**Observations & Thought Process**

**Exploring and Cleaning the Data:**

1. **Initial observations from basic statistical analysis of the data set :**
2. **Renewals :** Somewhere between 25-50% people do not renew
3. **Previous Renewals** : Majority of the customers have renewed only once
   * Business Insight - either most of the customers are new OR people are dropping out after year 1 which is where maximum focus should be to retain customers
4. **Main Income :** There is a very high variation in this variable
   * Business insight - shows reliance on few customers for income
5. **Finance Income :** The mean of Financing Income is about 60% of mean of Main Income - which tells us that policies paid in installment are a lucrative source of income
   * Business insight - installment plans could be encouraged
6. **Addon Income** : The mean of Addon Income is much lower than Financing Income
   * Business insight - assess if there is proportional value from investing into Addon products
7. **Gross Price vs. Last year gross price :** There is about 10% increase in the mean price offered to the customers
   * Business insight - i. check if the increase is higher than average increase in wages and inflation for the economy, and if so what are the reasons behind the sharp increase (is that due to riskier customer profile) ii. should see if this is a driver for renewal/ IPO
8. **Missing Value Treatment: Two categorical variables had missing values and they were dealt with as below:**
9. **Vehicle\_Liability\_Category** - As this is a categorical variable, we can impute it using the median of the variable. But as over 40% of data points are missing, we will replace them in same ratio of categories as in available data.
10. **Number\_Of\_Attempted\_Policy\_Ammendments -** <30% data points available, while remaining are "Null". This variable represents the number of amendments (e.g. additional drivers etc.) done to policy in the last year. Intuitively it seems unlikely that many people would do amendments within a year. Also we notice that there is no category for "0 (zero)" amendments. Based on both these points, i assume "Null" actually means "0(zero)", and is just represented as such in the data source. Hence , i will replace all "Null" values with "0"
11. **Outlier Treatment:**
12. Outlier was defined as anything that was in either lowest or highest quantile. These values were removed for numerical variables.

**Problem Statement:**

Which segments of customers perform better than others and why?

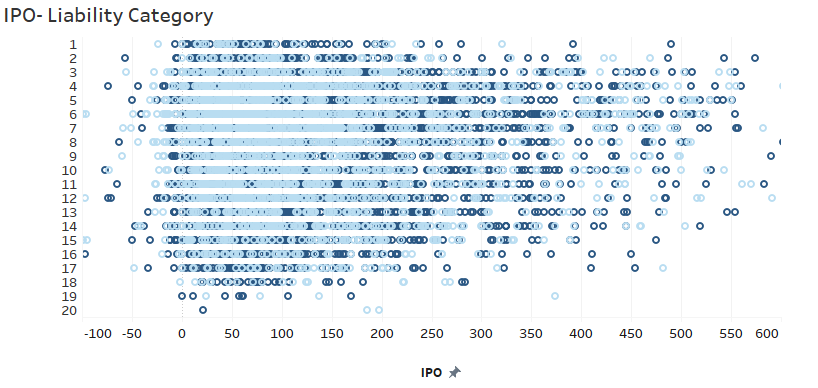
- In this case performance is defined as the expected income per offer (IPO).

- IPO= Sale \*(Main\_Income + Addon\_Income + Finance\_Income)

1. **New fields created:**
2. Calculated field for IPO as per above
3. **Total\_Income** – IPO as a calculation will be hugely dominated by Sale/ NO Sale. Although from a business point of view – it could be looked as the problem of achieving a sale vs. problem of generating higher revenue / per deal closed.

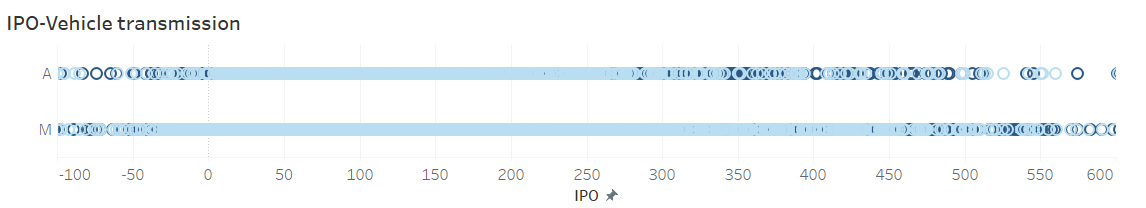
I just wanted to explore if there was any difference in variables driving Sale and Revenue – so I created this additional field.

