Project Name

Statistics and Exploratory Data Analysis Project -SET 1

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

Milestones

I. Two pillars of statistic - After completing solving the Part-A questions write in your own words (200 to 500 words) what is the importance of descriptive and inferential statistics and how each helps us to get better insights of the problem at hand

II. Exploratory data analysis - A thorough understanding of the dataset is the key to obtaining success in solving business problems using data science. After solving the EDA for the dataset, come up with your suggestions to the management for steps((200 to 500 words) to take to improve the business.

PART-A (Concept Based)--25 points

The following are the ages of 30 customers who ordered an EV scooter from Zen Automotives.ee.

42	44	62	35	20
30	56	20	23	41
55	22	31	27	66
21	18	24	42	25
32	50	31	26	36
39	40	18	36	22

Use this data for answering questions 1-13.

Q1. Compute the mean, median and the mode of the data
Q2. Compute the range , variance and standard deviation of customer ages
Q3. Find the mean deviation for the data . The mean deviation is defined as below.

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

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

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

$where\ MD\ is\ the\ median\ and\ s\ the\ standard\ deviation$

The value of the coefficient if 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 50.
Q10. What is the probability that a person ordering an EV scooter is above 50 years old?
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 age of customers. 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 person ordering an EV is above 50 years age. What is the probability that out of a random sample of 10 buyers exactly 6 are above 50 years of age?
Q15. A study claims that 10% of all customers for an EV scooter are above 50 years of age. Using the Normal approximation of a Binomial distribution, find the probability that in a random sample of 300 prospective customers exactly 25 will be above 50 years of age.
[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 $P(X=z) = P(z-0.5 < X < z+0.5)$]
Q16. Compute a 95% Confidence Interval for the true mean age of the population of EV scooter buyers for the dataset 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 proportion of people who own an EV in the population. A recent study showed that 20% of people interviewed had an EV. 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 proportion of executives who own an EV. She wants to be 90% confident and accurate within 5% of true proportion. Find the minimum sample size necessary.

Q19. A researcher claims that currently 20% of the population are owning EVs. Test his claim with an alpha =0.05 if out of a random sample of 30 two-wheeler owners only 5 own an EV.

Q20. Assume you are working for a Consumer Protection Agency that looks at complaints raised by customers for the transportation industry. Say you have been receiving complaints about the mileage of the latest EV launched by the Zen Automotives. Zen allows you to test randomly 40 of its new EVs to test mileage. Zen claims that the new EVs get a mileage of 96 kmpl on the highway. Your results show a mean of 91.3 kmpl and a standard deviation of 14.4.

a. Show why you support Zen's claim using the P-value obtained.
b. After more complaints you decide to test the variability of the mileage on the highway. On questioning Zen's quality control engineer , you find that they are claiming a standard deviation of 7.2. Test the claim about the standard deviation. [Hint :use the Chi-square test for variance/standard deviation]
c. Write a summary of results and the action that Zen must take to remedy the complaints.
d. What is your position on performing the test for variability along with the test for means?

Q21. Write a report comparing and contrasting Descriptive vs Inferential Statistics in about 200-500 words and how each helps to get more insight into the data at hand.

PART-B (Dataset based)--25 points

• **DOMAIN*:*** Sports
• **CONTEXT*:*** Company X manages the men's top professional basketball division of the American league system.The dataset contains information on all the teams that have participated in all the past tournaments. It has data about how many baskets each team scored, conceded, how many times they came within the first 2 positions,how many tournaments they have qualified, their best position in the past, etc.
• **DATA DESCRIPTION*:*** Basketball.csv - The data set contains information on all the teams so far participated in all the past tournaments.
• ATTRIBUTE INFORMATION:
1. Team: Team's name
2. Tournament: Number of played tournaments.
3. Score: Team's score so far.
4. PlayedGames: Games played by the team so far.
5. WonGames: Games won by the team so far.
6. DrawnGames: Games drawn by the team so far.
7. LostGames: Games lost by the team so far.
8. BasketScored: Basket scored by the team so far.
9. BasketGiven: Basket scored against the team so far.
10. TournamentChampion: How many times the team was a champion of the tournaments so far.
11. Runner-up: How many times the team was a runners-up of the tournaments so far.
12. TeamLaunch: Year the team was launched on professional basketball.
13. HighestPositionHeld: Highest position held by the team amongst all the tournaments played.

• **PROJECT OBJECTIVE*:*** Company's management wants to invest on proposals on managing some of the best teams in the league. The analytics department has been assigned with a task of creating a report on the performance shown by

the teams. Some of the older teams are already in contract with competitors. Hence Company X wants to understand which teams they can approach which will be a deal win for them.

Steps and tasks: [Total Score: 25 points]

1. Read the data set, clean the data and prepare a final dataset to be used for analysis.[5 points]
2. Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get data driven insights on recommending which teams they can approach which will be a deal win for them.. Also as a data and statistics expert you have to develop a detailed performance report using this data.[15 points]

Hint:

Use statistical techniques and visualization techniques to come up with useful metrics and reporting. Find out the best performing team, oldest team, team with highest goals, team with lowest performance etc. and many more.

These are just random examples. Please use your best analytical approach to build this report. You can mix match columns to create new ones which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns. Use graphical interactive libraries to enable you to publish interactive plots in python like plotly.

3. Please include any improvements or suggestions to the association management on quality, quantity, variety, velocity,veracity etc. on the data points collected by the association to perform a better data analysis in future.Submit a 200-500 words report to the management[5-points]