Project #2 (Visualizing Kiggle 2020 Survey Data)

Ivan Bayingana

1/31/2022

```
library(scales)
library(data.table)
library(ggplot2)
library(sqldf)
```

```
df <- read.csv("kaggle_survey_2020_responses.csv",na.strings=c("","NA"))
```

Cleaning

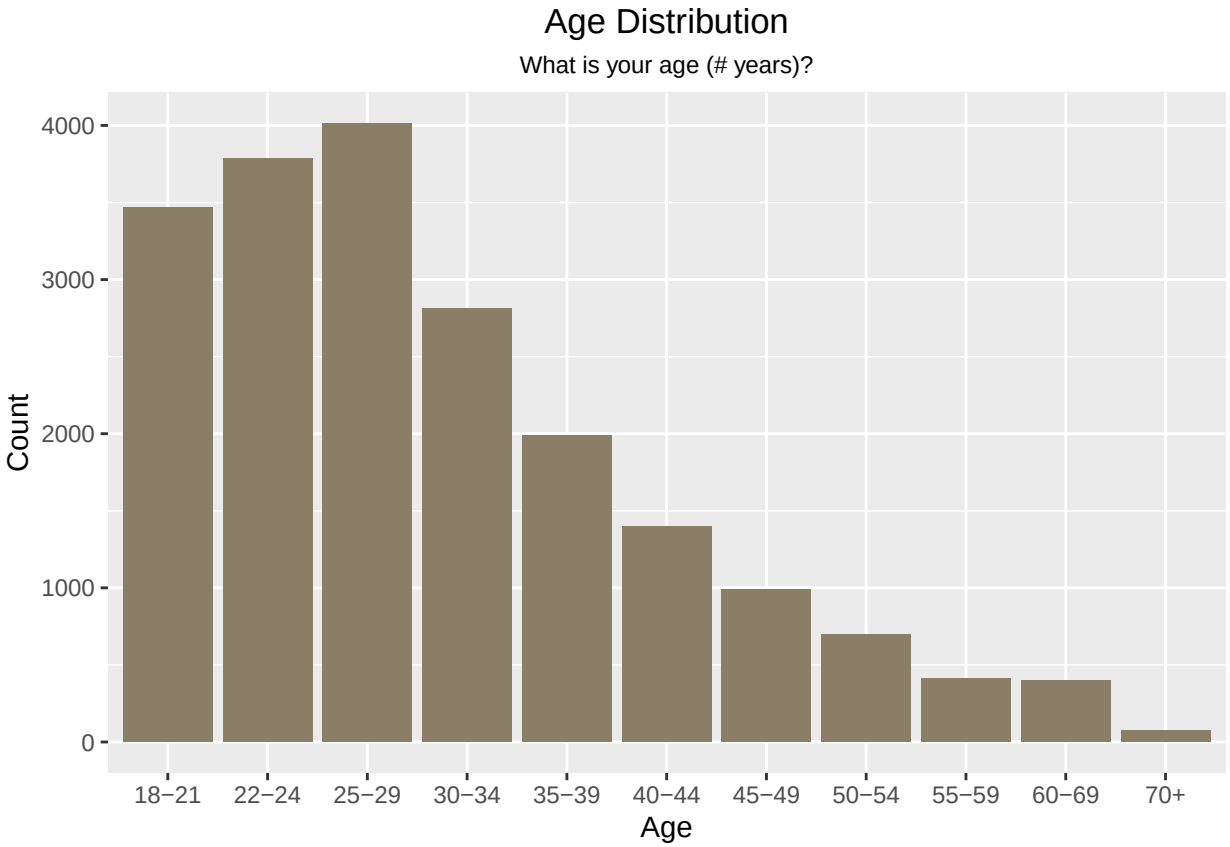
```
# the first row contains the questions, let's assign it to its own variable since...
#...it's not part of the responses
questions = df[1,]
df = df[-1,]
questions[1,2]
```

```
## [1] "What is your age (# years)?"
```

