R Project Regression

Opening the data

```
df <- read.csv("Levels_Fyi_Salary_Data.csv")
names(df)
```

```
##  [1] "timestamp"               "company"
##  [3] "level"                   "title"
##  [5] "totalyearlycompensation" "location"
##  [7] "yearsofexperience"       "yearsatcompany"
##  [9] "tag"                     "basesalary"
## [11] "stockgrantvalue"         "bonus"
## [13] "gender"                  "otherdetails"
## [15] "cityid"                  "dmaid"
## [17] "rowNumber"               "Masters_Degree"
## [19] "Bachelors_Degree"        "Doctorate_Degree"
## [21] "Highschool"              "Some_College"
## [23] "Race_Asian"              "Race_White"
## [25] "Race_Two_Or_More"        "Race_Black"
## [27] "Race_Hispanic"           "Race"
## [29] "Education"
```

```
head(df)
```

```
##           timestamp   company level                        title
## 1  6/7/2017 11:33:27    Oracle    L3              Product Manager
## 2 6/10/2017 17:11:29      eBay  SE 2            Software Engineer
## 3 6/11/2017 14:53:57    Amazon    L7              Product Manager
## 4  6/17/2017 0:23:14     Apple    M1 Software Engineering Manager
## 5 6/20/2017 10:58:51 Microsoft    60            Software Engineer
## 6 6/21/2017 17:27:47 Microsoft    63            Software Engineer
##   totalyearlycompensation         location yearsofexperience yearsatcompany
## 1                  127000  Redwood City, CA               1.5            1.5
## 2                  100000 San Francisco, CA               5.0            3.0
## 3                  310000       Seattle, WA               8.0            0.0
## 4                  372000     Sunnyvale, CA               7.0            5.0
## 5                  157000 Mountain View, CA               5.0            3.0
## 6                  208000       Seattle, WA               8.5            8.5
##    tag basesalary stockgrantvalue bonus gender otherdetails cityid dmaid
## 1 <NA>     107000           20000 10000   <NA>         <NA>   7392   807
## 2 <NA>          0               0     0   <NA>         <NA>   7419   807
## 3 <NA>     155000               0     0   <NA>         <NA>  11527   819
## 4 <NA>     157000          180000 35000   <NA>         <NA>   7472   807
## 5 <NA>          0               0     0   <NA>         <NA>   7322   807
## 6 <NA>          0               0     0   <NA>         <NA>  11527   819
##   rowNumber Masters_Degree Bachelors_Degree Doctorate_Degree Highschool
```

```
## 1         1             0                 0                 0           0
## 2         2             0                 0                 0           0
## 3         3             0                 0                 0           0
## 4         7             0                 0                 0           0
## 5         9             0                 0                 0           0
## 6        11             0                 0                 0           0
##   Some_College Race_Asian Race_White Race_Two_Or_More Race_Black Race_Hispanic
## 1            0          0          0                0          0             0
## 2            0          0          0                0          0             0
## 3            0          0          0                0          0             0
## 4            0          0          0                0          0             0
## 5            0          0          0                0          0             0
## 6            0          0          0                0          0             0
##   Race Education
## 1 <NA>      <NA>
## 2 <NA>      <NA>
## 3 <NA>      <NA>
## 4 <NA>      <NA>
## 5 <NA>      <NA>
## 6 <NA>      <NA>
```

```
#link for dataset https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries
```

Cleaning the data: I removed the races columns because they were redundant since there was another column which simply had a race value. I also removed other columns such as rows since I decided to use only companies from FAANG, and most of which have headquarters in selective areas (Silicon Valley and Seattle for Microsoft)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#get rid of non necessary columns
df <- df %>%
mutate(timestamp = NULL)
df <- df %>%
mutate(Race_Asian = NULL)
df <- df %>%
mutate(Race_White = NULL)
df <- df %>%
mutate(Race_Two_Or_More = NULL)
df <- df %>%
mutate(Race_Black = NULL)
```

```r
df <- df %>%
mutate(Race_Hispanic = NULL)
df <- df %>%
mutate(Masters_Degree = NULL)
df <- df %>%
mutate(Bachelors_Degree = NULL)
df <- df %>%
mutate(Doctorate_Degree = NULL)
df <- df %>%
mutate(Highschool = NULL)
df <- df %>%
mutate(Some_College = NULL)
df <- df %>%
mutate(otherdetails = NULL)
df <- df %>%
mutate(dmaid = NULL)
df <- df %>%
mutate(level = NULL)
df <- df %>%
mutate(title = NULL)
df <- df %>%
mutate(tag = NULL)
df <- df %>%
mutate(rowNumber = NULL)
df <- df %>%
mutate(cityid = NULL)
df <- df %>%
mutate(location = NULL)

sapply(df, function(x) sum(is.na(x)==TRUE))
```

```
##                 company totalyearlycompensation       yearsofexperience
##                       0                       0                       0
##          yearsatcompany              basesalary         stockgrantvalue
##                       0                       0                       0
##                   bonus                  gender                    Race
##                       0                   19540                   40215
##               Education
##                   32272
```

