Rajiv Kumar

Wednesday, January 14, 2016

### Background:

1. ~700k rows of ranking data (App\_ID Country Date Hour Device Rank)

2. ~ 3.5k rows of download data (App\_ID Country Date Device Downloads)

The rank of an app for a given (country, date, hour, device) combination represents its position in an ordered list of app\_ids associated with the combination.

For example, in USA, on 5/13/14 at 20:00:00 hours, the apps on iPhone top free charts had the following order [454, 1549, 2171]. App 454 was rank 1, app 1549 was rank 2 and app 2171 was rank 3.

We'd like you to try and answer the following questions for these data sets (below):

#### 1):What are your initial observations about the ranking and download datasets?

### Answer:

First of all lets load the data sets.

#read the given download info csv file  
df\_app\_downloads <- read.csv(file= file.path("E:","Sensor tower/ST\_DA\_exercise/download\_info.csv"))  
  
  
#read the given ranking info csv file  
df\_app\_ranking<-read.csv(file= file.path("E:","Sensor tower/ST\_DA\_exercise/ranking\_info.csv"))

Data is loaded in to dataframe. Now, we will first analzye the structure of the data.

str(df\_app\_downloads)

## 'data.frame': 3514 obs. of 5 variables:  
## $ App.ID : int 39 51 132 136 148 151 157 179 191 208 ...  
## $ Country : Factor w/ 3 levels "Great Britain",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ Date : Factor w/ 14 levels "05/13/2014","05/14/2014",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Device : Factor w/ 2 levels "iPad","iPhone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Downloads: int 3493 6471 4771 4218 30823 3659 664 11692 2964 1069 ...

str(df\_app\_ranking)

## 'data.frame': 682400 obs. of 6 variables:  
## $ App.ID : int 1549 973 1969 1900 1483 558 2171 2056 1959 313 ...  
## $ Country: Factor w/ 3 levels "Great Britain",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ Date : Factor w/ 14 levels "05/13/2014","05/14/2014",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Hour : Factor w/ 24 levels "00:00:00","01:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Device : Factor w/ 2 levels "iPad","iPhone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Rank : int 1 2 3 4 5 6 7 8 9 10 ...

Now, we have the basic understanding of our data set.We will now see the different statistics of our data set.

summary(df\_app\_downloads)

## App.ID Country Date Device   
## Min. : 21 Great Britain:1176 05/24/2014: 265 iPad :1738   
## 1st Qu.: 617 South Africa :1281 05/26/2014: 264 iPhone:1776   
## Median :1140 USA :1057 05/23/2014: 263   
## Mean :1075 05/25/2014: 263   
## 3rd Qu.:1543 05/22/2014: 260   
## Max. :2206 05/21/2014: 256   
## (Other) :1943   
## Downloads   
## Min. : 1   
## 1st Qu.: 22   
## Median : 278   
## Mean : 1620   
## 3rd Qu.: 1483   
## Max. :80080   
## NA's :73

summary(df\_app\_ranking)

## App.ID Country Date   
## Min. : 1 Great Britain:211200 05/18/2014: 56000   
## 1st Qu.: 495 South Africa :244800 05/17/2014: 54400   
## Median :1024 USA :226400 05/21/2014: 54400   
## Mean :1033 05/22/2014: 54400   
## 3rd Qu.:1526 05/25/2014: 54400   
## Max. :2211 05/19/2014: 52800   
## (Other) :356000   
## Hour Device Rank   
## 01:00:00: 32000 iPad :341200 Min. : 1.0   
## 02:00:00: 31200 iPhone:341200 1st Qu.:100.8   
## 03:00:00: 31200 Median :200.5   
## 05:00:00: 31200 Mean :200.5   
## 11:00:00: 31200 3rd Qu.:300.2   
## 04:00:00: 30400 Max. :400.0   
## (Other) :495200

To answer the question, from the first level analysis (without doing any plotting or further investigation), i have following observations about the data sets:

