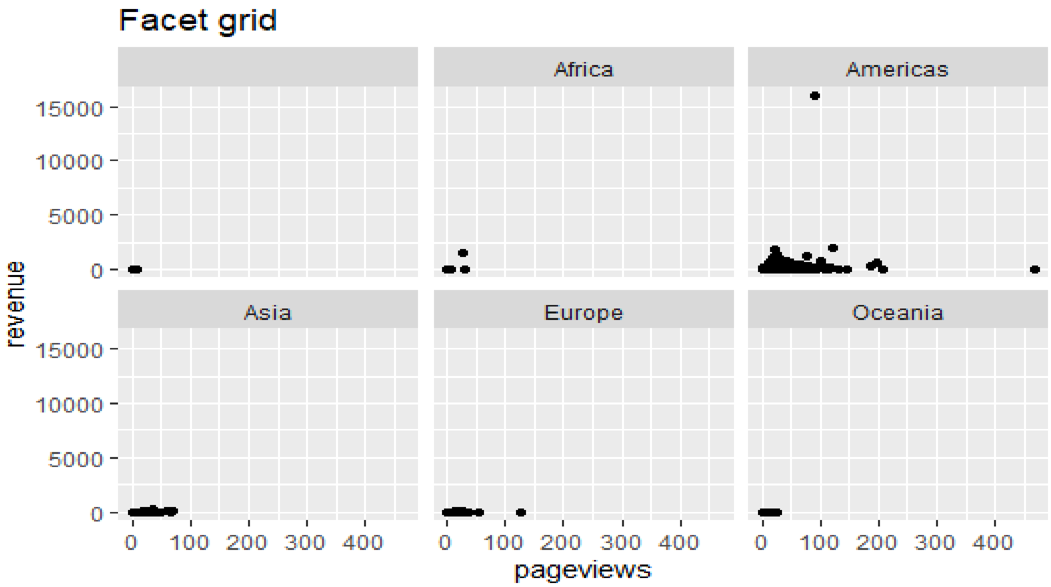
(Team:001-Mega) IDA-HW 4&5 [Group 2] Meghnath Reddy & Shan Peng

Part A) Exploratory Data Analysis (EDA)

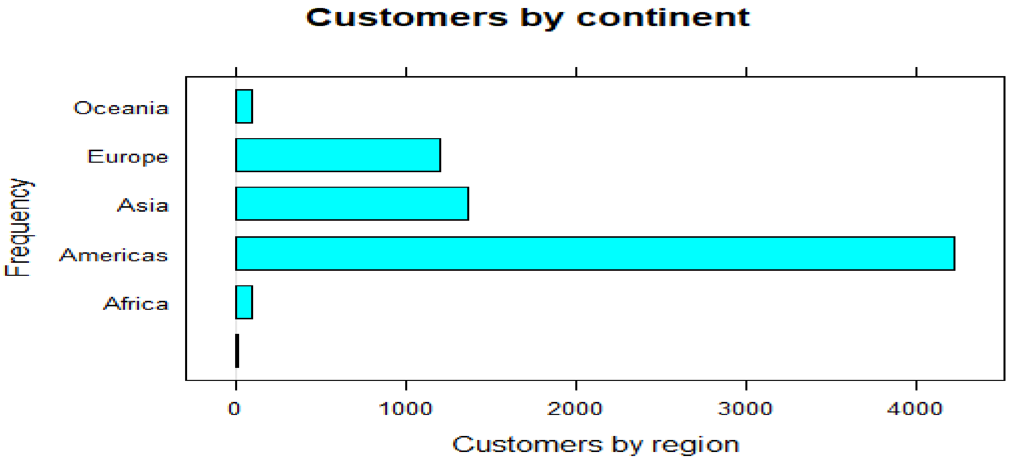
1) Which Continent is generating more revenue?

```
ggplot(data = dfSample) + geom_point(mapping = aes(x = pageviews, y = revenue))+ facet_wrap(~ continent,
nrow = 2)+labs(title="Facet grid") #Facet grid
```

