OSEMN PROJECT : Do Amazon customers rate Samsung phones with larger screens more favorably?

Nishant Kuhar

11/28/2016

Overview

Amazon.com, simply known as **Amazon**, is an American electronic commerce and cloud computing company. It is the largest Internet-based retailer in the world by total sales and market capitalization. Amazon.com started as an online bookstore, later diversifying to sell DVDs, Blu-rays, CDs, video downloads/streaming, MP3 downloads/streaming, audiobook downloads/streaming, software, video games, electronics, apparel, furniture, food, toys and jewelry. The company also produces consumer electronics—notably, Amazon Kindle e-readers, Fire tablets, and Fire TV—and is the world's largest provider of cloud infrastructure services (IaaS).[15] Amazon also sells certain low-end products like USB cables under its in-house brand AmazonBasics. [Source : https://en.wikipedia.org/wiki/Amazon.com]

The project mainly focus on whether samsung mobile phones with larger screens are favoured more by a customer on Amazom.com, or less.

Samsung as we know is a Korean multinational conglomerate company which manufacturers televisions, washing machines, cell phones, etc. But it also deals with steel manufacturing, it is also involved with military operation in South Korea, Samsung is the second largest in ship building industry as well. Samsung was started in March 1, 1938. Samsung has also a small town called Samsung Town in the Gangnam Station area in Seoul.

In this project we will consider 5 different models from Samsung and compare them to determine which model are more preferable the smaller screen smartphones or the larger screen smartphone. To determine this we need to gather the data for customer ratings of Samsung cell phones.

The Project Data retrived for Amazon API till July,2014

Obtain Data

Citation for the data

Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015

Inferring networks of substitutable and complementary products J. McAuley, R. Pandey, J. Leskovec Knowledge Discovery and Data Mining, 2015

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(data.table)
ratingcsvurl <- "http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/ratings_Cell_Phones_an
fp <- file.path(getwd(), "ratings.csv")
download.file(ratingcsvurl, fp)
df <- data.table(read.csv(fp))
```

```
colnames(df) <- c("ReviewerId","ProductId", "Rating", "date")
df$date <- as.POSIXct(df$date, origin="1970-01-01")
df$date <- as.Date(as.POSIXct(df$date, origin="1970-01-01"))
head(df)
```

```
##        ReviewerId  ProductId Rating       date
## 1:  A1YX2RBMS1L9L 0110400550      5 2012-11-22
## 2: A180NNPPKWCCU0 0110400550      5 2013-07-18
## 3:  A3HVRXV0LVJN7 0110400550      5 2013-01-13
## 4: A292527VPX98P8 0110400550      1 2012-11-26
## 5: A1BJGDS0L1I06I 0110400550      1 2013-01-30
## 6:  ANG01NK4RXCI9 0110400550      2 2014-01-08
```

```
dim(df)
```

```
## [1] 3447248       4
```

The above data represents the cutomer ratings of the **Cell Phone and Acessories Category in Amazon**

The **ReviwerId** is the customer id given by amazon to its users, the **ProductId** represents the ASIN number of each product at Amazon, **Rating** represents the Amazon rating given by the users to specific product on a scale of 1 to 5, **date** represents the time and place where the rating is given.

The library function in the above code install the packages that are further needed in the code. Rvest package is used to extract data from various websites which are represented in various forms. The data.table package help us to convert the data from a website or webservice to readable data in a table format and manipulate it as well.

The **file.path** function assigns a file path in the same directory for extracting the data file, **download.file** download the data from the .csv file, **data.table** this function converts the data to a readable table format, the ** as.POSIXct and as.Date ** functions are used to convert the unix time date format to english laguage date format readable by a normal human being, ** colnames() ** is used to assign specific names to the column which we desire.

Variables used in Code

- ratingcsvurl : A variable to store the url.
- fp : A variable to store file path of the downloaded csv file
- df : a data frame which stores the data in tabular form.

Scubbing and obtaing valuable data

In this section we will be obtaing Data into a new data frame and look at 5 popular models of Samsung with large and small screen devices. The devices I chosed for obtaining this data are, * Product Name - ASIN (Product ID) * Samsung Galaxy Note 3 - B00FJ8YCZM * Samsung Galaxy S5 - B00IZ1XVAC * Samsung Galaxy S3 - B008HTJLF6 * Samsung Galaxy Note 2 Titanium - B009Z1MNF0 * Samsung Galaxy 4 Zoom - B00G9G0OP0

