Course Project

By: David Catalan Perez, Vandyck Buabeng, Kristen Hanold, & Reuben Akipogu

CIS 4730 (Summer 2021) – Professor Zhang

Introduction

We are deploying a new phone product, and we need to examine the current market to ensure its success and survivability.

Use

- A data set that has user reviews for current phones in the market.

Clean and extract

- Data to generate tables and figures for analysis.

Present

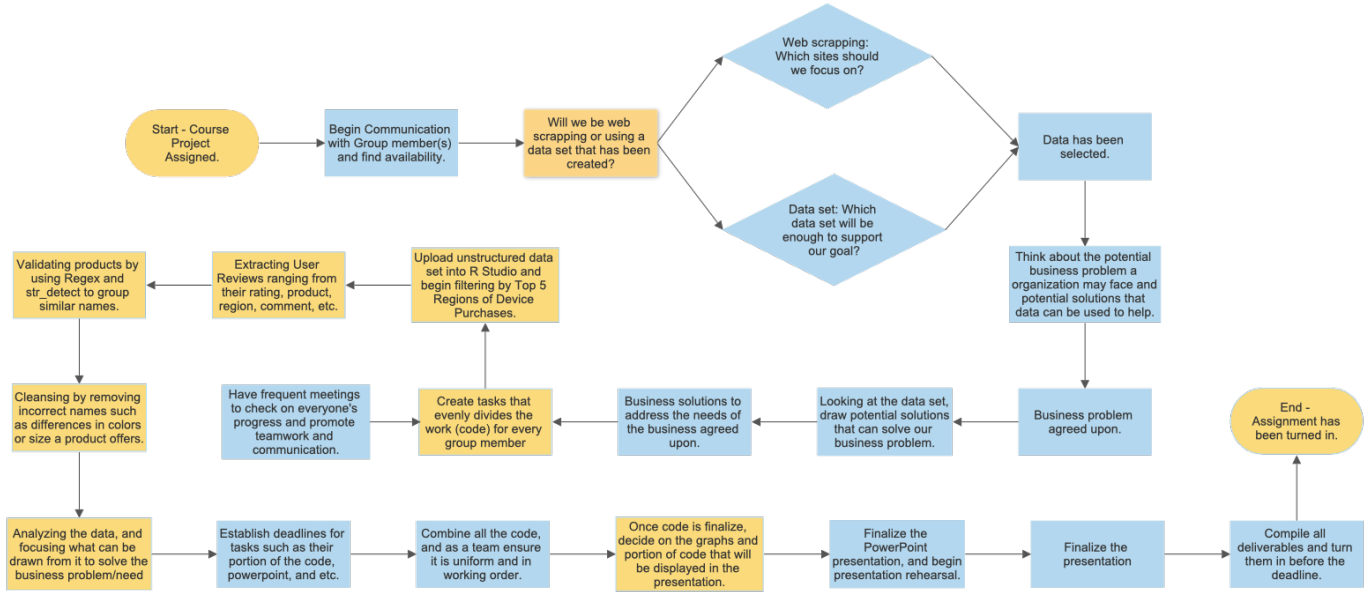
- A viable solution from the data that will assist the business in a successful launch.

Methodology

- Our Data Set

- Source: Kaggle
- Size: 374,910 User Reviews
- Sample Size: 246, 810 Reviews
- Data ranges from 2014 to 2017.

- Step-by-Step Workflow Diagram



Finding the Top 5 Countries

- After importing our data to R, we will find the top five countries with the most reviews.
- Once we find our top five regions, we can filter the data to focus on those regions for now.

```r
# Acquiring the Data - Reading Phone Reviews File
User_Reviews = read_csv("phone_user_review_file_1.csv")

# Filtering the Data - Finding the Top 5 Regions in Device Purchases
User_Reviews_Top5_Regions = User_Reviews %>%
  group_by(country) %>%
  summarise(Number_of_Region_Purchases = n()) %>%
  arrange(desc(Number_of_Region_Purchases))

User_Reviews_Top5_Regions <- top_n(User_Reviews_Top5_Regions, 5)
```

```r
# Extracting the Data - Getting United States Reviews
User_Reviews_US = User_Reviews %>% filter(country == "us")

# Extracting the Data - Getting India Reviews
User_Reviews_IN = User_Reviews %>% filter(country == "in")

# Extracting the Data - Getting Italy Reviews
User_Reviews_IT = User_Reviews %>% filter(country == "it")

# Extracting the Data - Getting Germany Reviews
User_Reviews_DE = User_Reviews %>% filter(country == "de")

# Extracting the Data - Getting France Reviews
User_Reviews_FR = User_Reviews %>% filter(country == "fr")
```

