**Exploratory data analysis on HP dataset**

**# Reading file**

data<-read.csv("3102\_UIDs\_Integrated\_data\_Training.csv",na.strings='NULL',stringsAsFactors = FALSE,colClasses =c("Actuals"="numeric"))

**# Columns of data**

names(data)

**Output:**

[1] "UID" "PPQD\_Bucket"

[3] "PPID" "Period"

[5] "RTM" "Country\_ISO\_Code"

[7] "Partner\_Name" "World\_Country\_Region"

[9] "QCV\_Segment" "combined\_with\_QCV"

[11] "Actuals" "Market\_Share\_Units"

[13] "Market\_Share\_Revenue" "Market\_Size\_Units"

[15] "Market\_Size\_Revenue" "NPS\_Overall\_Performance\_HP"

[17] "NPS\_Ease\_Conducting\_Business" "NPS\_Recommend\_HP"

[19] "MEI\_CPI\_Inflation" "MEI\_Exports"

[21] "MEI\_Fixed\_Investment" "MEI\_GDP"

[23] "MEI\_Government\_Growth" "MEI\_Imports"

[25] "MEI\_Industrial\_Production" "MEI\_Merchandise\_Exports"

[27] "MEI\_Merchandise\_Imports" "MEI\_Nominal\_Retail\_Sales"

[29] "MEI\_Private\_Consumption" "MEI\_Real\_Retail\_Sales"

[31] "MEI\_WPI\_Inflation" "MEI\_Price\_Index"

[33] "MEI\_Trade\_GDP\_Ratio" "MEI\_Merchandise\_Trade\_GDP\_Ratio"

[35] "MEI\_Real\_Nominal\_Sales" "Program\_Status"

[37] "PCOMP\_Achievement\_Bonus\_Perc" "PCOMP\_Sum\_Bonus\_CLC"

**# summary of data**

summary(data)

**# the number of unique values in each column**

sapply(data,function(x) length(unique(x)))

**output:**

UID PPQD\_Bucket

3026 11

PPID Period

1817 23

RTM Country\_ISO\_Code

3 102

Partner\_Name World\_Country\_Region

1829 104

QCV\_Segment combined\_with\_QCV

37 333

Actuals Market\_Share\_Units

50086 3485

Market\_Share\_Revenue Market\_Size\_Units

5016 3330

Market\_Size\_Revenue NPS\_Overall\_Performance\_HP

5838 89

NPS\_Ease\_Conducting\_Business NPS\_Recommend\_HP

86 73

MEI\_CPI\_Inflation MEI\_Exports

1613 1599

MEI\_Fixed\_Investment MEI\_GDP

1573 1615

MEI\_Government\_Growth MEI\_Imports

1545 1603

MEI\_Industrial\_Production MEI\_Merchandise\_Exports

1341 1516

MEI\_Merchandise\_Imports MEI\_Nominal\_Retail\_Sales

1516 903

MEI\_Private\_Consumption MEI\_Real\_Retail\_Sales

1548 904

MEI\_WPI\_Inflation MEI\_Price\_Index

1146 1631

MEI\_Trade\_GDP\_Ratio MEI\_Merchandise\_Trade\_GDP\_Ratio

1622 1628

MEI\_Real\_Nominal\_Sales Program\_Status

904 4

PCOMP\_Achievement\_Bonus\_Perc PCOMP\_Sum\_Bonus\_CLC

18827 10429

**# dropping below variables as 75% of them have missing values**

data$NPS\_Overall\_Performance\_HP<-NULL

data$NPS\_Ease\_Conducting\_Business<-NULL

data$NPS\_Recommend\_HP<-NULL

**# summarising actuals per quarter and finding quarters having null actuals**

f<-data %>%

group\_by(Period) %>%

summarise(Actuals=mean(Actuals,na.rm=TRUE))

f<-data.frame(f)

f

sum(is.na(f))

