

# DATA ANALYSIS USING R

In this project I'm analysing a data set of credit card payment which was cleaned using Pandas

*Loading required libraries*

```
library(readxl)
library(magrittr)
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## vforcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyrr    1.3.0
## v purrr    1.0.1
## -- Conflicts ----- tidyverse_conflicts() --
## x tidyrr::extract() masks magrittr::extract()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
library(ggplot2)
```

*Reading Excel file with Credit Card payment data*

```
creditCard = read_excel("C:/DATA_SET/DEFAULT_TO_CREDIT.xlsx")
```

*Here we can see what is the total bills for each of the categories*

```
creditCard %>%
  group_by(CLASSIFICATION) %>%
  drop_na() %>%
  summarise(TOTAL_BILL = sum(TOTAL_BILL))

## # A tibble: 4 x 2
##   CLASSIFICATION TOTAL_BILL
##   <chr>           <dbl>
## 1 DEFULTER        7236736002
## 2 STANDARD         248505764
## 3 SUBSTANDARD 1   166564155
## 4 SUBSTANDARD 2   444044215
```

*Here we can see what is the total bills paid by each of the categories*

```
creditCard %>%
  group_by(CLASSIFICATION) %>%
  drop_na() %>%
  summarise(TOTAL_BILL_PAID = sum(TOTAL_BILL_PAID))

## # A tibble: 4 x 2
##   CLASSIFICATION TOTAL_BILL_PAID
##   <chr>           <dbl>
## 1 DEFULTER        422411982
## 2 STANDARD         272730575
## 3 SUBSTANDARD 1   101962424
## 4 SUBSTANDARD 2   152436796
```

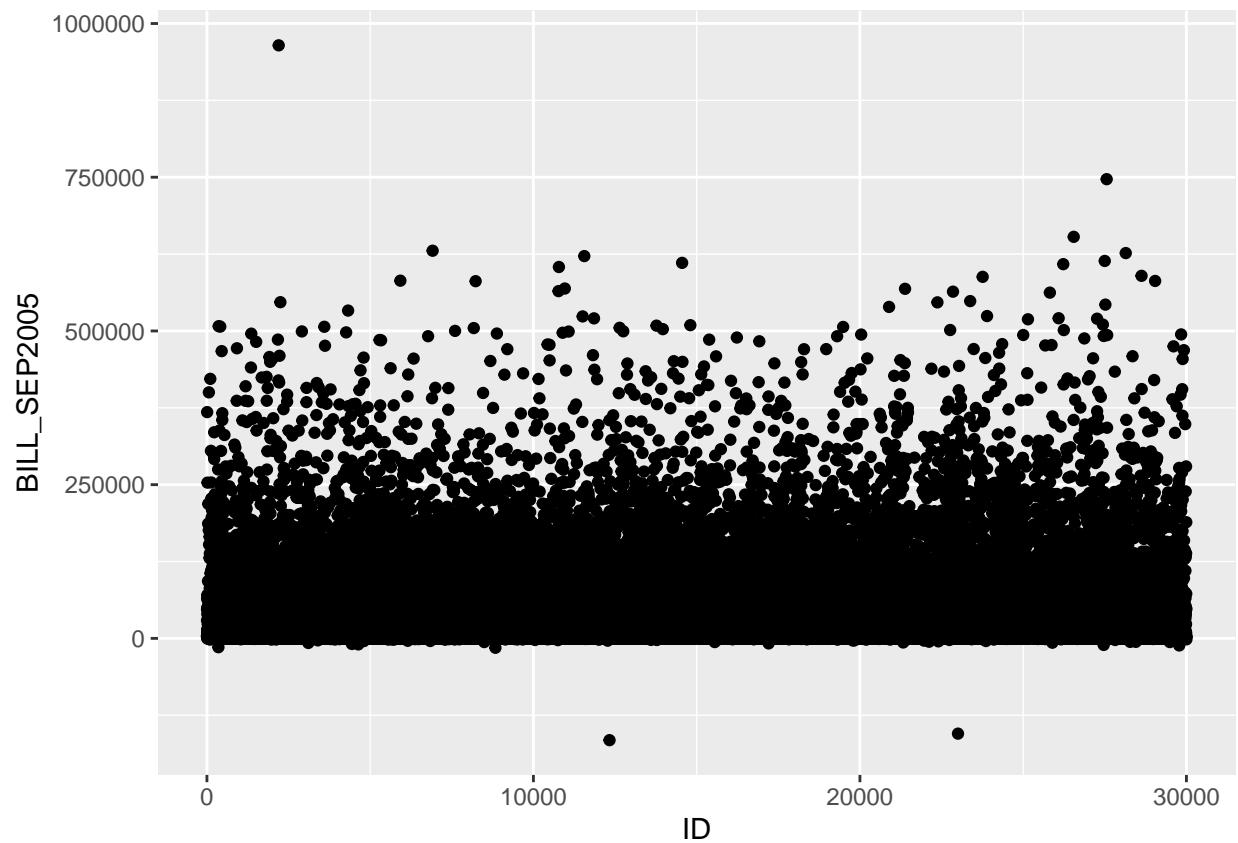
Here we get the total count for each of the classifications

```
ggplot(data = creditCard) +  
  geom_bar(mapping = aes(x=CLASSIFICATION,fill=CLASSIFICATION)) +  
  facet_wrap(~CLASSIFICATION)
```



