Final R Project

Parinyas Singh

3/3/2020

## Introduction

The World Health Organization defines ageism as “the stereotyping, prejudice, and discrimination against people based on their age.” Employers across various industries tend to show a negative attitude towards older workers, seeing them as less healthy, less skillful, and less productive than their younger counterparts. The problem of ageism has significantly affected the Tech industry, as the young and gifted developers and programmers, which led the way during the tech boom in the 1990s, now find themselves on the wrong side of age and falling prey to the problem of ageism. Although no significant research or studies conducted on the ageism problem in the tech industry and how it affects the salaries of the affected professionals. An analysis of data on this problem can shed some light on the factors that might affect it and provide some outlook on how different factors such as age, gender, sexual orientation, and several other factors affect the salaries of an IT professional.

# About the data

This dataset comes from a survey conducted by the popular website “Stack Overflow,” containing a wide range of data containing information about developers from various countries, their salaries, gender, education, the technology they work on, sexual orientation, age group, and other factors.

While the dataset contains a large amount of data, I will, however, concentrate on the data related to IT professionals of all age groups working in the United States of America. I will try to put various factors in the data set against the Annual salary to deduce how they might affect the salary.

## Loading libraries

library(readr)

## Warning: package 'readr' was built under R version 3.6.3

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

library(dplyr,warn.conflicts = FALSE)

## Warning: package 'dplyr' was built under R version 3.6.3

library(highcharter)

## Warning: package 'highcharter' was built under R version 3.6.3

library(knitr)

## Warning: package 'knitr' was built under R version 3.6.3

library(rmarkdown)

## Warning: package 'rmarkdown' was built under R version 3.6.3

library(rlang)

## Warning: package 'rlang' was built under R version 3.6.3

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.6.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

## Loading data

csvdata = read\_csv("D:/MSU/CIS663/R-Project/survey.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## Respondent = col\_double(),  
## AssessJob1 = col\_double(),  
## AssessJob2 = col\_double(),  
## AssessJob3 = col\_double(),  
## AssessJob4 = col\_double(),  
## AssessJob5 = col\_double(),  
## AssessJob6 = col\_double(),  
## AssessJob7 = col\_double(),  
## AssessJob8 = col\_double(),  
## AssessJob9 = col\_double(),  
## AssessJob10 = col\_double(),  
## AssessBenefits1 = col\_double(),  
## AssessBenefits2 = col\_double(),  
## AssessBenefits3 = col\_double(),  
## AssessBenefits4 = col\_double(),  
## AssessBenefits5 = col\_double(),  
## AssessBenefits6 = col\_double(),  
## AssessBenefits7 = col\_double(),  
## AssessBenefits8 = col\_double(),  
## AssessBenefits9 = col\_double()  
## # ... with 23 more columns  
## )

## See spec(...) for full column specifications.

## Loading data into dataframe

dfdata = data.frame(csvdata)

## Removing columns from dataframe that are not required

dfdata2 = subset(dfdata, select = -c(AssessJob1,AssessJob2,AssessJob3,AssessJob4,AssessJob5,AssessJob6,AssessJob7,AssessJob8,AssessJob9,AssessJob10,AssessBenefits1,AssessBenefits2,AssessBenefits3,AssessBenefits4,AssessBenefits5,AssessBenefits6,AssessBenefits7,AssessBenefits8,AssessBenefits9,AssessBenefits10,AssessBenefits11,JobContactPriorities1,JobContactPriorities2,JobContactPriorities3,JobContactPriorities4,JobContactPriorities5,JobEmailPriorities1,JobEmailPriorities2,JobEmailPriorities3,JobEmailPriorities4,JobEmailPriorities5,JobEmailPriorities6,JobEmailPriorities7,AdBlocker,AdBlockerDisable,AdBlockerReasons,AdsAgreeDisagree1,AdsAgreeDisagree2,AdsAgreeDisagree3,AdsActions,AdsPriorities1,AdsPriorities2,AdsPriorities3,AdsPriorities4,AdsPriorities5,AdsPriorities6,AdsPriorities7,StackOverflowRecommend,StackOverflowVisit,StackOverflowHasAccount,StackOverflowParticipate,StackOverflowJobs,StackOverflowDevStory,StackOverflowJobsRecommend,StackOverflowConsiderMember,HypotheticalTools1,HypotheticalTools2,HypotheticalTools3,HypotheticalTools4,HypotheticalTools5,Dependents,MilitaryUS,SurveyTooLong,SurveyEasy,ErgonomicDevices,Exercise,WakeTime,SkipMeals,AIDangerous,AIInteresting,AIResponsible,AIFuture,EthicsChoice,EthicsReport,EthicsResponsible,EthicalImplications,HoursOutside,Hobby,OpenSource,UpdateCV,HopeFiveYears,TimeAfterBootcamp,HackathonReasons,AgreeDisagree1,AgreeDisagree2,AgreeDisagree3,LanguageDesireNextYear,DatabaseDesireNextYear,PlatformDesireNextYear,FrameworkDesireNextYear,VersionControl,CheckInCode,Salary,JobSearchStatus,LastNewJob,CurrencySymbol,SelfTaughtTypes) )