***Conclusion:***

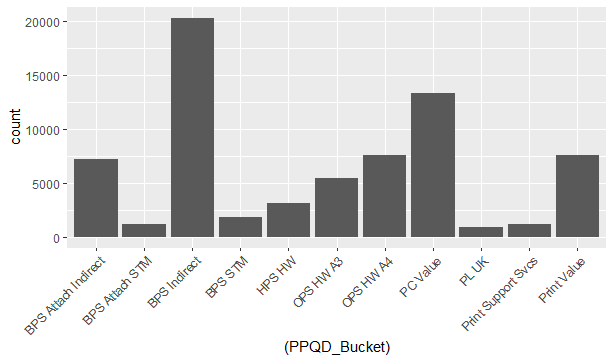
***Quarters 19-Q2 and 19-Q3 have all null values in actual variable.***

# **Distribution of sales metric (PPQD\_Bucket):**

library(ggplot2)

ggplot(data,aes(x=(PPQD\_Bucket))) + geom\_bar()+

theme(axis.text.x = element\_text(angle = 45, hjust = 1))



***We can see that the count of BPS Indirect is maximum.***

**Analysis at PPQD\_Bucket level:**

**# distribution of PPQD\_Bucket i.e. the buckets contribution to the actuals**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(PPQD\_Bucket) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v2<-v1[order(-v1$Actuals),]

v2

v3<-data.frame(v2)

v3

**output:**

PPQD\_Bucket Actuals

*<chr>* *<dbl>*

1 BPS Indirect 2470762260.

2 BPS STM 1182939314.

3 PC Value 950156716.

4 Print Value 403085398.

5 OPS HW A4 284280500.

6 BPS Attach Indirect 264535627.

7 OPS HW A3 200499985.

8 HPS HW 132313936.

9 BPS Attach STM 111929257.

10 Print Support Svcs 27076630.

11 PL UK 25665570.

***Conclusion:***

***BPS Indirect is contributing maximum towards sale of hp and***

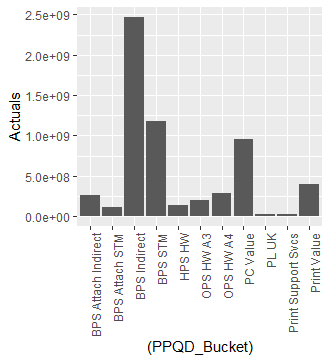
***PL UK is contributing minimum to the sale of hp.***

# plot for above

library(ggplot2)

ggplot(v3,aes(x=(PPQD\_Bucket),y=Actuals)) + geom\_bar(stat="identity")+

theme(axis.text.x = element\_text(angle = 90,hjust=1))



**Analysis at PPID level**

**# distribution of PPID i.e. the partner id’s contribution to actuals**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(PPID) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v2<-v1[order(-v1$Actuals),]

v2

v3<-data.frame(v2)

v3

**Output:**

PPID Actuals

*<chr>* *<dbl>*

1 1-10M99-18157 227992628.

2 1-10LT8-24280 123597210.

3 1-10M46-18672 114770838.

4 1-10LZ6-3123 83910449.

5 1-10LZ6-13516 82642380.

6 1-10M99-21300 68599182.

7 1-10M99-15982 62019695.

8 1-10MDH-9255 57874897.

9 France 50531924.

10 1-10LT8-17566 46261179.

# ... with 1,807 more rows

***Conclusion:***

***1-10M99-18157 partner id is contributing maximum towards actuals and***

***1-XNBHY6 Partner id is contributing minimum towards actuals.***

**# by dividing actuals of PPID by total actual sum**

total\_sum<-sum(data$Actuals,na.rm=TRUE)

total\_sum

library(dplyr)

v1<-data %>%

group\_by(PPID) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE)/total\_sum)

v2<-v1[order(-v1$Actuals),]

v2

v3<-data.frame(v2)

v3

**output:**

PPID Actuals

*<chr>* *<dbl>*

1 1-10M99-18157 0.0377

2 1-10LT8-24280 0.0204

3 1-10M46-18672 0.0190

4 1-10LZ6-3123 0.0139

5 1-10LZ6-13516 0.0137

6 1-10M99-21300 0.0113

7 1-10M99-15982 0.0102

8 1-10MDH-9255 0.00956

9 France 0.00835

10 1-10LT8-17566 0.00764

# ... with 1,807 more rows

**Analysis at country level**

**# distribution of countries i.e. order of countries giving sale to hp**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(World\_Country\_Region) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v2<-v1[order(-v1$Actuals),]

