**PGPDSE FT Mini-Project EDA and Statistics**

**Mod Clothing Dataset**

**Participant**

**Prakhar Purohit**

**Rupal Sanjay Nikum**

**Shachi Vaidya**

**Jitendra Dixit**

**Subham Pratik Khuntia**

**Industry Review**

ModCloth is the brainchild of husband-and-wife team Eric Koger and Susan Gregg Koger. When Susan collected too many vintage clothes to wear, Eric offered to help her put them online, which eventually inspired the couple to start the site in 2002. Today, the company has three offices—Pittsburgh, San Francisco and Los Angeles—456 employees (135 of them in San Francisco), around $48 million in funding and a reported $100 million in annual revenues.

ModCloth’s success comes largely from its interaction with customers. The site allows users to offer feedback to designers on how to make clothes better through its “Be the Buyer” program, and upload their own designs for production as part of the “Make the Cut” program. Photos of users wearing ModCloth clothes are displayed prominently on the site, as are images of ModCloth staffers in their chosen outfits of the day.

Online Fashion retailers have significantly increased in popularity over the last decade, making it possible for customers to explore hundreds of thousands of products without the need to visit multiple stores or stand in long queues for checkout. However, the customers still face several hurdles with current online shopping solutions. For example, customers often feel overwhelmed with the large selection of the assortment and brands. In addition, there is still a lack of effective suggestions capable of satisfying customers’ style preferences, or size and fit needs, necessary to enable them in their decision-making process. Moreover, in recent years social shopping in fashion has surfaced, thanks to platforms such as Instagram, providing a very interesting opportunity that allows to explore fashion in radically new ways. Such recent developments provides exciting challenges for Recommender Systems.

**Dataset and Domain**

Number of customers: 47,958

Number of products: 1,378

Number of transactions: 82,790

Field Description:

item\_id: unique product id

waist: waist measurement of customer

size: the standardized size of the product

quality: rating for the product

cup size: cup size measurement of customer

hips: hip measurement of customer

bra size: bra size of customer

category: the category of the product

bust: bust measurement of customer

height: height of the customer

length: feedback on the length of the product

fit: fit feedback

user\_id: a unique id for the customer

shoe size: shoe size of the customer

shoe width: shoe width of the customer

review\_text: review of customer

review\_summary: review summary

**Business Importance**:

1. Main Goals:

1. Using the core statistical theoretical concepts and knowledge to solve real time problem statements.

2. Visualize a real time industry scenario where one can use these statistical concepts.

3. Detailed data analysis and number crunching using statistics

4. Exhaustive report building using EDA and visualization techniques to help the business take decisions using insights from the data.

2. Customer Satisfaction:

* Client can do his shopping comfortably without any risk 3.
* Helping the customer with wide variety of their choices.
* Using EDA it is easy to understand requirement of customer choices.

3.Saving Customer Time:

* + Client doesn't need to go to shop for shopping .

4. Protecting The Customer choices:

* + It helps the customer to be satisfied and comfortable in his choices.

5. Part -I is concept based and walks you through various concepts of descriptive statistics, probability distributions and inferential statistics including confidence intervals and hypothesis testing.

Part -II on the other hand is dataset based and explore various data cleaning options, data analysis options and using EDA to derive deep and meaningful insights for the business

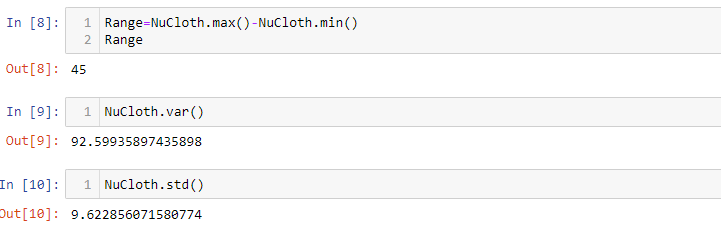
**Q1. Compute the mean, median and the mode of the data**

* Import dataset to workfield
* Here We use mean(),median (),mode() functions to find required result.



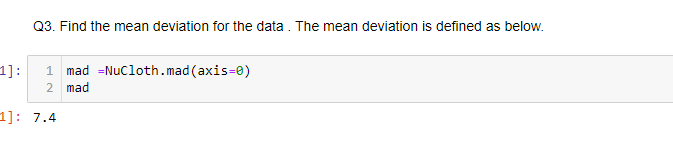
**Q2. Compute the range , variance and standard deviation of the data .**

* Here we use .var(),and .std () function which help us to find variance and standard deviation of given data.
* Variance and Standard Deviation are the two important measurements in statistics. Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data.
* The basic difference between both is standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.



**Q3. Find the mean deviation for the data . The mean deviation is defined as below.**

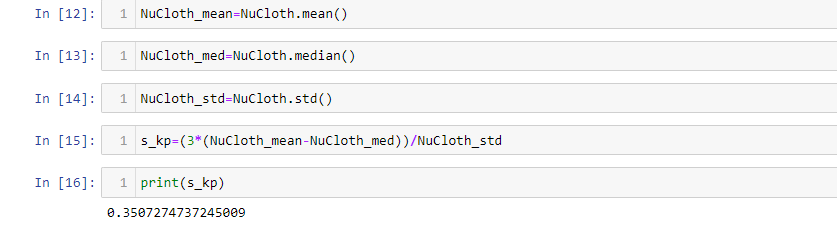
* .mad() function used to find mean absolute deviation.
* The mean absolute deviation of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.

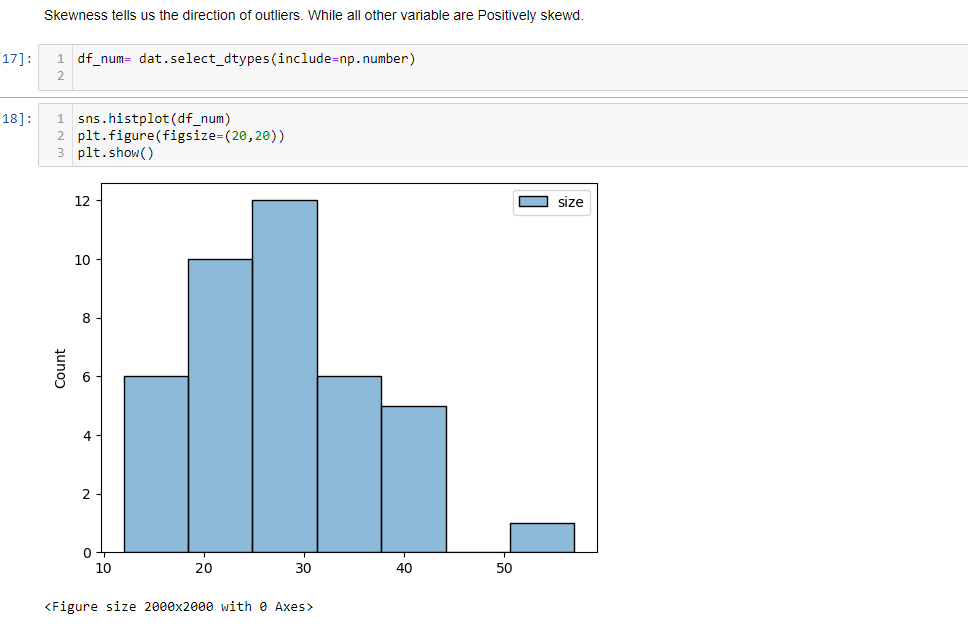


**Q4. Calculate the Pearson coefficient of skewness and comment on the skewness of the data[A measure to determine the skewness of a distribution is called the Pearson coefficient of skewness. The formula is**

**The value of the coefficient if skewness usually ranges from –3 to 3. When the distribution is symmetric, the coefficientis zero; when the distribution is positively skewed , the coefficient is positive, and when the distribution is negatively skewed the coefficient is negative.]**

