Aim : The project aims at visualising the data with the help of R language (ggplot2 for data visualisation) and applied machine algorithms like linear and multiple regression.

Our objective is to first visualise and then analyze the data. Based on our analyzation, we will predict the result using Machine learning algorithms i.e linear and multiple regression. Analysis of data is done by using R-language and the package used for visualisation is ggplot2.

Software Requirements

R-Studio - https://www.rstudio.com/products/rstudio/download/

ggplot2 - just type the following command to install ggplot2 package

```
> install.packages(ggplot2)
 > library(ggplot2) #this command is used to load the package if already installed
```

Introduction

The dataset in which we are performing the data analysis is a data-set provided by City Union Bank. It consists of all the ATM Transactions done by City Union bank and other banks from '1/1/2011' to '9/29/2017' at following five atm places.

- Big Street ATM
- Mount Road ATM
- Airport ATM
- KK Nagar ATM
- Christ College Road

Open the RStudio and load the .csv file.

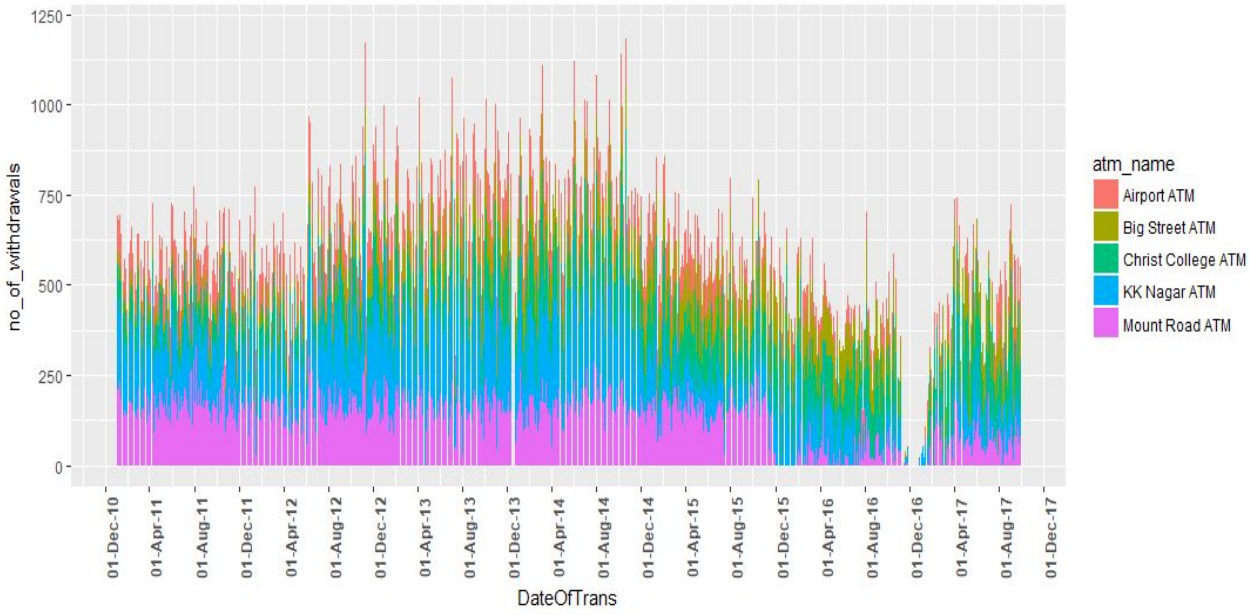
1.First is to directly set the working directory and store in our desired dataframe by following command :

```
> setwd("C:/Users/nits/Desktop") #set a new working directory where the sample file is downloaded

> data <- read.csv('atmdata.csv')

> atm <- data ;
```

We will begin our first visualisation as a bar graph between the transaction date and total amount withdrawn.The command used to create a bar graph using ggplot2 package is :

```
> p1 <- ggplot(atm,aes(x=DateOfTrans,y=total_amount_withdrawn,fill=atm_name))
 > p1 <- p1 + geom_bar(stat="identity") +
scale_x_date(date_labels="%d-%b-%y",date_breaks = "4 month")
 > p1 <- p1 + theme(axis.text.x = element_text(angle=90,face="bold"))
```

The output of the above code is :



