## **Business Analytics for Beginners Using R - Part I**

CAX Username: TapanGautam

Full Name: Tapan Gautam

Date:

**NOTE**: Please complete all sections in as much detail as possible and with supporting analysis. Each section should be answered. **Page Limit:** None.

--------------------------------------------------------------------------------------------------------------------------------------To be eligible for certification solvers need to submit a detailed report covering below mentioned sections:

--------------------------------------------------------------------------------------------------------------------------------------

Section 1: How the data was treated including missing value?

> setwd("E:\\Documents\\R language\\Crowd analytix and Kaggle\\CrowdAnalytix (Data Preparation and Exploration in R Part1)")

> company\_data<-read.csv(file="CAX\_Startup\_Data.csv",header=TRUE,as.is=T)

#Replacing blank cells and cells that contain name No Info by NA

> company\_data[company\_data=="No Info"]<-NA

> company\_data[company\_data==""]<-NA

#converting character into date format

> company\_data$Est..Founding.Date<-as.Date(company\_data$Est..Founding.Date,"%m/%d/%y")

> company\_data$Last.Funding.Date<-as.Date(company\_data$Last.Funding.Date,"%m/%d/%y")

> #Initiation of process for converting character variables in to numeric by grouping all relevant variables in vector col

> col<- c(3:5,10,11,18:23,25,61,66,68:70,72,74,88,92,94:96,98,99,102:116)

> for(i in col){company\_data[,i]<-as.numeric(company\_data[,i])}

> #performing calculations related to computation of error related to variables

> summation\_of\_error<-sapply(company\_data,function(x)sum(is.na(x)))

> percentage\_error<-as.data.frame (summation\_of\_error/nrow(company\_data)\*100)

> names<-row.names(percentage\_error)

> union\_of\_tables<-cbind(names,percentage\_error)

> row.names(union\_of\_tables)<-NULL

> colnames(union\_of\_tables)<-c("variable","percentage")

> #Extracting variables that contain error values less than 40% and encompassing them in new vector

>error\_less\_then\_40<-as.character(union\_of\_tables$variable[which(union\_of\_tables$percentage<40)])

> usable\_data<-company\_data[error\_less\_then\_40]

> #Extracting variables that contain error values more than 40% and encompassing them in new vector

> error\_more\_then\_40<-as.character(union\_of\_tables$variable[which(union\_of\_tables$percentage>40)])

> non\_usable\_data<-company\_data[error\_more\_then\_40]

> #Preparing excel sheet with edited data

> write.csv(usable\_data,"filtered\_data.csv",row.names=F)

>#preparing vector containing only numeric variables

> set\_of\_numeric\_variables<-usable\_data[,c(3:5,10,12:14,17:22,24,60,65,67:69,71,73,85,89,91:93,

+ 95,96,99:113)]

#forming data set of character variables

> names\_of\_numeric\_vector<-colnames(set\_of\_numeric\_variables)

> variable<-colnames(usable\_data) %in% names\_of\_numeric\_vector

> vector\_of\_charvariables<-usable\_data[!variable]

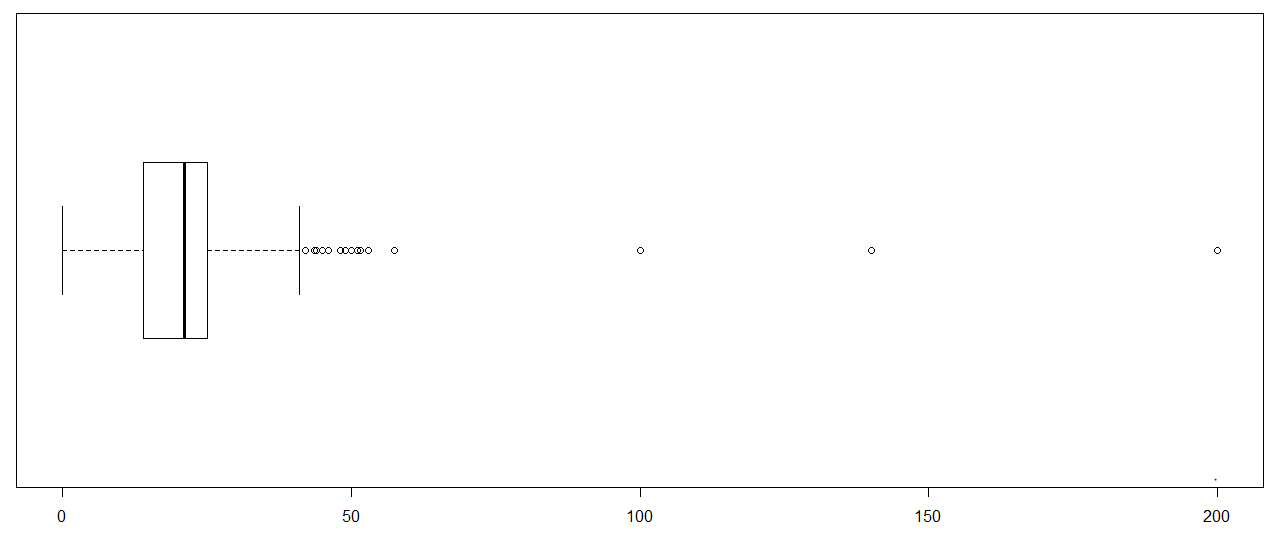
#detecting outliers from variable Skills score

