**Final Project**

## Things you will learn in this case-study:

1. Data Pre-processing
2. Exploratory Data Analysis
3. Data Manipulation

## Univariate, Bivariate and Multivariate Analysis using Visualization

1. Text Analytics

## Back Ground:

Sellers on online auction websites need to understand the characteristics of a successful item listing to maximize their revenue.  Buyers might also be interested in understanding which listings are less attractive so that they can choose the best deal. As an organization, ebay is interested in developing an understanding on this very important business element and this is where you come into picture.

## Overview of the problem:

In this project, you will play the role of Data Scientist for ebay, and you have been asked to do a thorough and detailed analysis on answering one question - **What makes an eBay listing successful?**

Post your report, the company will use the findings to develop an analytics model that will help buyers and sellers predict the sales success of a set of eBay listings.

**Data**

## **File descriptions**

The data comes from the Kaggle Competition.

## **Data fields**

* **sold: if the product was sold, it takes value ==1, otherwise it takes value ==0**
* **description** = The text description of the product provided by the seller.
* biddable = Whether this is an auction (biddable=1) or a sale with a fixed price (biddable=0).
* **startprice**= The start price (in US Dollars) for the auction (if biddable=1) or the sale price (if biddable=0).
* **condition** = The condition of the product (new, used, etc.)
* **cellular**= Whether the iPad has cellular connectivity (cellular=1) or not (cellular=0).
* **carrier**= The cellular carrier for which the iPad is equipped (if cellular=1); listed as "None" if cellular=0.
* **color**= The color of the iPad.
* **storage**= The iPad's storage capacity (in gigabytes).
* **productline**= The name of the product being sold.

## Objective:

Based on the knowledge you acquired throughout the course, you are expected to complete the Project by answering the below mentioned questions.

## You should do the following:

**Data Analysis and Preprocessing**

1. Load the required libraries and the data 'projectdata.csv' and call it p.
2. Look at the structure of the dataset
3. Change the description variable from factor to character
4. Re-look at the structure of the dataset
5. Explore the data - which variables have missing values and what is the count of these missing values
6. check the number of unique values in each column
7. what is the median price
8. add a variable priceclass, which takes value 1 if the price is above median value, or 0 otherwise
9. Re-look at the structure of the dataset
10. Identify and convert categorical variables to factor
11. Find Maximum value of all numerical variables
12. Find Maximum value of all categorical variables
13. use lapply function to Calculate Median of each of the numerical variables
14. use lapply function to Calculate standard deviation of each of the numerical variables
15. use the tapply function to calculate the median price according to sold
16. use the tapply function to calculate the median price according to biddable

**Data Manipulation and EDA**

1. Load the dplyr package
2. Print out a df with the columns startprice, condition, and sold.
3. Print out the columns biddable to sold
4. Add the new variable var1 which calculates the ratio of storage to startprice and save the result in p.
5. Arrange p by price
6. Arrange p so that condition is grouped
7. Arrange p so that biddable and sold is grouped
8. Definition of notbought - create a df of observations when ipad was notsold
9. Arrange notbought so that condition and sold is is grouped
10. Generate summary about startprice column of p. Summary should include min, max, mean, sd and IQR
11. Generate summary about storage column of housing

**Visualization**

1. Load the required package
2. Create a scatter plot between startprice and storage
3. In the above plot, add the color argument which should be dependent on the sold Variable
4. In the above plot where you had used the color argument, please add the smooth line using the geom\_smooth() function
5. Make a univariate histogram on startprice
6. In the above plot, add set binwidth to 200 in the geom layer
7. In the above plot, MAP ..density.. to the y aesthetic (i.e. in a second aes() function)
8. Now, In the above plot, plus SET the fill attribute to "#377EB8"
9. Draw a bar plot of sold, filled according to biddable
10. In the previous plot, Change the position argument to "stack""
11. Change the position argument to "fill""
12. Change the position argument to "dodge""
13. Now create a basic scatter plot between pce and psavert variables on econ\_2:
14. Separate rows according to sold
15. Separate columns according to biddable
16. Separate by both columns and rows

**Text Analytics**

1. load the required packages and libraries required for text analytics
2. Now, extract the relevant variable, the one containing the text. Please copy the following code as below

r1 = as.character(p$description)

#Set the seed to 100 for code reproducibility

set.seed(100)

# run the following command, 'sample = sample(r1, (length(r1)))', in your RStudio, now you are ready for Bag of Words

sample = sample(r1, (length(r1)))

1. Create a Corpus - which, in simple terms, is nothing but a collection of text documents.
2. Now, remove punctuations
3. Next, change the case of the word to lowercase so that same words are not counted as different because of lower or upper case.
4. Next, remove numbers
5. Next, remove whitespaces
6. Now, remove unhelpful terms, also referred as stopwords
7. Now, please carry out the process of stemming, motivated by the desire to represent words with different endings as the same word.
8. creat a document term matrix from the corpus
9. now create the data frame from the output of the above line
10. Create a word cloud and set random.order = TRUE:
11. Create a word cloud and set random.order = FALSE:
12. In the above word cloud, adjust the frequency level with min.freq parameter set at 5

**Text Analytics - Creating Word Cloud for Un Sold ipads**

1. craete a new dataframe from the original data 'p' which only includes those observations where the ipad was not sold

notsoldipads = subset(p, sold == 0)

n1 = as.character(notsoldipads$description)

#Set the seed to 100 for code reproducibility

set.seed(100)

#sample

sample2 = sample(n1, (length(n1)))Run the follwoing commands in your R Studio

1. #Bag of Words - Run the above codes

#1 - Create a Corpus

#2 - Remove punctuations

#3 - Convert to lowercase

#4 - Remove Numbers

#5 - Remove whitespaces

#6 - Remove stopwords

#7 - Perform Stemminga

1. create a document term matrix from the resultant corpus. Run the following codes

frequencies2 = DocumentTermMatrix(corpus2)

1. now create the data frame from the output of the above line

# Create three word clouds using the following three instructions

# WordCloud 1 - Create a word cloud and set random.order = TRUE.

# WordCloud 2 - Create a word cloud and set random.order = FALSE

# WordCloud 3 - In the above word cloud, adjust the frequency level with min.freq parameter set at 5

**Creating Word Cloud for Sold ipads**

1. Crraete a new dataframe from the original data 'p\_sold' which only includes those observations where the ipad was sold (p\_sold = subset(p,sold == 1))
2. Now, run the follwoing commands in your R Studio , extracting relevant positive tweets

p1 = as.character(p\_sold$description)

#Set the seed to 100 for code reproducibility

set.seed(100)

#sample

sample3 = sample(p1, (length(p1)))

1. Bag of Words - Run the above codes

#1 - Create a Corpus

#2 - Remove punctuations

#3 - Convert to lowercase

#4 - Remove Numbers

#5 - Remove whitespaces

#6 - Remove stopwords

#7 Perform Stemming

1. creat a document term matrix from the resultant corpus
2. now create the data frame from the output of the above line
3. Create three word clouds using the following three instructions

# WordCloud 1 - Create a word cloud and set random.order = TRUE.

# WordCloud 2 - Create a word cloud and set random.order = FALSE

# WordCloud 3 - In the above word cloud, adjust the frequency level with min.freq parameter set at 5

## Submission should include the following:

1. Answers to the above questions. Print the resultant output wherever applicable.
2. Approach and rationale should be documented.
3. R Code File.