Data science job salary

ilke

Veri setine erişmek için data-job-salary (https://www.kaggle.com/datasets/milanvaddoriya/data-science-job-salary) sitesini ziyaret edebilirsiniz.

1.Kütüphanelerin yüklenmesi

```
library(ggplot2)
library(tidyverse)
library(lubridate)
library(readxl)
library(funModeling)
library(gridExtra)
library(magrittr)
library(scales)
library(plotrix)
library(RColorBrewer)
library(readr)
library(maps)
library(highcharter)
library(dplyr)
library(tidyverse)
library(magrittr)
library(DataExplorer)
library(maps)
library(plotly)
library(DT)
library(tidytext)
library(gridExtra)
library(readxl)
library(ggplot2)
library(dplyr)
library(plotly)
library(tidyr)
library(d3Tree)
```

2. Verisetini yükleme

```
getwd()
```

```
## [1] "/home/ilke"
```

```
setwd("/home/ilke/Downloads")

df<- read.csv("datascience_salaries (1).csv",sep=",", header=TRUE,stringsAsFactors = FALSE)
```

3. Keşifci Veri Analizi(Veriseti genel yapısı hakkında bilgi edinme)

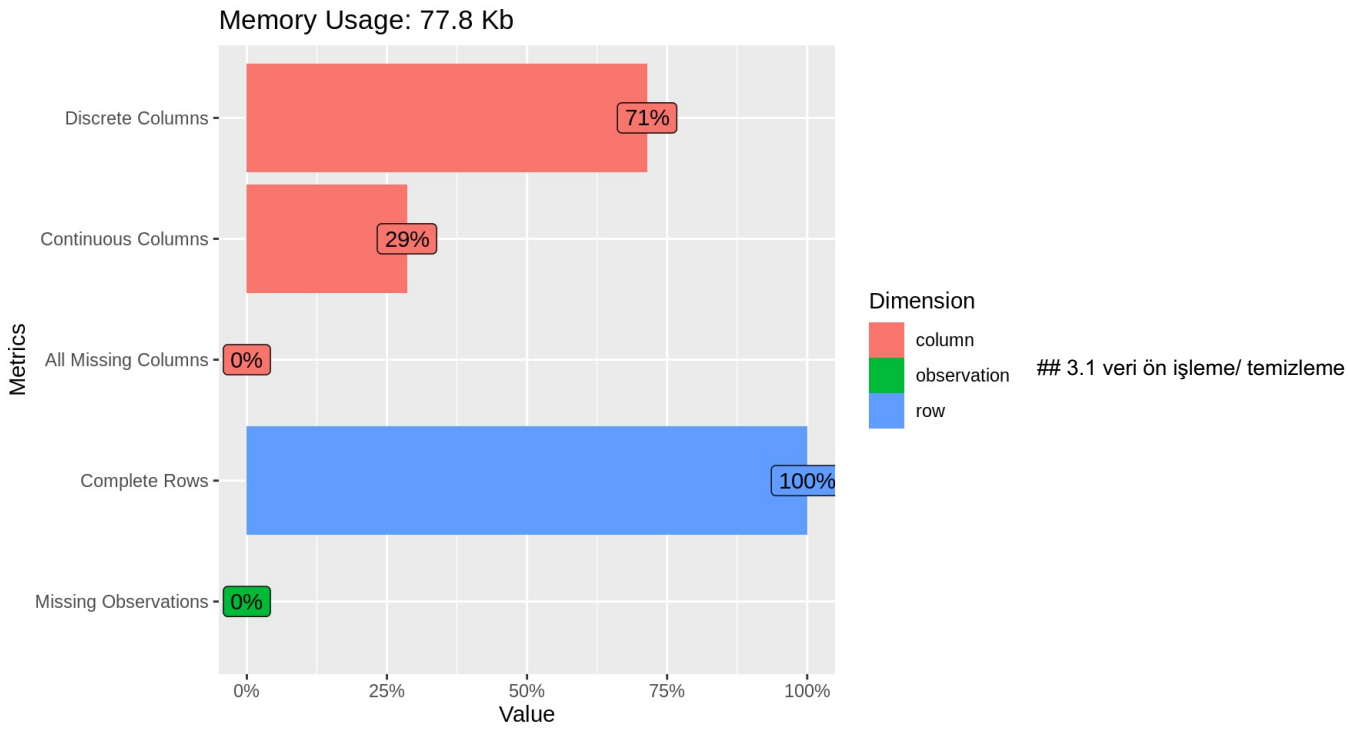
```
glimpse(df)
```

```
## 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, …
```

```
summary(df)
```

```
##        X           job_title           job_type         experience_level
##  Min.   :   0.0   Length:1171        Length:1171        Length:1171
##  1st Qu.: 364.5   Class :character   Class :character   Class :character
##  Median : 815.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 931.6
##  3rd Qu.:1504.5
##  Max.   :2259.0
##    location         salary_currency        salary
##  Length:1171        Length:1171        Min.   : 30000
##  Class :character   Class :character   1st Qu.: 45000
##  Mode  :character   Mode  :character   Median : 63000
##                                        Mean   : 64836
##                                        3rd Qu.: 68000
##                                        Max.   :228000
```

```
plot_intro(df)
```