>outlier\_detection<-set\_of\_numeric\_variables$Skills.score

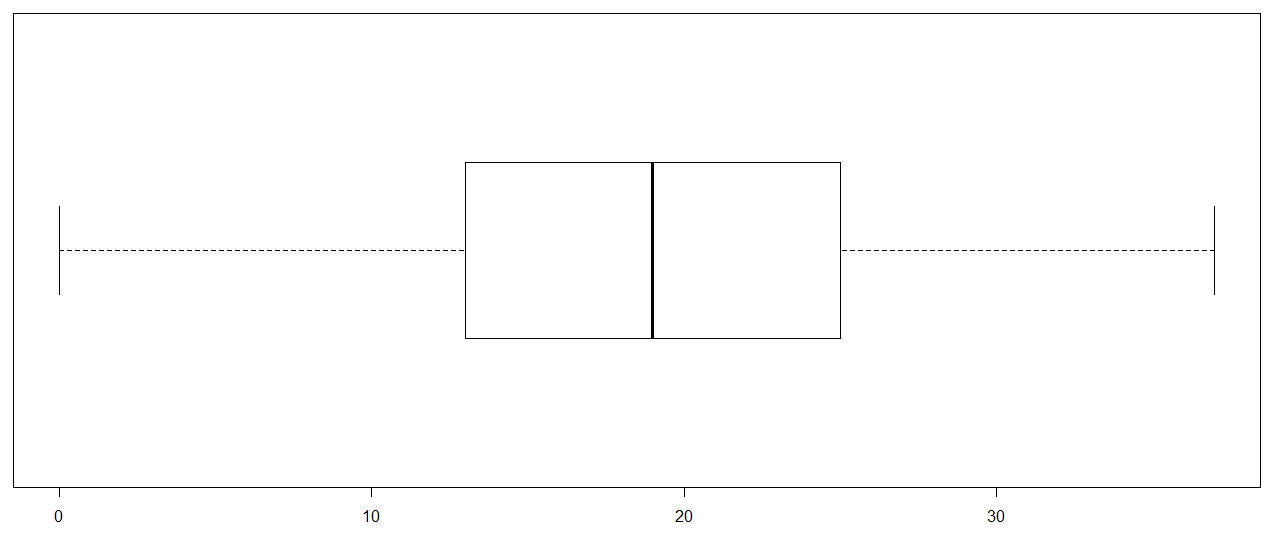
> hist(outlier\_detection)

> boxplot(outlier\_detection,horizontal=T)

> outlier\_detection\_1<-set\_of\_numeric\_variables$Skills.score[set\_of\_numeric\_variables$Skills.score>38]<-38



> boxplot(set\_of\_numeric\_variables$Skills.score,horizontal=T)



