DATA607_Project 3 - Data Science Skills

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

2021-10-17

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"))

# t3 <- tmp %>% group_by(Q6) %>%
#   mutate(
#    n = n()
#   ) %>%
#   select(Q6,n) %>%
#   distinct()

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 %>%
  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","")
  )
```

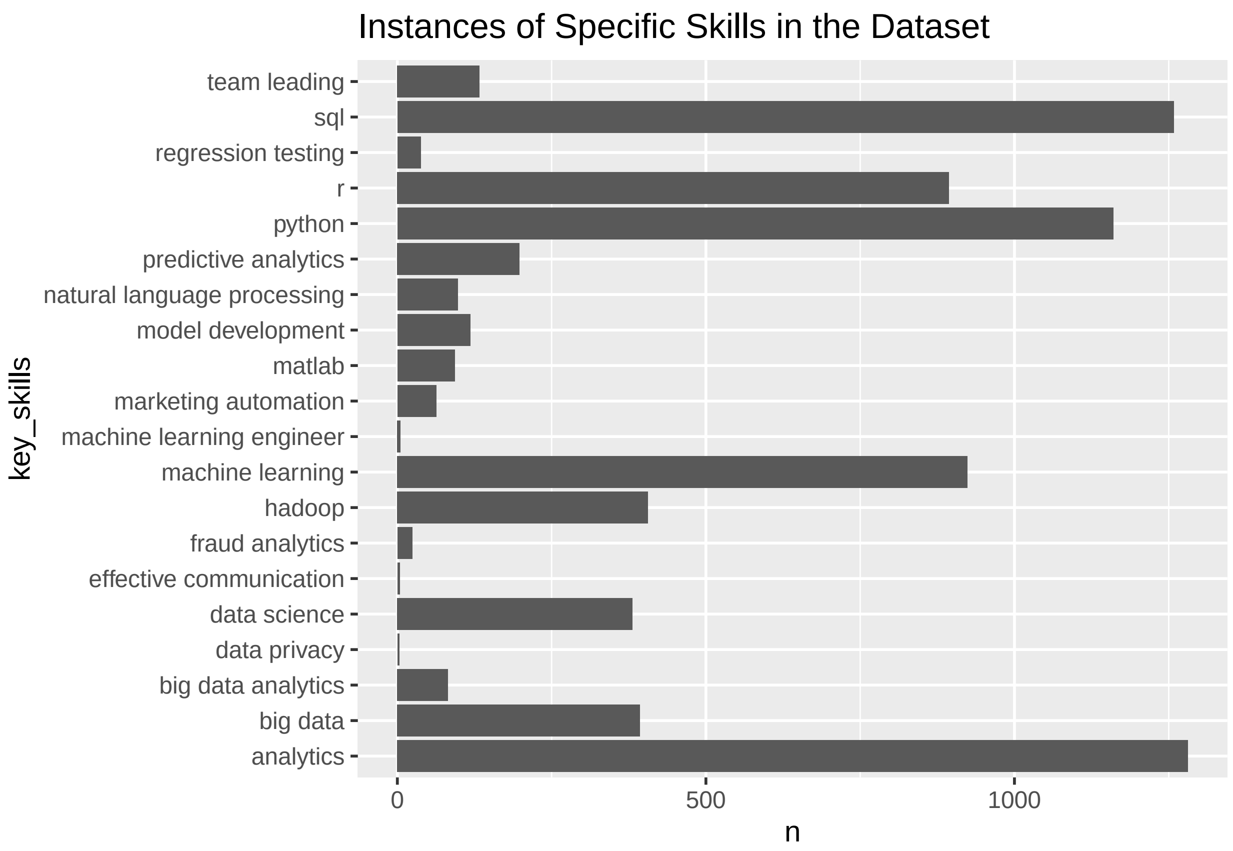
```r
tmp <- tmp %>%
  separate(
      Q9,
      c("min_salary" , "max_salary"),
      convert = TRUE,
      sep = "-"
  )