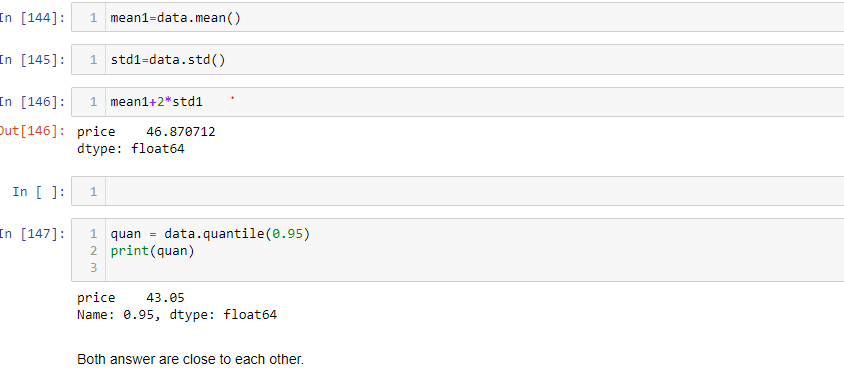
* Pearson’s coefficient of skewness is a method developed by Karl Pearson to find skewness in a sample using descriptive statistics like the mean and mode. Skewness is one measure of the shape of a set of data.





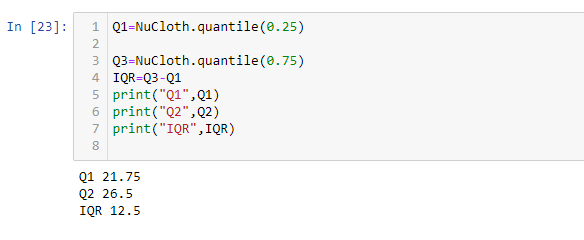
**Q5. Count the number of data values that fall within two standard deviations of the mean. Compare this with the answer from Chebyshev’s Theorem.**

* Chebyshev’s Theorem estimates the minimum proportion of observations that fall within a specified number of standard deviations from the mean. This theorem applies to a broad range of probability distributions. Chebyshev’s Theorem is also known as Chebyshev’s Inequality.
* Around 95% of values are within 2 standard deviations of the mean.
* So for this we use quantile method to find seconf=d standard deviation.



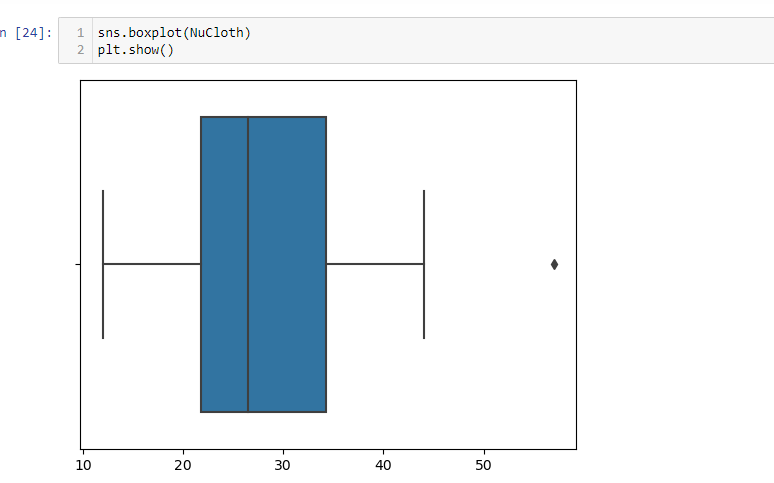
**Q6. Find the three quartiles and the interquartile range (IQR).**

* Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q1, Q2 and Q3, respectively.
* Q2 is nothing but the median, since it indicates the position of the item in the list and thus, is a positional average. To find quartiles of a group of data, we have to arrange the data in ascending order.
* The **interquartile range (IQR)** is the difference between the upper and lower quartile of a given data set and is also called a midspread. It is a measure of statistical distribution, which is equal to the difference between the upper and lower quartiles.

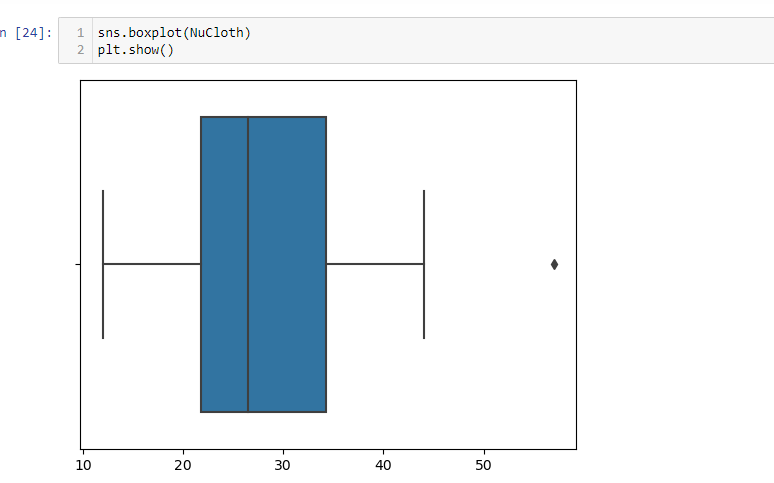


**Q7. Are there any outliers in the data set ?**

* An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

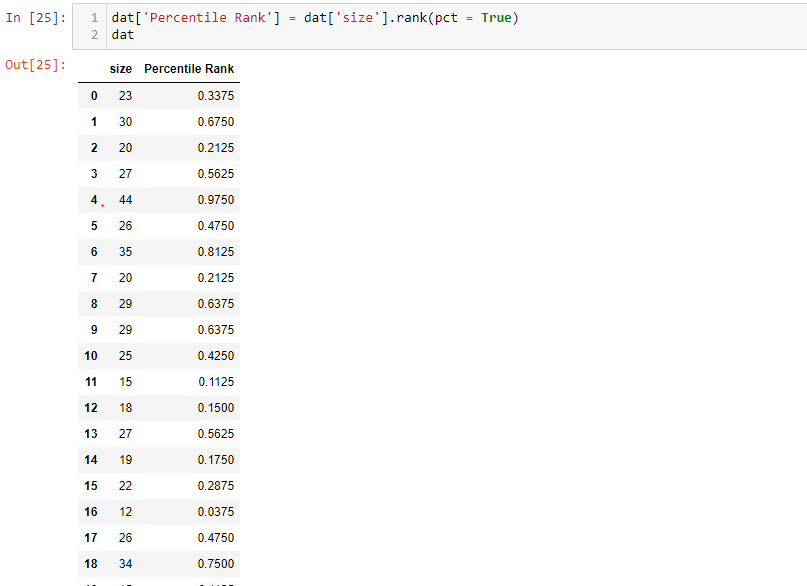


**Q8. Draw a boxplot of the dataset to confirm .**

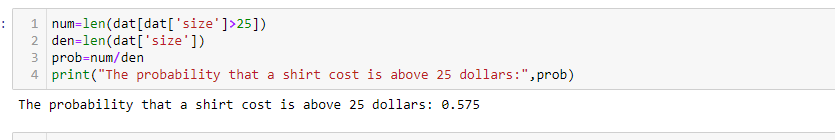


**Q9. Find the percentile rank of the datapoint 25.**

* Percentile rank is a common statistical measurement that you can use for everything from comparing standardized test scores to analysing weight distribution in a sample.
* Statisticians often use percentile rank to get an idea of how a particular assessment score or result compares with others in a set.