Device Rating Averages Based on Country

- We can find the averages of every reviewed device within all the selected countries from our filtered lists.
- We can sort from descending order to display the top-rated devices to the lowest.
- Examining this will allow us to see popular devices in that region.

```
# US Device and Average Rating
Device_Ratings_US = User_Reviews_US %>%
  group_by(product) %>%
  summarise(Number_of_Products = n(),
            Average_Product_Rating = mean(score)) %>%
  mutate(Product_Percentage = round((Number_of_Products /
                                     sum(Number_of_Products)), 2)) %>%
  arrange(desc(Number_of_Products))
```

```
> Device_Ratings_US
# A tibble: 1,898 x 4
   product                                Number_of_Produc… Average_Product_Ra… Product_Percenta…
   <chr>                                              <int>               <dbl>            <dbl>
 1 Samsung Galaxy S7 edge 32GB (Verizon)               1811                9.35             2.15
 2 Samsung Galaxy S7 edge 32GB (T-Mobile)              1729                9.42             2.05
 3 Samsung Galaxy S7 32GB (Verizon)                    1607                9.35             1.91
 4 Samsung Galaxy S7 32GB (T-Mobile)                   1532                9.38             1.82
 5 Samsung Galaxy S5 16GB (Verizon)                    1432                9.19             1.7
 6 Samsung Galaxy S7 edge 32GB (AT&T)                  1383                9.42             1.64
 7 Samsung Galaxy S5 16GB (T-Mobile)                   1165                9.03             1.38
 8 Samsung Galaxy S5 16GB (AT&T)                       1079                9.16             1.28
 9 Samsung Galaxy S6 edge+ 32GB (T-Mobile)             1040                9.45             1.23
10 Huawei Honor 5X Unlocked Smartphone, 16GB…          1005                8.31             1.19
# … with 1,888 more rows
```

Devices in Each Region

- We can discover the number of devices that are within each region.
- Examining these regions will allow us to see if there is a trend and potential competitor.
- Analyze what they do right and wrong that allows for a high usage from consumers (High v. Low rated devices).

```
# Counts the Number of Devices in the US
us <- User_Reviews_US %>%
  select(country, product) %>%
  summarise(product = n())

# Counts the Number of Devices in the India
In <- User_Reviews_IN %>%
  select(country, product) %>%
  summarise(product = n())

#Counts the Number of Devices in the Italy
it <- User_Reviews_IT %>%
  select(country, product) %>%
  summarise(product = n())

# Counts the Number of Devices in the Germany
de <- User_Reviews_DE %>%
  select(country, product) %>%
  summarise(product = n())

# Counts the Number of Devices in the France
fr <- User_Reviews_FR %>%
  select(country, product) %>%
  summarise(product = n())
```

```
> us
# A tibble: 1 x 1
  product
    <int>
1   84259
> In
# A tibble: 1 x 1
  product
    <int>
1   52821
> it
# A tibble: 1 x 1
  product
    <int>
1   45729
> de
# A tibble: 1 x 1
  product
    <int>
1   34264
> fr
# A tibble: 1 x 1
  product
    <int>
1   29737
```

```
> Ten_Rated_Devices_US
# A tibble: 1,676 x 3
# Groups:   product, country [1,676]
   country product                                                  score
   <chr>   <chr>                                                    <dbl>
 1 us      Samsung Galaxy S7 edge 32GB (T-Mobile)                      10
 2 us      Samsung Galaxy S7 edge 32GB (Verizon)                       10
 3 us      Samsung Galaxy S7 32GB (T-Mobile)                           10
 4 us      Samsung Galaxy S7 32GB (Verizon)                            10
 5 us      Samsung Galaxy S7 edge 32GB (AT&T)                          10
 6 us      Samsung Galaxy S5 16GB (Verizon)                            10
 7 us      Huawei Nexus 6P unlocked smartphone, 32GB Gold (US Warranty)  10
 8 us      Huawei Honor 5X Unlocked Smartphone, 16GB Dark Grey (US Warranty)  10
 9 us      Samsung Galaxy S6 edge+ 32GB (T-Mobile)                     10
10 us      Samsung Galaxy S5 16GB (AT&T)                               10
# … with 1,666 more rows
```

