Project1

Jiangying Zhong,Ziyi Zheng,Zheng Lu

10/18/2021

Contents

Import Libraries

```
library(dplyr)
library(ggplot2)
library(ggmap)
library(knitr)
library(tidyr)
library(tibble)
library(reshape2)
library(treemap)
```

Import Data

We utilized the user survey from Kaggle during 2017 to 2020 as our metadata for this project.For this part,we only remove the row where the questionnaire is located to get the user's answer data.

```
survey2020 <- read.csv("/Users/zhongjiangying/Desktop/NEU/IE5374/project1/2020/2020_kaggle_ds_and_ml_su
survey2020 <- survey2020[-1,]

survey2019 <- read.csv("/Users/zhongjiangying/Desktop/NEU/IE5374/project1/kaggle-survey-2019/multiple_c
survey2019 <- survey2019[-1,]

survey2018 <- read.csv("/Users/zhongjiangying/Desktop/NEU/IE5374/project1/2018/multipleChoiceResponses.
survey2018 <- survey2018[-1,]
```

Business Question 1

BQ1:Determine main users of Kaggle.

Data Wrangling

In this part, we select age and gender to determine the main users of Kaggle platform for each year from 2018 to 2020.For **age** variable,we firstly group the data according to age groups and then count each age group and arrange them in descending order.After that,we merge the data frames of these three years into one, denoted as *age_compare* which contains three columns:*age*:age group;*year*:year the survey was issued;*number*:number of users in the age group.For **gender** variable,we deal data in the same way and finally merge the data frames into *genders* which includes three columns:*gender*:gender group;*year*:year the survey was issued;*number*:number of users in the gender group.

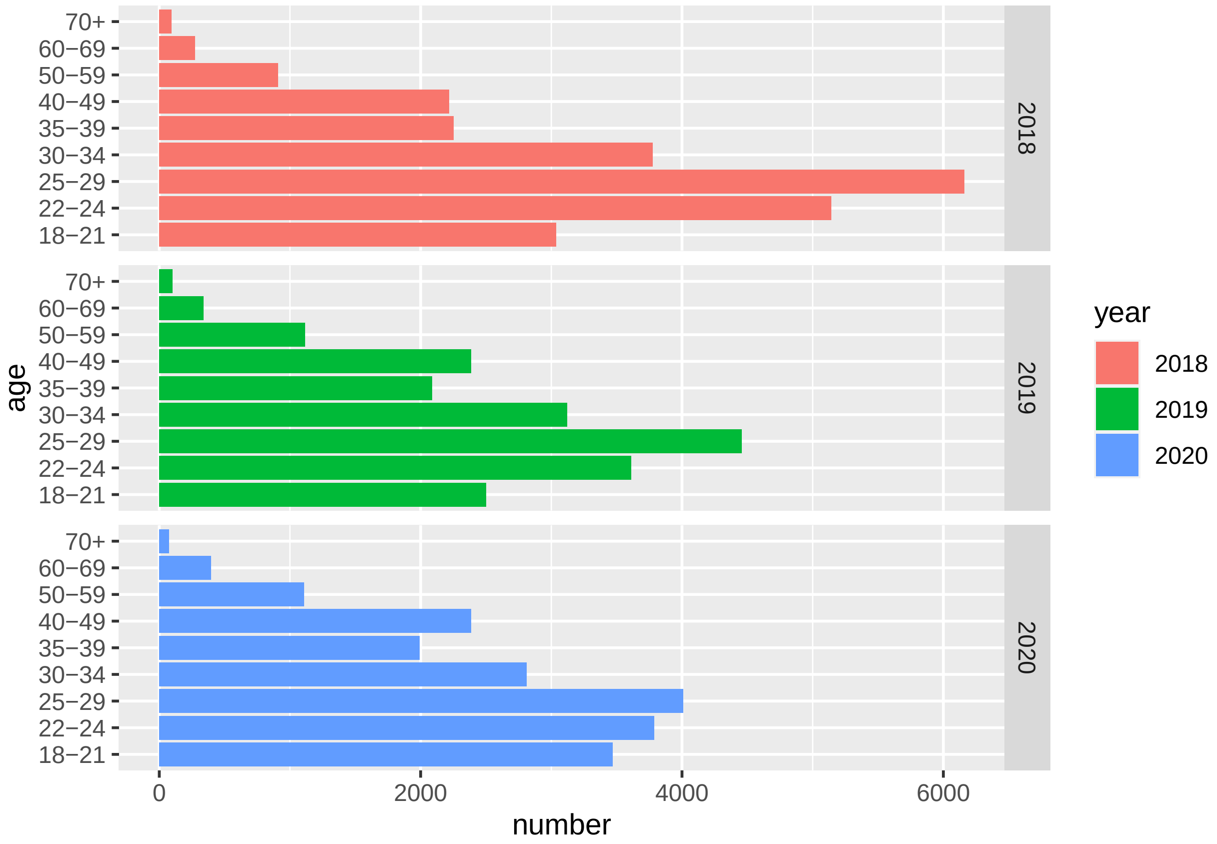
```
age_period2020 <- survey2020 %>%
  group_by(Q1) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(flag = case_when(Q1=='18-21' ~ '18-21',Q1=='22-24' ~ '22-24',Q1=='25-29' ~ '25-29' ,Q1=='30-34
  group_by(flag) %>%
  summarise(total_number = sum(number,na.rm = TRUE)) %>%
  arrange(flag) %>%
  mutate(year = '2020')
names(age_period2020) <- c('age','number','year')

age_period2019 <- survey2019 %>%
  group_by(Q1) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(flag = case_when(Q1=='18-21' ~ '18-21',Q1=='22-24' ~ '22-24',Q1=='25-29' ~ '25-29' ,Q1=='30-34
  group_by(flag) %>%
  summarise(total_number = sum(number,na.rm = TRUE)) %>%
  arrange(flag) %>%
  mutate(year = '2019')
names(age_period2019) <- c('age','number','year')

age_period2018 <- survey2018 %>%
  group_by(Q2) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(flag = case_when(Q2=='18-21' ~ '18-21',Q2=='22-24' ~ '22-24',Q2=='25-29' ~ '25-29' ,Q2=='30-34
  group_by(flag) %>%
  summarise(total_number = sum(number,na.rm = TRUE)) %>%
  arrange(flag) %>%
  mutate(year = '2018')
names(age_period2018) <- c('age','number','year')

age_period <- merge(age_period2018,age_period2019,by = 'age')
age_period <- merge(age_period,age_period2020,by = 'age')
names(age_period) <- c('age','number2018','2018','number2019','2019','number2020','2020')
age_compare <- data.frame(age = rep(age_period$age,3), year = rep(c('2018','2019','2020'),each = 9),numb
```

```
genders2020 <- survey2020 %>%
  group_by(Q2) %>%
  summarise(number = n()) %>%
  drop_na() %>%
  arrange(desc(number))
names(genders2020) <- c('gender','number')
genders2020[1,1] ='Male'
genders2020[2,1] = 'Female'

genders2019 <- survey2019 %>%
  group_by(Q2) %>%
  summarise(number = n()) %>%
  drop_na() %>%
  arrange(desc(number))
names(genders2019) <- c('gender','number')

genders2018 <- survey2018 %>%
  group_by(Q1) %>%
  summarise(number = n()) %>%
  drop_na() %>%
  arrange(desc(number))
names(genders2018) <- c('gender','number')

genders <- merge(genders2018,genders2019,by = 'gender')
genders <- merge(genders,genders2020[-5,],by = 'gender')
names(genders) <- c('gender','2018','2019','2020')
genders <- melt(genders,id = 'gender')
names(genders) <- c('gender','year','number')
```

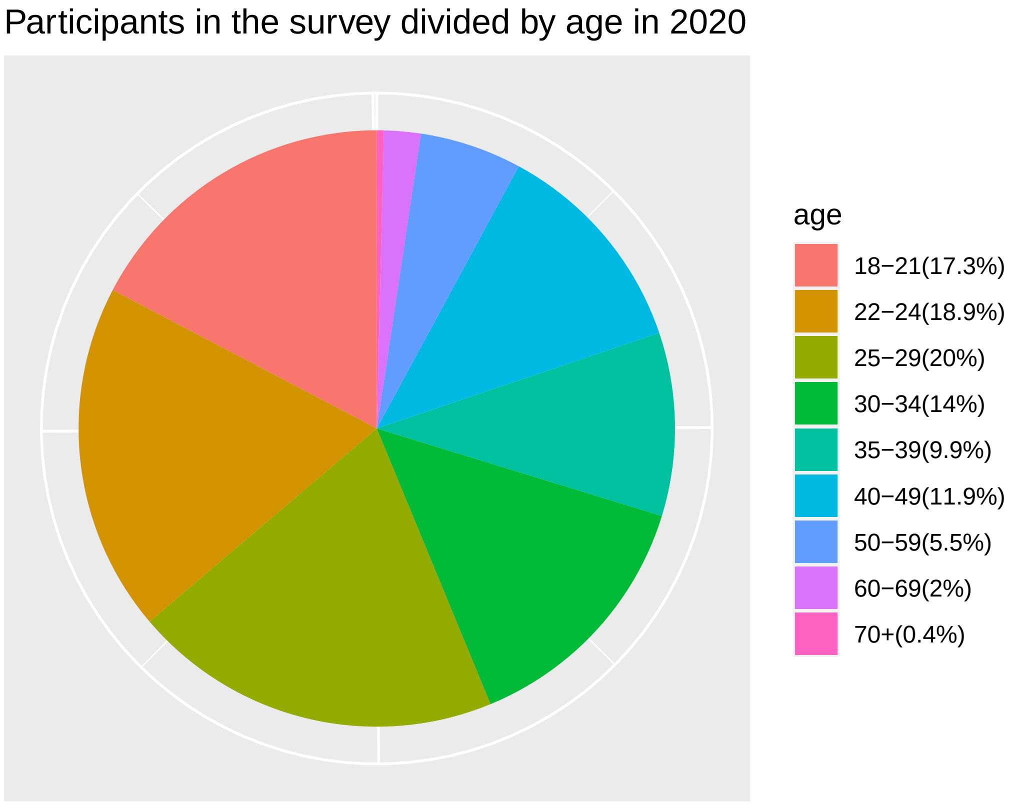
Data Visualization & Conclusion

Comparing the three-year trend,we find that the number of young people (18-30 years old) is on the rise, but the number of men has continuously been far more than that of women, almost three times the number of women,although the number of women users is climbing up.In 2020,age group '18-30' occupies 46.2%, and the number of men accounted for 78.8% while the number of women only accounted for 19.4%.Consequently,our suggestion is promoting more on those 18-30 years old and giving more chance for women.

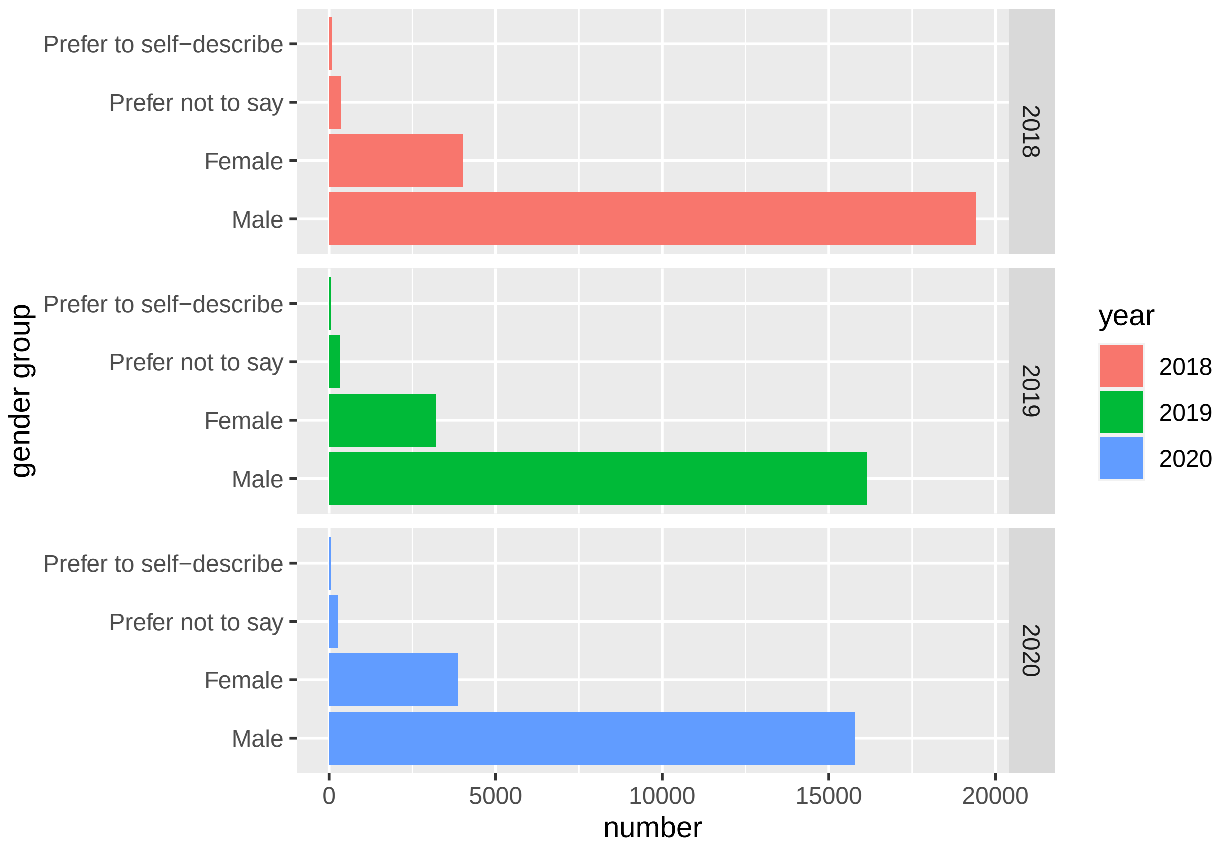
```
age_bargraph <- age_compare %>%
  ggplot(mapping = aes(x = age,fill = year,y = number)) +
  geom_col(position = 'dodge') +
  coord_flip() +
  facet_grid(year~.)
age_bargraph
```