**Output:**

World\_Country\_Region Actuals

*<chr>* *<dbl>*

1 France 608465038.

2 United Kingdom 535673435.

3 Russian Federation 479363417.

4 United Arab Emirates 387572221.

5 Germany 372392124.

6 Netherlands 318749972.

7 Italy 315776580.

8 Switzerland 302216893.

9 Turkey 260527651.

10 Poland 224515599.

# ... with 94 more rows

***Conclusion:***

***Thus we can see that France is contributing more towards sale of hp and***

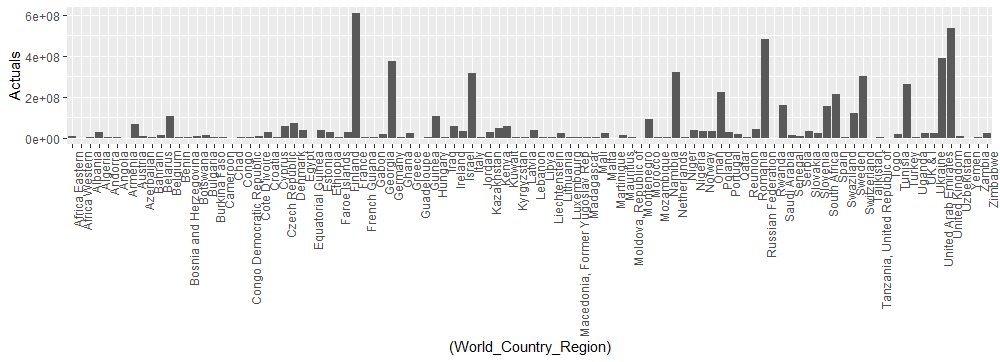
***Martinique is contributing very less.***

Below is the graph for this:

library(ggplot2)

ggplot(v3,aes(x=(World\_Country\_Region),y=Actuals)) + geom\_bar(stat="identity")+

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



***Highest contribution by: France***

***Lowest contribution by: Martinique***

**Analysis at QCV segment level**

**# distribution of channels i.e. order of QCV segments giving sale to hp**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(QCV\_Segment) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v2<-v1[order(-v1$Actuals),]

v3<-data.frame(v2)

**Output:**

QCV\_Segment Actuals

*<chr>* *<dbl>*

1 HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 674489421.

2 HPI EMEA UKI Channel 617337053.

3 HPI EMEA France Channel 608465038.

4 HPI EMEA CEMA EASTERN EUROPE Russia CHANNEL 479363417.

5 HPI EMEA D&A Germany Channel 372392124.

6 HPI EMEA GWE Netherlands Channel 318749972.

7 HPI EMEA Italy Channel 315776580.

8 HPI EMEA GWE Switzerland Channel 305206412.

9 HPI EMEA CEMA MESA TURKEY CHANNEL 260527651.

10 HPI EMEA CEMA\_CENTRAL EUROPE POLAND CHANNEL 224515599.

# ... with 27 more rows

***Conclusion:***

***Thus we can see that HPI EMEA CEMA MESA MIDDLE EAST CHANNELis more contributing towards sale of hp and HPI EMEA CEMA EASTERN EUROPE Azerbaijan is contributing very less.***

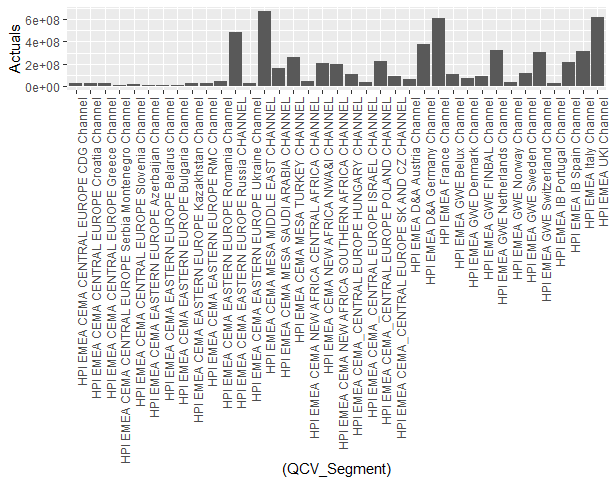
We can also depict it through graph. Below is the graph for this:

**# plot for channels contribution to sale for hp**

library(ggplot2)

ggplot(v3,aes(x=(QCV\_Segment),y=Actuals)) + geom\_bar(stat="identity")+

theme(axis.text.x = element\_text(angle = 90,hjust=1))



**Analysis of country and QCV segment**

**# number of distinct countries a channel belongs to**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(QCV\_Segment) %>%

summarise(n\_distinct(World\_Country\_Region))

v2<-v1[order(-v1$`n\_distinct(World\_Country\_Region)`),]

v3<-data.frame(v2)

**output:**

QCV\_Segment `n\_distinct(World\_Country\_Regi~

*<chr>* *<int>*

1 HPI EMEA CEMA NEW AFRICA NWA&I CHANNEL 23

2 HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 16

3 HPI EMEA CEMA NEW AFRICA SOUTHERN AFRICA CHANNEL 7

4 HPI EMEA CEMA EASTERN EUROPE RMC Channel 5

5 HPI EMEA CEMA NEW AFRICA CENTRAL AFRICA CHANNEL 5

6 HPI EMEA CEMA CENTRAL EUROPE CDG Channel 4

7 HPI EMEA GWE FINBAL Channel 4

8 HPI EMEA UKI Channel 3

9 HPI EMEA CEMA CENTRAL EUROPE Greece Channel 2

10 HPI EMEA CEMA CENTRAL EUROPE Serbia Montenegro Ch~ 2

# ... with 27 more rows

***Conclusion:***

***Thus we can see that HPI EMEA CEMA NEW AFRICA NWA&I CHANNEL belongs to maximum***

***number of distinct countries.***

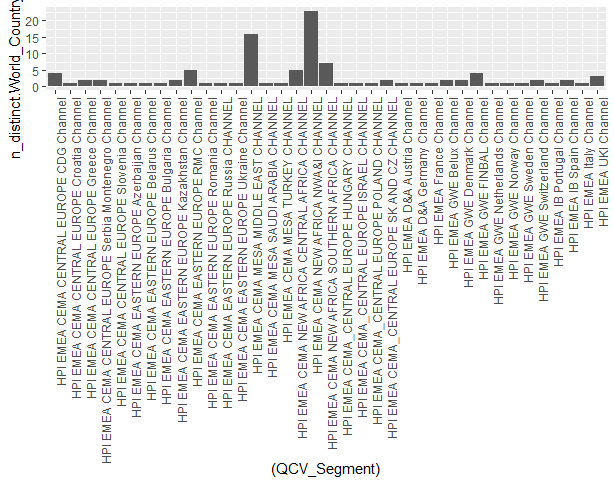
We can also depict it through graph. Below is the graph for this:

**# plot for distinct countries a channel belongs to**

library(ggplot2)

ggplot(v3,aes(x=(QCV\_Segment),y=`n\_distinct.World\_Country\_Region.`)) + geom\_bar(stat="identity")+

theme(axis.text.x = element\_text(angle = 90,hjust=1))



**Analysis at PPQD Bucket, PPID and country level**

**# distribution on PPQD Bucket, PPID and country level i.e. contribution towards revenue on this level**

**Code:**

v1<-data %>%

group\_by(PPQD\_Bucket,PPID,World\_Country\_Region) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v1

v2<-v1[order(-v1$Actuals),]

v2

v3<-data.frame(v2)