Visuals

1. Age distrubution

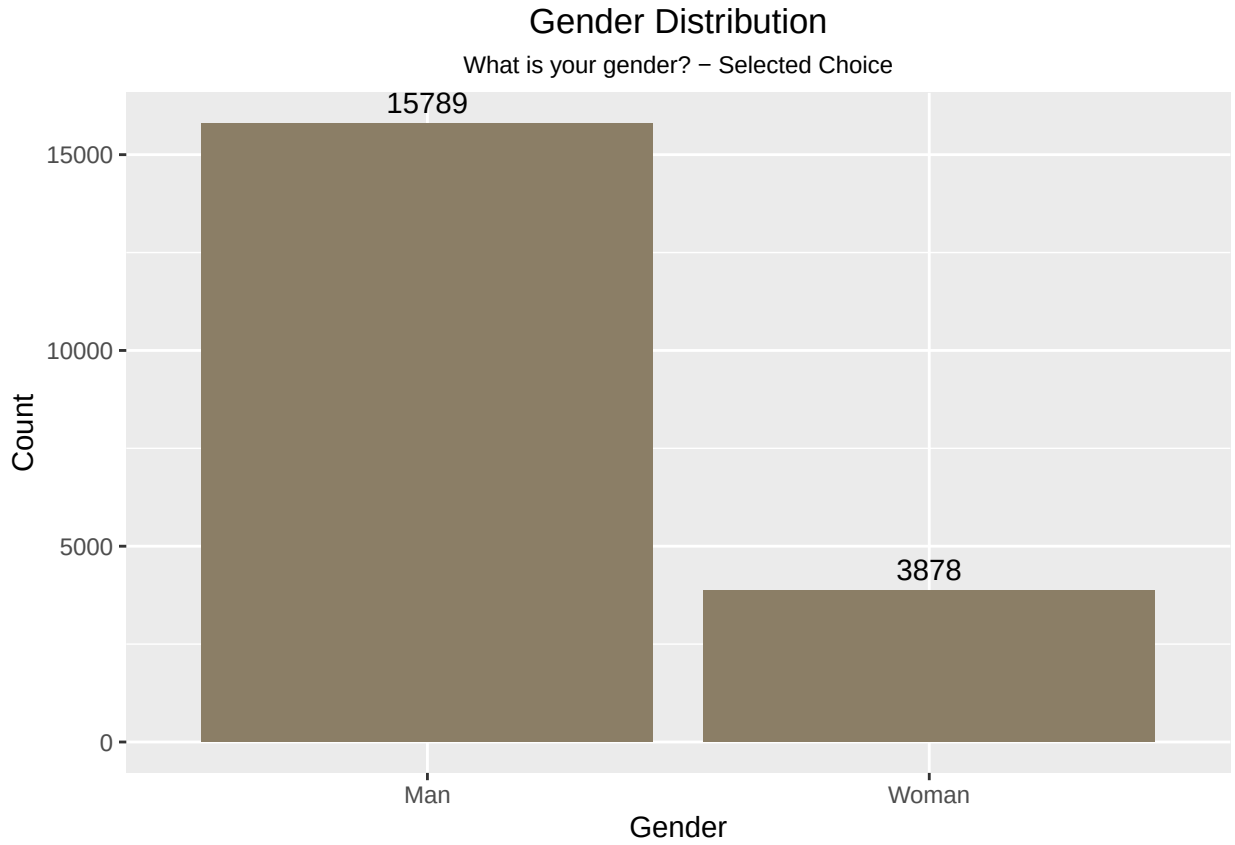
```
ggplot(df, aes(Q1)) +
  geom_bar(stat = "count",fill='wheat4')+
  labs(x='Age',y='Count',title = 'Age Distribution',subtitle=questions[1,2]) +
  theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))
```



It appears most people who filled out this survey are between the age of 22 and 29. The older people are the less likely they are of filling out this survey.

2. Gender

```
ggplot(df[df$Q2=='Man' | df$Q2=='Woman',], aes(Q2)) +
geom_bar(fill='wheat4')+
labs(x='Gender',y='Count',title = 'Gender Distribution',subtitle=questions[1,3]) +
theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
  geom_text(stat = "count",aes(label=after_stat(count)), vjust = -0.5)
```



Looks like more men filled out this survey. Women account for less than 5000 survey respondents compare that to 15000 men, you can see that there is a big difference.

3. Country

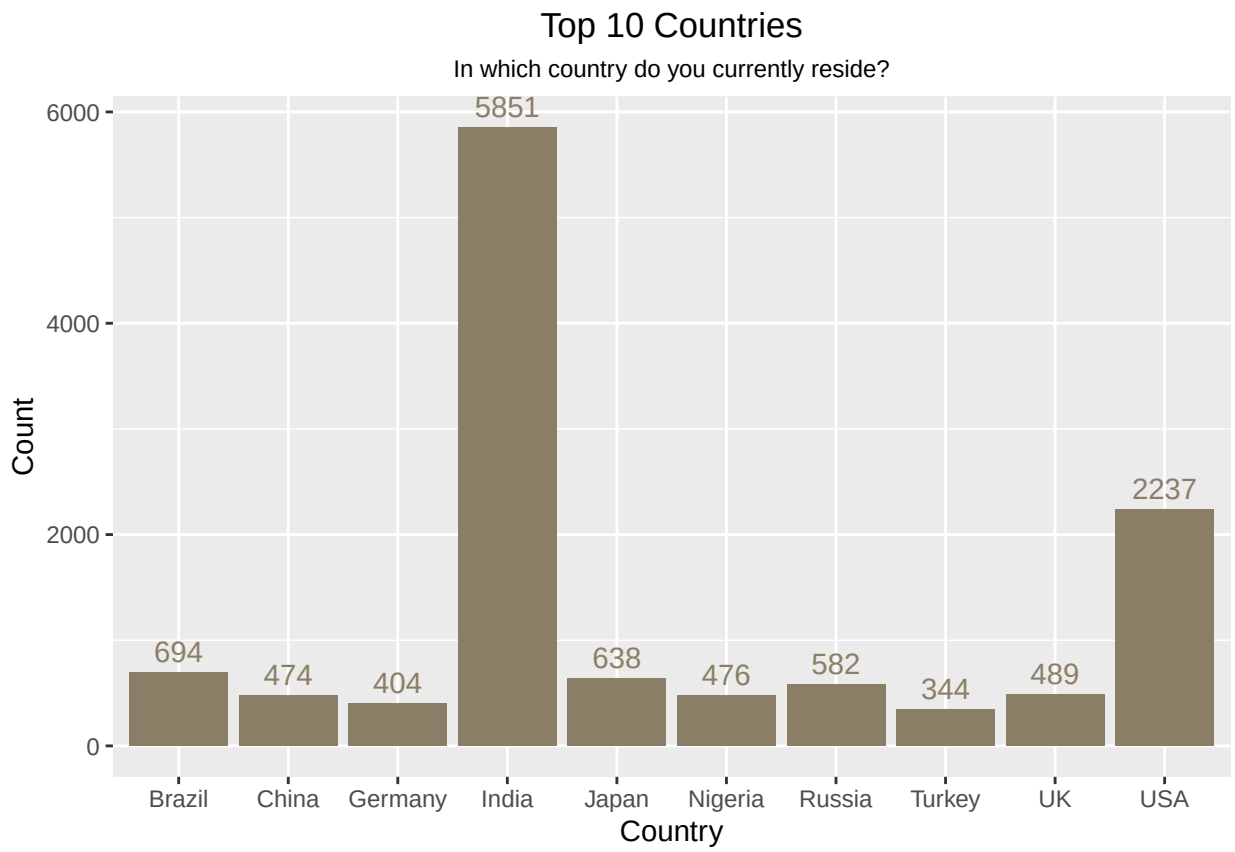
Plotting all the countries would result in an overcrowded graph, I'm only going to plot the top 10 countries instead.

```
df_country = sqldf("select Q3 as Country, count(Q3) as Count
                   from df
                   where Q3 != 'Other'
                   group by Q3
                   order by Count DESC
                   limit 10 ")


# I abbreviated the names because they were too long.
df_country[2,1] = "USA"
df_country[6,1] = "UK"
df_country
```

```
##    Country Count
## 1    India  5851
## 2      USA  2237
## 3   Brazil   694
```

```
## 4    Japan   638
## 5   Russia   582
## 6       UK   489
## 7  Nigeria   476
## 8    China   474
## 9  Germany   404
## 10  Turkey   344
```

```
ggplot(df_country, aes(x=Country, y=Count)) +
geom_bar(stat = "summary",fun=sum)+
stat_summary( fun=sum,geom="bar",fill='wheat4')+
labs(x='Country',y='Count',title = 'Top 10 Countries',subtitle=questions[1,4])+
  theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
geom_text(aes(label = Count), vjust = -0.5,color='wheat4')
```