## Filtering data for united states only

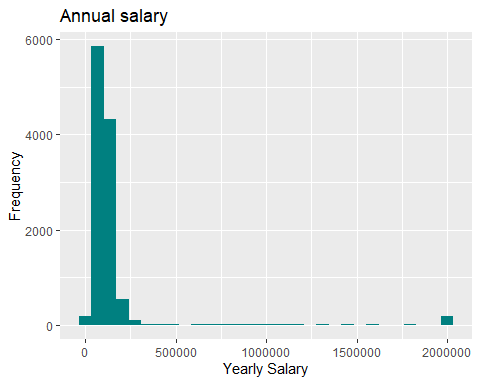
dfdata2<-dfdata2[(dfdata$Country=="United States"),]

## Graph showing yearly salary of developers with full time employment

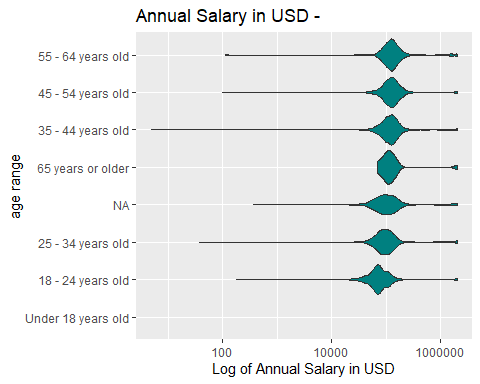
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>% ggplot() +  
 geom\_histogram(aes(ConvertedSalary),fill = "#008080") +  
 labs(x = "Yearly Salary",   
 y = "Frequency",  
 title = "Annual salary")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 4535 rows containing non-finite values (stat\_bin).

 ## Violin plot of salary against age group

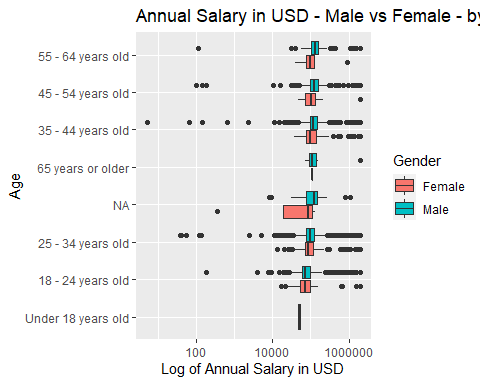
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 group\_by(Age) %>%   
 mutate(n = n()) %>%   
 summarise(med\_sal = median(ConvertedSalary, na.rm = T)) %>%   
 arrange(med\_sal) %>%   
 select(Age) %>% mutate(Age = factor(Age)) -> age\_range  
options(scipen=999)  
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 group\_by(Age) %>%   
 mutate(n = n()) %>%   
 arrange(desc(ConvertedSalary)) %>%   
 ungroup(Age) %>%   
 ggplot() +  
 geom\_violin(aes(Age,ConvertedSalary), fill = "#008080") +  
 scale\_x\_discrete(limits = age\_range$Age) +  
 coord\_flip() +  
 scale\_y\_log10() +   
 labs(x = "age range",   
 y = "Log of Annual Salary in USD",  
 title = "Annual Salary in USD - ")

 ## Annual salary plotted wrt Age in male and females

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(Gender %in% c('Male','Female')) %>%   
 group\_by(Age) %>%   
 mutate(n = n()) %>%   
 summarise(med\_sal = median(ConvertedSalary, na.rm = T)) %>%   
 arrange(med\_sal) %>%   
 select(Age) %>% mutate(Age = factor(Age)) -> countries\_salary  
options(scipen=999)  
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(Gender %in% c('Male','Female')) %>%   
 group\_by(Age) %>%   
 mutate(n = n()) %>%   
 #arrange(desc(ConvertedSalary)) %>%   
 ungroup(Age) %>%   
 ggplot() +  
 geom\_boxplot(aes(Age,ConvertedSalary, fill = Gender)) +  
 scale\_x\_discrete(limits = countries\_salary$Age) +  
 coord\_flip() +  
 #facet\_grid('Gender') +  
 scale\_y\_log10() +   
 labs(x = "Age",   
 y = "Log of Annual Salary in USD",  
 title = "Annual Salary in USD - Male vs Female - by Age")