v3

**Output:**

PPQD\_Bucket PPID World\_Country\_Region Actuals

*<chr>* *<chr>* *<chr>* *<dbl>*

1 BPS STM 1-10M99-18157 France 175242379.

2 BPS STM 1-10LT8-24280 United Kingdom 92139080

3 BPS STM 1-10LZ6-13516 United Arab Emirates 80520886.

4 BPS Indirect 1-10M99-15982 United Arab Emirates 62019695.

5 BPS STM 1-10LZ6-3123 United Arab Emirates 56956150

6 BPS STM 1-10M99-21300 France 52467359.

7 BPS Indirect 1-10M46-18672 Netherlands 51605263.

8 BPS STM France France 46881357.

9 BPS STM 1-10MDH-9255 United Arab Emirates 42075425.

10 PC Value 1-10M46-18672 Netherlands 41946288.

# ... with 3,016 more rows

***Conclusion:***

***PPQD Bucket: BPS STM +***

***PPID: 1-10M99-18157 +***

***Country: France***

***Is contributing maximum to actuals and***

***PPQD Bucket: PL UK +***

***PPID: 1-XNBHY6 +***

***Country: Turkey***

***Is contributing minimum to actuals.***

**If we consider qcv segment also along with PPQD Bucket, PPID and country**

**# distribution on PPQD Bucket, PPID,country and QCV segment level**

**Code:**

v1<-data %>%

group\_by(PPQD\_Bucket,PPID,World\_Country\_Region,QCV\_Segment) %>%

summarise(Actuals=sum(Actuals,na.rm=TRUE))

v1

v2<-v1[order(-v1$Actuals),]

v2

v3<-data.frame(v2)

v3

**Output:**

PPQD\_Bucket PPID World\_Country\_Region QCV\_Segment Actuals

*<chr>* *<chr>* *<chr>* *<chr>* *<dbl>*

1 BPS STM 1-10M99-18157 France HPI EMEA France Channel 175242379.

2 BPS STM 1-10LT8-24280 United Kingdom HPI EMEA UKI Channel 92139080

3 BPS STM 1-10LZ6-13516 United Arab Emirates HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 80520886.

4 BPS Indirect 1-10M99-15982 United Arab Emirates HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 62019695.

5 BPS STM 1-10LZ6-3123 United Arab Emirates HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 56956150

6 BPS STM 1-10M99-21300 France HPI EMEA France Channel 52467359.

7 BPS Indirect 1-10M46-18672 Netherlands HPI EMEA GWE Netherlands Channel 51605263.

8 BPS STM France France HPI EMEA France Channel 46881357.

9 BPS STM 1-10MDH-9255 United Arab Emirates HPI EMEA CEMA MESA MIDDLE EAST CHANNEL 42075425.

10 PC Value 1-10M46-18672 Netherlands HPI EMEA GWE Netherlands Channel 41946288.

# ... with 3,016 more rows

***Conclusion:***

***PPQD Bucket: BPS STM +***

***PPID: 1-10M99-18157 +***

***Country: France +***

***QCV segment : HPI EMEA France Channel***

***Is contributing maximum to actuals and***

***PPQD Bucket: PL UK +***

***PPID: 1-XNBHY6 +***

***Country: Turkey +***

***QCV segment : HPI EMEA CEMA MESA TURKEY CHANNEL***

***Is contributing minimum to actuals.***

**Analysis of partners and products**

**# which partner is having more products**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(PPID) %>%

summarise(n\_distinct(PPQD\_Bucket))

v2<-v1[order(-v1$`n\_distinct(PPQD\_Bucket)`),]

v2

v3<-data.frame(v2)

**Output:**

PPID `n\_distinct(PPQD\_Bucket)`

*<chr>* *<int>*

1 1-10LQI-17174 8

2 9.91E+11 8

3 1-10LT8-17735 7

4 1-10LT8-22536 7

5 1-10M99-12763 7

6 1-10LZ6-11299 6

7 1-10LZ6-1612 6

8 1-10LQI-1199 5

9 1-10LQI-12378 5

10 1-10LQI-4152 5

# ... with 1,807 more rows

***Conclusion:***

***Partner Id- 1-10LQI-17174 (Partner name- B.D.F. SPA\_HPI|500953567)***

***and 9.91E+11 have more number of products.***

**# difference between market size units and market share units for different**

**Products**

**Code:**

library(dplyr)

v1<-data %>%

group\_by(PPQD\_Bucket) %>%

summarise(Market\_Size\_Units=sum(Market\_Size\_Units,na.rm=TRUE),Market\_Share\_Units=sum(Market\_Share\_Units,na.rm=TRUE),difference=sum(Market\_Size\_Units,na.rm=TRUE)-sum(Market\_Share\_Units,na.rm=TRUE))

v1

v2<-v1[order(-v1$Market\_Size\_Units),]

v2

v3<-data.frame(v2)

