DATA607_Project 3 - Data Science Skills

Team 3: David Simbandumwe, Thomas Buonora, Charles Ugiagbe, Jaya Veluri

2021-10-19

Introduction

```
usr <- keyring::key_list("DATA607")[1,2]
pwd <-  keyring::key_get("DATA607", usr)
con = dbConnect(MySQL(), user=usr, password=pwd, dbname='DATA607', host='localhost')


rs = dbSendQuery(con, "select *
            from SkillsMeta")
ds_skills_list_df = fetch(rs, n=-1)


dbDisconnect(con)
```

```
## Warning: Closing open result sets
```

```
## [1] TRUE
```

Read data

```
# read skills data from csv
skills_df <- read_csv( file = "/Users/dsimbandumwe/dev/cuny/data_607_T3/DATA607Team3/source/Final_Train
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 19802 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): experience, job_description, job_desig, job_type, key_skills, locat...
## dbl (2): ...1, company_name_encoded

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
names(skills_df)[1] <- "id"


# build a temporary data frame
tmp <- skills_df %>%
    select(id, key_skills) %>%
    separate_rows(
        key_skills,
        convert = TRUE,
        sep = "\\,"
    )


tmp <- tmp %>%
    mutate(
        key_skills = str_to_lower(key_skills),
        key_skills = str_replace_all(key_skills, "\\.{3}", "" ),
        key_skills = str_trim(key_skills)
    )

tmp <- tmp %>%
  right_join(ds_skills_list_df, by="key_skills" ) %>%
  rename(
      key_skills_id = id.y,
      id = id.x
  )

# build a list of user ids that have data science skills
id_df <- tmp %>%
    select(id) %>%
    distinct()


# join the temporary dataframe with the original dataframe
skills_df <- skills_df %>%
    inner_join( id_df,by="id" ) %>%
    right_join(tmp, by="id") %>%
    select(-c(job_description,job_desig,key_skills.x)) %>%
    rename(
      key_skills = key_skills.y
    )


# update the salary
skills_df <- skills_df %>%
    separate(
        salary,
        c("min_salary" , "max_salary"),
        convert = TRUE,
        sep = "to"
    )
```

```
#write out csv file
write.csv(skills_df, "/Users/dsimbandumwe/dev/cuny/data_607_T3/DATA607Team3/output/skillsOutput.csv", r
```

```
emp_df <-  read.csv(
  "/Users/dsimbandumwe/dev/cuny/data_607_T3/DATA607Team3/source/multipleChoiceResponses.csv",
  header=T,sep=","
)

tmp <- emp_df %>% select(Q1, Q2, Q3, Q4, Q6, Q8, Q9, starts_with("Q13"), starts_with("Q16"))
#tmp <- tmp %>% select(-c("Q13_OTHER_TEXT","Q16_OTHER_TEXT"))

tmp <- tmp %>%
  filter(grepl("Data Scientist",Q6)) %>%
  mutate (
     id = row_number()
  )


tmp <- slice(tmp,-(1:1))
tmp <- tmp %>% pivot_longer(
                starts_with("Q13") | starts_with("Q16"),
                names_to = "q",
                values_to = "ans"
              )

tmp <- tmp %>%
    separate_rows(
        ans,
        convert = TRUE,
        sep = "\\/"
    )


tmp <- tmp %>%
  mutate (
    ans = str_squish(ans),
    ans = str_to_lower(ans)
  )


tmp <- tmp %>%
  right_join(ds_skills_list_df, by=c("ans" = "key_skills") ) %>%
  rename(
      key_skills_id = id.y,
      id = id.x
  )


tmp <- tmp %>%
  filter (ans != "" & Q9 != "") %>%
  filter(!grepl("I do not",Q9)) %>%
  mutate (
    Q9 = str_replace(Q9, "\\+",""),
```

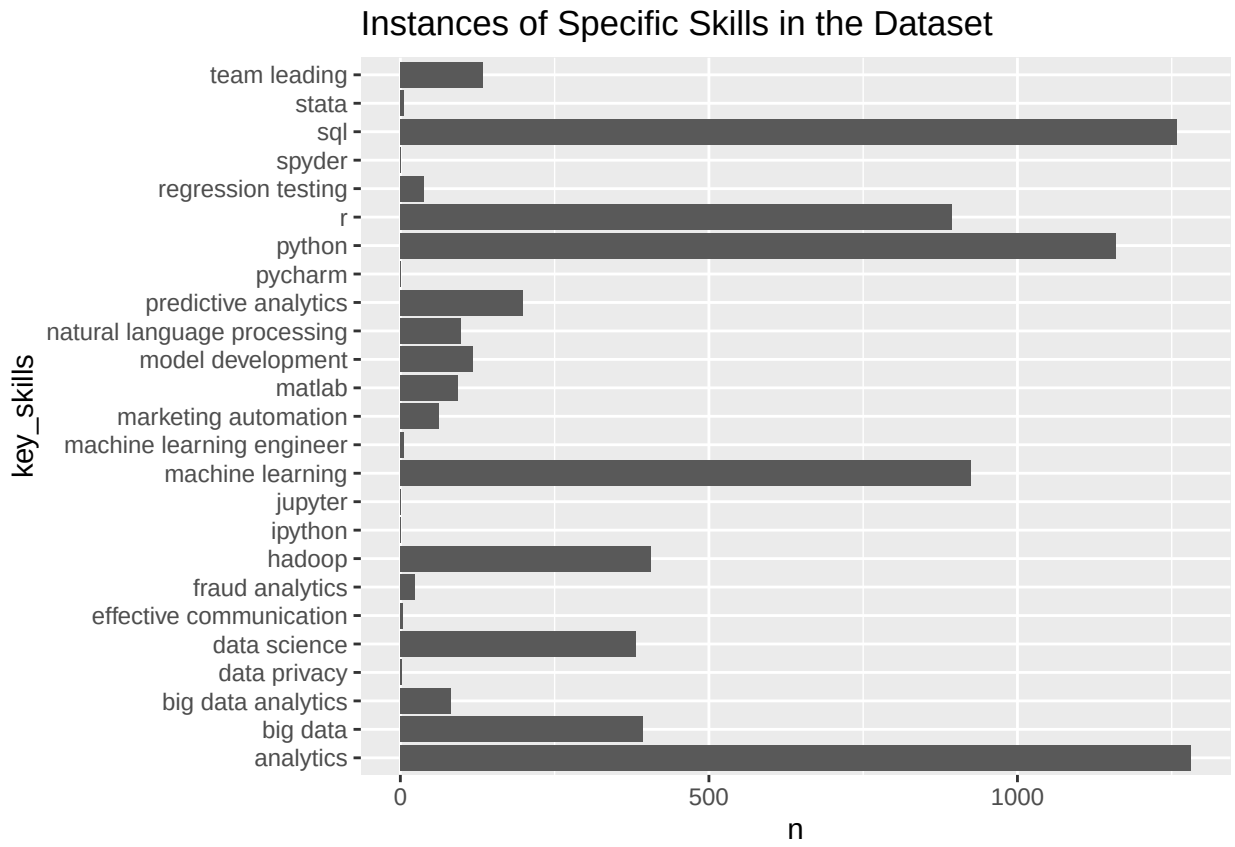
```
    Q9 = str_replace(Q9, ",000","")
  )


tmp <- tmp %>%
  separate(
      Q9,
      c("min_salary" , "max_salary"),
      convert = TRUE,
      sep = "-"
  )
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 37 rows [10452,
## 10453, 10454, 10455, 10456, 10457, 10632, 10633, 10634, 10635, 11562, 11563,
## 11564, 11565, 11566, 11567, 11568, 11569, 11570, 11969, ...].
```

```
tmp <- tmp %>%
  transform(
    min_salary = as.numeric(min_salary),
    max_salary = ifelse(is.na(as.numeric(max_salary)),2000,as.numeric(max_salary))
  )

emp_df <- tmp %>%
    rename(
      gender = Q1,
      age = Q2,
      location = Q3,
      education = Q4,
      title = Q6,
      experience = Q8
    ) %>%
  select (id, q, ans, key_skills_id, min_salary, max_salary, gender, age, location, education, title, e


write.csv(emp_df,"/Users/dsimbandumwe/dev/cuny/data_607_T3/DATA607Team3/output/multipleChoiceOutput.csv
```

