Applications

```
setwd("~/R/")
```

```
Jobs <- read.csv("rows.csv")
```

Introduction

Our dataset was retrieved from Kaggle and contains the different type of job positions that people in Los Angeles, California applied for. The categories are specified by the names of the jobs and the date that they were listed, in addition to the genders of the people who applied, and the races of the people that applied. We originally started off with 187 observations and 14 columns.

Motivation

In many corporate settings, there is a definite lack of representation of people of color in higher ranking positions. We were curious to see if this hindered more people of color from applying for these positions if they felt they would not receive the job. After we finish college, a large portion of us will probably start out in corporate positions. Analyzing this data can have an impact on our future decision making when applying for various positions.

Reseach Question

On average, do Hispanics apply for high ranking positions (ie. senior and chief titles)as much as African American people?

Null Hypothesis

We believe that African Americans apply to higher ranking positions more than Hispanics.

Alternative Hypothesis

African Americans are not applying to higher ranking positions as much as Hispanics.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven)
library(sjmisc)
```

```
##
## Attaching package: 'sjmisc'
```

```
## The following object is masked from 'package:purrr':
##
##     is_empty
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
## The following object is masked from 'package:tibble':
##
##     add_case
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(dplyr)
library(ggplot2)
library(tidyselect)
```

Data Wrangling

After viewing our observations and variables, we wanted to remove the “Fiscal.Year” column because the years were already listed in the Job Description column.

```
Jobs <- select(Jobs, c(Job.Number, Job.Description, Apps.Received, Female, Male, Unknown_Gender, Black, Hispanic, Asian, Caucasian, American.Indian..Alaskan.Native, Filipino, Unknown_Ethnicity))
glimpse(Jobs)
```

```
## Rows: 187
## Columns: 13
## $ Job.Number                       <chr> "9206 OP 2014/04/18", "1223 P 2013/08/~
## $ Job.Description                  <chr> "311 DIRECTOR 9206", "ACCOUNTING CLERK~
## $ Apps.Received                    <int> 54, 648, 51, 48, 40, 161, 102, 702, 10~
## $ Female                           <int> 20, 488, 13, 9, 15, 89, 53, 430, 3, 46~
## $ Male                             <int> 31, 152, 37, 38, 24, 66, 48, 240, 101,~
## $ Unknown_Gender                   <int> 3, 8, 1, 1, 1, 6, 1, 32, 1, 19, 8, 2, ~
## $ Black                            <int> 25, 151, 8, 21, 3, 12, 22, 96, 24, 197~
## $ Hispanic                         <int> 18, 204, 12, 14, 7, 36, 18, 173, 44, 2~
## $ Asian                            <int> 1, 123, 9, 3, 7, 20, 14, 84, 2, 197, 1~
## $ Caucasian                        <int> 6, 62, 20, 7, 19, 73, 37, 211, 27, 116~
## $ American.Indian..Alaskan.Native <int> 0, 3, 0, 0, 1, 0, 2, 5, 0, 3, 1, 1, 1,~
## $ Filipino                         <int> 0, 79, 0, 1, 1, 6, 4, 40, 6, 103, 10, ~
## $ Unknown_Ethnicity                <int> 4, 26, 2, 2, 2, 14, 5, 93, 2, 52, 25, ~
```

The next step was shortening the variable names.

```
colnames(Jobs) <- c('JobNum', 'JobDesc', 'AppsRec', 'Female', 'Male', 'UnknownGen', 'Black', 'Hispanic', 'Asian', 'white', 'AIAN', 'Fil', 'UnknownEth')
```

Then we wanted to specifically look at who was applying for higher ranking positions. First we made a new data frame that specified “Top Jobs”

