EXPLORATORY DATA ANALYSIS

-SALARY ANALYSIS

**DATA SOURCE-** DATA.WORLD

**DATASET DESCRIPTION:**

The dataset tells about the details of the data science jobs locating in foreign countries, which has the attributes of 6 and has rows about 1200 .The detail of attribute are mentioned as fallows,

Job Title: Profile of the person

Job type: Is the person working in the internship or full time

Experience level: based on their working years

Location: Location of the person where he works

Salary currency: currency of salary they earning

Salary: Earnings of a person based on their category

**ASSUMPTION:**

* First, let us assume the salary is high on the job type
* Form the above dataset, we can analyze the person salary based on their job title
* Let, the salary may be falls high in range who works in the mid level or as an senior experience level
* From the above assume, we subset the data to get a clear idea about the person who having their salary is high

**PROBLEM:**

Here, we are going to find the salary segments according to their working category.

**SOURCE CODE:**

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(lattice)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2  
## ──

## ✔ ggplot2 3.4.1 ✔ purrr 1.0.1  
## ✔ tibble 3.1.8 ✔ stringr 1.5.0  
## ✔ tidyr 1.3.0 ✔ forcats 1.0.0  
## ✔ readr 2.1.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(plyr)

## ------------------------------------------------------------------------------  
## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)  
## ------------------------------------------------------------------------------  
##   
## Attaching package: 'plyr'  
##   
## The following object is masked from 'package:purrr':  
##   
## compact  
##   
## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

df <- read.csv('datascience\_salaries.csv')  
head(df)

## X job\_title job\_type experience\_level location salary\_currency  
## 1 0 Data scientist Full Time Senior New York City USD  
## 2 2 Data scientist Full Time Senior Boston USD  
## 3 3 Data scientist Full Time Senior London USD  
## 4 4 Data scientist Full Time Senior Boston USD  
## 5 5 Data scientist Full Time Senior New York City USD  
## 6 6 Data scientist Full Time Senior London USD  
## salary  
## 1 149000  
## 2 120000  
## 3 68000  
## 4 120000  
## 5 149000  
## 6 68000

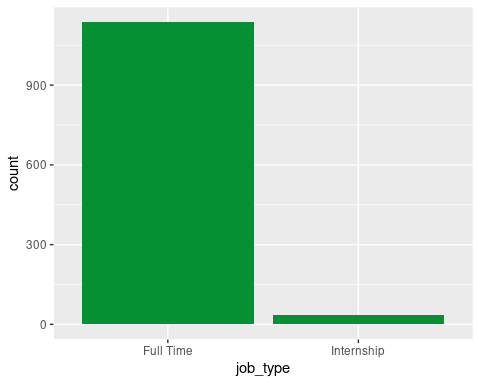
glimpse(df)

## Rows: 1,171  
## Columns: 7  
## $ X <int> 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 17, 18, 20, 22…  
## $ job\_title <chr> "Data scientist", "Data scientist", "Data scientist",…  
## $ job\_type <chr> "Full Time", "Full Time", "Full Time", "Full Time", "…  
## $ experience\_level <chr> "Senior", "Senior", "Senior", "Senior", "Senior", "Se…  
## $ location <chr> "New York City", "Boston", "London", "Boston", "New Y…  
## $ salary\_currency <chr> "USD", "USD", "USD", "USD", "USD", "USD", "USD", "USD…  
## $ salary <int> 149000, 120000, 68000, 120000, 149000, 68000, 69000, …

colSums(is.na(df))

## X job\_title job\_type experience\_level   
## 0 0 0 0   
## location salary\_currency salary   
## 0 0 0

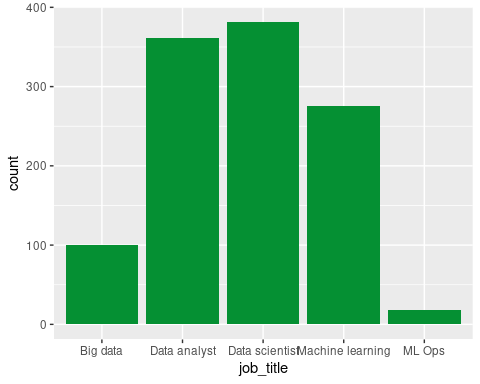
df$salary\_currency[df$salary\_currency=='EUR']= 'USD'  
df$salary\_currency[df$salary\_currency== 'GBP']= 'USD'  
  
df %>% ggplot(aes(job\_type))+  
 geom\_bar(stat="Count",fill="#059033")