```
df2 <- df[is.element(df$ProductId, 'B00FJ8YCZM') |
            is.element(df$ProductId, 'B00IZ1XVAC') |
            is.element(df$ProductId, 'B008HTJLF6') |
            is.element(df$ProductId, 'B009Z1MNF0') |
            is.element(df$ProductId, 'B00G9G0OP0')]
head(df2)
```

```
##        ReviewerId  ProductId Rating       date
## 1: A10IFQ6YWAZ4QD B008HTJLF6      5 2013-04-14
## 2: A11SL1I6C688VJ B008HTJLF6      2 2012-09-03
## 3:  AL67SANRYGW1C B008HTJLF6      5 2013-05-10
## 4: A1KOBA858HIGUB B008HTJLF6      4 2013-06-24
## 5: A11S2DLFPQZYAX B008HTJLF6      5 2012-12-10
## 6: A32HWPIWO1DDO6 B008HTJLF6      5 2012-10-12
```

```
class(df2)
```

```
## [1] "data.table" "data.frame"
```

```
str(df2)
```

```
## Classes 'data.table' and 'data.frame':   450 obs. of  4 variables:
##  $ ReviewerId: Factor w/ 2261044 levels "A00000922W28P2OCH6JSE",..: 12842 34068 2016142 346151 33837
##  $ ProductId : Factor w/ 319677 levels "0110400550","0195866479",..: 157348 157348 157348 157348 157
##  $ Rating    : num  5 2 5 4 5 5 5 5 5 5 ...
##  $ date      : Date, format: "2013-04-14" "2012-09-03" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(df2)
```

```
##           ReviewerId       ProductId       Rating     
##  A10IFQ6YWAZ4QD:  1   B008HTJLF6:165   Min.   :1.000  
##  A10LWBOIZCF2QT:  1   B00FJ8YCZM:130   1st Qu.:4.000  
##  A10ONQZBIDKENR:  1   B00IZ1XVAC: 88   Median :5.000  
##  A117P0J6GE8DO4:  1   B009Z1MNF0: 58   Mean   :4.209  
##  A11CIWA4UYKENH:  1   B00G9G0OP0:  9   3rd Qu.:5.000  
##  A11CQKYW4LH41J:  1   0110400550:  0   Max.   :5.000  
##  (Other)       :444   (Other)   :  0                  
##       date           
##  Min.   :2012-07-12  
##  1st Qu.:2013-02-09  
##  Median :2013-12-06  
##  Mean   :2013-09-22  
##  3rd Qu.:2014-04-17  
##  Max.   :2014-07-22  
## 
```

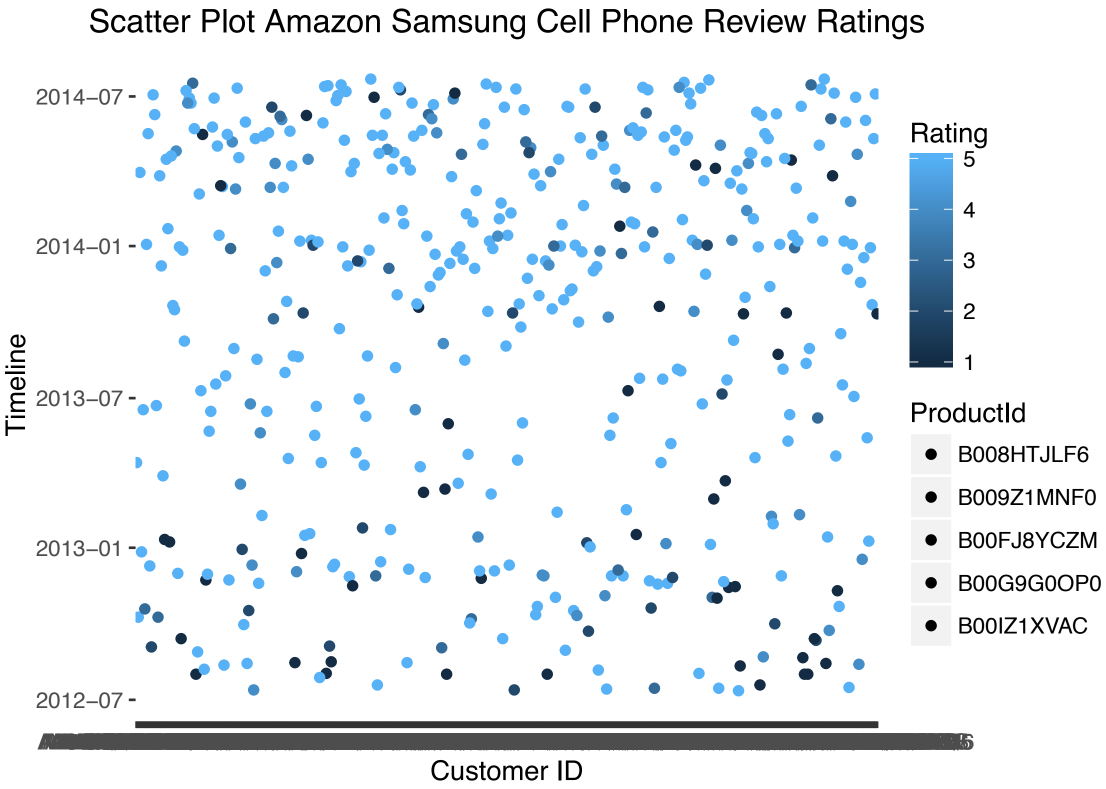
In the above code we determine a data frame **df2** from an existing data frame **df**. To obtain data from the larger data table we use **is.element()** to compare and extract 5 models we desire and their ratings by using the or operator.

The **head()** function, gives us the first six rows of the data frame, in this case **df2**. The **class()** function, it tells us what is the type of data frame, in this case it is a table. The **str()** function, it tells about the structure of the data frame. What is the data type of the columns like if it's a string ("a string of charaters and numbers"), number ("a number value"), date ("a date data type")
The **summary()** function gives us the summary of a data frame, the count of values in columns, minimum, maximum, median, and mean of the data frame. In the following data which we have gathered for our research the mean rating of the data is 4.209.

Scatter Plot of the scrubbed data

The following scatter plot represents the data frame **df2** with all the customers on the x- axis and the timeline on the y- axis, with ratings as the color differences and the product as the fill color difference. This data truly represents the anonymity and human behavior in rating their products.

