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

Before starting work on the data, required libraries must be added to propery load, cleanse and execute the data.

library(readr) # It provides a fast and friendly way to read rectangular data like csv, tsv, and fwf.

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

library(ggplot2) # It is used to declaratively creating graphics, based on The Grammar of Graphics.

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

library(dplyr,warn.conflicts = FALSE) # It is a grammar of data manipulation.

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

library(highcharter) #create various charts with the same style like scatter, bubble, etc.

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

library(knitr) # It an engine for dynamic report generation with R.

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

library(rmarkdown) # It is a file format for making dynamic documents with R.

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

library(rlang) # It provides tools to work with core language features of R and the tidyverse

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

library(tidyr) # It helps in creating tidy data.

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

library(tidyverse) # It is an opinionated collection of R packages designed for data science.

## 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 into R studio

After loading the required libraries, the data set is loaded into Rstudio via read\_csv command. The data set contains the survey data from popular site stack developers of IT professionals of different age groups, working on various different technologies. The data also contains the salary of each individual that took the survey. The survey contains data from professionals from all over the world.

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

After the data has been Loaded into Rstudio, it can now be converted into a data frame to use further operations in order to manipulate and cleanse the data, so it can be used further and represented graphically.

dfdata = data.frame(csvdata)

## Removing columns from Dataframe that are not required

As the data has been converted into a data frame, further, we can drop columns from the data that are not required for the prediction and visualization of the objective.

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

After unnecessary columns have been dropped from the data frame, the data can be further filtered according to the country of origin. Survey Takers from the only United States Of America are taken into account, and other countries are to be filtered out.

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

## Visualisation of Data

As the data has been cleaned and filtered, it can now be used to visualize various aspects related to IT developers such as Age, Gender, Education, Salary, etc.

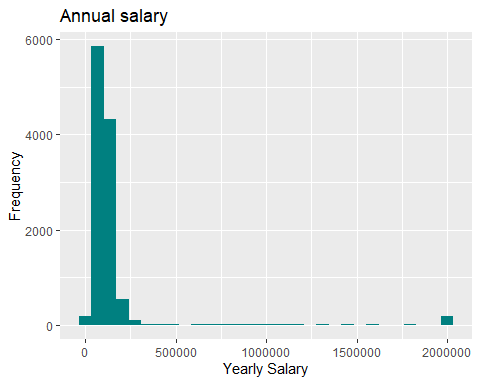
## Graph showing yearly salary of developers with Full-Time Employment

Through this histogram, we try to visualize the salary range and frequency of the salary amount that occurs within the data. As we can see in the graph, the maximum frequency of data is within USD 0 to USD 500,000, even though the data is a bit spread apart, as the data ranges from USD 0 to USD 2,000,000.

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`.

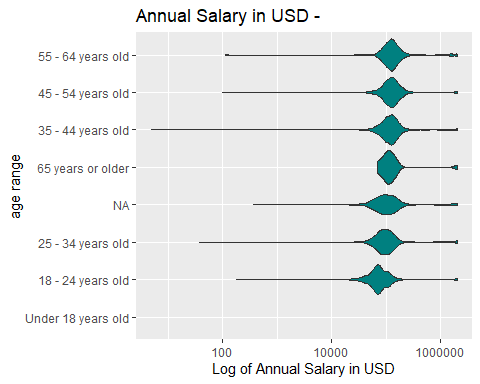
## Warning: Removed 4535 rows containing non-finite values (stat\_bin).



## Violin Plot of Salary with respect to Age Group

Through this Violin plot, we try to visualize the salary with respect to the various age groups of the survey takers. The age group of 55 to 64 year old tends to have a higher probability of having a median salary of around USD 500,000. The age groups of 18 to 24 and 25 to 34 tend to be close to USD 80,000 to USD 100,000.

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 - ")



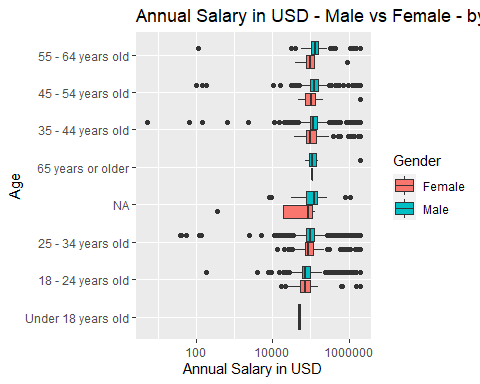
## Box-plot plotting Annual Salary with respect to Age in Males and Females

Through this Box-plot, we try to visualize the Annual Salary with respect to age in male and female developers in the IT industry. According to the data provided across all age ranges, males tend to have higher salaries the women; also, the number of males working in the IT industry far outweigh the number of 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 = "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).



