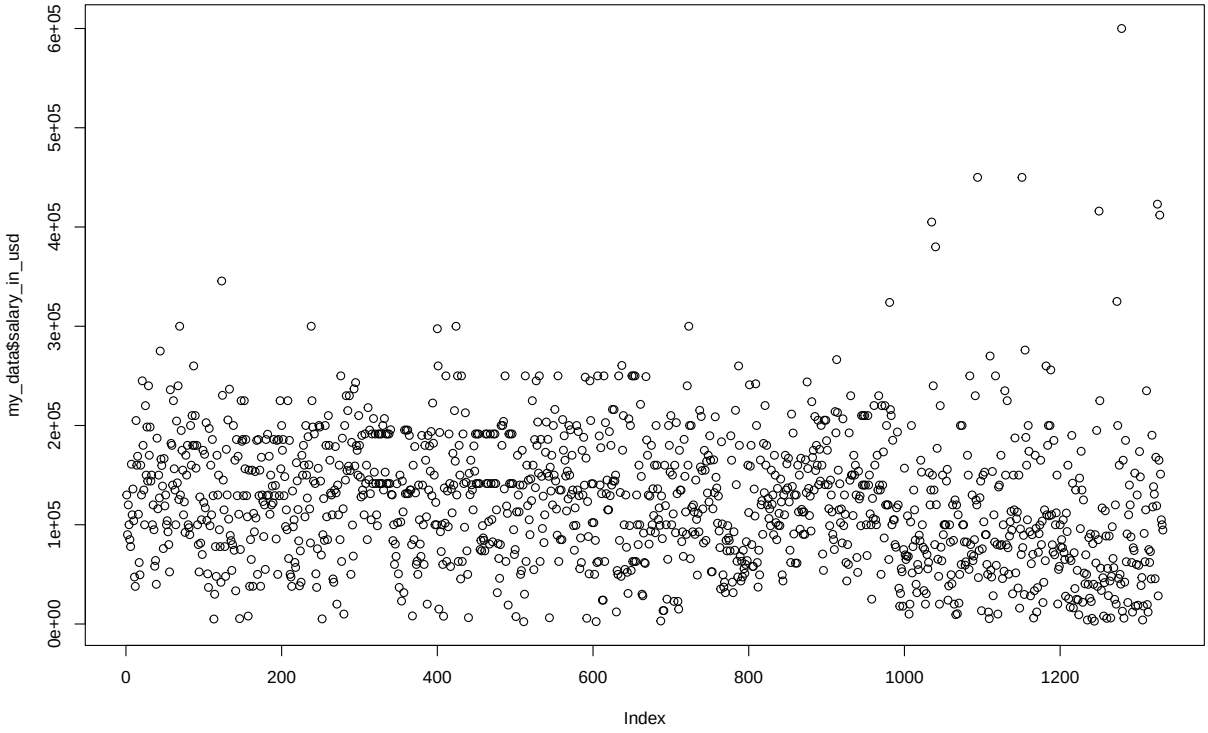
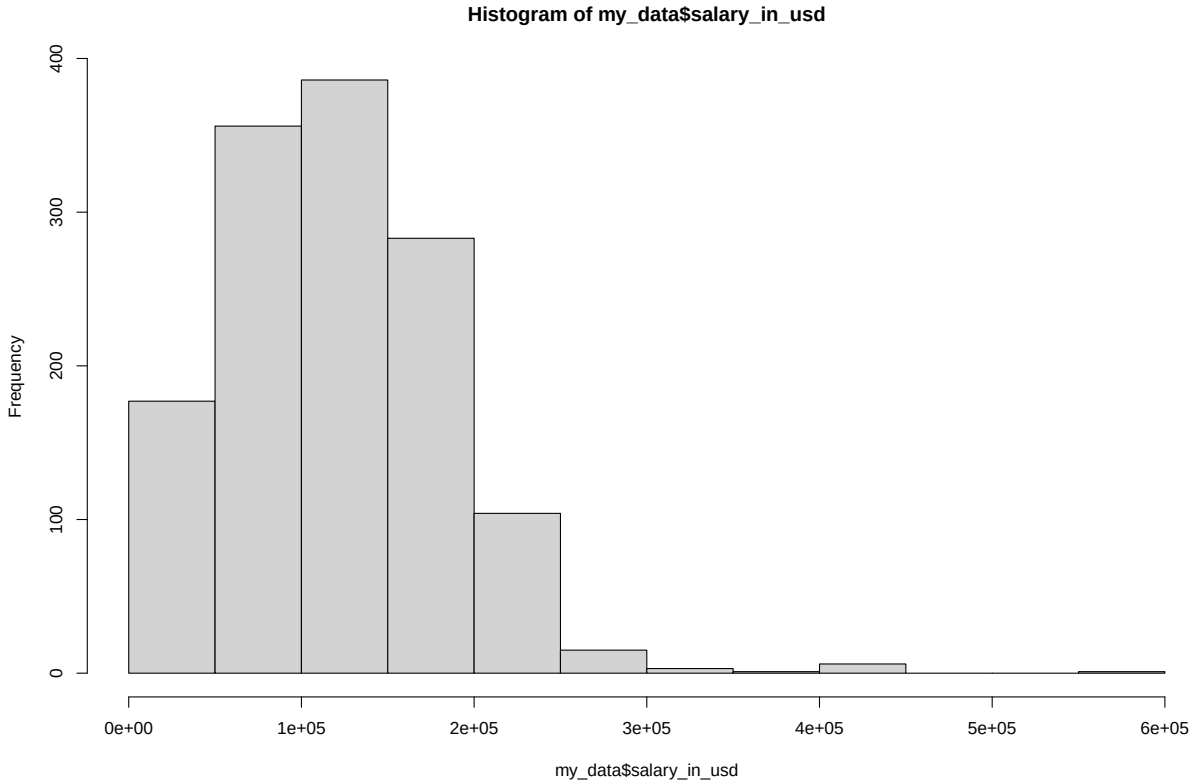