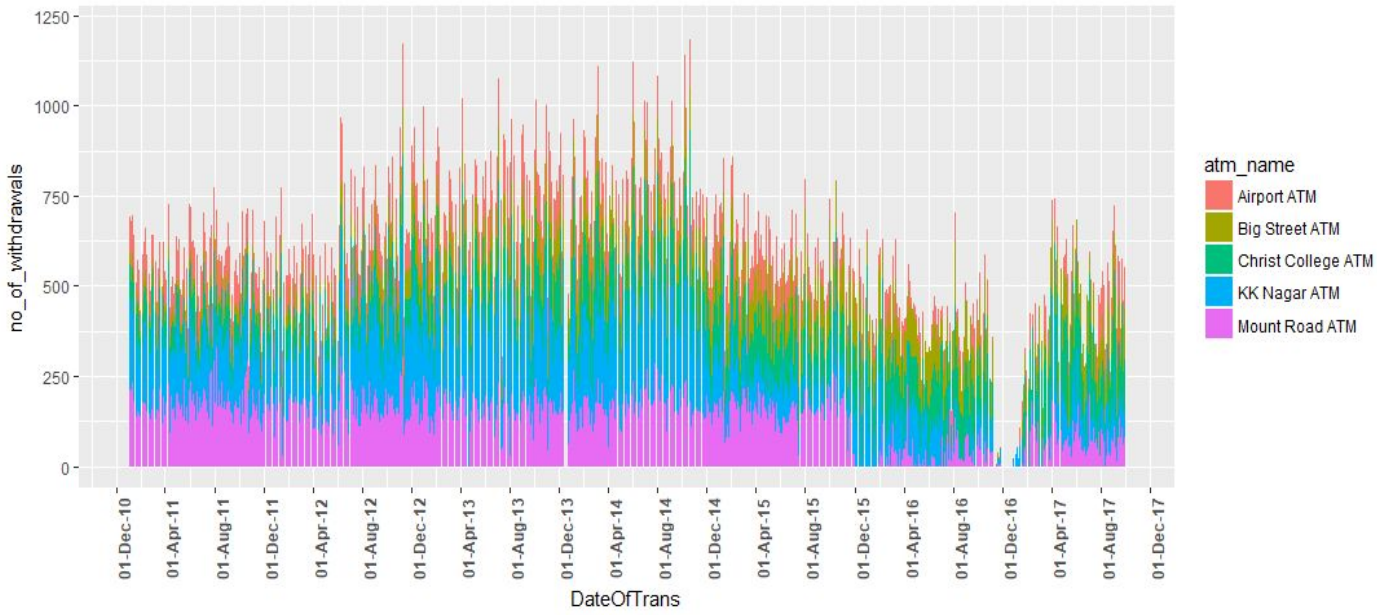
Similarly the code for bar graph between the transaction date and total number of atm card withdraws is :

```
 > p1 <- ggplot(atm,aes(x=DateOfTrans,y=no_of_withdrawals,fill=atm_name))
```

```
  > p1 <- p1 + geom_bar(stat="identity") +
```

```
scale_x_date(date_labels="%d-%b-%y",date_breaks = "4 month")
```

```
  > p1 <- p1 + theme(axis.text.x = element_text(angle=90,face="bold"))
```