### Summary on data treatment and missing values

* In order to make sure consistency first of all by using “str” function it was identified that which variables are treated in numeric and character format by R. It was identified that R was treating date format variables in to character. Thus, by performing relevant function in date format variables Est. Founding Date and Last Funding Date is converted.
* Apart from this, all blank cells and cells that contain “No Info” were filled and replaced by NA. In this way missing values are filled by NA. There were few variables that contain numbers but treated like character by R. Some of these variables were year of founding and age of company in years. In order to convert all relevant columns in to numeric variable loop was run.
* In loop through indexing we instruct R to assume all elements of vector col as column [i] and convert them in to numeric format. In this regard as.numeric function is used.
* Our main target is to remove those variables from dataset that contain NA values extremely. This is because if any variable contain higher number of NA values then on the basis of that variable accurate prediction cannot be made. In this regard NA values in each column are summed by using SAPPLY function. We use function(x) to locate NA values in entire table and then by sum(is.na(x))) it is instructed to R to sum NA values.
* Finally, by using which statement like VLOOKUP which is used in excel we specify that only those variables in data frame that contain error values less than 40 must be included in the vector.
* Separate vector for numeric and character variables is prepared. First of all vector of numeric variables is prepared simply. Then from vector usable\_data by using %in% similar numeric variables identified. These identified numeric variables included in another vector namely “variable”. At last from usable data which contain both numeric and character variables by using “!variable” all numeric variables were removed from vector “usable data”. Thus, only character variable remain at end in vector “vector\_of\_charvariables”.
* In order to detect outliers boxplot chart is used and under this points at which there are outliers are identified. Step by step through indexing values less then specific number are contained in vector. In this way outliers are removed.

### Section 2: Details of additional features created

#creation of additional features

> vector\_of\_charvariables$Investors.count<-length(strsplit(vector\_of\_charvariables$Investors, "|",fixed=T))

> for (i in (1:length(vector\_of\_charvariables$Investors)))

> {if(is.na(vector\_of\_charvariables$Investors[i])==T){vector\_of\_charvariables$Investors.count[i]<-NA}

+ else{lst<-strsplit(vector\_of\_charvariables$Investors[i],"|", fixed=T)

+ vector\_of\_charvariables$Investors.count[i]<-length(lst[[1]])}}

***Summary of additional features created***

Additional feature is created to separate names of individuals from each other in each and every single cell that comes under column labeled “Investors”. In this regard loop is prepared under which we define range of data by (1:length(vector\_of\_charvariables$Investors). After this IF statement is prepared under which we instruct R to split all names in single cell. Further, we instruct R to count by considering split names in each cell. In this way, entire additional feature work.

### Section 3: Graphical presentation of your exploration and its interpretation

#creation of vector containing variable dependent company status

>set\_of\_numeric\_variables$Dependent.Company.Status<-vector\_of\_charvariables$Dependent.Company.Status

#Coding of variable continent of company

> vector\_of\_charvariables$Continent.of.company[vector\_of\_charvariables$Continent.of.company=="North America"]<-1

> vector\_of\_charvariables$Continent.of.company[vector\_of\_charvariables$Continent.of.company=="Asia"]<-2

> vector\_of\_charvariables$Continent.of.company[vector\_of\_charvariables$Continent.of.company=="Europe"]<-3

> vector\_of\_charvariables$Continent.of.company[vector\_of\_charvariables$Continent.of.company=="South America"]<-4

#formulation of mode function

> Mode <- function(x) {

+ u <- unique(x)

+ u[which.max(tabulate(match(x, u)))]

+ }

#computation of mode for variable continent of company

> vector\_of\_charvariables$Continent.of.company[is.na(vector\_of\_charvariables$Continent.of.company)]<-Mode(vector\_of\_charvariables$Continent.of.company)

#loading of library ggplot 2 and plotting of continent of company and dependent company status

> library(ggplot2)

>ggplot(vector\_of\_charvariables,aes(x=Continent.of.company,y=Dependent.Company.Status,fill="blue"))+geom\_bar(stat = "identity")