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 16 rows [4836,
## 4837, 4912, 4913, 5343, 5344, 5345, 5346, 5347, 5531, 5532, 5533, 5744, 5745,
## 5822, 5823].
```

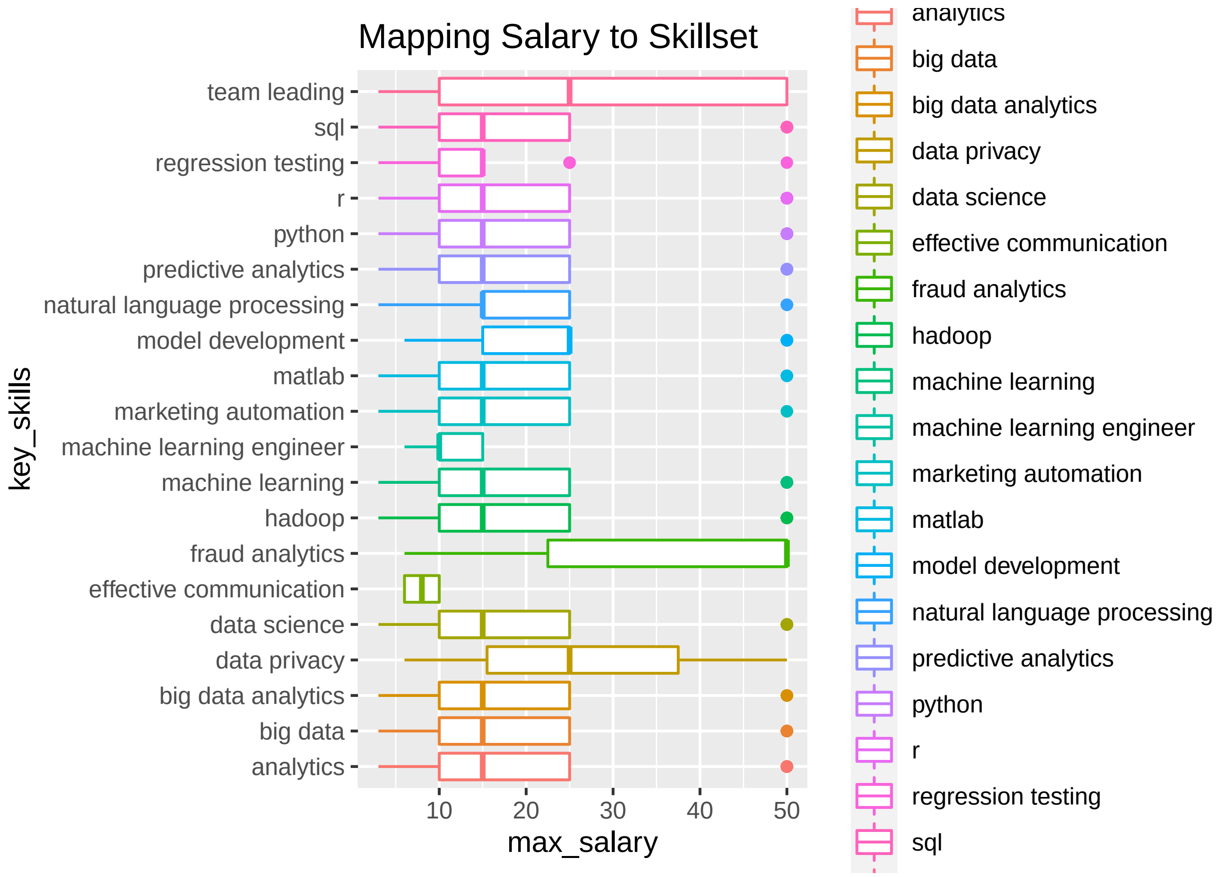
```r
tmp <- tmp %>%
  transform(
    min_salary = as.numeric(min_salary),
    max_salary = as.numeric(max_salary)
  )


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

```r
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" )
```



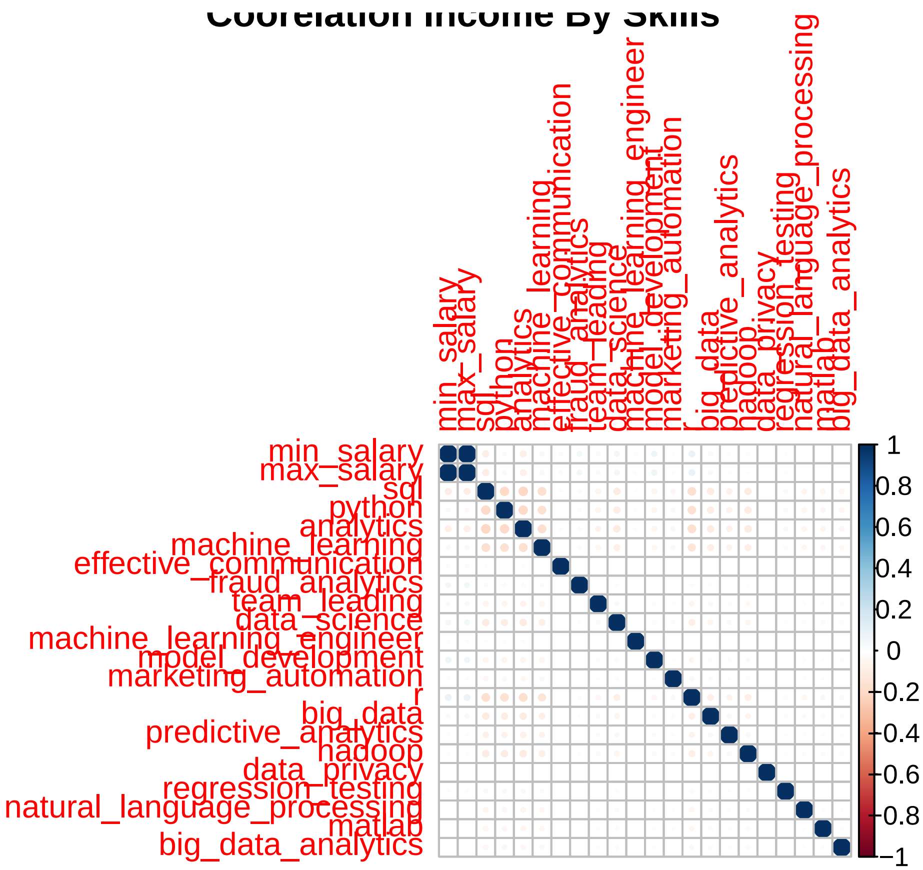
```
skills_df %>%
  ggplot() +
  geom_boxplot(mapping = aes(y=key_skills, x=max_salary, color=key_skills)) +
  labs (title = "Mapping Salary to Skillset" )
```



```
# 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
```

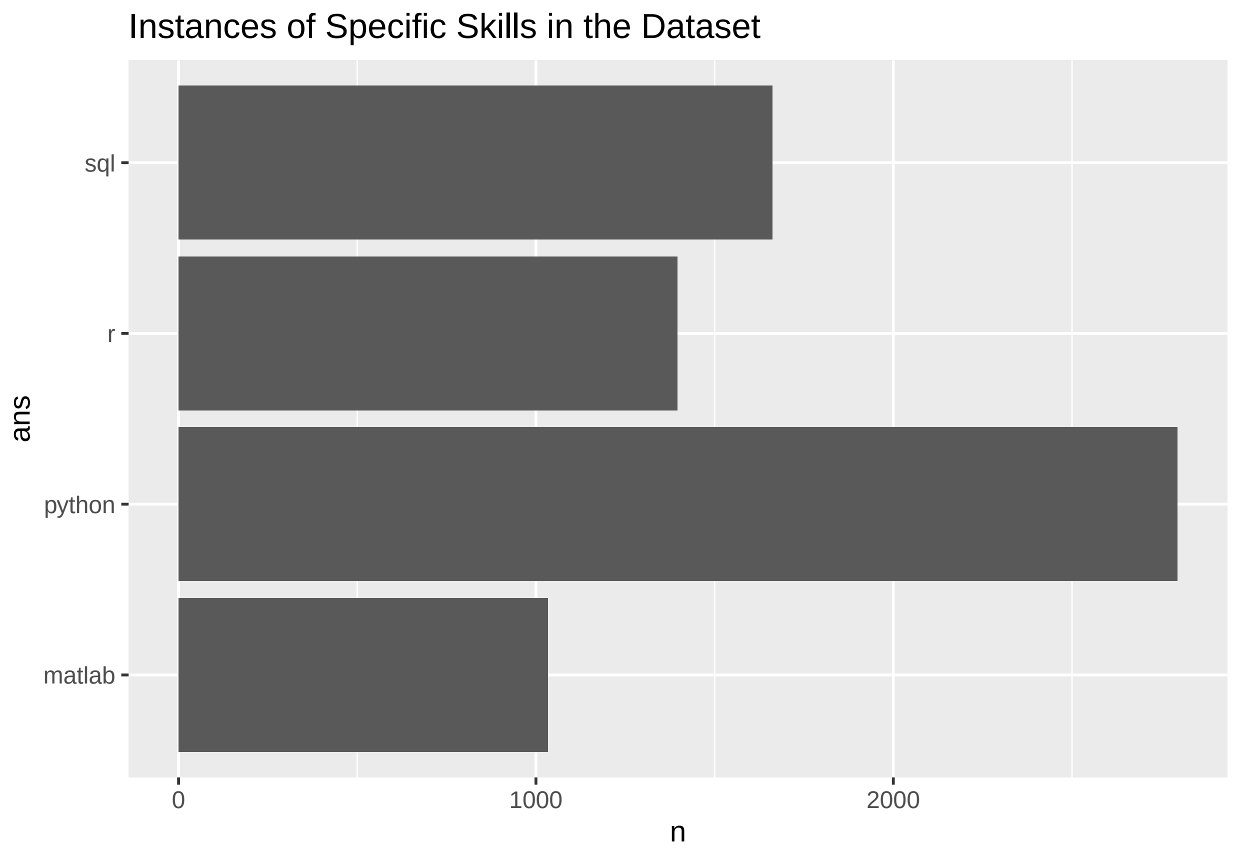
```
##                               min_salary max_salary   sql python analytics
## min_salary                          1.00       0.98 -0.11  -0.04     -0.06
## max_salary                          0.98       1.00 -0.11  -0.04     -0.04
## sql                              -0.11      -0.11  1.00  -0.19     -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
##                             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
##                             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
##                             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
##                             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
##                             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
##                             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
##                             matlab big_data_analytics
## min_salary                    0.00               0.00
## max_salary                   -0.01               0.00
## sql                          -0.05              -0.05
## python                        -0.05          -0.04
## analytics                     -0.05          -0.05
## machine_learning              -0.04          -0.04
## effective_communication        0.00           0.00
## fraud_analytics               -0.01          -0.01
## team_leading                  -0.01          -0.01