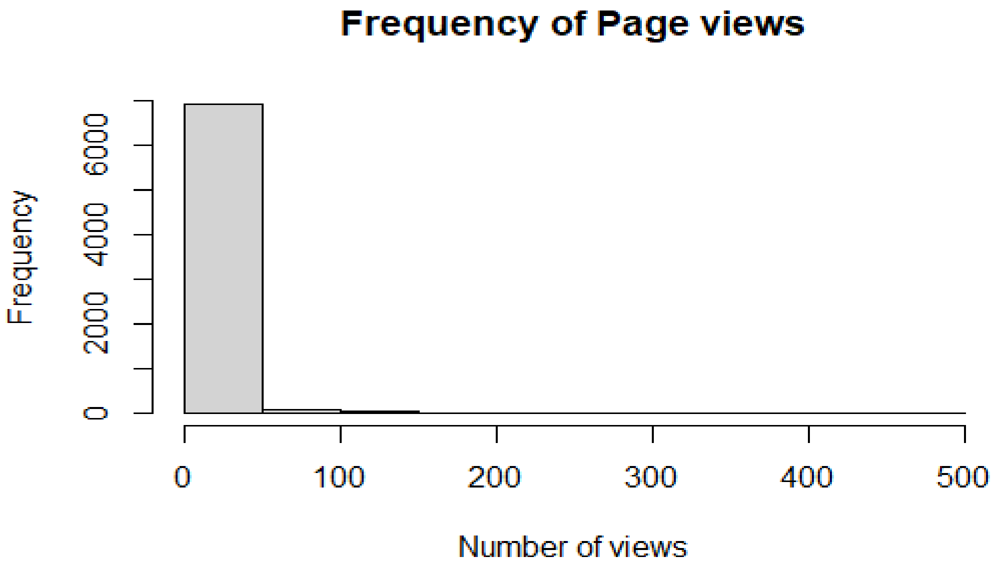


Insight: Many page views are coming from America which results in more income whereas Oceania has fewer views obviously due to lack of internet and population.

Let's verify that by creating a plot showing customers group by Continent.