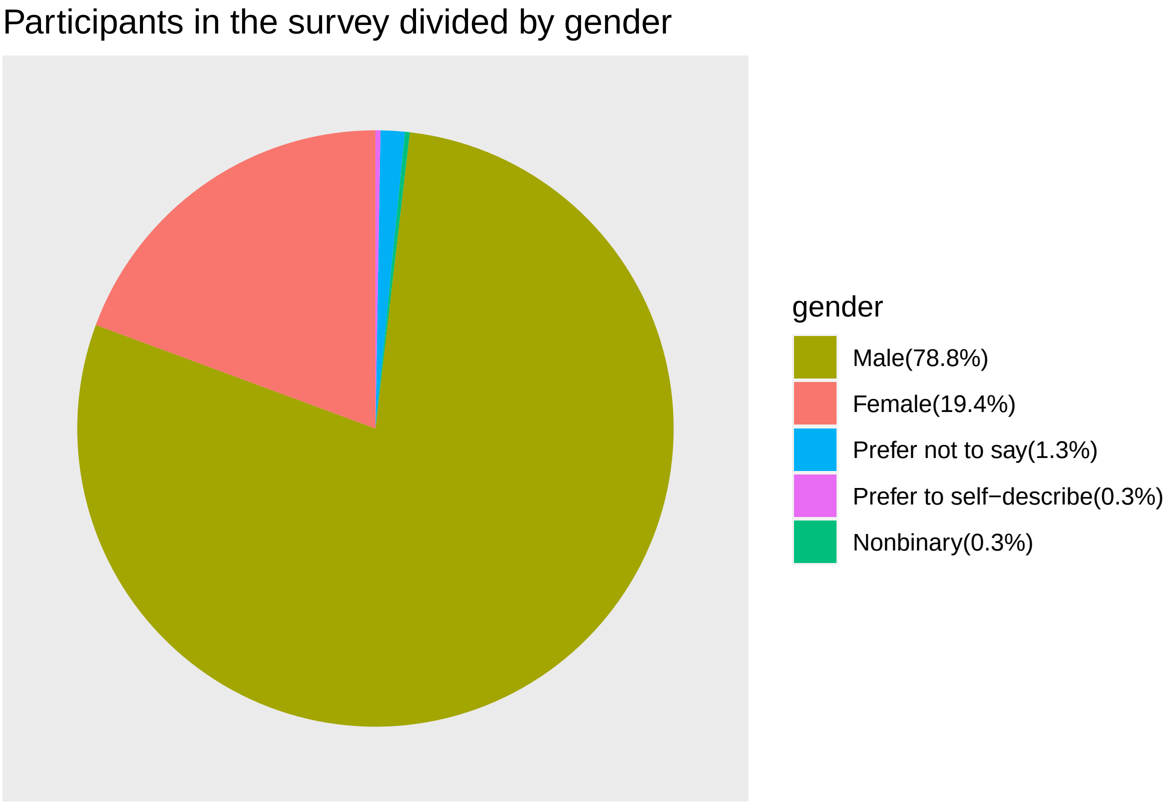
```
label_value <- paste('(', round(age_period2020$number/sum(age_period2020$number) * 100, 1), '%)', sep =
label <- paste(age_period2020$age, label_value, sep = '')
age_piechart <- ggplot(age_period2020,aes(x = '',y = number,fill = age)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  labs(x = '', y = '', title = 'Participants in the survey divided by age in 2020') +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_blank()) +
  scale_fill_discrete(breaks = age_period2020$age,labels = label)
age_piechart
```


Participants in the survey divided by age in 2020

```
gender_bargraph <-genders %>%
  ggplot(mapping = aes(x = reorder(gender,-number),fill = year,y = number)) +
  geom_col(position = 'dodge') +
  xlab('gender group') +
  ylab('number') +
  coord_flip() +
  facet_grid(year~.)
gender_bargraph
```



```
label_value1 <- paste('(', round(genders2020$number/sum(genders2020$number) * 100, 1), '%)', sep = '')
label1 <- paste(genders2020$gender, label_value1, sep = '')
gender_piechart <- ggplot(genders2020,aes(x = '',y = number,fill = gender)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  labs(x = '', y = '', title = 'Participants in the survey divided by gender') +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_blank()) +
  theme(panel.grid = element_blank()) +
  scale_fill_discrete(breaks = genders2020$gender,labels = label1)
gender_piechart
```



Business Question 2

BQ2:Identify potential customer status of Kaggle.

Data Wrangling

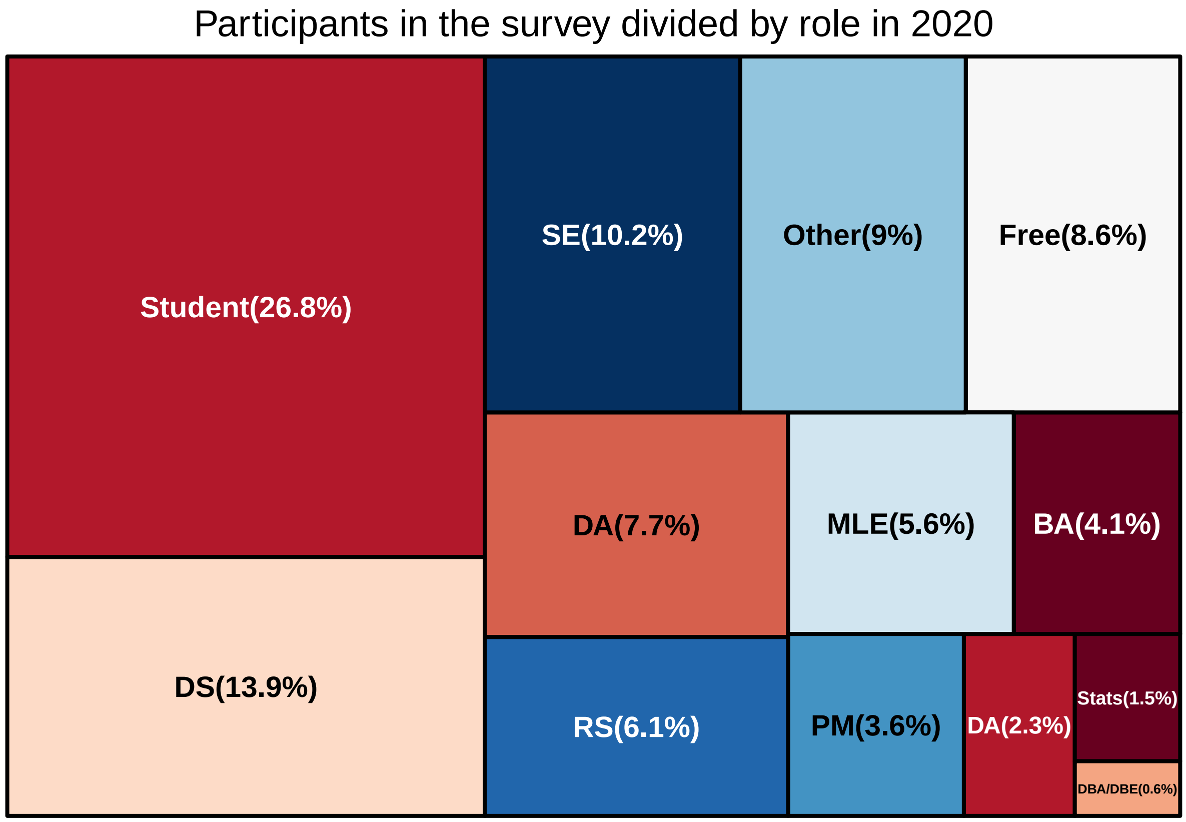
In this part, we select role(column5) to determine potential customer status of Kaggle platform in 2020.We group the data according to column5 and then count each group, remove the NA values and arrange them in descending order. The data frame called *potential_status* contains two columns:*status*:user's current role;*number*:number of users in the group.

```
potential_status <- survey2020%>%
  group_by(Q5) %>%
  summarise(number_of_users = n()) %>%
  drop_na() %>%
  arrange(desc(number_of_users)) %>%
  mutate(
    flag=
      case_when(Q5=='Student' ~ 'Student',Q5=='Business Analyst' ~ 'BA',Q5=='Data Analyst' ~ 'DA',Q5=='
  select(flag, number_of_users)
names(potential_status) <- c('status','number')
```

Data Visualization & Conclusion

From the data,we find that students occupies the most at 26.8%,following by Data Scientist(13.9%),Software Engineer(10.2%).However,Database Analyst/Database Engineer occupies the least at 0.6%.Hence,in order to increase the stickiness of the Kaggle users,we should focus more on Students, Data Scientist and Software Engineer such as providing more relevant competitions.

```r
label_value2 <- paste('(', round(potential_status$number/sum(potential_status$number) * 100, 1), '%)', 
label2 <- paste(potential_status$status, label_value2, sep = '')
potential_status <- mutate(potential_status,percent=label2)
potential_status_treemap <- treemap(potential_status,index=c("percent"),vSize="number", type="index",ti
```



Business Question 3

BQ3:Clarify market demanding on Data Science and Machine Learning.

Data Wrangling

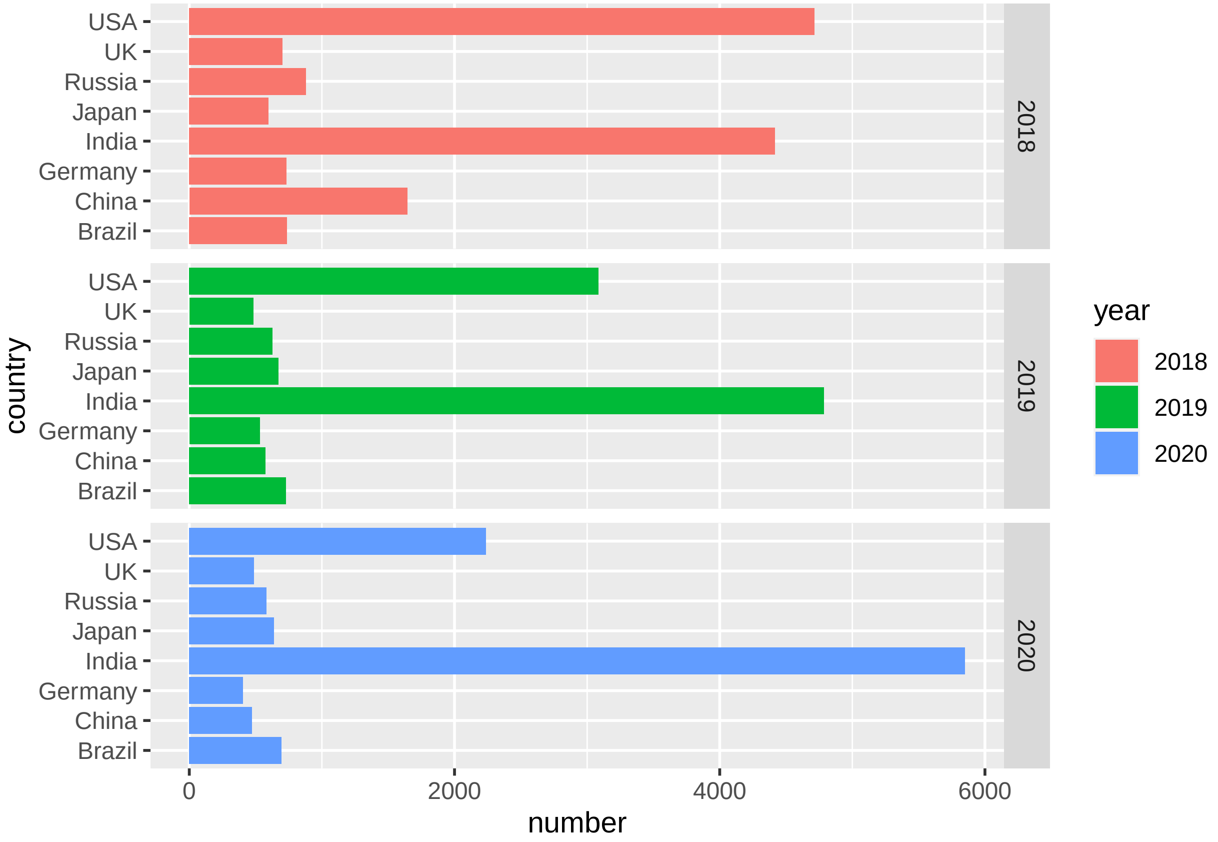
For each year in 2018-2020 period,we group the data by country,count the number of each group and then arrange them in descending order.After that,we select the top eleven countries,excluding the variable whose value equals NA.We then combine the data frames into one called **country** using inner_join function so that the intersection of these three would remain.In data frame **country**,there are three columns:*country*:country that user belongs to;*year*:year the survey was issued;*number*:number of users in the group.

```
country2020 <- survey2020 %>%
  group_by(Q3) %>%
  summarise(number=n()) %>%
  arrange(desc(number))
names(country2020) <- c('country','number')
country2019 <- survey2019 %>%
  group_by(Q3) %>%
  summarise(number=n()) %>%
  arrange(desc(number))
names(country2019) <- c('country','number')
country2018 <- survey2018 %>%
  group_by(Q3) %>%
  summarise(number=n()) %>%
  arrange(desc(number))
names(country2018) <- c('country','number')

c1 <- country2018[1:11,]
c2 <- country2019[1:11,]
c3 <- country2020[1:11,]
country <- inner_join(c1,c2,by='country')
country <- inner_join(country,c3,by='country')
names(country) <- c('country','2018','2019','2020')
country <- melt(country,id = 'country')
names(country) <- c('region','year','number')
country <- country %>%
  filter(region!='Other') %>%
  mutate(country = case_when(region=='United States of America' ~ 'USA',region=='India' ~ 'India',regio
  select(country,year,number)
```