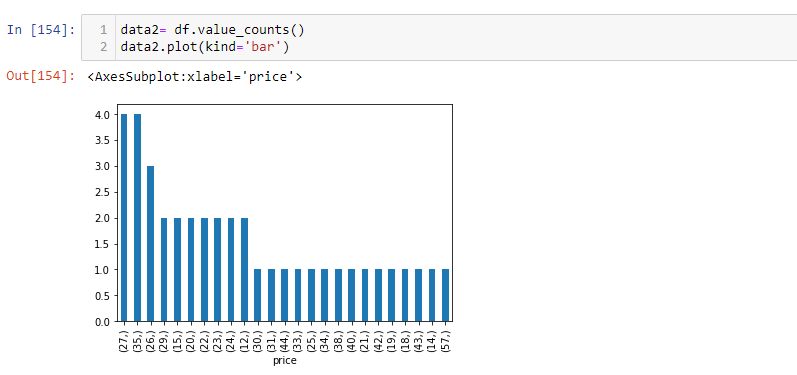


**Q10. What is the probability that a shirt cost is above 25 dollars?**



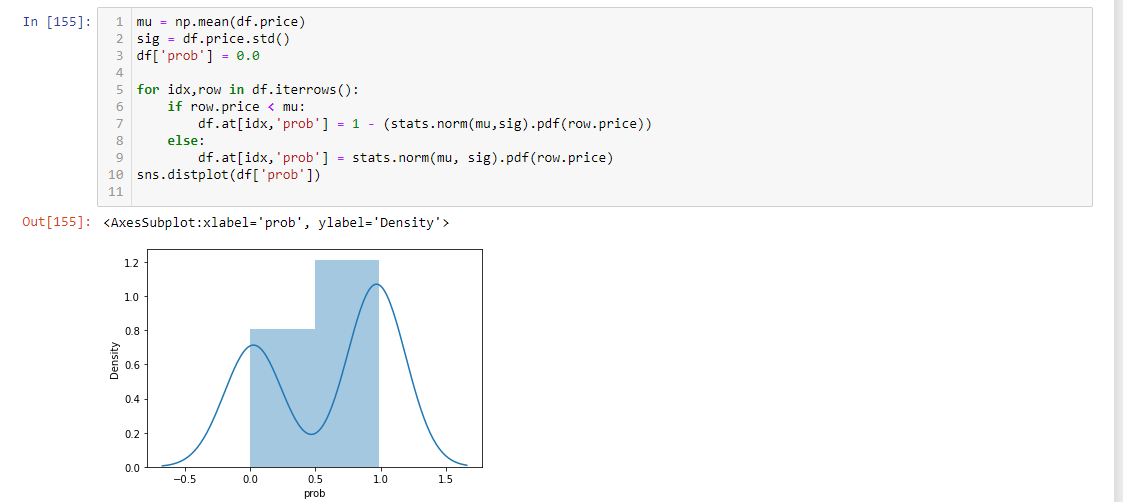
**Q11. Create a frequency distribution for the data and visualize it appropriately.**

* Frequency distributions are visual displays that organise and present frequency counts so that the information can be interpreted more easily.

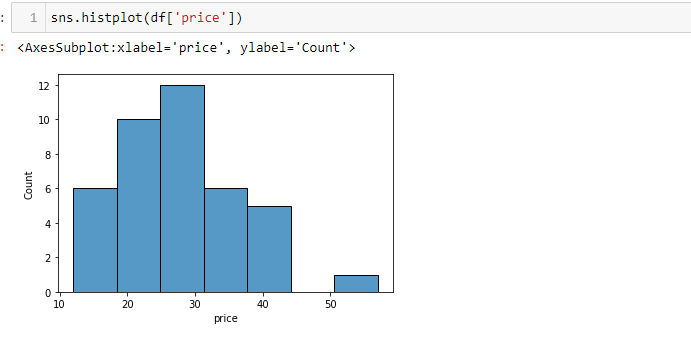


**Q12. Create a probability distribution of the data and visualize it appropriately.**

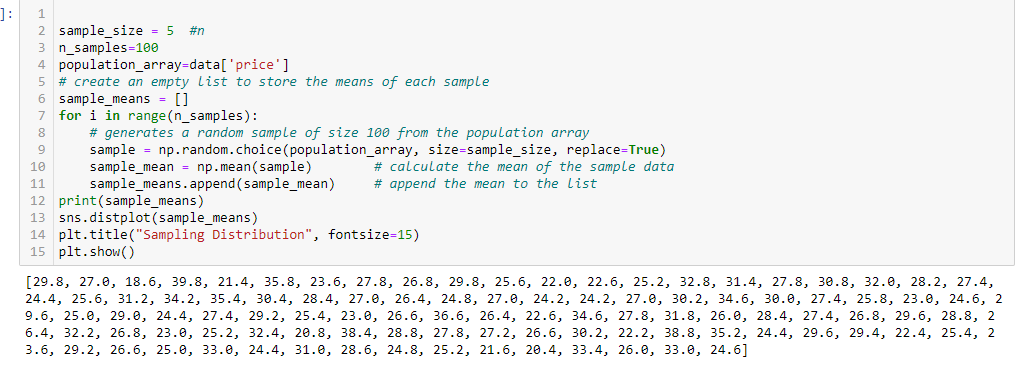
* Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics

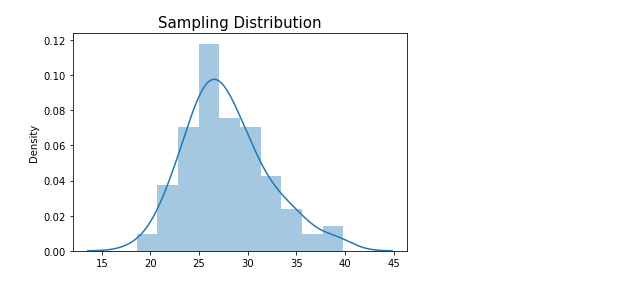


**Q13. What is the shape of the distribution of this dataset? Create an appropriate graph to determine that. Take 100 random samples with replacement from this dataset of size 5 each. Create a sampling distribution of the mean shirt prices. Compare with other sampling distributions of sample size 10, 15, 20, 25, 30. State your observations. Does it corroborate the Central Limit Theorem?**



A distribution is said to be positively skewed when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).



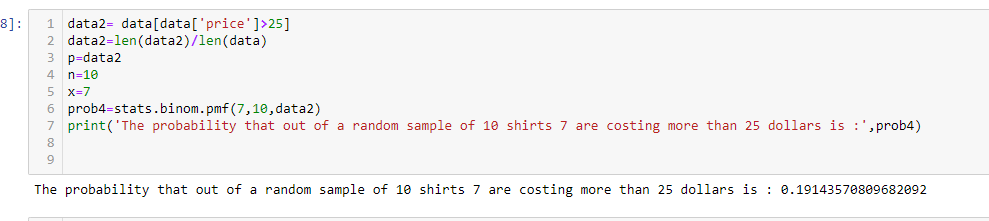


**Q14. Treat this dataset as a binomial distribution where p is the probability**

**that a shirt costs above 25 dollars. What is the probability that out of a random**

**sample of 10 shirts 7 are costing more than 25 dollars?**

* pmf function is a part of Python's SciPy library and is used to model probabilistic experiments with the help of binomial distribution.
* Pmf stands for **probability mass function.**



**Q15. NuCloth Claims that 60% of all shirts in their website cost less than 25**

**dollars . Using the Normal approximation of a Binomial distribution, find the**

**probability that in a random sample of 10 shirts 7 of them will cost less than 25**

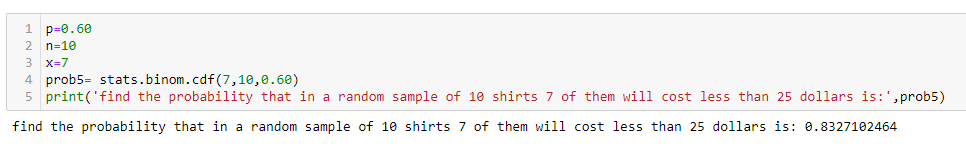
**dollars.**

**[ Note that the normal distribution can be used to approximate a binomial**

**distribution if np>=5 and nq>=5 with the following correction for continuity**

**P(X=z) = P(z-0.5 < X < z+0.5 ) ]**

* Binomcdf stands for binomial cumulative probability.

