

Project Name

EDA and Statistical Analysis of Mod Clothing Dataset

Overview

This Statistics and EDA project is designed to train and test you on basic Data Exploratory and Statistical techniques used in the industry today. Apart from bringing you to speed with basic descriptive and

inferential methods, you will also deep dive into a dataset and perform thorough cleaning and analysis in order to draw useful business insights from the data. This will expose you to what data scientists do most often—Exploratory Data Analysis.

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

Specifications

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

PART-A (Concept Based)--25 points

The following data represents the price in dollars for branded shirts in a website NuCloth

23 30 20 27 44 26 35 20 29 29
25 15 18 27 19 22 12 26 34 15
27 35 26 43 35 14 24 12 23 31
40 35 38 57 22 42 24 21 27 33

Use this data for answering following questions where relevant

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

Q2. Compute the range , variance and standard deviation of the data Q3.

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

$$\text{Mean deviation} = \frac{\sum |X - \bar{X}|}{n}$$

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

$$\text{Skewness} = \frac{3(\bar{X} - MD)}{s}$$

where MD is the median and s the standard deviation

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

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.

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

Q7. Are there any outliers in the data set ?

Q8. Draw a boxplot of the dataset to confirm .

Q9. Find the percentile rank of the datapoint 25.

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

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

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?

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?

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 $np \geq 5$ and $nq \geq 5$ with the following correction for continuity $P(X=z) = P(z-0.5 < X < z+0.5)$]

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)

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.

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.

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.

- What hypothesis would he use to compare the shirt prices of NuCloth vs Supermarket
- What are critical values to be used?
- What statistical test will be used to compare these prices?
- Complete the test and obtain the P-value.
- Summarize his conclusion based on the P-value.

PART-B (Dataset Based)--25 points

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.

2. Change the height column datatype to float after converting the values as shown Below. 5.7

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 the missing 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.

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.

Statistical Analysis:


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

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

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

Check the normality of the data before the above test. Alpha = .05

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

- 
5. Check if the length feature affects the review summary with a 95% confidence interval.
 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