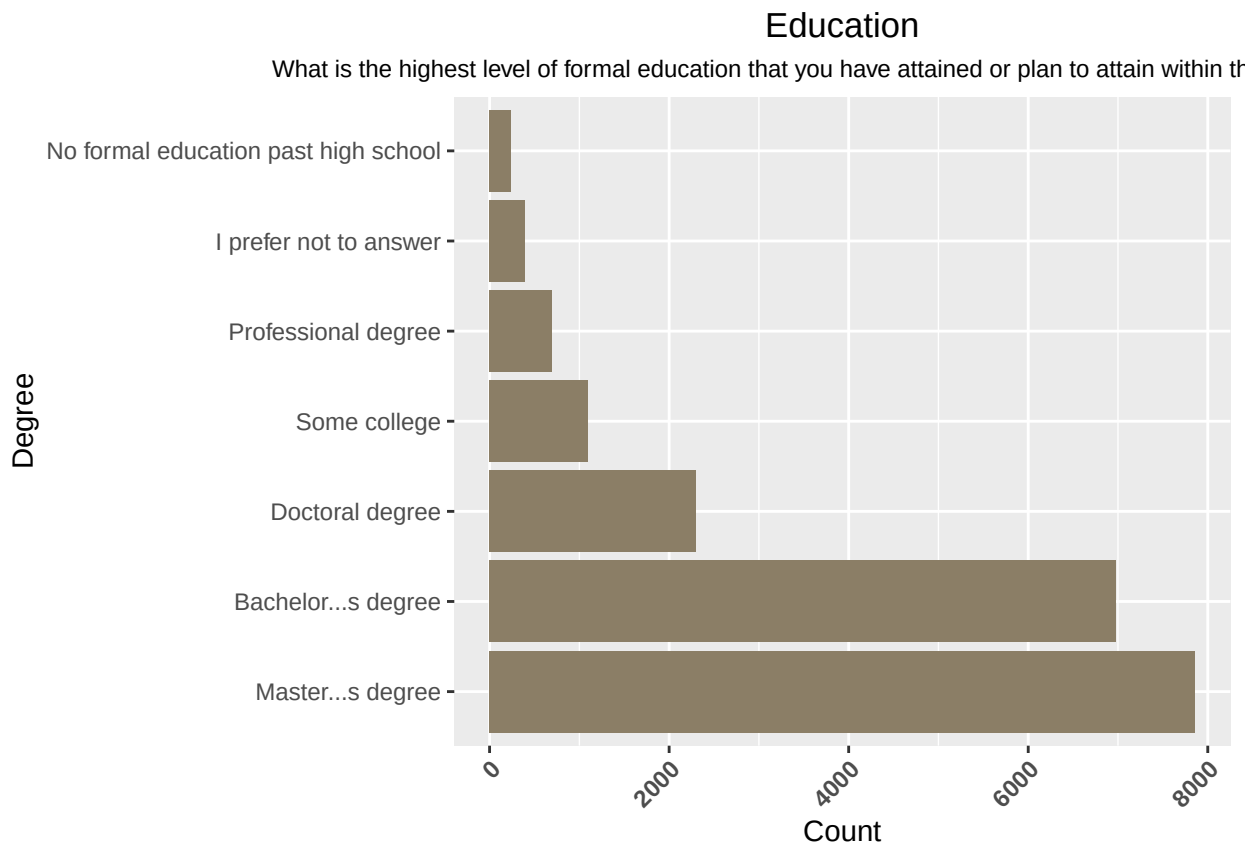


India has by far the most survey respondents followed by the USA. All other countries have less than 700 survey respondents .