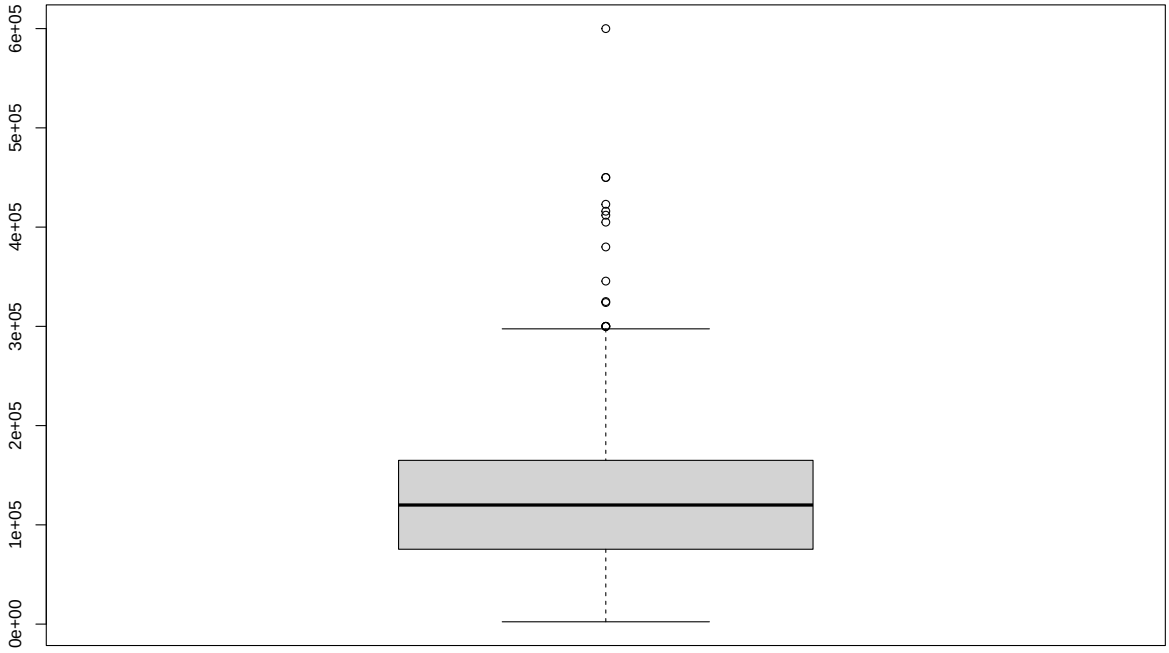
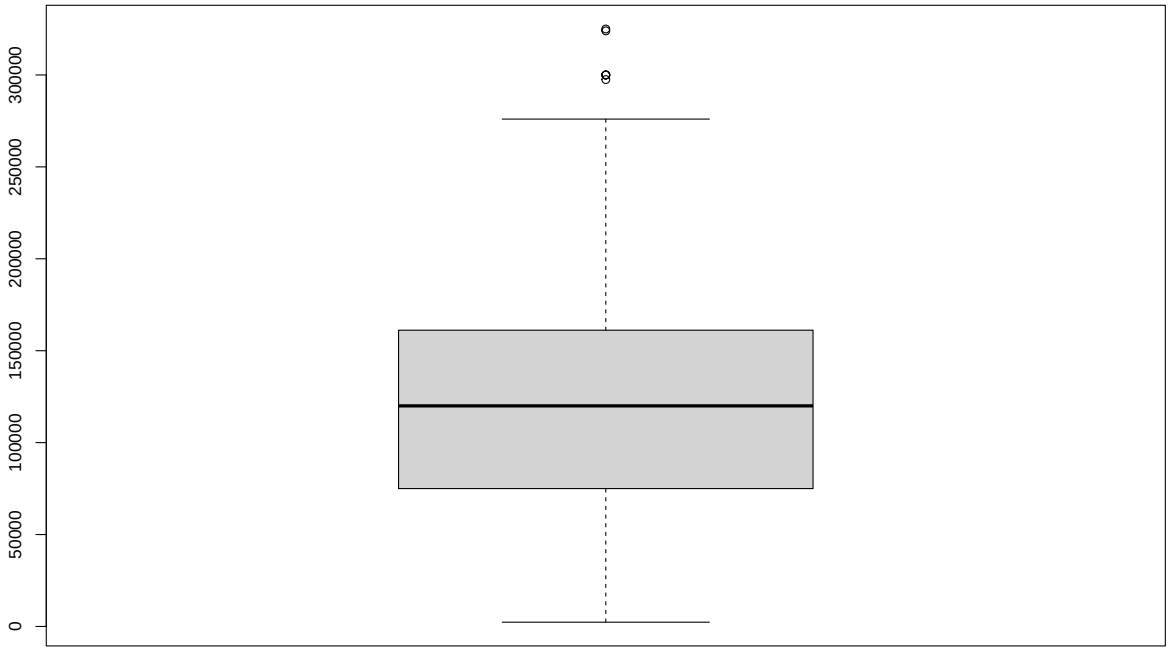
Main