## Warning: Transformation introduced infinite values in continuous y-axis

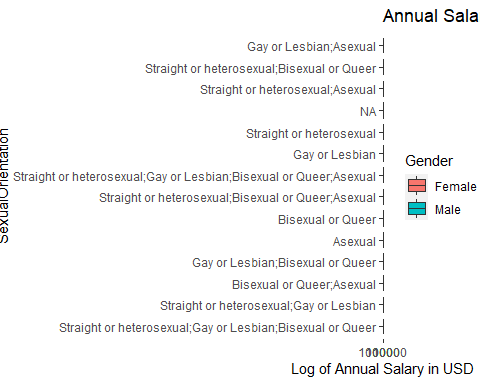
## Warning: Removed 1707 rows containing non-finite values (stat\_boxplot).

 ## Annual salary of developers wrt sexual orientation in males and females

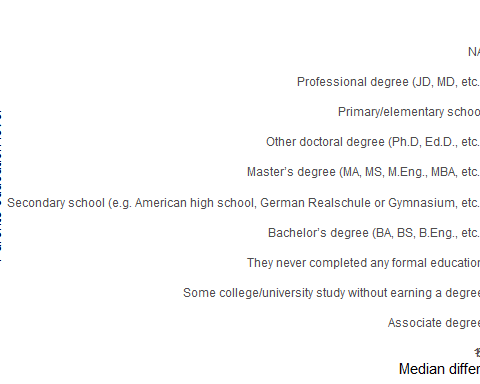
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(Gender %in% c('Male','Female')) %>%   
 group\_by(SexualOrientation) %>%   
 mutate(n = n()) %>%   
 summarise(med\_sal = median(ConvertedSalary, na.rm = T)) %>%   
 arrange(med\_sal) %>%   
 select(SexualOrientation) %>% mutate(SexualOrientation = factor(SexualOrientation)) -> avg\_salary  
options(scipen=999)  
dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(Gender %in% c('Male','Female')) %>%   
 group\_by(SexualOrientation) %>%   
 mutate(n = n()) %>%   
 #arrange(desc(ConvertedSalary)) %>%   
 ungroup(SexualOrientation) %>%   
 ggplot() +  
 geom\_boxplot(aes(SexualOrientation,ConvertedSalary, fill = Gender)) +  
 scale\_x\_discrete(limits = avg\_salary$SexualOrientation) +  
 coord\_flip() +  
 #facet\_grid('Gender') +  
 scale\_y\_log10() +   
 labs(x = "SexualOrientation",   
 y = "Log of Annual Salary in USD",  
 title = "Annual Salary in USD - Male vs Female - by SexualOrientation")

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1707 rows containing non-finite values (stat\_boxplot).

 ## Differnce in salaries of developers wrt thier parents education level

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 group\_by(EducationParents) %>%   
 mutate(n = n()) %>%   
 summarise(med\_sal = median(ConvertedSalary, na.rm = T)) %>%   
 arrange(med\_sal) %>%   
 mutate(med\_sal = med\_sal - median(med\_sal)) %>%   
 ggplot() +  
 geom\_bar(aes(reorder(EducationParents,med\_sal),med\_sal), stat = 'identity', fill = "#FFB935") +  
 coord\_flip() +   
 labs(x = 'Parents education level',  
 y = 'Median difference in salaries',  
 title = 'Difference in salaries according to parents education')

 ## Median Salary of a Developer according to Age group

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(!is.na(Age)) %>%  
 select(Age,ConvertedSalary,Gender) %>%  
 mutate(Age = str\_split(Age, pattern = ";")) %>%  
 unnest(Age) %>%  
 group\_by(Age) %>%  
 summarise(Median\_Salary = median(ConvertedSalary,na.rm = TRUE)) %>%  
 arrange(desc(Median\_Salary)) %>%  
 ungroup() %>%  
 mutate(Age = reorder(Age,Median\_Salary)) %>%  
 head(20) %>%   
 hchart('column',hcaes('Age','Median\_Salary')) %>%   
 hc\_colors(c("#829c02", "9c0228")) %>%   
 hc\_title(text = "Median Salary according to Age group of Developer")

## Warning: `parse\_quosure()` is deprecated as of rlang 0.2.0.  
## Please use `parse\_quo()` instead.  
## This warning is displayed once per session.

