**# Step 1 Importing packages and libraries for calculation and storing data**

*import numpy as np*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*%matplotlib inline*

*import seaborn as sns*

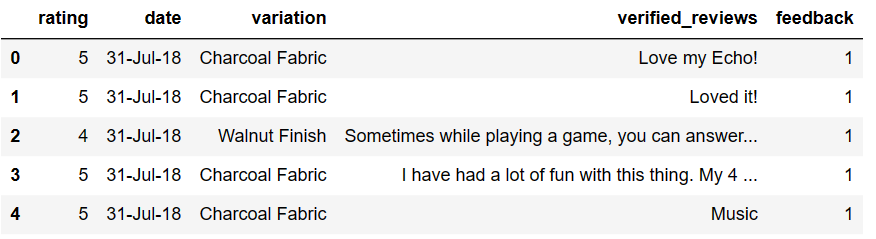
**# Step 2 Importing Amazone Alexa Dataset and display First five Records**

*df\_reviews=pd.read\_csv('amazon\_alexa.tsv', sep='\t')*

**# Step 3 Display top records**

*df\_reviews.head()*

**Table 1 Top five records of the dataset with columns and rows**



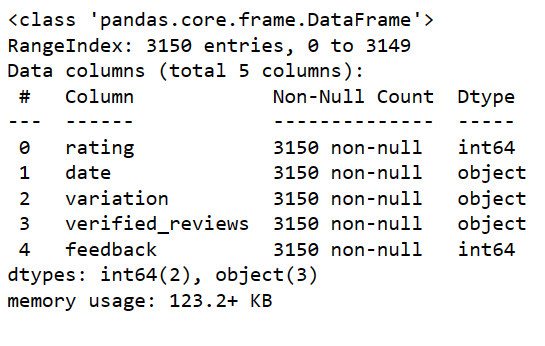
**# Step 4 Display no of rows and column**

*df\_reviews.shape*

**# Step 5 Display Structure of the file**

*df\_reviews.info()*

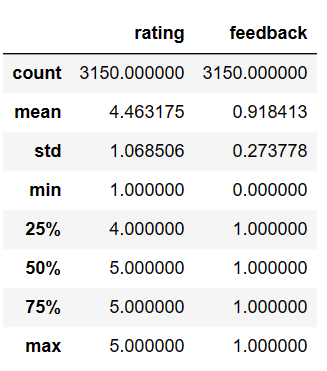
**Table 2 Structural information of the Amazon Alexa**