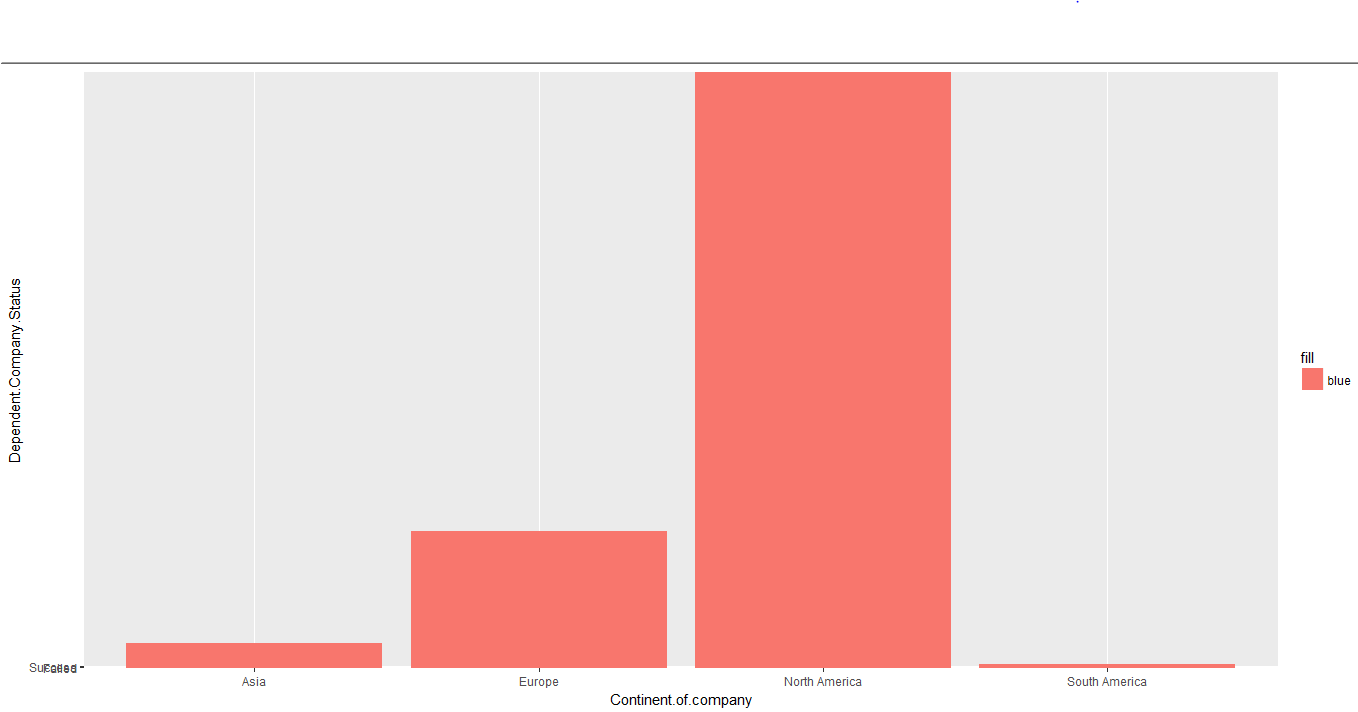


Figure 1Chart 1

#plotting of variables continent of company and skills score

> ggplot(vector\_of\_charvariables,aes(x=Continent.of.company,y=set\_of\_numeric\_variables$Skills.score,fill=Continent.of.company))+geom\_boxplot()