3.1 veri ön işleme/ temizleme

```
## [1] 0
```

```
df$X <- NULL   #gereksiz satır silme
```

```
#%99unun maaşı USD türünden girili. Euro ve diğer olanları çıkardım.
df <-df %>%
  filter(salary_currency == "USD")
```

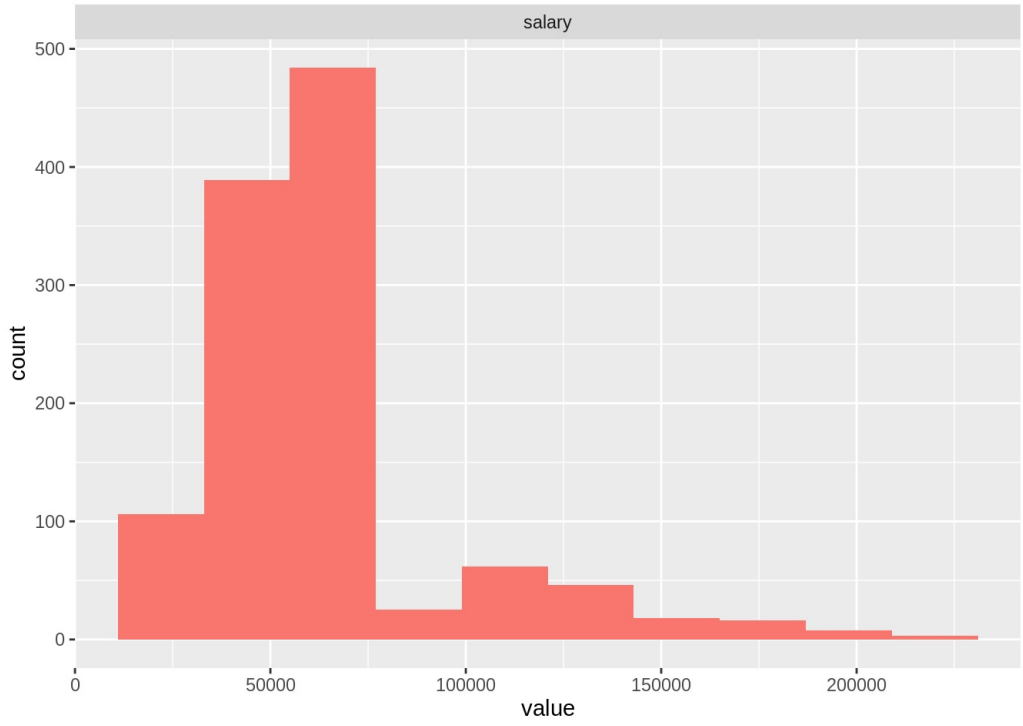
3.2 Genel istatistik/grafik

3.1.1 Sürekli Değişkenlerin Özet Bazı İstatistikleri

```
profiling_num(df)
```

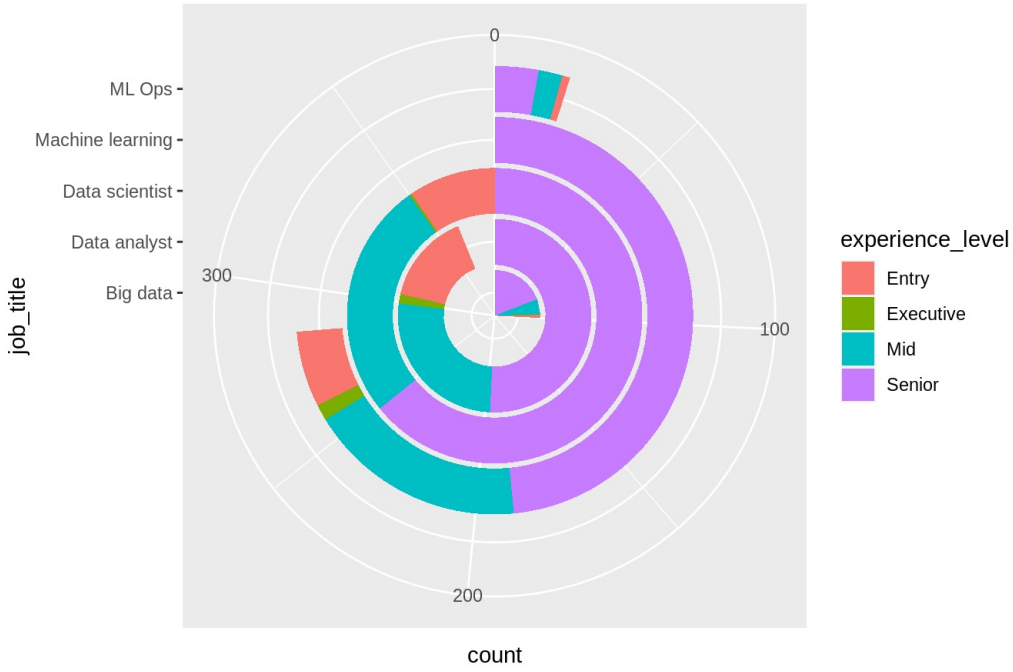