1. **Price Change** – In addition to the price offered this year, I also wanted to see if the change from last year had any impact
2. **Feature Selection :**
3. **Numerical variables:** I checked for any correlation between the independent variables and dropped all variables which were correlated more than 85%
4. **Categorical variables :** I drew Bivariate scatter plots between categorical independent variables and IPO in Tableau, based on which I took the below actions:
   * **Vehicle Liability Category** – Visual inspection showed that IPO is spread across in value and density similarly across some classes. Based on this observation, these classes were clubbed together



* + **3 categorical variables** are being dropped as the spread in values and density are same across classes , hence these variables do not add any information for the model - Vehicle\_Fuel\_Type', 'Vehicle\_Transmission' & 'Data\_Quality

Eg below:



1. **One Hot Coding :** Majority of the features in the data set are Categorical – so I created dummy variables from those using the one hot coding method
2. **Feature Reduction:** I decided to use K means clustering for segmentation of the customers. I built the model with selected features but due to high dimensionality, it gave a very poor performance. So I took the below steps before building the model again (and again ☺)
3. **Linear Regression:** As I wanted to maintain the interpretability of model, I first used Linear Regression to reduce the number of features**. I tried different variations in Linear regression model:** 
   * With IPO as dependent variable :
     1. WE get an acceptable R2 of 73% but that’s when we include Sale\_1 as an independant feature
     2. If Sale\_1 is excluded as a dependent variable, the R2 of model falls to upper 30s%
     3. This is expected as Sale/ No Sale (1/0) has a dominant impact on IPO
   * AS IPO is derived based on two main features - Sale and Total Income , and as Linear Regression is better suited to predict continuous variables - I decided to use this model to select features that impact Total Income.
   * The RMSE of test and train data set were similar which means there is no overfitting
   * Durbin-Watson value showed there is low/ no auto-correlation in the model
   * As the model quality was good, we dropped any features whose coefficient were lesser than 3 and kept the remaining

I built the K means clustering model again based on reduced features from Linear regression – I still couldn’t get a good performing model.

So I had to use PCA to reduce features.

1. **PCA:**  I standardized the numerical features before building components using PCA.
   * Even though we have standardized the numerical variables and one hot coded the categorical ones – I thought that clubbing both of them to create components would make the numerical values dominate them.
   * I tried to build components individually for Numerical/ Categorical hot coded variables – but that meant building almost 14-15 components to keep ~80% explained variance. This would not have solved the problem of reducing variables.
   * So I finally built components on combined numerical and categorical data – about 6 components gave 78% explained variation
2. **Final K Means for Segmentation: The components from PCA were given as inputs to the K means model and we see that 4 clusters is able to show some distinguishing characters for each.**

**Cluster 1:**

* IPO (Standardized value) : 113
* Almost equal number of Married and Single
* Mostly license type 4
* Lowest Claim discount