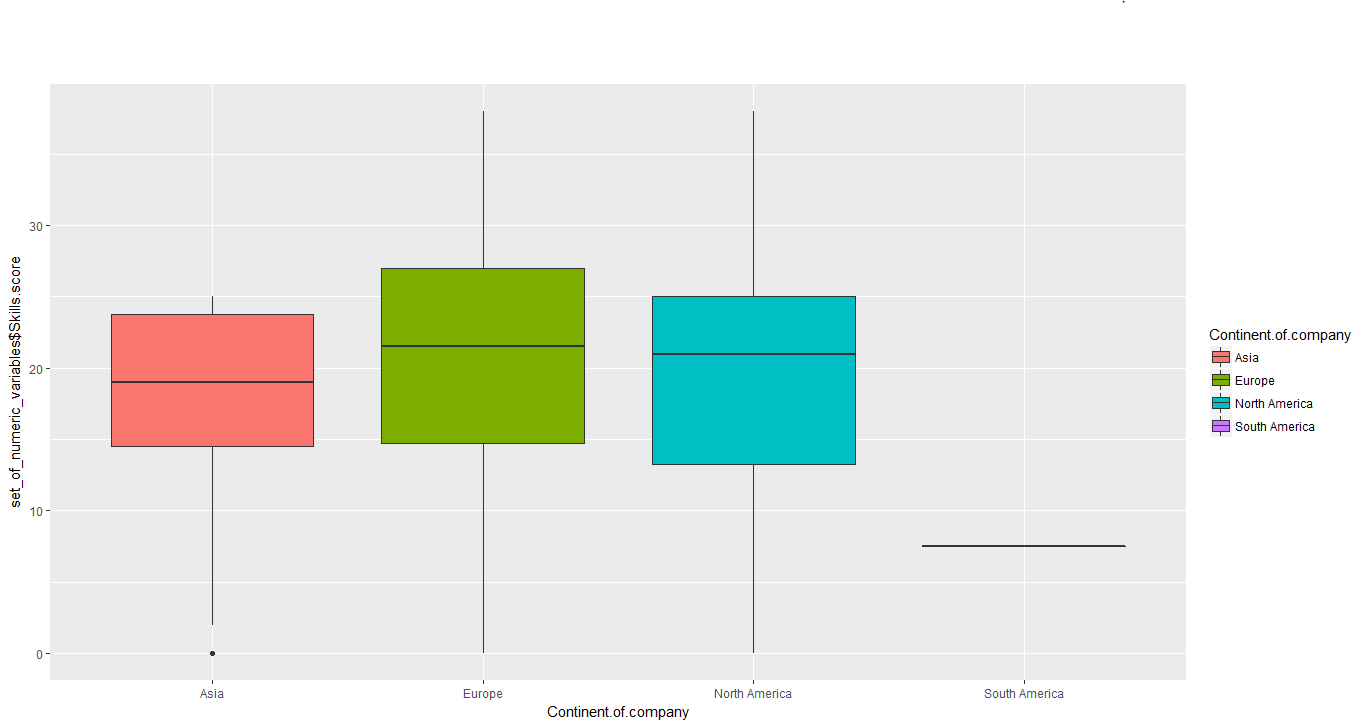


Figure 2Boxplot chart

### #tabulation of product or service providers and company status variable

### > counts<-table(vector\_of\_charvariables$Product.or.service.company,vector\_of\_charvariables$Continent.of.company)

### #charting of product or service providers and company status variable

> barplot(counts,main="Chart on Products and services offered across different continents",xlab="Continents",ylab="Products and services",col=c("Red","blue","green"),legend=rownames(counts))

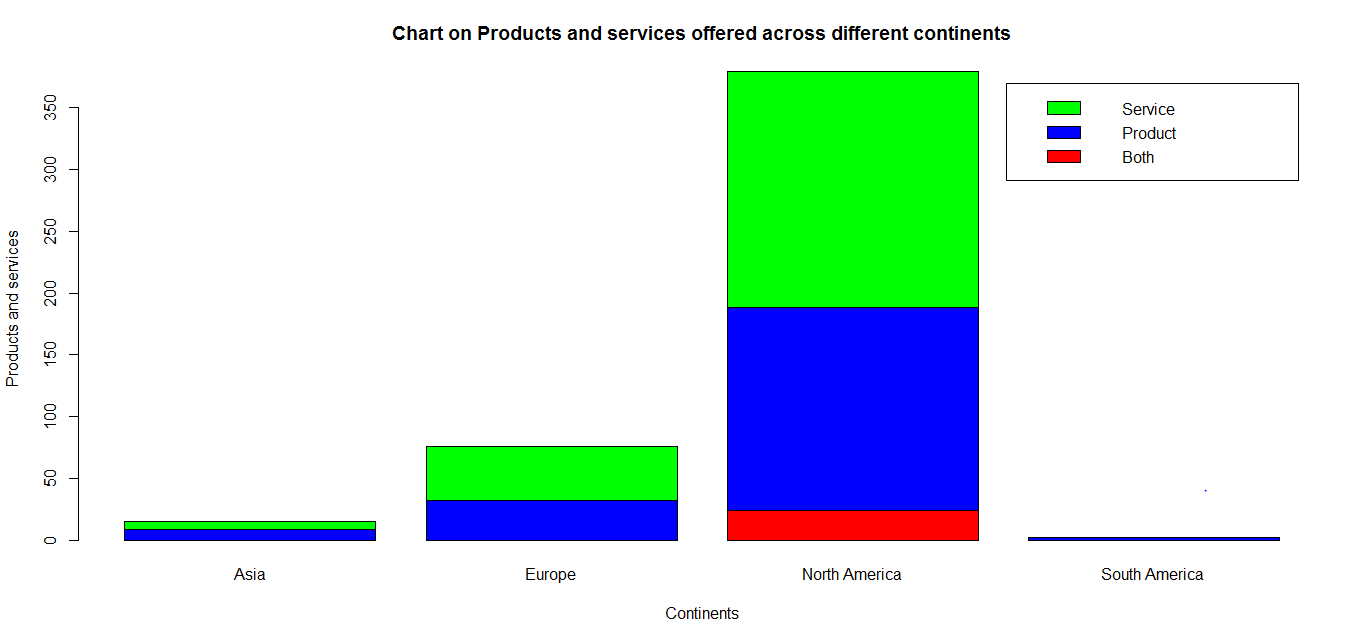


Figure 3Chart on products and services offered across different continents

>#replacing NA values by mode value for variables cloud and platform based sieve product and average size of companies worked in past