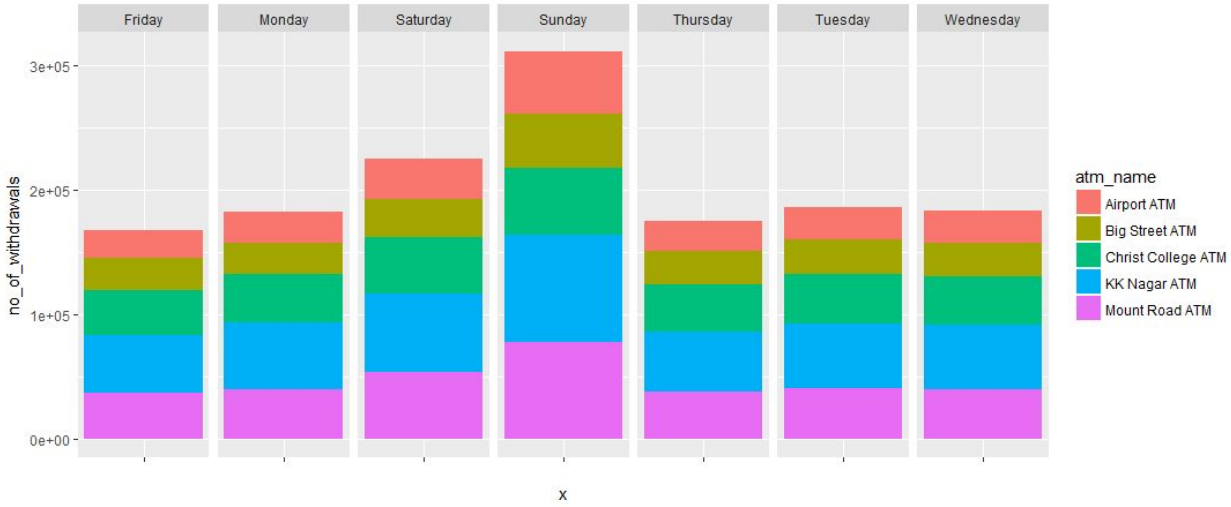
Analysis Observed :-

1. It is clearly observed from the graphs that from date between Oct-2016 to jan-2016 , there is no cards and cash withdrawals. The bar graph at that series of points tends to almost zero. The only reason is the *2016 Indian banknote demonetisation* .
2. One can figured out that the amount of transaction or cash withdrawal is maximum from KK Nagar ATM. Filling duration or capacity of cash needs to be increased therefore in KK Nagar ATM.
3. It can be also seen that the card withdrawals and atm transaction is maximum on month October of every year. One possible reason might be the more number of holidays as compared to any other month.

The next visualisation is based on our usage of facets. Facets are very useful in splitting the data according to some field. Here we will observe the no of card withdrawals on the basis of weekdays i.e Sunday,Monday,etc.

The following code will help to create the facet:

```
> p <- ggplot(atm,aes(x="factors",y=no_of_withdrawals,fill=atm_name))
> p = p + geom_bar(stat = "identity" )
> p = p + facet_grid(facets = ~ weekday)
```

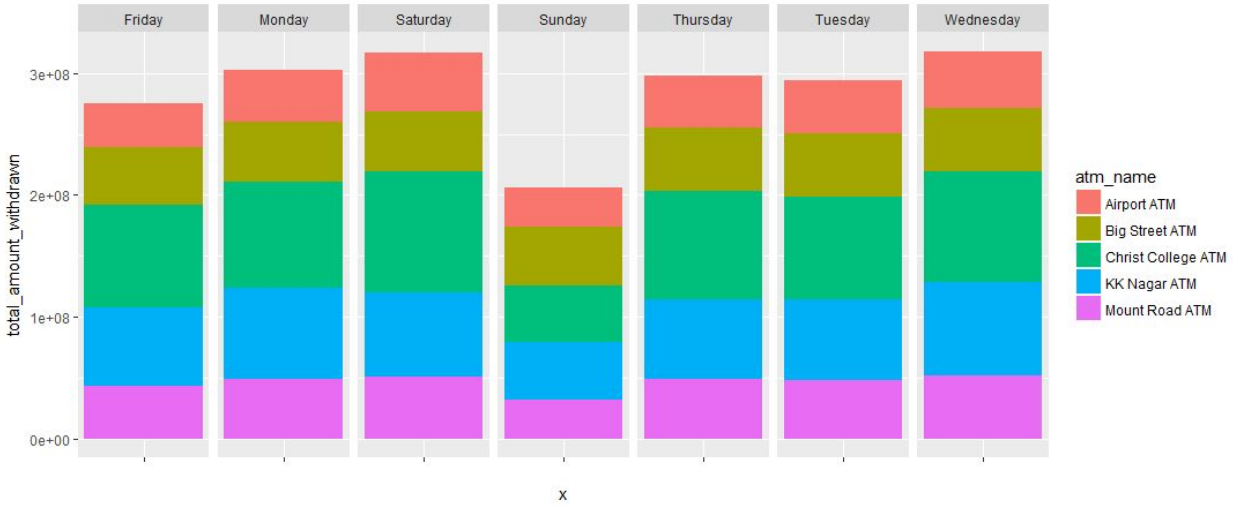


OBSERVATIONS :-

1. Obviously On sundays the amount of withdrawals are more because of holiday.
2. *KK Nagar ATM* still has the maximum number of withdrawals on almost all days due to its dense population.