3.1.2 Genel Histogram

```
plot_num(df)  #genel histogram
```




```
#veri setiyle ilgili genel bir önizleme için
d3tree(list(root = df2tree(rootname ='title',
                           struct = as.data.frame(df)),
                           layout = 'collapse'))
```

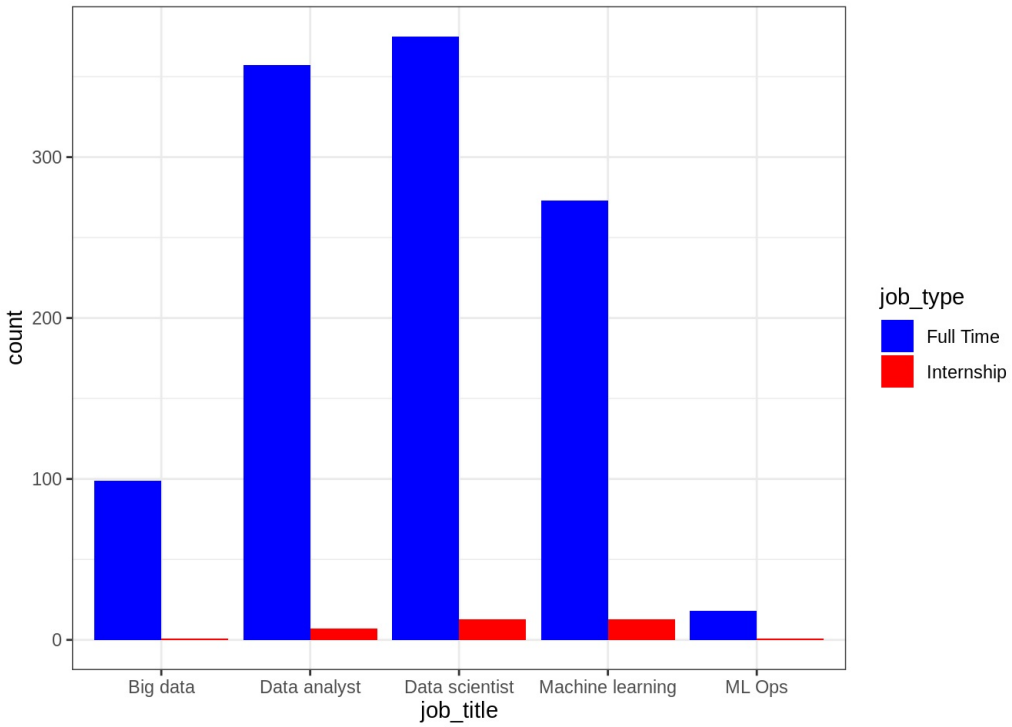

```
ggplot(df,aes(job_title, fill=experience_level))+
  geom_bar()+
  coord_polar(theta = "y")
```



