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()  
## )

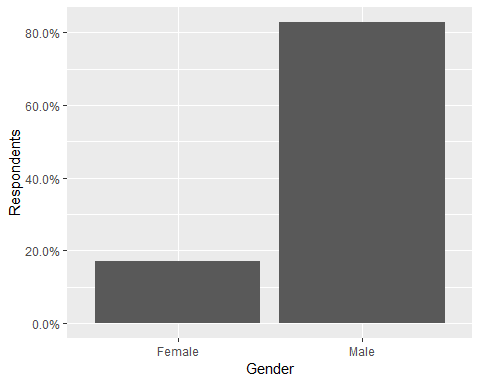
## See spec(...) for full column specifications.

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)

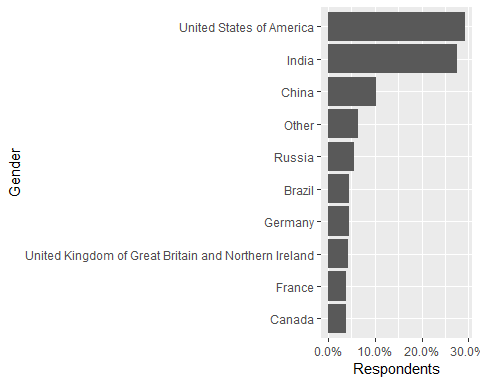


We can see from the plot above, that more than 80% of the respondents were Men. It would also be interesting to see the respondents’ home countries.

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

## Selecting by count

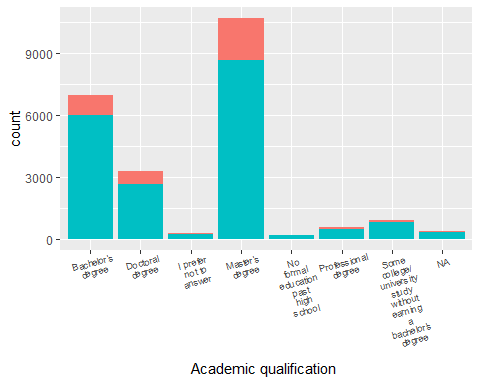
ggplot(data=countries, aes(x = reorder(Q3,count) , y = count/sum(count))) + geom\_bar(stat = "identity") + xlab("Gender") +ylab("Respondents") + coord\_flip() + scale\_y\_continuous( labels = scales::percent)



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

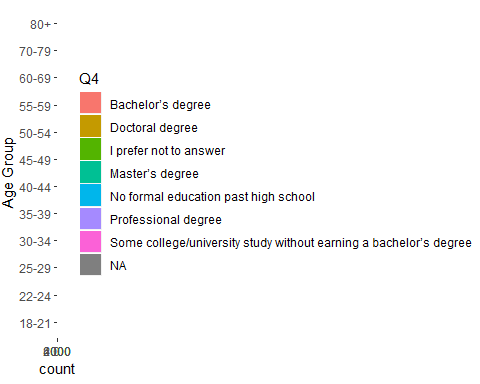
Let’s also check their academic backgrounds. Are people looking to study Data Science after obtaining a graduate degree ?

survey1 %>% filter(Q1 == "Female" | Q1 == "Male") %>% 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))

 Most Kaggle users seem to have either a Bachelor’s Degree or a Master’s Degree. It seems that people with an advanced degree are interested in pursuing a career in data science.

Let’s try to understand how old the users are and what their academic backgrounds are.

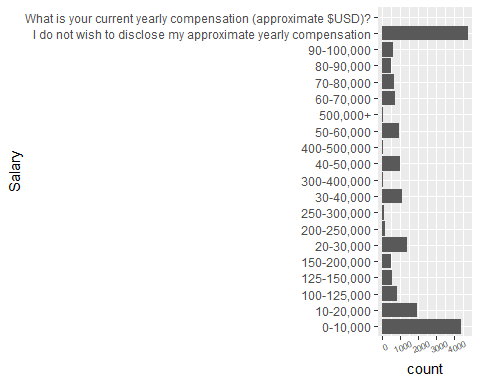
survey1 %>% filter(Q1 == "Female" | Q1 == "Male") %>% ggplot(aes(Q2)) + geom\_bar(aes(fill = Q4)) +xlab("Age Group") + coord\_flip()



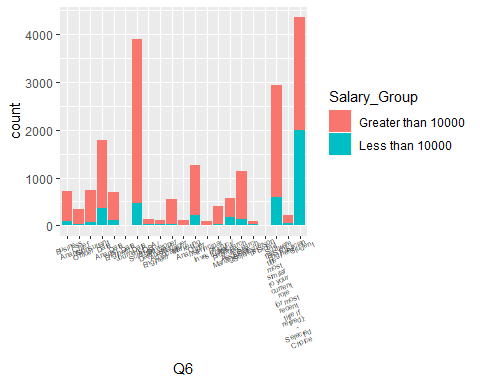
Kaggle is used by people in the 25-29 age group , followed closely by people in the 22-24 age group. Most of these users have a Master’s degree.

Let’s look at how much people are earning.

survey1 %>% filter(Q9 != " ") %>% ggplot(aes(Q9)) +geom\_bar()+ theme(axis.text.x = element\_text(angle = 20,hjust = 0.5,size = 6)) +xlab("Salary") + coord\_flip()

 Most people earn between 0 - 10000. According do this number, we can assume that most users are students. Let’s confirm this by looking at their job titles.

dat1 <- survey1 %>% filter(Q9 != " ") %>% mutate(Salary\_Group = ifelse(Q9 == "0-10,000","Less than 10000","Greater than 10000"))  
ggplot(data=dat1,aes(x = Q6),size = 20) + geom\_bar(aes(fill = Salary\_Group)) +theme(axis.text.x = element\_text(angle = 20,size = 5,hjust = 0.5)) + scale\_x\_discrete(labels = function(x) str\_wrap(x, width = 8))

 It is evident that most people who earn Less than 10000 are students.Now, let’s dig a little deeper into what the most popular Machine Learning Frameworks are.

role <- survey1 %>%  
 group\_by(Q6) %>%  
 filter(!is.na(Q6)) %>%  
 summarise(RoleCount = n())   
  
lang <- survey1 %>%  
 filter(!is.na(Q17)) %>%  
 group\_by(Q6,Q17) %>%  
 summarise(Count = n()) %>%  
 filter(Count > 1)  
  
ip <- inner\_join(role,lang,by=c("Q6"))  
  
ip <- ip %>%  
 mutate(Percentage = round(Count/RoleCount\*100,1))  
  
ip %>%  
 ggplot(aes(x = Q6, y = Q17)) +  
 geom\_tile(aes(fill = Percentage),width = 5) +   
 geom\_text(aes(label = Percentage),position = position\_dodge(width=0.9), size=2) +  
 theme\_bw() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1,size = 6)) +scale\_fill\_gradientn(colors = c("white","pink","purple"))