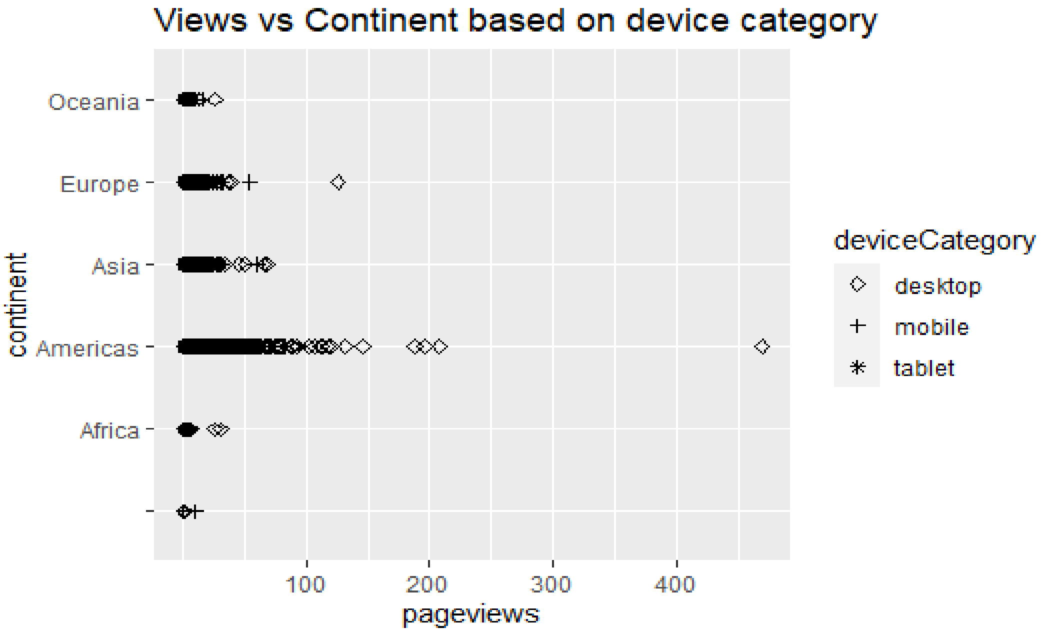


2) **What is the range of frequency of pageviews?**



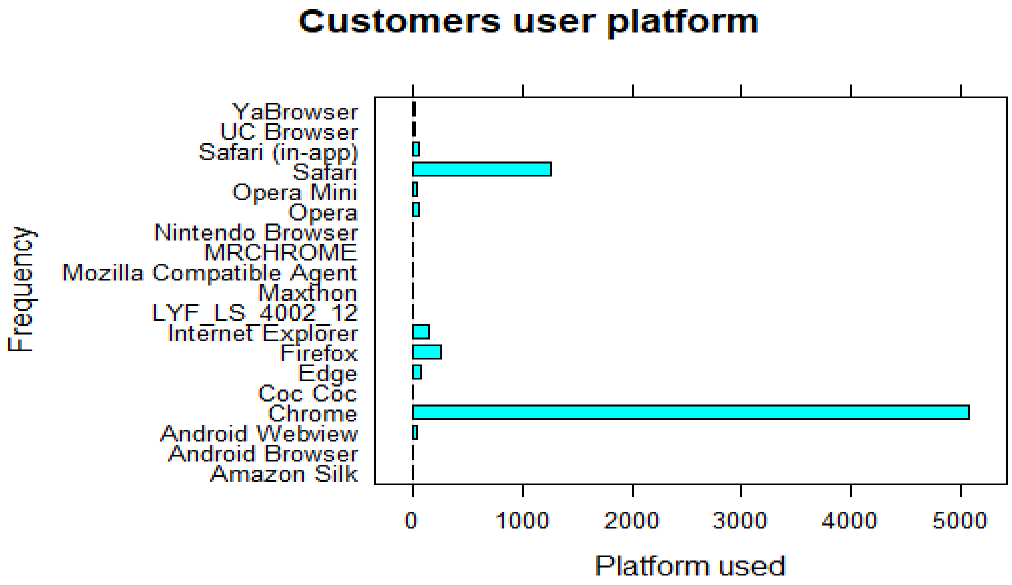
Insight: Majority of pageviews seems to be lying under 50 views.

3) **What kind of device is used across the Continents and Is there any relationship between increase in views to type of device used?**



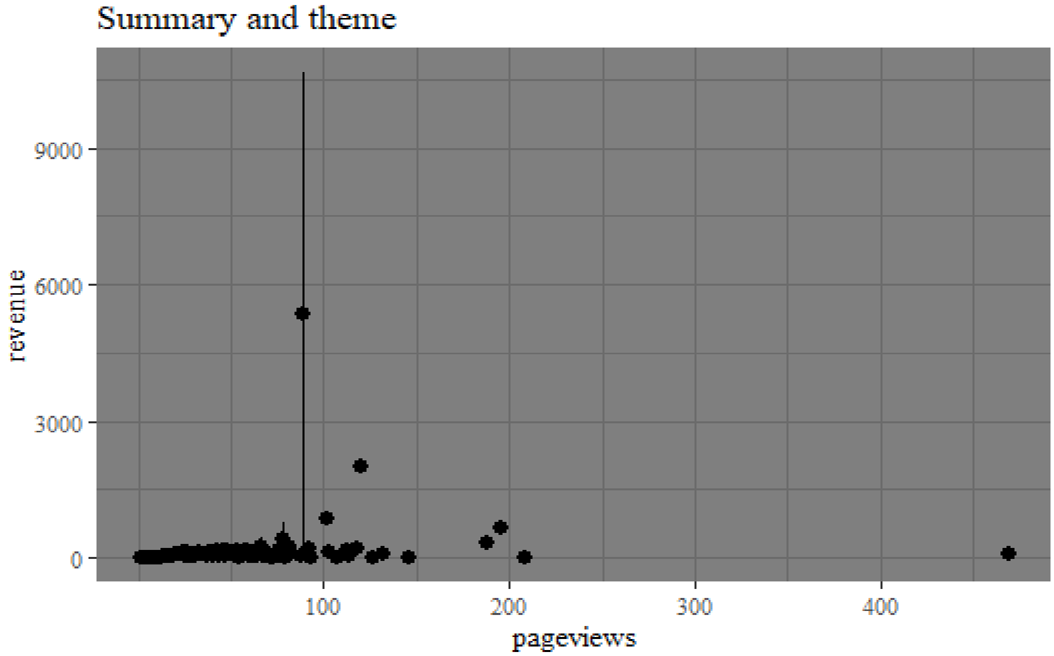
Insight: Desktop seems to generate more page views.

4) What is most common kind of platform used by Customer?



Insight: As expected most of the customers prefer Chrome as a medium to purchase followed by Safari. Some of the browser platforms seem to have a smaller number of the customer base the reason being that platform might be popular or specific to that region/continent.