```
library(ggplot2)
ggplot(df2,
  aes(x = ReviewerId, y = date ,fill = ProductId, color = Rating)) +
  geom_point() +
  xlab("Customer ID") + ylab("Timeline") +
    ggtitle("Scatter Plot Amazon Samsung Cell Phone Review Ratings")
```



Funcionality of the code

- ggplot2 : R package to plot beautiful graphs
- aes : Aesthetics of a graj which tells us the x-axis, y- axis and the color of the graph on which it will be differentiating the values.
- geom_point() : this defines the type of the graph which in this case is a scatter plot.
- theme() : this defines the custom theme of the Graph
- xlab() : this function provides title to the x- axis
- ylab() : this Function provides title to the y axis
- labs() : this function helps to give labels such as title of the graph.

Result

Determing whether larger screens are more favorable in Samsung Mobile Phones

To determine the following we will narrow down our research to 3 cell phones which have the screen sizes of small, medium and large. For this we will be considering the following models from our above data frame **df2** ** Model name - ASIN (Product ID) *Samsung Galaxy Note 3 - B00FJ8YCZM* Samsung Galaxy S5 - B00IZ1XVAC *Samsung Galaxy S3 - B008HTJLF6

```
df3 <- df[is.element(df$ProductId, 'B00FJ8YCZM') |
            is.element(df$ProductId, 'B00IZ1XVAC') |
            is.element(df$ProductId, 'B008HTJLF6')]
str(df3)
```

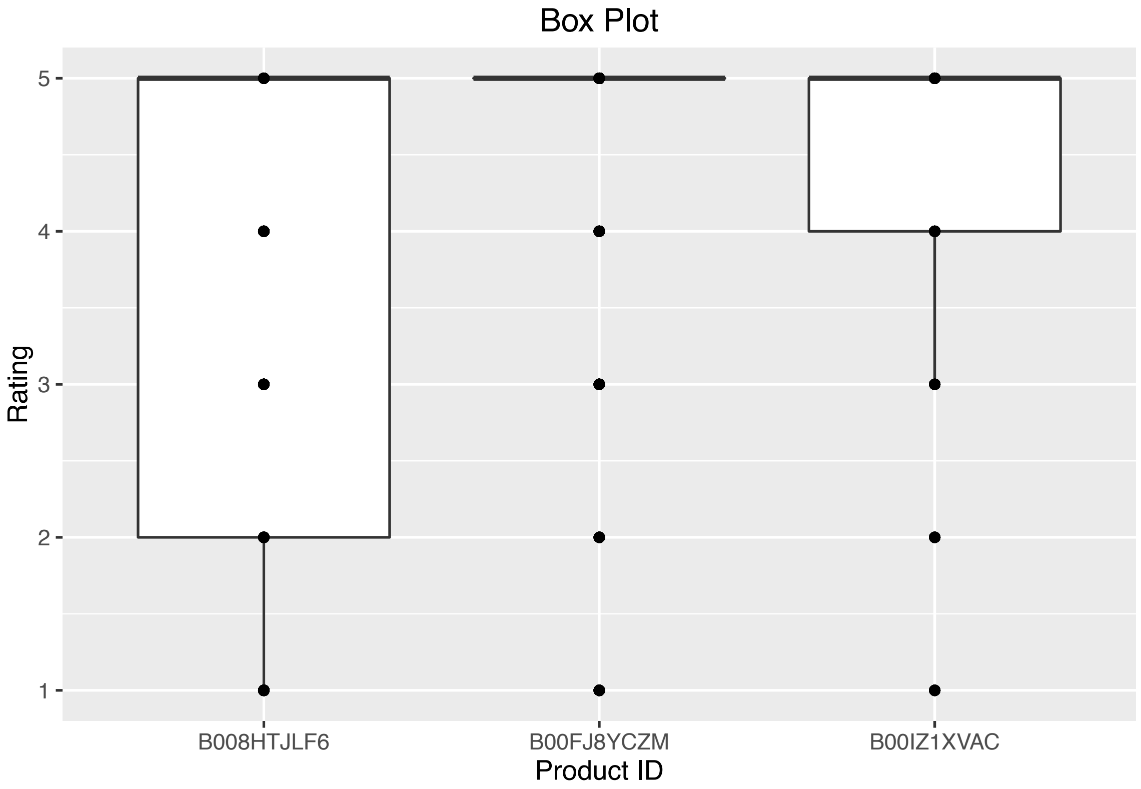
```
## Classes 'data.table' and 'data.frame':   383 obs. of  4 variables:
##  $ ReviewerId: Factor w/ 2261044 levels "A00000922W28P2OCH6JSE",..: 12842 34068 2016142 346151 33837
##  $ ProductId : Factor w/ 319677 levels "0110400550","0195866479",..: 157348 157348 157348 157348 157
##  $ Rating    : num  5 2 5 4 5 5 5 5 5 5 ...
##  $ date      : Date, format: "2013-04-14" "2012-09-03" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Now we have a data frame **df3** with only the three products which we will be evaluating the results.

We will plot a bar graph and a box plot to represent the above gathered data graphically and determing which model is prefered more by the consumers and which is not.

Box Plot

```
quartermonthly <- quarter(df3$date)
ggplot(df3,
       aes(x = ProductId, y = Rating )) +
  geom_boxplot() +
  geom_point() +
   xlab("Product ID") + ylab("Rating") +
    ggtitle("Box Plot")
```



In this above code we are plotting boxplot and point graph to determine the outcome of customer ratings for large screen phones and the small screen phones. The above graph is plot between Product id (x- axis), Customer Rating.