## data_science                  -0.03          -0.02
## machine_learning_engineer      0.00           0.00
## model_development             -0.01          -0.01
## marketing_automation          -0.01          -0.01
## r                             -0.04          -0.04
## big_data                      -0.03          -0.02
## predictive_analytics          -0.02          -0.02
## hadoop                        -0.03          -0.02
## data_privacy                   0.00           0.00
## regression_testing            -0.01          -0.01
## natural_language_processing  -0.01          -0.01
## matlab                         1.00          -0.01
## big_data_analytics            -0.01           1.00
##
## n= 7556
##
##
## P
##                             min_salary max_salary sql    python analytics
## min_salary                             0.0000     0.0000 0.0017 0.0000
## max_salary                  0.0000                0.0000 0.0015 0.0005
## sql                         0.0000     0.0000            0.0000 0.0000
## python                      0.0017     0.0015     0.0000        0.0000
## analytics                   0.0000     0.0005     0.0000 0.0000
## machine_learning            0.0022     0.0053     0.0000 0.0000 0.0000
## effective_communication     0.0455     0.0838     0.3713 0.3943 0.3661
## fraud_analytics             0.0000     0.0000     0.0283 0.0367 0.0266
## 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   0.1599     0.1622     0.3175 0.3409 0.3123
## model_development           0.0000     0.0000     0.0000 0.0000 0.0000
## marketing_automation        0.2043     0.1458     0.0004 0.0007 0.0003
## r                           0.0000     0.0000     0.0000 0.0000 0.0000
## big_data                    0.0066     0.0217     0.0000 0.0000 0.0000
## predictive_analytics        0.0514     0.0472     0.0000 0.0000 0.0000
## hadoop                      0.8325     0.2988     0.0000 0.0000 0.0000
## data_privacy                0.4249     0.3341     0.4388 0.4607 0.4338
## regression_testing          0.1068     0.0758     0.0057 0.0085 0.0052
## natural_language_processing 0.9438     0.5234     0.0000 0.0000 0.0000
## matlab                      0.7624     0.6057     0.0000 