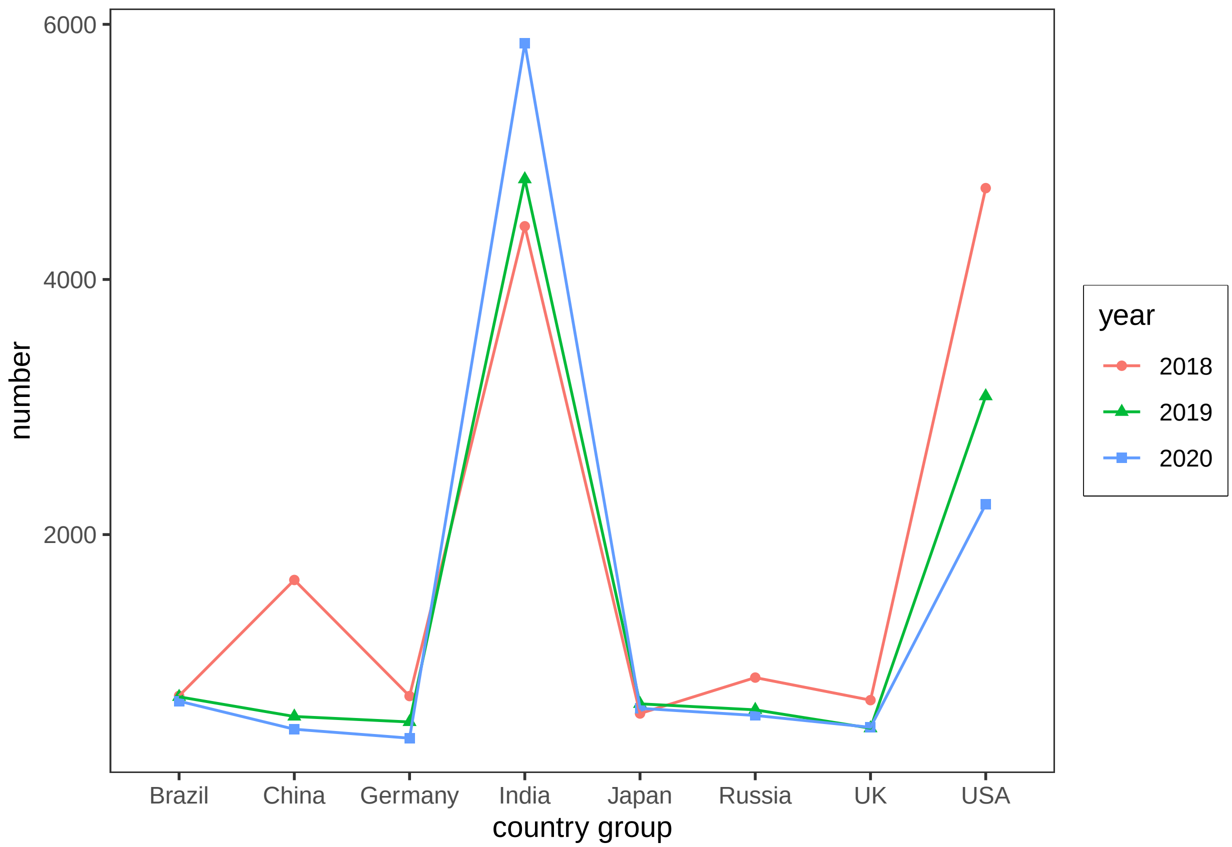
Data Visualization & Conclusion

From the figure,we can see that the number of users in USA and India has been far more than that in other countries, although the number of users in USA has declined from 2018 to 2020.During the same period,Brazil,Germany,UK,Japan and Russia have basically stabilized the number of users in the past three years.However,The number of users in China has decreased to a large extent.Since the number of users in these eight countries far exceeds that of other countries in three years, we recommend that Kaggle keeps concentrating on developing these markets. At the same time, we believe that Kaggle needs to make appropriate adjustments to the US and Chinese markets to reduce the loss of users.

```
country_bargraph <-country %>%
  ggplot(mapping = aes(x = country,fill = year,y = number)) +
  geom_col(position = 'dodge') +
  coord_flip() +
  facet_grid(year~.)
country_bargraph
```



```
country_linegraph <- ggplot(country,aes(x = country,y = number,group = year,color = year,shape = year))
  geom_point() +
  geom_line() +
  xlab('country group') +
  ylab("number") +
  theme_bw() +
  theme(panel.grid.major = element_line(colour = NA)) +
  theme(panel.background = element_rect(fill = 'transparent',colour = NA)) +
  theme(panel.grid.minor = element_blank()) +
  theme(legend.box.background = element_rect(colour = 'black'))
country_linegraph
```



Business Question 4

BQ4:Explore the coding importance in Data Science and Machine Learning.

Data Wrangling

We used the years for coding to make the assumption about the coding years, “Totally Fresh” means that “I have never written code”, “Beginners” means that “ < 1 years & 2 years”, “Junior” means that “3-5 years”, “Senior” means that “5-10 years” and “Experienced” means that “10-20 years & 20+ years”. Therefore, we introduced the Q6 and Q15 to explore the coding experience. The final data frame **coding_period** contains three columns:*programming_time*:how well user exposure to programming is;*year*:year the survey was issued;*number*:number of users in the group.

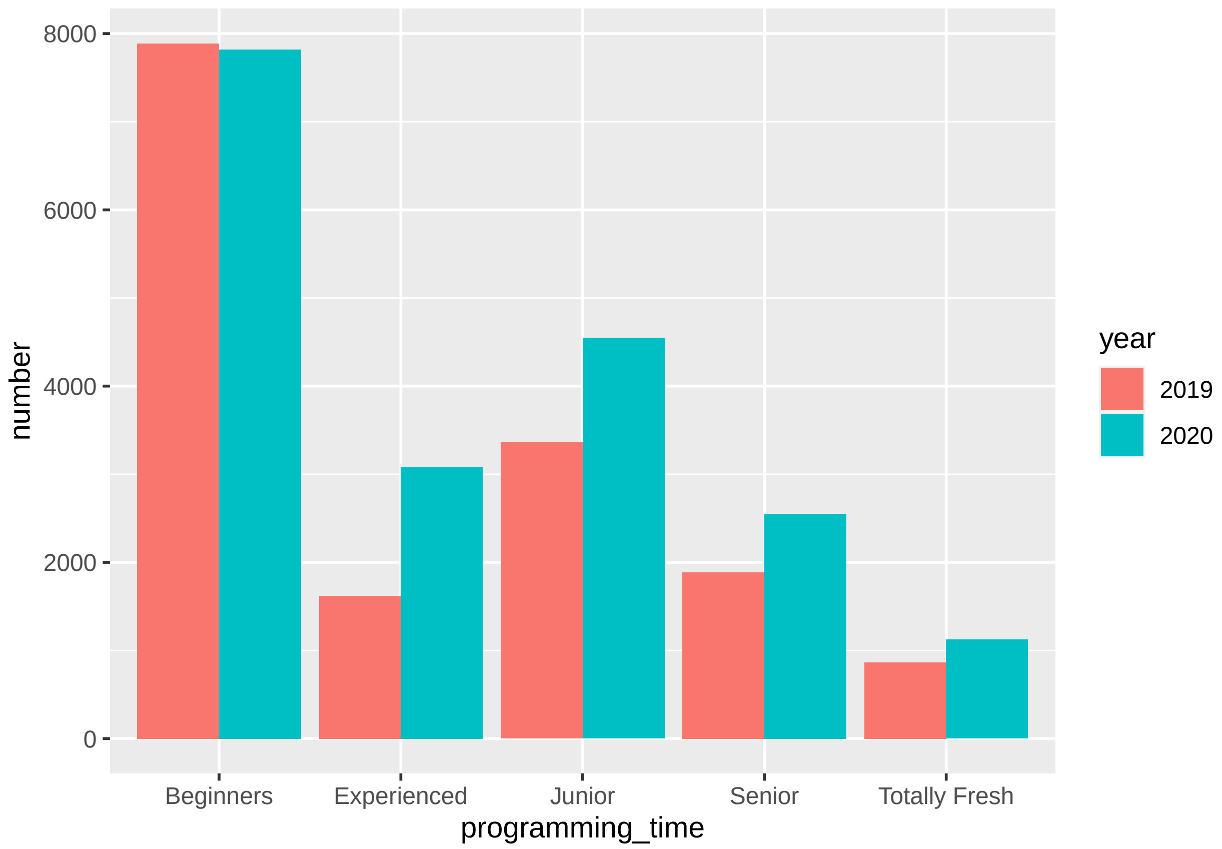
```
coding_period2020 <- survey2020 %>%
  group_by(Q6) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(flag = case_when(Q6=='I have never written code' ~ 'Totally Fresh',Q6=='1-2 years' ~ 'Beginners
  select(flag,number) %>%
  group_by(flag) %>%
  summarise(total=sum(number,na.rm = TRUE))
names(coding_period2020) <- c('programming_time','number')
```

```r
coding_period2019 <- survey2019 %>%
  group_by(Q15) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(flag = case_when(Q15=='I have never written code' ~ 'Totally Fresh',Q15=='1-2 years' ~ 'Beginne
  select(flag,number) %>%
  group_by(flag) %>%
  summarise(total=sum(number,na.rm = TRUE))
names(coding_period2019) <- c('programming_time','number')

coding_period <- merge(coding_period2019,coding_period2020,by='programming_time')
names(coding_period) <- c('programming_time','2019','2020')
coding_period <- melt(coding_period,id = 'programming_time')
names(coding_period) <- c('programming_time','year','number')
```