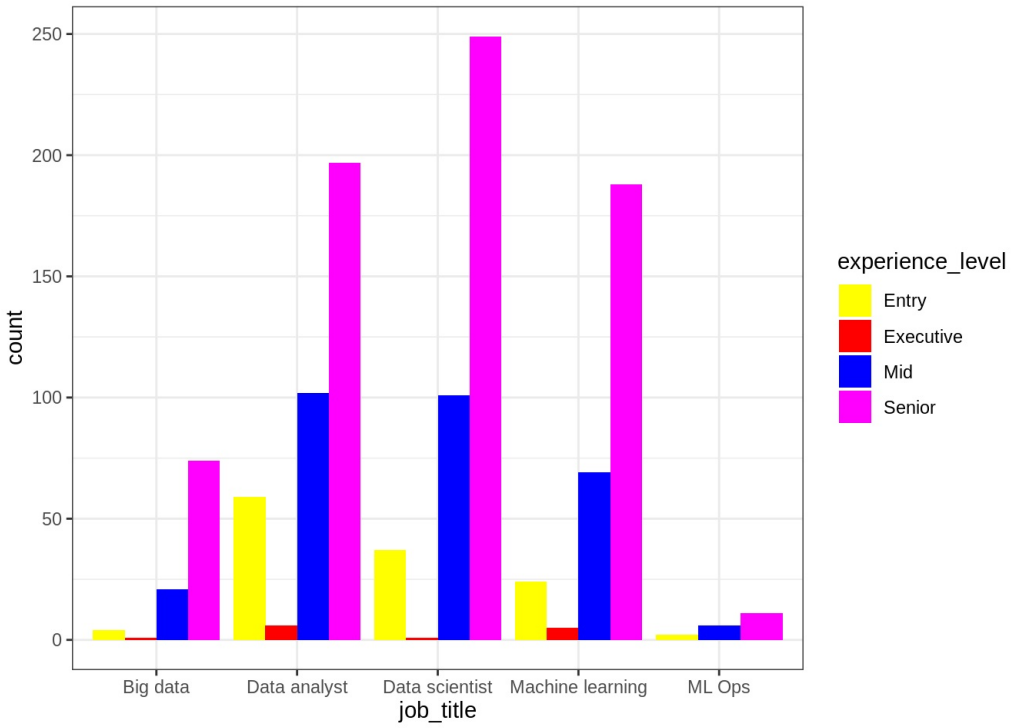

```
title_count<- df %>% group_by(job_title) %>% tally() %>% arrange(n, decreasing=T)
title_count
```

```
## # A tibble: 5 × 2
##   job_title            n
##   <chr>            <int>
## 1 ML Ops              19
## 2 Big data           100
## 3 Machine learning   286
## 4 Data analyst       364
## 5 Data scientist     388
```

```
#mesleklerde intern dağılımı nasıl?
#Big data'da intern olarak çalışan yok
ggplot(data = df) +
  geom_bar(mapping = aes(x = job_title, fill = job_type), position = "dodge") + scale_fill_manual(values = c("blue"
,"red"))+
  theme_bw()
```




```
#mesleklere göre deneyimleri görebileceğimiz grafik
ggplot(data = df) +
  geom_bar(mapping = aes(x = job_title, fill = experience_level), position = "dodge") + scale_fill_manual(values =
c("yellow", "red", "blue","magenta"))+
  theme_bw()
```