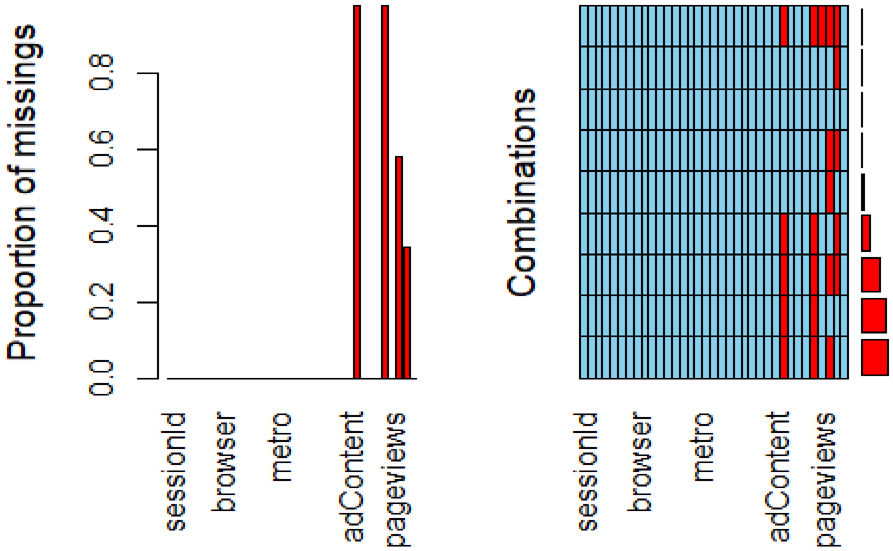
5) Does more pageviews result in more revenue?



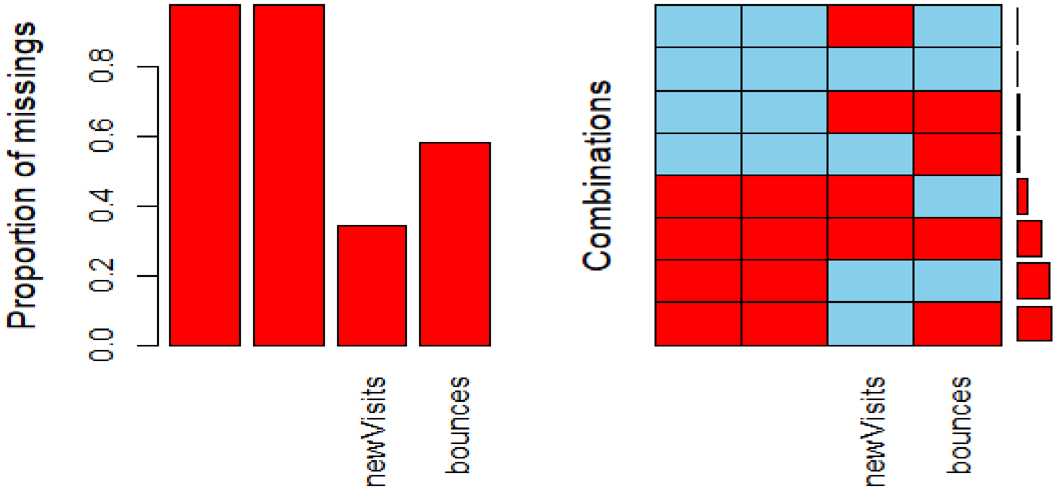
Insight: Even the one with more than 150page views seems to generate almost the same revenue as that of less than 10 views. So, more views more revenue is a misconception.

6) Which feature is having the most frequent Missing values and is there any correlation in missingness of those features?

Insight: adwordsClickInfo.isVideoAd, adwordsClickInfo.page have SAME number of NA'S (68260). some pattern of missingness exists between newVisits and Bounces.