0.0000 0.0000
## big_data_analytics          0.7172     0.7267     0.0000 0.0001 0.0000
##                             machine_learning effective_communication
## min_salary                  0.0022           0.0455
## max_salary                  0.0053           0.0838
## sql                         0.0000           0.3713
## python                      0.0000           0.3943
## analytics                   0.0000           0.3661
## machine_learning                             0.4553
## effective_communication     0.4553
## fraud_analytics             0.0670           0.9101
## team_leading                0.0000           0.7889
## data_science                0.0000           0.6449
## machine_learning_engineer   0.4038           0.9590
## model_development           0.0000           0.8011
## marketing_automation        0.0029           0.8545
## r                           0.0000           0.4637
## big_data                    0.0000           0.6394
## predictive_analytics        0.0000           0.7428
## hadoop                      0.0000           0.6336
## data_privacy                0.5179           0.9682
## regression_testing          0.0211           0.8869
## natural_language_processing 0.0002           0.8186
## matlab                      0.0003           0.8233
## big_data_analytics          0.0007           0.8341
##                             fraud_analytics team_leading data_science
## min_salary                  0.0000          0.0000       0.0000
## max_salary                  0.0000          0.0000       0.0000
## sql                         0.0283          0.0000       0.0000
## python                      0.0367          0.0000       0.0000
## analytics                   0.0266          0.0000       0.0000
## machine_learning            0.0670          0.0000       0.0000
## effective_communication     0.9101          0.7889       0.6449
## fraud_analytics                             0.5114       0.2582
## team_leading                0.5114                       0.0073
## data_science                0.2582          0.0073
## machine_learning_engineer   0.8995          0.7647       0.6063
## model_development           0.5366          