        quartile
##  Africa       :  5   Data Scientist:465     Low        :325
##  Americas     : 26   Data Engineer :413     Medium_low :325
##  Asia         : 64   Data Analyst  :220     Medium_high:324
##  Europe       :233   ML Engineer   :119     High       :324
##  North America:958   Data Architect: 39
##  Oceania      : 12   BI Analyst    : 17
##                      (Other)       : 25
```

```
summary(lm(formula = salary_in_usd ~ work_year + experience_level + job_title_grouped + remote_ratio + c
           data = data_no_outlier))
```

```
##
## Call:
## lm(formula = salary_in_usd ~ work_year + experience_level + job_title_grouped +
##     remote_ratio + company_size + company_location, data = data_no_outlier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -147403  -23652   -2076   24894  148165
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       56037.0    30561.8   1.834 0.066961 .
## work_year2021                     -3286.7     6224.7  -0.528 0.597589
## work_year2022                     -2151.1     6069.0  -0.354 0.723067
## experience_levelEX                89770.7     7986.1  11.241  < 2e-16 ***
## experience_levelMI                18002.6     4778.5   3.767 0.000173 ***
## experience_levelSE                45917.7     4697.2   9.776  < 2e-16 ***
## job_title_groupedBI Analyst       -6229.4    19771.8  -0.315 0.752765
## job_title_groupedData Analyst    -14386.1    17304.3  -0.831 0.405935
## job_title_groupedData Architect   25308.0    18455.9   1.371 0.170542
## job_title_groupedData Engineer    10874.0    17230.4   0.631 0.528097
## job_title_groupedData Manager     21602.1    21927.5   0.985 0.324740
```

```
## job_title_groupedData Scientist   18768.8    17206.3   1.091 0.275571
## job_title_groupedData Specialist    437.1    24357.2   0.018 0.985687
## job_title_groupedML Engineer      25053.8    17613.7   1.422 0.155164
## remote_ratioPR                   -10508.1     5276.9  -1.991 0.046666 *
## remote_ratioFR                    -1574.1     2675.5  -0.588 0.556412
## company_sizeM                     -3277.7     3375.9  -0.971 0.331790
## company_sizeS                    -19687.5     5387.0  -3.655 0.000268 ***
## company_locationAL               -62435.9    49168.8  -1.270 0.204387
## company_locationAR                12074.3    49163.4   0.246 0.806037
## company_locationAS               -10065.1    38905.6  -0.259 0.795906
## company_locationAT               -12350.2    31165.9  -0.396 0.691973
## company_locationAU                15958.6    28664.3   0.557 0.577807
## company_locationBE                -1671.7    32516.3  -0.051 0.959007
## company_locationBR               -34468.4    27440.7  -1.256 0.209318
## company_locationCA                13767.2    25605.4   0.538 0.590905
## company_locationCH               -14089.3    38917.3  -0.362 0.717390
## company_locationCL               -47909.7    49020.0  -0.977 0.328589
## company_locationCN               -12002.0    38853.4  -0.309 0.757448
## company_locationCO               -42174.4    49444.6  -0.853 0.393847
## company_locationCZ               -40896.7    38897.1  -1.051 0.293278
## company_locationDE                 3278.2    25720.7   0.127 0.898601
## company_locationDK                 3615.9    38740.4   0.093 0.925651
## company_locationEE               -40643.0    50659.2  -0.802 0.422545
## company_locationEG               -63005.6    48944.3  -1.287 0.198236
## company_locationES               -37754.3    26040.3  -1.450 0.147358
## company_locationFI               -31489.8    48974.0  -0.643 0.520351
## company_locationFR               -13918.7    26822.3  -0.519 0.603909
## company_locationGB                 2084.2    25180.5   0.083 0.934047
## company_locationGR               -18575.6    27466.7  -0.676 0.498981
## company_locationHN               -19966.1    48994.8  -0.408 0.683701
## company_locationHR               -60128.9    48836.4  -1.231 0.218473
## company_locationHU               -46565.4    49213.4  -0.946 0.344236
## company_locationID               -29853.0    38868.7  -0.768 0.442606
## company_locationIE               -35159.8    48705.7  -0.722 0.470505
## company_locationIL                34389.0    49134.7   0.700 0.484127
## company_locationIN               -46593.3    25783.2  -1.807 0.070989 .
## company_locationIQ                52391.5    49176.9   1.065 0.286920
## company_locationIR               -72775.2    49146.1  -1.481 0.138919
## company_locationIT               -34234.7    49102.1  -0.697 0.485801
## company_locationJP                40582.6    30023.6   1.352 0.176723
## company_locationKE               -15987.2    50222.0  -0.318 0.750288
## company_locationLU                -5536.4    35158.7  -0.157 0.874902
## company_locationMD               -43939.4    48991.9  -0.897 0.369964
## company_locationMT               -42749.9    49222.3  -0.869 0.385287
## company_locationMX               -60671.8    32302.8  -1.878 0.060589 .
## company_locationMY               -31080.6    49107.8  -0.633 0.526914
## company_locationNG                26450.2    34726.8   0.762 0.446408
## company_locationNL               -23621.3    28823.8  -0.820 0.412657
## company_locationNZ                42366.8    49184.8   0.861 0.389198
## company_locationPH               -13020.4    48979.0  -0.266 0.790411
## company_locationPK               -28683.9    38969.4  -0.736 0.461834
## company_locationPL               -40165.9    32629.9  -1.231 0.218577
## company_locationPR                54584.3    32439.3   1.683 0.092697 .
## company_locationPT               -44876.5    28295.7  -1.586 0.113001
```