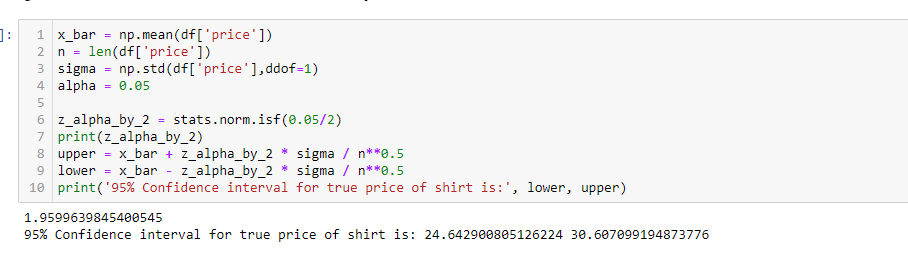


**Q16. Compute a 95% Confidence Interval for the true price of a shirt in the**

**NuCloth website using appropriate distribution.( State reasons as to why did**

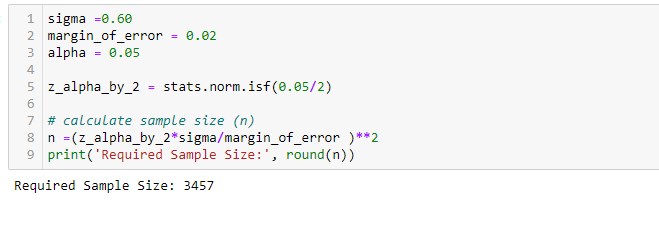
**you use a z or t distribution)**

* It is used to get the values of the inverse survival function.

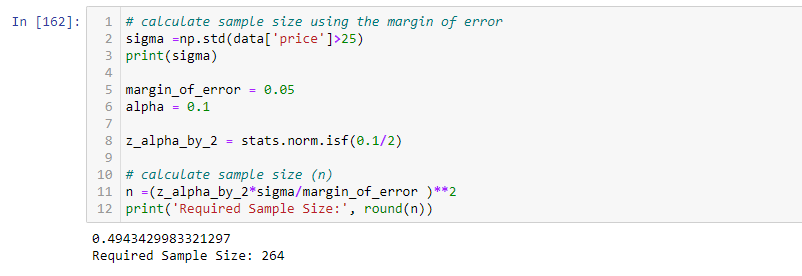


**Q17. A data scientist wants to estimate with 95% confidence the true proportion of shirts having price greater than 25 dollars in the NuCloth website. A recent study showed that 60% of all shirts have a price greater than 25 dollars. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.**

* Margin of error, also called confidence interval, tells you how much you can expect your survey results to reflect the views from the overall population.

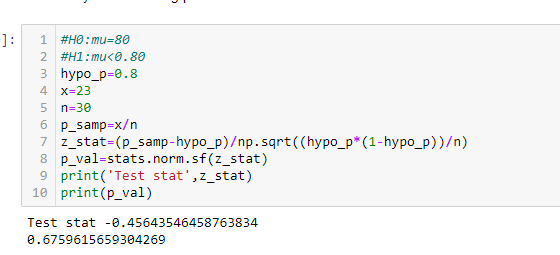


**Q18. The same data scientist wants to estimate the true proportion of shirts having price greater than 25 dollars. She wants to be 90% confident and accurate within 5% of true proportion. Find the minimum sample size necessary.**



**Q19. NuCloth claims that currently 80% of all shirts have prices greater than 25 dollars . Test this claim with an alpha =0.05 if out of a random sample of 30 shirts only 23 are having prices above 25 dollars.**

* A two-proportion Z-test is a statistical hypothesis test used to determine whether two proportions are different from each other.



**Q20. A data scientist is researching the hypothesis that the average price of a shirt in NuCloth is higher than the supermarket. So he collects data from NuCloth and the supermarket that the average shirt price of shirts in NuCloth is 36 dollars vs 34 dollars in the supermarket. The standard deviations are 7.6 and 5.8 respectively.Suppose the data scientist got these values after randomly selecting 100 shirts from each place.**

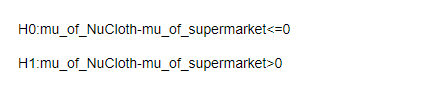
1. **What hypothesis would he use to compare the shirt prices of NuCloth vsSupermarket**

**b. What are critical values to be used?**

**c. What statistical test will be used to compare these prices?**

**d. Complete the test and obtain the P-value.**

**e. Summarize his conclusion based on the P-value.**





**STATS MINI PROJECTS**

This dataset contains self-reported clothing-fit feedback from customers as well as other

side information like reviews, ratings, product categories, catalog sizes, customers’

measurements (etc.) from 2 websites:

1. Mod Cloth

2. Rent the runway

1. Mod Cloth sells women’s vintage clothing and accessories, from which the

curator of the dataset collected data from three categories: dresses, tops, and bottoms.

2. Rent The Run Way is a unique platform that allows women to rent clothes for

various occasions; they collected data from several categories.

**Problem Statement:**

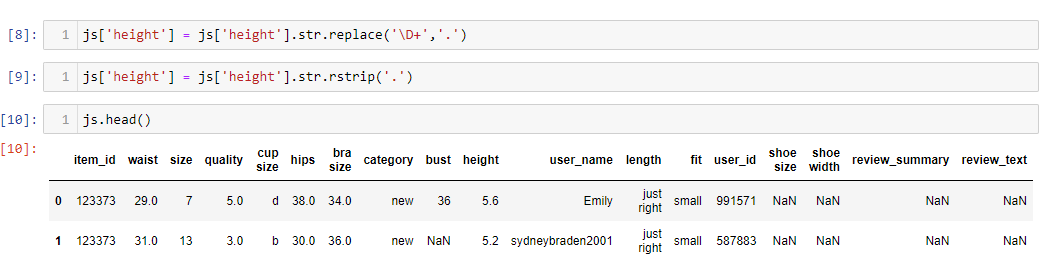
**Let’s assume you are working as a data scientist in a newly started textile company. You have given a dataset that consists of most of the features related to the measurements, clothes types, Product ID, User ID, etc. Now you suppose to work the dataset to identify the patterns to understand the customer’s preferred brands and how the company needs to brand its products based on the patterns that you will be found them out. Explore all the EDA concepts you learned and use a statistical test to ensure that your identification is true or false.**

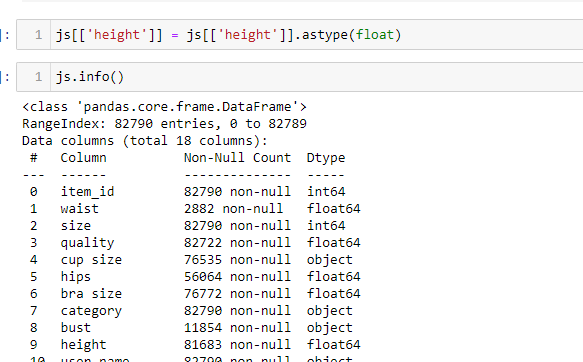
**Questions:**

1. **Read the datasets, Check the data types and Change the data types appropriately.**



