A company has customer data that contains 8 columns of customer details and another table having name customer\_policy data contains the policy details of the customer.

The company intends to offer some discount in premium for certain customers. To do that they ask their Data scientist team to get some information. Hence, following tasks DS team decided to perform:

1. Add the column names to both datasets:

i. Column Name for customer details table:

**customer\_id,**

**Gender,**

**age,**

**driving licence present,**

**region code,**

**previously insured,**

**vehicle age**

**and vehicle damage, in respective order.**

ii. Column Name for customer\_policy table:

**customer\_id,**

**annual premium (in Rs),**

**sales channel code,**

**vintage and response.**

2. Checking and Cleaning Data Quality:

i. Null values

* Generate a summary of count of all the null values column wise
* Drop Null values for customer\_id because central tendencies for id’s is not feasible.
* Replace all null values for numeric columns by mean.
* Replace all null values for Categorical value by mode.

ii. Outliers

* Generate a summary of count of all the outliers column wise
* Replace all outlier values for numeric columns by mean.

(Hint1: for outlier treatment use IQR method as follows:

For example: for a column X calculate Q1 = 25th percentile and Q3 = 75th percentile then IQR = Q3 – Q1 ) then to check outlier, anything lower than a Q1 – 1.5IQR or greater than Q3 + 1.5 IQR would be an outlier

Hint2: For getting percentile value, explore pd.describe() function)

iii. White spaces

* Remove white spaces

iv. case correction(lower or upper, any one)

v. Convert nominal data (categorical) into dummies

* for future modeling use if required

vi. Drop Duplicates (duplicated rows)

3. Create a Master table for future use. Join the customer table and customer\_policy table to get a master table using **customer\_id** in both tables.

(Hint: use pd.merge() function)

4. Company needs some important information from the master table to make decisions for future growth.They needs following information:

 i. Gender wise average annual premium

ii. Age wise average annual premium

iii. Is your data balanced between the genders?

          (Hint: Data is balanced if number of counts in each group is approximately same)

iv. Vehicle age wise average annual premium.

5. Is there any relation between Person Age and annual premium?

Hint: use correlation function (Correlation describes the relationship between two variables).

Correlation coefficient < -0.5           - Strong negative relationship

Correlation coefficient > 0.5            -  Strong positive relationship

0.5 < Correlation coefficient < 0.5   - There is no relationship.