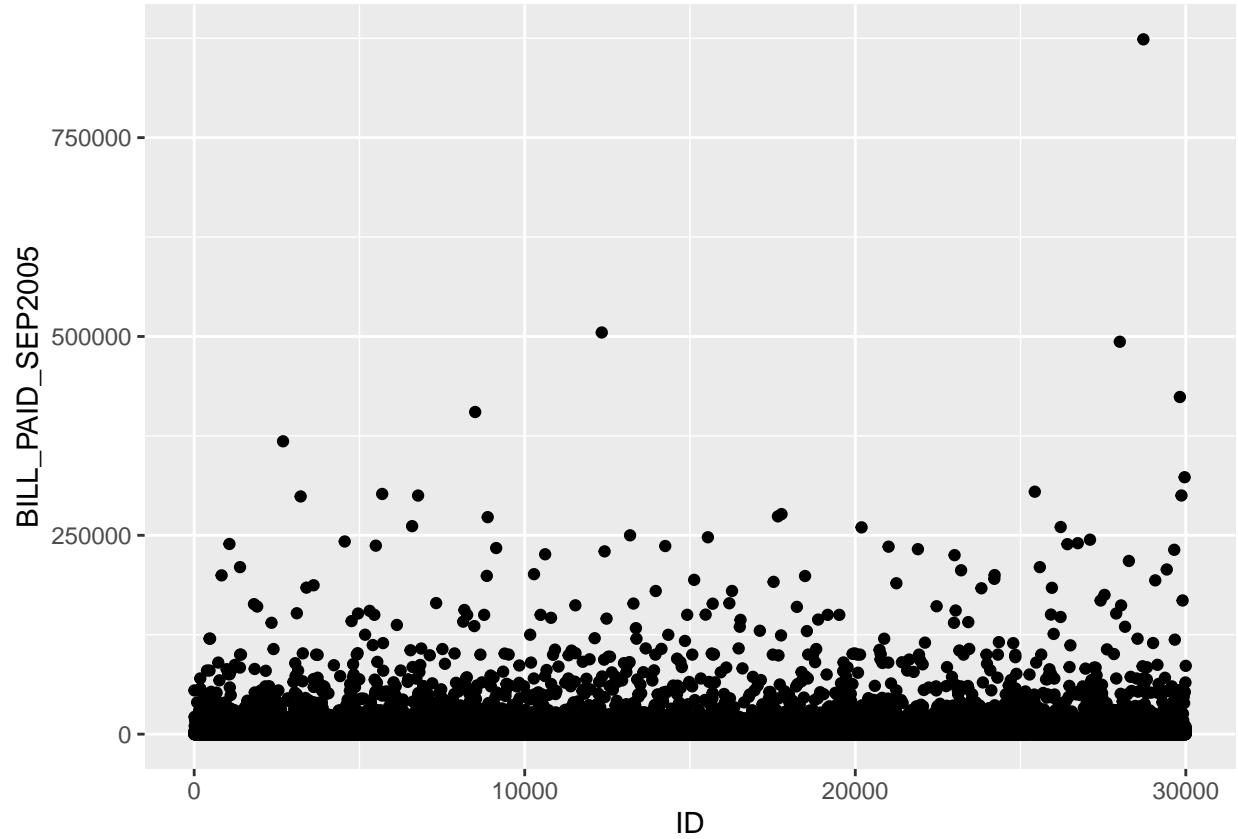
Here we can see Bill for month of September month for particular ID

```
ggplot(data = creditCard) +  
  geom_point(mapping = aes(x = ID,y = BILL_SEP2005))
```