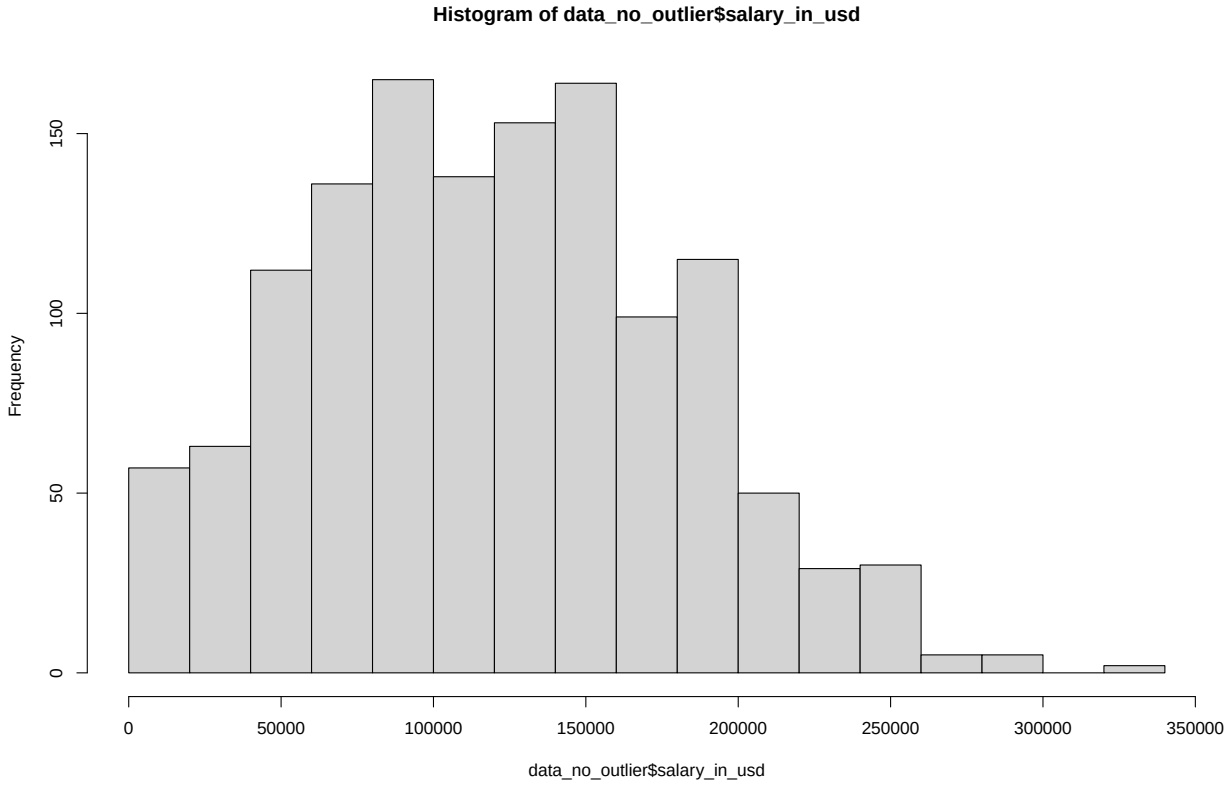
2022-12-11

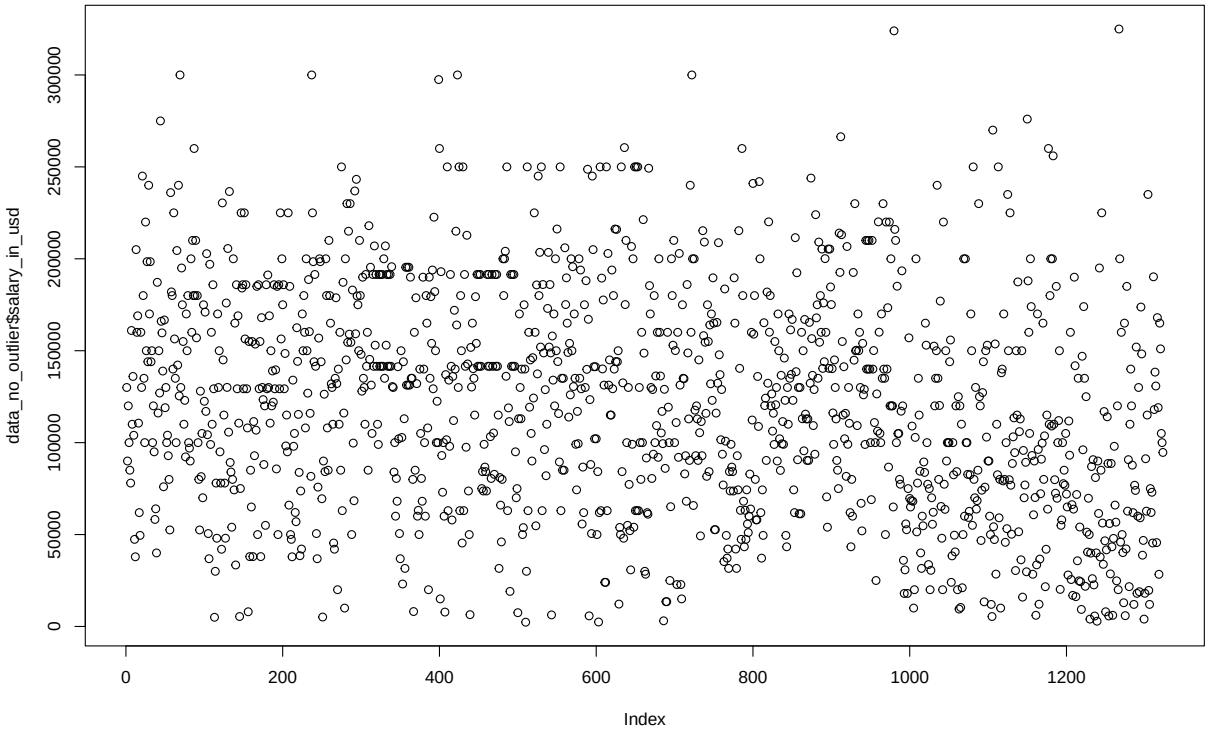



Histogram of my_data$salary_in_usd
Frequency
0
100
200
300
400
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
my_data$salary_in_usd

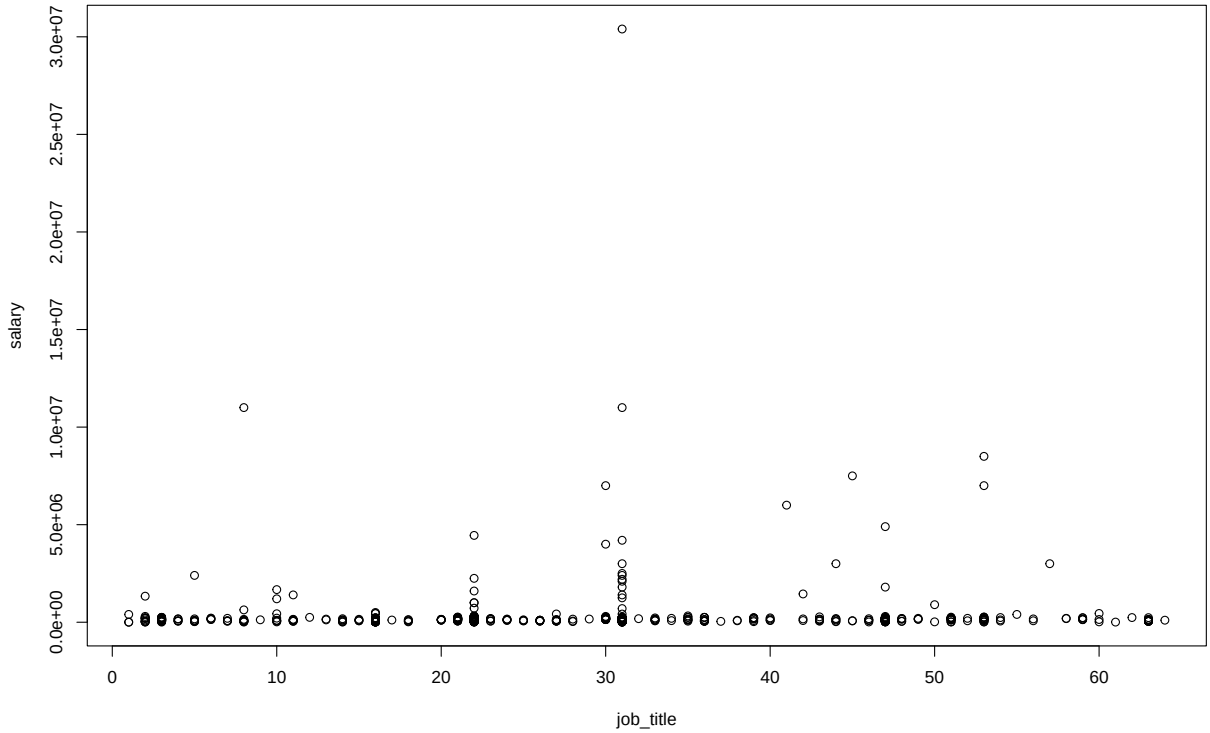

0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05


0
50000
100000
150000
200000
250000
300000


Histogram of data_no_outlier$salary_in_usd
Frequency
0
50
100
150
0
50000
100000
150000
200000
250000
300000
350000
data_no_outlier$salary_in_usd




```
##    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
```