```r
#remove any row with NA
df1 <- na.omit(df)
#data then becomes only 20k rows which is too much removed

#dataset only including the FAANG companies
df2 <- df[which(df$company == "Facebook" | df$company=="Apple"| df$company=="Amazon"| df$company=="Netf

#FAANG With NAs all omitted (most filtered dataset)
df3 <- na.omit(df2)

df3$company <- as.factor(df3$company)
df3$company <- as.factor(df3$company)
df3$Race <- as.factor(df3$Race)
```

```r
df3$Race <- as.factor(df3$Race)
df3$gender <- as.factor(df3$gender)
df3$gender <- as.factor(df3$gender)
df3$Education <- as.factor(df3$Education)
df3$Education <- as.factor(df3$Education)
```

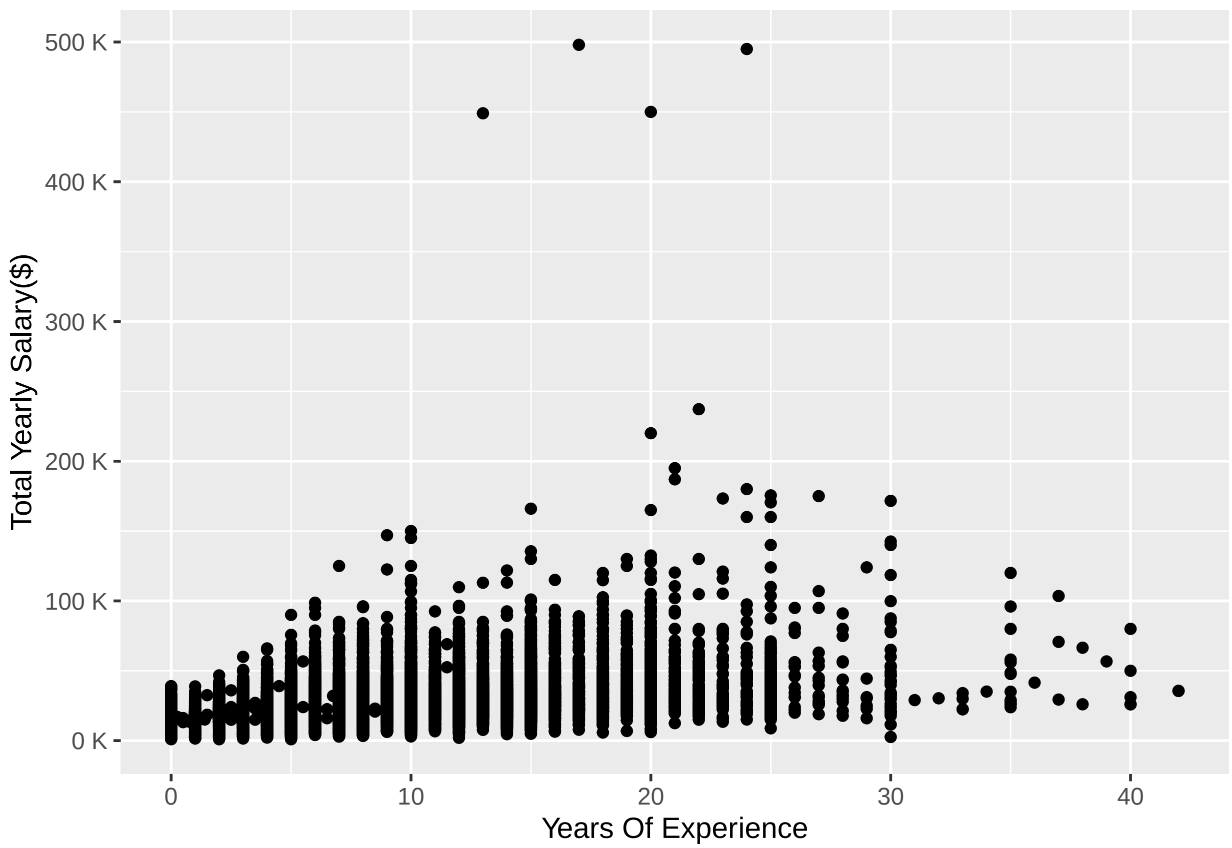
Data Exploration: From the graphs we can see that there is actually a little cluster around the beginning of the graphs, showing that many of those who work at FAANG usually have not worked there for an extremely long time. the target was the Totalyearlycompensation. We can see that the yearsatcompany and yearsofexperience were good predictors for total yearly salary. EducationPHD was another good predictor for Totalyearlycompensation.