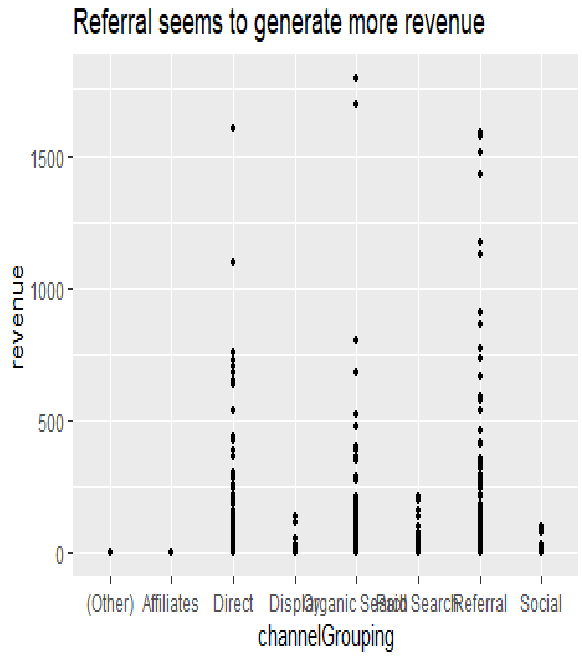


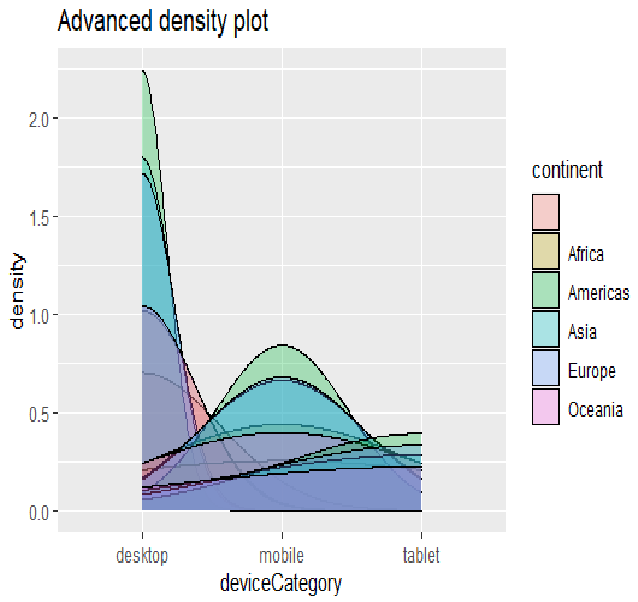
Relation between top 4 missing values:

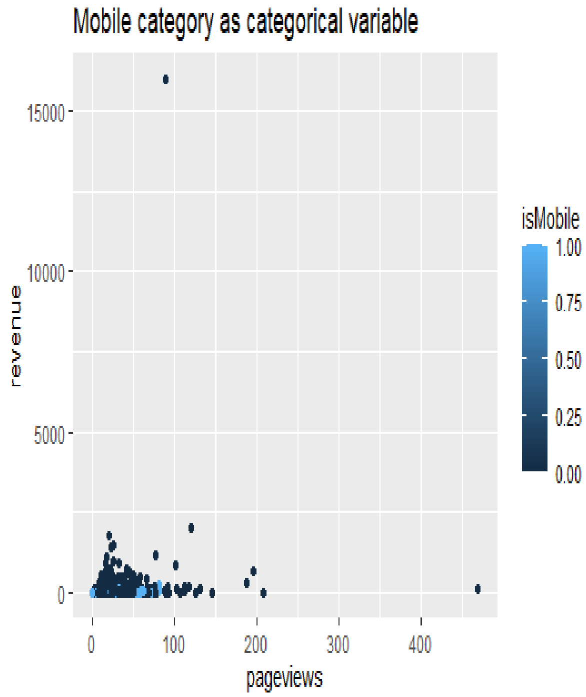


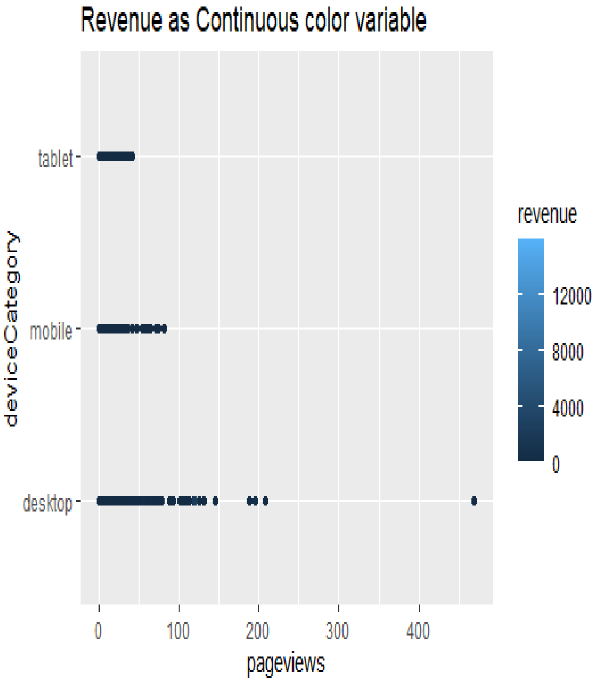
Only 264 out of 70071 observation have no missing values when all the features are considered.