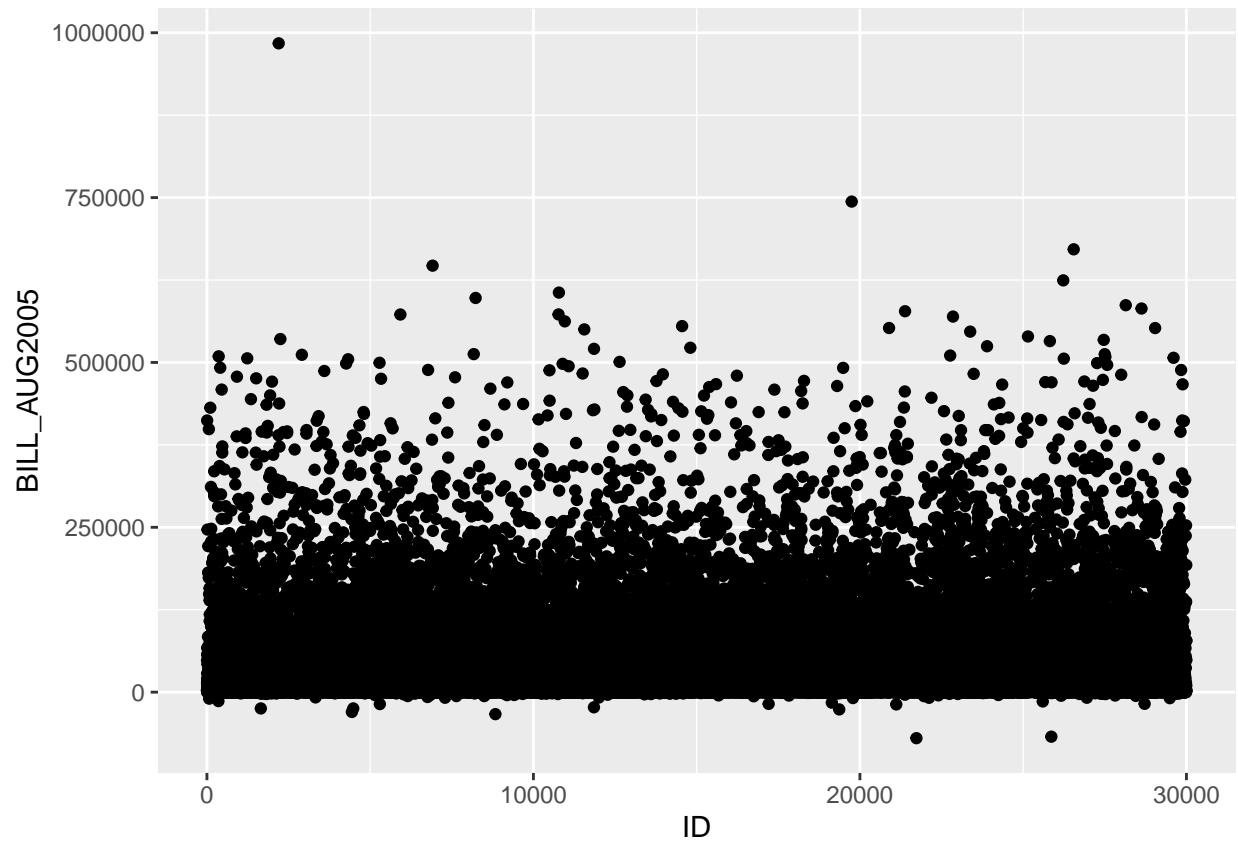
Here we can see actual Bill paid for month of September month for particular ID

```
ggplot(data = creditCard) +  
  geom_point(mapping = aes(x = ID, y = BILL_PAID_SEP2005))
```