```r
summary(df3)
```

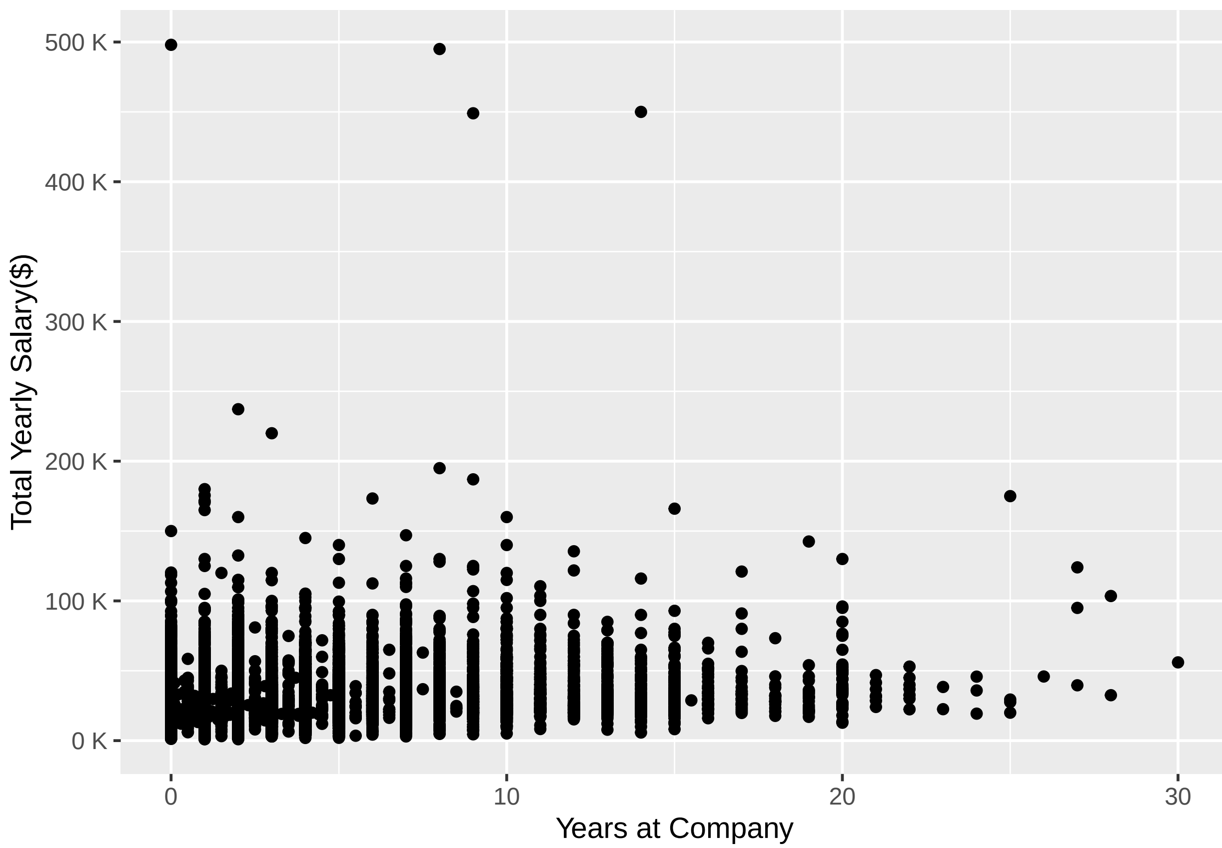
```
##       company     totalyearlycompensation yearsofexperience yearsatcompany
##  Amazon   :2583   Min.   :  13000         Min.   : 0.000    Min.   : 0.000
##  Apple    : 603   1st Qu.: 165000         1st Qu.: 3.000    1st Qu.: 0.000
##  Facebook :1001   Median : 216000         Median : 6.000    Median : 2.000
##  Google   :1314   Mean   : 247497         Mean   : 7.279    Mean   : 2.532
##  Microsoft:1619   3rd Qu.: 296000         3rd Qu.:10.000    3rd Qu.: 4.000
##  Netflix  :  80   Max.   :4980000         Max.   :39.000    Max.   :28.000
##    basesalary      stockgrantvalue      bonus            gender
##  Min.   : 10000   Min.   :     0   Min.   :     0   Female:1329
##  1st Qu.:120000   1st Qu.: 22000   1st Qu.: 10000   Male  :5840
##  Median :148000   Median : 45000   Median : 20000   Other :  31
##  Mean   :148198   Mean   : 73588   Mean   : 24017
##  3rd Qu.:170000   3rd Qu.: 95000   3rd Qu.: 30000
##  Max.   :893000   Max.   :954000   Max.   :555000
##           Race                   Education
##  Asian      :3821   Bachelor's Degree:3393
##  Black      : 261   Highschool       : 108
##  Hispanic   : 427   Master's Degree  :3156
##  Two Or More: 268   PhD              : 445
##  White      :2423   Some College     :  98
##
```