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

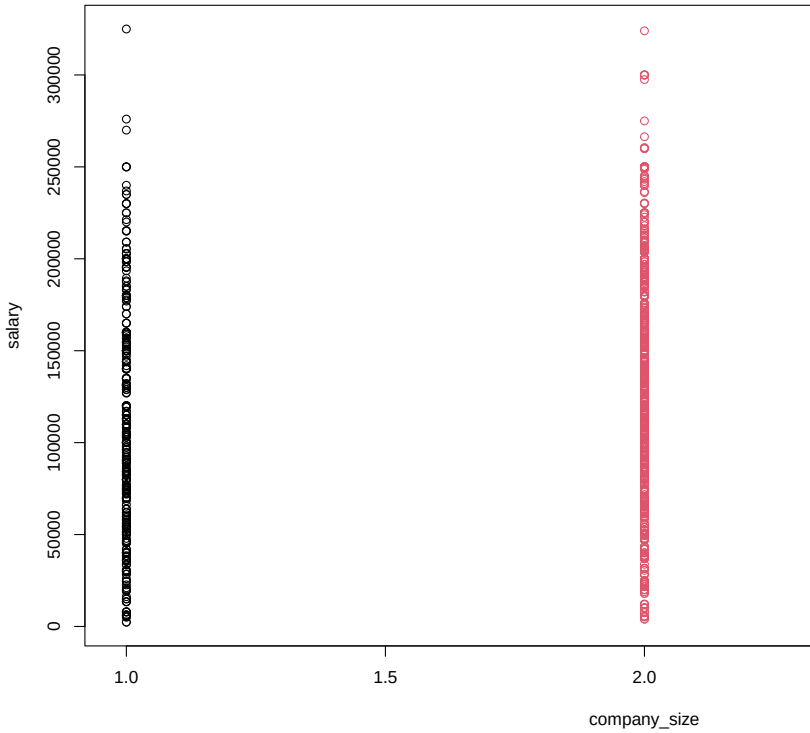
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.00   63.00   51.53   63.00   64.00
```

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

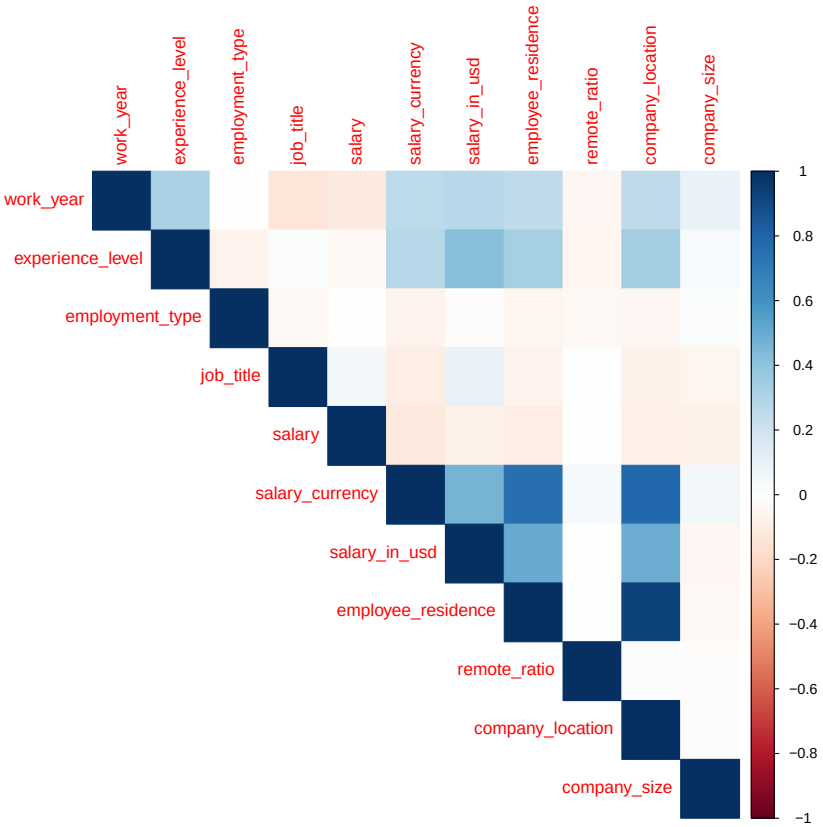
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.845   2.000   3.000
```

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

```
##    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



DATA CLEANING

```
##   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
```

```
## 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
```

```
## [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"
```

```
## '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 ...
```

```
##   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
```

```
##  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
```

```
##
## Call:
## lm(formula = salary_in_usd ~ work_year + experience_level + job_title +
##     remote_ratio + company_size + company_location, data = data_no_outlier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -127645  -22634   -1509   21389  144902
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  -74844.8    48092.9  -1.556
## work_year2021                                 -5585.7     6245.8  -0.894
## work_year2022                                   518.2     6148.6   0.084
## experience_levelEX                            76753.1     8735.7   8.786
```