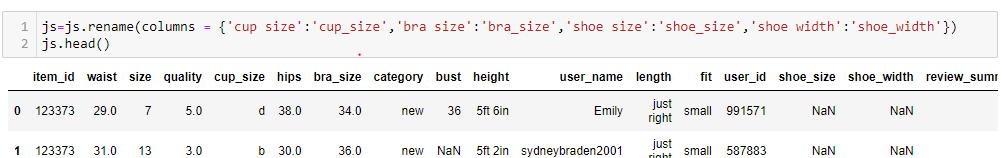
1. **Change the height column datatype to float after converting the values as shown**





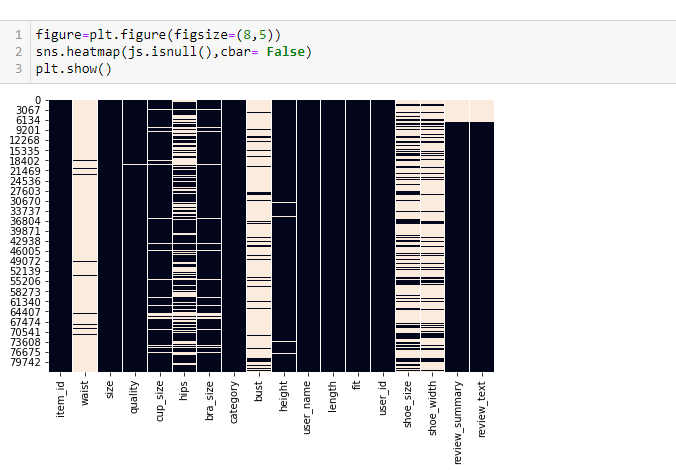
**3. Rename the names of the columns which have space in between the column.**

**Ex: shoe size as shoe\_size etc.**

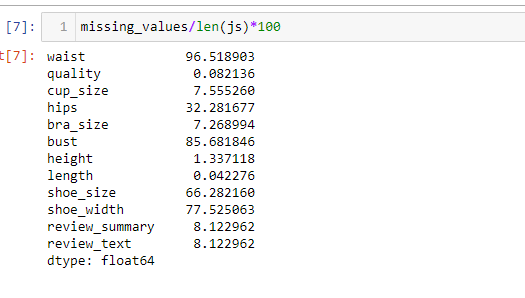


**4. Check the missing values and Identify the distribution of the variables to impute themissing values. Explain based on your analysis decide the features which can be dropped and Which can be imputed. And also explain the reason to choose the metric that you have chosen to impute the null values.**

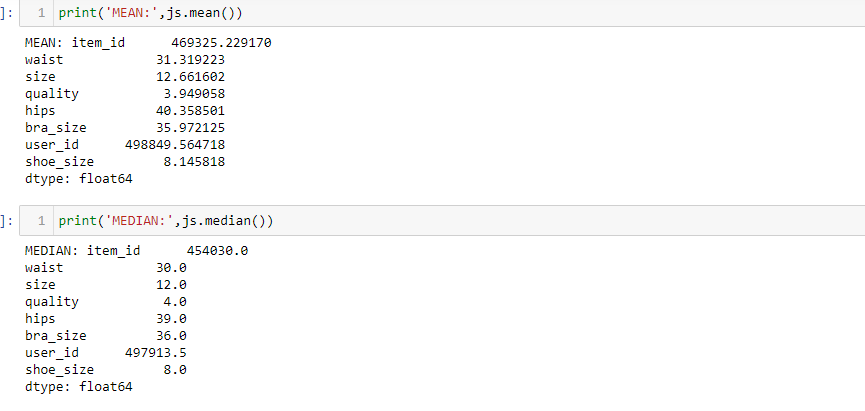
**Note: Kindly copy the datasets and work on the new data frame.**



**4.1 Check the count and percentage of missing values.**

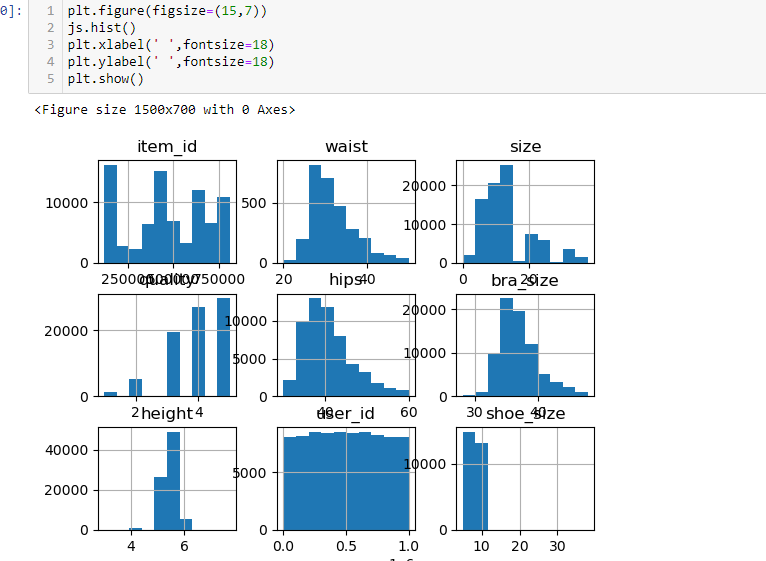


**4.2 check the Mean and Median.**



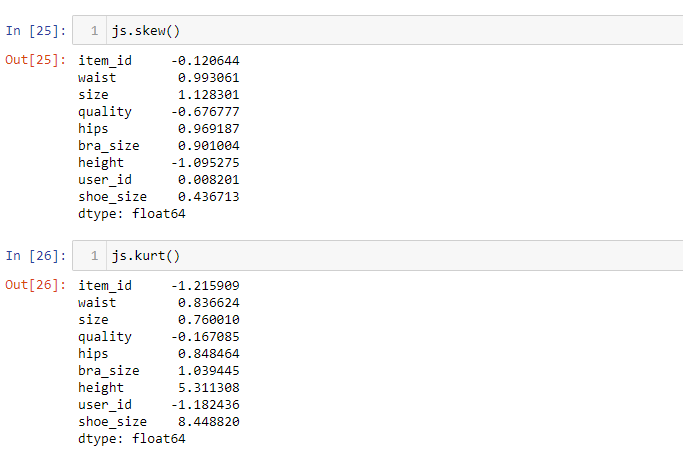
**4.3 Check the distribution of the variables using Histogram or Dist plot or KDE and boxplot etc.**

**Kindly explore at least two.**

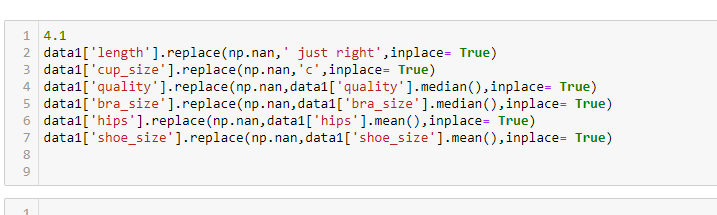


**4.4 Check the Skewness and Kurtosis.**

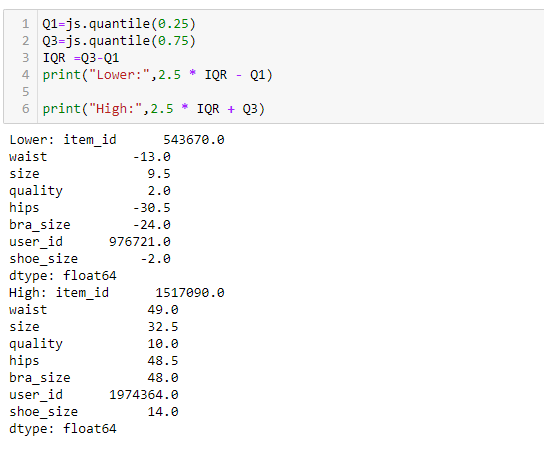
**Explain what the Skewness and Kurtosis describe.**