v3

**Output:**

PPQD\_Bucket Market\_Size\_Units Market\_Share\_Units difference

*<chr>* *<dbl>* *<dbl>* *<dbl>*

1 BPS Indirect 696327. 617195. 79131.

2 HPS HW 212511. 163431. 49080.

3 BPS STM 64400. 54514. 9886.

4 PL UK 0 0 0

5 Print Support Svcs 0 0 0

6 BPS Attach STM 17736. 39354. -21618.

7 OPS HW A3 23440. 55432. -31993.

8 Print Value 91882. 197147. -105265.

9 BPS Attach Indirect 108526. 245074. -136548.

10 OPS HW A4 197321. 371356. -174035.

11 PC Value 251415. 493613. -242199.

***Conclusion:***

***Maximum difference between market size unit and market share unit is for BPS indirect i.e. its potential of sale is more than how much it sale.***

**# Program status distribution as per partners**

library(dplyr)

v1<-data %>%

group\_by(Partner\_Name) %>%

count(Program\_Status)

v2<-v1[order(-v1$Program\_Status),]

v3<-data.frame(v2)

**Output:**

Partner\_Name Program\_Status n

*<chr>* *<int>* *<int>*

1 4S BILGI TEKNOLOJILERI ANONIM SIRKETI\_HPI|500954254 3 51

2 A & O SYSTEMS + SERVICES IRE LTD\_HPI|500861014 3 7

3 ABAST GRUP\_HPI|500953393 3 2

4 ABDULLA FOUAD COMPANY - DAMMAM\_HPI|500954633 3 44

5 ACSN\_HPI|500953437 3 4

6 Active Computers Systems S.A.\_HPI|500954585 3 28

7 ADVANCED COMPUTER TECHNOLOGY\_HPI|500954613 3 8

8 Advania AB\_HPI|500953229 3 6

9 Al-Falak Electronic Equipment & Supplies Co.\_HPI|500865683 3 8

10 AL MANARAH COMMUNICATIONS & INFORMATION TECHNO PI|500864758 3 15

***Thus we can see that ‘4S BILGI TEKNOLOJILERI ANONIM SIRKETI\_HPI|500954254’***

***Distributor got maximum number of program status 3 i.e. platinum.***

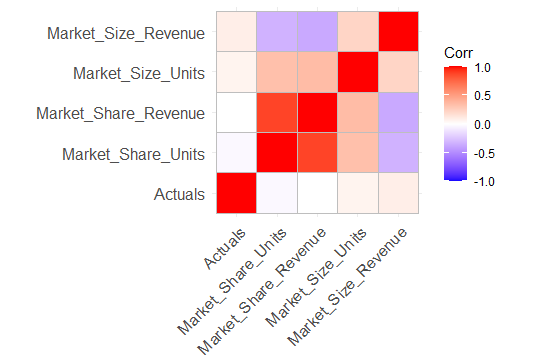
**# Correlation plot**

mydata<-data1[,c(11:15)]

corr<-cor(mydata)

library(ggcorrplot)

ggcorrplot(corr)



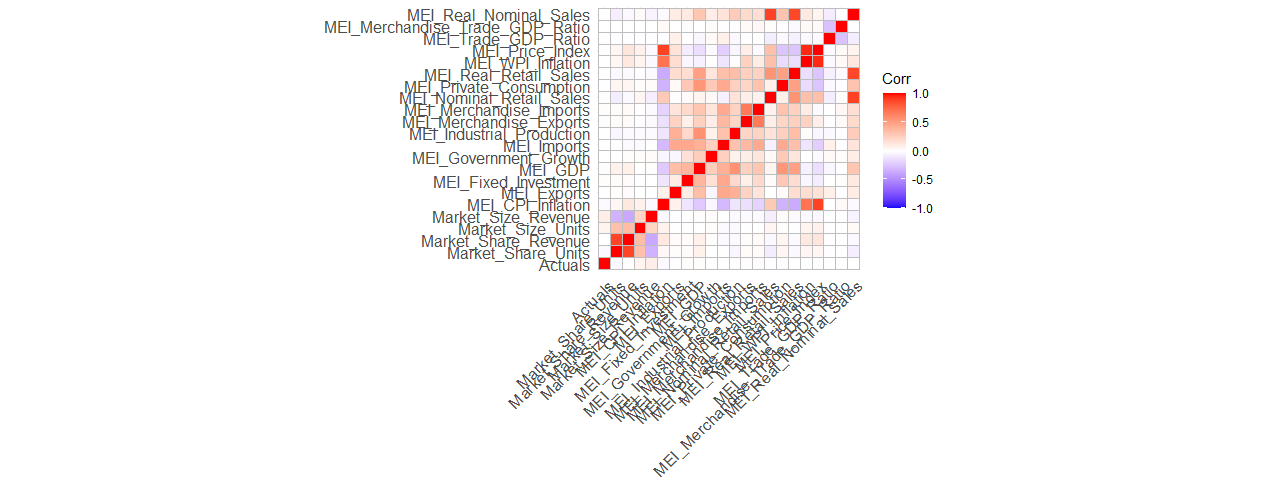
# correlation between all numerical variables(including MEI variables also)