```
TopJobs <- Jobs[17:23,]
str(TopJobs)
```

```
## 'data.frame':    7 obs. of  13 variables:
##  $ JobNum    : chr  "1260 P 2013/09/20" "9230 OP 2014/4/11" "9182 P 2014/06/20" "7945 P 2013/
12/20" ...
##  $ JobDesc   : chr  "CHIEF CLERK PERSONNEL 1260" "CHIEF FINANCIAL OFFICER 9230" "CHIEF MANAGE
MENT ANALYST 9182" "CHIEF OF AIRPORT PLANNING 7945" ...
##  $ AppsRec   : int  39 57 143 5 13 13 14
##  $ Female    : int  35 22 78 3 2 4 1
##  $ Male      : int  4 34 54 2 11 9 13
##  $ UnknownGen: int  0 1 11 0 0 0 0
##  $ Black     : int  16 10 19 0 1 1 1
##  $ Hispanic  : int  12 7 32 3 7 2 0
##  $ Asian     : int  2 11 15 0 2 3 7
##  $ white     : int  4 21 42 2 1 4 3
##  $ AIAN      : int  0 0 0 0 0 0 1
##  $ Fil       : int  3 4 15 0 1 1 2
##  $ UnknownEth: int  2 4 20 0 1 2 0
```

Then we made another one that specified “Senior Jobs”

```
SeniorJobs <- Jobs[92:114,]
```

Then we combined these two data frames as “Chief Jobs” to see who all applied for these two categories. After wrangling this data we used the glimpse function to view all of the new changes.

```
ChiefJobs <- full_join(TopJobs, SeniorJobs)
```

```
## Joining, by = c("JobNum", "JobDesc", "AppsRec", "Female", "Male", "UnknownGen", "Black", "His
panic", "Asian", "white", "AIAN", "Fil", "UnknownEth")
```

##Exploratory Analysis Here we will explore the data by looking at the head, structure, and summary of the data.

```
head(ChiefJobs)
```

```
##              JobNum                           JobDesc AppsRec Female Male
## 1 1260 P 2013/09/20        CHIEF CLERK PERSONNEL 1260      39     35    4
## 2 9230 OP 2014/4/11      CHIEF FINANCIAL OFFICER 9230      57     22   34
## 3 9182 P 2014/06/20     CHIEF MANAGEMENT ANALYST 9182     143     78   54
## 4 7945 P 2013/12/20    CHIEF OF AIRPORT PLANNING 7945       5      3    2
## 5 7271 P 2013/11/08 CHIEF OF DRAFTING OPERATIONS 7271      13      2   11
## 6 1741 P 2013/10/25      CHIEF PERSONNEL ANALYST 1741      13      4    9
##   UnknownGen Black Hispanic Asian white AIAN Fil UnknownEth
## 1          0    16       12     2     4    0   3          2
## 2          1    10        7    11    21    0   4          4
## 3         11    19       32    15    42    0  15         20
## 4          0     0        3     0     2    0   0          0
## 5          0     1        7     2     1    0   1          1
## 6          0     1        2     3     4    0   1          2
```

```
names(ChiefJobs)
```

```
##  [1] "JobNum"     "JobDesc"    "AppsRec"    "Female"     "Male"
##  [6] "UnknownGen" "Black"      "Hispanic"   "Asian"      "white"
## [11] "AIAN"       "Fil"        "UnknownEth"
```

The mean function is used to display the averages of the different ethnic groups that applied for higher positions.