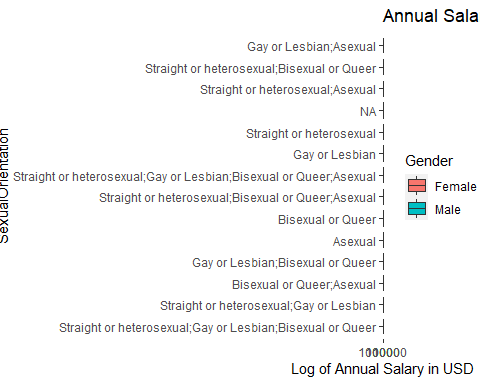
## Box-plot plotting Annual Salary with respect to sexual orientation in males and females

Through this Box-plot, we try to visualize the Annual Salary with respect to sexual orientation in male and female developers in the IT industry. According to the data provided across all age ranges, the data is highly concentrated near the straight or heterosexual line, a significant number of survey takers tend to fall in this category. However, when we look at data for bisexual or queer, the number of females tends to outnumber the number of males in the IT industry, also earning more than males compared to any other type of sexual orientation.

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

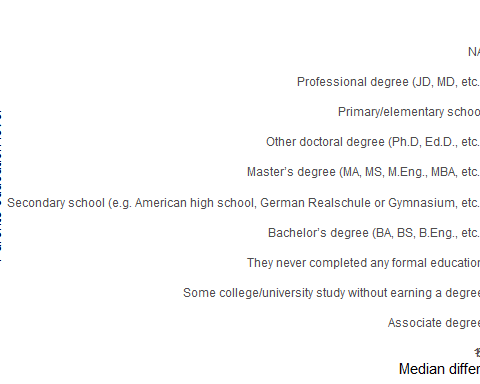
## Warning: Removed 1707 rows containing non-finite values (stat\_boxplot).



## Differnce in salaries of developers with respect to thier parents education level

Through this graph, we try to deduce if there is any relationship between how much a developer might earn with respect to the education level of their parents. This graph shows that developers tend to have higher salaries if their parents have some professional degree or have a doctorate. While on the other hand, developers tend to have lower salaries if their parents own an associate degree, have some college degree, or have no formal education.

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

Through this histogram, we plot the median salaries of developers across different age groups. When the histogram is analyzed, we notice that the age group of 45-54-year-old ager group tends to earn a median salary of 125k USD; also the age group of 55 - 64 years old falls in the same median salary group. However, developers above the age of 65 tend to have lower median salaries than people aged between 45 - 64 years old. Young people between the age group of 18 - 43 tend to earn the lowest salaries.

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 various types of Developer according to Gender

Through this chart, we plot the median salaries across various types of developers in males and females. Both males and females working as an engineering manager tend to earn the highest median salaries across all types of developing job roles. In most job roles, males tend to earn significantly more than females working in the same job roles. However, in the roles of C-Suite executives and DevOps specialists, female salaries tend to outweigh the salaries of their male counterparts. Also, to be noted that some job roles such as embedded applications, enterprise application developers, backend developer, and test developers have negligible females working.

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

Through this chart, we plot the number of males and females developers across various age groups that are working in the IT industry. In males, the highest concentration of males working in the IT industry is of the age group 25-34-year-old, with the second-highest being in the age group of 35 - 44 years old, the lowest number of males tend to fall in the age group of 65 years and older. For female developers like males, the highest age group is 25 - 34 years old; however, the number of females in the age groups of 18 - 24 years old and 35 - 44 years old is approximately the same without much difference. Also, in females, the lowest number of developers working is in the age group of 65 years and older.

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

Through this chart, we plot the median salaries of developers in males and females according to their undergrad major. The highest median salaries for males tend to be for developers with the undergrad majors in an engineering discipline and natural science; however, females with the same undergrad majors tend to earn significantly lower than their male counterparts. For females, with engineering disciple and computer science, undergrad majors tend to earn the highest median salaries. Developers with undergrad majors in fine arts, health science, and web development contain an almost negligible amount of females, and even the males with these undergrad majors tend to have some of the lowest median salaries.

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 Formal Education in Males and Females

Through this chart, we plot the median salaries of developers in males and females according to the formal education they have acquired. For males, the developers with a doctorate tend to earn the highest of median salaries, with developers with a master’s degree earning the second-highest median salaries. Male developers with an associate degree tend to earn the lowest salaries across all types of formal education. In female developers, doctoral and professional degree holders tend to earn the highest of median salaries across all types of formal educations. Also, same as the males, the females with an associate degree tend to earn the lowest of median salaries across all formal education types.

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 Developers according Hours on Computer in Males and Females

Through this chart, we plot the median salaries of developers in males and females according to the number of hours they spend on their computers. Male developers spending about 5 to 12 hours on their computers tend to earn the highest median salaries, and male developers spending about 1 to 4 hours on the computer tend to earn the lowest salaries. In the case of female developers, the developers spending about 5 to 8 hours tend to earn the maximum salaries, even more than female developers spending about 9 to 12 hours on the computer. However, in terms of the amount of time spent on a computer, females tend to earn significantly lower than their male counterparts.

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

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