## Warning: `as\_data\_frame()` is deprecated as of tibble 2.0.0.  
## Please use `as\_tibble()` instead.  
## The signature and semantics have changed, see `?as\_tibble`.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

## Median Salary of Developer according to Gender

dfdata %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(!is.na(DevType)) %>%  
 filter(!is.na(Gender), Gender %in% c('Male','Female')) %>%   
 select(DevType,ConvertedSalary,Gender) %>%  
 mutate(DevType = str\_split(DevType, pattern = ";")) %>%  
 unnest(DevType) %>%  
 group\_by(DevType,Gender) %>%  
 summarise(Median\_Salary = median(ConvertedSalary,na.rm = TRUE)) %>%  
 arrange(desc(Median\_Salary)) %>%  
 ungroup() %>%  
 mutate(DevType = reorder(DevType,Median\_Salary)) %>%  
 head(20) %>%   
 hchart('column',hcaes('DevType','Median\_Salary', group = 'Gender')) %>%   
 hc\_colors(c("#7afd30", "#fe4e0d")) %>%   
 hc\_title(text = "Median Salary of Developer according to Gender")

## Male vs Female by Age groups in IT industry

dfdata2 %>%   
 filter(!is.na(Gender)) %>%  
 filter(!is.na(Age)) %>%   
 filter(Gender %in% c('Male','Female')) %>%   
 select(Gender, Age) %>%  
 mutate(Gender = str\_split(Gender, pattern = ";")) %>%  
 unnest(Gender) %>%  
 group\_by(Gender, Age) %>%  
 summarise(Count = n()) %>%  
 arrange(desc(Count)) %>%  
 ungroup() %>%  
 mutate(Gender = reorder(Gender,Count)) %>%  
 hchart('column',hcaes('Gender','Count', group = 'Age')) %>%   
 hc\_title(text = 'Male vs Female by Age groups in IT industry') %>%   
 hc\_yAxis(type = 'logarithmic')

## Median Salary of Developer according to Gender and Undergrad Major

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(!is.na(UndergradMajor)) %>%  
 filter(!is.na(UndergradMajor), Gender %in% c('Male','Female')) %>%   
 select(UndergradMajor,ConvertedSalary,Gender) %>%  
 mutate(UndergradMajor = str\_split(UndergradMajor, pattern = ";")) %>%  
 unnest(UndergradMajor) %>%  
 group\_by(UndergradMajor,Gender) %>%  
 summarise(Median\_Salary = median(ConvertedSalary,na.rm = TRUE)) %>%  
 arrange(desc(Median\_Salary)) %>%  
 ungroup() %>%  
 mutate(UndergradMajor = reorder(UndergradMajor,Median\_Salary)) %>%  
 head(20) %>%   
 hchart('column',hcaes('UndergradMajor','Median\_Salary', group = 'Gender')) %>%   
 hc\_colors(c("#7afd30", "#fe4e0d")) %>%   
 hc\_title(text = "Median Salary of Developer according to Gender and Undergrad Major")

## Median Salary of Developer according to Gender and Formal Education

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(!is.na(FormalEducation)) %>%  
 filter(!is.na(FormalEducation), Gender %in% c('Male','Female')) %>%   
 select(FormalEducation,ConvertedSalary,Gender) %>%  
 mutate(FormalEducation = str\_split(FormalEducation, pattern = ";")) %>%  
 unnest(FormalEducation) %>%  
 group\_by(FormalEducation,Gender) %>%  
 summarise(Median\_Salary = median(ConvertedSalary,na.rm = TRUE)) %>%  
 arrange(desc(Median\_Salary)) %>%  
 ungroup() %>%  
 mutate(FormalEducation = reorder(FormalEducation,Median\_Salary)) %>%  
 head(20) %>%   
 hchart('column',hcaes('FormalEducation','Median\_Salary', group = 'Gender')) %>%   
 hc\_colors(c("#7afd30", "#fe4e0d")) %>%   
 hc\_title(text = "Median Salary of Developer according to Gender and Formal Education")

## Median Salary of Developer according to Gender and Sexual Orientation

dfdata2 %>% filter(Employment %in% 'Employed full-time') %>%   
 filter(!is.na(SexualOrientation)) %>%  
 filter(!is.na(SexualOrientation), Gender %in% c('Male','Female')) %>%   
 select(SexualOrientation,ConvertedSalary,Gender) %>%  
 mutate(SexualOrientation = str\_split(SexualOrientation, pattern = ";")) %>%  
 unnest(SexualOrientation) %>%  
 group\_by(SexualOrientation,Gender) %>%  
 summarise(Median\_Salary = median(ConvertedSalary,na.rm = TRUE)) %>%  
 arrange(desc(Median\_Salary)) %>%  
 ungroup() %>%  
 mutate(SexualOrientation = reorder(SexualOrientation,Median\_Salary)) %>%  
 head(20) %>%   
 hchart('column',hcaes('SexualOrientation','Median\_Salary', group = 'Gender')) %>%   
 hc\_colors(c("#7afd30", "#fe4e0d")) %>%   
 hc\_title(text = "Median Salary of Developer according to Gender and Sexual Orientation")

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.