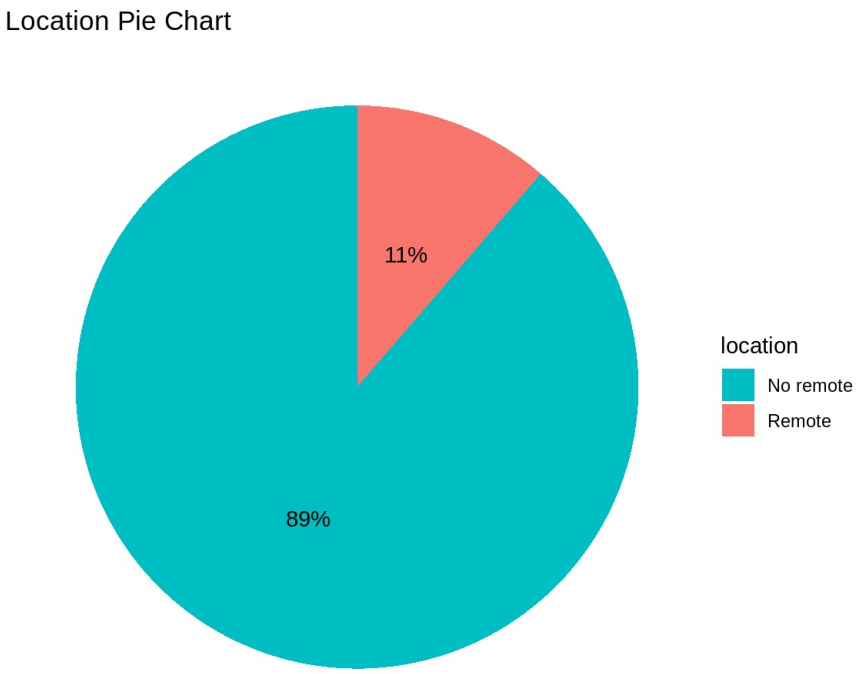

```
#verisetine çalışma yeri olarak remote/no remote ifade edecek şekilde remote=0, no remote=1 olacak şekilde yeni bir
sütuna eklendi.
ds <- df %>% mutate(remote = ifelse(grepl("remote",tolower(location) ), "1", "0"))
```

```
#131 kişi remote olarak çalışmaktadır.
remote_counts <- count(ds, remote)
remote_counts
```

```
##   remote    n
## 1      0 1026
## 2      1  131
```

```
write.csv(ds, file = "/home/ilke/Downloads/ödev.csv", row.names = FALSE)
#remote eklenmiş halini kaydetme
```