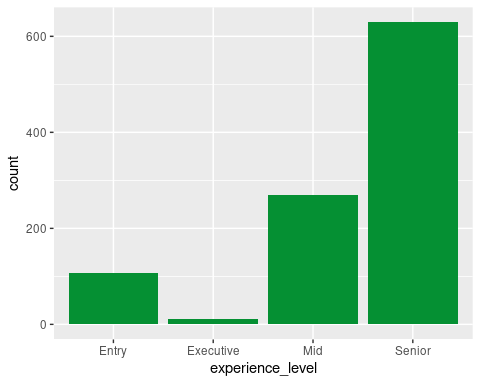
df2=subset(df,job\_type=="Full Time")  
head(df2)

## X job\_title job\_type experience\_level location salary\_currency  
## 1 0 Data scientist Full Time Senior New York City USD  
## 2 2 Data scientist Full Time Senior Boston USD  
## 3 3 Data scientist Full Time Senior London USD  
## 4 4 Data scientist Full Time Senior Boston USD  
## 5 5 Data scientist Full Time Senior New York City USD  
## 6 6 Data scientist Full Time Senior London USD  
## salary  
## 1 149000  
## 2 120000  
## 3 68000  
## 4 120000  
## 5 149000  
## 6 68000

df2 %>% ggplot(aes(job\_title))+  
 geom\_bar(stat="Count",fill="#059033")



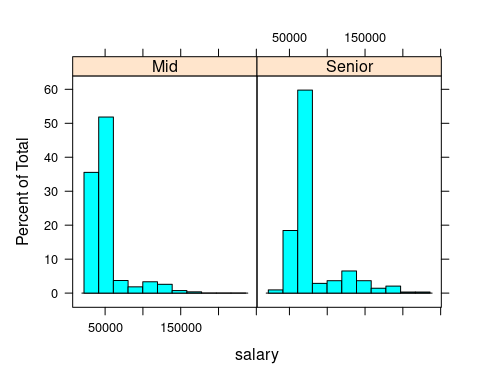
d3=subset(df2,job\_title=="Data analyst")  
d4=subset(df2,job\_title=="Data scientist")  
d5=subset(df2,job\_title=="Machine learning")  
  
d6=rbind(d3,d4,d5)  
d6 %>% ggplot(aes(experience\_level))+  
 geom\_bar(stat="Count",fill="#059033")



d7=subset(d6,experience\_level=='Mid')  
d8=subset(d6,experience\_level=="Senior")  
d9=rbind(d7,d8)  
  
count(d9,'experience\_level')

## experience\_level freq  
## 1 Mid 270  
## 2 Senior 629

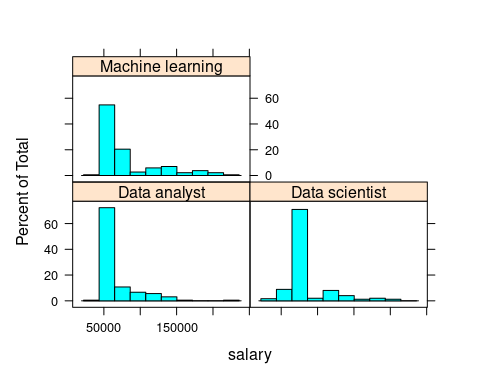
histogram(~salary|experience\_level,data=d9)



d10=subset(d9,experience\_level=="Senior")  
  
count(d10,"job\_title")

## job\_title freq  
## 1 Data analyst 195  
## 2 Data scientist 248  
## 3 Machine learning 186

histogram(~salary|job\_title,data=d10)



head(d10)

## X job\_title job\_type experience\_level location salary\_currency  
## 395 500 Data analyst Full Time Senior Copenhagen USD  
## 397 503 Data analyst Full Time Senior Stuttgart USD  
## 399 506 Data analyst Full Time Senior Berlin USD  
## 401 509 Data analyst Full Time Senior Jakarta USD  
## 402 510 Data analyst Full Time Senior Chicago USD  
## 403 512 Data analyst Full Time Senior Lexington USD  
## salary  
## 395 62000  
## 397 62000  
## 399 63000  
## 401 63000  
## 402 63000  
## 403 63000

**INFERENCE:**

* + Generally, the salary issued to the people who has high experience on their working field
  + Majority of the higher salary are lies under the Machine Learning job
  + So, the person who are aiming for higher salary, the can prefer the Machine Learning jobs

**INSIGHTS:**

* + By comparing both experience level of mid and senior level histogram analysis, the distribution shows that the senior level has the more right tailed compared to the mid-level experience of person
  + From the histogram analysis, we can conclude that most of the people who are working as a senior level of experience having the salary more
  + Since the histogram of job title, Machine Learning having their high paying salary compared to other type of jobs
  + Average salary of above 50000 is earned by the data science jobs. But especially the Data Scientist and Machine Learning having their salary more than 150000

BY

DHARSHANRAJ S

22CSEG06

Ist M.sc DATA ANALYICS