🤘 Similarly the code for facet on the total amount withdrawn based on weekdays is :

```
> p <- ggplot(atm,aes(x="factors",y=total_amount_withdrawn,fill=atm_name))
> p = p + geom_bar(stat = "identity" )
> p = p + facet_grid(facets = ~ weekday);
```

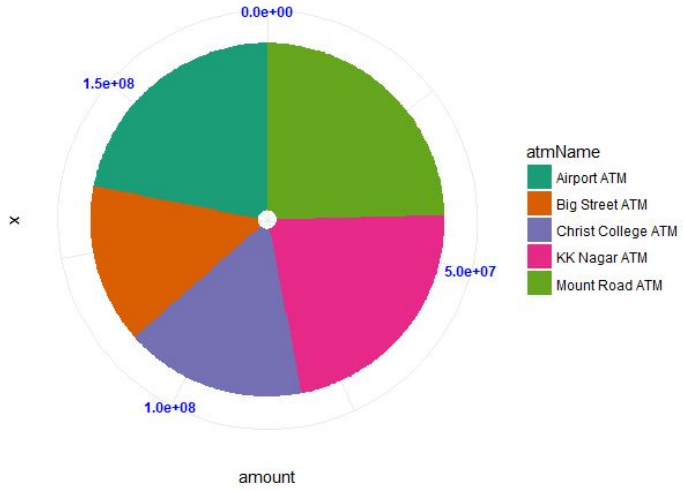


OBSERVATIONS :

1. Here is an interesting observation that the transaction is maximum on saturday and wednesday instead of Sunday and Monday. It is probably due to the starting weekend course of a new week.
2. The amount withdrawn is much more from Christ College ATM. Another visualisation one can made is the sudden increase of cash withdrawn on saturday from Christ College ATM. The reason may be the more amount of withdrawals from students as students take cash for the weekend.

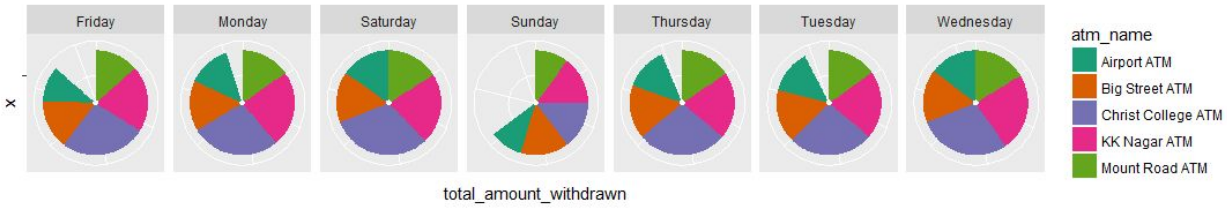
The above two plots can be also visualised through pie chart. One of the main problem occurs with the pie charts is that they take more time in loading/creation , hence the following example is reduced to minimum number of instruction so that pie chart will not take too long to load. The command for creating a pie chart based on the atm name and amount withdrawal is :

```
> tmp = subset(atm, weekday == "Sunday") #509 entries only
 > pc <- data.frame(atmName = tmp$atm_name, amount = tmp$total_amount_withdrawn
)
 > bp <- ggplot(pc,aes(x="",y=amount,fill = atmName)) + geom_bar(stat = "identity")
 > pie <- bp + coord_polar("y",start=0)
 > pie <- pie + scale_fill_brewer(palette = "Dark2") + theme_minimal() +
theme(axis.text.x = element_text(color =     "blue",face="bold"));
```



Analysis are same as that of above analysis observed. Bar chart with facets win over pie charts in this aspect. The visualisation and cration of facets is also possible in pie chart. Again it took double of the current time to display the resulted chart. Therefore we have again done the slicing by taking records dated from "2015-1-1" to "2017-9-27". Though it take too long But then also we will form it 😝 The command for creating facets in pie chart is :

```
> tmp <- data
 > tmp$transaction_date <- as.Date(tmp$transaction_date, format=
"%Y-%m-%d");#formating the date in "%Y-%m-%d" format
 > k <- subset(tmp,transaction_date > "2015-1-1" & transaction_date < "2017-9-27");
 > pl <- ggplot(data=k,aes(x="",y=total_amount_withdrawn,fill = atm_name)) +
geom_bar(stat="identity")
 > pl <- pl + facet_grid(facets = ~ weekday )
 > pl <- pl + coord_polar(theta="y")
 > pl <- pl + scale_fill_brewer(palette = "Dark2") + theme(axis.text.x =
element_blank(),axis.text.y = element_blank())
```