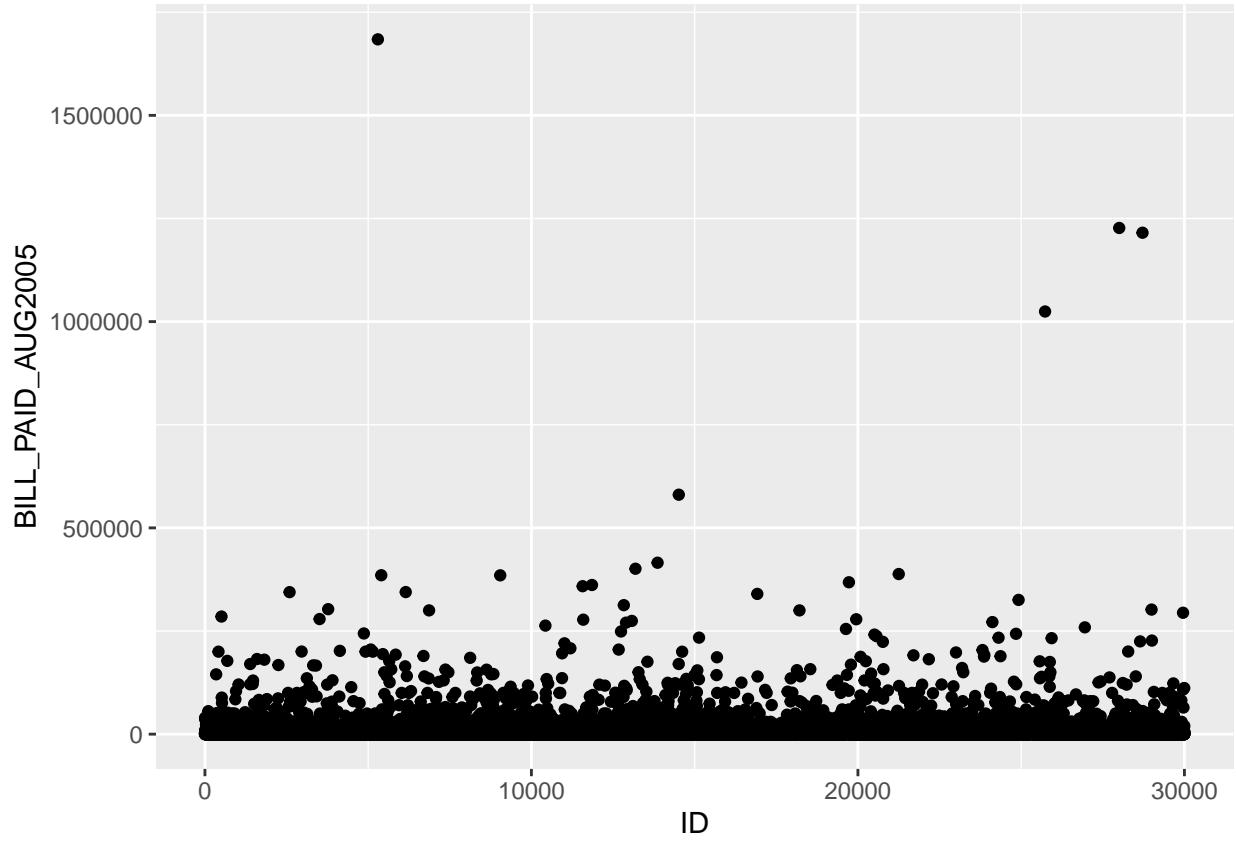
Here we can see Bill for month of August month for particular ID

```
ggplot(data = creditCard) +  
  geom_point(mapping = aes(x = ID, y = BILL_AUG2005))
```