4. Education

```
library(stringr)
```

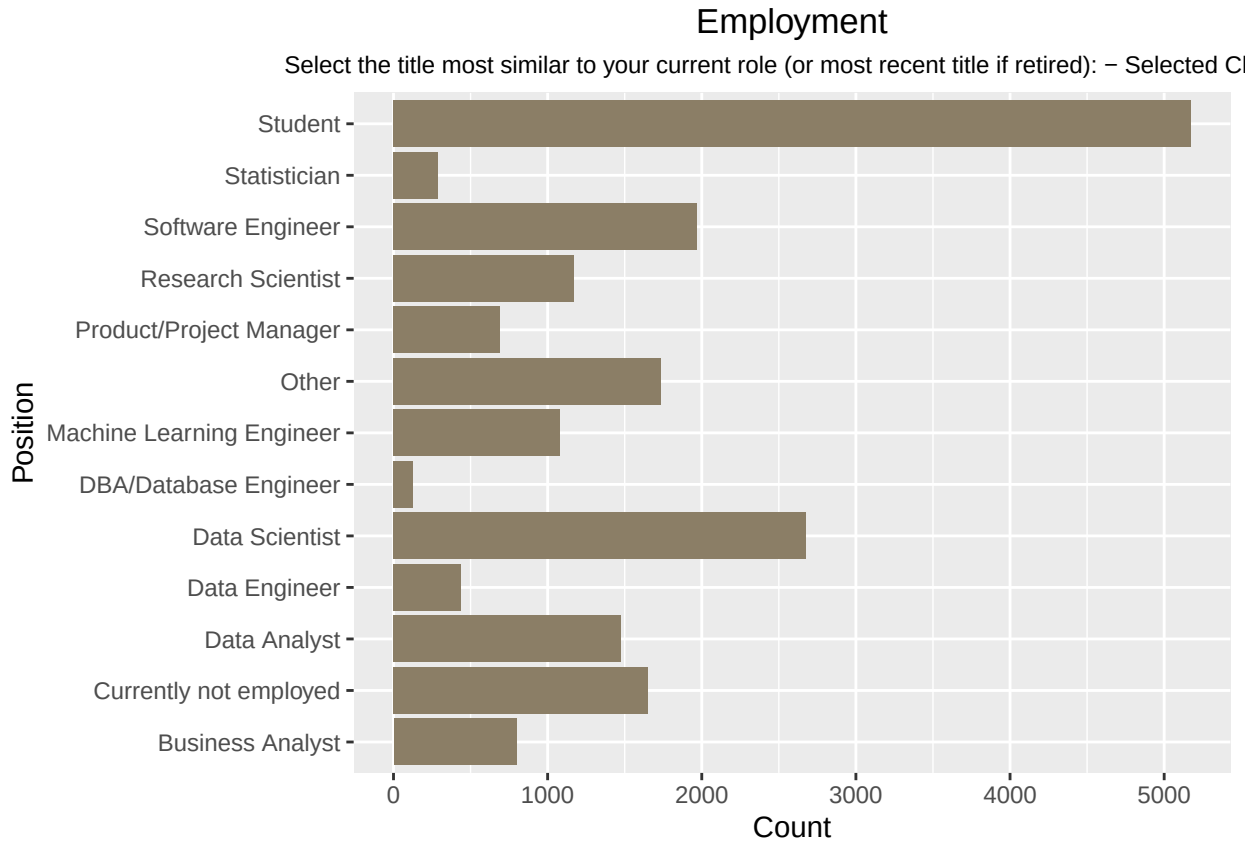
```
ggplot(df, aes(x=reorder(Q4,Q4,function(x)-length(x)))) +
geom_bar(stat = "count",fill='wheat4')+
labs(x='Degree',y='Count',title = 'Education',subtitle=questions[1,5]) +
theme(plot.title=element_text(hjust=0.5),
      plot.subtitle=element_text(hjust=0.5,size=9),
      axis.text.x = element_text(face = "bold",angle =45,hjust = 1))+
scale_x_discrete(na.translate = FALSE)+
coord_flip()
```



Most respondents have at least a Bachelor or Masters degree, very few have no formal education past high school.