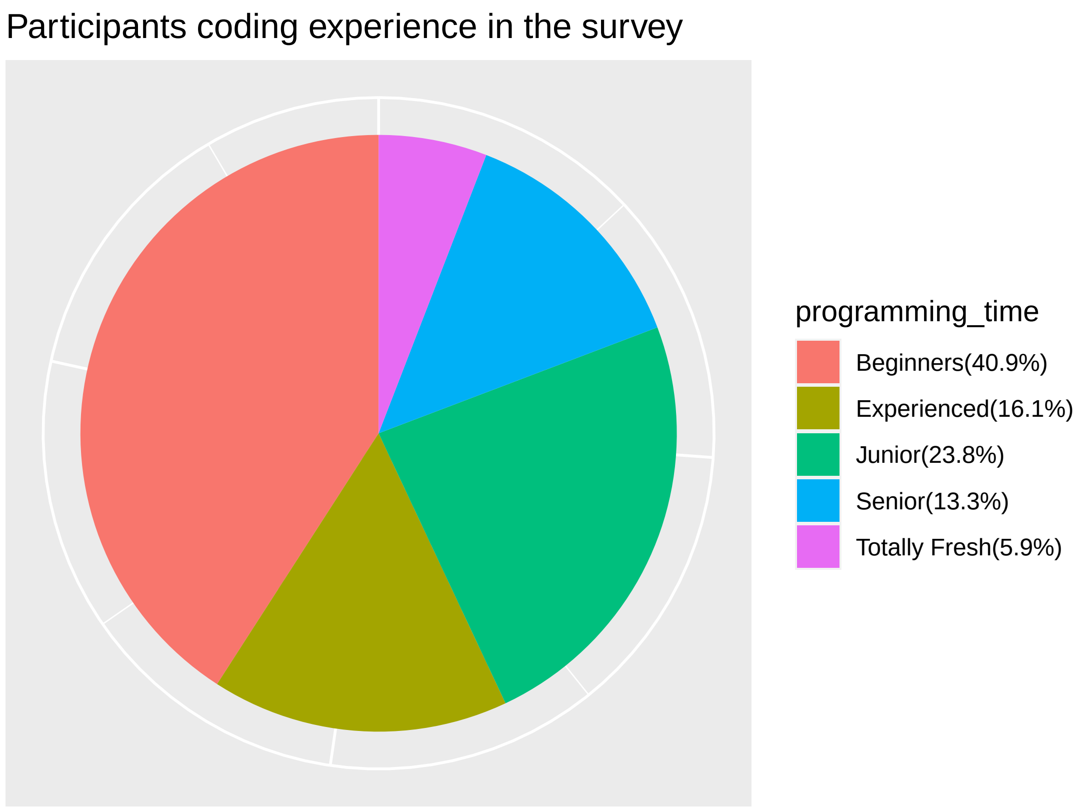
Data Visualization & Conclusion

According to the graph, we observed that the most are beginners which means that users only have coding experience for 1 and 2 years. The next level is Junior, which means that users have coding experience for 3-5 years. And the least number is totally fresh users. For the beginners, the numbers in 2019 and 2020 are similar. For the rest, the numbers in 2020 are much more than the numbers in 2019. The population of beginners accounts for almost half of the total population. It is worth noting that the number of beginners is much higher than the others in both 2019 and 2020.Based on these data, we have reason to believe that programming is becoming more and more important in daily work and life, and Kaggle should focus on this development, such as providing programming courses.

```r
coding_bargraph <- coding_period %>%
  ggplot(mapping = aes(x = programming_time,fill = year,y = number)) +
  geom_col(position = 'dodge')
coding_bargraph
```



```
label_value3 <- paste('(', round(coding_period2020$number/sum(coding_period2020$number) * 100, 1), '%)'
label3 <- paste(coding_period2020$programming_time, label_value3, sep = '')
coding_piechart <-
  ggplot(coding_period2020,aes(x = '',y = number,fill = programming_time)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  labs(x = '', y = '', title = 'Participants coding experience in the survey') +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_blank()) +
  scale_fill_discrete(breaks = coding_period2020$programming_time,labels = label3)
coding_piechart
```



Business Question 5

BQ5:Discover the most popular coding language in Data Science and Machine Learning.

Data Wrangling

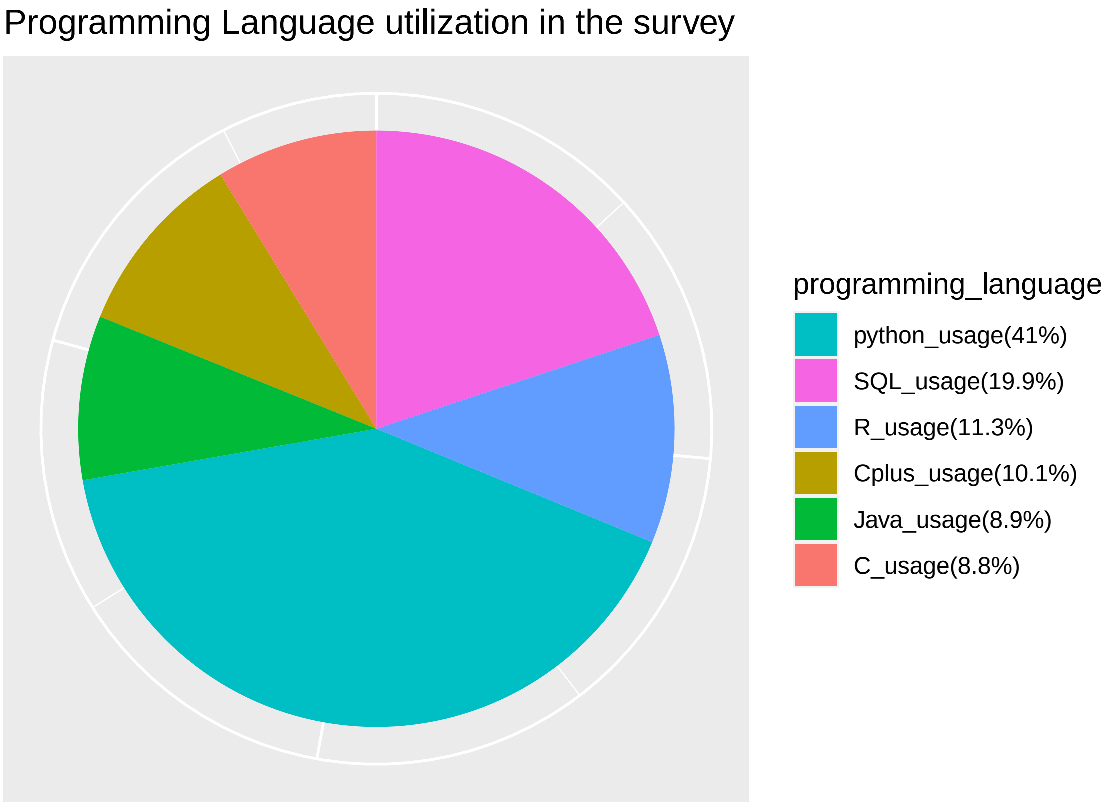
In order to obtain the result about what is the most popular coding language in Data Science and Machine Learning, we drew the data frame for different coding languages to compare.Since the problem is a multiple choice question, we must first integrate the 8-12 columns and then process and analyze the data.Based on this, we use a for loop statement to integrate 8-12 columns of data into a data frame and use the name of the programming language as the column name of the new data frame.Finally, we transpose the data frame and select the top six to make it easier for us to process.The final data frame is called **language__usage**, with three columns respectively:*programming__language*:name of the coding language;*number*:number of people using the programming language.

```
vec <- c()
for(i in 8:20){
  temp <- filter(survey2020,is.na(survey2020[,i]) == FALSE)
  temp1 <- data.frame(temp[,i])
  ans <- summarise(temp1,number = n())
  vec <- c(vec,ans)
}
names(vec) <- c('python_usage','R_usage','SQL_usage','C_usage','Cplus_usage','Java_usage','Javascript_u
language_usage <- as.data.frame(vec)
language_usage <- t(language_usage)
language_usage <- data.frame(programming_language = rownames(language_usage),number = language_usage[,1
language_usage <- language_usage %>%
  arrange(desc(number)) %>%
  slice(1:6)
```

Data Visualization & Conclusion

According to the bar graph, we realized that the top two coding languages are python and SQL. In addition, more than twice as many users utilize python as the SQL users. The population of R, Cplus, Java and C users are similar. Therefore, the most popular coding language in Machine Learning and Data Science is Python.

```
label_value4 <- paste('(', round(language_usage$number/sum(language_usage$number) * 100, 1), '%)', sep
label4 <- paste(language_usage$programming_language, label_value4, sep = '')
language_piechart <-
  ggplot(language_usage,aes(x = '',y = number,fill = programming_language)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  labs(x = '', y = '', title = 'Programming Language utilization in the survey') +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_blank()) +
  scale_fill_discrete(breaks = language_usage$programming_language,labels = label4)
language_piechart
```



Business Question 6

BQ6:Evaluate the usage of Kaggle IDE.

Data Wrangling

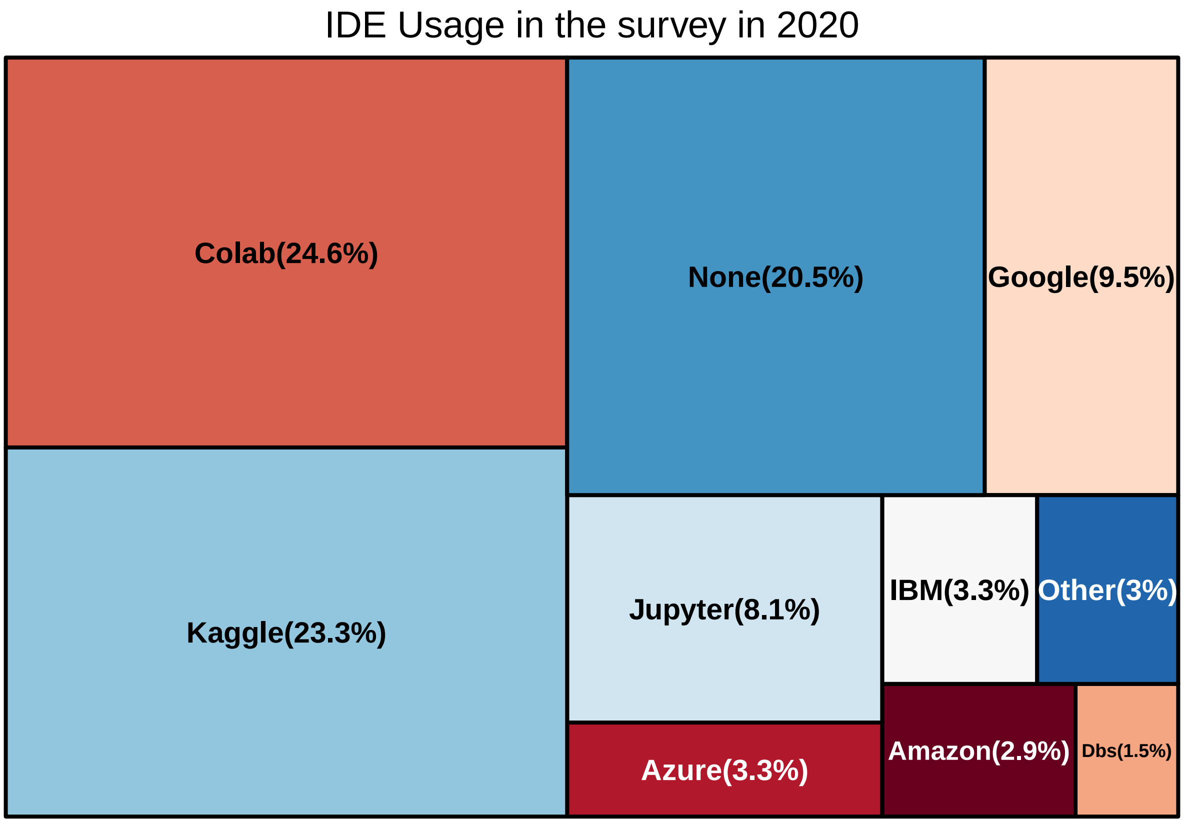
We compare the usage of Kaggle, Colab, Azure, Jupyter, IBM, Amazon, Google, Dbs, etc. IDEs. And present result in tree map form. We combine *Amazon Sagemaker* and *Amazon EMR* as *Amazon*; combine *Google AI* and *Google Datalab* as *Google*; combine *Paperspace*, *Code Ocean*, *Other* as *Other*.In the final **IDE_usage** data frame,there are two columns:*IDE*:name of IDEs;*number*:number of people using the IDE.

```
subsurvey <- select(survey2020,Q10_Part_1,Q10_Part_2,Q10_Part_3,Q10_Part_4,Q10_Part_5,Q10_Part_6,Q10_Pa
list1 <- list()
for (i in 1:14) {
  list1 <- append(list1,table(subsurvey[,i]))
}
names(list1) <- c('Kaggle','Colab','Azure','Paperspace','Jupyter','Code Ocean','IBM','Amazon Sagemaker'
IDE_usage <- as.data.frame(list1) #
IDE_usage <- t(IDE_usage)
IDE_usage <- data.frame(IDE = rownames(IDE_usage), number = IDE_usage[,1])
row1 <- c('Amazon',sum(IDE_usage$number[8:9]))
row2 <- c('Google',sum(IDE_usage$number[10:11]))
row3 <- c('Other',sum(IDE_usage$number[c(4,6,14)]))
IDE_usage <- rbind(IDE_usage,row1,row2,row3)
IDE_usage <- IDE_usage[c(-14,-11:-8,-6,-4),]
IDE_usage$number <- as.integer(IDE_usage$number)
```

Data Visualization & Conclusion

From the tree map, we found that Colab has the most users, next goes by Kaggle. However, in the survey, not using any IDE in the participants are also very common. The Top3 accounted for a large proportion of participants.To maintain and expand existing advantages, Kaggle can develop more functions to meet developers' needs, like reducing the memory usage of IDE, installing plugins and extensions, etc.

```
label_value5 <- paste('(', round(IDE_usage$number/sum(IDE_usage$number) * 100, 1), '%)', sep = '')
label5 <- paste(IDE_usage$IDE, label_value5, sep = '')
IDE_usage <- mutate(IDE_usage,percent=label5)
IDE_treemap <- treemap(IDE_usage,index=c("percent"),vSize="number", type="index",title='IDE Usage in the
```



Business Question 7

BQ7:Estimate the correlation between Data Science and Machine Learning.