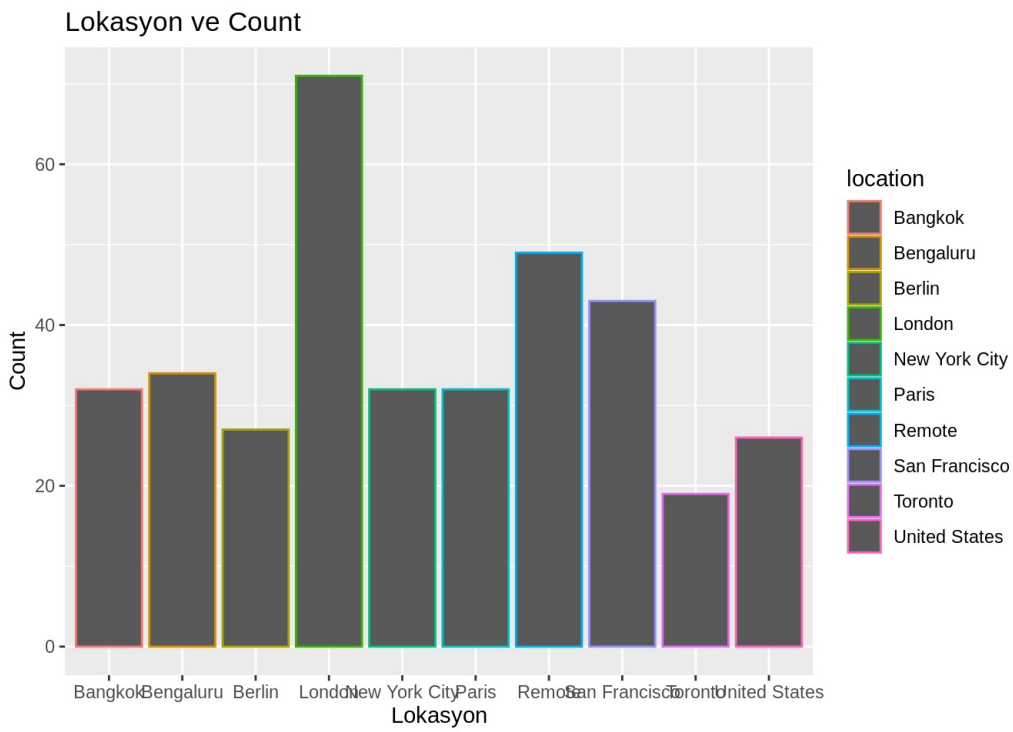
```
# çalışanların ½11'i remote olarak çalışmaktadır.
data <- c(131, 1026)
data_percent <- prop.table(data) * 100
library(ggplot2)
data_df <- data.frame(remote = c("Remote", "No remote"), count = c(131, 1026))

ggplot(data = data_df, aes(x = "", y = count, fill = remote)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  ggtitle("Location Pie Chart") +
  scale_fill_manual(values = c("#00BFC4", "#F8766D")) +
  theme_void() +
  labs(fill = "location", title = "Location Pie Chart") +
  geom_text(aes(label = paste0(round(data_percent), "%")), position = position_stack(vjust = 0.5))
```




```
#en çok çalışılan lokasyonlara bakalım
top_10_locations <- df %>% group_by(location) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>%
  top_n(10)
top_10_locations
```

```
## # A tibble: 10 × 2
##    location      count
##    <chr>         <int>
##  1 London           71
##  2 Remote           49
##  3 San Francisco    43
##  4 Bengaluru        34
##  5 Bangkok          32
##  6 New York City    32
##  7 Paris            32
##  8 Berlin           27
##  9 United States    26
## 10 Toronto          19
```

```
ggplot(data = top_10_locations, aes(x = location, y = count, colour = location)) +
  geom_bar(stat = "identity") +
  ggtitle("Lokasyon ve Count") +
  xlab("Lokasyon") +
  ylab("Count")
```




```
#Çalışma tipine göre sıklıklar ve maaş istatistikleri
salary_by_jobtype <- df %>% group_by(job_type) %>%
  summarise(count= n(),
            min_salary=min(salary),
            max_salary=max(salary),
            mean_salary=mean(salary)) %>%
  arrange(desc(mean_salary)) %>%
  mutate(rank = rank(-mean_salary))
salary_by_jobtype
```

```
## # A tibble: 2 × 6
##   job_type   count min_salary max_salary mean_salary  rank
##   <chr>      <int>      <int>      <int>       <dbl> <dbl>
## 1 Full Time   1122      30000     228000      65360.     1
## 2 Internship    35      30000     135000      54629.     2
```

```
#mesleğe göre meslek sıklıkları ve maaş istatistikleri
salary_by_job <- df %>% group_by(job_title) %>%
  summarise(count= n(),
            min_salary=min(salary),
            max_salary=max(salary),
            mean_salary=mean(salary),
            median_salary=median(salary)) %>%
  arrange(desc(mean_salary)) %>%
  mutate(rank = rank(-mean_salary))
salary_by_job
```

```
## # A tibble: 5 × 7
##   job_title        count min_salary max_salary mean_salary median_salary  rank
##   <chr>            <int>      <int>      <int>       <dbl>         <dbl> <dbl>
## 1 ML Ops              19      35000     228000      81263.         62000     1
## 2 Machine learning   286      30000     228000      68503.         51000     2
## 3 Data scientist     388      30000     196000      67454.         68000     3
## 4 Data analyst       364      30000     225000      60973.         63000     4
## 5 Big data           100      30000     178000      57440          53500     5
```

```
#deneyim seviyesine göre maaş istatistikleri
salary_by_exp <- df %>% group_by(experience_level) %>% summarise(count= n(),
            min_salary=min(salary),
            max_salary=max(salary),
            mean_salary=mean(salary)) %>%
  arrange(desc(mean_salary)) %>%
  mutate(rank = rank(-mean_salary))
salary_by_exp
```

```
## # A tibble: 4 × 6
##   experience_level count min_salary max_salary mean_salary  rank
##   <chr>            <int>      <int>      <int>       <dbl> <dbl>
## 1 Executive           13      41000     175000      76077.     1
## 2 Senior             719      30000     228000      75403.     2
## 3 Mid                299      30000     160000      51813.     3
## 4 Entry              126      30000     140000      36111.     4
```