```
mean(ChiefJobs$AppsRec, na.rm=TRUE)
```

```
## [1] 129.6
```

```
mean(ChiefJobs$Female, na.rm=TRUE)
```

```
## [1] 66.06667
```

```
mean(ChiefJobs$UnknownGen, na.rm=TRUE)
```

```
## [1] 3.033333
```

```
mean(ChiefJobs$Black, na.rm=TRUE)
```

```
## [1] 29.96667
```

```
mean(ChiefJobs$Hispanic, na.rm=TRUE)
```

```
## [1] 41.5
```

```
mean(ChiefJobs$Asian, na.rm=TRUE)
```

```
## [1] 16.53333
```

```
mean(ChiefJobs$white, na.rm=TRUE)
```

```
## [1] 23.5
```

```
mean(ChiefJobs$AIAN, na.rm=TRUE)
```

```
## [1] 0.9333333
```

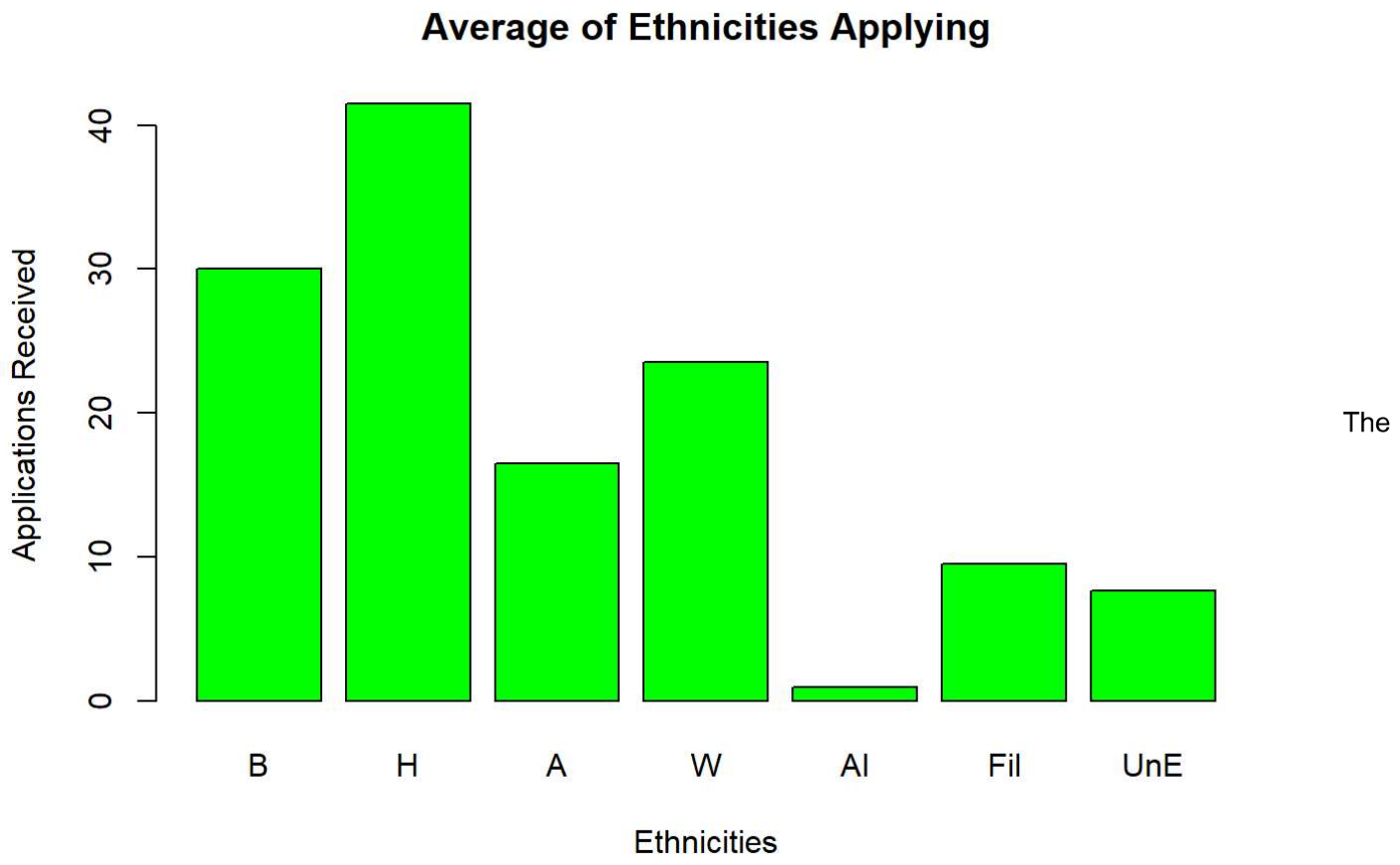
```
mean(ChiefJobs$Fil, na.rm=TRUE)
```