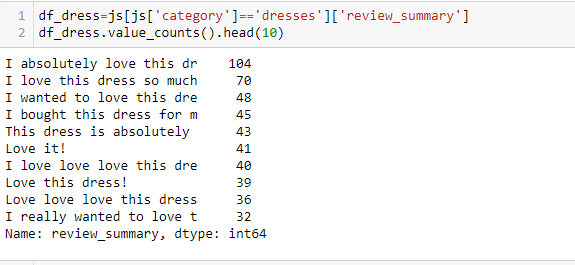
**4.5 Based on the above approach impute the missing values with the right metric. Or If you want to get some analysis before imputing missing values .feel free to explore the analysis.**



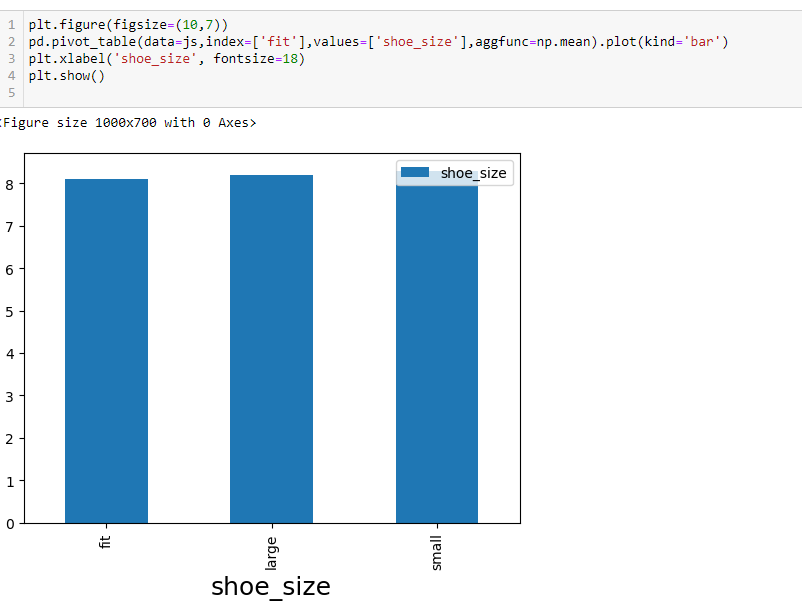
**5. Find the outliers which are below and above 2.5 \* IQR - Q1 and 2.5 \* IQR + Q3.**



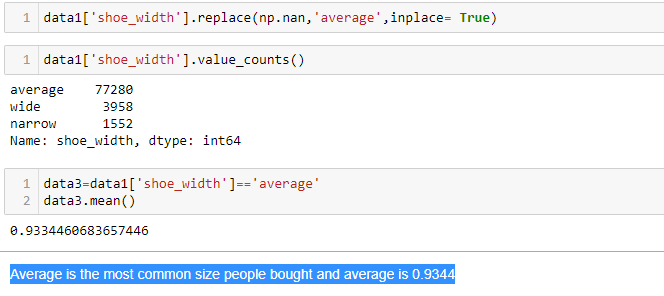
**6. Check for the category dress review and visualize the top 10 reviews using any relevant plot. Identify the negative reviews if there are any.**



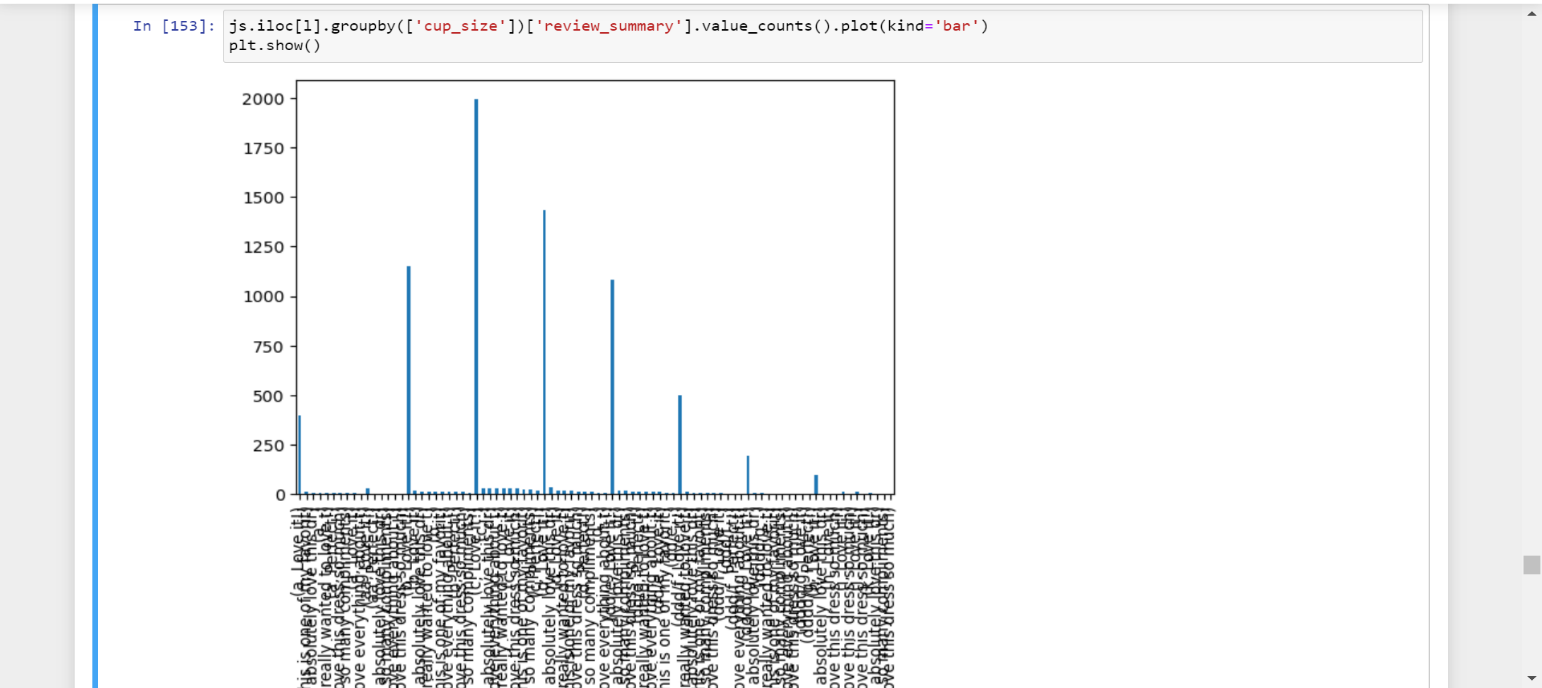
**7. Find out the average shoe size for the different fits of the customer. Visualize using any relevant plot. Explain that, Is there any significant different shoe sizes for different fits?**



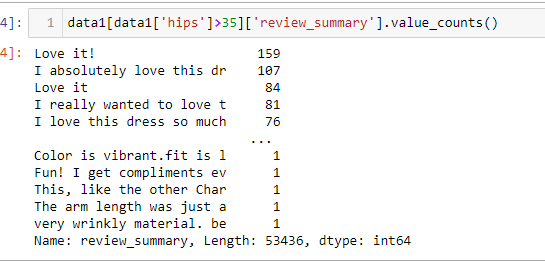
**8. Identify the customer’s common shoe width and average size for those who purchased the maximum quality. Is the mode of shoe width affect the user review? Visualize using the appropriate plot.**



**9. Extract the records belonging to the top 10 reviews, and then find the review summary for the different cup sizes. The basic analysis explains what you would try to infer. Try to use visualization.**