Some interesting visual finding









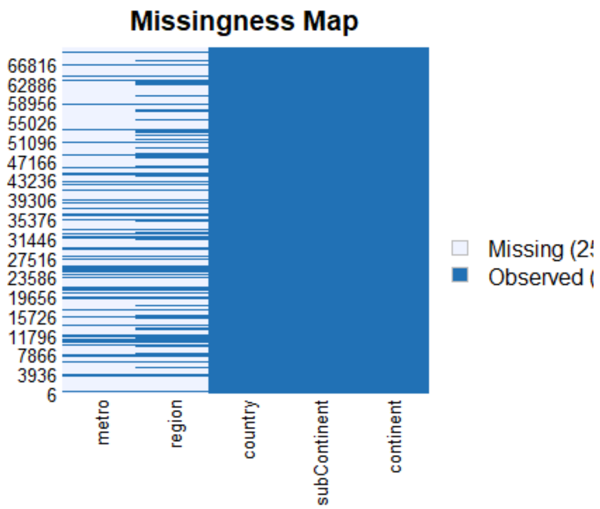
Insight: Referral seems to be popular means of drawing customer attention whereas Organic search generated the highest revenue. Desktop has lot of pageviews and seems to bring in more revenue. The highest revenue generated by Mobile devices seems to than 250$ and most of the views are below 100.

DSA5103_001_Group2_PartB

Missing Value Imputation - 1

In train data, the variances of address for customers include continent, subcontinent, country, region and metro. We change the empty value of these variances into Null value, and generate a missing value map. We find out the missing value in country, subcontinent, and continent is much less than metro, region. So, in our later model creating, we prefer to use the variances of country, subcontinent and continent as predictors.

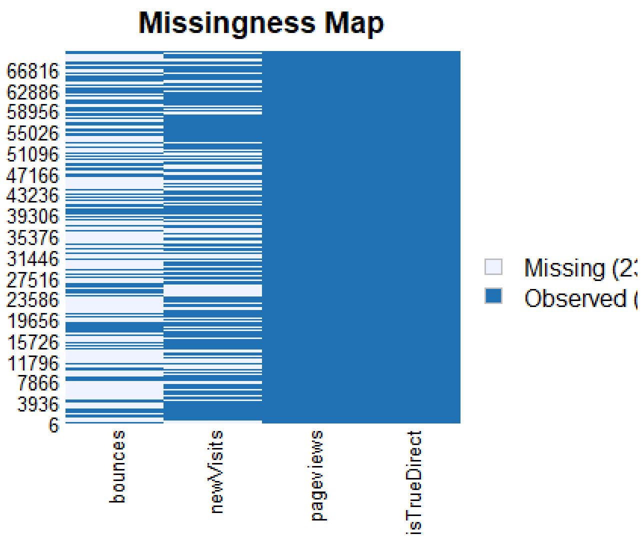
```
#find out the missing value of address of customers
train[train$continent=="",'continent'] <- NA
train[train$subContinent=="",'subContinent'] <- NA
train[train$country=="",'country'] <- NA
train[train$region=="",'region'] <- NA
train[train$metro=="",'metro'] <- NA
missmap(train[,12:16])
```



Missing Value Imputation - 2

By showing the missing value map for the variances of newVisits, isTrueDirect, bounces and pageviews, we find out the missing value of these variances should be equal to the meaning of zero. So we change all missing value of these variances into 0 and use these variances as predictors to create our model.

```
#find out the missing value of newVisit, isTrueDirect, pageviews, bounces whi
ch are all binary and transform
missmap(train[,c("newVisits","isTrueDirect", "bounces", "pageviews")])
```



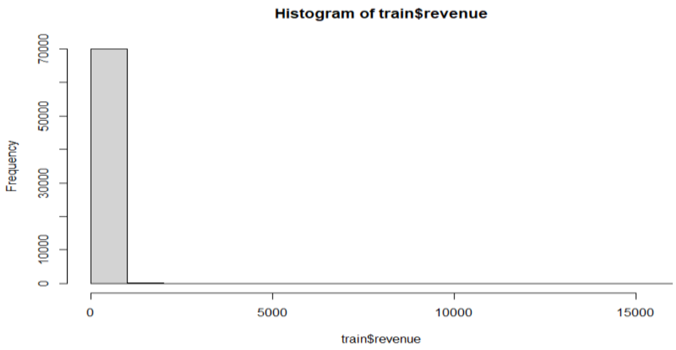
The imputation and transformation of missing value in the variances of newVisits, isTrueDirect, bounces and pageviews.

```
train[is.na(train$newVisits),"newVisits"] <- 0
train[is.na(train$isTrueDirect),"isTrueDirect"] <- 0
train[is.na(train$pageviews),"pageviews"] <- 0
train[is.na(train$bounces),"bounces"] <- 0
```