#### df\_app\_downloads data set:

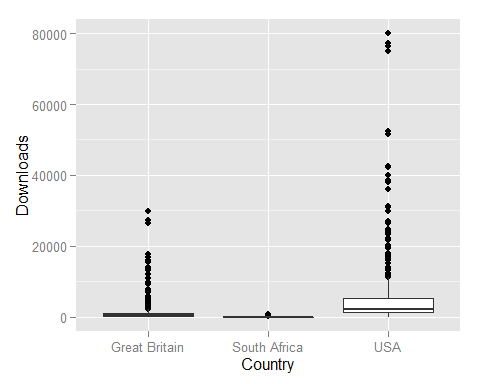
1. Data is highly skewed in terms of # of downloads. Mean is much greater than the median, which means there are high chances of outliers in the data.
2. Max value for # of downloads is 80080, which is way greater than IQR but given the business scenario these high number could be just the high number of downloads for top ranked apps.
3. There are 73 Missing values present in Downloads column, this concern needs to be addressed before we reach the modeling stage.
4. Data is approximately evenly distributed among 3 countries.
5. Data is approximately evenly distributed among device type.
6. Data is approximately evenly distributed among each date.

#### df\_app\_ranking data set:

1. Data is approximately evenly distributed among 3 countries.
2. Data is evenly distributed among device type.
3. Data is approximately evenly distributed among each date.
4. Data is approximately evenly distributed among each Hour.
5. Number of Records with given ranking are equally distributed.

Presence of outliers in Downloads column can be seen from Boxplot

ggplot(aes(x=Country,y=Downloads),data=df\_app\_downloads)+geom\_boxplot()



We can see there are outliers and extreme values prsent in downloads column, especially for USA. However, it is too early to declare these values as outliers because these can be just high number of downloads for top ranked apps. We will see the relation as we go further in our analysis.

#### 2): What patterns do you see in the download dataset?

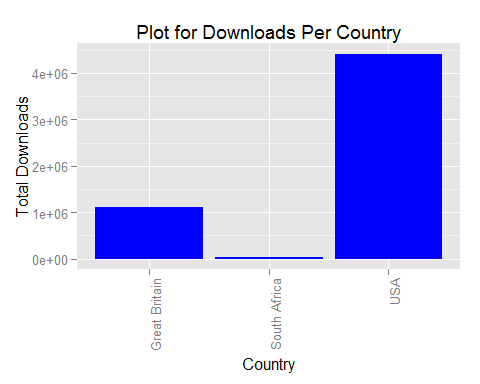
### Answer:

To answer the question, let’s do some exploratory analysis and see if there is any possibility of patterns in downloads dataset.

From the analysis, i found that:

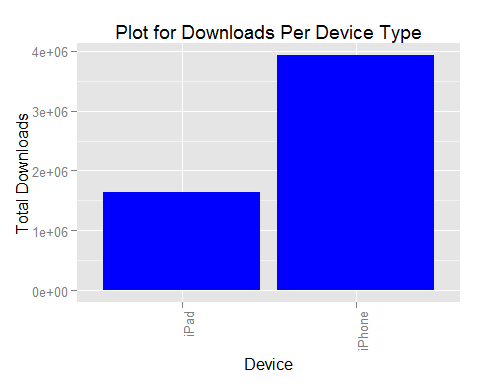
1. Number of downloads are highest in USA then in Great Britain and lowest in SouthAfrica. We can see that from following plot.