>vector\_of\_charvariables$Cloud.or.platform.based.serive.product[is.na(vector\_of\_charvariables$Cloud.or.platform.based.serive.product)]<-Mode(vector\_of\_charvariables$Cloud.or.platform.based.serive.product)

> vector\_of\_charvariables$Average.size.of.companies.worked.for.in.the.past[is.na(vector\_of\_charvariables$Average.size.of.companies.worked.for.in.the.past)]<-Mode(vector\_of\_charvariables$Average.size.of.companies.worked.for.in.the.past)

# coding of cloud and platform based firms

> vector\_of\_charvariables$Cloud.or.platform.based.serive.product[vector\_of\_charvariables$Cloud.or.platform.based.serive.product=="Platform"]<-1

> vector\_of\_charvariables$Cloud.or.platform.based.serive.product[vector\_of\_charvariables$Cloud.or.platform.based.serive.product=="Cloud"]<-2

### > vector\_of\_charvariables$Cloud.or.platform.based.serive.product[vector\_of\_charvariables$Cloud.or.platform.based.serive.product=="Both"]<-3

> vector\_of\_charvariables$Cloud.or.platform.based.serive.product[vector\_of\_charvariables$Cloud.or.platform.based.serive.product=="none"]<-4

> vector\_of\_charvariables$Cloud.or.platform.based.serive.product[vector\_of\_charvariables$Cloud.or.platform.based.serive.product=="cloud"]<-2

# Creation of table for tabulation of variable cloud or platform based service and average size of companies

> counts\_size<-table(vector\_of\_charvariables$Cloud.or.platform.based.serive.product,vector\_of\_charvariables$Average.size.of.companies.worked.for.in.the.past)

#barplot chart of cloud or platform based firms and size of companies

> barplot(counts\_size,main="Firms providing cloud or platform services in different size categories",xlab="Size of firms",ylab="Cloud or platform based serive/product",col=c("Red","blue","green","yellow"),legend=rownames(counts\_size))

