Analyzing Stackoverflow developer survey

## Overview

Stackoverflow recently released the anonymized results of their 2019 annual developer survey. The goal of this project is to analyse various trends for developers worldwide.

#### Data Wrangling

The dataset has 90,000 rows. Our goal is to filter R responses so that we can drill down further and analyze the trends related to R developers.

#reading necessary libraries  
require(tidyverse)  
require(countrycode)  
require(highcharter)  
require(ggplot2)  
  
#reading the data  
data <- read\_csv("data/survey\_results\_public.csv")  
attach(data)  
  
#Let's check who all are developers  
devTable <- as.data.frame(table(data$MainBranch))

Let’s create a braplot of the profession of the respondents.

hchart(devTable,hcaes(x=Var1,y=Freq),type="column",name="Count",color="#9B6ED8") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot of occupation of participants",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

So we can notice that that most of the participants are developers.

Now let’s see how many developers have developement just as a hobby.

table(data$Hobbyist)

##   
## No Yes   
## 17626 71257

table(OpenSourcer)

## OpenSourcer  
## Less than once a month but more than once per year   
## 20561   
## Less than once per year   
## 24972   
## Never   
## 32295   
## Once a month or more often   
## 11055

Let’s check the Employment of the respondents.

round(prop.table(table(Employment))\*100,2)

## Employment  
## Employed full-time   
## 73.92   
## Employed part-time   
## 5.13   
## Independent contractor, freelancer, or self-employed   
## 9.76   
## Not employed, and not looking for work   
## 4.36   
## Not employed, but looking for work   
## 6.42   
## Retired   
## 0.41

We can notice that 74 % of the respondents were ***Full time employees***.

Now let’s check to which country the respondents belonged.

#Creating a data frame of the Country and its frequency.  
countryDf <- data %>% count(Country) %>%   
 top\_n(10)  
  
#arranging in descending order of count  
countryDf<-arrange(countryDf,desc(n))  
   
hchart(countryDf,hcaes(x=Country,y=n),type="column",name="Count") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot of Country of respondents",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

From the above plot we can notice that most of the respondents were from US,India followed by Germany.

table(Student)

## Student  
## No Yes, full-time Yes, part-time   
## 65816 15769 5429

We can notice that most of the respondents are not Students.

### Let’s see which country had most respondents as Students ?

To achieve this let’s create a Dataframe of countries with highest number of Student respondents.

StudentDf <- data %>% filter(Student=="Yes, full-time") %>%   
 count(Country) %>%   
 top\_n(30) %>%   
 arrange(desc(n))  
  
StudentDf

## # A tibble: 30 x 2  
## Country n  
## <chr> <int>  
## 1 India 2944  
## 2 United States 2053  
## 3 Germany 1197  
## 4 United Kingdom 556  
## 5 Canada 539  
## 6 France 364  
## 7 Poland 350  
## 8 Netherlands 340  
## 9 Turkey 317  
## 10 Pakistan 295  
## # … with 20 more rows

So most of the Students were from **India**,not from US.

Let’s check the Education level for the major countries like India and US.

EdlevelIndia<- data %>% filter(Country == "India") %>%   
 group\_by(EdLevel) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
#ignoring NA values  
EdlevelIndia <- na.omit(EdlevelIndia)  
   
#let's create a Barplot  
hchart(EdlevelIndia,hcaes(x=EdlevelIndia$EdLevel,y=EdlevelIndia$count),type="column",color="red",name="Count") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot Education level of Indian Respondents",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

Let’s check the results for USA.

EdlevelUS<- data %>% filter(Country == "United States") %>%   
 group\_by(EdLevel) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
#ignoring NA values  
EdlevelIndia <- na.omit(EdlevelIndia)  
  
# let's create a Barplot  
hchart(EdlevelUS,hcaes(x=EdlevelUS$EdLevel,y=EdlevelUS$count),type="column",color="Green",name="Count") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot Education level of American Respondents",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

Let’s check the results for Germany.