```
## company_locationRO                    -7841.3    49367.2  -0.159 0.873825
## company_locationRU                     1686.4    39907.4   0.042 0.966301
## company_locationSG                   -23126.3    34875.1  -0.663 0.507379
## company_locationSI                   -38166.4    39386.2  -0.969 0.332723
## company_locationTH                   -61230.0    49132.5  -1.246 0.212921
## company_locationTR                   -66202.9    31328.1  -2.113 0.034784 *
## company_locationUA                   -56545.0    49212.8  -1.149 0.250784
## company_locationUS                    44434.9    24743.4   1.796 0.072769 .
## company_locationVN                   -64241.5    49379.8  -1.301 0.193515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42080 on 1224 degrees of freedom
## Multiple R-squared:  0.5395, Adjusted R-squared:  0.512
## F-statistic: 19.64 on 73 and 1224 DF,  p-value: < 2.2e-16
```

Vemos que no funciona lm, porque está intentando clasificar, y no un problema de regresión lineal, ya que la variable es categórica. Por tanto, intentaremos con un Naive Bayes.

```r
set.seed(123)
trainIndex=createDataPartition(data_no_outlier$quartile, p=0.7)$Resample1

train=data_no_outlier[trainIndex, ]
test=data_no_outlier[-trainIndex, ]
NBclassfier=naiveBayes(quartile~., data=train)
```

```r
printALL=function(model){
  trainPred=predict(model, newdata = train, type = "class")
  trainTable=table(train$quartile, trainPred)
  testPred=predict(NBclassfier, newdata=test, type="class")
  testTable=table(test$quartile, testPred)
  trainAcc=(trainTable[1,1]+trainTable[2,2]+trainTable[3,3])/sum(trainTable)
  testAcc=(testTable[1,1]+testTable[2,2]+testTable[3,3])/sum(testTable)
  message("Contingency Table for Training Data")
  print(trainTable)
  message("Contingency Table for Test Data")
  print(testTable)
  message("Accuracy")
  print(round(cbind(trainAccuracy=trainAcc, testAccuracy=testAcc),4))
}
printALL(NBclassfier)
```

```
## Contingency Table for Training Data
```

```
##              trainPred
##               Low Medium_low Medium_high High
##   Low         210         18           0    0
##   Medium_low   16        206           6    0
##   Medium_high   0          8         214    5
##   High          0          0           8  219
```

```
## Contingency Table for Test Data
```

```
##              testPred
##               Low Medium_low Medium_high High
##   Low          92          5           0    0
##   Medium_low   12         80           5    0
##   Medium_high   0          1          92    4
##   High          0          0           5   92
```