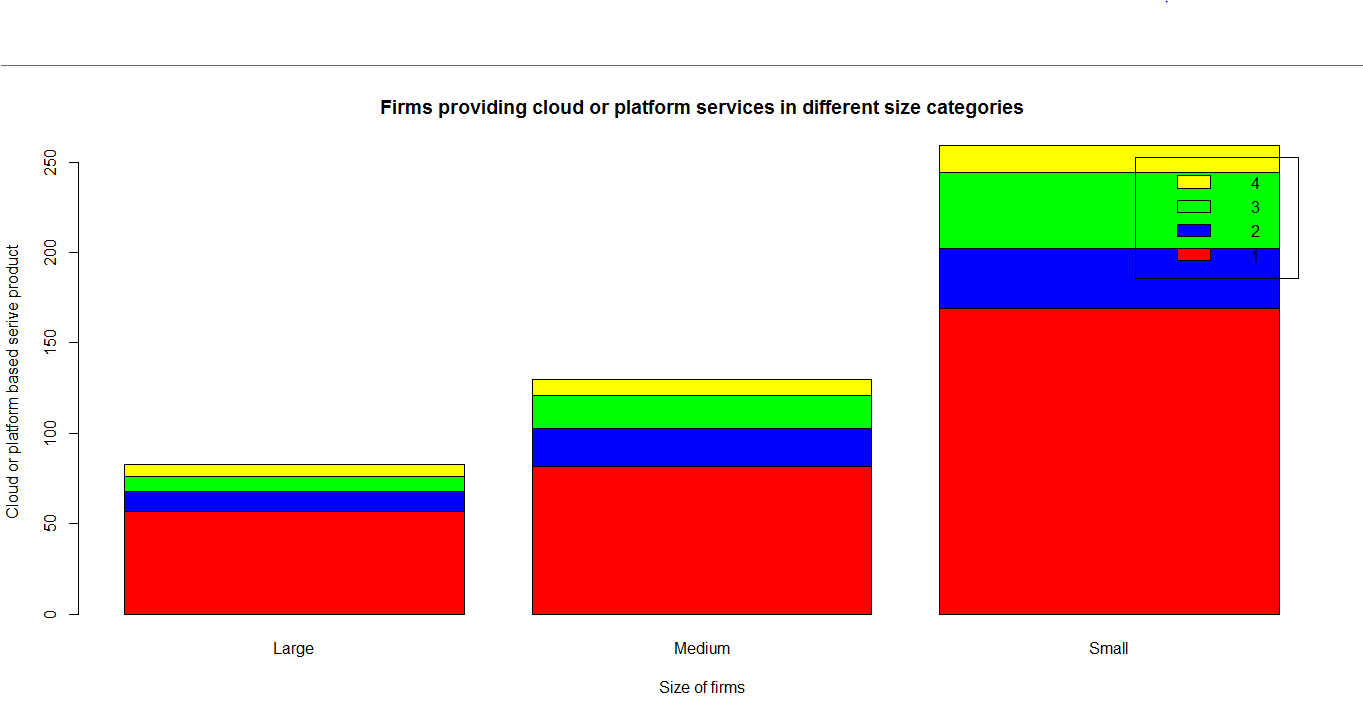


Figure 4Firms providing cloud or platform services in different size categories

### #tabulation of variables dependent company status and cloud or platform based services

### > tab<-table(vector\_of\_charvariables$Dependent.Company.Status,vector\_of\_charvariables$Cloud.or.platform.based.serive.product)

# barplot of dependent company status and cloud or platform based services

> barplot(tab,main="Products and services classification on the basis of success as well as failure",xlab="Products and services",ylab="Dependent-Company Status",col=c("red","green"),legend=rownames(tab))

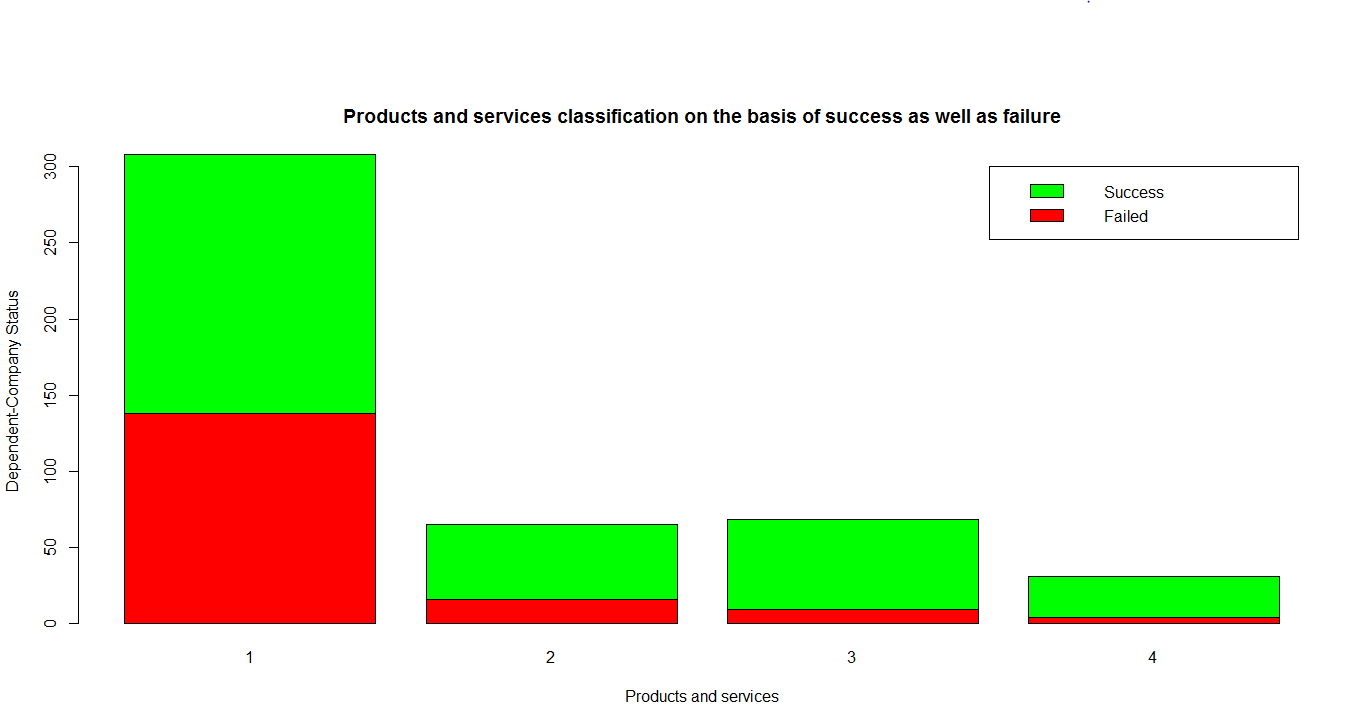


Figure 5Product and services providers across different continents

### Section 4: Details of statistical tests conducted and its interpretation

#formulation of mode function

> Mode <- function(x) {

+ u <- unique(x)

