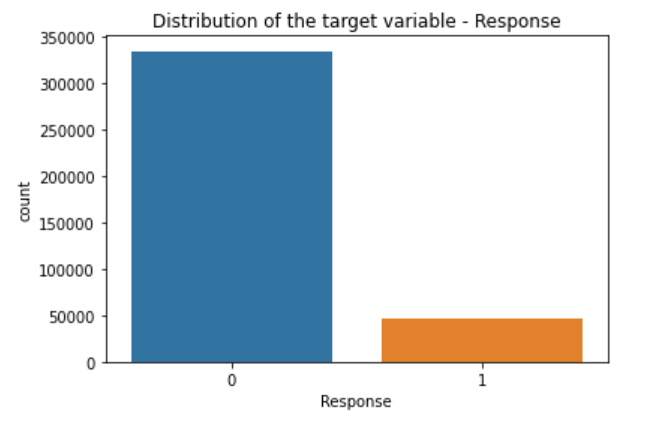
**IST 718 Big Data Analytics**

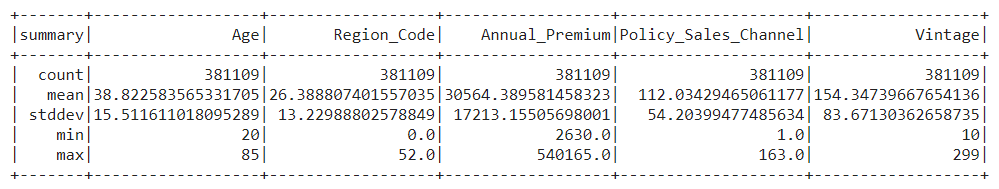
**Health Insurance Cross Sell Prediction**