**# 6 Describing Descriptive Business Analytics**

*df\_reviews.describe()*

**Table 3 Descriptive Statistics of the dataset**

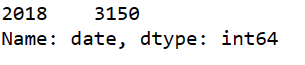


**# 7 Converting date from object to date datatype and displaing year**

*df\_reviews['date']=pd.to\_datetime(df\_reviews['date'])*

**# 8 Display year**

*df\_reviews['date'].dt.year.value\_counts()*



**# 9 Displaying mimimum date**

*df\_reviews['date'].min()*



**# 10 Displaying Maximum date**

*df\_reviews['date'].max()*



**# 11 Displaying data graphically**

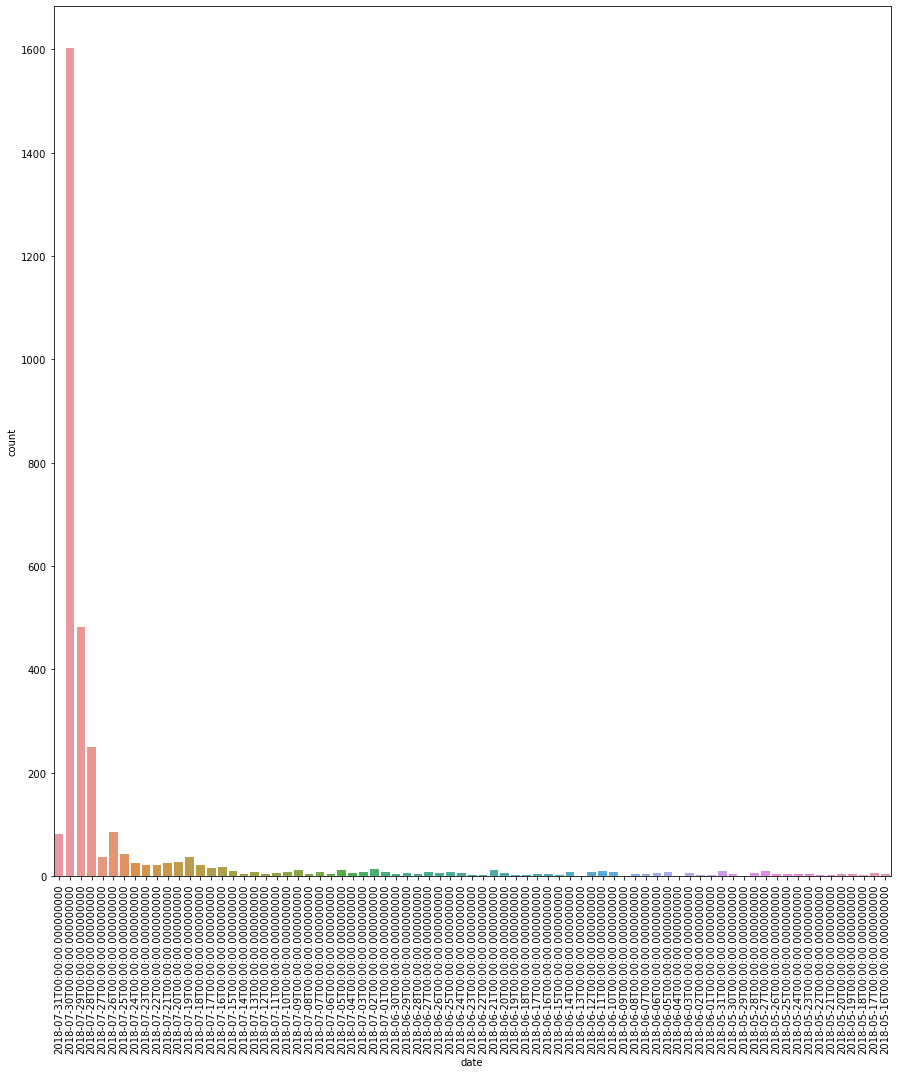
*plt.figure(figsize=(15,16))*

*sns.countplot(x='date',data=df\_reviews)*

*plt.xticks(rotation=90)*

*plt.show();*

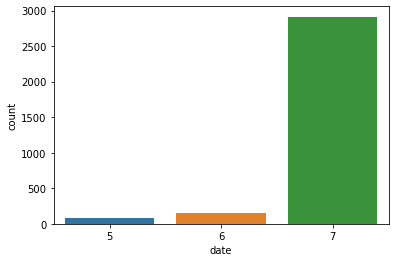
**Figure 1 Display of three months daily sales**



**# 12 Displaying three months sales**

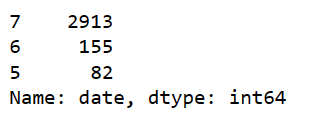
*sns.countplot(df\_reviews['date'].dt.month)*

**Figure 2 Display barplot of total monthly Sales**



**# 13 Displaying three months sales in numbers**

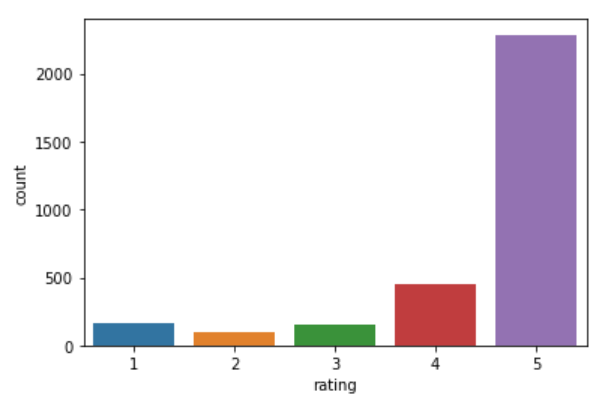
*df\_reviews['date'].dt.month.value\_counts()*