# Group the data set by Country  
df\_groupby\_country <- df\_app\_downloads %>% group\_by(Country)%>% summarise(total=sum(Downloads,na.rm=TRUE))  
  
ggplot(aes(x=Country,y=total),  
 data=df\_groupby\_country)+   
 geom\_bar(stat="identity",fill="Blue")+  
 ylab("Total Downloads")+  
 xlab("Country")+  
 ggtitle("Plot for Downloads Per Country")+  
 theme(axis.text.x=element\_text(angle=90, hjust=1))



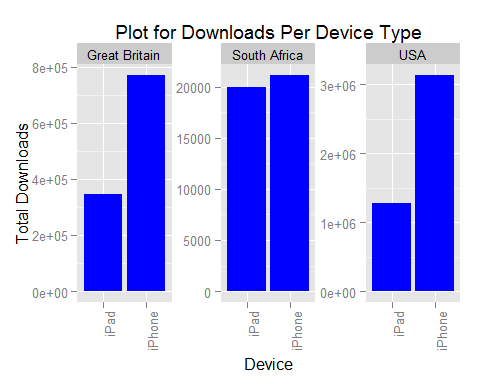
1. There are more number of downloads on iphone as compare to ipad in each country we can see this from followin plot;

# Group the data set by Device Type  
df\_groupby\_Device <- df\_app\_downloads %>% group\_by(Device)%>% summarise(total=sum(Downloads,na.rm=TRUE))  
  
ggplot(aes(x=Device,y=total),  
 data=df\_groupby\_Device)+   
 geom\_bar(stat="identity",fill="Blue")+  
 ylab("Total Downloads")+  
 xlab("Device")+  
 ggtitle("Plot for Downloads Per Device Type")+  
 theme(axis.text.x=element\_text(angle=90, hjust=1))



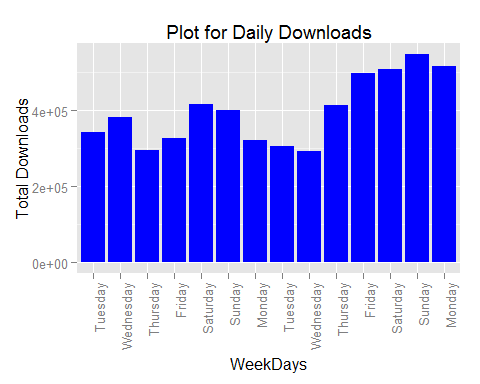
The above pattern holds true for each country as shown in below plot.Another interesting fact here is that for south africa number of downloads on ipad and iphone are very close.

# Group the data set by Device Type and country  
df\_groupby\_Device\_country <- df\_app\_downloads %>% group\_by(Device,Country)%>% summarise(total=sum(Downloads,na.rm=TRUE))  
  
ggplot(aes(x=Device,y=total),  
 data=df\_groupby\_Device\_country)+   
 geom\_bar(stat="identity",fill="Blue")+  
 ylab("Total Downloads")+  
 xlab("Device")+  
 ggtitle("Plot for Downloads Per Device Type")+  
 theme(axis.text.x=element\_text(angle=90, hjust=1))+  
 facet\_wrap(~Country,scales="free")