```
## experience_levelMI                                  16527.0     4785.2   3.454
## experience_levelSE                                  40865.8     4771.2   8.565
## job_titleAI Scientist                              156788.4    43279.6   3.623
## job_titleAnalytics Engineer                        136922.9    41421.6   3.306
## job_titleApplied Data Scientist                    156980.3    44926.3   3.494
## job_titleApplied Machine Learning Scientist        106268.4    44148.4   2.407
## job_titleApplied Scientist                         162130.1    45680.5   3.549
## job_titleBI Analyst                                166804.4    46719.5   3.570
## job_titleBI Data Analyst                           107412.7    42805.3   2.509
## job_titleBig Data Architect                        134366.0    58265.5   2.306
## job_titleBig Data Engineer                         140025.8    43989.6   3.183
## job_titleBusiness Data Analyst                     133740.3    44163.8   3.028
## job_titleCloud Data Architect                      252398.3    57727.1   4.372
## job_titleCloud Data Engineer                       163844.3    52629.5   3.113
## job_titleComputer Vision Engineer                  109593.7    46593.3   2.352
## job_titleComputer Vision Software Engineer         158142.7    47522.6   3.328
## job_titleData Analyst                              111959.0    41016.8   2.730
## job_titleData Analytics Consultant                 126234.8    57789.8   2.184
## job_titleData Analytics Engineer                   123281.9    45062.5   2.736
## job_titleData Analytics Manager                    119070.1    43556.6   2.734
## job_titleData Architect                            154209.9    41420.2   3.723
## job_titleData Engineer                             142624.8    40950.7   3.483
## job_titleData Engineering Manager                  150610.6    44061.7   3.418
## job_titleData Manager                              125991.1    45008.9   2.799
## job_titleData Operations Analyst                    79315.9    45690.5   1.736
## job_titleData Operations Engineer                   89495.8    43482.9   2.058
## job_titleData Science Consultant                   138878.1    43145.6   3.219
## job_titleData Science Engineer                     145678.4    47845.2   3.045
## job_titleData Science Lead                         181504.7    58285.4   3.114
## job_titleData Science Manager                      176013.7    41532.7   4.238
## job_titleData Scientist                            144232.3    40950.4   3.522
## job_titleData Scientist Lead                       171896.0    57654.1   2.982
## job_titleData Specialist                           131252.4    44056.3   2.979
## job_titleDirector of Data Engineering              175229.0    49882.4   3.513
## job_titleDirector of Data Science                  183083.4    43843.1   4.176
## job_titleETL Developer                             138487.5    42951.3   3.224
## job_titleFinance Data Analyst                      110192.3    57879.0   1.904
## job_titleFinancial Data Analyst                    112012.8    50258.8   2.229
## job_titleHead of Data                              179322.5    45005.8   3.984
## job_titleHead of Data Science                      131409.6    46654.2   2.817
## job_titleHead of Machine Learning                  126667.1    59055.0   2.145
## job_titleLead Data Analyst                         131331.3    47112.0   2.788
## job_titleLead Data Engineer                        161346.7    44787.8   3.602
## job_titleLead Data Scientist                       156515.0    44535.6   3.514
## job_titleLead Machine Learning Engineer            150098.7    47674.4   3.148
## job_titleMachine Learning Developer                143728.8    47464.2   3.028
## job_titleMachine Learning Engineer                 157026.0    41141.5   3.817
## job_titleMachine Learning Infrastructure Engineer 166145.8    45130.4   3.681
## job_titleMachine Learning Manager                  158352.5    47259.0   3.351
## job_titleMachine Learning Research Engineer        130975.3    50502.5   2.593
## job_titleMachine Learning Scientist                182094.6    42442.6   4.290
## job_titleMarketing Data Analyst                    215555.5    53009.2   4.066
## job_titleML Engineer                               152114.2    42389.4   3.588
## job_titleNLP Engineer                               62375.4    53032.6   1.176
```

```
## job_titlePower BI Developer                  137966.2    58311.5   2.366
## job_titlePrincipal Data Analyst              152855.0    50121.2   3.050
## job_titlePrincipal Data Architect            116672.2    58314.1   2.001
## job_titlePrincipal Data Engineer             188042.0    49823.3   3.774
## job_titlePrincipal Data Scientist            200590.6    44144.5   4.544
## job_titleProduct Data Analyst                131718.5    50294.1   2.619
## job_titleProduct Data Scientist               33343.1    66743.1   0.500
## job_titleResearch Engineer                   254318.9    57821.4   4.398
## job_titleResearch Scientist                  151169.1    42103.3   3.590
## remote_ratioPR                               -9851.7     5364.2  -1.837
## remote_ratioFR                               -2453.4     2626.9  -0.934
## company_sizeM                                -1084.0     3431.4  -0.316
## company_sizeS                               -20210.4     5533.3  -3.653
## company_locationAL                           76650.6    62930.3   1.218
## company_locationAR                           14821.0    47862.4   0.310
## company_locationAS                          -21300.6    39602.8  -0.538
## company_locationAT                          -12005.2    31016.7  -0.387
## company_locationAU                           14807.4    28483.5   0.520
## company_locationBE                            1084.4    31959.4   0.034
## company_locationBR                          -34128.5    27359.8  -1.247
## company_locationCA                           15837.4    25549.3   0.620
## company_locationCH                           -9962.5    38925.8  -0.256
## company_locationCL                          -37837.4    47771.7  -0.792
## company_locationCN                          -12515.9    38300.6  -0.327
## company_locationCO                          -43815.8    48147.8  -0.910
## company_locationCZ                           -1660.6    41959.6  -0.040
## company_locationDE                            1602.2    25689.4   0.062
## company_locationDK                          -45434.4    42666.9  -1.065
## company_locationEE                          -67273.1    51280.0  -1.312
## company_locationEG                          -97957.6    49045.0  -1.997
## company_locationES                          -34677.9    25872.5  -1.340
## company_locationFI                          -30331.1    47801.6  -0.635
## company_locationFR                          -12170.3    26760.0  -0.455
## company_locationGB                            1120.2    24992.9   0.045
## company_locationGR                          -18709.3    27301.7  -0.685
## company_locationHN                          -33190.3    55766.7  -0.595
## company_locationHR                          -54765.2    47611.1  -1.150
## company_locationHU                          -40327.8    48074.4  -0.839
## company_locationID                          -26367.4    38026.0  -0.693
## company_locationIE                          -32465.4    47439.0  -0.684
## company_locationIL                           42267.7    47861.8   0.883
## company_locationIN                          -42274.9    25663.4  -1.647
## company_locationIQ                           66763.9    53761.1   1.242
## company_locationIR                          -71183.8    47843.9  -1.488
## company_locationIT                          -32206.8    47844.5  -0.673
## company_locationJP                           34846.2    30220.4   1.153
## company_locationKE                            4953.7    49345.3   0.100
## company_locationLU                           -2203.4    35386.9  -0.062
## company_locationMD                          -37911.9    50452.4  -0.751
## company_locationMT                          -40500.6    48008.5  -0.844
## company_locationMX                          -53981.8    32416.4  -1.665
## company_locationMY                          -27452.3    47828.7  -0.574
## company_locationNG                           18998.2    34766.0   0.546
## company_locationNL                          -20387.8    28562.3  -0.714
```