```r
library(ggplot2)
library(scales)

ggplot(df2,aes(x=yearsofexperience,y=totalyearlycompensation))+geom_point()+ scale_y_continuous(labels =
```




```
ggplot(df2,aes(x=yearsatcompany,y=totalyearlycompensation))+geom_point()+ scale_y_continuous(labels = la
```



Linear Regression

```
set.seed(1234)
i <- sample(1:nrow(df3),nrow(df3)*.75,replace=FALSE)
train <- df3[i,]
test <- df3[-i,]

lm1 <- lm(totalyearlycompensation~. , data=train)
summary(lm1)
```

```
##
## Call:
## lm(formula = totalyearlycompensation ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158330   -5657    -153    5010 3540361
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.678e+04  3.148e+03  -5.329 1.03e-07 ***
## companyApple             -2.209e+03  2.963e+03  -0.745 0.456026
## companyFacebook          -2.814e+02  2.599e+03  -0.108 0.913783
## companyGoogle            -4.513e+02  2.240e+03  -0.201 0.840343
## companyMicrosoft         -6.118e+03  2.129e+03  -2.873 0.004080 **
## companyNetflix           -3.731e+04  1.070e+04  -3.488 0.000491 ***
```

```
## yearsofexperience        -1.855e+02  1.692e+02  -1.096 0.273025
## yearsatcompany            9.396e+02  2.875e+02   3.268 0.001089 **
## basesalary                1.110e+00  2.178e-02  50.977  < 2e-16 ***
## stockgrantvalue           9.214e-01  1.150e-02  80.148  < 2e-16 ***
## bonus                     1.320e+00  3.056e-02  43.200  < 2e-16 ***
## genderMale                1.515e+03  1.977e+03   0.766 0.443516
## genderOther              -1.090e+04  1.066e+04  -1.023 0.306515
## RaceBlack                 8.279e+03  4.220e+03   1.962 0.049861 *
## RaceHispanic             -1.659e+03  3.328e+03  -0.499 0.618084
## RaceTwo Or More          -3.069e+03  4.046e+03  -0.759 0.448136
## RaceWhite                -2.573e+03  1.746e+03  -1.473 0.140725
## EducationHighschool       1.162e+03  6.194e+03   0.188 0.851224
## EducationMaster's Degree  8.089e+02  1.657e+03   0.488 0.625525
## EducationPhD             -3.761e+03  3.250e+03  -1.157 0.247166
## EducationSome College    -7.079e+02  6.907e+03  -0.102 0.918369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55710 on 5379 degrees of freedom
## Multiple R-squared:  0.8594, Adjusted R-squared:  0.8588
## F-statistic:  1643 on 20 and 5379 DF,  p-value: < 2.2e-16
```

```r
predLin <- predict(lm1,newdata=test)
corr <- cor(predLin,test$totalyearlycompensation)
cat("The correlation is  ",corr)
```

```
## The correlation is   0.7957884
```

kNN

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
train$company <- as.integer(train$company)
test$company <- as.integer(test$company)
train$Race <- as.integer(train$Race)
test$Race <- as.integer(test$Race)
train$gender <- as.integer(train$gender)
test$gender <- as.integer(test$gender)
train$Education <- as.integer(train$Education)
test$Education <- as.integer(test$Education)


fit <- knnreg(train[,3:10],train[,2],k=2)
predictions <- predict(fit, test[,3:10])

corr2 <- cor(predictions, test$totalyearlycompensation)
mse <- mean((predictions - test$totalyearlycompensation)^2)

cat("The correlation for kNN was ", corr2)
```

```
## The correlation for kNN was  0.9836536
```

Result Analysis: The best performing algorithm was the kNN for this dataset. The linear regression had a significantly lower accuracy than the kNN for this dataset. This is likely due to the linear regression assuming that the relationship would be linear. However salaries are not linear and have a variety of factors that influence them. This is why kNN was a much better predictor, because it was able to check the neighbors and see how far off they were. From this data we were able to learn that yearsofexperience, yearsatcompany, and EducationPHD were all very indicitive of what the totalyearly compensation would be.