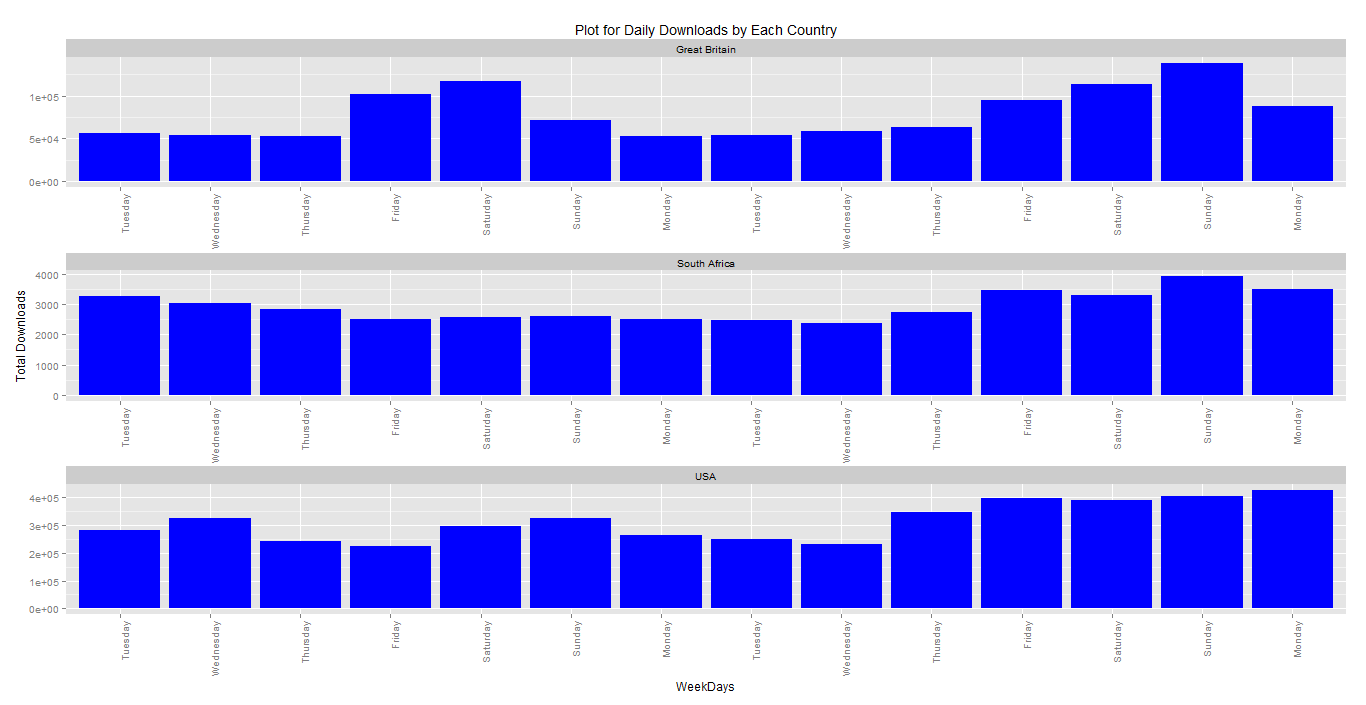
1. When i grouped the total downloads by date then i observed that downloads are kind of increasing as we are approaching weekends.

# Group the data set by date  
df\_groupby\_date <- df\_app\_downloads %>% group\_by(Date)%>% summarise(total=sum(Downloads,na.rm=TRUE))  
  
ggplot(aes(x=Date,y=total),  
 data=df\_groupby\_date)+   
 geom\_bar(stat="identity",fill="Blue")+  
 ylab("Total Downloads")+  
 xlab("WeekDays")+  
 scale\_x\_discrete(labels=format(as.Date(as.POSIXlt(strptime(df\_groupby\_date$Date, "%m/%d/%Y"))),"%A"))+  
 ggtitle("Plot for Daily Downloads")+  
 theme(axis.text.x=element\_text(angle=90, hjust=1))



Above observation holds true for each Country. We can see this from following plot

# Group the data set by Date and country  
df\_groupby\_date\_country <- df\_app\_downloads %>% group\_by(Date,Country)%>% summarise(total=sum(Downloads,na.rm=TRUE))  
  
ggplot(aes(x=Date,y=total),  
 data=df\_groupby\_date\_country)+   
 geom\_bar(stat="identity",fill="Blue")+  
 ylab("Total Downloads")+  
 xlab("WeekDays")+  
 scale\_x\_discrete(labels=format(as.Date(as.POSIXlt(strptime(df\_groupby\_date$Date, "%m/%d/%Y"))),"%A"))+  
 ggtitle("Plot for Daily Downloads by Each Country")+  
 theme(axis.text.x=element\_text(angle=90, hjust=1))+  
 facet\_wrap(~Country,scales="free")



However, with only 2 weeks data and without any further investigation, it is very hard to confirm any of these patterns.

#### 3) Which segments should you divide the data into before analyzing the relationship between ranks and downloads?

#### Answer:

Some of the possible ways to segment the Data are:

1. By Date
2. By DeviceType
3. By Country

#### 4) What is the relationship between ranks and downloads?

#### Answer :

There is one to many relationship between downloads and ranks. Downloads are aggegated on daily basis. However, ranks are aggregated on hourly basis and there are entries for each country. So for each record in download data set there are multiple records in ranking data set.