Identifying Key Players in All 5 Regions

- With the identification of the top and highly-rated devices in each region, we can examine the key players.
- Using str_detect, we can search through the data set to find specific products (e.g., Samsung Galaxy, Apple iPhone, etc.).
- We can see which regions these key players are dominating or fallen behind from competitors.

```
# Show individuals in the Top 5 Countries that uses a Samsung Galaxy.
Samsung = "Samsung Galaxy"

Samsung_Galaxy = User_Reviews_Top_Products %>%
  group_by(country) %>%
  filter(str_detect(product, Samsung)) %>%
  summarise(Device_Purchases = sum(Total_Products)) %>%
  arrange(desc(Device_Purchases))
```

```
> Samsung_Galaxy
# A tibble: 5 x 2
  country Device_Purchases
  <chr>              <int>
1 us                 48160
2 de                 10954
3 fr                  9722
4 in                  7066
5 it                  3109
```

Finding Popular Device Source Locations

- We need to examine these purchased devices and find where consumers buy them primarily from.
- Identifying these device sources will enable us to adjust strategies to ensure the availability of our product.
- We'll see where we should focus our efforts when it comes time to sell a product.

```
# Finding the Most Popular Locations Where Devices were Bought.
device_source_us = User_Reviews %>%
  group_by(source) %>%
  summarise(num_of_device = n()) %>%
  mutate(source_percentage = round((num_of_device / sum(num_of_device) *
                                      100), 2)) %>%
  arrange(desc(num_of_device))
head(device_source_us, 5)
```

```
# A tibble: 5 x 3
  source     num_of_device source_percentage
  <chr>              <int>             <dbl>
1 Amazon            225034              60.0
2 Samsung            31872               8.5
3 KIESKEURIG         18858               5.03
4 Bondfaro            7097               1.89
5 Yandex              6674               1.78
```

Using Regex to Dig Deeper

- We will use Regex (Regular Expressions) to go through thousands of reviews to find specific keywords.
- These keywords will help determine what is essential to a product and improve where competitors fall short.
- Keywords such as “Best,” “Issues,” ”Love,” etc., can tell us much about a device and see what matters to consumers.

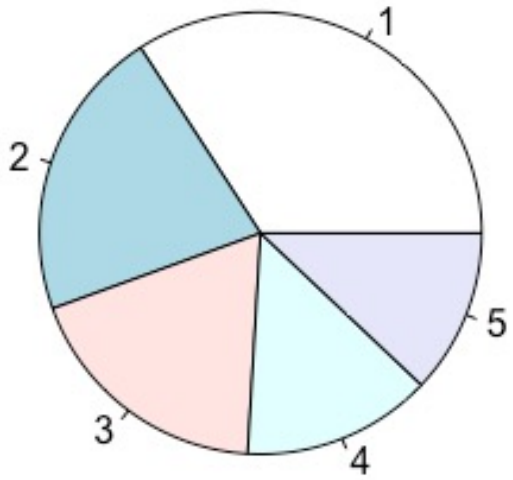
```
# Finding Keyword "best" in Reviews and Displaying the top 5 devices
reviews_keywords_Best <- User_Reviews %>%
  select(extract, product) %>%
  filter(str_detect(extract, "[Bb]est")) %>%
  count(product) %>%
  arrange(desc(n))
head(reviews_keywords_Best, 5)

# Using Regex to confirm str_detect received the correct number.
Phone_Reviews = User_Reviews$extract
my_regex_Best = "[Bb]est"
Best = (stringr::str_extract_all(Phone_Reviews, my_regex_Best))
Best[lengths(Best) == 0] <- NA_character_
Best = Best[!is.na(Best)]
Best_Match = length(Best)
Best_Match
```

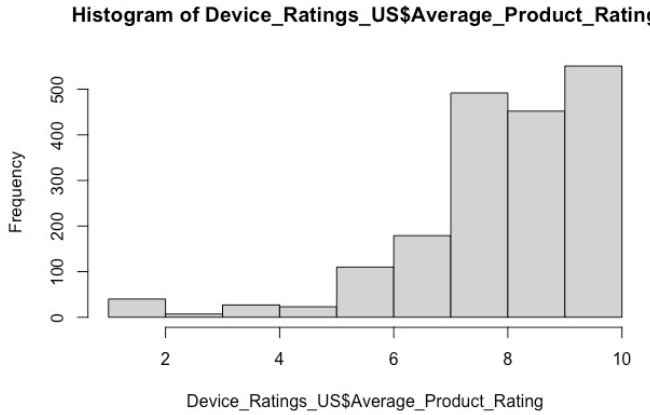
```
# A tibble: 5 x 2
  product                                                  n
  <chr>                                                <int>
1 OnePlus 3 (Graphite, 64 GB)                            679
2 OnePlus 3 (Soft Gold, 64 GB)                           645
3 Samsung Galaxy S7 edge 32GB (T-Mobile)                 326
4 OnePlus 3T (Gunmetal, 6GB RAM + 64GB memory)           314
5 Samsung Galaxy S7 edge 32GB (Verizon)                  297
```