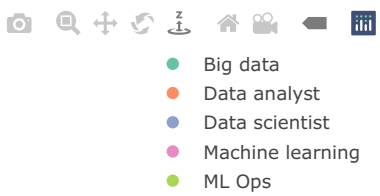
```
#meslek ve çalışma tiplerine göre maaş ortalamaları
a <- df %>%
  select(job_title, salary, job_type) %>%
  group_by(job_title, job_type) %>%
  summarise(avg_income = mean(salary))
a
```

```
## # A tibble: 10 × 3
## # Groups:   job_title [5]
##    job_title        job_type   avg_income
##    <chr>            <chr>           <dbl>
##  1 Big data         Full Time      57717.
##  2 Big data         Internship     30000
##  3 Data analyst     Full Time      61210.
##  4 Data analyst     Internship     48857.
##  5 Data scientist   Full Time      67936
##  6 Data scientist   Internship     53538.
##  7 Machine learning Full Time      69168.
##  8 Machine learning Internship     54538.
##  9 ML Ops           Full Time      78278.
## 10 ML Ops           Internship    135000
```

```
#remote çalışma durumuna göre maaş istatistikleri
salary_by_remote <- ds %>% group_by(remote) %>%
  summarise(count= n(),
            min_salary=min(salary),
            max_salary=max(salary),
            mean_salary=mean(salary)) %>%
  arrange(desc(mean_salary)) %>%
  mutate(rank = rank(-mean_salary))
salary_by_remote
```

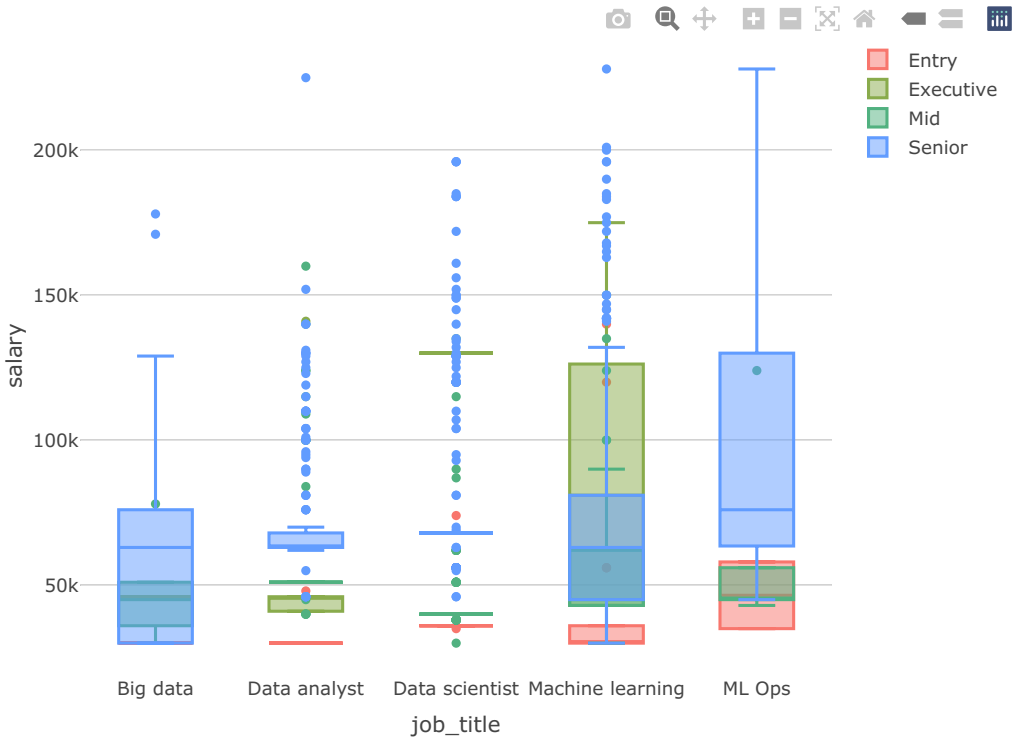
```
## # A tibble: 2 × 6
##   remote count min_salary max_salary mean_salary  rank
##   <chr>  <int>      <int>      <int>       <dbl> <dbl>
## 1 1        131      30000     225000      73008.     1
## 2 0       1026      30000     228000      64018.     2
```

```
#meslek ve deneyimlere göre maaşın saçılım grafiği
saçılım<- plot_ly(ds , x= ~experience_level , y = ~salary , z= ~job_title , color = ~job_title , text = ~salary ) %
>%
  add_markers() %>%
  layout(
    scene = list(xaxis = list(title = "LEVEL"),
                 yaxis= list(title = "SALARY"),
                 zaxis = list (title = "TİTLE"))
  )
saçılım
```