Edlevelger<- data %>% filter(Country == "Germany") %>%   
 group\_by(EdLevel) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
#let's create a Barplot  
hchart(Edlevelger,hcaes(x=Edlevelger$EdLevel,y=Edlevelger$count),type="column",color="purple",name="Count") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot Education level of german Respondents",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

### Let’s analyze the UnderGraduate majors of the Participants

table(UndergradMajor)

## UndergradMajor  
## A business discipline (ex. accounting, finance, marketing)   
## 1841   
## A health science (ex. nursing, pharmacy, radiology)   
## 323   
## A humanities discipline (ex. literature, history, philosophy)   
## 1571   
## A natural science (ex. biology, chemistry, physics)   
## 3232   
## A social science (ex. anthropology, psychology, political science)   
## 1352   
## Another engineering discipline (ex. civil, electrical, mechanical)   
## 6222   
## Computer science, computer engineering, or software engineering   
## 47214   
## Fine arts or performing arts (ex. graphic design, music, studio art)   
## 1233   
## I never declared a major   
## 976   
## Information systems, information technology, or system administration   
## 5253   
## Mathematics or statistics   
## 2975   
## Web development or web design   
## 3422

As expected most of them are **Computer Science Graduates followed by Another Engineering disclipine and Information Systems**.

Country and the under-graduate programmes in which the participants were enrolled.

#for India  
UGIndia<- data %>% filter(Country == "India") %>%   
 group\_by(UndergradMajor) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
UGIndia<-na.omit(UGIndia)  
  
UGIndia

## # A tibble: 12 x 2  
## UndergradMajor count  
## <chr> <int>  
## 1 Computer science, computer engineering, or software engineering 5680  
## 2 Another engineering discipline (ex. civil, electrical, mechanical) 870  
## 3 Information systems, information technology, or system administra… 556  
## 4 Web development or web design 542  
## 5 Mathematics or statistics 93  
## 6 A business discipline (ex. accounting, finance, marketing) 74  
## 7 A natural science (ex. biology, chemistry, physics) 67  
## 8 I never declared a major 43  
## 9 A health science (ex. nursing, pharmacy, radiology) 14  
## 10 A humanities discipline (ex. literature, history, philosophy) 12  
## 11 A social science (ex. anthropology, psychology, political science) 9  
## 12 Fine arts or performing arts (ex. graphic design, music, studio a… 7

#for US  
UGUS<- data %>% filter(Country == "United States") %>%   
 group\_by(UndergradMajor) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
UGUS<-na.omit(UGUS)  
  
UGUS

## # A tibble: 12 x 2  
## UndergradMajor count  
## <chr> <int>  
## 1 Computer science, computer engineering, or software engineering 10747  
## 2 Another engineering discipline (ex. civil, electrical, mechanical) 1344  
## 3 Information systems, information technology, or system administra… 1186  
## 4 A natural science (ex. biology, chemistry, physics) 961  
## 5 Mathematics or statistics 895  
## 6 A humanities discipline (ex. literature, history, philosophy) 718  
## 7 A social science (ex. anthropology, psychology, political science) 695  
## 8 Fine arts or performing arts (ex. graphic design, music, studio a… 667  
## 9 A business discipline (ex. accounting, finance, marketing) 615  
## 10 Web development or web design 493  
## 11 I never declared a major 345  
## 12 A health science (ex. nursing, pharmacy, radiology) 87

#for Germany  
UGGermany<- data %>% filter(Country == "Germany") %>%   
 group\_by(UndergradMajor) %>%   
 summarise(count=n()) %>%   
 arrange(desc(count))  
  
UGGermany<-na.omit(UGGermany)  
  
UGGermany

## # A tibble: 12 x 2  
## UndergradMajor count  
## <chr> <int>  
## 1 Computer science, computer engineering, or software engineering 2919  
## 2 Another engineering discipline (ex. civil, electrical, mechanical) 349  
## 3 A natural science (ex. biology, chemistry, physics) 266  
## 4 Information systems, information technology, or system administra… 234  
## 5 Mathematics or statistics 198  
## 6 Web development or web design 123  
## 7 A business discipline (ex. accounting, finance, marketing) 88  
## 8 A humanities discipline (ex. literature, history, philosophy) 84  
## 9 A social science (ex. anthropology, psychology, political science) 84  
## 10 I never declared a major 57  
## 11 Fine arts or performing arts (ex. graphic design, music, studio a… 43  
## 12 A health science (ex. nursing, pharmacy, radiology) 20

We see almost similar trends for all the 3 major countries.