Analyzing the Data

- Number of Reviews in each Filtered Region
- Highly-Rated Devices




```
> User_Reviews_Top5_Regions
# A tibble: 5 x 2
  country Number_of_Region_Purchases
  <chr>                        <int>
1 us                           84259
2 in                           52821
3 it                           45729
4 de                           34264
5 fr                           29737
```

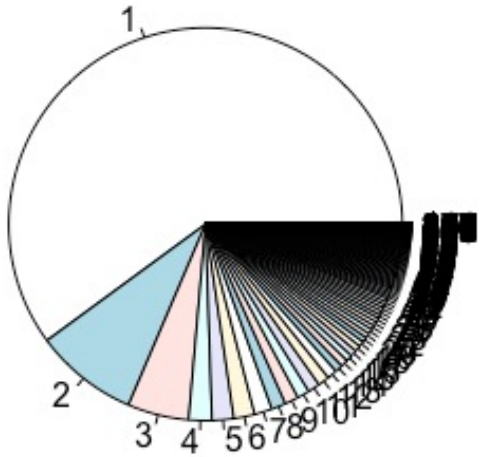



```
> Device_Ratings_US
# A tibble: 1,898 x 4
   product                                  Number_of_Produc… Average_Product_Ra… Product_Percenta…
   <chr>                                                <int>               <dbl>            <dbl>
 1 Samsung Galaxy S7 edge 32GB (Verizon)                 1811                9.35             2.15
 2 Samsung Galaxy S7 edge 32GB (T-Mobile)                1729                9.42             2.05
 3 Samsung Galaxy S7 32GB (Verizon)                      1607                9.35             1.91
 4 Samsung Galaxy S7 32GB (T-Mobile)                     1532                9.38             1.82
 5 Samsung Galaxy S5 16GB (Verizon)                      1432                9.19             1.7
 6 Samsung Galaxy S7 edge 32GB (AT&T)                    1383                9.42             1.64
 7 Samsung Galaxy S5 16GB (T-Mobile)                     1165                9.03             1.38
 8 Samsung Galaxy S5 16GB (AT&T)                         1079                9.16             1.28
 9 Samsung Galaxy S6 edge+ 32GB (T-Mobile)               1040                9.45             1.23
10 Huawei Honor 5X Unlocked Smartphone, 16GB…            1005                8.31             1.19
# … with 1,888 more rows
```

Analyzing the Data

- Popular Device Source Location
- Domain (URL) Purchases


```
> Region_Purchases
# A tibble: 253 x 3
   domain        Domain_Purchases Domain_Percentage
   <chr>                    <int>             <dbl>
 1 amazon.com               46695              0.12
 2 amazon.in                45058              0.12
 3 amazon.it                43601              0.12
 4 samsung.com              31772              0.08
 5 amazon.de                26673              0.07
 6 amazon.es                22699              0.06
 7 amazon.fr                22086              0.06
 8 amazon.co.uk             15191              0.04
 9 kieskeurig.be             9805              0.03
10 kieskeurig.nl             9053              0.02
# … with 243 more rows
```




```
> device_source_us
# A tibble: 217 x 3
   source         num_of_device source_percentage
   <chr>                  <int>             <dbl>
 1 Amazon                225034              60.0
 2 Samsung                31872               8.5
 3 KIESKEURIG             18858               5.03
 4 Bondfaro                7097               1.89
 5 Yandex                  6674               1.78
 6 Otto.de                 6453               1.72
 7 Cissa Magazine          5426               1.45
 8 Argos                   4414               1.18
 9 Verkkokauppa            4248               1.13
10 Flipkart                3781               1.01
# … with 207 more rows
```

Conclusion

We found the key players in their regions, what they did right, and what consumers want for our business problem. Utilizing R allowed us to take and extract information from the data set to support a business need.

As a team, we were able to develop and test our technical skills and work as one unit, supporting one another along the way.

We used our collective knowledge of the R labs to set tasks that would allow us to explore the data on a deeper level.

Learned how we could use unstructured data to tackle business problems to find insights that solve them.

Thank You!

Any Questions?