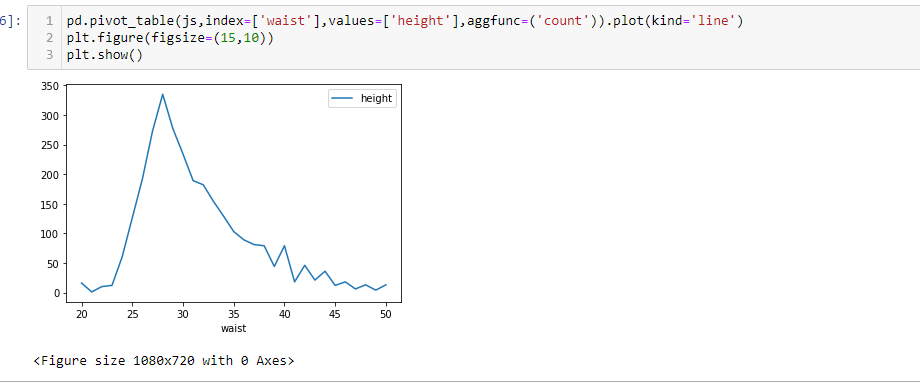
****

**10. Identify the most common review that we got from the customer whose hips size is greater than 35. Find out what kind of inference you can make.**



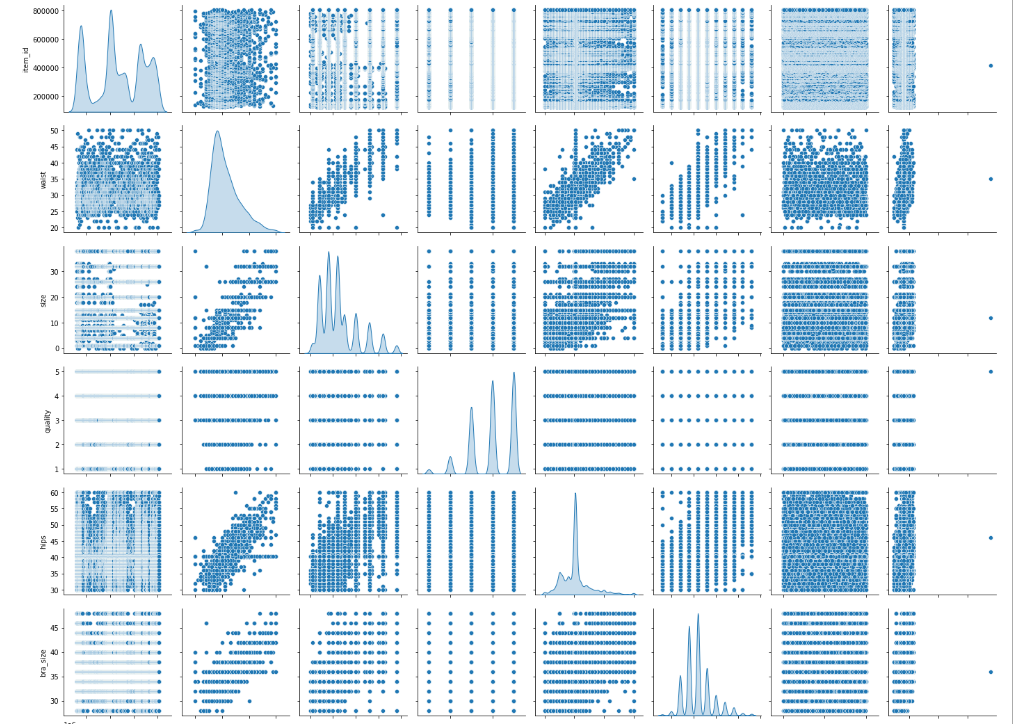
Mostly customer of hips size greater than 35 loves the product they purchase.So that the ModClot can have more number of cloths variety which can use ful for this customer.

**11. What is the relationship between height and weight? Describe what kind of relationship it has.**



**12. Plot the pair plot for the numerical plot. Explain according to your problem statement how the pair plot would help you.**



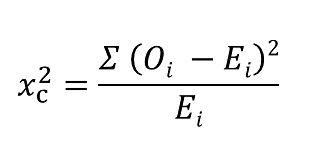


The default pairplot shows scatter plots between variables on the upper and lower triangle and histograms along the diagonal.

We can also see that the distribution of hips,bra\_size,size is heavily skewed to the right

**Statistical Analysis:**

* For solving all questions below we use chi\_test,Anova.
* A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.
* A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.
* Formula For Chi-Square Test



Where

c = Degrees of freedom

O = Observed Value

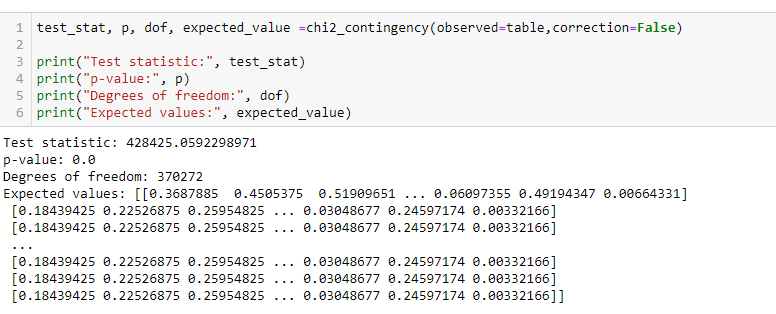
E = Expected Value

The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation. The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid. These tests are frequently used to compare observed data with data that would be expected to be obtained if a particular hypothesis were true.

* ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples.  Analysing variance tests the hypothesis that the means of two or more populations are equal.
* In a regression study, analysts use the ANOVA test to determine the impact of independent variables on the dependent variable.
* The most common method of performing an ANOVA test is one-way ANOVA. The one-way ANOVA means that the analysis of variance has one independent variable.
* You can use the one-way ANOVA to see if there are any significant differences between the means of your independent variables. When you know how each independent variable's mean differs from the others, you can figure out which of them is linked to your dependent variable and start to figure out what's driving that behaviour.
* The two-way analysis of variance is a variation of the one-way analysis. There are two independent variables in this equation (hence the name two-way). Factors are the two independent variables in a two-way ANOVA. The concept is that the dependent variable is influenced by two variables, or factors.

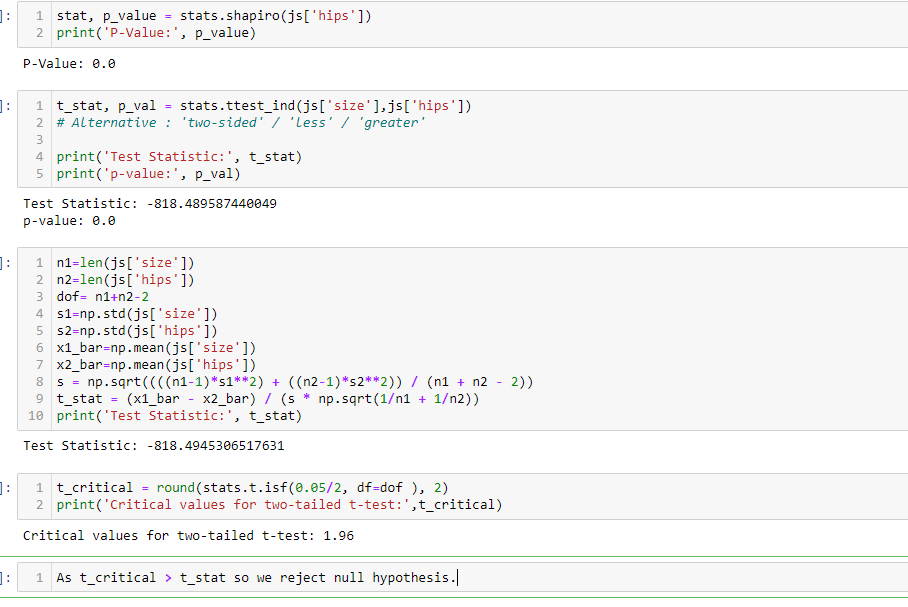
1. **Test the claim that the category feature and review summary have any relationship among them. The level of significance is 5%.**