**# 14 Graphical Display of User Rating**

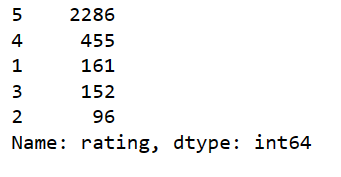
*sns.countplot(x='rating',data=df\_reviews)*

**Figure 3 Display sentiment rating of users**



**# 15 Display numerical values of ratings**

*df\_reviews.rating.value\_counts()*

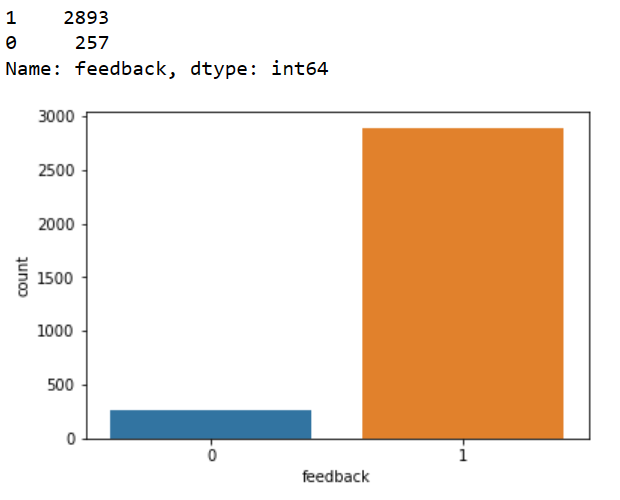


**# Step 16 Display graph and Numerical values of User feedback for Amazon Alexa products**

*sns.countplot(x='feedback', data=df\_reviews)*

*df\_reviews.feedback.value\_counts()*

**Figure 4 Display User Feedback**



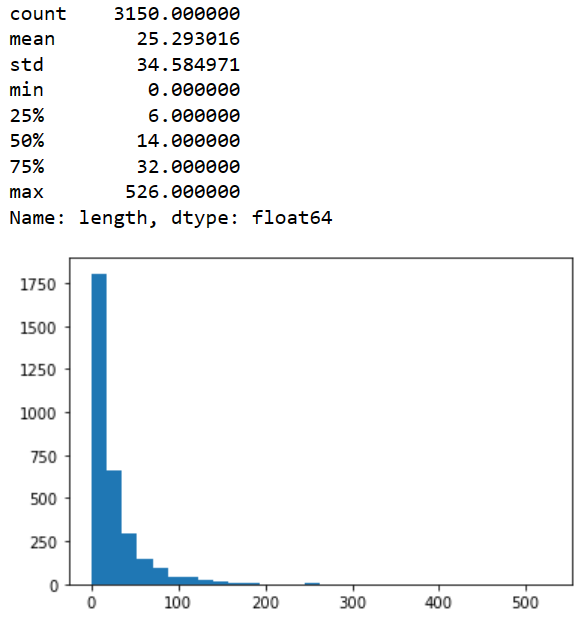
**# Step 17 Length of Review in number and graphs**

*df\_reviews['length']=df\_reviews['verified\_reviews'].apply(lambda x:len(x.split()))*