mydata<-data1[,c(11:32)]

corr<-cor(mydata)

library(ggcorrplot)

ggcorrplot(corr)



**# Outlier Treatment**

data1<-data

**# summary of actual column before capping**

summary(data1$Actuals)

**output:**

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

-1670875 997 13916 95258 60354 17600100 6052

**# outlier treatment for continuous variable - "Actuals"**

boxplot(data1$Actuals)



**# an outlier is considered so if it is below the first quartile – 1.5·IQR or above third quartile + 1.5·IQR.**

qnt <- quantile(data1$Actuals, probs=c(.25, .75), na.rm = T)

qnt

**output:**

25% 75%

996.7675 60354.2800

caps <- quantile(data1$Actuals, probs=c(.05, .95), na.rm = T)

caps

**output:**

|  |
| --- |
| 5% 95%  0.0 387804.8 |
|  |
| |  | | --- | |  | |

H <- 1.5 \* IQR(data1$Actuals, na.rm = T)

H

**Output:**

[1] 89036.27

**# capping values**

data1$Actuals[data1$Actuals < caps[1]] <- caps[1]

data1$Actuals[data1$Actuals > caps[2]] <- caps[2]

**#summary of actual column after capping**

summary(data1$Actuals)

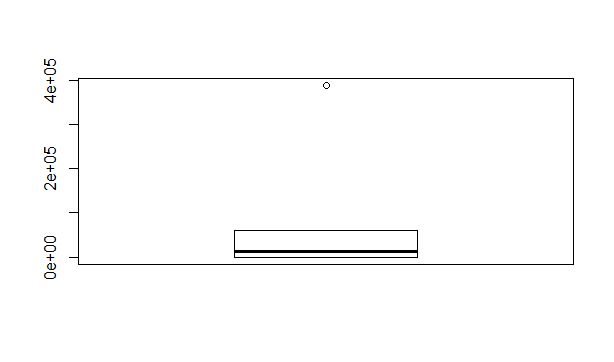
**Output:**

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.0 249.4 13915.9 106418.8 60354.3 387804.8 6052

**# Boxplot after outlier treatment**

boxplot(data1$Actuals)



**# Missing value treatment**

**# Country\_ISO\_Code has 0.13% of missing value, thus imputing it through mode**

val2 <- unique(data1$Country\_ISO\_Code[!is.na(data1$Country\_ISO\_Code)])

mode2 <- val2[which.max(tabulate(match(data1$Country\_ISO\_Code, val2)))]

mode2

data1$Country\_ISO\_Code[is.na(data1$Country\_ISO\_Code)] <- mode2

apply(data1,2,function(x)(sum(is.na(x))))

summary(data1$Actuals)

**# imputing missing values in Actuals column with median**

median1<-median(data1$Actuals,na.rm=TRUE)

median1

# imputed with median

data1<-data1 %>%

mutate(Actuals

=replace(Actuals,is.na(Actuals),median(Actuals,na.rm=TRUE)))

**# imputed missing values in Market\_Share\_Units with median**

data1<-data1 %>%

mutate(Market\_Share\_Units

=replace(Market\_Share\_Units,is.na(Market\_Share\_Units),median(Market\_Share\_Units,na.rm=TRUE)))

**# imputed missing values in Market\_Share\_Revenue with median**

data1<-data1 %>%

mutate(Market\_Share\_Revenue

=replace(Market\_Share\_Revenue,is.na(Market\_Share\_Revenue),median(Market\_Share\_Revenue,na.rm=TRUE)))