5. Employment

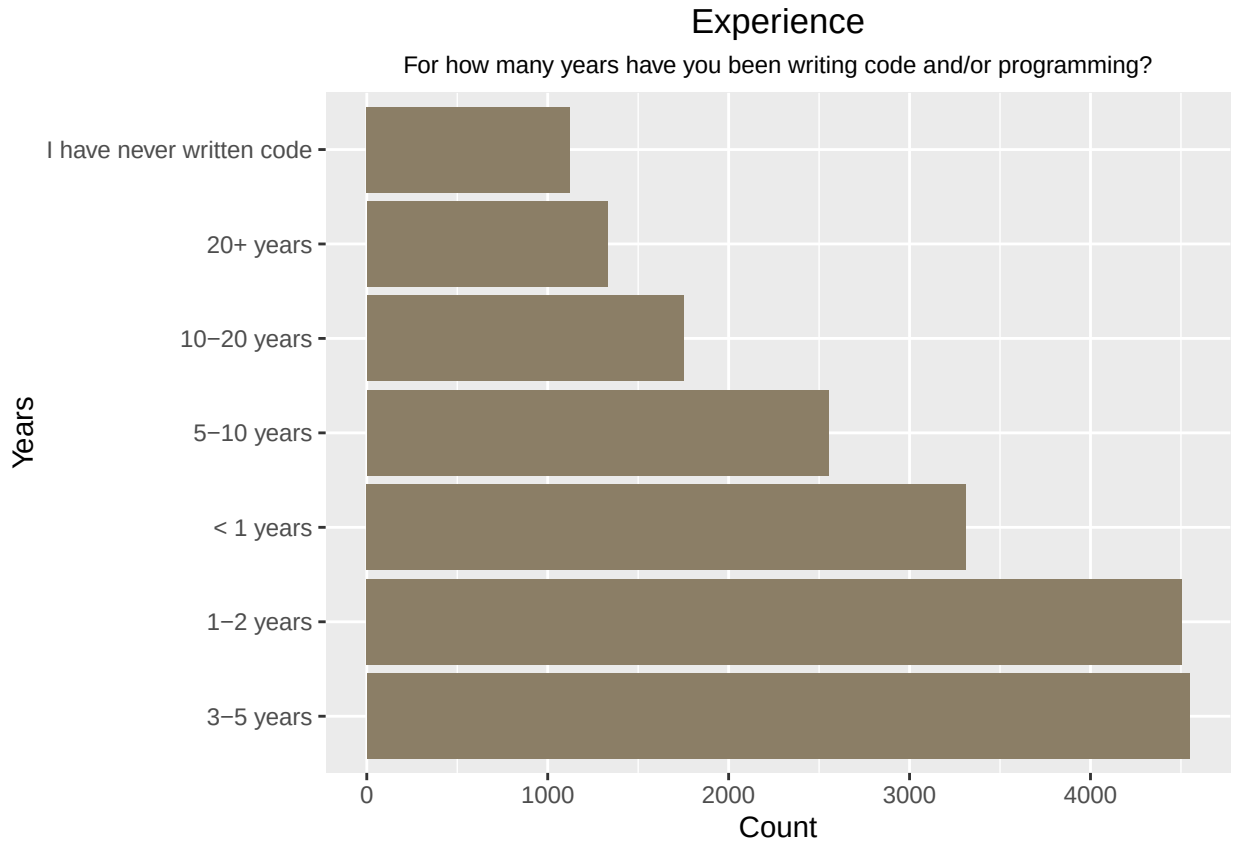
```
ggplot(df, aes(Q5)) + geom_bar(stat = "count",fill='wheat4')+
labs(x='Position',y='Count',title = 'Employment',subtitle=questions[1,6]) +
  theme(plot.title=element_text(hjust=0.5),
        plot.subtitle=element_text(hjust=0.5,size=9))+
scale_x_discrete(na.translate = FALSE)+
coord_flip()
```



We see that most respondents are students which make sense considering the age distribution graph.

6.Experience

```
ggplot(df, aes(x=reorder(Q6,Q6,function(x)-length(x)))) +
geom_bar(stat = "count",fill='wheat4')+
labs(x='Years',y='Count',title = 'Experience',subtitle=questions[1,7])+
theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
scale_x_discrete(na.translate = FALSE)+
coord_flip()
```



Here the graph shows most people have less than 5 years of experience, again this makes sense considering the age distribution graph shows the majority of respondents are between the age of 22 and 29.

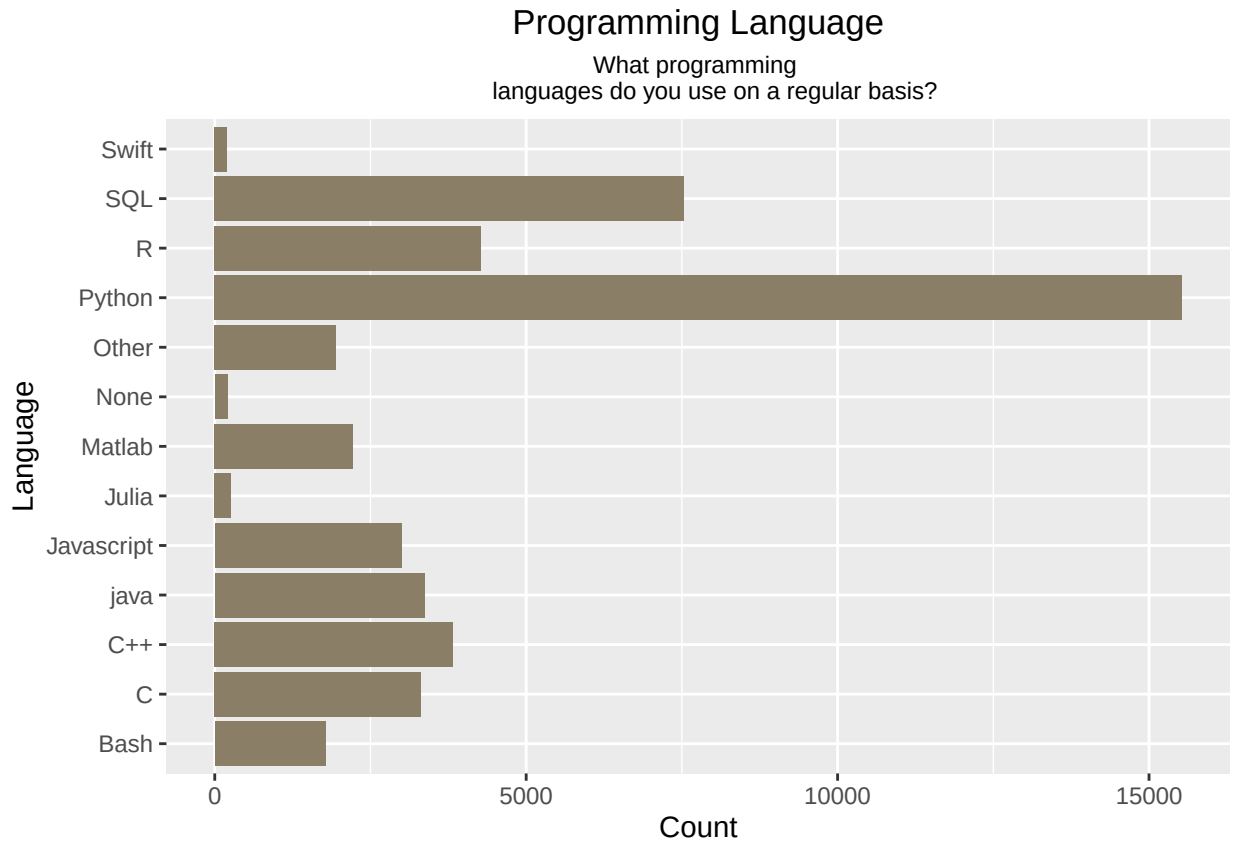
7. programming languages

```
Program_lang = sqldf("select count(Q7_Part_1) as Python,
                     count(Q7_Part_2) as R,
                     count(Q7_Part_3) as SQL,
                     count(Q7_Part_4) as C,
                     count(Q7_Part_5) as 'C++',
                     count(Q7_Part_6) as java,
                     count(Q7_Part_7) as Javascript,
                     count(Q7_Part_8) as Julia,
                     count(Q7_Part_9) as Swift,
                     count(Q7_Part_10) as Bash,
                     count(Q7_Part_11) as Matlab,
                     count(Q7_Part_12) as None,
                     count(Q7_OTHER) as Other from df")

programming_languages = transpose(Program_lang,keep.names="Language")

programming_languages
```

```
##       Language    V1
```

```
## 1      Python 15530
## 2           R  4277
## 3         SQL  7535
## 4           C  3315
## 5         C++  3827
## 6        java  3367
## 7  Javascript  2995
## 8       Julia   262
## 9       Swift   198
## 10       Bash  1776
## 11     Matlab  2217
## 12       None   206
## 13      Other  1945
```

```
ggplot(programming_languages, aes(x=Language, y=V1)) +
  geom_bar(stat = "summary",fun=sum)+
  stat_summary( fun=sum,geom="bar",fill='wheat4')+

labs(x='Language',y='Count',title = 'Programming Language',subtitle='What programming
     languages do you use on a regular basis?')+
theme(plot.title=element_text(hjust=0.5),
        plot.subtitle=element_text(hjust=0.5,size=9))+
coord_flip()
```