*df\_reviews.head()*

*plt.hist(x='length',data=df\_reviews,bins=30)*

*df\_reviews.length.describe()*



**#Step 18 Generating Wordcloud**

*neg=df\_reviews[df\_reviews['feedback']==0]*

*from wordcloud import WordCloud, STOPWORDS*

*import matplotlib.pyplot as plt*

**#Step 19 Generating Wordcloud**

*neg=df\_reviews[df\_reviews['feedback']==0]*

*from wordcloud import WordCloud, STOPWORDS*

*import matplotlib.pyplot as plt*

*text =neg['verified\_reviews'].values*

*wordcloud=WordCloud(*

*width=3000,*

*height=2000,*

*background\_color='black',*

*stopwords = STOPWORDS).generate(str(text))*

*fig=plt.figure(*

*figsize=(40,30),*

*facecolor='k',*

*edgecolor='k')*

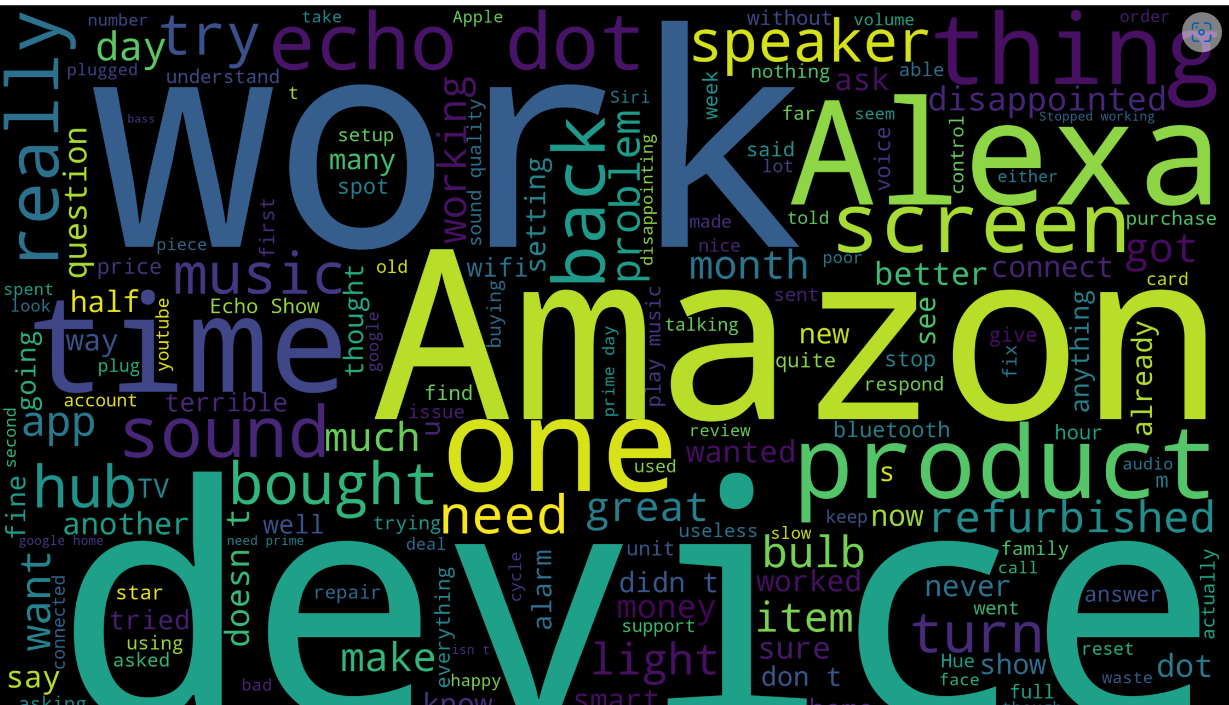
*plt.imshow(wordcloud,interpolation='bilinear')*