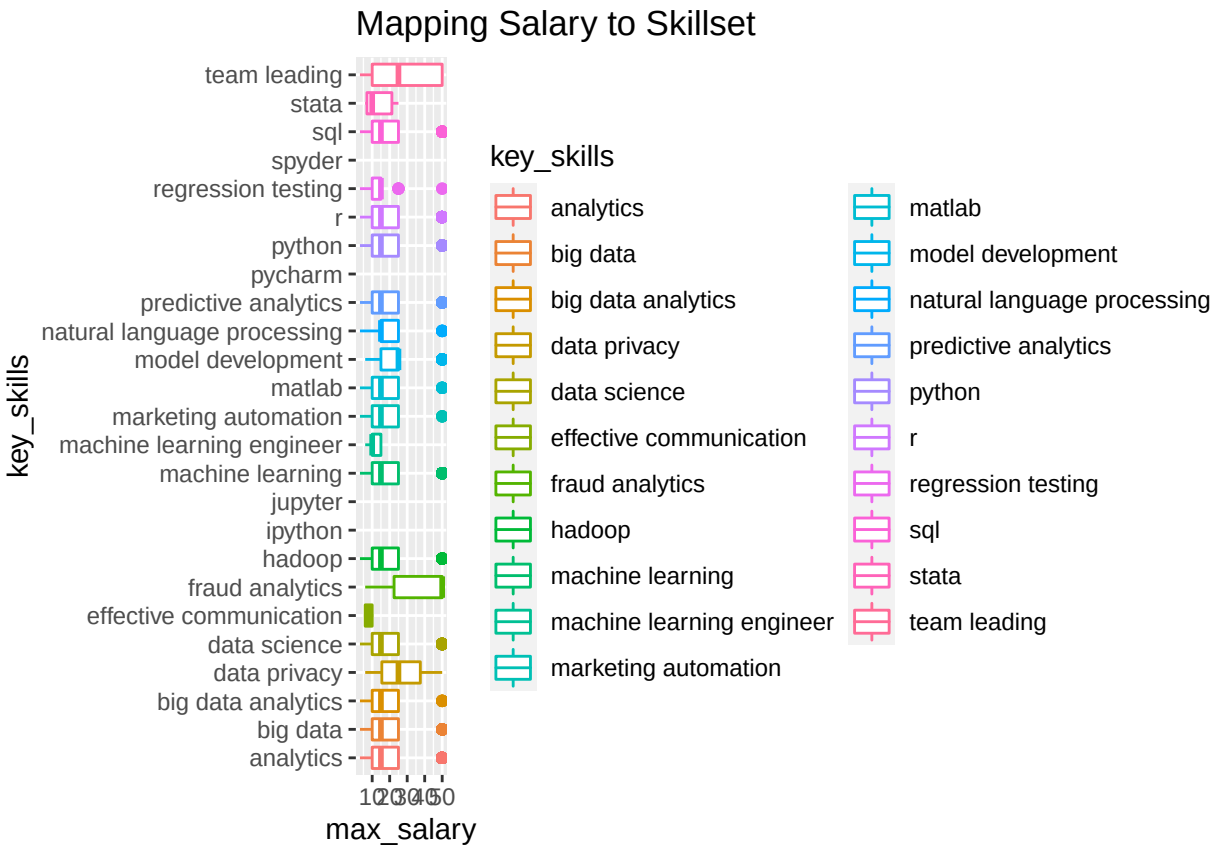
Analysis Open Roles

```
s1 <- skills_df %>% group_by(key_skills) %>%
  mutate(
    n = n()
  ) %>%
  select(key_skills, n) %>%
  distinct()

s1 %>%
  ggplot(aes( y=key_skills, x=n)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title = "Instances of Specific Skills in the Dataset" )
```




```
skills_df %>%
  ggplot() +
  geom_boxplot(mapping = aes(y=key_skills, x=max_salary, color=key_skills)) +
  labs (title = "Mapping Salary to Skillset" )
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```




```
# create a wide dataframe for correlation
t <- skills_df %>%
  mutate(
    flag = 100
  ) %>%
  select(-c(key_skills_id,job_type)) %>%
  mutate (
   key_skills = str_replace_all(key_skills, " ", "_"),
   key_skills = str_squish(key_skills),
   row = row_number()
  )
t <- t %>%
  pivot_wider(
    names_from = key_skills,
    values_from = flag,
    values_fill = 0
  )
t <- t %>% select(-c(id, experience, location, row, company_name_encoded))


t.rcorr = rcorr(as.matrix(t))
t.rcorr
```

```
##                            min_salary max_salary   sql python analytics
## min_salary                       1.00       0.98 -0.11  -0.04     -0.06
## max_salary                       0.98       1.00 -0.11  -0.04     -0.04
```

```
## sql                           -0.11      -0.11  1.00  -0.19     -0.20
## python                        -0.04      -0.04 -0.19   1.00     -0.19
## analytics                     -0.06      -0.04 -0.20  -0.19      1.00
## machine_learning               0.04       0.03 -0.17  -0.16     -0.17
## effective_communication       -0.02      -0.02 -0.01  -0.01     -0.01
## fraud_analytics                0.07       0.08 -0.03  -0.02     -0.03
## team_leading                   0.05       0.07 -0.06  -0.06     -0.06
## data_science                   0.06       0.06 -0.10  -0.10     -0.10
## machine_learning_engineer     -0.02      -0.02 -0.01  -0.01     -0.01
## model_development              0.06       0.06 -0.06  -0.05     -0.06
## marketing_automation          -0.01      -0.02 -0.04  -0.04     -0.04
## r                              0.08       0.07 -0.16  -0.16     -0.17
## big_data                       0.03       0.03 -0.10  -0.10     -0.11
## predictive_analytics           0.02       0.02 -0.07  -0.07     -0.07
## hadoop                         0.00      -0.01 -0.11  -0.10     -0.11
## data_privacy                   0.01       0.01 -0.01  -0.01     -0.01
## regression_testing            -0.02      -0.02 -0.03  -0.03     -0.03
## natural_language_processing    0.00      -0.01 -0.05  -0.05     -0.05
## matlab                         0.00      -0.01 -0.05  -0.05     -0.05
## big_data_analytics             0.00       0.00 -0.05  -0.04     -0.05
## stata                         -0.01      -0.01 -0.01  -0.01     -0.01
## jupyter                         NaN        NaN -0.01   0.00     -0.01
## ipython                         NaN        NaN -0.01   0.00     -0.01
## pycharm                         NaN        NaN -0.01   0.00     -0.01
## spyder                          NaN        NaN -0.01   0.00     -0.01
##                           machine_learning effective_communication
## min_salary                            0.04                    -0.02
## max_salary                            0.03                    -0.02
## sql                                  -0.17                    -0.01
## python                               -0.16                    -0.01
## analytics                            -0.17                    -0.01
## machine_learning                      1.00                    -0.01
## effective_communication              -0.01                     1.00
## fraud_analytics                      -0.02                     0.00
## team_leading                         -0.05                     0.00
## data_science                         -0.09                    -0.01
## machine_learning_engineer            -0.01                     0.00
## model_development                    -0.05                     0.00
## marketing_automation                 -0.03                     0.00
## r                                    -0.14                    -0.01
## big_data                             -0.09                    -0.01
## predictive_analytics                 -0.06                     0.00
## hadoop                               -0.09                    -0.01
## data_privacy                         -0.01                     0.00
## regression_testing                   -0.03                     0.00
## natural_language_processing          -0.04                     0.00
## matlab                               -0.04                     0.00
## big_data_analytics                   -0.04                     0.00
## stata                                -0.01                     0.00
## jupyter                               0.00                     0.00
## ipython                               0.00                     0.00
## pycharm                               0.00                     0.00
## spyder                                0.00                     0.00
##                           fraud_analytics team_leading data_science
```