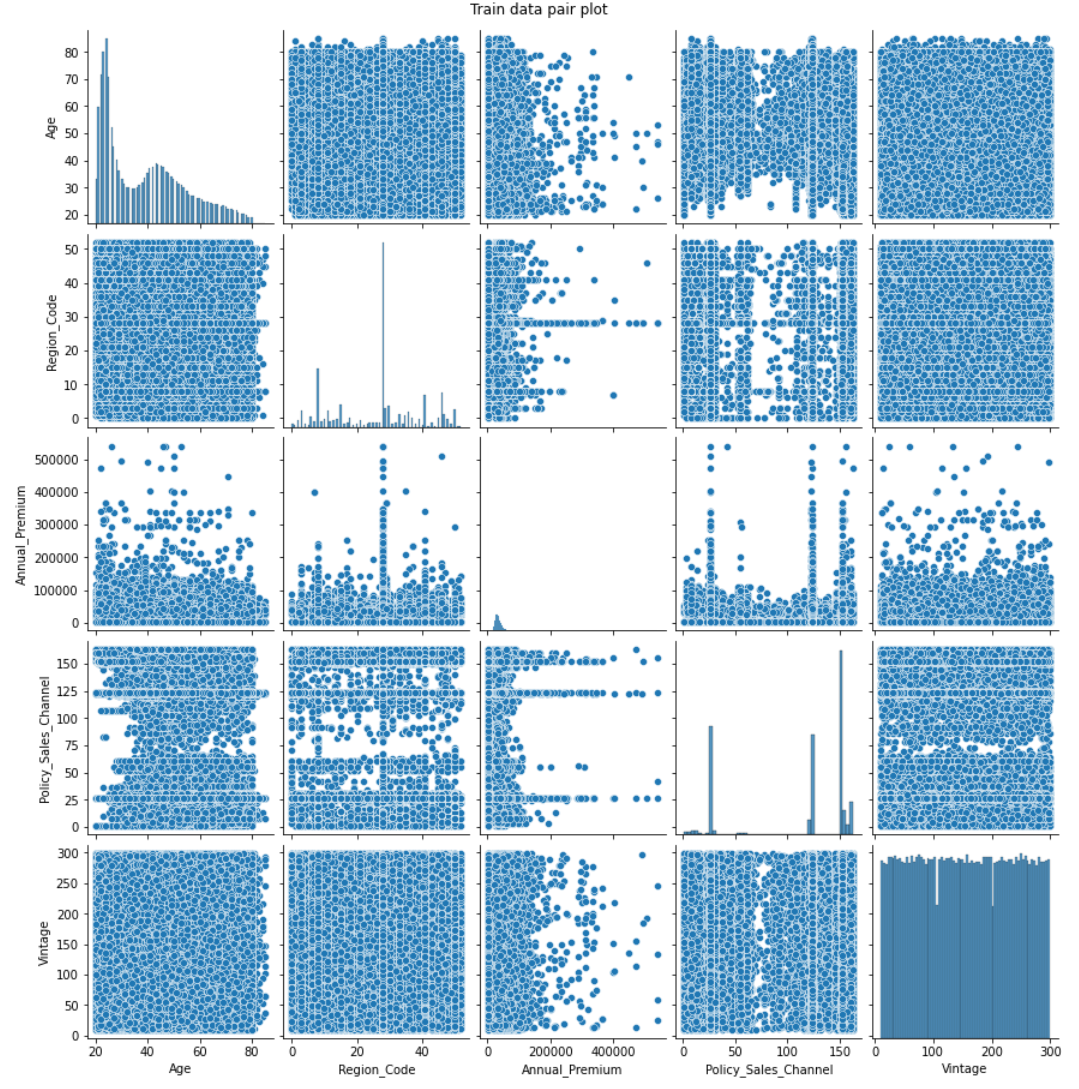
**Project Proposal**

* **Project Title:** Health Insurance Cross Sell Prediction
* **Group Number**: Group 6
* **Group Members:** Aatish Suman, Adheesh Phadnis, Sharvil Turbadkar, Snehal Ghatpande
* **Objective:** The objective of the project is to build a model to predict whether the policyholders from the past year will also be interested in purchasing Vehicle Insurance provided by the company.
* **Data Set Description:**
* **Overview/ Description:** The goal of the project is to build a model to predict whether a policyholder would be interested in Vehicle Insurance. It would be extremely helpful for the company because it can aid the company to plan its communication strategy to reach out to those customers and optimize its business model and revenue. To predict whether the customer would be interested in Vehicle insurance we are using data about the demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.
* **Number of rows and columns:** The training data set consists of 381109 rows and 12 columns while the test data set consists of 127037 and 11 columns.
* **Sample predictors:** id (unique customer ID), Gender (gender of the customer), Age (age of the customer), Driving\_License (whether or not the customer has a driving license), Region\_Code (region code of the customer), Previously\_Insured (whether or not the customer already has vehicle insurance), Vehicle\_Age (age of customer’s vehicle), Vehicle\_Damage (whether or not the customer’s vehicle has been damaged in the past), Annual\_Premium (the amount customer pays as premium annually), Policy\_Sales\_Channel (channel used to reach out to the customer), Vintage (number of days the customer has been associated with the company) are the sample predictors.
* **Link to the dataset:** <https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv>
* **Preliminary Data Exploration:** The customers’ responses to whether or not they want to buy vehicle insurance from the company are heavily imbalanced. Accuracy is not a good metric as data is heavily imbalanced with a lot of customers planning to not pursue a vehicle insurance