#### 5) What is the best measure to use to calculate an average daily rank for an app from its hourly ranks? (Mean, Median, Harmonic Mean, etc.)

#### Answer:

I have used weighted average to get the daily average of ranks.

The ranking average is calculated as follows, where:

w = weight of ranked position.

x = response count

x1w1 + x2w2 + x3w3 ... xnwn divided by Total

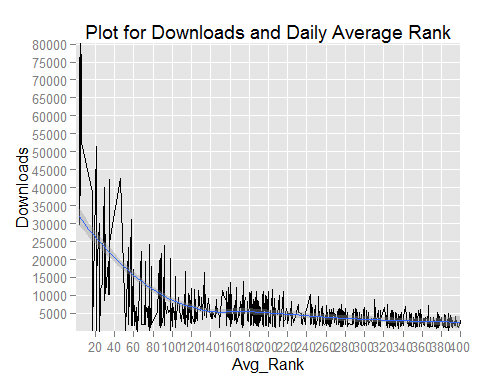
(Weights are applied in reverse). For example : If 05/13/2014 has 3 records with ranking (301,233,186) then weights are (186,233,301) i.e Highest rank is given the lowest weight and lowest rank is given the highest weigth.

Now, Let’s prepare our data for modeling stage. In this step, we will segment the data for USA and join both the data sets by App.ID and Date.

#segment both the data set to include only USA  
df\_app\_downloads\_seg\_by\_country<-subset(df\_app\_downloads,df\_app\_downloads$Country =="USA")  
df\_app\_downloads\_seg\_by\_country<-subset(df\_app\_downloads\_seg\_by\_country,select=-Device)  
df\_app\_ranking\_seg\_by\_country<-subset(df\_app\_ranking,df\_app\_ranking$Country =="USA")  
  
#drop Hour column  
df\_app\_ranking\_seg\_by\_country<-subset(df\_app\_ranking\_seg\_by\_country,select=c(-Hour,-Device))  
  
#Group ranking data set by app.id and Date so that we can join this with downloads data set  
gp\_df\_app\_ranking\_seg\_by\_country<-df\_app\_ranking\_seg\_by\_country %>% group\_by(App.ID,Date)%>% summarise(Avg\_Rank=  
weighted.mean(sort(Rank,decreasing=TRUE),sort(Rank)))  
  
# Join both the data sets on App.id and Date  
df\_joined\_by\_app\_id<-dplyr::inner\_join(df\_app\_downloads\_seg\_by\_country,gp\_df\_app\_ranking\_seg\_by\_country,by=c("App.ID","Date"))

Also, Lets try to plot downloads and Rank and see if we find something interesting

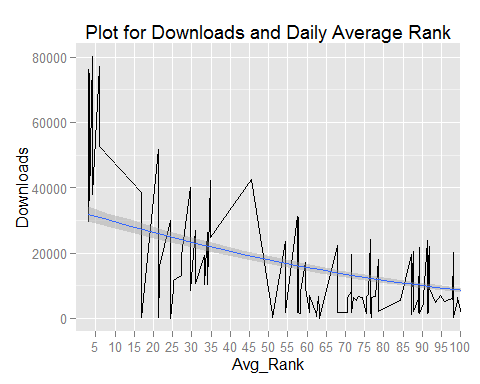
plot<-ggplot(aes(x=Avg\_Rank,y=Downloads),data=df\_joined\_by\_app\_id)+  
 geom\_line()+  
 geom\_smooth()+  
 scale\_y\_discrete(breaks=seq(0,80000,5000))+  
 scale\_x\_discrete(breaks=seq(0,400,20))+  
 ggtitle("Plot for Downloads and Daily Average Rank")  
 suppressMessages(print(plot))



We can see that for top 10 Apps the number of downloads are highest. Number of downloads starts decreasing as the app rank deteriorate.This was kind of expected.