*plt.axis('off')*

*plt.tick\_params(axis='both',labelsize=14)*

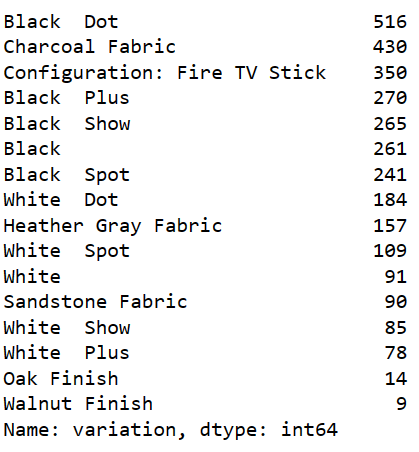
*plt.show()*

**Figure 5 Word cloud of Sentiment**

****

**# Step 20 Sentiment Variations count**

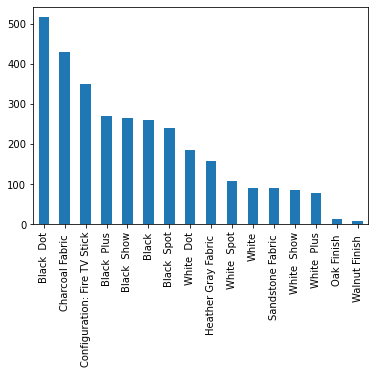
*df\_reviews['variation'].value\_counts()*



**Step 21 Plot of sentiment variation count**

*df\_reviews['variation'].value\_counts().plot.bar()*

**Figure 6 Sentiment Variation Count Plot**

****

**# Step 22 Sentiment Variation Distribution in Alexa**

*sns.countplot(x='variation',data=df\_reviews)*

*plt.title('Variation Distribution in Alexa')*

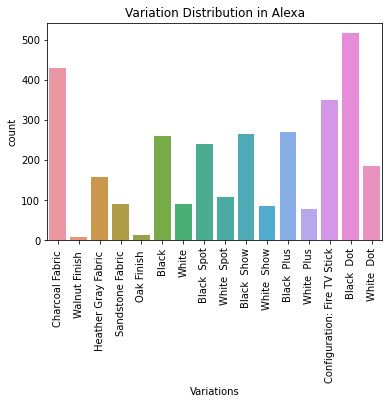
*plt.xlabel('Variations')*

*plt.ylabel('count')*

*plt.xticks(rotation='vertical')*

*plt.show()*

**Figure 7 Sentiment Variation Distribution in Alexa**



**5. CLAIMS**

**We Claim:**

1) Black Dot and charcoal fabric have been identified by more than 400 users of Amazon Alexa as the main cause of their negative sentiment.

2) In the synthesis of the word cloud there is use of jupyter note book, python coding with inventors loop, and use of modern libraries, therefore the reported method is scientific and hence bring true result of the sentiment.

3) The majority (87 Percentage) of the user of Amazon Alexa are highly satisfied by the utilities of the product and have higher sentiment then 4-point rating on a scale of 5.

4) 92 percentage of the users have positive feedback for the product based on the feedback values 2893(2893/3150) and bar chart created in figure 4.

5) Average positive sentiment rating is 4.46/5 = 89.20 percentage which indicates that the users of the products are having positive experiences with the features of the product and its utilities.

**Python – Pearson’s Chi-Square Test**

In this article, we will perform Pearson’s Chi-Square test using a mathematical approach and then using Python’s [**SciPy**](https://www.geeksforgeeks.org/data-analysis-with-scipy/)module. It is an important statistic test in data science for categorical column selection. generally in data science projects, we select only those columns which are important and are not [correlated](https://www.geeksforgeeks.org/exploring-correlation-in-python/) with each other.

**Pearson’s Chi-Square**

 Pearson’s Chi-Square is a statistical hypothesis test for independence between categorical variables. We will perform this chi-square test first using a mathematical approach and then using Python’s scipy module.

Let us know some terms before we understand the chi-square distribution

**The Contingency Table**

The Contingency table (also called crosstab) is used in statistics to summarise the relationship between several categorical variables. Here, we are taking a table that shows the number of men and women buying different types of pets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | dog | cat | bird | total |
| men | 207 | 282 | 241 | 730 |
| women | 234 | 242 | 232 | 708 |
| total | 441 | 524 | 473 | 1438 |

**Null Hypothesis**

A [null Hypothesis](https://www.geeksforgeeks.org/difference-between-null-and-alternate-hypothesis/) is a general statistical statement or assumption about a population parameter that is assumed to be true Until we have sufficient evidence to reject it.

It is generally denoted by Ho.