Data Wrangling

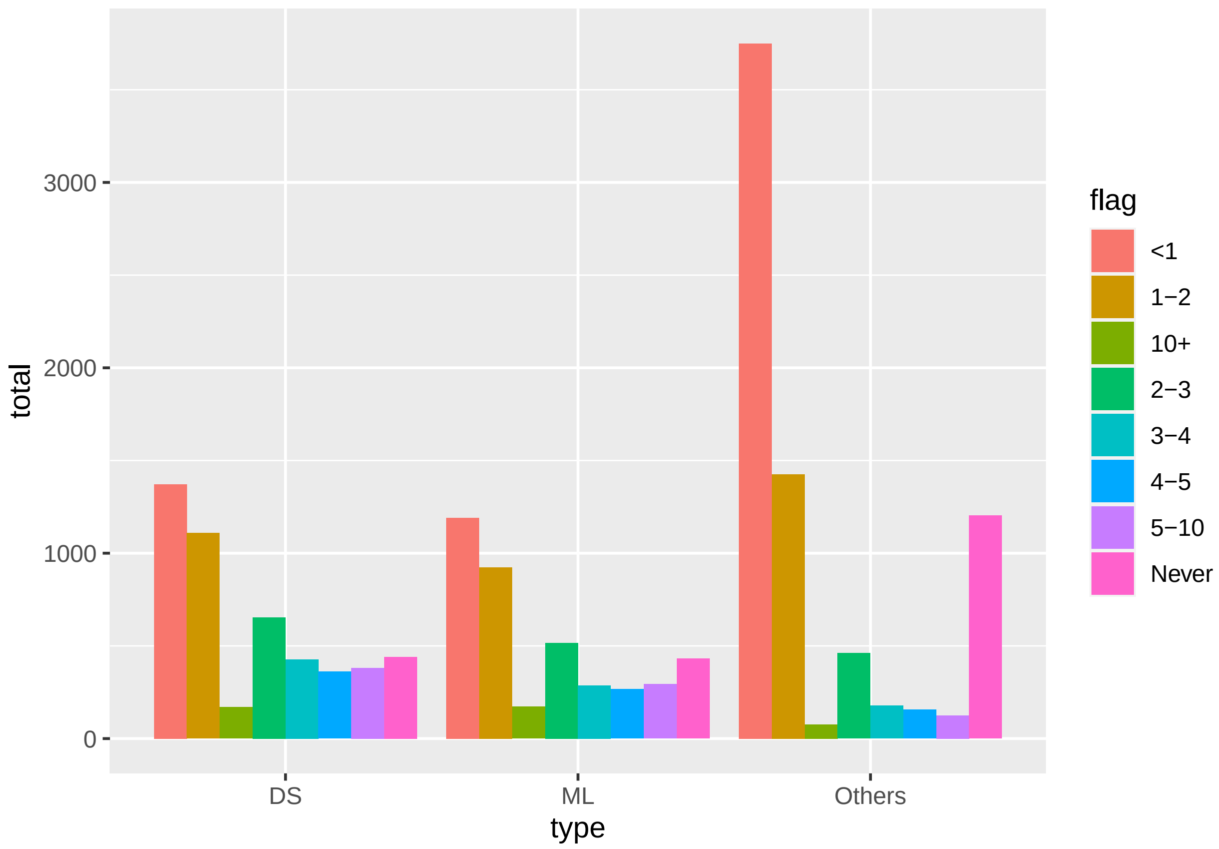
In this part, we managed data about Machine learning in 2019 and 2020 and grouped them by Q5, which is the title that is most similar to their current role. For question 23, we concluded the roles that are important at their works and summarized the parts that related to machine learning in 2019 and 2020. Then we combined them together to observe the tendency.Also,due to the subtle differences between 2019 and 2020 options, we have made the following assumptions:using '<1' equals to 'under 1 year';using '1-2' equals to '1-2 years';using '2-10' equals to the totality of '2-3 years','3-4 years','4-5 years' and '5-10 years';using '10+' equals to the totlity of '10-20 years', '20 or more years' and '10-15 years'. We ultimately get a data frame 'ML_period' with four columns:*label*:the length users utilize machine learning methods;*type*:type of their jobs;*number*:number of users in the group;*year*:year the survey was issued.

```
ML_period2020 <- survey2020 %>%
  filter(is.na(Q15)==FALSE) %>%
  mutate(flag=case_when(Q15=='I do not use machine learning methods'~'Never',Q15=='Under 1 year'~'<1',Q
  group_by(flag,Q5) %>%
  summarise(number=n()) %>%
  mutate(type=case_when(Q5=='Business Analyst'~'DS',Q5=='Data Analyst'~'DS',Q5=='Data Engineer'~'DS',Q5=
  group_by(flag,type) %>%
  summarise(total=sum(number))

ML_period2019 <- survey2019 %>%
  mutate(flag=case_when(Q23=='< 1 years'~'<1',Q23=='1-2 years'~'1-2',Q23=='2-3 years'~'2-3',Q23=='3-4 ye
  group_by(flag,Q5) %>%
  summarise(number=n()) %>%
  mutate(type=case_when(Q5=='Business Analyst'~'DS',Q5=='Data Analyst'~'DS',Q5=='Data Engineer'~'DS',Q5=
  group_by(flag,type) %>%
  summarise(total=sum(number))

ML_period <- merge(ML_period2019,ML_period2020,by=c('flag','type'))
names(ML_period) <- c('flag','type','2019','2020')
ML_period <- melt(ML_period,id=c('flag','type'))
names(ML_period) <- c('flag','type','year','number')
ML_period <- ML_period %>%
  mutate(label=case_when(flag=='<1'~'<1',flag=='1-2'~'1-2',flag=='2-3'~'2-10',flag=='3-4'~'2-10',flag==
  group_by(label) %>%
  select(label,type,number,year)
```

```
subsurvey_activities_2020 <- survey2020[,111:118]
list_temp <- list()
for (i in 1:length(subsurvey_activities_2020)) {
  list_temp <- append(list_temp,table(subsurvey_activities_2020[,i]))
}
names(list_temp) <- c('Analyze', 'Build data infrastructure', 'Explore ML','Improvement','Iteration','R
activities_2020 <- t(as.data.frame(list_temp))
activities_2020 <- data_frame(activities = rownames(activities_2020), number = activities_2020[,1])

subsurvey_activities_2019 <- survey2019[,12:19]
list_temp_1 <- list()
for (i in 1:length(subsurvey_activities_2019)) {
  list_temp_1 <- append(list_temp_1,table(subsurvey_activities_2019[,i]))
}
names(list_temp_1) <- c('Analyze', 'Build data infrastructure', 'Explore ML','Improvement','Iteration',
activities_2019 <- t(as.data.frame(list_temp_1))
activities_2019 <- data_frame(activities = rownames(activities_2019), number = activities_2019[,1])

Activities <- inner_join(activities_2019,activities_2020, by = 'activities')
colnames(Activities) <- c('activities',2019, 2020)
Activities <- melt(Activities, id = 'activities')
names(Activities) <- c('activities', 'year', 'number')
```

##Data Visualiation & Conclusion For data visualization, we draw a bar graph for data science and machine learning to compare the years that people have been used. According to these graphs, they showed that there are more data scientists than machine learning engineers. In addition, from 2019 to 2020, there is an increasing tendency to show that increasing populations are more likely to use both machine learning and data science.

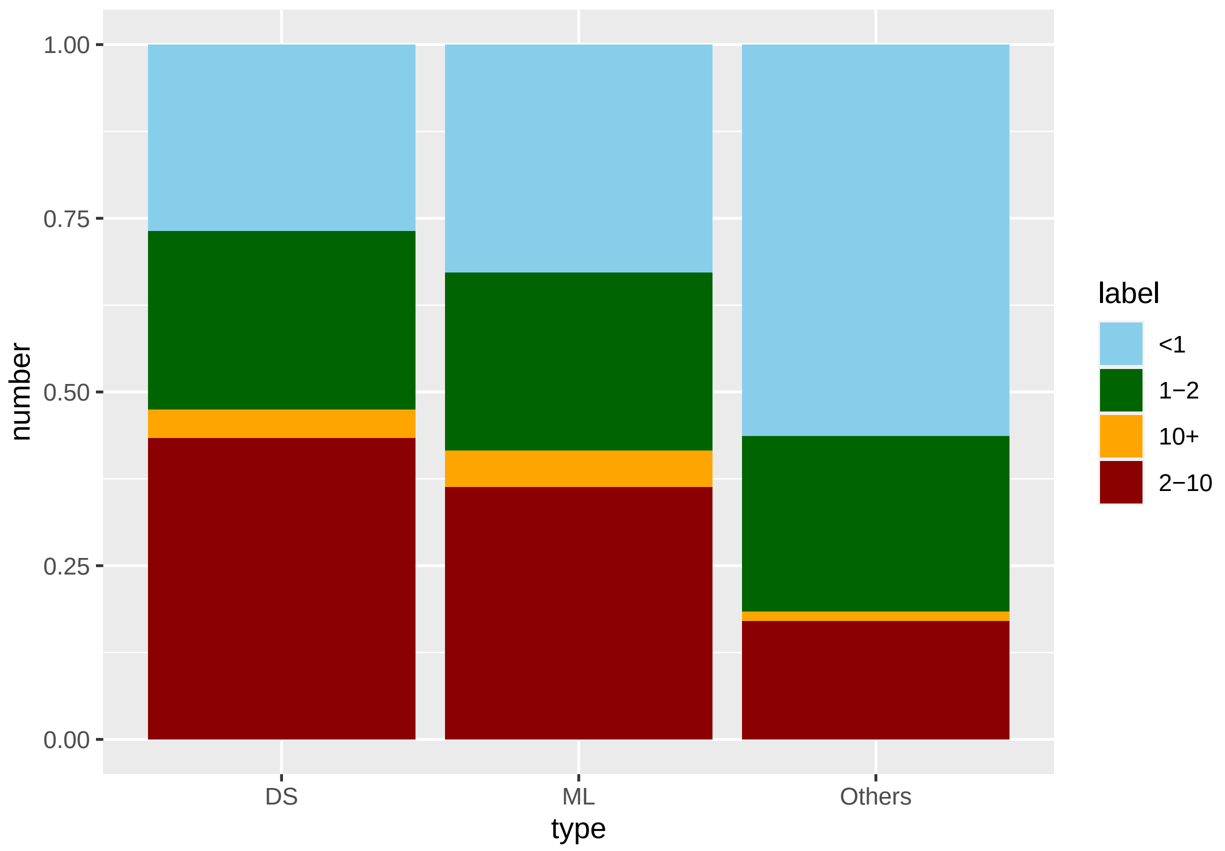
```
ML_bargraph2020 <- ML_period2020 %>%
  ggplot(mapping = aes(x = type,fill = flag,y = total)) +
  geom_col(position = 'dodge')
ML_bargraph2020
```



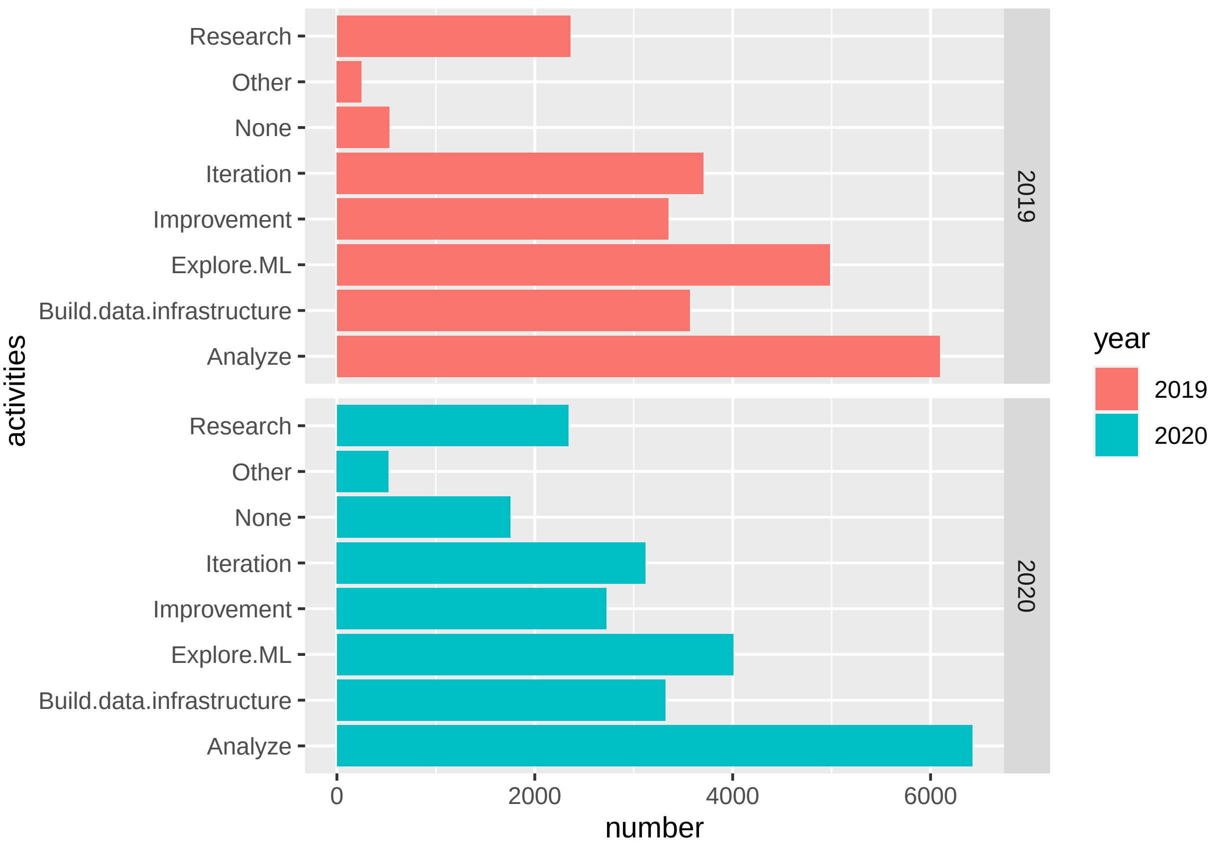
```
ML_bargraph <-ML_period %>%
  ggplot(mapping = aes(x = type,fill = label,y = number)) +
  geom_col(position = 'fill') +
  scale_fill_manual(values = c('#87CEEB','darkgreen','orange','darkred'))
  facet_grid(year~.)
```

```
## <ggproto object: Class FacetGrid, Facet, gg>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map_data: function
##     params: list
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train_scales: function
##     vars: function
##     super:  <ggproto object: Class FacetGrid, Facet, gg>
```

```
ML_bargraph
```



```
Activities_bar_chart <- Activities %>%
  ggplot(mapping = aes(x = activities, fill = year, y = number)) +
  geom_col(position = 'dodge') +
  coord_flip() +
  facet_grid(year~.)
Activities_bar_chart
```



Business Question 8

BQ8:Predict the overview of Data Science and Machine Learning.