```
## min_salary                          0.07         0.05         0.06
## max_salary                          0.08         0.07         0.06
## sql                                -0.03        -0.06        -0.10
## python                             -0.02        -0.06        -0.10
## analytics                          -0.03        -0.06        -0.10
## machine_learning                   -0.02        -0.05        -0.09
## effective_communication             0.00         0.00        -0.01
## fraud_analytics                     1.00        -0.01        -0.01
## team_leading                       -0.01         1.00        -0.03
## data_science                       -0.01        -0.03         1.00
## machine_learning_engineer           0.00         0.00        -0.01
## model_development                  -0.01        -0.02        -0.03
## marketing_automation               -0.01        -0.01        -0.02
## r                                  -0.02        -0.05        -0.08
## big_data                           -0.01        -0.03        -0.05
## predictive_analytics               -0.01        -0.02        -0.04
## hadoop                             -0.01        -0.03        -0.05
## data_privacy                        0.00         0.00         0.00
## regression_testing                  0.00        -0.01        -0.02
## natural_language_processing        -0.01        -0.02        -0.03
## matlab                             -0.01        -0.01        -0.03
## big_data_analytics                 -0.01        -0.01        -0.02
## stata                               0.00         0.00        -0.01
## jupyter                             0.00         0.00         0.00
## ipython                             0.00         0.00         0.00
## pycharm                             0.00         0.00         0.00
## spyder                              0.00         0.00         0.00
##                          machine_learning_engineer model_development
## min_salary                                  -0.02              0.06
## max_salary                                  -0.02              0.06
## sql                                         -0.01             -0.06
## python                                      -0.01             -0.05
## analytics                                   -0.01             -0.06
## machine_learning                            -0.01             -0.05
## effective_communication                      0.00              0.00
## fraud_analytics                              0.00             -0.01
## team_leading                                 0.00             -0.02
## data_science                                -0.01             -0.03
## machine_learning_engineer                    1.00              0.00
## model_development                            0.00              1.00
## marketing_automation                         0.00             -0.01
## r                                           -0.01             -0.05
## big_data                                    -0.01             -0.03
## predictive_analytics                         0.00             -0.02
## hadoop                                      -0.01             -0.03
## data_privacy                                 0.00              0.00
## regression_testing                           0.00             -0.01
## natural_language_processing                  0.00             -0.01
## matlab                                       0.00             -0.01
## big_data_analytics                           0.00             -0.01
## stata                                        0.00              0.00
## jupyter                                      0.00              0.00
## ipython                                      0.00              0.00
## pycharm                                      0.00              0.00
```

```
## spyder                                        0.00             0.00
##                            marketing_automation     r big_data
## min_salary                                -0.01  0.08     0.03
## max_salary                                -0.02  0.07     0.03
## sql                                       -0.04 -0.16    -0.10
## python                                    -0.04 -0.16    -0.10
## analytics                                 -0.04 -0.17    -0.11
## machine_learning                          -0.03 -0.14    -0.09
## effective_communication                    0.00 -0.01    -0.01
## fraud_analytics                           -0.01 -0.02    -0.01
## team_leading                              -0.01 -0.05    -0.03
## data_science                              -0.02 -0.08    -0.05
## machine_learning_engineer                  0.00 -0.01    -0.01
## model_development                         -0.01 -0.05    -0.03
## marketing_automation                       1.00 -0.03    -0.02
## r                                         -0.03  1.00    -0.09
## big_data                                  -0.02 -0.09     1.00
## predictive_analytics                      -0.02 -0.06    -0.04
## hadoop                                    -0.02 -0.09    -0.06
## data_privacy                               0.00 -0.01     0.00
## regression_testing                        -0.01 -0.03    -0.02
## natural_language_processing               -0.01 -0.04    -0.03
## matlab                                    -0.01 -0.04    -0.03
## big_data_analytics                        -0.01 -0.04    -0.02
## stata                                      0.00 -0.01    -0.01
## jupyter                                    0.00  0.00     0.00
## ipython                                    0.00  0.00     0.00
## pycharm                                    0.00  0.00     0.00
## spyder                                     0.00  0.00     0.00
##                            predictive_analytics hadoop data_privacy
## min_salary                                 0.02   0.00         0.01
## max_salary                                 0.02  -0.01         0.01
## sql                                       -0.07  -0.11        -0.01
## python                                    -0.07  -0.10        -0.01
## analytics                                 -0.07  -0.11        -0.01
## machine_learning                          -0.06  -0.09        -0.01
## effective_communication                    0.00  -0.01         0.00
## fraud_analytics                           -0.01  -0.01         0.00
## team_leading                              -0.02  -0.03         0.00
## data_science                              -0.04  -0.05         0.00
## machine_learning_engineer                  0.00  -0.01         0.00
## model_development                         -0.02  -0.03         0.00
## marketing_automation                      -0.02  -0.02         0.00
## r                                         -0.06  -0.09        -0.01
## big_data                                  -0.04  -0.06         0.00
## predictive_analytics                       1.00  -0.04         0.00
## hadoop                                    -0.04   1.00         0.00
## data_privacy                               0.00   0.00         1.00
## regression_testing                        -0.01  -0.02         0.00
## natural_language_processing               -0.02  -0.03         0.00
## matlab                                    -0.02  -0.03         0.00
## big_data_analytics                        -0.02  -0.02         0.00
## stata                                      0.00  -0.01         0.00
## jupyter                                    0.00   0.00         0.00
```

```
## ipython                                  0.00   0.00         0.00
## pycharm                                  0.00   0.00         0.00
## spyder                                   0.00   0.00         0.00
##                           regression_testing natural_language_processing
## min_salary                             -0.02                         0.00
## max_salary                             -0.02                        -0.01
## sql                                    -0.03                        -0.05
## python                                 -0.03                        -0.05
## analytics                              -0.03                        -0.05
## machine_learning                       -0.03                        -0.04
## effective_communication                 0.00                         0.00
## fraud_analytics                         0.00                        -0.01
## team_leading                           -0.01                        -0.02
## data_science                           -0.02                        -0.03
## machine_learning_engineer               0.00                         0.00
## model_development                      -0.01                        -0.01
## marketing_automation                   -0.01                        -0.01
## r                                      -0.03                        -0.04
## big_data                               -0.02                        -0.03
## predictive_analytics                   -0.01                        -0.02
## hadoop                                 -0.02                        -0.03
## data_privacy                            0.00                         0.00
## regression_testing                      1.00                        -0.01
## natural_language_processing            -0.01                         1.00
## matlab                                 -0.01                        -0.01
## big_data_analytics                     -0.01                        -0.01
## stata                                   0.00                         0.00
## jupyter                                 0.00                         0.00
## ipython                                 0.00                         0.00
## pycharm                                 0.00                         0.00
## spyder                                  0.00                         0.00
##                           matlab big_data_analytics stata jupyter ipython
## min_salary                  0.00               0.00 -0.01     NaN     NaN
## max_salary                 -0.01               0.00 -0.01     NaN     NaN
## sql                        -0.05              -0.05 -0.01   -0.01   -0.01
## python                     -0.05              -0.04 -0.01    0.00    0.00
## analytics                  -0.05              -0.05 -0.01   -0.01   -0.01
## machine_learning           -0.04              -0.04 -0.01    0.00    0.00
## effective_communication     0.00               0.00  0.00    0.00    0.00
## fraud_analytics            -0.01              -0.01  0.00    0.00    0.00
## team_leading               -0.01              -0.01  0.00    0.00    0.00
## data_science               -0.03              -0.02 -0.01    0.00    0.00
## machine_learning_engineer   0.00               0.00  0.00    0.00    0.00
## model_development          -0.01              -0.01  0.00    0.00    0.00
## marketing_automation       -0.01              -0.01  0.00    0.00    0.00
## r                          -0.04              -0.04 -0.01    0.00    0.00
## big_data                   -0.03              -0.02 -0.01    0.00    0.00
## predictive_analytics       -0.02              -0.02  0.00    0.00    0.00
## hadoop                     -0.03              -0.02 -0.01    0.00    0.00
## data_privacy                0.00               0.00  0.00    0.00    0.00
## regression_testing         -0.01              -0.01  0.00    0.00    0.00
## natural_language_processing -0.01             -0.01  0.00    0.00    0.00
## matlab                      1.00              -0.01  0.00    0.00    0.00
## big_data_analytics         -0.01               1.00  0.00    0.00    0.00
```