### Let’s see what type of developers were there ?

#keeping only the respondents from top 10 countries.  
DevTypeDf<- data %>% filter(Country %in% countryDf$Country) %>%   
 select(DevType) %>%   
 group\_by(DevType) %>%   
 #creating a new column which has the count  
 summarise(Count = n()) %>%   
 filter(Count > 100) %>%   
 arrange(desc(Count))  
  
DevTypeDf<-na.omit(DevTypeDf)  
  
DevTypeDf

## # A tibble: 52 x 2  
## DevType Count  
## <chr> <int>  
## 1 Developer, full-stack 5864  
## 2 Developer, back-end 3024  
## 3 Developer, back-end;Developer, front-end;Developer, full-stack 1486  
## 4 Developer, front-end 1454  
## 5 Developer, back-end;Developer, full-stack 1324  
## 6 Developer, mobile 1292  
## 7 Student 939  
## 8 Developer, front-end;Developer, full-stack 676  
## 9 Developer, desktop or enterprise applications 666  
## 10 Developer, back-end;Developer, desktop or enterprise applications… 563  
## # … with 42 more rows

So most of the respondents from the top 10 countries were **Full stack developers** and **backend developers**.

### Let’s check if Indian,American and German respondents had what job ?

DevIndia<- data %>% filter(Country == "India") %>%   
 select(DevType) %>%   
 group\_by(DevType) %>%   
 summarise(Count=n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(10)  
  
  
DevUs<- data %>% filter(Country == "United States") %>%   
 select(DevType) %>%   
 group\_by(DevType) %>%   
 summarise(Count=n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(10)  
  
  
  
DevGer<- data %>% filter(Country == "Germany") %>%   
 select(DevType) %>%   
 group\_by(DevType) %>%   
 summarise(Count=n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(10)  
  
DevGer<- data %>% filter(Country == "Germany") %>%   
 select(DevType) %>%   
 group\_by(DevType) %>%   
 summarise(Count=n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(10)

Creating the Plot:

hchart(DevIndia,hcaes(x=DevIndia$DevType,y=DevIndia$Count),type="column") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot of top developers in India",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

hchart(DevUs,hcaes(x=DevUs$DevType,y=DevUs$Count),type="column") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot of top developers in US",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

hchart(DevGer,hcaes(x=DevGer$DevType,y=DevGer$Count),type="column") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Barplot of top developers in Germany",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

### Let’s check the Job satisfaction of the Respondents:

This is a really interesting feature to analyze. We can analyze which specific Job roles were most satisfied or in the list of top 10 countries how many respondents were satisfied with their Jobs etc.

table(JobSat)

## JobSat  
## Neither satisfied nor dissatisfied Slightly dissatisfied   
## 8720 10752   
## Slightly satisfied Very dissatisfied   
## 24207 4857   
## Very satisfied   
## 22452

Let’s create a helper function

# Drilldown(dataset,Filter\_by\_var,filter\_by\_val,group\_by\_var)  
DrillDown <- function(data,fil\_var,val,gp\_var) {  
   
 attach(data)  
 #Initializing new columns for the dataframe data  
 data["fil\_var"]<-NA  
 data["gp\_var"]<-NA  
   
 #copying the values of args to the initialized columns which would be used in the data aggregation below.  
 data$fil\_var<-fil\_var  
 data$gp\_var<-gp\_var  
 df <- data %>% filter(fil\_var==val) %>%   
 group\_by(gp\_var) %>%   
 summarise(Count=n()) %>%  
 arrange(desc(Count))  
   
 df<-na.omit(df)  
 names(df)[1] <- c("Satisfaction")  
   
 df  
  
   
   
   
}

#### Job satisfaction and Career satisfaction of Respondents from specific countries:

# !diagnostics off  
JobSatIndia <- DrillDown(data,Country,"India",JobSat)  
  
JobSatUS <- DrillDown(data,Country,"United States",JobSat)  
  
JobSatUk <- DrillDown(data,Country,"Argentina",JobSat)  
  
JobSatChina <- DrillDown(data,Country,"China",JobSat)   
  
JobSatJapan <- DrillDown(data,Country,"Japan",JobSat)   
  
  
   
#Checking the career Satisfaction of different countries  
CarSatIndia <- DrillDown(data,Country,"India",CareerSat)  
  
CarSatUS <- DrillDown(data,Country,"United States",CareerSat)   
  
CarSatUk <- DrillDown(data,Country,"United Kingdom",CareerSat)   
  
CarSatChina <- DrillDown(data,Country,"China",CareerSat)   
CarSatJapan <- DrillDown(data,Country,"Japan",CareerSat)

For almost all the countries the Job and career satisfactions were same. Most of the people were **SLightly Satisfied** .