Resolution of outliers

The predict result of revenue is the most important variance for us to create a prediction model. The outliers of the variance: revenue has strong influence in the model, so we need to check the status of the variance: revenue. From the visualized histogram, we can see most value of revenue stays in the range from 0 to 1000. So the outliers of revenue should be removed by using the method of outliers.

```
#outlier
hist(train$revenue)
```



```
grubbs.test(train$revenue)

##
##  Grubbs test for one outlier
##
## data:  train$revenue
## G = 160.45368, U = 0.63257, p-value < 2.2e-16
## alternative hypothesis: highest value 15980.79 is an outlier

outlier(train$revenue)

## [1] 15980.79

train <- train[train$revenue!=outlier(train$revenue),]
```

Transformation of dummy variances

For the category variances, such as continent and subcontinent, we need to transform these category variances into dummy variances as predictors in creating a prediction model. The method of dummyVars in the packet: caret can be used to transform category variances.

```
#category variance transform into dummy variances
travarience <- train[1:5,c("custId","continent")]
dmy <- dummyVars(~.,data=travarience,fullRank = TRUE)
traf <- data.frame(predict(dmy,newdata = travarience))
traf

##   custId continent.Africa continent.Americas continent.Asia continent.Euro
pe
## 1   1795                0                  0              1               
0
## 2   1797                0                  1              0               
0
## 3   1799                0                  0              1               
0
## 4   1800                1                  0              0               
0
## 5   1801                0                  1              0               
0
##   continent.Oceania
## 1                 0
## 2                 0
## 3                 0
## 4                 0
## 5                 0
```

Feature engineering and Feature extraction

We can create a prediction model from the perspective of sessionId or custId. If we choose to create a model of sessionId, firstly the rows of data are too large which is over 70,000, and then the error of final results in the model will be aggregated to more by computing the target revenue. So we will create a model from custId and then we need to extract the related features in the data . By grouping with custId, we aggregate the sum value of revenue for each customer, and other related features: pageviews, newVisits, isTrueDirect, bounces, and continent. By constructing a new dataset from the perspective of custId, we will use the new dataset to create different prediction models.

```
#feature extraction and transformation
trainmd <- train %>% group_by(custId) %>%
                    summarise(sumRevenue = sum(revenue),
                              sumViews = sum(pageviews),
                              newVisits = max(newVisits),
                              sumDirect = sum(isTrueDirect),
                              sumBounce = sum(bounces),
                              groupContinent = continent)
```

Dimension reduction using PCA

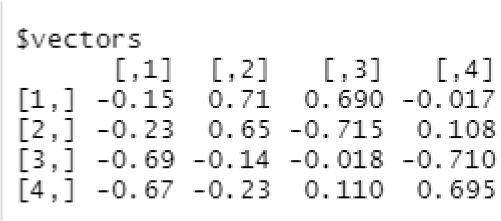
To remove the duplicated row.

unique_rows<-unique_rows[!duplicated(sales),] #find and remove the du plicate row

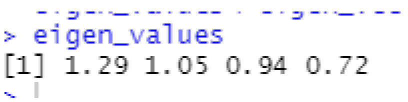
Since PCA works better with numerical value we avoid Type attribute.

my_num_data <- unique_rows[, sapply(unique_rows, is.numeric)] #remove non-numeric

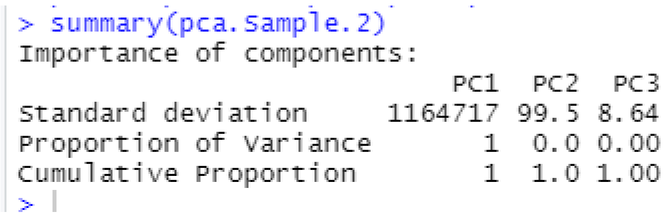
Eigen Vectors:



```
$vectors
      [,1]  [,2]   [,3]   [,4]
[1,] -0.15  0.71  0.690 -0.017
[2,] -0.23  0.65 -0.715  0.108
[3,] -0.69 -0.14 -0.018 -0.710
[4,] -0.67 -0.23  0.110  0.695
```