```
## stata                          0.00              0.00  1.00    0.00    0.00
## jupyter                        0.00              0.00  0.00    1.00    0.00
## ipython                        0.00              0.00  0.00    0.00    1.00
## pycharm                        0.00              0.00  0.00    0.00    0.00
## spyder                         0.00              0.00  0.00    0.00    0.00
##                             pycharm spyder
## min_salary                      NaN    NaN
## max_salary                      NaN    NaN
## sql                           -0.01  -0.01
## python                         0.00   0.00
## analytics                     -0.01  -0.01
## machine_learning               0.00   0.00
## effective_communication        0.00   0.00
## fraud_analytics                0.00   0.00
## team_leading                   0.00   0.00
## data_science                   0.00   0.00
## machine_learning_engineer      0.00   0.00
## model_development              0.00   0.00
## marketing_automation           0.00   0.00
## r                              0.00   0.00
## big_data                       0.00   0.00
## predictive_analytics           0.00   0.00
## hadoop                         0.00   0.00
## data_privacy                   0.00   0.00
## regression_testing             0.00   0.00
## natural_language_processing    0.00   0.00
## matlab                         0.00   0.00
## big_data_analytics             0.00   0.00
## stata                          0.00   0.00
## jupyter                        0.00   0.00
## ipython                        0.00   0.00
## pycharm                        1.00   0.00
## spyder                         0.00   1.00
##
## n
##                             min_salary max_salary  sql python analytics
## min_salary                        7562       7562 7562   7562      7562
## max_salary                        7562       7562 7562   7562      7562
## sql                               7562       7562 7566   7566      7566
## python                            7562       7562 7566   7566      7566
## analytics                         7562       7562 7566   7566      7566
## machine_learning                  7562       7562 7566   7566      7566
## effective_communication           7562       7562 7566   7566      7566
## fraud_analytics                   7562       7562 7566   7566      7566
## team_leading                      7562       7562 7566   7566      7566
## data_science                      7562       7562 7566   7566      7566
## machine_learning_engineer         7562       7562 7566   7566      7566
## model_development                 7562       7562 7566   7566      7566
## marketing_automation              7562       7562 7566   7566      7566
## r                                 7562       7562 7566   7566      7566
## big_data                          7562       7562 7566   7566      7566
## predictive_analytics              7562       7562 7566   7566      7566
## hadoop                            7562       7562 7566   7566      7566
## data_privacy                      7562       7562 7566   7566      7566
```

```
## regression_testing                7562       7562 7566   7566      7566
## natural_language_processing       7562       7562 7566   7566      7566
## matlab                            7562       7562 7566   7566      7566
## big_data_analytics                7562       7562 7566   7566      7566
## stata                             7562       7562 7566   7566      7566
## jupyter                           7562       7562 7566   7566      7566
## ipython                           7562       7562 7566   7566      7566
## pycharm                           7562       7562 7566   7566      7566
## spyder                            7562       7562 7566   7566      7566
##                            machine_learning effective_communication
## min_salary                             7562                     7562
## max_salary                             7562                     7562
## sql                                    7566                     7566
## python                                 7566                     7566
## analytics                              7566                     7566
## machine_learning                       7566                     7566
## effective_communication                7566                     7566
## fraud_analytics                        7566                     7566
## team_leading                           7566                     7566
## data_science                           7566                     7566
## machine_learning_engineer              7566                     7566
## model_development                      7566                     7566
## marketing_automation                   7566                     7566
## r                                      7566                     7566
## big_data                               7566                     7566
## predictive_analytics                   7566                     7566
## hadoop                                 7566                     7566
## data_privacy                           7566                     7566
## regression_testing                     7566                     7566
## natural_language_processing            7566                     7566
## matlab                                 7566                     7566
## big_data_analytics                     7566                     7566
## stata                                  7566                     7566
## jupyter                                7566                     7566
## ipython                                7566                     7566
## pycharm                                7566                     7566
## spyder                                 7566                     7566
##                            fraud_analytics team_leading data_science
## min_salary                            7562         7562         7562
## max_salary                            7562         7562         7562
## sql                                   7566         7566         7566
## python                                7566         7566         7566
## analytics                             7566         7566         7566
## machine_learning                      7566         7566         7566
## effective_communication               7566         7566         7566
## fraud_analytics                       7566         7566         7566
## team_leading                          7566         7566         7566
## data_science                          7566         7566         7566
## machine_learning_engineer             7566         7566         7566
## model_development                     7566         7566         7566
## marketing_automation                  7566         7566         7566
## r                                     7566         7566         7566
## big_data                              7566         7566         7566
## predictive_analytics                  7566         7566         7566
```

```
## hadoop                                7566         7566         7566
## data_privacy                          7566         7566         7566
## regression_testing                    7566         7566         7566
## natural_language_processing           7566         7566         7566
## matlab                                7566         7566         7566
## big_data_analytics                    7566         7566         7566
## stata                                 7566         7566         7566
## jupyter                               7566         7566         7566
## ipython                               7566         7566         7566
## pycharm                               7566         7566         7566
## spyder                                7566         7566         7566
##                           machine_learning_engineer model_development
## min_salary                                     7562              7562
## max_salary                                     7562              7562
## sql                                            7566              7566
## python                                         7566              7566
## analytics                                      7566              7566
## machine_learning                               7566              7566
## effective_communication                        7566              7566
## fraud_analytics                                7566              7566
## team_leading                                   7566              7566
## data_science                                   7566              7566
## machine_learning_engineer                      7566              7566
## model_development                              7566              7566
## marketing_automation                           7566              7566
## r                                              7566              7566
## big_data                                       7566              7566
## predictive_analytics                           7566              7566
## hadoop                                         7566              7566
## data_privacy                                   7566              7566
## regression_testing                             7566              7566
## natural_language_processing                    7566              7566
## matlab                                         7566              7566
## big_data_analytics                             7566              7566
## stata                                          7566              7566
## jupyter                                        7566              7566
## ipython                                        7566              7566
## pycharm                                        7566              7566
## spyder                                         7566              7566
##                           marketing_automation    r big_data
## min_salary                                7562 7562     7562
## max_salary                                7562 7562     7562
## sql                                       7566 7566     7566
## python                                    7566 7566     7566
## analytics                                 7566 7566     7566
## machine_learning                          7566 7566     7566
## effective_communication                   7566 7566     7566
## fraud_analytics                           7566 7566     7566
## team_leading                              7566 7566     7566
## data_science                              7566 7566     7566
## machine_learning_engineer                 7566 7566     7566
## model_development                         7566 7566     7566
## marketing_automation                      7566 7566     7566
## r                                         7566 7566     7566
```