Let's zoom the plot a little bit to see the trend in top 100 apps.

plot\_zoom<-ggplot(aes(x=Avg\_Rank,y=Downloads),data=df\_joined\_by\_app\_id)+  
 geom\_line()+  
 geom\_smooth()+  
 scale\_x\_discrete(breaks=seq(0,100,5))+  
 coord\_cartesian(xlim=c(0,100))+  
 ggtitle("Plot for Downloads and Daily Average Rank")  
suppressMessages(print(plot\_zoom))

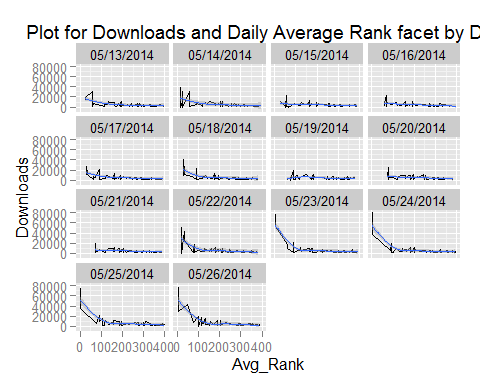


From the graph we can see that there are pikes and valleis in downloads based on ranking but downloads are kind of higher in better ranking.

From the plot, we can see that there is gradual decrease in the number of downloads as the rank of the app deteroriate. Towards the end, our smoother kinds of level off, which is as expected.

The trend is true for each date as shwon in below plot.

p<-ggplot(aes(x=Avg\_Rank,y=Downloads),data=df\_joined\_by\_app\_id)+  
 geom\_line()+  
 geom\_smooth()+  
 facet\_wrap(~Date)+  
 ggtitle("Plot for Downloads and Daily Average Rank facet by Date")  
suppressMessages(print(p))



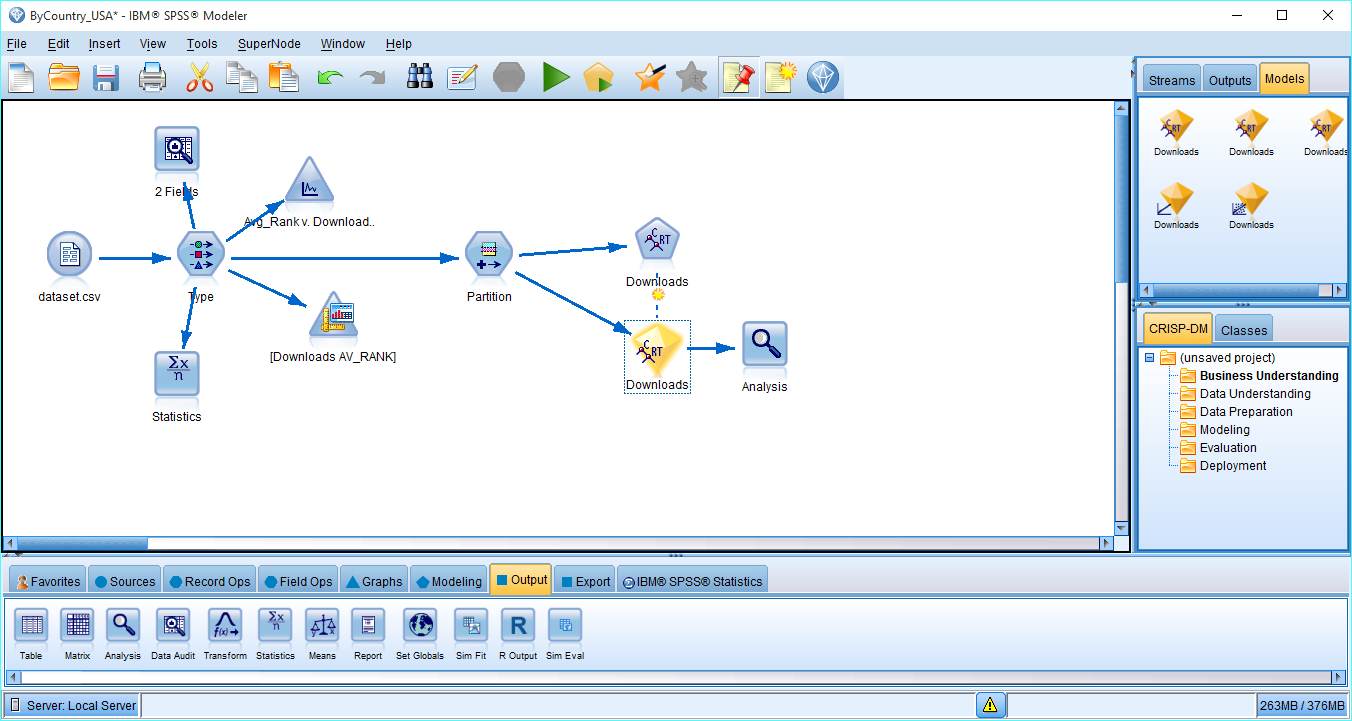
Lets write the data set in to a csv file. I will use spss to build and evaluate the model.