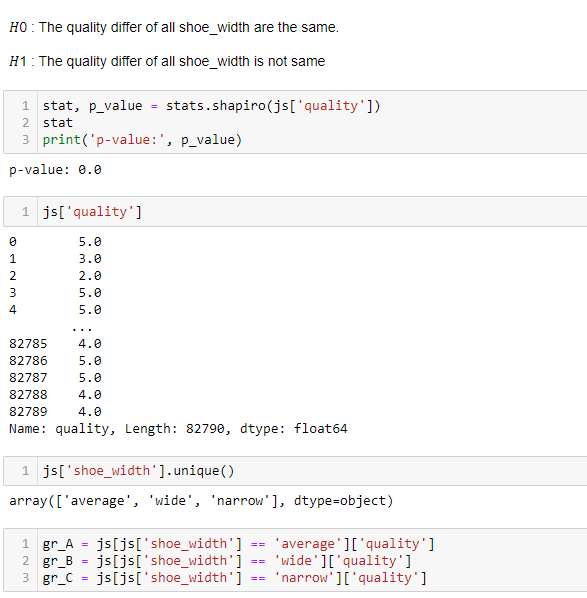


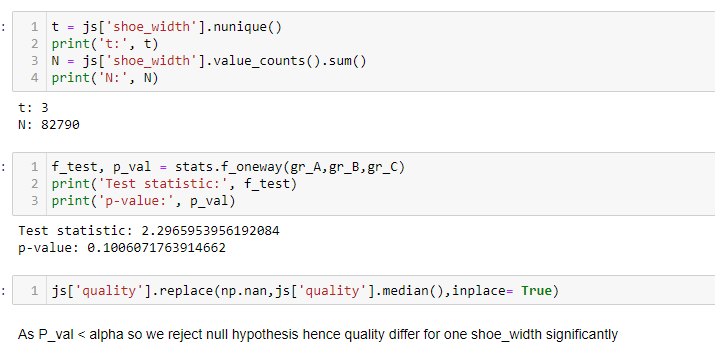
**As p\_val < alpha so we reject null hypothesis hence the variables category and review summary are dependent.**

1. **Test statistically whether the size and hips have any relationship using 0.05 alpha. Before the above test, Test the normality test.**



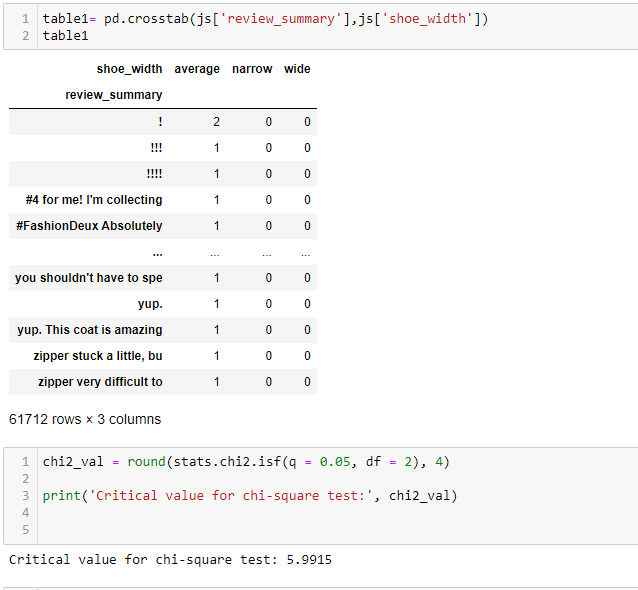
1. **Does the quality significantly differ for any one shoe width? Test the test with 96%confidence intervals.**

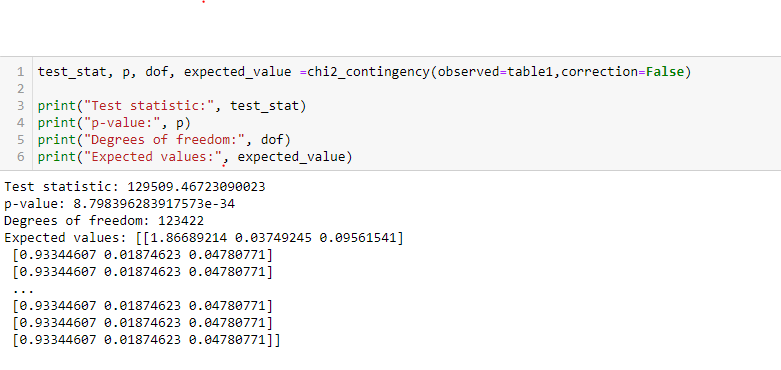




**As P\_val < alpha so we reject null hypothesis hence quality differ for one shoe\_width significantly**

**4. Check if the shoe width feature affects the review summary with a 99% confidence interval.**

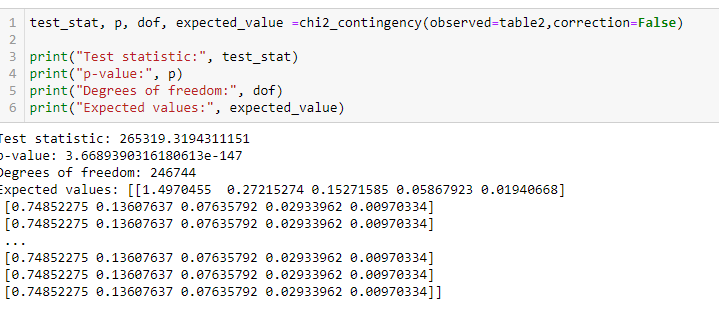




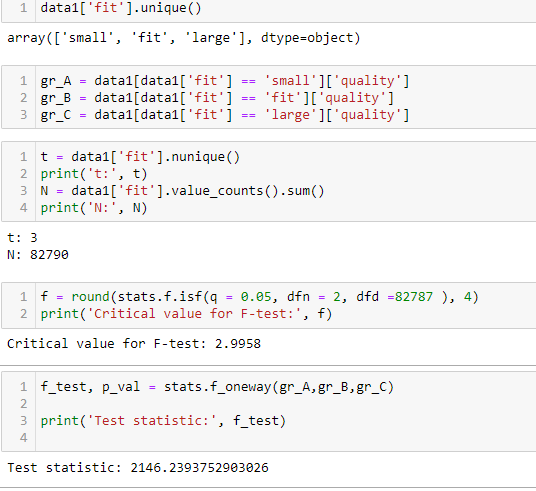
**As P-Val is < alpha so we reject null hypothesis and shoe\_width and reviw\_summary are dependent variable**

**5. Check if the length feature affects the review summary with a 95% confidenceinterval.**



**P\_val < alpha so we reject null hypothesis hence length and review summary are dependent variables.**

**6. Does the average quality significantly differ for the different fits? Kindly test the relevant hypothesis test by having 0.05 alpha.Check the normality of the data before the above test. Alpha = .05**



as t\_stat > critical value so we fail to reject null hypothesis hence average quality for different is same.

Conclusion:

So finally, there is a lot more in EDA than We just covered here. To sum up the EDA, we can say that it is really helpful to know data before you use it to train your model with it. For analysis part we use different visualization technique. We perform various data cleaning techniques and tried to gain some insight knowledge from the data.

We applied various statistical test which provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process