Extensive EDA - Kaggle Survey 2018

Bhavani

June 19, 2019

Loading Libraries

library(readr)  
library(purrr)  
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

library(ggplot2)

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

library(stringr)

Load the data

survey1 <- read\_csv("C:Users/BhavaniC/Downloads/multipleChoiceResponses.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character()  
## )

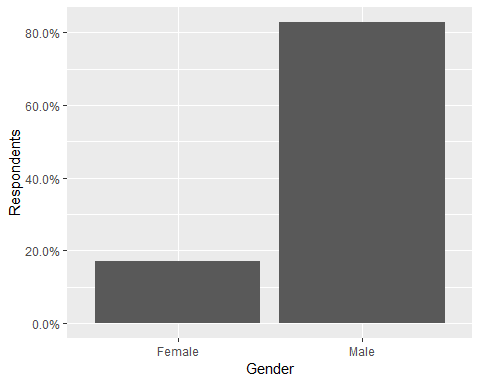
## See spec(...) for full column specifications.

survey2 <- read\_csv("C:Users/BhavaniC/Downloads/freeFormResponses.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character()  
## )  
## See spec(...) for full column specifications.

Let’s look at the Gender of the Respondents.

gender <- survey1 %>% filter(Q1 == "Male"| Q1 == "Female") %>% group\_by(Q1) %>% summarise(count = n())  
  
ggplot(data=gender, aes(x = Q1 , y = count/sum(count))) + geom\_bar(stat = "identity") + xlab("Gender") +ylab("Respondents") + scale\_y\_continuous(labels = scales::percent)

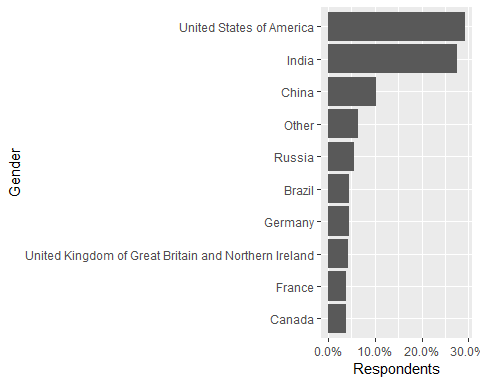


We can see from the plot above, that more than 80% of the respondents were Men. It would also be interesting to see the respondents’ home countries.

countries <- survey1 %>% group\_by(Q3) %>% summarise(count = n()) %>% arrange(desc(count)) %>% top\_n(10)

## Selecting by count

ggplot(data=countries, aes(x = reorder(Q3,count) , y = count/sum(count))) + geom\_bar(stat = "identity") + xlab("Gender") +ylab("Respondents") + coord\_flip() + scale\_y\_continuous( labels = scales::percent)



Most kaggle users seem to be from USA or India. These two countries are dominant.

Let’s also check their academic backgrounds. Are people looking to study Data Science after obtaining a graduate degree ?

survey1 %>%   
filter(Q4!="") %>%  
filter(Q1=="Male"|Q1=="Female")%>%  
ggplot(aes(Q4))+   
geom\_bar(aes(fill=Q1))+  
theme(axis.text.x = element\_text(angle = 15,hjust = 0.5, size =7))+  
theme(legend.position="none")+xlab("Academic qualification ")+  
scale\_x\_discrete(labels = function(x) str\_wrap(x, width = 8))