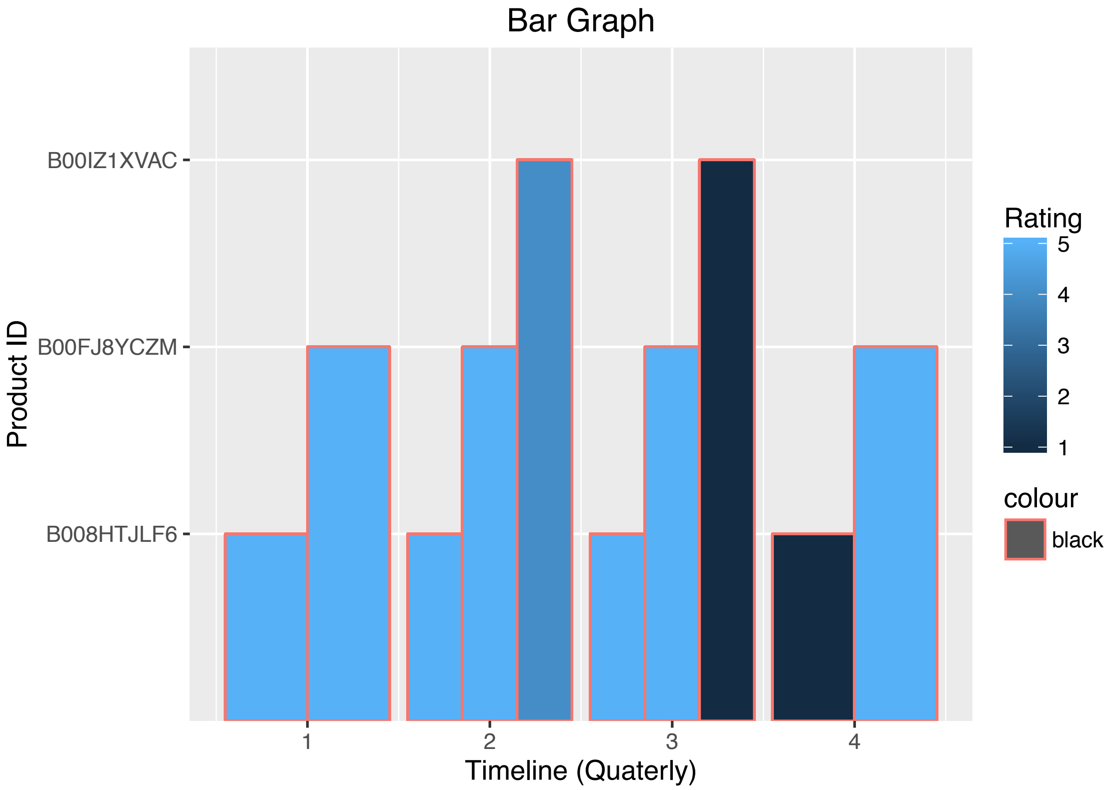
In the graph we can see the large screen phone in this case Note 3 has most liking because it is given 5 rating throughout.

Code Functionality

geom_boxplot : Function which defines the type of graph which we will be plotting, in this case box plot quarter() : function extracts Quarter from the Date and timeline

Bar- Graph

```
ggplot(df3,
       aes(x = quartermonthly, y = ProductId ,fill = Rating, color="black")) +
  geom_bar(stat = "identity", position = position_dodge()) +
  xlab("Timeline (Quaterly)") + ylab("Product ID") +
    ggtitle("Bar Graph")
```



In the above code we are plotting a bar graph with x-axis as a quartermonthly timeline for the ratings and y-axis as the Samsung Product. From the Bar Graph we can determine that Galaxy Note 3 is rated more and people buy it more often than the others and always good reviews and ratings are posted by the customers.

Code Functionality

geom_bar : Function which defines the type of graph which we will be plotting. position_dodge : Dodging things with different widths. *stat : It defines the graph needs to be plot on which factors.* Quarter() : function extracts Quarter from the Date and timeline

Parameteric test on ratings of Galaxy Note3(B00FJ8YCZM) and Galaxy S3(B008HTJLF6)

Choosing these two models for the because of the timeline they were introduced in market in the same time frame.

```
t.test(df3[is.element(df3$ProductId,'B00FJ8YCZM')]$Rating,
       df3[is.element(df3$ProductId,'B008HTJLF6')]$Rating)
```

```
##
##  Welch Two Sample t-test
##
## data:  df3[is.element(df3$ProductId, "B00FJ8YCZM")]$Rating and df3[is.element(df3$ProductId, "B008HT
## t = 4.0764, df = 291.17, p-value = 5.907e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3373124 0.9671165
## sample estimates:
## mean of x mean of y
##  4.446154  3.793939
```

Result of t-test

This t.test defines that the mean of "Ratings" in case of Larger Screens Galaxy note3 is higher than the smaller screens. The p-value is also greater than 0.05 this means we can consider this hypothesis.

Conclusions Drawn

As we can see in the above graphs the product id represents the product as *Model name - ASIN (Product ID)* Samsung Galaxy Note 3 - B00FJ8YCZM *Samsung Galaxy S5 - B00IZ1XVAC* Samsung Galaxy S3 - B008HTJLF6

From the above graphs we can observe that "Galxy Note 3" (ASIN - B00FJ8YCZM) which has the largest screen of the three models over the quarter is rated between 4-5 from the users over the timeline, wheras " Galaxy S3 and S5" with small and medium screens range ratings between 1-5 over the timeline for S3 and 1-3 for S5, this data shows us the difference in the liking of a Product.

So, now we can conclude that Larger Screen Samsung Mobile Phones are more favored by Amazon Customers