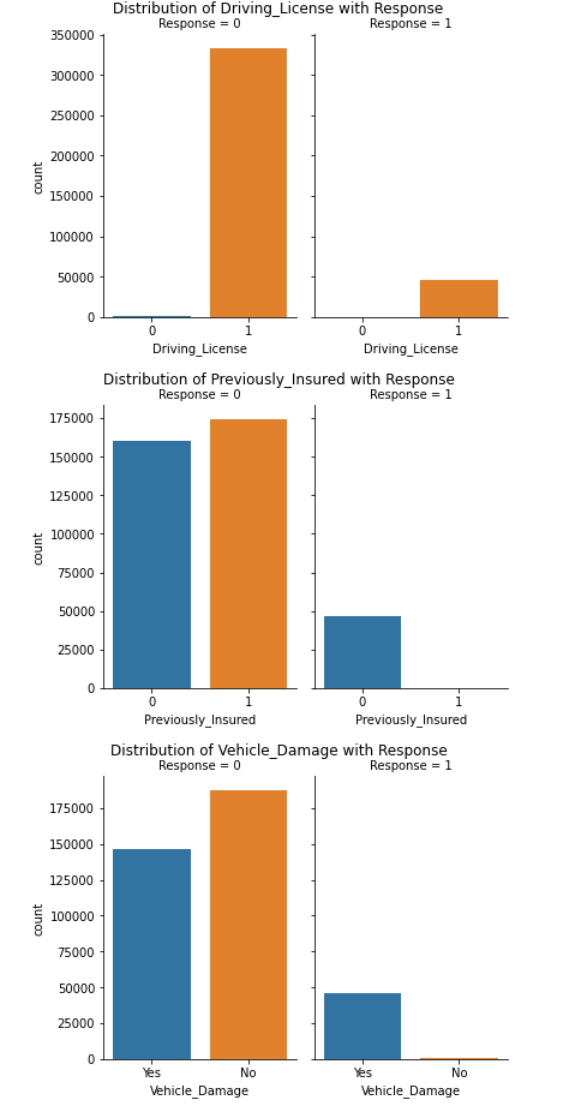


The min mean and max values for each numerical variable are comparable, which reduces the possibility of erroneous data. The data does not have null values. The unit of measurement of different variables are different, as a result scaling might be required. Annual Premium in dollars is much higher than vintage and Age as both scales have different units of measurement .Some Attributes like Annual premium have very high values which need to scaled to recue impact on target variable .



There doesn't seem to be any correlation between the numerical variables like Age, Region\_Code, Annual\_Premium, Policy\_Sales\_Channel and Vintage. Annual\_Premium is heavily left skewed. Age is left skewed. 

All interested customers (Response=1) have a driving license (Driving\_License=1) and already have a vehicle insurance (Previously\_Insured=1), and almost all of them have damaged their vehicle in the past (Vehicle\_Damaged=Yes).These 3 columns have high predictive power.



Multicollinearity does not seem to be an issue as no columns have high correlation amongst each other

There are many young customers that have driving license who do not want to vehicle insurance which clearly explains why annual premium is right skewed. Vehicle Age and Vintage are on left skewed thus showcasing that customers are loyal and that the vehicle insurance firm provider has started offering insurance whose target audience is usually young

* **Predictions:** We plan to use various machine learning algorithms to predict whether a current policyholder would be interest in purchasing Vehicle insurance from the company using information demographics, vehicle ownership and damage, and policy. This can aid the business to optimize its business model and revenue and plan an effective communication strategy to reach out to these customers tailored to those particular customers.
* **Inference:**

Which variables in the dataset play a critical role in predicting the target variable. Specifically, what characteristic of the customer makes him more likely to buy Vehicle insurance. This adds value to the project as it highlights the characteristics of the customer which narrows down the search strategy for the insurance company.

Classify the customers based on machine learning algorithms which concretely provides a model for the insurance company to classify the customers.

From the univariate analysis we plan to gain information regarding the distributions of the individual features which might lead to feature engineering for specific algorithms, which in turn will lead to a better model for the company. From the univariate analysis we can also infer the relevance of each feature to target variable. For example, we can see if male/female tend more to subscribe for vehicle insurance etc.

From multivariate variate analysis we can plan to gain information about the correlation between the variables which is vital for dimensionality reduction, feature engineering etc. to build better models.

* **Non-Spark Packages:**

**For Modelling we would use**

* xgboost
* sklearn
* RandomForestRegressor
* GBTRegressor
* RegressionEvaluator
* VectorAssembler
* StandardScaler
* MinMaxScaler
* Vector