**Alternate Hypothesis**

 The Alternate Hypothesis is considered as competing of the null hypothesis. It is generally denoted by **H1**. The general goal of our hypothesis testing is to test the Alternative hypothesis against the null hypothesis.

**P-Value**

A P-value is used as a measure of evidence against the null hypothesis. If it is greater than our level of significance then we will accept our null hypothesis.

**Chi-Square Mathematical Approach**

The aim of this chi-square test is to conclude whether the two variables( gender and choice of pet ) are related to each other not.

**Null hypothesis:**We start by defining our null hypothesis (**H0)**which states that there is *no relation* between the variables.

**Alternate** **hypothesis:**It would state that there is a significant relationship between the two variables.

We will verify our hypothesis using these methods:

**Using p-value:**

We will define a significant factor to determine whether the relation between the variables is of considerable significance. Generally, a significant factor or *alpha value* of 0.05 is chosen. This alpha value denotes the probability of erroneously rejecting **H0** when it is true. A lower alpha value is chosen in cases when we expect more precision. If the p-value for the test comes out to be strictly greater than the alpha value, then we will accept our **H0**.

**Using chi-square value:**

If our calculated value of chi-square is less than or equal to the tabular(also called **critical**) value of chi-square, then we will accept our H0.

**Expected Values Table**

Next, we prepare a similar table of calculated(or expected) values. To do this we need to calculate each item in the new table as:

The expected values table :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | dog | cat | bird | total |
| men | 223.87343533 | 266.00834492 | 240.11821975 | 730 |
| women | 217.12656467 | 257.99165508 | 232.88178025 | 708 |
| total | 441 | 524 | 473 | 1438 |

**Chi-Square Table:** We prepare this table by calculating for element item through this formula.

The chi-square table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | observed (o) | calculated (c) | (o-c)^2 / c |
|  | 207 | 223.87343533 | 1.2717579435607573 |
|  | 282 | 266.00834492 | 0.9613722161954465 |
|  | 241 | 240.11821975 | 0.003238139990850831 |
|  | 234 | 217.12656467 | 1.3112758457617977 |
|  | 242 | 257.99165508 | 0.991245364156322 |
|  | 232 | 232.88178025 | 0.0033387601600580606 |
| Total |  |  | **4.542228269825232** |

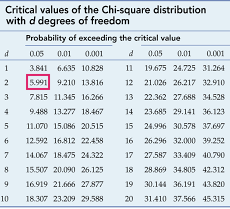
From this table, we obtain the total of the last column, which gives us the calculated value of chi-square.  Here the calculated value of chi-square is **4.542228269825232**

Now, we need to find the [**critical** value](https://www.geeksforgeeks.org/how-to-find-the-z-critical-value-in-python/) of the chi-square distribution. We can obtain this from the chi-square distribution table. To use this table, we need to know the [**degrees of freedom**](https://www.geeksforgeeks.org/degrees-of-freedom-formula/) for the dataset.

The degrees of freedom is defined as : **(no. of rows – 1) \* (no. of columns – 1).**

Hence, the degrees of freedom is **(2-1) \* (3-1) = 2**

Now, let us look at the table and find the value corresponding to **2**degrees of freedom and a **0.05** significance factor



*chi-square distribution table*

The tabular or critical value of chi-square here is  **5.991**

Hence

So here, we will accept our null hypothesis **H0**, that is our variables do not have a significant relation.

Next, let us see how to perform this chi-square test in Python.

**Performing the test using Python (scipy. stats) :**

SciPy is an Open Source Python library, which is used in mathematics, engineering, scientific and technical computing.

**Installation:**To install scipy in our notebook, we will use this command.

pip install scipy

The **chi2\_contingency()** function of **scipy.stats** module takes the contingency table element in 2d array format and it returns a tuple containing *test statistics*,  *p-value*, *degrees of freedom,* and *expected table*(the one we created from the calculated values) in that order.  Here, we need to compare the obtained p-value with an alpha value of 0.05.