0.1428       0.0116
## marketing_automation        0.6528          0.2861       0.0663
## r                           0.0723          0.0000       0.0000
## big_data                    0.2505          0.0064       0.0000
## predictive_analytics        0.4209          0.0563       0.0010
## hadoop                      0.2424          0.0056       0.0000
## data_privacy                0.9221          0.8167       0.6898
## regression_testing          0.7272          0.4082       0.1545
## natural_language_processing 0.5739          0.1823       0.0217
## matlab                      0.5839          0.1940       0.0253
## big_data_analytics          0.6073          0.2230       0.0359
##                             machine_learning_engineer model_development
## min_salary                  0.1599                    0.0000
## max_salary                  0.1622                    0.0000
## sql                         0.3175                    0.0000
## python                      0.3409                    0.0000
## analytics                   0.3123                    0.0000
## machine_learning            0.4038                    0.0000
## effective_communication     0.9590                    0.8011
## fraud_analytics             0.8995                    0.5366
## team_leading                0.7647                    0.1428
## data_science                0.6063                    0.0116
## machine_learning_engineer                             0.7782
## model_development           0.7782
## marketing_automation        0.8375                    0.3155
## r                           0.4126                    0.0000
## big_data                    0.6004                    0.0103
## predictive_analytics        0.7137                    0.0725
## hadoop                      0.5941                    0.0091
## data_privacy                0.9644                    0.8273
## regression_testing          0.8737                    0.4364
## natural_language_processing 0.7977                    0.2095
## matlab                      0.8029                    0.2217
## big_data_analytics          0.8148                    0.2515
##                             marketing_automation r      