Data Wrangling

To predict the overview of DS and ML, we concluded the percentages of both participants' incorporation ML into their business and data scientists in participants' companies in 2018-2020. We also compare the change of their salary within 3 years.We divided the participants into 4 levels according to the level of incorporation of ML, from high to low indicating the degree of incorporation from much to little.The data frame 'ml_incorporation' has three columns:*level*:how well users incorporate machine learning;*year*:year the survey was issued;*number*:number of users in the group.Also,'ds_per' has three columns:*size*:the company size;*year*:year the survey was issued;*number*:number of users in the group.For the salary, we make an assumption that salary within '0-999','1000-1999','2000-2999','3000-3999','4000-4999', '5000-7499' and '7500-9999' belongs to group '<10';'10000-14999','15000-19999','20000-24999' and '25000-29999' belongs to group '10-30';'30000-39999','40000-49999','50000-59999','60000-69999','70000-79999','80000-89999' and '90000-99999' belongs to group '50-100';'100000-124999','125000-149999','150000-199999','200000-249999','250000-299999' and '300000-500000' belongs to group '100-500';'> 500000' belongs to group '>500'.The data frame 'Salary' has three columns:*salary*:users annual salary;*year*:year the survey was issued;*number*:number of users in the group.

```
###Does your current employer incorporate machine learning methods into their business?
ml_incorporation_2020 <- survey2020 %>%
```

```r
  group_by(Q22) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit()
names(ml_incorporation_2020) <- c('ml_incorporation','number')
ml_incorporation_2020 <- ml_incorporation_2020 %>%
  mutate(percentage_2020 = paste(round(number/sum(number) * 100, 1), '%', sep = '')) %>%
  select(ml_incorporation,percentage_2020)

ml_incorporation_2019 <- survey2019 %>%
  group_by(Q8) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit()
names(ml_incorporation_2019) <- c('ml_incorporation','number')
ml_incorporation_2019 <- ml_incorporation_2019 %>%
  mutate(percentage_2019 = paste(round(number/sum(number) * 100, 1), '%', sep = '')) %>%
  select(ml_incorporation,percentage_2019)

ml_incorporation_2018 <- survey2018 %>%
  group_by(Q10) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit()
names(ml_incorporation_2018) <- c('ml_incorporation','number')
ml_incorporation_2018 <- ml_incorporation_2018 %>%
  mutate(percentage_2018 = paste(round(number/sum(number) * 100, 1), '%', sep = '')) %>%
  select(ml_incorporation,percentage_2018)

ml_incorporation <- merge(ml_incorporation_2018,ml_incorporation_2019,by = 'ml_incorporation')
ml_incorporation <- merge(ml_incorporation,ml_incorporation_2020,by = 'ml_incorporation')
names(ml_incorporation) <- c('ml_incorporation','2018','2019','2020')

ml_incorporation<- ml_incorporation %>%
  filter(ml_incorporation != 'I do not know') %>%
  mutate(flag = case_when(ml_incorporation=='No (we do not use ML methods)' ~ 'Level0',ml_incorporation=
  arrange(flag) %>%
  select(flag,`2018`,`2019`,`2020`)

ml_incorporation <- data.frame(level = rep(ml_incorporation$flag,3), year = rep(c('2018','2019','2020')
```

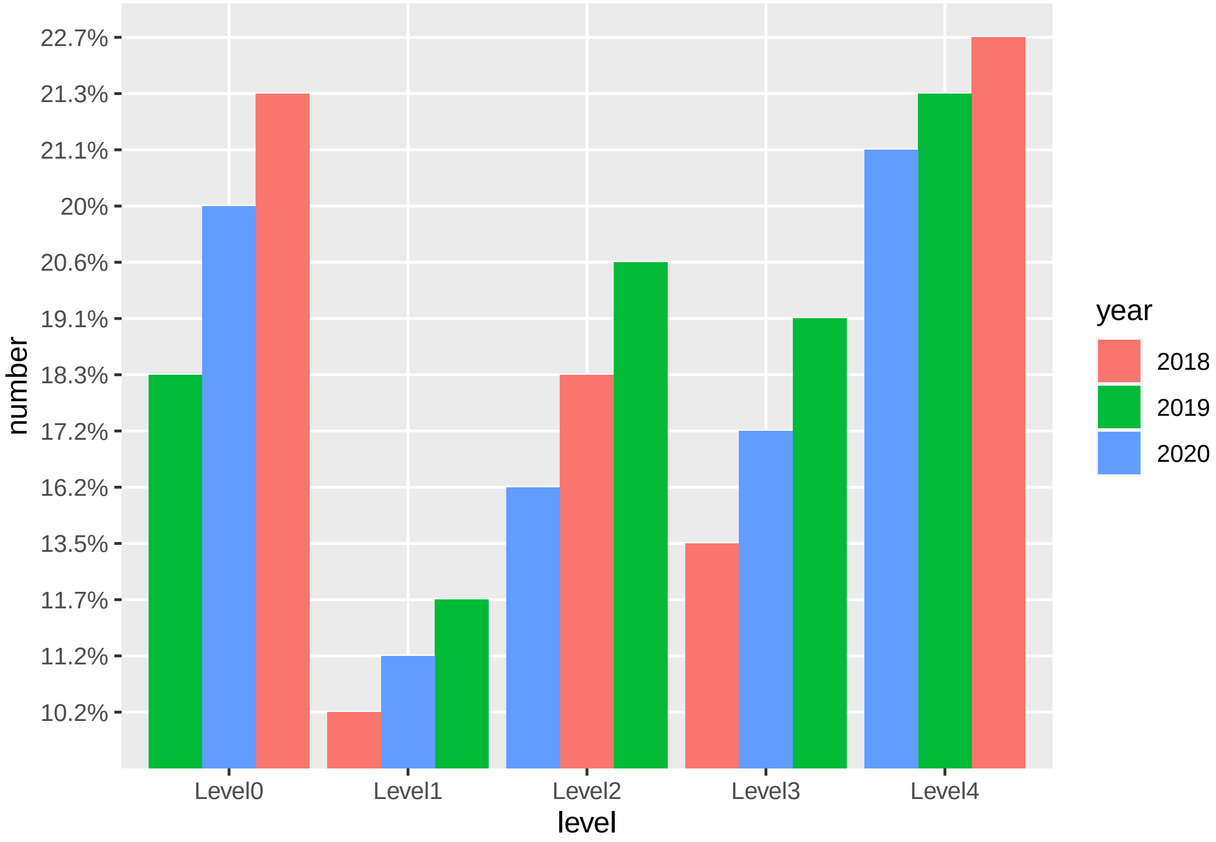
```r
### Approximately how many individuals are responsible for data science workloads at your place of busi
### What is the size of the company where you are employed?
data_science_teams_2020 <- survey2020 %>%
  select(Q20,Q21) %>%
  filter(is.na(Q20) == FALSE) %>%
  filter(is.na(Q21) == FALSE) %>%
  group_by(Q20,Q21) %>%
  summarise(numer=n()) %>%
  arrange(Q20)
names(data_science_teams_2020) <- c('size1','data science team','number')
data_science_teams_2020 <- data_science_teams_2020 %>%
  mutate(size = case_when(size1=='0-49 employees' ~ '0-49',size1=='10,000 or more employees' ~ '10,000+
  select(size,`data science team`,number)
data_science_teams_2020 <- data_science_teams_2020[,-1]
data_science_teams_2020 <- data_science_teams_2020 %>%
  filter(`data science team` != '0') %>%
  mutate(per = number/sum(number)) %>%
  group_by(size) %>%
  mutate(total=sum(number),percentage=sum(per)) %>%
  select(size,total,percentage)
data_science_teams_2020 <- data_science_teams_2020[!duplicated(data_science_teams_2020),]

ds2020_per <- data_science_teams_2020 %>%
  select(size,percentage)

data_science_teams_2019 <- survey2019 %>%
  select(Q6,Q7) %>%
  filter(is.na(Q6) == FALSE) %>%
  filter(is.na(Q7) == FALSE) %>%
  group_by(Q6,Q7) %>%
  summarise(number=n()) %>%
  arrange(Q6)
names(data_science_teams_2019) <- c('size1','data science team','number')
data_science_teams_2019 <- data_science_teams_2019 %>%
  mutate(size = case_when(size1=='0-49 employees' ~ '0-49',size1=='> 10,000 employees' ~ '10,000+',size
data_science_teams_2019 <- data_science_teams_2019[,-1]
data_science_teams_2019 <- data_science_teams_2019 %>%
  filter(`data science team` != '0') %>%
  mutate(per = number/sum(number)) %>%
  group_by(size) %>%
  mutate(total=sum(number),percentage=sum(per)) %>%
  select(size,total,percentage)
data_science_teams_2019 <- data_science_teams_2019[!duplicated(data_science_teams_2019),]

ds2019_per <- data_science_teams_2019 %>%
  select(size,percentage)

ds_per <- merge(ds2019_per,ds2020_per,by = 'size')
names(ds_per) <- c('size','2019','2020')
ds_per <- data.frame(size = rep(ds_per$size,2), year = rep(c('2019','2020'),each = 5),number = c(ds_per
```

```
salary2020 <- survey2020 %>%
  group_by(Q24) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(salary = case_when(Q24=='$0-999' ~ '<10',Q24=='1,000-1,999' ~ '<10',Q24=='2,000-2,999' ~ '<10'
  group_by(salary) %>%
  summarise(value = sum(number)) %>%
  select(salary,value)

salary2019 <- survey2019 %>%
  group_by(Q10) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(salary = case_when(Q10=='$0-999' ~ '<10',Q10=='1,000-1,999' ~ '<10',Q10=='2,000-2,999' ~ '<10'
  group_by(salary) %>%
  summarise(value = sum(number)) %>%
  select(salary,value)

salary2018 <- survey2018 %>%
  group_by(Q9) %>%
  summarise(number=n()) %>%
  arrange(desc(number)) %>%
  na.omit() %>%
  mutate(salary = case_when(Q9=='0-10,000' ~ '<10',Q9=='10-20,000' ~ '10-30',Q9=='20-30,000' ~ '10-30',
  filter(salary != 'NA') %>%
  group_by(salary) %>%
  summarise(value = sum(number)) %>%
  select(salary,value)
salary <- merge(salary2018,salary2019,by = 'salary')
salary <- merge(salary,salary2020,by = 'salary')
names(salary) <- c('salary','2018','2019','2020')
salary <- melt(salary,id = 'salary')
names(salary) <- c('salary','year','number')
```