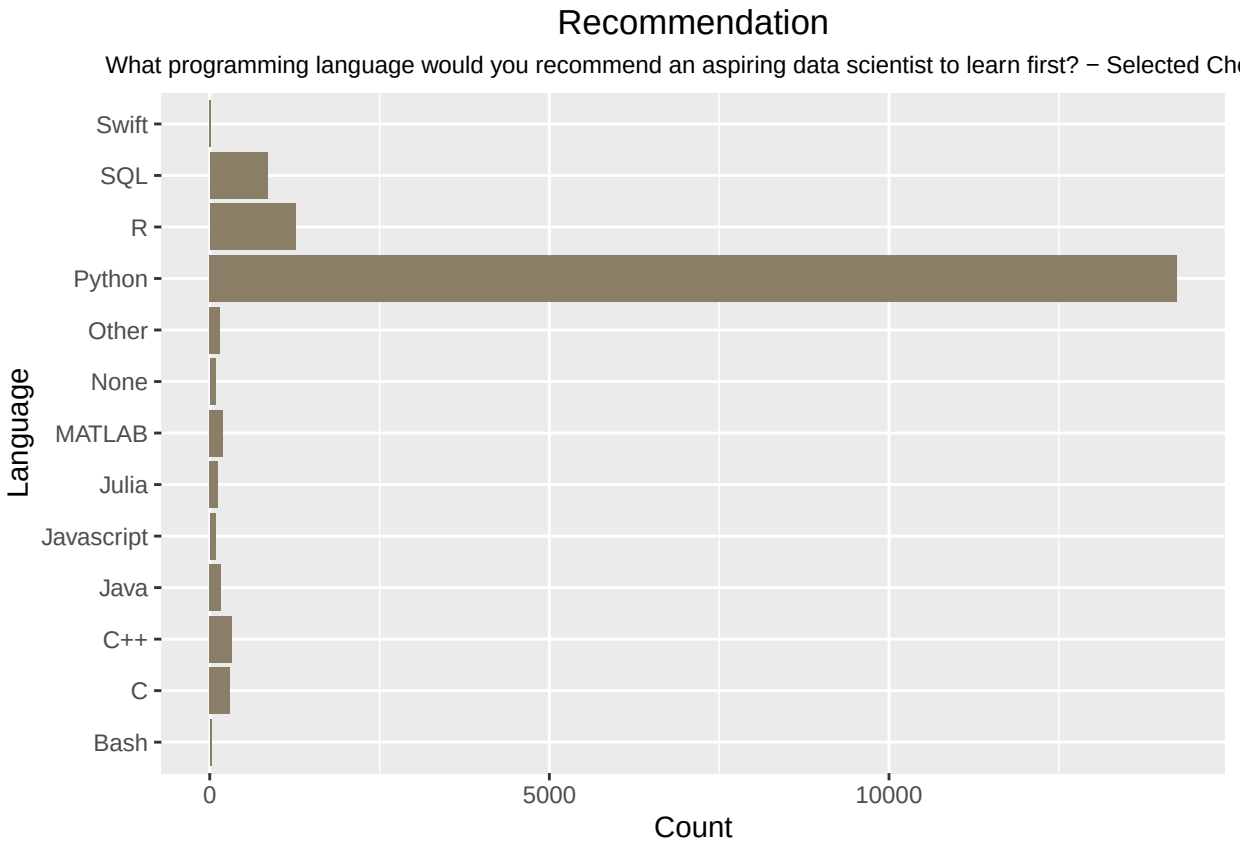


Python is by far the most used programming language compared to everything else available, SQL comes in second place followed by R. It will be interesting to see if some of the new programming language like Julia

can catch up to the popularity of Python.