Here we can see actual Bill paid for month of August month for particular ID

```
ggplot(data = creditCard) +  
  geom_point(mapping = aes(x = ID, y = BILL_PAID_AUG2005))
```



*Finding the difference between Total Bill and Total Bill Paid*

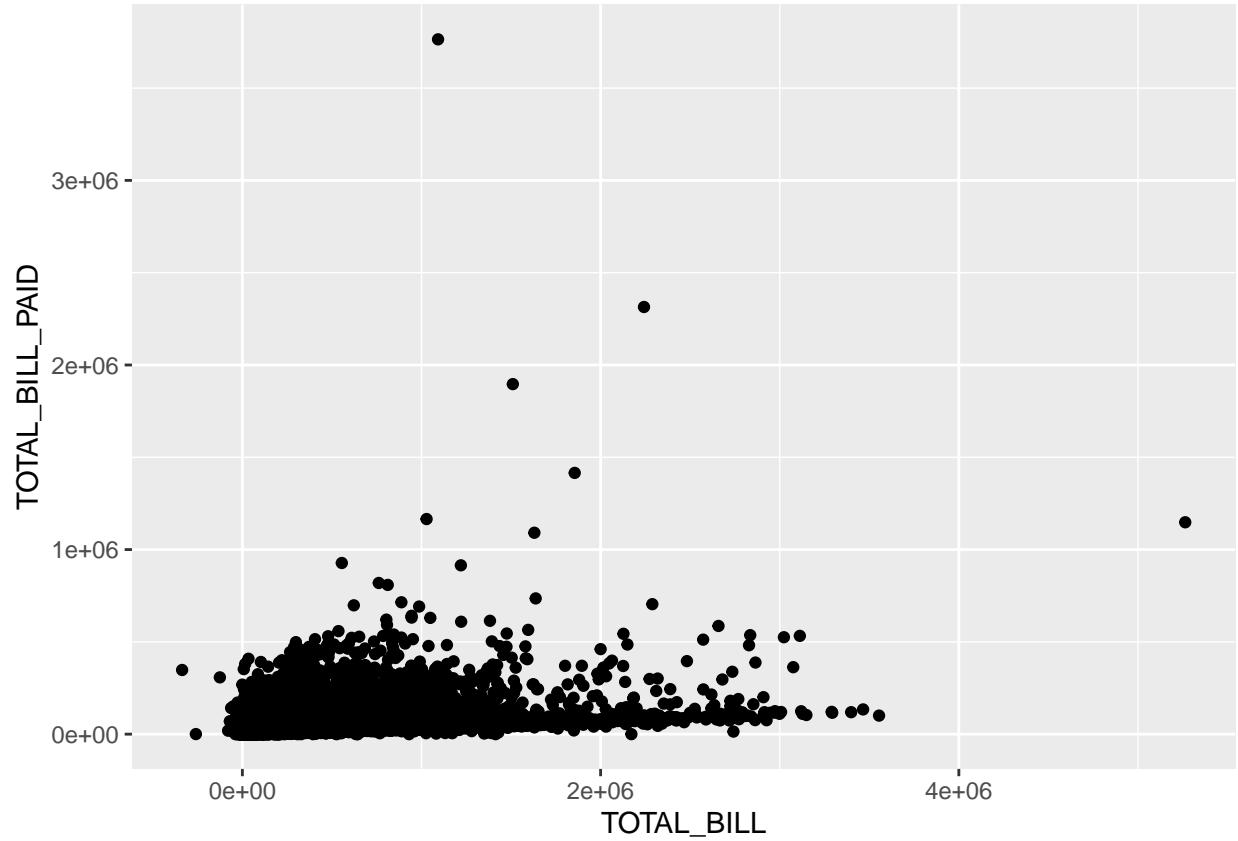
```

sumx = sum(creditCard$TOTAL_BILL)
sumy = sum(creditCard$TOTAL_BILL_PAID)
diff = sumx - sumy
print(diff)

## [1] 7146308359

ggplot(data = creditCard) +
  geom_jitter(mapping = aes(x = TOTAL_BILL,y = TOTAL_BILL_PAID))

```



## Conclusion

We can analize that there is a difference between Total Bill and Tottal Bill Paid