big_data
## min_salary                  0.2043               0.0000 0.0066
## max_salary                  0.1458               0.0000 0.0217
## 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     0.8545               0.4637 0.6394
## fraud_analytics             0.6528               0.0723 0.2505
## team_leading                0.2861               0.0000 0.0064
## data_science                0.0663               0.0000 0.0000
## machine_learning_engineer   0.8375               0.4126 0.6004
## model_development           0.3155               0.0000 0.0103
## marketing_automation                             0.0035 0.0619
## r                           0.0035                      0.0000
## big_data                    0.0619               0.0000
## predictive_analytics        0.1911               0.0000 0.0008
## hadoop                      0.0575               0.0000 0.0000
## data_privacy                0.8738               0.5257 0.6849
## regression_testing          0.5710               0.0236 0.1478
## natural_language_processing 0.3610               0.0003 0.0196
## matlab                      0.3737               0.0004 0.0230
## big_data_analytics          0.4039               0.0008 0.0330
##                             predictive_analytics hadoop data_privacy
## min_salary                  0.0514               0.8325 0.4249
## max_salary                  0.0472               0.2988 0.3341
## sql                         0.0000               0.0000 0.4388
## python                      0.0000               0.0000 0.4607
## analytics                   0.0000               0.0000 0.4338
## machine_learning            0.0000               0.0000 0.5179
## effective_communication     0.7428               0.6336 0.9682
## fraud_analytics             0.4209               0.2424 0.9221
## team_leading                0.0563               0.0056 0.8167
## data_science                0.0010               0.0000 0.6898
## machine_learning_engineer   0.7137               0.5941 0.9644
## model_development           0.0725               0.0091 0.8273
## marketing_automation        0.1911               0.0575 0.8738
## r                           0.0000               