+ u[which.max(tabulate(match(x, u)))]

+ }

# computation of mode for variable continent of company

> vector\_of\_charvariables$Continent.of.company[is.na(vector\_of\_charvariables$Continent.of.company)]<-Mode(vector\_of\_charvariables$Continent.of.company)

#Chi square test

> input<-table(vector\_of\_charvariables$Dependent.Company.Status,vector\_of\_charvariables$Cloud.or.platform.based.serive.product)

> chisq.test(input)

Pearson's Chi-squared test

data: input

X-squared = 36.802, df = 4, p-value = 1.979e-07

#filling of NA values by median

> set\_of\_numeric\_variables$Skills.score[is.na(set\_of\_numeric\_variables$Skills.score)]<-median(set\_of\_numeric\_variables$Skills.score,na.rm=T)

#t-test for identifying significant mean difference

> t.test(set\_of\_numeric\_variables$Skills.score~vector\_of\_charvariables$Dependent.Company.Status,data=vector\_of\_charvariables)

Welch Two Sample t-test

data: set\_of\_numeric\_variables$Skills.score by vector\_of\_charvariables$Dependent.Company.Status

t = -1.0421, df = 443.39, p-value = 0.2979

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.3949410 0.7351589

sample estimates:

mean in group Failed mean in group Success

19.81437 20.64426

### Section 5: Summary of your exploration

***Interpretation of figure 1***

* In order to analyze dataset charting of variables is done. It can be observed from the chart 1 that in case of North America there are higher numbers of success and failure companies.
* In case of Asia and South America there are few firms that come in category of success or failure. Apart from North America, Europe is the continent where number of firms is large.
* It can be said that in case of North America and Europe there are large number of firms then Asia and South America.

***Interpretation of figure 2***

* Skill score is very low in case of South America but in other relevant geographic area its value is almost same as reflected by boxplot chart. In case of North America skill score value change at rapid pace.
* Same thing is not observed in case of Asia and South America. It can be said that in case of North America there is high skill score then Asia, Europe and South America.

***Interpretation of figure 3***

* In North America as reflected by stacked bar chart there are almost equal number of firms that are producing products and services (see figure 3). Contrary to this in case of Asia continent most of firms are producing products and there are few firms that provide services to their customers.
* Trend of North America and Europe is same in terms of production of products and services. In case of South America only products are produced. It can be concluded that in equal proportion products are produced and services are provided by firms in North America and Europe.
* There are few firms that are producing product and provide service to the customers in South America. Interesting fact is that all these firms are in large number located in North America.

***Interpretation of figure 4***

* Stacked bar chart is given in sheet and from seen it can be seen that in large size firms majority of them are providing platform based services. As there are few firms that are providing cloud, both or no service.
* In case of medium size a firm same trend exist but number of firms that provide cloud and both services as well as none of them increase slightly. Number of firms in these three mentioned categories slightly increases more in small firm category.
* However, still proportion of firms providing platform based services is high across large, medium and small size firms are high.

***Interpretation of figure 5***

* In case of North America proportion of successful firms is slightly more then failed firms. However, in case of Asia number of successful firms is much higher than failure firms. In case of Europe also same trend is observed.
* In case of South America also same trend persists but numbers of successful firms are low in case of mentioned geographic area then North America. It can be said that there are good business friendly environment in Asia, Europe and South America then North America.

***Chi square test***

***Chi square test:*** It is a test that is used to explore relationship between two categorical variables. In present case two categorical variables are cloud or platform based firms and dependent company status.

H0: There is no significant association between cloud, platform or firms that are providing or none of these services in terms of success and failure status.

H1: cloud, platform or firms that are providing or none of these services in terms of success and failure status.

Chi square test level of significance is 1.97>0.05 reflecting that there is no significant relationship between cloud, platform or firms that are providing or none of these services in terms of success and failure status. Hence, it can be said that null hypothesis is accepted.

***T-test***

***T-test:*** It is statistical tool that is used to identify mean difference between two variables which are company status and skill score in below given case.

H0: There is no significant mean difference between company status and skill score.

H1: There is significant mean difference between company status and skill score.

T test level of significance value is 0.2979>0.05 which indicate that there is no significant mean difference between variables which are company status and skill score. Means that in case of success and failure of firms no significant difference is observed between them in terms of skill score. It can be assumed that there was almost similar skill score across success and failed firms. It can be said that because level of significance is greater than 0.05and due to this reason there is no significant mean difference between company status and skills score.