**# imputed missing values in Market\_Size\_Units with median**

data1<-data1 %>%

mutate(Market\_Size\_Units

=replace(Market\_Size\_Units,is.na(Market\_Size\_Units),median(Market\_Size\_Units,na.rm=TRUE)))

**# imputed missing values in Market\_Size\_Revenue with median**

data1<-data1 %>%

mutate(Market\_Size\_Revenue

=replace(Market\_Size\_Revenue,is.na(Market\_Size\_Revenue),median(Market\_Size\_Revenue,na.rm=TRUE)))

# imputed missing values in MEI variables with mean of countries

# all values of MEI variables are missing for particular countries thus taking qcv segment in account for that and computing average of all countries in particular segment

data1<-data1 %>%

group\_by(QCV\_Segment) %>%

mutate(MEI\_CPI\_Inflation=replace(MEI\_CPI\_Inflation,is.na(MEI\_CPI\_Inflation),mean(MEI\_CPI\_Inflation,na.rm=TRUE)),

MEI\_Exports=replace(MEI\_Exports,is.na(MEI\_Exports),mean(MEI\_Exports,na.rm=TRUE)),

MEI\_Fixed\_Investment=replace(MEI\_Fixed\_Investment,is.na(MEI\_Fixed\_Investment),mean(MEI\_Fixed\_Investment,na.rm=TRUE)),

MEI\_GDP=replace(MEI\_GDP,is.na(MEI\_GDP),mean(MEI\_GDP,na.rm=TRUE)),

MEI\_Government\_Growth=replace(MEI\_Government\_Growth,is.na(MEI\_Government\_Growth),mean(MEI\_Government\_Growth,na.rm=TRUE)),

MEI\_Imports=replace(MEI\_Imports,is.na(MEI\_Imports),mean(MEI\_Imports,na.rm=TRUE)),

MEI\_Industrial\_Production=replace(MEI\_Industrial\_Production,is.na(MEI\_Industrial\_Production),mean(MEI\_Industrial\_Production,na.rm=TRUE)),

MEI\_Merchandise\_Imports=replace(MEI\_Merchandise\_Imports,is.na(MEI\_Merchandise\_Imports),mean(MEI\_Merchandise\_Imports,na.rm=TRUE)),

MEI\_Nominal\_Retail\_Sales=replace(MEI\_Nominal\_Retail\_Sales,is.na(MEI\_Nominal\_Retail\_Sales),mean(MEI\_Nominal\_Retail\_Sales,na.rm=TRUE)),

MEI\_Private\_Consumption=replace(MEI\_Private\_Consumption,is.na(MEI\_Private\_Consumption),mean(MEI\_Private\_Consumption,na.rm=TRUE)),

MEI\_Real\_Retail\_Sales=replace(MEI\_Real\_Retail\_Sales,is.na(MEI\_Real\_Retail\_Sales),mean(MEI\_Real\_Retail\_Sales,na.rm=TRUE)),

MEI\_WPI\_Inflation=replace(MEI\_WPI\_Inflation,is.na(MEI\_WPI\_Inflation),mean(MEI\_WPI\_Inflation,na.rm=TRUE)),

MEI\_Price\_Index=replace(MEI\_Price\_Index,is.na(MEI\_Price\_Index),mean(MEI\_Price\_Index,na.rm=TRUE)),

MEI\_Trade\_GDP\_Ratio=replace(MEI\_Trade\_GDP\_Ratio,is.na(MEI\_Trade\_GDP\_Ratio),mean(MEI\_Trade\_GDP\_Ratio,na.rm=TRUE)),

MEI\_Merchandise\_Trade\_GDP\_Ratio=replace(MEI\_Merchandise\_Trade\_GDP\_Ratio,is.na(MEI\_Merchandise\_Trade\_GDP\_Ratio),mean(MEI\_Merchandise\_Trade\_GDP\_Ratio,na.rm=TRUE)),

MEI\_Real\_Nominal\_Sales=replace(MEI\_Real\_Nominal\_Sales,is.na(MEI\_Real\_Nominal\_Sales),mean(MEI\_Real\_Nominal\_Sales,na.rm=TRUE)))

**# Final data-frame after EDA and data cleaning is data1**