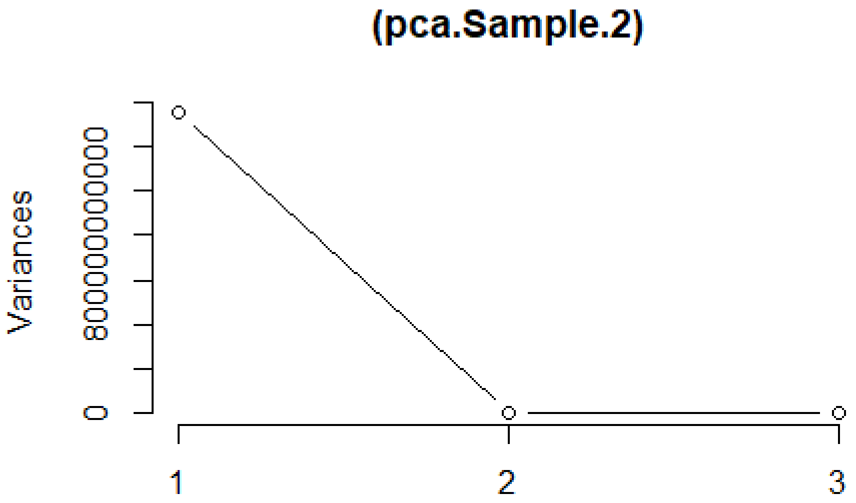


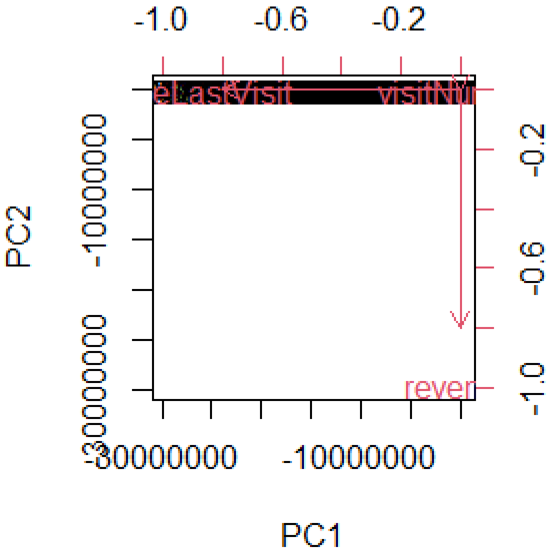
```
> eigen_values
[1] 1.29 1.05 0.94 0.72
```



```
> summary(pca.Sample.2)
Importance of components:
                           PC1  PC2  PC3
Standard deviation     1164717 99.5 8.64
Proportion of Variance       1  0.0 0.00
Cumulative Proportion        1  1.0 1.00
> |
```

Insight: Most of the variability is accounted at PC1 but Standard deviation is very high and Decreases by the time we reach PC3.





DSA5103_001_Group2_PartC

Part C Summary of data preparation and Explanation

In the train dataset, we found there are too many variances and values to select and choose. We need to determine which variances are most likely useful as predictors in the model. Firstly, the variances of date and time can be excluded because there are different values of these variances in different sessions and it is not suitable to compute in the prediction model. Same reason happens in the variances of networkDomain and referalPath. There are too many missing values in the variances of adWords., and it is too hard to impute these missing values. The category variances of channelGrouping, browser and medium are likely to use in the model. And the binary variances of isMobile, isTrueDirect, bounces and newVisits are also likely to use in the model. Finally, through the analysis of data preparation step, we summarize the chosen of variances as below:

1) The importance variance of pageviews is strongly related to the prediction results of revenue, and the missing values of pageviews will be imputed as zero to use in the model;

2) The binary variances of isTrueDirect, bounces and newVisits are chosen to add in the model. The missing values of these variances are imputed as zero to use in the model;

3) The category variance of continent is chosen as a predictor in the model. There are little missing values in the variance of continent, and the variance of continent is highly related to revenue. Finally, we transform the category variance into the dummy variances for computing in the model.

And the explanation of our choice of the best model we think is described as below:

1) Grouping by the variance of custId, we use sum of pageviews in the model.

2) For the binary variances, newVisits is computed as maximum, so the real new visit customer could be counted in the model. The variances of isTrueDirect and bounces is computed as sum to get the visit frequencies of customers. These binary variances are multiplied as predictors in the model.

3) The category variance of continent is grouped by custId and as dummy variances in the model to computed.

4) These three parts of predictors are plus in the liner model to predictor the results.

Part D) Modeling

Model	Method	Package	Hyperparameter	Selection	R^2	RMSE
OLS	lm	stats	NA	NA	**0.434**	**1.531**
OLS	lm without intercept	stats	NA	NA	**0.568**	**1.531**
MARS	Earth	Earth	Degree	5	**0.734**	**1.05**
PLS	Partial least Square	pls	**NA**	**Validation=CV**		**406**

ii) The modeling problem was faced due to the null values and non-numeric parameters always resulted in errors. Our best R-squared value was found for the MARS model using the Earth package. By changing the hyperparameter for different values of degree. The R-squared value kept on changing and the better result was found at 5th degree. So, by chopping and changing we found different results. Same kind of behavior was found in PLS model by changing the tune length and number with different values.

iii) Submitted the results in Kaggle and got the best result of R squared value as **0.76577.**