0.0000 0.5257
## big_data                    0.0008               0.0000 0.6849
## predictive_analytics                             0.0007 0.7763
## hadoop                      0.0007                      0.6798
## data_privacy                0.7763               0.6798
## regression_testing          0.3108               0.1409 0.9020
## natural_language_processing 0.1022               0.0176 0.8426
## matlab                      0.1115               0.0208 0.8467
## big_data_analytics          0.1353               0.0300 0.8560
##                             regression_testing natural_language_processing
## min_salary                  0.1068             0.9438
## max_salary                  0.0758             0.5234
## sql                         0.0057             0.0000
## python                      0.0085             0.0000
## analytics                   0.0052             0.0000
## machine_learning            0.0211             0.0002
## effective_communication     0.8869             0.8186
## fraud_analytics             0.7272             0.5739
## team_leading                0.4082             0.1823
## data_science                0.1545             0.0217
## machine_learning_engineer   0.8737             0.7977
## model_development           0.4364             0.2095
## marketing_automation        0.5710             0.3610
## r                           0.0236             0.0003
## big_data                    0.1478             0.0196
## predictive_analytics        0.3108             0.1022
## hadoop                      0.1409             0.0176
## data_privacy                0.9020             0.8426
## regression_testing                             0.4788
## natural_language_processing 0.4788
## matlab                      0.4903             0.2661
## big_data_analytics          0.5175             0.2967
##                             matlab big_data_analytics
## min_salary                  0.7624 0.7172
## max_salary                  0.6057 0.7267
## sql                         0.0000 0.0000
## python                      0.0000 0.0001
## analytics                   0.0000 0.0000
## machine_learning            0.0003 0.0007
## effective_communication     