```
## big_data                                7566 7566     7566
## predictive_analytics                    7566 7566     7566
## hadoop                                  7566 7566     7566
## data_privacy                            7566 7566     7566
## regression_testing                      7566 7566     7566
## natural_language_processing             7566 7566     7566
## matlab                                  7566 7566     7566
## big_data_analytics                      7566 7566     7566
## stata                                   7566 7566     7566
## jupyter                                 7566 7566     7566
## ipython                                 7566 7566     7566
## pycharm                                 7566 7566     7566
## spyder                                  7566 7566     7566
##                          predictive_analytics hadoop data_privacy
## min_salary                               7562   7562         7562
## max_salary                               7562   7562         7562
## sql                                      7566   7566         7566
## python                                   7566   7566         7566
## analytics                                7566   7566         7566
## machine_learning                         7566   7566         7566
## effective_communication                  7566   7566         7566
## fraud_analytics                          7566   7566         7566
## team_leading                             7566   7566         7566
## data_science                             7566   7566         7566
## machine_learning_engineer                7566   7566         7566
## model_development                        7566   7566         7566
## marketing_automation                     7566   7566         7566
## r                                        7566   7566         7566
## big_data                                 7566   7566         7566
## predictive_analytics                     7566   7566         7566
## hadoop                                   7566   7566         7566
## data_privacy                             7566   7566         7566
## regression_testing                       7566   7566         7566
## natural_language_processing              7566   7566         7566
## matlab                                   7566   7566         7566
## big_data_analytics                       7566   7566         7566
## stata                                    7566   7566         7566
## jupyter                                  7566   7566         7566
## ipython                                  7566   7566         7566
## pycharm                                  7566   7566         7566
## spyder                                   7566   7566         7566
##                          regression_testing natural_language_processing
## min_salary                             7562                        7562
## max_salary                             7562                        7562
## sql                                    7566                        7566
## python                                 7566                        7566
## analytics                              7566                        7566
## machine_learning                       7566                        7566
## effective_communication                7566                        7566
## fraud_analytics                        7566                        7566
## team_leading                           7566                        7566
## data_science                           7566                        7566
## machine_learning_engineer              7566                        7566
## model_development                      7566                        7566
```

```
## marketing_automation                    7566                    7566
## r                                       7566                    7566
## big_data                                7566                    7566
## predictive_analytics                    7566                    7566
## hadoop                                  7566                    7566
## data_privacy                            7566                    7566
## regression_testing                      7566                    7566
## natural_language_processing             7566                    7566
## matlab                                  7566                    7566
## big_data_analytics                      7566                    7566
## stata                                   7566                    7566
## jupyter                                 7566                    7566
## ipython                                 7566                    7566
## pycharm                                 7566                    7566
## spyder                                  7566                    7566
##                            matlab big_data_analytics stata jupyter ipython
## min_salary                   7562               7562  7562    7562    7562
## max_salary                   7562               7562  7562    7562    7562
## sql                          7566               7566  7566    7566    7566
## python                       7566               7566  7566    7566    7566
## analytics                    7566               7566  7566    7566    7566
## machine_learning             7566               7566  7566    7566    7566
## effective_communication      7566               7566  7566    7566    7566
## fraud_analytics              7566               7566  7566    7566    7566
## team_leading                 7566               7566  7566    7566    7566
## data_science                 7566               7566  7566    7566    7566
## machine_learning_engineer    7566               7566  7566    7566    7566
## model_development            7566               7566  7566    7566    7566
## marketing_automation         7566               7566  7566    7566    7566
## r                            7566               7566  7566    7566    7566
## big_data                     7566               7566  7566    7566    7566
## predictive_analytics         7566               7566  7566    7566    7566
## hadoop                       7566               7566  7566    7566    7566
## data_privacy                 7566               7566  7566    7566    7566
## regression_testing           7566               7566  7566    7566    7566
## natural_language_processing  7566               7566  7566    7566    7566
## matlab                       7566               7566  7566    7566    7566
## big_data_analytics           7566               7566  7566    7566    7566
## stata                        7566               7566  7566    7566    7566
## jupyter                      7566               7566  7566    7566    7566
## ipython                      7566               7566  7566    7566    7566
## pycharm                      7566               7566  7566    7566    7566
## spyder                       7566               7566  7566    7566    7566
##                            pycharm spyder
## min_salary                    7562   7562
## max_salary                    7562   7562
## sql                           7566   7566
## python                        7566   7566
## analytics                     7566   7566
## machine_learning              7566   7566
## effective_communication       7566   7566
## fraud_analytics               7566   7566
## team_leading                  7566   7566
## data_science                  7566   7566
```

```
## machine_learning_engineer      7566   7566
## model_development              7566   7566
## marketing_automation           7566   7566
## r                              7566   7566
## big_data                       7566   7566
## predictive_analytics           7566   7566
## hadoop                         7566   7566
## data_privacy                   7566   7566
## regression_testing             7566   7566
## natural_language_processing    7566   7566
## matlab                         7566   7566
## big_data_analytics             7566   7566
## stata                          7566   7566
## jupyter                        7566   7566
## ipython                        7566   7566
## pycharm                        7566   7566
## spyder                         7566   7566
##
## P
##                              min_salary max_salary sql    python analytics
## min_salary                                 0.0000     0.0000 0.0018 0.0000
## max_salary                   0.0000                   0.0000 0.0016 0.0005
## sql                          0.0000     0.0000                0.0000 0.0000
## python                       0.0018     0.0016     0.0000        0.0000
## analytics                    0.0000     0.0005     0.0000 0.0000
## machine_learning             0.0021     0.0051     0.0000 0.0000 0.0000
## effective_communication      0.0456     0.0839     0.3717 0.3947 0.3665
## fraud_analytics              0.0000     0.0000     0.0284 0.0368 0.0267
## team_leading                 0.0000     0.0000     0.0000 0.0000 0.0000
## data_science                 0.0000     0.0000     0.0000 0.0000 0.0000
## machine_learning_engineer    0.1602     0.1624     0.3179 0.3412 0.3126
## model_development            0.0000     0.0000     0.0000 0.0000 0.0000
## marketing_automation         0.2054     0.1465     0.0004 0.0007 0.0003
## r                            0.0000     0.0000     0.0000 0.0000 0.0000
## big_data                     0.0064     0.0213     0.0000 0.0000 0.0000
## predictive_analytics         0.0507     0.0466     0.0000 0.0000 0.0000
## hadoop                       0.8386     0.3021     0.0000 0.0000 0.0000
## data_privacy                 0.4245     0.3337     0.4392 0.4611 0.4342
## regression_testing           0.1073     0.0761     0.0058 0.0085 0.0053
## natural_language_processing  0.9468     0.5256     0.0000 0.0000 0.0000
## matlab                       0.7652     0.6080     0.0000 0.0000 0.0000
## big_data_analytics           0.7198     0.7290     0.0000 0.0001 0.0000
## stata                        0.2401     0.2810     0.2739 0.2971 0.2687
## jupyter                                            0.6552 0.6705 0.6517
## ipython                                            0.6552 0.6705 0.6517
## pycharm                                            0.6552 0.6705 0.6517
## spyder                                             0.6552 0.6705 0.6517
##                              machine_learning effective_communication
## min_salary                   0.0021           0.0456
## max_salary                   0.0051           0.0839
## sql                          0.0000           0.3717
## python                       0.0000           0.3947
## analytics                    0.0000           0.3665
## machine_learning                              0.4556
```