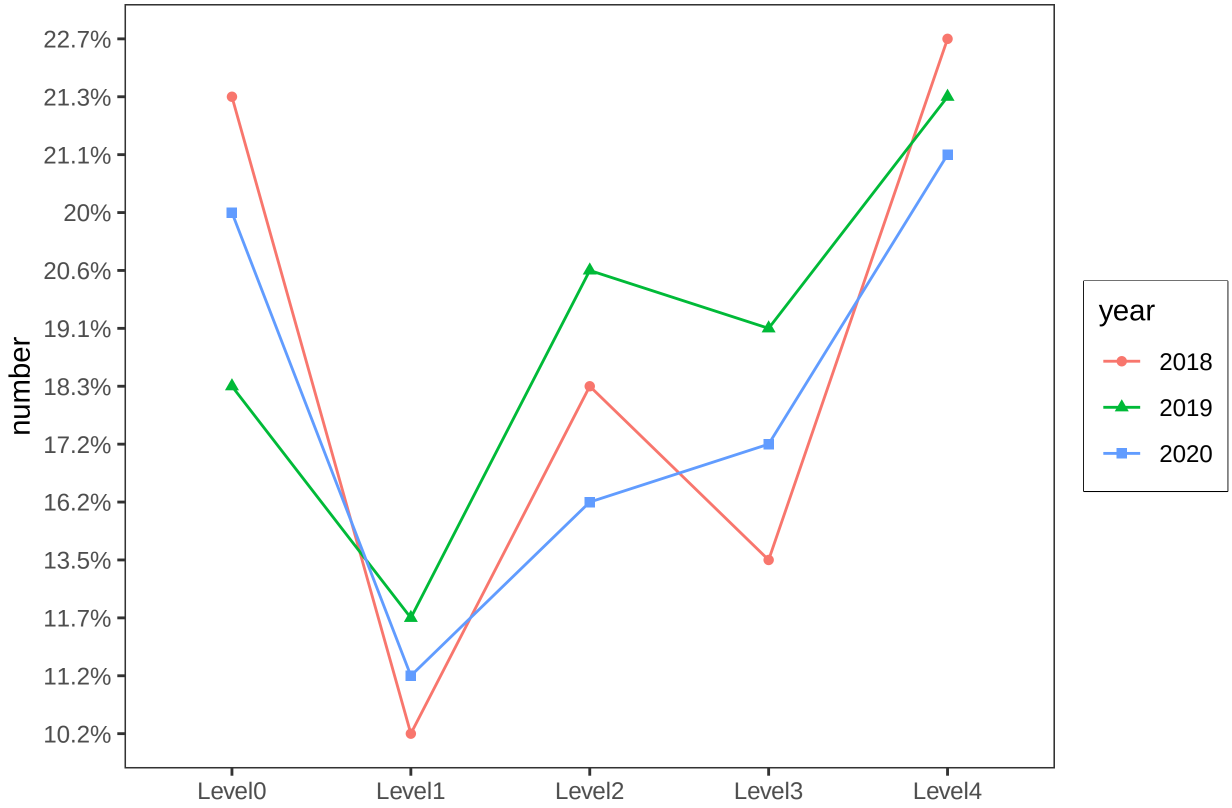
Data Visualization & Conclusion

For data visualization, we draw a bar graph for data science and machine learning to compare the years that people have been used. According to these graphs, they showed that there are more data scientists than machine learning engineers. In addition, from 2019 to 2020, there is an increasing tendency to show that increasing populations are more likely to use both machine learning and data science.

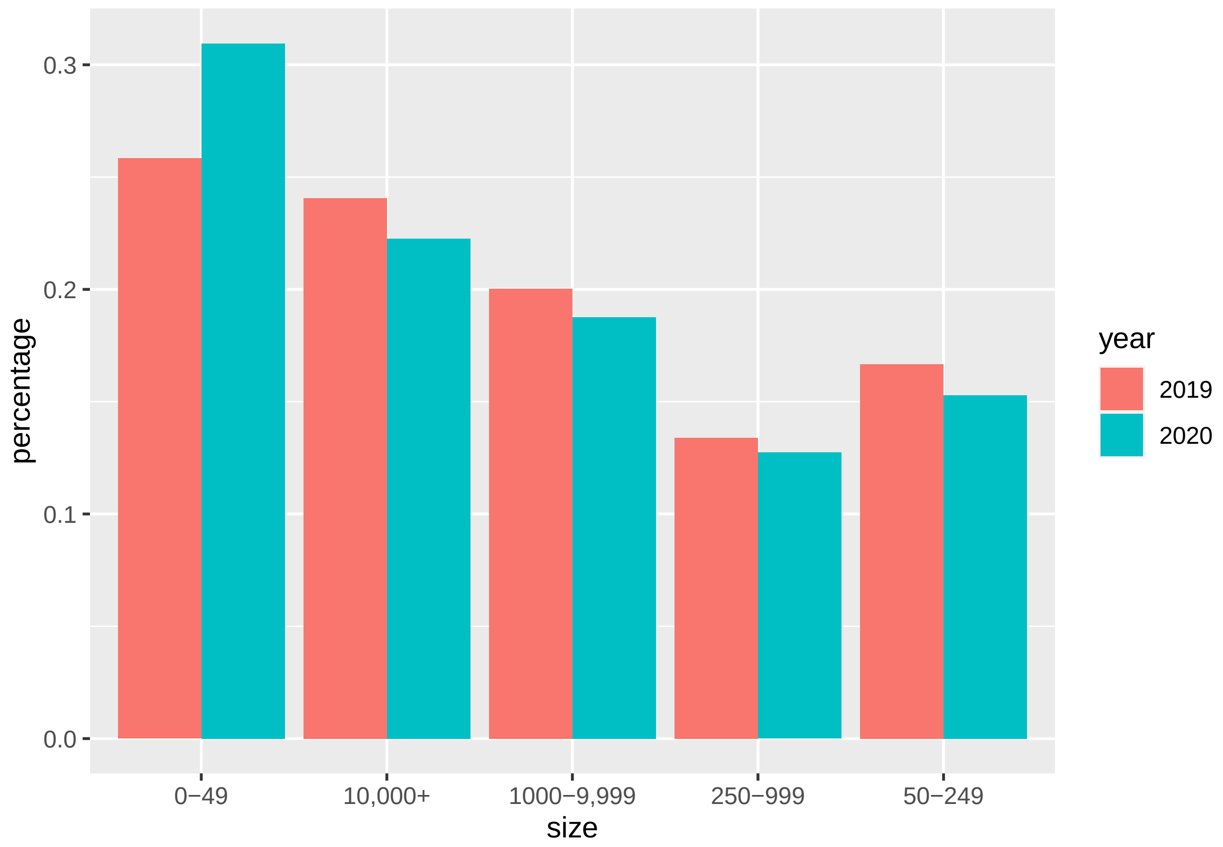
```
ml_incorporation_bargraph <- ml_incorporation %>%
  ggplot(mapping = aes(x = level,fill = year,y = number)) +
  geom_col(position = 'dodge')
ml_incorporation_bargraph
```



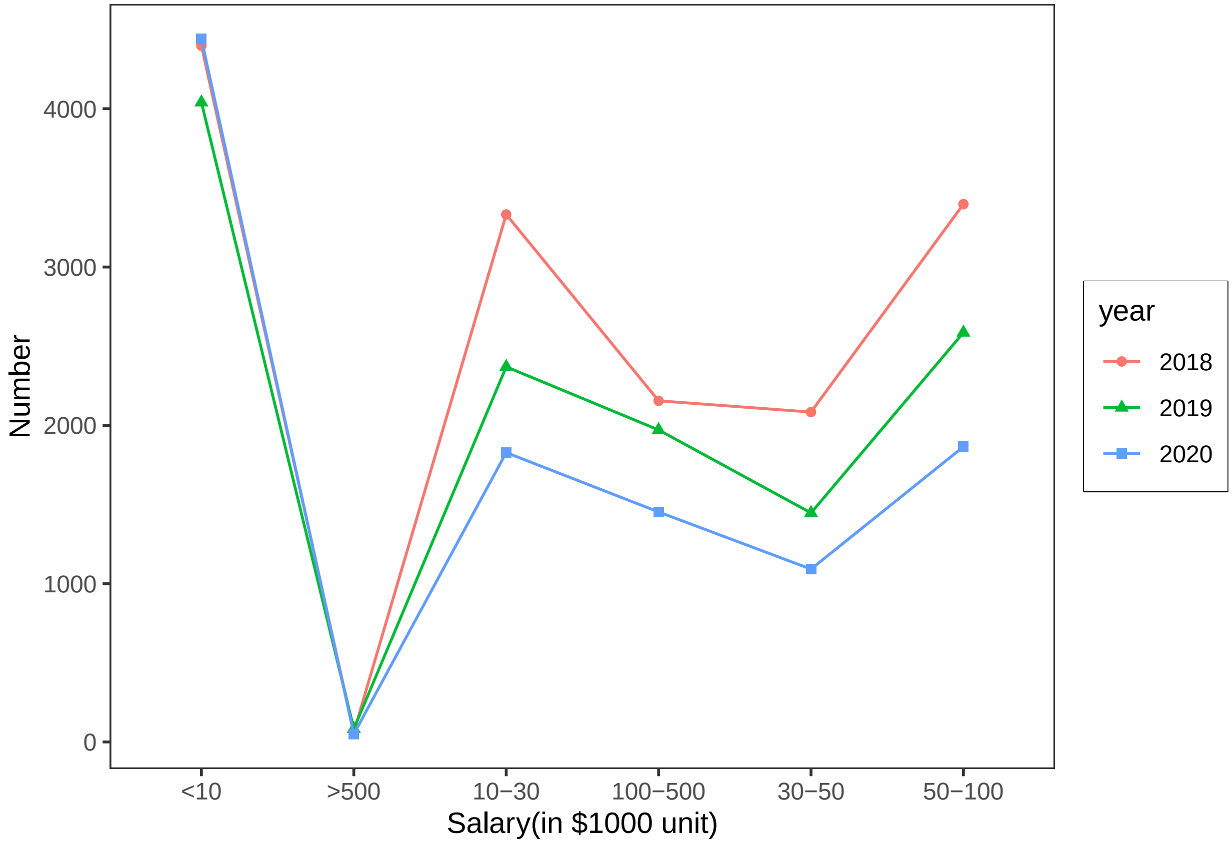
```
ml_incorporation_linegraph <- ggplot(ml_incorporation,aes(x = level,y = number,group = year,color=year,
  geom_point() +
  geom_line() +
  xlab('') +
  ylab("number") +
  theme_bw() +
  theme(panel.grid.major = element_line(colour = NA)) +
  theme(panel.background = element_rect(fill = 'transparent',colour = NA)) +
  theme(panel.grid.minor = element_blank()) +
  theme(legend.box.background = element_rect(colour = 'black'))
ml_incorporation_linegraph
```



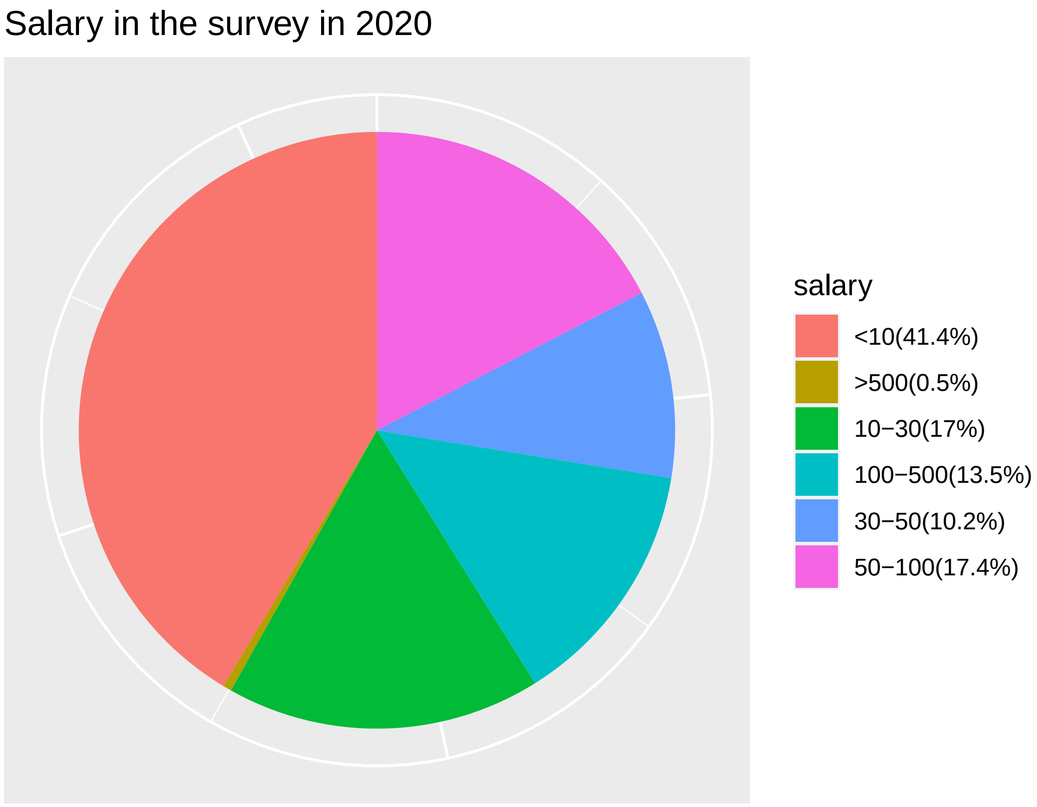
```
ds_per_bargraph <- ds_per %>%
  ggplot(mapping = aes(x = size,fill = year,y = number)) +
  ylab('percentage') +
  geom_col(position = 'dodge')
ds_per_bargraph
```



```r
salary_linegraph <- ggplot(salary,aes(x = salary,y = number,group = year,color=year,shape = year)) +
  geom_point() +
  geom_line() +
  xlab('Salary(in $1000 unit)') +
  ylab("Number") +
  theme_bw() +
  theme(panel.grid.major = element_line(colour = NA)) +
  theme(panel.background = element_rect(fill = 'transparent',colour = NA)) +
  theme(panel.grid.minor = element_blank()) +
  theme(legend.box.background = element_rect(colour = 'black'))
salary_linegraph
```



```
label_value9 <- paste('(', round(salary2020$value/sum(salary2020$value) * 100, 1), '%)', sep = '')
label9 <- paste(salary2020$salary, label_value9, sep = '')
salary_piechart <-
  ggplot(salary2020,aes(x = '',y = value,fill = salary)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  labs(x = '', y = '', title = 'Salary in the survey in 2020') +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_blank()) +
  scale_fill_discrete(breaks = salary2020$salary,labels = label9)
salary_piechart
```



Business Question 9

BQ9:Compare market share of Kaggle within recent years.