Whereas in Country like US, most of the respondents were **Very Satisfied**.

Only **China** had most of the respondents which were **Slightly Satisfied** and **Slightly Dissatisfied** .

#### We can also analyze the top Job Factors

TOPJF <- data %>% group\_by(JobFactors) %>%   
 summarise(Count=n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(11)  
  
TOPJF <- na.omit(TOPJF)  
  
  
hchart(TOPJF,hcaes(x=TOPJF$JobFactors,y=TOPJF$Count),type="column",color="#FDE725") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Job Factors which matter the most",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

JfIndia <- DrillDown(data,Country,"India",JobFactors)  
  
JfIndia <- JfIndia %>%   
 top\_n(10)

So in India respondents mostly cared for factors such as ***‘Languages, frameworks, and other technologies I’d be working with;Office environment or company culture;Opportunities for professional development’***

Let’s see for the US respondents

JfUS <- DrillDown(data,Country,"United States",JobFactors)  
  
JfUS <- JfUS %>%   
 top\_n(10)  
  
  
JfChina <- DrillDown(data,Country,"China",JobFactors)  
  
JfChina <- JfChina %>%   
 top\_n(10)  
  
  
  
JfJapan <- DrillDown(data,Country,"Japan",JobFactors)  
  
JfJapan <- na.omit(JfJapan) %>%   
 top\_n(10)

#### Let’s check the Weekly Working hours of the Respondents

mean(na.omit(WorkWeekHrs))

## [1] 42.1272

CountryWorkHrs<- data %>% select(Country,WorkWeekHrs) %>%   
 group\_by(Country) %>%   
 summarise(Mean = mean(na.omit(WorkWeekHrs)))  
  
  
# Finding the top 20 countries with highest average weekly work hours.  
TopCountryWork<- CountryWorkHrs %>% arrange(desc(CountryWorkHrs$Mean)) %>%   
 top\_n(20)  
  
  
  
hchart(TopCountryWork,hcaes(x=TopCountryWork$Country,y=TopCountryWork$Mean),type="column",color="#ADE554") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Highest weekly work hours",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

So the average working hours were **42** hours.

We can notice that the Country where the weekly work hours were highest is **Finland** followed by **Norway**.

### Let’s see the challenges that the Respondents face at work ?

WorkChallTop <- data %>% group\_by(WorkChallenge) %>%   
 select(WorkChallenge) %>%   
 summarise(Count = n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(15)  
  
  
WorkChallTop <- na.omit(WorkChallTop)  
  
hchart(WorkChallTop,hcaes(x=WorkChallTop$WorkChallenge,y=WorkChallTop$Count),type="column",color="#FEA554") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Top Work challenges",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

The top Work challenges included things like :

1. Being tasked with non-development work.
2. Not enough people for the workload.
3. Distracting work environment.
4. Meetings. 5)Non-work commitments (parenting, school work, hobbies, etc.)
5. Toxic Work environment.

### The top languages the respondents have worked with ?

TopLang <- data %>% group\_by(LanguageWorkedWith) %>%   
 summarise(Count = n()) %>%   
 arrange(desc(Count)) %>%   
 top\_n(20)  
  
TopLang<-na.omit(TopLang)  
  
  
hchart(TopLang,hcaes(x=TopLang$LanguageWorkedWith,y=TopLang$Count),type="column",color="#EAF253") %>%   
 hc\_exporting(enabled = TRUE) %>%  
 hc\_title(text="Top Programming languages used By developers",align="center") %>%  
 hc\_add\_theme(hc\_theme\_elementary())

So we can notice that the most used languages are **HTML/CSS;JavaScript;PHP;SQL;C#**.