```
## effective_communication     0.4556
## fraud_analytics             0.0672           0.9102
## team_leading                0.0000           0.7890
## data_science                0.0000           0.6451
## machine_learning_engineer   0.4042           0.9590
## model_development           0.0000           0.8012
## marketing_automation        0.0029           0.8546
## r                           0.0000           0.4641
## big_data                    0.0000           0.6396
## predictive_analytics        0.0000           0.7430
## hadoop                      0.0000           0.6339
## data_privacy                0.5182           0.9682
## regression_testing          0.0212           0.8870
## natural_language_processing 0.0002           0.8188
## matlab                      0.0003           0.8234
## big_data_analytics          0.0007           0.8342
## stata                       0.3608           0.9551
## jupyter                     0.7092           0.9817
## ipython                     0.7092           0.9817
## pycharm                     0.7092           0.9817
## spyder                      0.7092           0.9817
##                            fraud_analytics team_leading data_science
## min_salary                 0.0000          0.0000       0.0000
## max_salary                 0.0000          0.0000       0.0000
## sql                        0.0284          0.0000       0.0000
## python                     0.0368          0.0000       0.0000
## analytics                  0.0267          0.0000       0.0000
## machine_learning           0.0672          0.0000       0.0000
## effective_communication    0.9102          0.7890       0.6451
## fraud_analytics                            0.5117       0.2586
## team_leading               0.5117                       0.0074
## data_science               0.2586          0.0074
## machine_learning_engineer  0.8996          0.7648       0.6066
## model_development          0.5369          0.1431       0.0117
## marketing_automation       0.6530          0.2864       0.0665
## r                          0.0725          0.0000       0.0000
## big_data                   0.2508          0.0065       0.0000
## predictive_analytics       0.4213          0.0565       0.0010
## hadoop                     0.2427          0.0056       0.0000
## data_privacy               0.9222          0.8168       0.6900
## regression_testing         0.7274          0.4085       0.1547
## natural_language_processing 0.5741         0.1826       0.0218
## matlab                     0.5842          0.1943       0.0255
## big_data_analytics         0.6076          0.2233       0.0360
## stata                      0.8901          0.7431       0.5726
## jupyter                    0.9550          0.8936       0.8179
## ipython                    0.9550          0.8936       0.8179
## pycharm                    0.9550          0.8936       0.8179
## spyder                     0.9550          0.8936       0.8179
##                            machine_learning_engineer model_development
## min_salary                 0.1602                    0.0000
## max_salary                 0.1624                    0.0000
## sql                        0.3179                    0.0000
## python                     0.3412                    0.0000
```

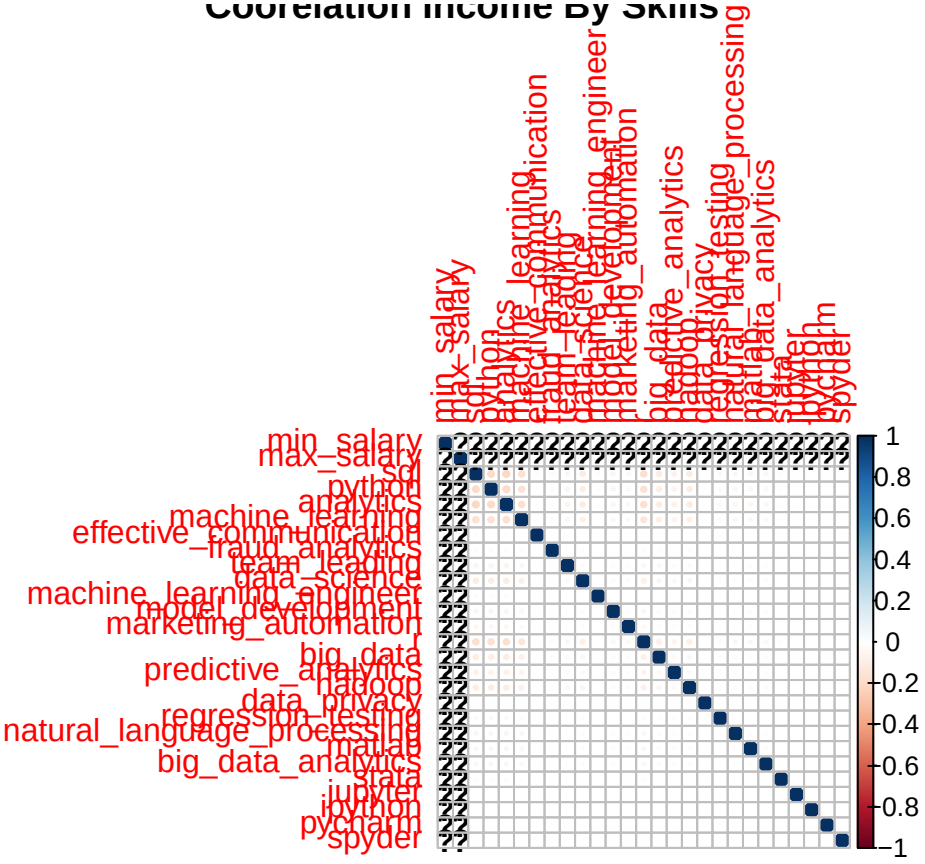
```
## analytics                   0.3126                    0.0000
## machine_learning            0.4042                    0.0000
## effective_communication     0.9590                    0.8012
## fraud_analytics             0.8996                    0.5369
## team_leading                0.7648                    0.1431
## data_science                0.6066                    0.0117
## machine_learning_engineer                             0.7783
## model_development           0.7783
## marketing_automation        0.8376                    0.3158
## r                           0.4130                    0.0000
## big_data                    0.6006                    0.0104
## predictive_analytics        0.7139                    0.0727
## hadoop                      0.5943                    0.0091
## data_privacy                0.9645                    0.8274
## regression_testing          0.8738                    0.4367
## natural_language_processing 0.7978                    0.2098
## matlab                      0.8030                    0.2220
## big_data_analytics          0.8149                    0.2518
## stata                       0.9498                    0.7578
## jupyter                     0.9795                    0.8998
## ipython                     0.9795                    0.8998
## pycharm                     0.9795                    0.8998
## spyder                      0.9795                    0.8998
##                             marketing_automation r      big_data
## min_salary                  0.2054               0.0000 0.0064
## max_salary                  0.1465               0.0000 0.0213
## sql                         0.0004               0.0000 0.0000
## python                      0.0007               0.0000 0.0000
## analytics                   0.0003               0.0000 0.0000
## machine_learning            0.0029               0.0000 0.0000
## effective_communication     0.8546               0.4641 0.6396
## fraud_analytics             0.6530               0.0725 0.2508
## team_leading                0.2864               0.0000 0.0065
## data_science                0.0665               0.0000 0.0000
## machine_learning_engineer   0.8376               0.4130 0.6006
## model_development           0.3158               0.0000 0.0104
## marketing_automation                             0.0035 0.0621
## r                           0.0035                      0.0000
## big_data                    0.0621               0.0000
## predictive_analytics        0.1914               0.0000 0.0008
## hadoop                      0.0577               0.0000 0.0000
## data_privacy                0.8739               0.5261 0.6852
## regression_testing          0.5713               0.0237 0.1481
## natural_language_processing 0.3613               0.0003 0.0197
## matlab                      0.3740               0.0004 0.0231
## big_data_analytics          0.4042               0.0009 0.0331
## stata                       0.8224               0.3698 0.5663
## jupyter                     0.9270               0.7144 0.8149
## ipython                     0.9270               0.7144 0.8149
## pycharm                     0.9270               0.7144 0.8149
## spyder                      0.9270               0.7144 0.8149
##                             predictive_analytics hadoop data_privacy
## min_salary                  0.0507               0.8386 0.4245
## max_salary                  0.0466               0.3021 0.3337
```

```
## sql                          0.0000               0.0000 0.4392
## python                       0.0000               0.0000 0.4611
## analytics                    0.0000               0.0000 0.4342
## machine_learning             0.0000               0.0000 0.5182
## effective_communication      0.7430               0.6339 0.9682
## fraud_analytics              0.4213               0.2427 0.9222
## team_leading                 0.0565               0.0056 0.8168
## data_science                 0.0010               0.0000 0.6900
## machine_learning_engineer    0.7139               0.5943 0.9645
## model_development            0.0727               0.0091 0.8274
## marketing_automation         0.1914               0.0577 0.8739
## r                            0.0000               0.0000 0.5261
## big_data                     0.0008               0.0000 0.6852
## predictive_analytics                              0.0007 0.7765
## hadoop                       0.0007                      0.6800
## data_privacy                 0.7765               0.6800
## regression_testing           0.3111               0.1412 0.9021
## natural_language_processing 0.1024               0.0177 0.8427
## matlab                       0.1117               0.0209 0.8468
## big_data_analytics           0.1356               0.0302 0.8561
## stata                        0.6879               0.5596 0.9611
## jupyter                      0.8698               0.8118 0.9841
## ipython                      0.8698               0.8118 0.9841
## pycharm                      0.8698               0.8118 0.9841
## spyder                       0.8698               0.8118 0.9841
##                             regression_testing natural_language_processing
## min_salary                  0.1073             0.9468
## max_salary                  0.0761             0.5256
## sql                         0.0058             0.0000
## python                      0.0085             0.0000
## analytics                   0.0053             0.0000
## machine_learning            0.0212             0.0002
## effective_communication     0.8870             0.8188
## fraud_analytics             0.7274             0.5741
## team_leading                0.4085             0.1826
## data_science                0.1547             0.0218
## machine_learning_engineer   0.8738             0.7978
## model_development           0.4367             0.2098
## marketing_automation        0.5713             0.3613
## r                           0.0237             0.0003
## big_data                    0.1481             0.0197
## predictive_analytics        0.3111             0.1024
## hadoop                      0.1412             0.0177
## data_privacy                0.9021             0.8427
## regression_testing                             0.4790
## natural_language_processing 0.4790
## matlab                      0.4906             0.2664
## big_data_analytics          0.5178             0.2970
## stata                       0.8618             0.7790
## jupyter                     0.9434             0.9088
## ipython                     0.9434             0.9088
## pycharm                     0.9434             0.9088
## spyder                      0.9434             0.9088
##                             matlab big_data_analytics stata  jupyter ipython
```