8. Recommendation

```
ggplot(df, aes(Q8)) + geom_bar(stat = "count",fill='wheat4')+
  labs(x='Language',y='Count',title = 'Recommendation',subtitle=questions[1,21])+
  theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
  scale_x_discrete(na.translate = FALSE)+
  coord_flip()
```



The popularity of python is undeniable here, almost all people in this survey recommend aspiring data scientist to learn python first.

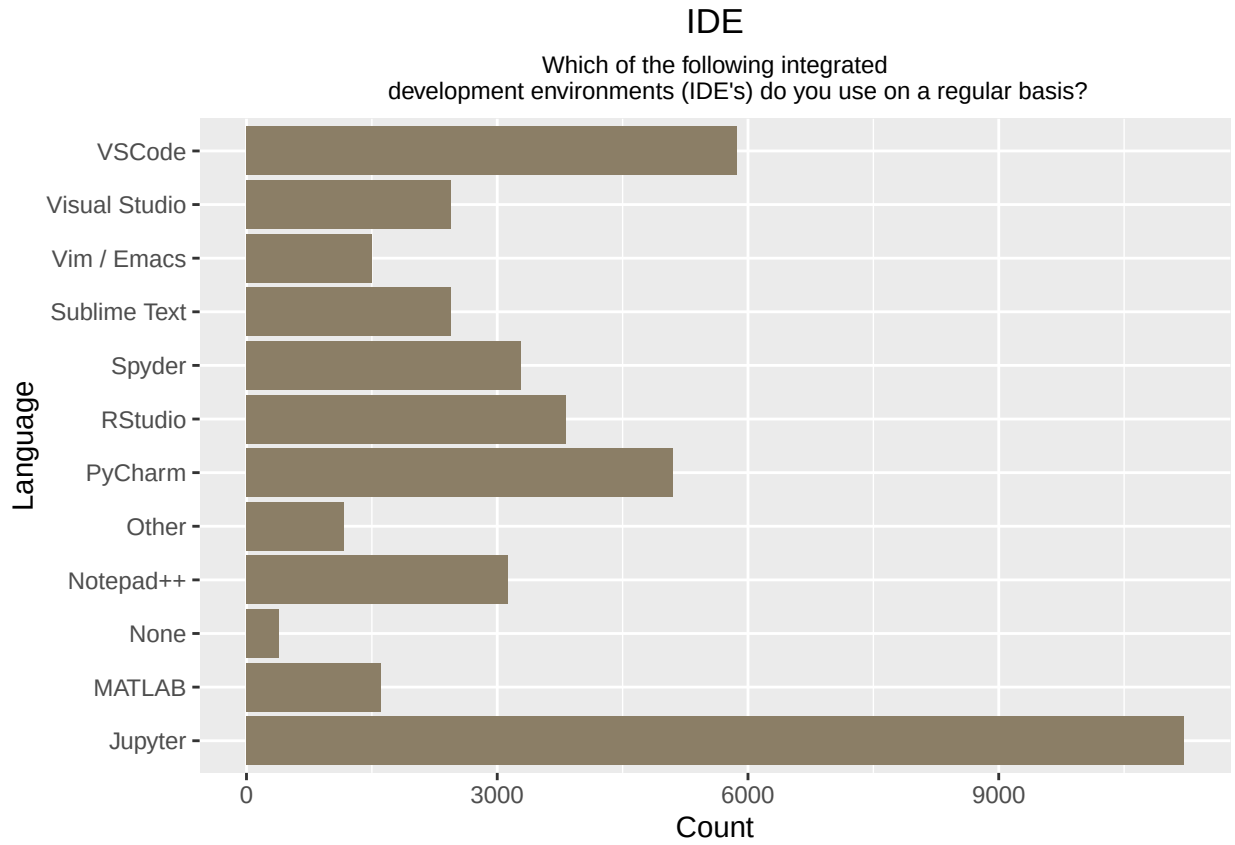
9. Environment

```
query = sqldf("select count(Q9_Part_1) as Jupyter,
              count(Q9_Part_2) as RStudio ,
              count(Q9_Part_3) as 'Visual Studio',
              count(Q9_Part_4) as VSCode,
              count(Q9_Part_5) as 'PyCharm',
              count(Q9_Part_6) as Spyder,
              count(Q9_Part_7) as 'Notepad++'  ,
```

```
              count(Q9_Part_8) as 'Sublime Text'  ,
              count(Q9_Part_9) as  'Vim / Emacs' ,
              count(Q9_Part_10) as 'MATLAB' ,
              count(Q9_Part_11) as None,
              count(Q9_OTHER) as Other from df")

env = transpose(query,keep.names ='Environment')
env
```

```
##      Environment    V1
## 1        Jupyter 11211
## 2        RStudio  3826
## 3  Visual Studio  2445
## 4         VSCode  5873
## 5        PyCharm  5099
## 6         Spyder  3290
## 7      Notepad++  3132
## 8   Sublime Text  2452
## 9    Vim / Emacs  1502
## 10        MATLAB  1604
## 11          None   386
## 12         Other  1162
```

```
ggplot(env, aes(x=Environment, y=V1)) +
  geom_bar(stat = "summary",fun=sum)+
  stat_summary( fun=sum,geom="bar",fill='wheat4')+

  labs(x='Language',y='Count',title = 'IDE',subtitle="Which of the following integrated
       development environments (IDE's) do you use on a regular basis?")+
  theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
  coord_flip()
```



It looks like most people use Jupyter as their IDE on a regular basis.

