Be Practical and Knowledgeable

PS SOFTECH

CASE STUDY:
DATA EXPLORATORY ANALYSIS AND HYPOTHESIS TESTING FOR INSURANCE CLAIMS DATA

1. Import claims_data.csv and cust_data.csv which is provided to you and combine the two datasets appropriately to create a 360-degree view of the data. Use the same for the subsequent questions.
2. Perform a data audit for the datatypes and find out if there are any mismatch within the current datatypes of the columns and their business significance.
3. Convert the column claim_amount to numeric. Use the appropriate modules/attributes to remove the $ sign.
4. Of all the injury claims, some of them have gone unreported with the police. Create an alert flag (1,0) for all such claims.
5. One customer can claim for insurance more than once and in each claim, multiple categories of claims can be involved. However, customer ID should remain unique.
   Retain the most recent observation and delete any ***duplicated*** records in the data based on the customer ID column.
6. Check for missing values and impute the missing values with an appropriate value. (mean for continuous and mode for categorical)
7. Calculate the age of customers in years. Based on the age, categorize the customers according to the below criteria

Children	< 18
Youth	18-30
Adult	30-60
Senior	> 60

8. What is the average amount claimed by the customers from various segments?
9. What is the total claim amount based on incident cause for all the claims that have been done at least 20 days prior to 1st of October, 2018.

10. How many adults from TX, DE and AK claimed insurance for driver related issues and causes?
11. Draw a pie chart between the aggregated value of claim amount based on gender and segment. Represent the claim amount as a percentage on the pie chart.
12. Among males and females, which gender had claimed the most for any type of driver related issues? E.g. This metric can be compared using a bar chart
13. Which age group had the maximum fraudulent policy claims? Visualize it on a bar chart.
14. Visualize the monthly trend of the total amount that has been claimed by the customers. Ensure that on the “month” axis, the month is in a chronological order not alphabetical order.
15. What is the average claim amount for gender and age categories and suitably represent the above using a facetted bar chart, one facet that represents fraudulent claims and the other for non-fraudulent claims.

Based on the conclusions from exploratory analysis as well as suitable statistical tests, answer the below questions. Please include a detailed write-up on the parameters taken into consideration, the Hypothesis testing steps, conclusion from the p-values and the business implications of the statements.

16. Is there any similarity in the amount claimed by males and females?
17. Is there any relationship between age category and segment?
18. The current year has shown a significant rise in claim amounts as compared to 2016-17 fiscal average which was $10,000.
19. Is there any difference between age groups and insurance claims?
20. Is there any relationship between total number of policy claims and the claimed amount?