0.8233 0.8341
## fraud_analytics             0.5839 0.6073
## team_leading                0.1940 0.2230
## data_science                0.0253 0.0359
## machine_learning_engineer   0.8029 0.8148
## model_development           0.2217 0.2515
## marketing_automation        0.3737 0.4039
## r                           0.0004 0.0008
## big_data                    0.0230 0.0330
## predictive_analytics        0.1115 0.1353
## hadoop                      0.0208 0.0300
## data_privacy                0.8467 0.8560
## regression_testing          0.4903 0.5175
## natural_language_processing 0.2661 0.2967
## matlab                             0.3095
## big_data_analytics          0.3095
```

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



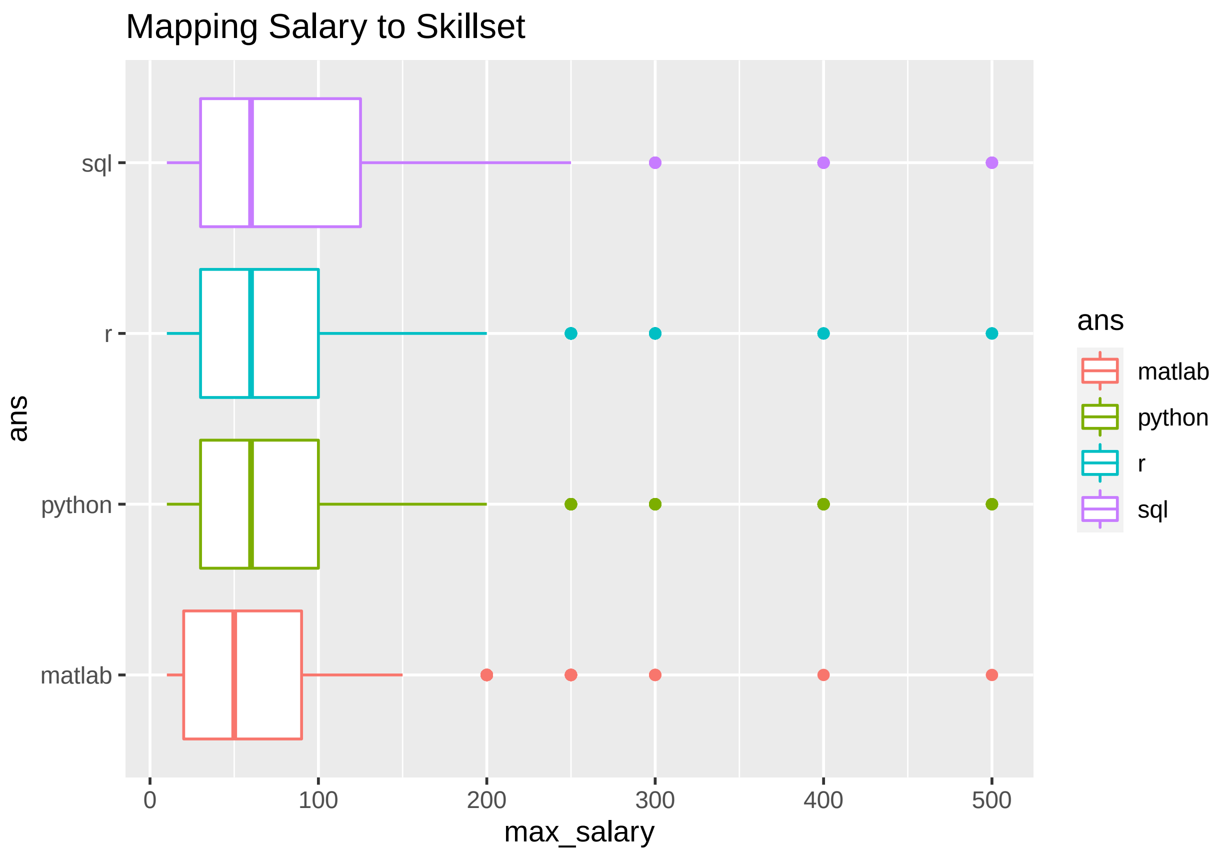
Analytis Job Seekers

```
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" )
```

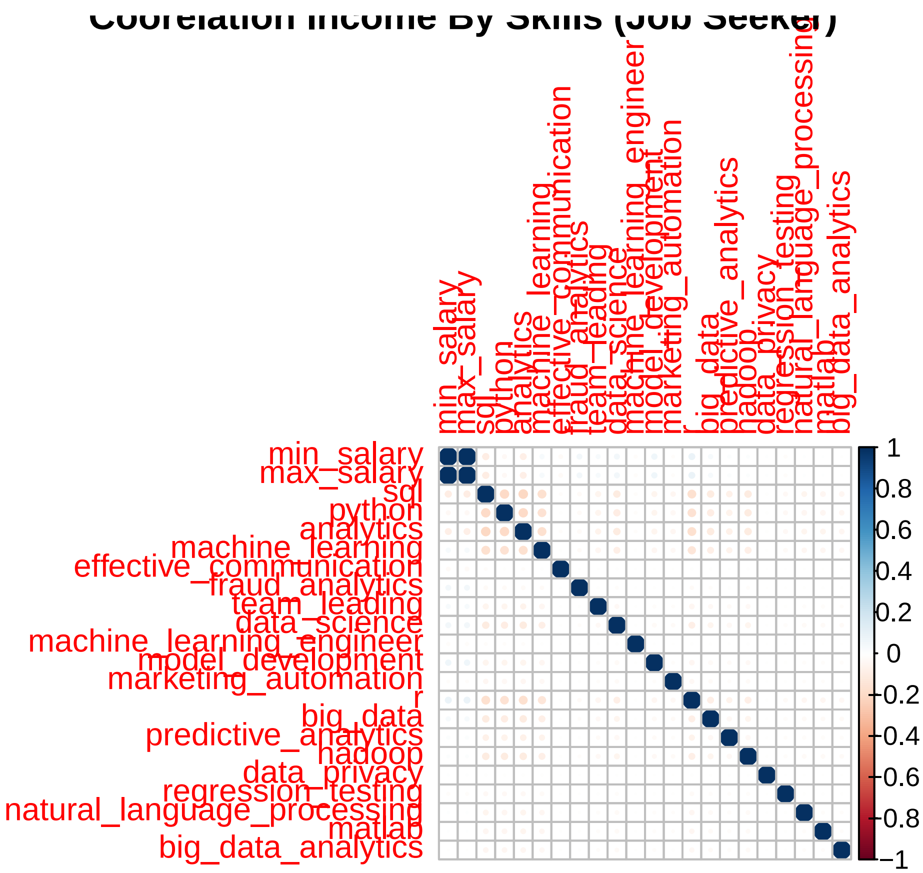
```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



```
# 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     r
## min_salary       1.00       0.99 -0.04  -0.07  -0.01  0.05  0.02
## max_salary       0.99       1.00 -0.03  -0.07  -0.01  0.05  0.02
## row             -0.04      -0.03  1.00  0.01  0.00 -0.01  0.00
## matlab          -0.07      -0.07  0.01  1.00 -0.35 -0.24 -0.21
## python          -0.01      -0.01  0.00 -0.35  1.00 -0.47 -0.42
## sql              0.05       0.05 -0.01 -0.24 -0.47  1.00 -0.28
## r                0.02       0.02  0.00 -0.21 -0.42 -0.28  1.00
##
## n= 6869
##
##
## P
##            min_salary max_salary row    matlab python sql    r
## min_salary            0.0000     0.0009 0.0000 0.3132 0.0000 0.0898
## max_salary 0.0000                0.0044 0.0000 0.3733 0.0000 0.1538
## row        0.0009     0.0044            0.3982 0.8818 0.2958 0.8563
## matlab     0.0000     0.0000     0.3982        0.0000 0.0000 0.0000
## python     0.3132     0.3733     0.8818 0.0000        0.0000 0.0000
## sql        0.0000     0.0000     0.2958 0.0000 0.0000        0.0000
## r          0.0898     0.1538     0.8563 0.0000 0.0000 0.0000
```

```
t2_cor = cor(t, method = c("spearman"))
corrplot(t_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)