```
## min_salary                  0.7652 0.7198            0.2401
## max_salary                  0.6080 0.7290            0.2810
## sql                         0.0000 0.0000            0.2739 0.6552  0.6552
## python                      0.0000 0.0001            0.2971 0.6705  0.6705
## analytics                   0.0000 0.0000            0.2687 0.6517  0.6517
## machine_learning            0.0003 0.0007            0.3608 0.7092  0.7092
## effective_communication     0.8234 0.8342            0.9551 0.9817  0.9817
## fraud_analytics             0.5842 0.6076            0.8901 0.9550  0.9550
## team_leading                0.1943 0.2233            0.7431 0.8936  0.8936
## data_science                0.0255 0.0360            0.5726 0.8179  0.8179
## machine_learning_engineer   0.8030 0.8149            0.9498 0.9795  0.9795
## model_development           0.2220 0.2518            0.7578 0.8998  0.8998
## marketing_automation        0.3740 0.4042            0.8224 0.9270  0.9270
## r                           0.0004 0.0009            0.3698 0.7144  0.7144
## big_data                    0.0231 0.0331            0.5663 0.8149  0.8149
## predictive_analytics        0.1117 0.1356            0.6879 0.8698  0.8698
## hadoop                      0.0209 0.0302            0.5596 0.8118  0.8118
## data_privacy                0.8468 0.8561            0.9611 0.9841  0.9841
## regression_testing          0.4906 0.5178            0.8618 0.9434  0.9434
## natural_language_processing 0.2664 0.2970            0.7790 0.9088  0.9088
## matlab                             0.3098            0.7846 0.9112  0.9112
## big_data_analytics          0.3098                   0.7976 0.9166  0.9166
## stata                       0.7846 0.7976                   0.9775  0.9775
## jupyter                     0.9112 0.9166            0.9775         0.9908
## ipython                     0.9112 0.9166            0.9775 0.9908
## pycharm                     0.9112 0.9166            0.9775 0.9908  0.9908
## spyder                      0.9112 0.9166            0.9775 0.9908  0.9908
##                             pycharm spyder
## min_salary
## max_salary
## sql                         0.6552  0.6552
## python                      0.6705  0.6705
## analytics                   0.6517  0.6517
## machine_learning            0.7092  0.7092
## effective_communication     0.9817  0.9817
## fraud_analytics             0.9550  0.9550
## team_leading                0.8936  0.8936
## data_science                0.8179  0.8179
## machine_learning_engineer   0.9795  0.9795
## model_development           0.8998  0.8998
## marketing_automation        0.9270  0.9270
## r                           0.7144  0.7144
## big_data                    0.8149  0.8149
## predictive_analytics        0.8698  0.8698
## hadoop                      0.8118  0.8118
## data_privacy                0.9841  0.9841
## regression_testing          0.9434  0.9434
## natural_language_processing 0.9088  0.9088
## matlab                      0.9112  0.9112
## big_data_analytics          0.9166  0.9166
## stata                       0.9775  0.9775
## jupyter                     0.9908  0.9908
## ipython                     0.9908  0.9908
## pycharm                             0.9908
```

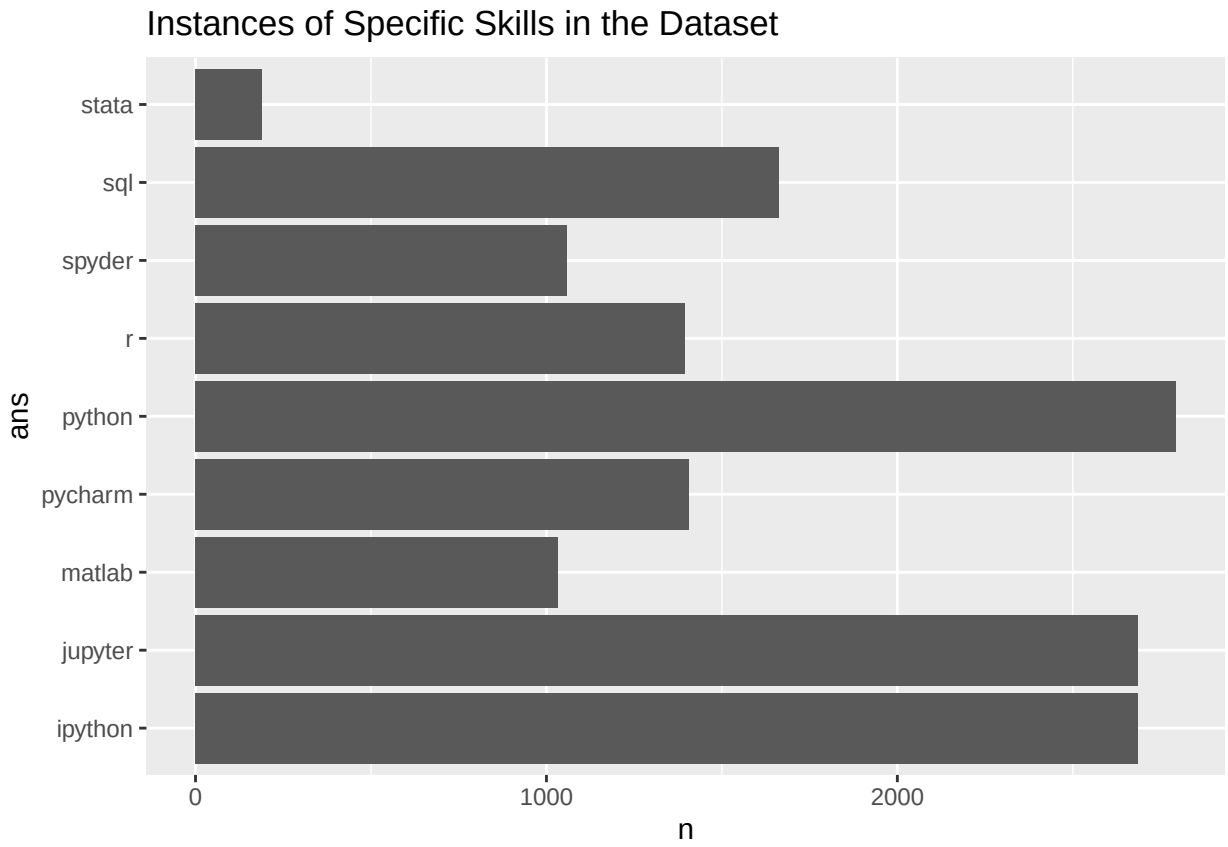
```
## spyder                        0.9908
```

```
t_cor = cor(t, method = c("spearman"))
corrplot(t_cor, title="Coorelation Income By Skills")
```



Analytis Job Seekers