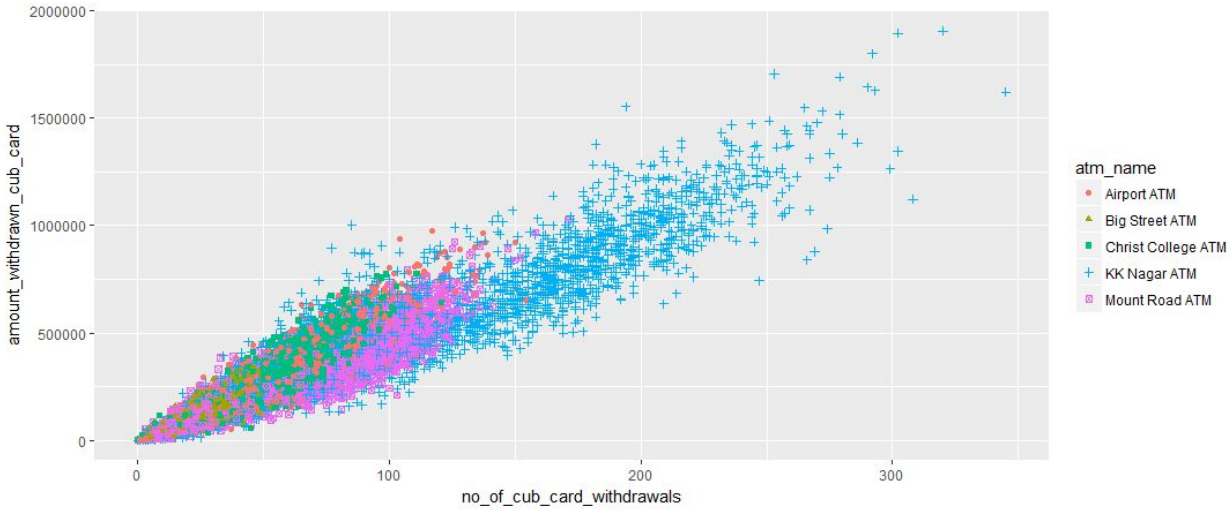
- The only observation different from the bar facets are :- it shows the days when the demonetisation period was there or minimum number of transaction takes place. The increasing order of withdrawal is Sunday < Friday < Tuesday < Thursday < Monday < Saturday == Wednesday , and clearly the demonetisation was forced upon sunday.
- The next visualisation is on the scatter plot which is very useful for implementing the machine algorithms. Scatter plot is basically placing a point which describes the y-axis value on a particular x-axis value. The scatter plot is done on the number of cub cards withdraw and amount withdrawn from the cub card only. the command for creating a scatter plot is
- `> p <-`

  `ggplot(atm,aes(x=no_of_cub_card_withdrawals,y=amount_withdrawn_cub_card,color = atm_name))`

  `> p <- p + geom_point(aes(color=atm_name,shape=atm_name)) # geom_point is used for plotting a point.`

if the field name(atm_name) is small or abbreviated then we can also use

`> p + geom_text(aes(label=atm_name),size=1);`



OBSERVATIONS :

1. It is obvious that as the amount of cub card is more withdrawn the cash transcation from cub card becomes more.
2. Nos of card withdrawals as well as total amount withdrawn are both maximum from KK Nagar ATM

Linear Regression

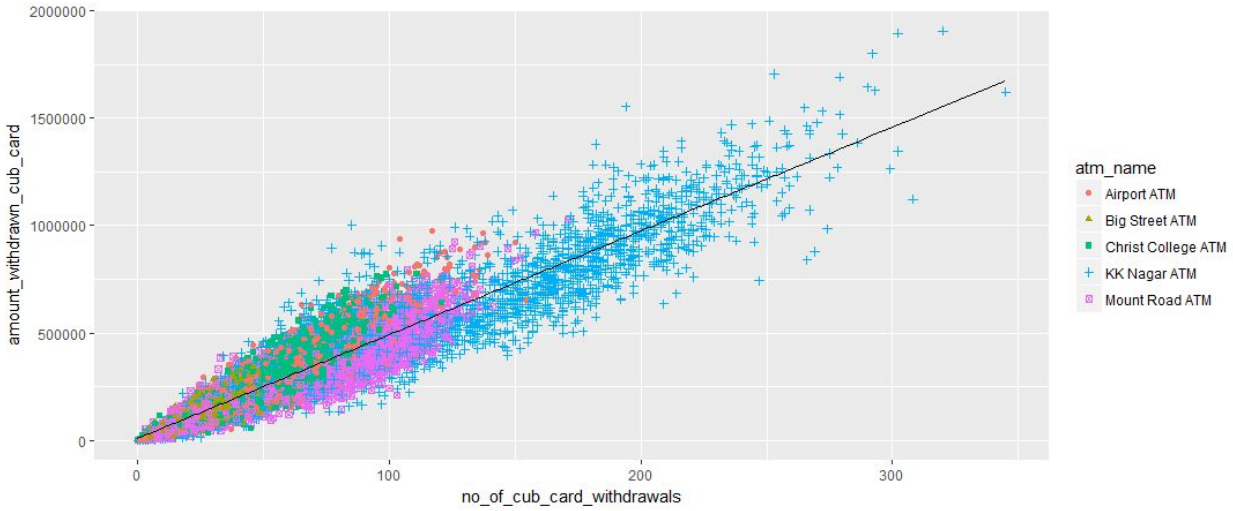
- Now let us apply the linear regression in this graph using lm() function. lm() will take a relation which in our case will be amount_withdrawn_cub_card ~ no_of_cub_card_withdrawals. We will take the x and y values differently so that we will not face problem while predicting the data. The below code will scatter plot the above scenario but with regression line in black color.
- `> x <- atm$no_of_cub_card_withdrawals`