```
## [1] 9.533333
```

```
mean(ChiefJobs$UnknownEth, na.rm=TRUE)
```

```
## [1] 7.633333
```

```
ChiefJobbar <- c(30,41.5,16.5,23.5,0.933,9.53,7.63)
barplot(ChiefJobbar,
main = "Average of Ethnicities Applying",
xlab = "Ethnicities",
ylab = "Applications Received",
names.arg = c("B","H","A","W","AI", "Fil","UnE"),
col = "green")
```



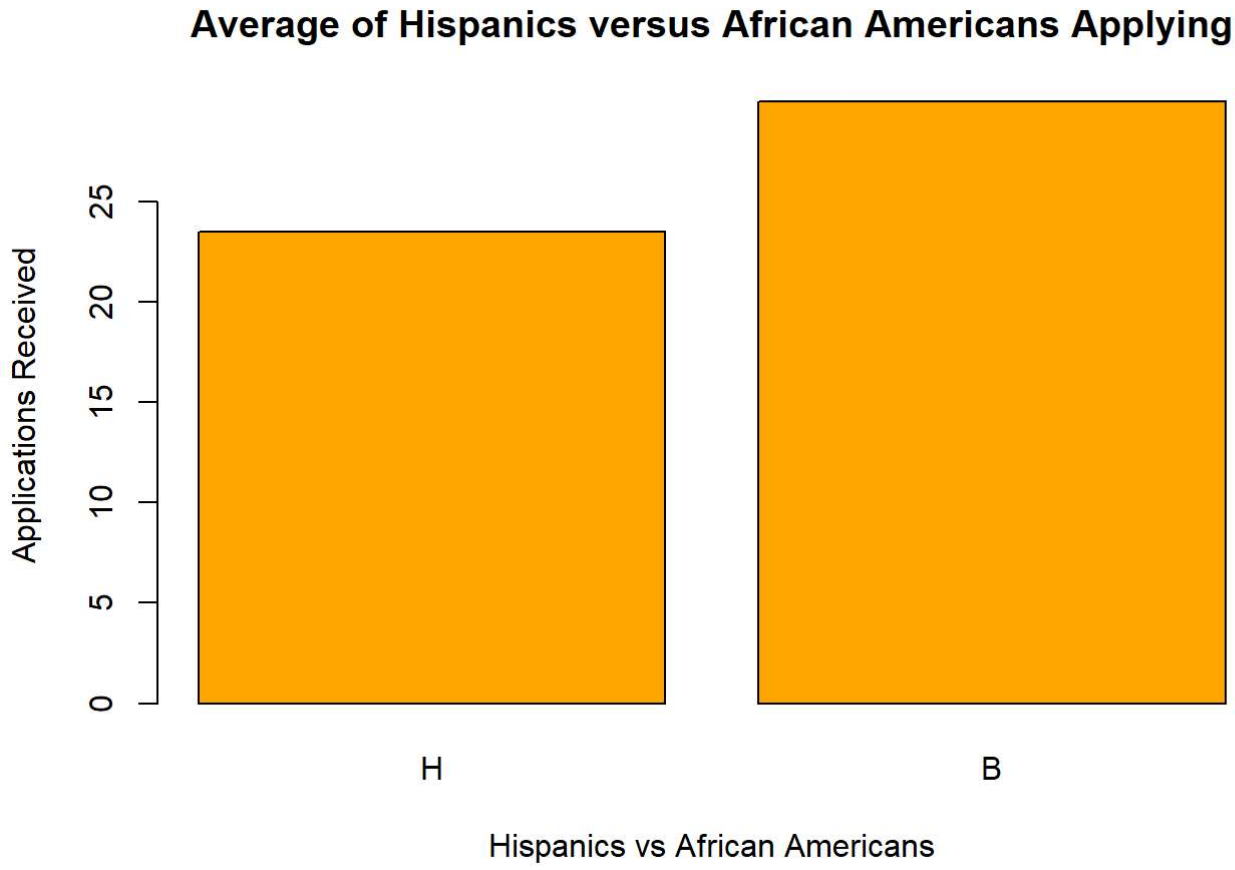
The str function shows us the characteristics of the variables and the number of observations.