```
## company_locationNZ                                  27694.4    51259.3   0.540
## company_locationPH                                    269.6    47786.6   0.006
## company_locationPK                                 -31830.4    39357.7  -0.809
## company_locationPL                                 -41361.9    32188.0  -1.285
## company_locationPR                                  54672.8    31830.4   1.718
## company_locationPT                                 -45924.9    28258.8  -1.625
## company_locationRO                                  -5186.6    50758.5  -0.102
## company_locationRU                                  11279.2    41128.1   0.274
## company_locationSG                                  10571.1    41991.4   0.252
## company_locationSI                                 -46456.6    39103.6  -1.188
## company_locationTH                                 -45844.8    53466.8  -0.857
## company_locationTR                                 -62638.6    30877.9  -2.029
## company_locationUA                                 -47948.3    47944.6  -1.000
## company_locationUS                                  47018.2    24607.5   1.911
## company_locationVN                                 -58717.8    48102.4  -1.221
##                                                   Pr(>|t|)
## (Intercept)                                       0.119918
## work_year2021                                     0.371338
## work_year2022                                     0.932845
## experience_levelEX                                 < 2e-16 ***
## experience_levelMI                                0.000573 ***
## experience_levelSE                                 < 2e-16 ***
## job_titleAI Scientist                             0.000304 ***
## job_titleAnalytics Engineer                       0.000976 ***
## job_titleApplied Data Scientist                   0.000493 ***
## job_titleApplied Machine Learning Scientist       0.016235 *
## job_titleApplied Scientist                        0.000402 ***
## job_titleBI Analyst                               0.000371 ***
## job_titleBI Data Analyst                          0.012230 *
## job_titleBig Data Architect                       0.021279 *
## job_titleBig Data Engineer                        0.001495 **
## job_titleBusiness Data Analyst                    0.002513 **
## job_titleCloud Data Architect                     1.34e-05 ***
## job_titleCloud Data Engineer                      0.001896 **
## job_titleComputer Vision Engineer                 0.018831 *
## job_titleComputer Vision Software Engineer        0.000903 ***
## job_titleData Analyst                             0.006436 **
## job_titleData Analytics Consultant                0.029132 *
## job_titleData Analytics Engineer                  0.006317 **
## job_titleData Analytics Manager                   0.006357 **
## job_titleData Architect                           0.000206 ***
## job_titleData Engineer                            0.000514 ***
## job_titleData Engineering Manager                 0.000652 ***
## job_titleData Manager                             0.005206 **
## job_titleData Operations Analyst                  0.082838 .
## job_titleData Operations Engineer                 0.039793 *
## job_titleData Science Consultant                  0.001322 **
## job_titleData Science Engineer                    0.002380 **
## job_titleData Science Lead                        0.001890 **
## job_titleData Science Manager                     2.43e-05 ***
## job_titleData Scientist                           0.000445 ***
## job_titleData Scientist Lead                      0.002928 **
## job_titleData Specialist                          0.002950 **
## job_titleDirector of Data Engineering             0.000460 ***
```

```
## job_titleDirector of Data Science                 3.19e-05 ***
## job_titleETL Developer                            0.001298 **
## job_titleFinance Data Analyst                     0.057176 .
## job_titleFinancial Data Analyst                   0.026022 *
## job_titleHead of Data                             7.18e-05 ***
## job_titleHead of Data Science                     0.004934 **
## job_titleHead of Machine Learning                 0.032166 *
## job_titleLead Data Analyst                        0.005395 **
## job_titleLead Data Engineer                       0.000328 ***
## job_titleLead Data Scientist                      0.000458 ***
## job_titleLead Machine Learning Engineer           0.001683 **
## job_titleMachine Learning Developer               0.002514 **
## job_titleMachine Learning Engineer                0.000142 ***
## job_titleMachine Learning Infrastructure Engineer 0.000242 ***
## job_titleMachine Learning Manager                 0.000832 ***
## job_titleMachine Learning Research Engineer       0.009621 **
## job_titleMachine Learning Scientist               1.93e-05 ***
## job_titleMarketing Data Analyst                   5.09e-05 ***
## job_titleML Engineer                              0.000346 ***
## job_titleNLP Engineer                             0.239766
## job_titlePower BI Developer                       0.018142 *
## job_titlePrincipal Data Analyst                   0.002342 **
## job_titlePrincipal Data Architect                 0.045649 *
## job_titlePrincipal Data Engineer                  0.000169 ***
## job_titlePrincipal Data Scientist                 6.09e-06 ***
## job_titleProduct Data Analyst                     0.008934 **
## job_titleProduct Data Scientist                   0.617469
## job_titleResearch Engineer                        1.19e-05 ***
## job_titleResearch Scientist                       0.000344 ***
## remote_ratioPR                                    0.066526 .
## remote_ratioFR                                    0.350519
## company_sizeM                                     0.752122
## company_sizeS                                     0.000271 ***
## company_locationAL                                0.223461
## company_locationAR                                0.756875
## company_locationAS                                0.590779
## company_locationAT                                0.698786
## company_locationAU                                0.603259
## company_locationBE                                0.972939
## company_locationBR                                0.212501
## company_locationCA                                0.535460
## company_locationCH                                0.798045
## company_locationCL                                0.428495
## company_locationCN                                0.743892
## company_locationCO                                0.362996
## company_locationCZ                                0.968438
## company_locationDE                                0.950282
## company_locationDK                                0.287158
## company_locationEE                                0.189819
## company_locationEG                                0.046023 *
## company_locationES                                0.180395
## company_locationFI                                0.525865
## company_locationFR                                0.649340
## company_locationGB                                0.964258
```