  `> y <- atm$amount_withdrawn_cub_card`

  `> relation <- lm(y ~ x)`

  `> summary(relation) # to know about other factors affecting the graph.`

  `> p <- p + geom_line(aes(y=predict(relation)),color="black") # to show the regression line through graph`



Now for predicting the data we will use this regression line. Suppose we have to predict that what will be the total amount withdraw from cub card when the number of withdrawal is 345. The actual amount withdraw from cub card when card withdrawal is 345 was 1662290. You can check the actual value of amount withdraw by the following code :

```
 > k = subset(atm,no_of_cub_card_withdrawals==345)
  > k = k$amount_withdrawn_cub_card # it equals to 1662290
```

Now lets predict the same result by using the linear regression. We will make dataframe with same frames. The below small snippet will give you the following result :-

```
> l <- data.frame(x <- 345)
 > k <- predict(relation,l)# now k is equal to 1675879
```

Result came out to be 1675879 which is 8% margin. Therefore now we will apply multiple regression which contains more dependent variable for a particular condition.

Multiple regression

Multiple Regression is used for analyzing the relationship between several independent or predictor variables and a dependent or criterion variable. In our scenario the several independent variables are number of cub card withdrawals and number of other card withdrawals. The dependent or criterion variable will be the total amount withdrawn on that day. In the end we will predict the total amount withdrawn by putting the values of cub cards and other cards in the multiple regression equation. We will first create a data frame with factors containing cub card withdrawals , other card withdrawals and total amount withdrawan from the atm. Then we will form a relation between the factors using lm() function which will be :

formula = total_amount_withdrawn ~ no_of_cub_card_withdrawals + no_of_other_card_withdrawals

With this relation we will get the coefficients as well slopes for different factors. The following code will store the coefficients and slopes on variables a,xCubCard and xOther_card.

```
> mult <- tmp[,c("no_of_cub_card_withdrawals","no_of_other_card_withdrawals","total_amount_withdrawn")]
 > relation <- lm(data=mult,total_amount_withdrawn ~ no_of_cub_card_withdrawals + no_of_other_card_withdrawals) #applying multiple regressio and then creating an equation
 print(relation)
```

The result of print(relation) will be

Call:

lm(formula = total_amount_withdrawn ~ no_of_cub_card_withdrawals + no_of_other_card_withdrawals, data = mult)

Coefficients:

(Intercept) no_of_cub_card_withdrawals

-14002 5171

no_of_other_card_withdrawals

3351

We will now store the intercepts and slope in particular variables.

```
> a <- coef(relation)[1]
 > xCub_card <- coef(relation)[2]
 > xOther_card <- coef(relation)[3]
```

Form a multiple regression equation by :-

```
> Y <- a + xCub_card*x1 + xOther_card*x2
```

Let us now predict the result when number of cub card withdrawal = 27 and number of other card withdrawal = 52. The amount withdrawan comes out to be 305100. Now lets check it by multiple regression by following command :

```
> x1 <- 27
> x2 <- 52
> Y <- a + xCub_card*x1 + xOther_card*x2
> Y
 (Intercept)
 299879.2
```

Hence the difference between the actual and observed value is quite less. At last just share whatever predictions and visualisation you can do by the given dataset. Eventually we will be able to get the more accurate results.

Thank you.

Project Partner : (Maneesh Bhakuni 3241 &
Mohit Kumar 3246)