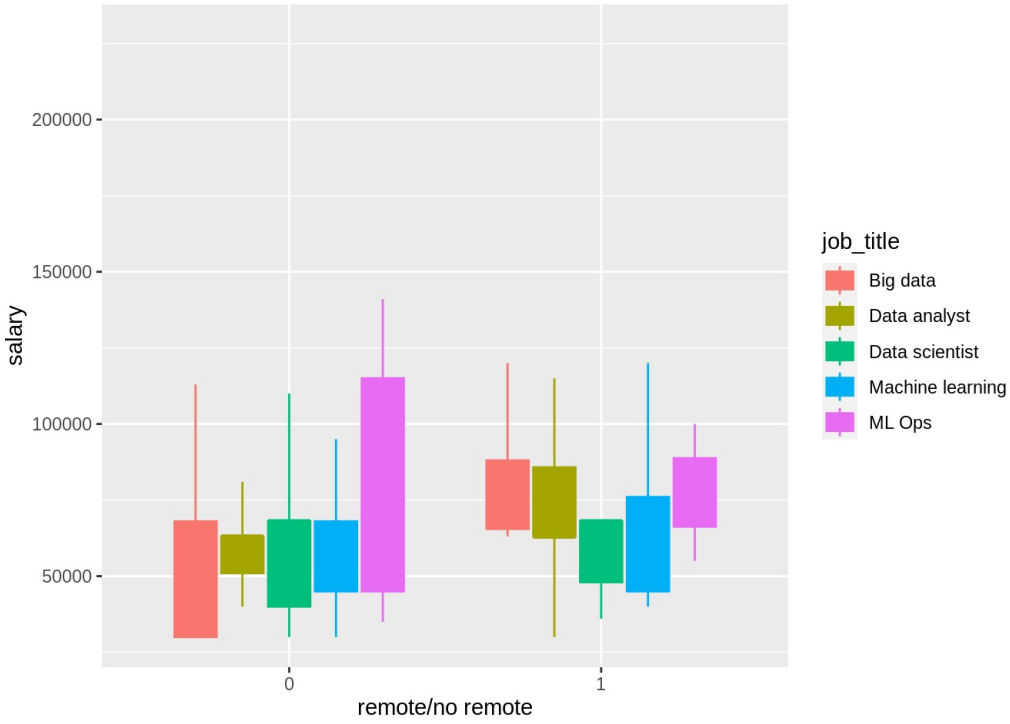


WebGL is not supported by your browser - visit https://get.webgl.org for more info

```
p1 <- plot_ly(ds, y = ~salary, color = ~experience_level, colors = c('#F8766D', '#00BA38','#619CFF'), legendgroup = ~job_title) %>%
  add_boxplot(x = ~job_title)

p1
```




```
#mesleklere göre remote/no remeto maaş
ggplot(ds, aes(x = remote, y = salary, fill =job_title, colour = job_title)) +
geom_boxplot(outlier.colour = NA) + xlab("remote/no remote") + ylab("salary")
```


200000
150000
100000
50000
salary
0
1
remote/no remote
job_title
Big data
Data analyst
Data scientist
Machine learning
ML Ops