```
str(ChiefJobs)
```

```
## 'data.frame':    30 obs. of  13 variables:
##  $ JobNum    : chr  "1260 P 2013/09/20" "9230 OP 2014/4/11" "9182 P 2014/06/20" "7945 P 2013/
12/20" ...
##  $ JobDesc   : chr  "CHIEF CLERK PERSONNEL 1260" "CHIEF FINANCIAL OFFICER 9230" "CHIEF MANAGE
MENT ANALYST 9182" "CHIEF OF AIRPORT PLANNING 7945" ...
##  $ AppsRec   : int  39 57 143 5 13 13 14 88 175 1824 ...
##  $ Female    : int  35 22 78 3 2 4 1 0 89 1442 ...
##  $ Male      : int  4 34 54 2 11 9 13 87 84 360 ...
##  $ UnknownGen: int  0 1 11 0 0 0 0 1 2 22 ...
##  $ Black     : int  16 10 19 0 1 1 1 7 68 569 ...
##  $ Hispanic  : int  12 7 32 3 7 2 0 22 49 705 ...
##  $ Asian     : int  2 11 15 0 2 3 7 4 9 167 ...
##  $ white     : int  4 21 42 2 1 4 3 46 30 168 ...
##  $ AIAN      : int  0 0 0 0 0 0 1 1 0 10 ...
##  $ Fil       : int  3 4 15 0 1 1 2 2 11 131 ...
##  $ UnknownEth: int  2 4 20 0 1 2 0 6 8 74 ...
```

The barplot code is used to show the difference in the averages between the whites compared to blacks that apply for higher positions. As displayed, the barplot proves the hypothesis to be correct, and also shows that blacks apply even more for these positions.

```
ChiefJobsbar <- c(23.5,29.96667)
barplot(ChiefJobsbar,
main = "Average of Hispanics versus African Americans Applying",
xlab = "Hispanics vs African Americans",
ylab = "Applications Received",
names.arg = c("H","B"),
col = "orange")
```



Linear Regression

First for linear regression, the pearson method was used to find the correlation coefficient, which is about 0.99 for the two variables in the table, proving that they are strongly correlated.

The lm() function was used to obtain the Least Squares Estimate, which shows us the intercept and slope values, then the summary() function extracts more information such as the standard error.

```
res <- cor.test(ChiefJobs$Hispanic, ChiefJobs$Black,
                method = "pearson")
res
```

```
##
##  Pearson's product-moment correlation
##
## data:  ChiefJobs$Hispanic and ChiefJobs$Black
## t = 52.737, df = 28, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9894064 0.9976472
## sample estimates:
##       cor
## 0.9950038
```