Data Wrangling

To compare market share of Kaggle within recent years, we can analyze data from 2018 to 2020 about users' regular notebook products basis and their favorite media sources, and present outputs in graphical form.Finally,'platform_usage_top6' has three columns:*platform*:platform that users usually use;*year*:year the survey was issued;*number*:number of users using this platform and 'media_top7' has three columns:*media*:media that uses usually use;*year*:year the survey was issued;*number*:number of users using this media.

```
subsurvey_2020_platform <- survey2020[,232:243]
list_temp_3 <- list()
for (i in 1:ncol(subsurvey_2020_platform)) {
  list_temp_3 <- append(list_temp_3,table(subsurvey_2020_platform[,i]))
}
names(list_temp_3) <- c('Coursera','edX','Kaggle','DataCamp','Fast.ai','Udacity','Udemy','LinkedIn','Azu
platform_usage_2020 <- t(as.data.frame(list_temp_3))
platform_usage_2020 <- data.frame(platform = rownames(platform_usage_2020),number = platform_usage_2020
platform_usage_2020 <- arrange(platform_usage_2020,desc(number))
```

```
subsurvey_2019_platform <- survey2019[,36:47]
list_temp_2 <- list()
for (i in 1:ncol(subsurvey_2019_platform)) {
  list_temp_2 <- append(list_temp_2,table(subsurvey_2019_platform[,i]))
}
names(list_temp_2) <- c('Udacity','Coursera','edX','DataCamp','DataQuest','Kaggle','Fast.ai','Udemy','Li
platform_usage_2019 <- t(as.data.frame(list_temp_2))
platform_usage_2019 <- data.frame(platform = rownames(platform_usage_2019),number = platform_usage_2019
platform_usage_2019 <- arrange(platform_usage_2019, desc(number))

subsurvey_2018_platform <-survey2018[,292:304]
list_temp_1 <- list()
for (i in 1:ncol(subsurvey_2018_platform)) {
  list_temp_1 <- append(list_temp_1,table(subsurvey_2018_platform[,i]))
}
names(list_temp_1) <- c('Udacity','Coursera','edX','DataCamp','DataQuest','Kaggle','Fast.ai','Google','U
platform_usage_2018 <- t(as.data.frame(list_temp_1))
platform_usage_2018 <- data.frame(platform = rownames(platform_usage_2018),number = platform_usage_2018
platform_usage_2018 <- arrange(platform_usage_2018, desc(number))

platform1 <- platform_usage_2018[1:7,]
platform2 <- platform_usage_2019[1:7,]
platform3 <- platform_usage_2020[1:7,]

platform_usage_top6 <- inner_join(platform1,platform2, by = 'platform')
platform_usage_top6 <- inner_join(platform_usage_top6,platform3, by = 'platform')
names(platform_usage_top6) <- c('platform', 2018,2019,2020)
platform_usage_top6 <- melt(platform_usage_top6, id = 'platform')
names(platform_usage_top6) <- c('platform', 'year', 'number')
```

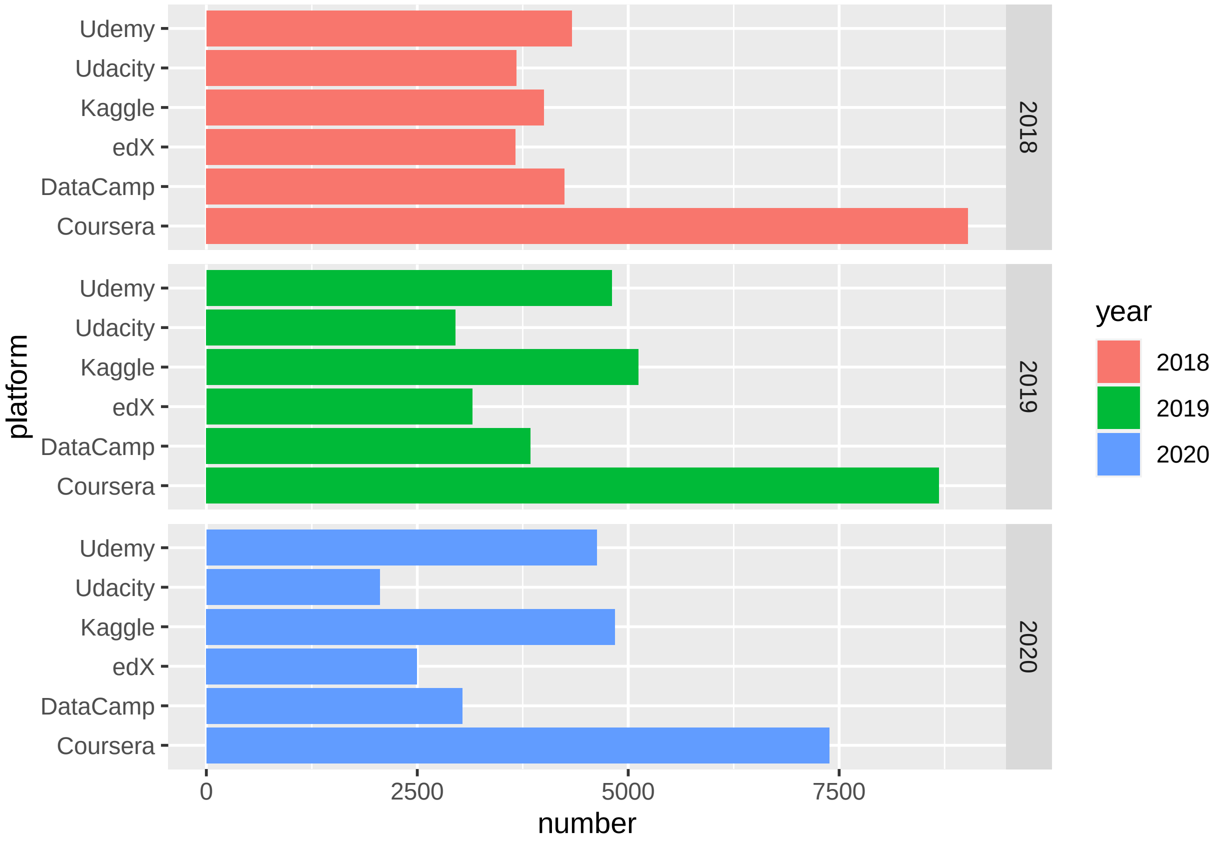
```
subsurvey_2020_media <- survey2020[,245:256]
list_temp_a <- list()
for (i in 1:ncol(subsurvey_2020_media)) {
  list_temp_a <- append(list_temp_a,table(subsurvey_2020_media[,i]))
}
names(list_temp_a) <- c('Twitter','Email','Reddit','Kaggle','Course forums','Youtube','Podcasts','Blogs
media_2020 <- t(as.data.frame(list_temp_a))
media_2020 <- data.frame(media = rownames(media_2020),number = media_2020[,1])
media_2020 <- arrange(media_2020,desc(number))

subsurvey_2019_media <- survey2019[,23:34]
list_temp_b <- list()
for (i in 1:ncol(subsurvey_2019_media)) {
  list_temp_b <- append(list_temp_b,table(subsurvey_2019_media[,i]))
}
names(list_temp_b) <- c('Twitter','Hacker','Reddit','Kaggle','Course forums','Youtube','Podcasts','Blog
media_2019 <- t(as.data.frame(list_temp_b))
media_2019 <- data.frame(media = rownames(media_2019),number = media_2019[,1])
media_2019 <- arrange(media_2019,desc(number))

media1 <- media_2019[1:8,]
media2 <- media_2020[1:8,]
```

```
media_top7 <- inner_join(media1,media2, by = 'media')
names(media_top7) <- c('media', 2019,2020)
media_top7 <- melt(media_top7, id = 'media')
names(media_top7) <- c('media', 'year', 'number')
```

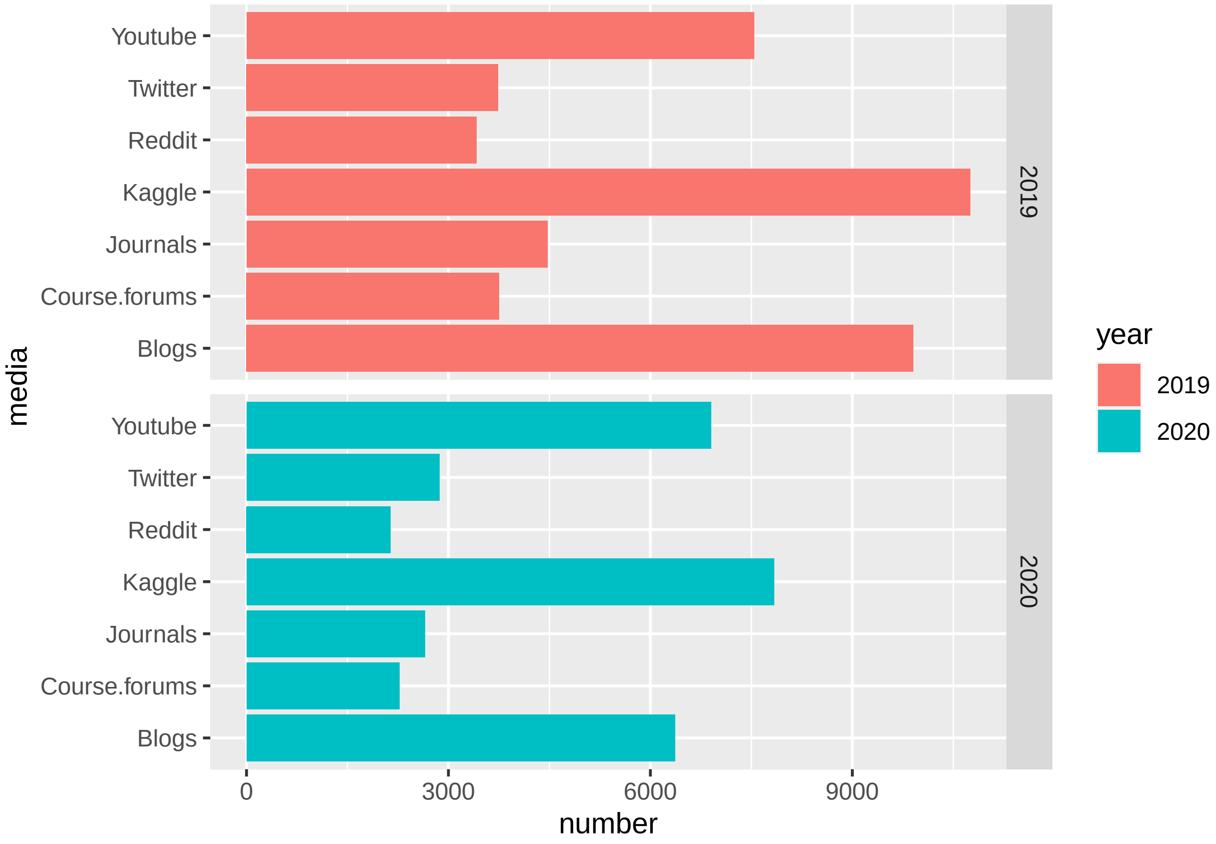
Data Visualization & Conclusion

According to the top6-platform-usage bar chart, we can see that Coursera has been top1 for continuous 3 years, but its market share keeps decreasing. Kaggle increased from top3 to top2 in 2020, and its usage is steady. Usage of Udemy, Udacity, edX, DataCamp are reducing.Since there were no related questions in the survey in 2018, we skipped it. According to the top7-media-sharing bar chart, we can find that Kaggle has the largest market share, Youtube and Blogs are at the 2nd and 3rd place in 2020. The disparity between Kaggle and Youtube is decreasing. All the media usages keep reducing.To increase market share, Kaggle could regularly update existing functions and also enhance the functionality of its products. In addition, Kaggle might increase advertising volume in social media applications.

```
platform_bargraph <- platform_usage_top6 %>%
  ggplot(mapping = aes(x = platform,fill = year,y = number)) +
  geom_col(position = 'dodge') +
  coord_flip() +
  facet_grid(year~.)
platform_bargraph
```



```
media_bargraph <- media_top7 %>%
  ggplot(mapping = aes(x = media,fill = year,y = number)) +
  geom_col(position = 'dodge') +
  coord_flip() +
  facet_grid(year~.)
media_bargraph
```


Youtube
Twitter
Reddit
Kaggle
Journals
Course.forums
Blogs
media
number
year