```
## Accuracy
```

```
##      trainAccuracy testAccuracy
## [1,]        0.6923       0.6804
```

#CLUSTERING

##Manual Distance matrix

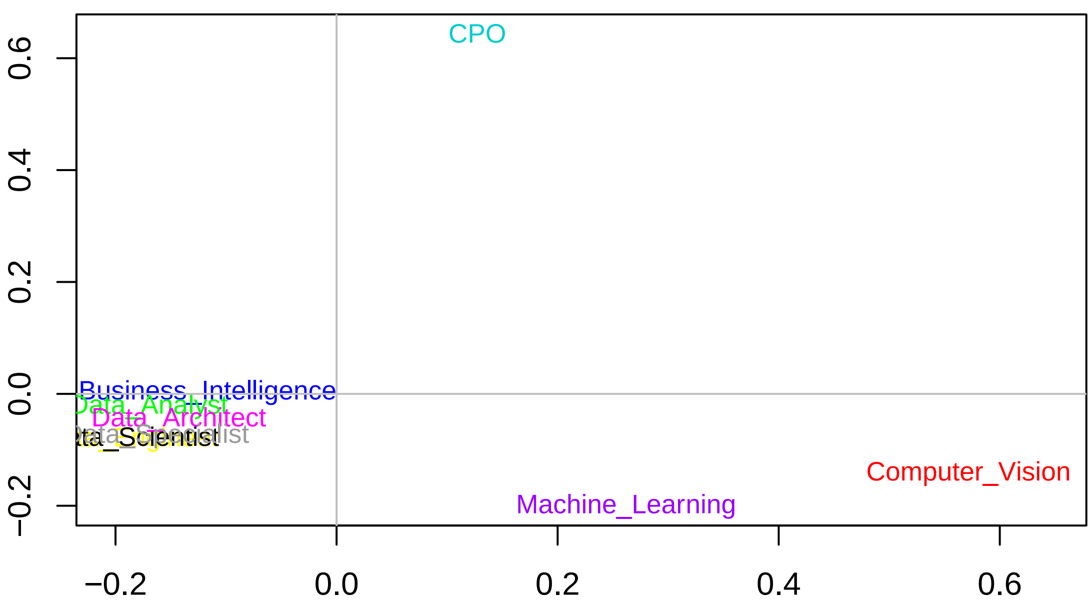
```
# Creamos los vectores que formarÃ¡n la matriz de distancia (0 son iguales los trabajos, 1 son opuestos
Machine_Learning <- c(0.25, 0.5,    0.5,    0.4,    0.4,    0.85,   0.5,    0.5, 0)
Data_Specialist <- c(0.75,  0.4,    0.3,    0.3,    0.3,    0.8,    0.15, 0, 0.5)
Data_Scientist <- c(0.75,   0.3,    0.15,   0.25,   0.3,    0.8, 0, 0.15, 0.5)
CPO <- c(0.9,   0.75,   0.75,   0.75,   0.8, 0, 0.8,0.8,0.85)
Data_Engineer <- c(0.8, 0.4,    0.25,   0.1, 0,0.8,0.3,0.3,0.4)
Data_Architect <- c(0.75,   0.4,    0.25, 0,0.1,0.75,0.25,0.3,0.4)
Data_Analyst <- c(0.75, 0.3, 0,0.25,0.25,0.75,0.15,0.3,0.5)
Business_Intelligence <- c(0.75, 0, 0.3,0.4,0.4,0.75,0.3,0.4,0.5)
Computer_Vision <- c(0, 0.75,0.75,0.75,0.8,0.9,0.75,0.75,0.25)

D <- c(Computer_Vision, Business_Intelligence, Data_Analyst, Data_Architect, Data_Engineer, CPO, Data_S

My_Matrix <- matrix(D, byrow=TRUE, nrow=9)
rownames(My_Matrix) <- c("Computer_Vision", "Business_Intelligence", "Data_Analyst", "Data_Architect",
colnames(My_Matrix) <- c("Computer_Vision", "Business_Intelligence", "Data_Analyst", "Data_Architect",