```
## company_locationGR                                0.493303
## company_locationHN                                0.551850
## company_locationHR                                0.250271
## company_locationHU                                0.401718
## company_locationID                                0.488194
## company_locationIE                                0.493883
## company_locationIL                                0.377353
## company_locationIN                                0.099768 .
## company_locationIQ                                0.214536
## company_locationIR                                0.137063
## company_locationIT                                0.500981
## company_locationJP                                0.249118
## company_locationKE                                0.920053
## company_locationLU                                0.950361
## company_locationMD                                0.452540
## company_locationMT                                0.399057
## company_locationMX                                0.096128 .
## company_locationMY                                0.566097
## company_locationNG                                0.584855
## company_locationNL                                0.475493
## company_locationNZ                                0.589106
## company_locationPH                                0.995499
## company_locationPK                                0.418825
## company_locationPL                                0.199043
## company_locationPR                                0.086128 .
## company_locationPT                                0.104400
## company_locationRO                                0.918630
## company_locationRU                                0.783944
## company_locationSG                                0.801282
## company_locationSI                                0.235059
## company_locationTH                                0.391375
## company_locationTR                                0.042726 *
## company_locationUA                                0.317479
## company_locationUS                                0.056284 .
## company_locationVN                                0.222451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40660 on 1171 degrees of freedom
## Multiple R-squared:  0.5887, Adjusted R-squared:  0.5444
## F-statistic:  13.3 on 126 and 1171 DF,  p-value: < 2.2e-16
```

CLUSTERING

Manual Distance matrix

First of all, for us to be able to apply clustering techniques to our Job Titles, we need to create a Distance Matrix which will be used as an input in the clustering methods. To generate this Distance Matrix, we have had to investigate and do some research to deeply understand what these Job Titles consist of so we can actually know how similar they are one to another. Once we have done this first step, we have created a distance scale in which the value 0 represents total similarity between 2 Job Titles (therefore, their distance will be 0) and the value 1 represents that 2 Job Titles are completely opposite. So, with this distance scale

and the research done previously, the Distance Matrix generated is the following:

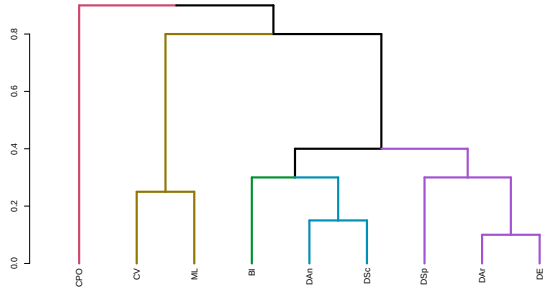
Note that: CV = “Computer Vision”, BI = “Business Intelligence”, DAn = “Data Analyst”, DAr = “Data Architect”, DE = “Data Engineer”, CPO = “Chief Product Officer”, DSc = “Data Scientist”, DSp = “Data Specialist”, ML = “Machine Learning”

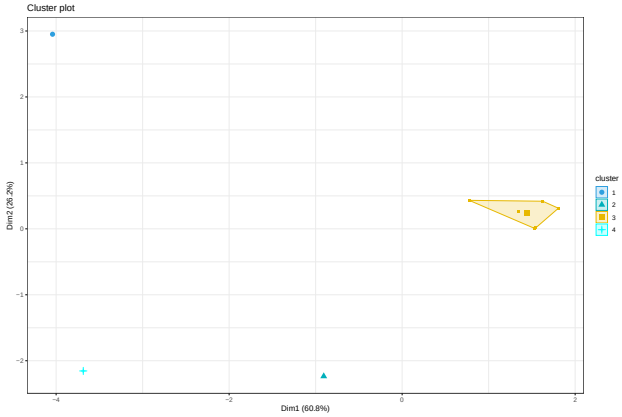
	CV	BI	DAn	DAr	DE	CPO	DSc	DSp	ML
CV	0.00	0.75	0.75	0.75	0.80	0.90	0.75	0.75	0.25
BI	0.75	0.00	0.30	0.40	0.40	0.75	0.30	0.40	0.50
DAn	0.75	0.30	0.00	0.25	0.25	0.75	0.15	0.30	0.50
DAr	0.75	0.40	0.25	0.00	0.10	0.75	0.25	0.30	0.40
DE	0.80	0.40	0.25	0.10	0.00	0.80	0.30	0.30	0.40
CPO	0.90	0.75	0.75	0.75	0.80	0.00	0.80	0.80	0.85
DSc	0.75	0.30	0.15	0.25	0.30	0.80	0.00	0.15	0.50
DSp	0.75	0.40	0.30	0.30	0.30	0.80	0.15	0.00	0.50
ML	0.25	0.50	0.50	0.40	0.40	0.85	0.50	0.50	0.00

Hierarchical Clustering and K-Means methods using the Distance matrix of the Job Titles

With the Hierarchical clustering method, we are able to visualize the clusters formed by the 2 closest Job Titles in each iteration.

Finally, to assure that the clustering results are correct, we use the K-Means method so we have 2 clustering techniques that most likely produce the same result.





Once the K-Means algorithm has finished, this plot shows us the the results are: Cluster 1: “CV”, Cluster 2: “ML”, Cluster3: "“BI”, “DAn”, “DAr”, “DE”, “DSc”, “DSp” and Cluster4: “CPO”.