```
lit <- lm(Black ~ Hispanic, data = ChiefJobs)
lit
```

```
##
## Call:
## lm(formula = Black ~ Hispanic, data = ChiefJobs)
##
## Coefficients:
## (Intercept)     Hispanic
##     -3.6221       0.8094
```

Machine Learning

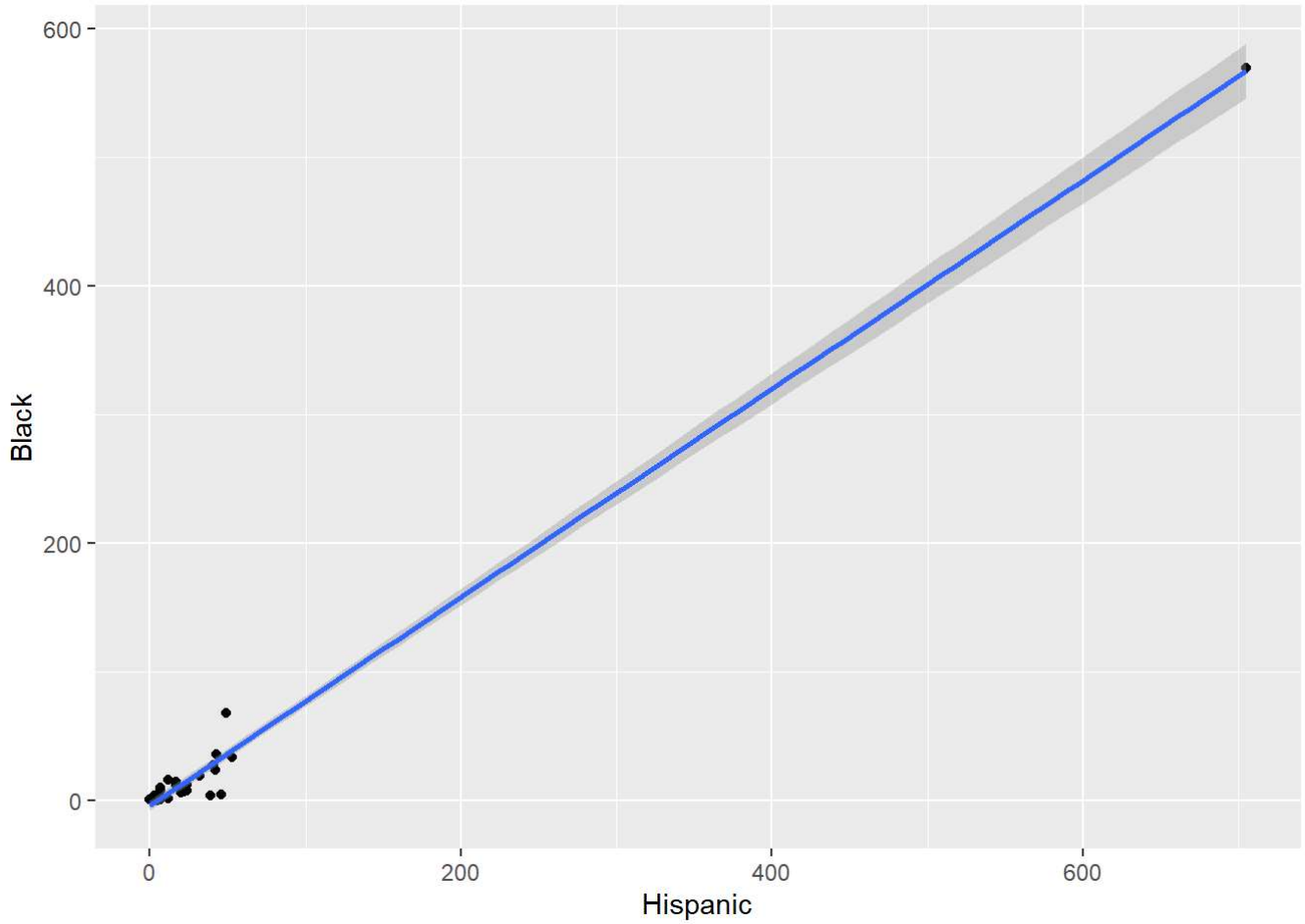
Based on our data, will we be able to see who out of African Americans and Hispanics is apply for higher ranking positions more?

```
summary(lit)
```

```
##
## Call:
## lm(formula = Black ~ Hispanic, data = ChiefJobs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.609  -4.018   1.933   4.818  31.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.62210    2.01146  -1.801   0.0825 .
## Hispanic     0.80937    0.01535  52.737   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.45 on 28 degrees of freedom
## Multiple R-squared:   0.99,  Adjusted R-squared:  0.9897
## F-statistic:  2781 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
ggplot(ChiefJobs, aes(x = Hispanic, y = Black)) +
  geom_point() +
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



This scatter plot was created using the ggplot() function with the geom_point() and geom_smooth() layers added. It shows a positive correlation between the Hispanic and Black variables with the confidence intervals of the regression line also presented. This plot shows the prediction of Y, the amount of black applications, when we know X, the amount of Hispanic applications, and it proves that African American people usually apply for higher job positions at a smaller rate or even less than Hispanic people.

Conclusions

In conclusion, we were able to find that Hispanics are applying to these higher ranking jobs more than African Americans. In fact, on average, they are apply more than any other ethnic group or race.

Future Work

For future work, we would find more data to see the success rate of Hispanics with there applications. Then,hopefully we will be able to look further into our dataset by seeing if there is any information on how many people overall are actually successful when it comes to apply for these higher ranking positions. Comparing the amount that are applying in accordance to race to the amount who are getting these higher ranking jobs.

Acknowledgements

Thank you EdX for the courses that provided us with appropriate information to help us with the project and gain the necessary knowledge needed in order to proceed with the project.

Thanks to Kaggle for supply us with this very informative data set.

We would like to acknowledge the Clark Atlanta University Provost's Summer Data Science Initiative. This project was funded in part by the National Science Foundation Grate # 1912256, the Atlanta University Data Science Initiative, and by the UNCF Career Pathways Initiative.