My_Matrix
```

```
##                       Computer_Vision Business_Intelligence Data_Analyst
## Computer_Vision                  0.00                  0.75         0.75
## Business_Intelligence            0.75                  0.00         0.30
## Data_Analyst                     0.75                  0.30         0.00
## Data_Architect                   0.75                  0.40         0.25
## Data_Engineer                    0.80                  0.40         0.25
## CPO                              0.90                  0.75         0.75
## Data_Scientist                   0.75                  0.30         0.15
## Data_Specialist                  0.75                  0.40         0.30
## Machine_Learning                 0.25                  0.50         0.50
##                       Data_Architect Data_Engineer  CPO Data_Scientist
## Computer_Vision                 0.75          0.80 0.90           0.75
## Business_Intelligence           0.40          0.40 0.75           0.30
## Data_Analyst                    0.25          0.25 0.75           0.15
## Data_Architect                  0.00          0.10 0.75           0.25
## Data_Engineer                   0.10          0.00 0.80           0.30
## CPO                             0.75          0.80 0.00           0.80
## Data_Scientist                  0.25          0.30 0.80           0.00
```

```
## Data_Specialist                   0.30          0.30 0.80             0.15
## Machine_Learning                  0.40          0.40 0.85             0.50
##                         Data_Specialist Machine_Learning
## Computer_Vision                    0.75             0.25
## Business_Intelligence              0.40             0.50
## Data_Analyst                       0.30             0.50
## Data_Architect                     0.30             0.40
## Data_Engineer                      0.30             0.40
## CPO                                0.80             0.85
## Data_Scientist                     0.15             0.50
## Data_Specialist                    0.00             0.50
## Machine_Learning                   0.50             0.00
```

##Plotted Distance Matrix

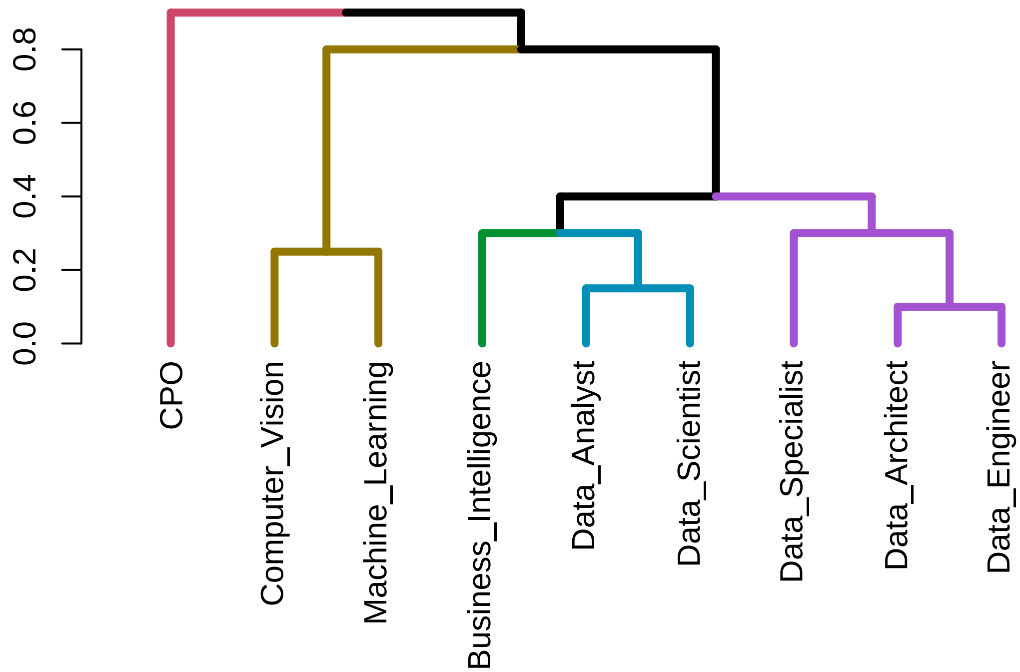
```
Distance_Matrix <- as.dist(My_Matrix)