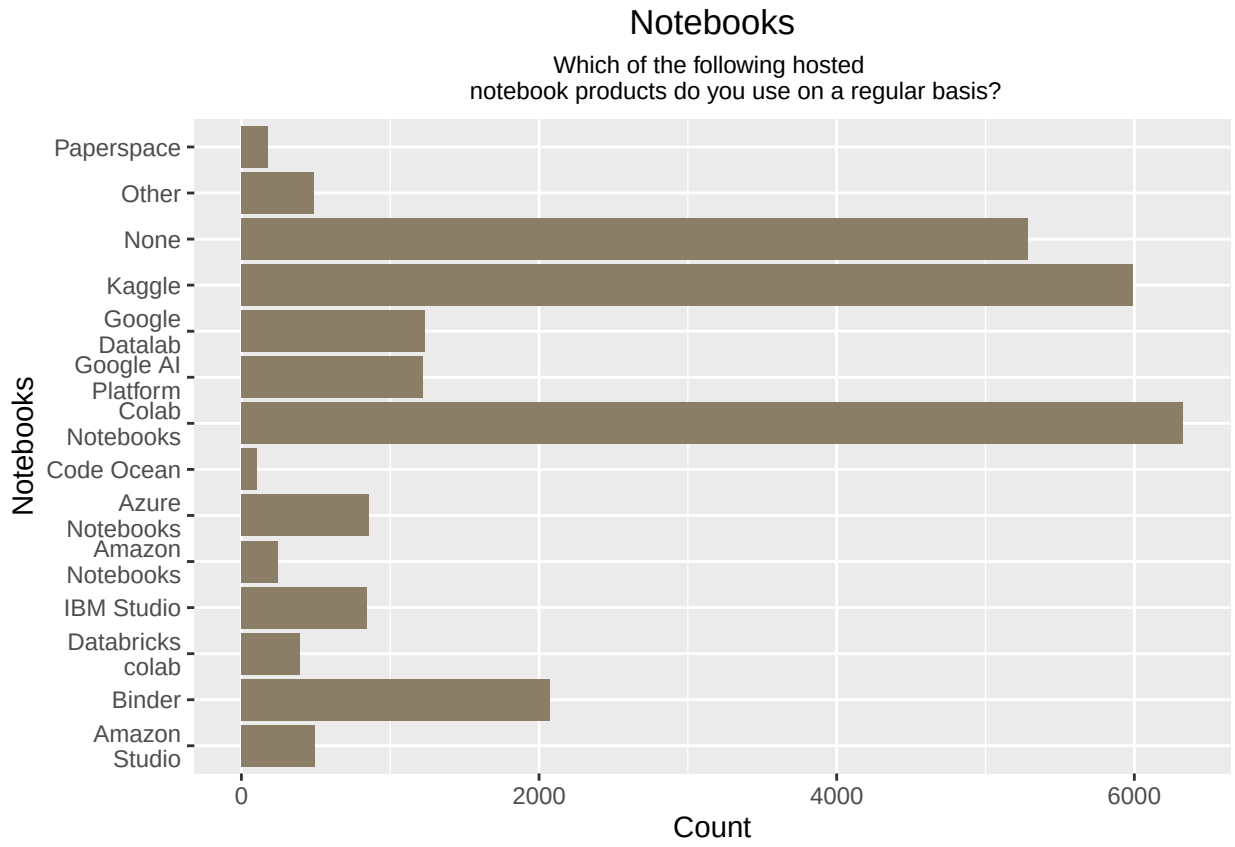
10. Notebooks

```
query = sqldf("select count(Q10_Part_1) as  Kaggle ,
              count(Q10_Part_2) as 'Colab Notebooks' ,
              count(Q10_Part_3) as 'Azure Notebooks',
              count(Q10_Part_4) as  'Paperspace'
              ,count(Q10_Part_5) as ' Binder',
              count(Q10_Part_6) as 'Code Ocean',
              count(Q10_Part_7) as ' IBM Studio'
              ,count(Q10_Part_8) as ' Amazon Studio'  ,
              count(Q10_Part_9) as  'Amazon Notebooks' ,
              count(Q10_Part_10) as 'Google AI Platform' ,
              count(Q10_Part_11) as 'Google Datalab',
              count(Q10_Part_12) as ' Databricks colab',
              count(Q10_Part_13) as None,
              count(Q10_OTHER) as Other from df")

notebooks = transpose(query,keep.names ='Notebooks')
notebooks
```

```
##             Notebooks   V1
## 1              Kaggle 5992
```

```
## 2     Colab Notebooks 6329
## 3     Azure Notebooks  857
## 4          Paperspace  180
## 5              Binder 2072
## 6           Code Ocean  105
## 7           IBM Studio  846
## 8        Amazon Studio  497
## 9     Amazon Notebooks  245
## 10 Google AI Platform 1218
## 11     Google Datalab 1231
## 12   Databricks colab  394
## 13               None 5282
## 14              Other  485
```

```
ggplot(notebooks, aes(x=Notebooks, y=V1)) +
  geom_bar(stat = "summary",fun=sum)+
  stat_summary( fun=sum,geom="bar",fill='wheat4')+

  labs(x='Notebooks',y='Count',title = 'Notebooks',subtitle="Which of the following hosted
       notebook products do you use on a regular basis?")+
  theme(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5,size=9))+
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))+
  coord_flip()
```



Unsurprisingly, Colab Notebooks and Kaggle which use a jupyter IDE are the most used hosted notebook products.

END