```
s1 <- emp_df %>% group_by(ans) %>%
  mutate(
    n = n()
  ) %>%
  select(ans, n) %>%
  distinct()

s1 %>%
  ggplot(aes( y=ans, x=n)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title = "Instances of Specific Skills in the Dataset" )
```



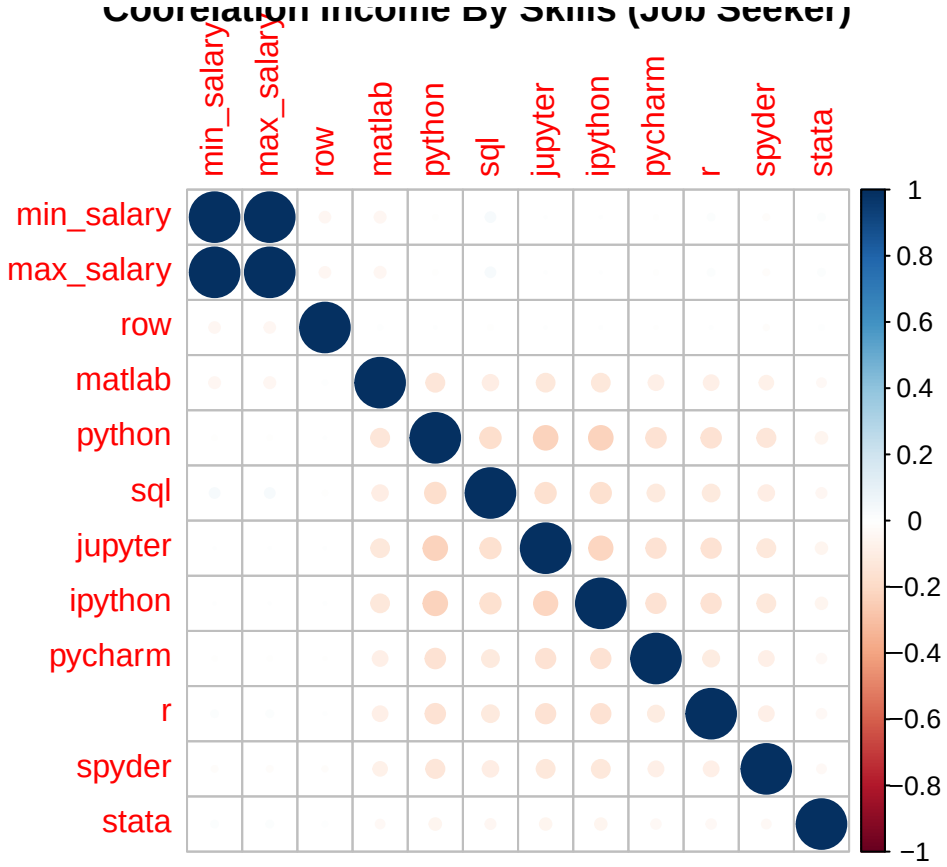

```
emp_df %>%
  ggplot() +
  geom_boxplot(mapping = aes(y=ans, x=max_salary, color=ans)) +
  labs (title = "Mapping Salary to Skillset" )
```


```
# create a wide dataframe for correlation
t2 <- emp_df %>%
  mutate(
    flag = 100
  ) %>%
  mutate (
   ans = str_replace_all(ans, " ", "_"),
   ans = str_squish(ans),
   row = row_number()
  )
t2 <- t2 %>%
  pivot_wider(
    names_from = ans,
    values_from = flag,
    values_fill = 0
  )
t2 <- t2 %>% select(-c(id, q, key_skills_id ,experience, title, age, gender, location, education ))
t2 <- t2 %>% drop_na()


t2.rcorr = rcorr(as.matrix(t2))
t2.rcorr
```

```
##            min_salary max_salary   row matlab python   sql jupyter ipython
## min_salary       1.00       0.83 -0.01  -0.04  -0.01  0.03    0.00    0.00
## max_salary       0.83       1.00  0.03  -0.02  -0.01  0.02    0.00    0.00
```

```
## row            -0.01       0.03  1.00   0.01   0.00 -0.01    0.00    0.00
## matlab         -0.04      -0.02  0.01   1.00  -0.13 -0.10   -0.13   -0.13
## python         -0.01      -0.01  0.00  -0.13   1.00 -0.17   -0.23   -0.23
## sql             0.03       0.02 -0.01  -0.10  -0.17  1.00   -0.17   -0.17
## jupyter         0.00       0.00  0.00  -0.13  -0.23 -0.17    1.00   -0.22
## ipython         0.00       0.00  0.00  -0.13  -0.23 -0.17   -0.22    1.00
## pycharm        -0.01      -0.01  0.00  -0.09  -0.15 -0.11   -0.15   -0.15
## r               0.01       0.01  0.00  -0.09  -0.15 -0.11   -0.15   -0.15
## spyder         -0.01      -0.01 -0.01  -0.08  -0.13 -0.10   -0.13   -0.13
## stata           0.02       0.02  0.01  -0.03  -0.05 -0.04   -0.05   -0.05
##            pycharm     r spyder stata
## min_salary   -0.01  0.01  -0.01  0.02
## max_salary   -0.01  0.01  -0.01  0.02
## row           0.00  0.00  -0.01  0.01
## matlab       -0.09 -0.09  -0.08 -0.03
## python       -0.15 -0.15  -0.13 -0.05
## sql          -0.11 -0.11  -0.10 -0.04
## jupyter      -0.15 -0.15  -0.13 -0.05
## ipython      -0.15 -0.15  -0.13 -0.05
## pycharm       1.00 -0.10  -0.09 -0.04
## r            -0.10  1.00  -0.09 -0.04
## spyder       -0.09 -0.09   1.00 -0.03
## stata        -0.04 -0.04  -0.03  1.00
##
## n= 14913
##
##
## P
##            min_salary max_salary row    matlab python sql    jupyter ipython
## min_salary            0.0000     0.4621 0.0000 0.4298 0.0000 0.8348  0.8348
## max_salary 0.0000                0.0002 0.0122 0.3652 0.0255 0.8186  0.8186
## row        0.4621     0.0002            0.3728 0.8139 0.3891 0.8264  0.8161
## matlab     0.0000     0.0122     0.3728        0.0000 0.0000 0.0000  0.0000
## python     0.4298     0.3652     0.8139 0.0000        0.0000 0.0000  0.0000
## sql        0.0000     0.0255     0.3891 0.0000 0.0000        0.0000  0.0000
## jupyter    0.8348     0.8186     0.8264 0.0000 0.0000 0.0000         0.0000
## ipython    0.8348     0.8186     0.8161 0.0000 0.0000 0.0000 0.0000
## pycharm    0.2372     0.3828     0.7864 0.0000 0.0000 0.0000 0.0000  0.0000
## r          0.0704     0.2660     0.7686 0.0000 0.0000 0.0000 0.0000  0.0000
## spyder     0.1497     0.2735     0.1991 0.0000 0.0000 0.0000 0.0000  0.0000
## stata      0.0057     0.0039     0.3593 0.0002 0.0000 0.0000 0.0000  0.0000
##            pycharm r      spyder stata
## min_salary 0.2372  0.0704 0.1497 0.0057
## max_salary 0.3828  0.2660 0.2735 0.0039
## row        0.7864  0.7686 0.1991 0.3593
## matlab     0.0000  0.0000 0.0000 0.0002
## python     0.0000  0.0000 0.0000 0.0000
## sql        0.0000  0.0000 0.0000 0.0000
## jupyter    0.0000  0.0000 0.0000 0.0000
## ipython    0.0000  0.0000 0.0000 0.0000
## pycharm            0.0000 0.0000 0.0000
## r          0.0000         0.0000 0.0000
## spyder     0.0000  0.0000        0.0001
## stata      0.0000  0.0000 0.0001
```

```
t2_cor = cor(t2, method = c("spearman"))
corrplot(t2_cor, title="Coorelation Income By Skills (Job Seeker)")
```



Conclusions

Based on our analysis we can identify a few skills that do not correlate with income or salary. However given how low the overal levels of correlation it is difficult to come to any additional conclusions.

- **Open Roles**
- marketing automation 0.2043 min and 0.1458 max
- data privacy 0.4249 min and 0.3341 max
- matlab 0.7624 min and 0.6057 max
- big data analytics min 0.7172 and 0.1767 max
- and interestingly enough pyhton has a slightly negative correlation
- **Job Seekers**
- python 0.3131 min and 0.3132 max

The results could be a factor of our limited datasets or it could also be caused by additional factors that impact salary that are not included in the data. Some items could include: - where you received your college education - industry - geography within the course grained location (NY, SFO markets)