# Write to CSV file  
write.csv(df\_joined\_by\_app\_id,file= file.path("E:","OPT/JobSearch/Assesment/Sensor tower/ST\_DA\_exercise/dataset.csv"))

#### 6) Create a model to predict downloads from average daily rank.

#### Answer:

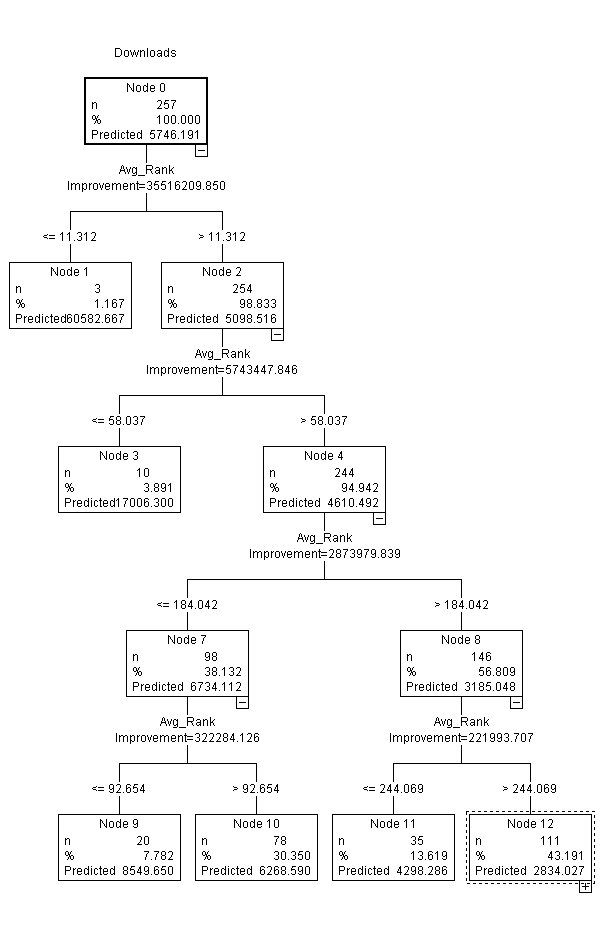
After trying quick couple of regression models, i found that CRT should work fine for this scenario. Below are the results from IBM spss.



Below are the tree rules.



Below is the regression tree:



**Analysis**:

Below are the analysis results from the model.

We can see that difference in Mean Error on training and test data is not very huge, which means our model is not over fitting a lot on training data.

By looking at the MAE, ME and min and MAX errors we can see that this model does not explain the relationship between Downloads and Ranks very well. However, I am sure more predictors can help building a better model.

Note: Even though, there might be a better model with same predictor but for the sake of simplicity, I have not gone in to trying each and every model.