mds.coor <- cmdscale(Distance_Matrix)
plot(mds.coor[,2], mds.coor[,2], type="n", xlab="", ylab="")
text(jitter(mds.coor[,1]), jitter(mds.coor[,2]), rownames(mds.coor), cex=0.8, col = c("#FF0000", "#0000
abline(h=0,v=0,col="gray75")
```



##Hierarchical Clustering using the Distance matrix of the Job Titles

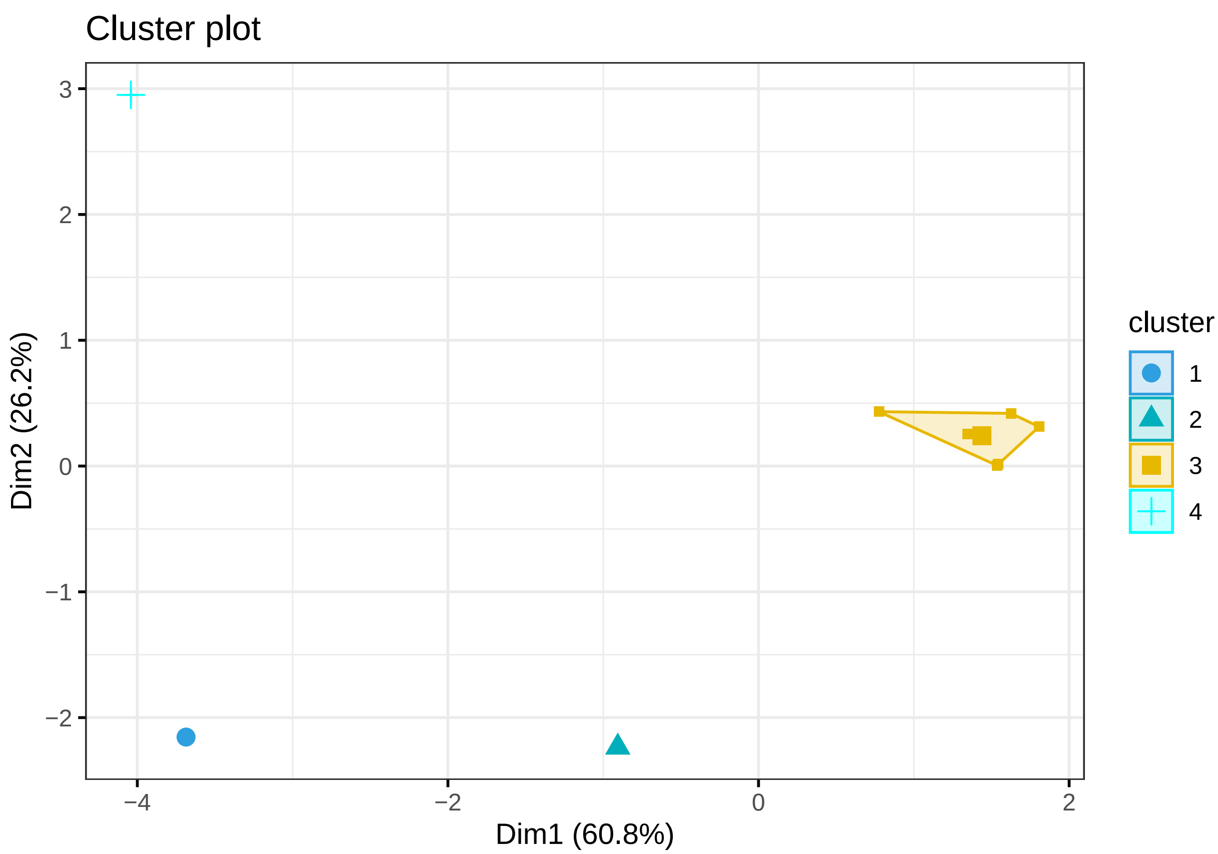
```
hc <- hclust(Distance_Matrix)
dend <-set(as.dendrogram(hc), "branches_lwd", 4)
```

```
d1=color_branches(dend,k=5, col = c(3,1,1,4,1))
d2=color_branches(d1,k=5) # auto-coloring 5 clusters of branches.
par(mar = c(9, 4, 4, 2) + 0.1)
plot(d2, lwd=2)
```



##K-Means using Distance matrix of the Job Titles

```
kmeans.re <- kmeans(Distance_Matrix, centers = 4, nstart = 20)

fviz_cluster(kmeans.re, Distance_Matrix,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#00FFFF"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
             )
```


Cluster plot
Dim2 (26.2%)
Dim1 (60.8%)
3
2
1
0
−1
−2
−4
−2
0
2
cluster
1
2
3
4