**Cluster 2:**

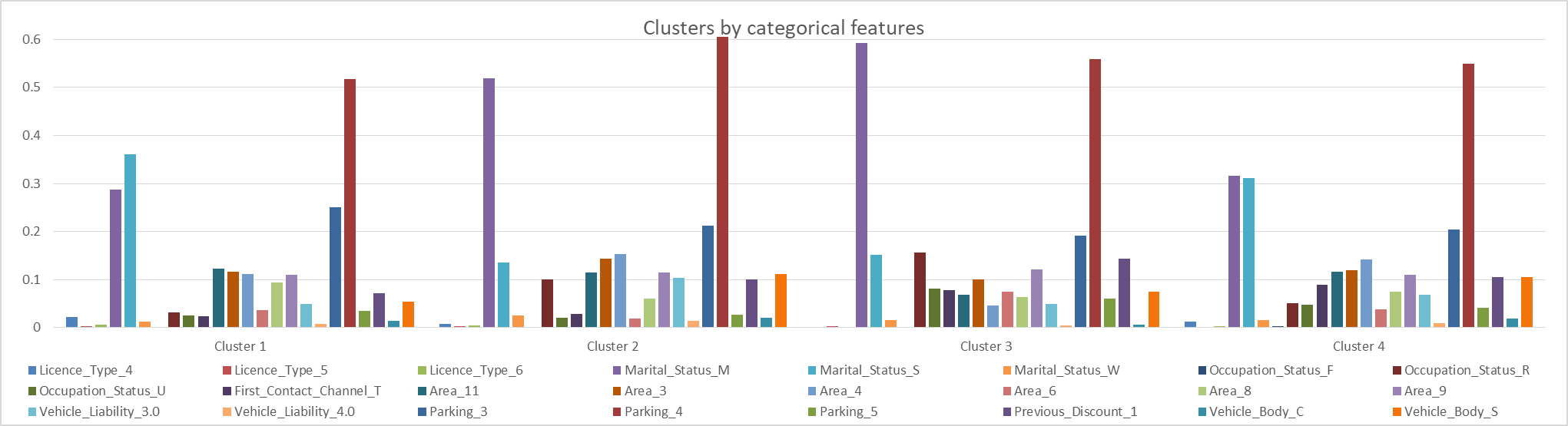
* Lowest IPO (please note though that this is standardized value of IPO): 56
* Highest vehicle liability type 2
* Very high Married than Single
* Highest in Geog area 4 and least in area 6

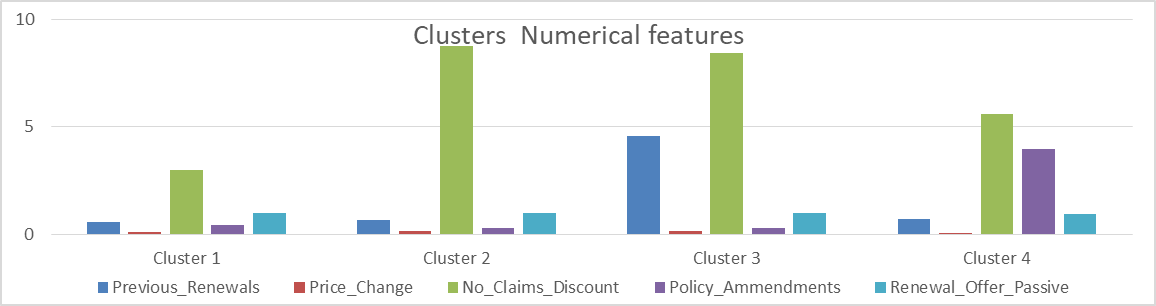
**Cluster 3:**

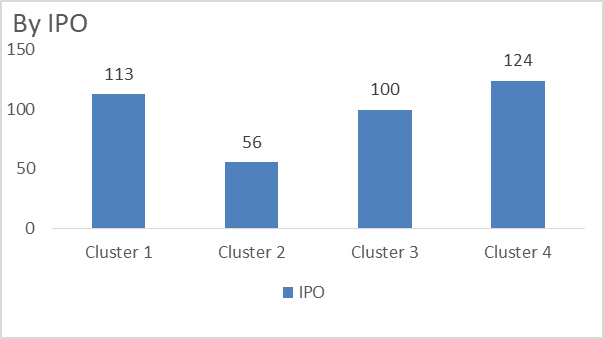
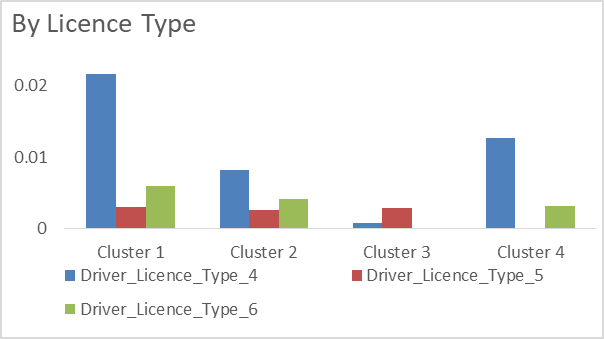
* Highest occupation R
* Highest previous renewals

**Cluster 4:**

* Highest IPO 124 (do note though that this is standardized value of IPO)
* Highest policy amendments
* Equal Single and Married , although lower than Cluster 1
* No one from license type 5
* Lowest price change since last year

****

****

** ** 