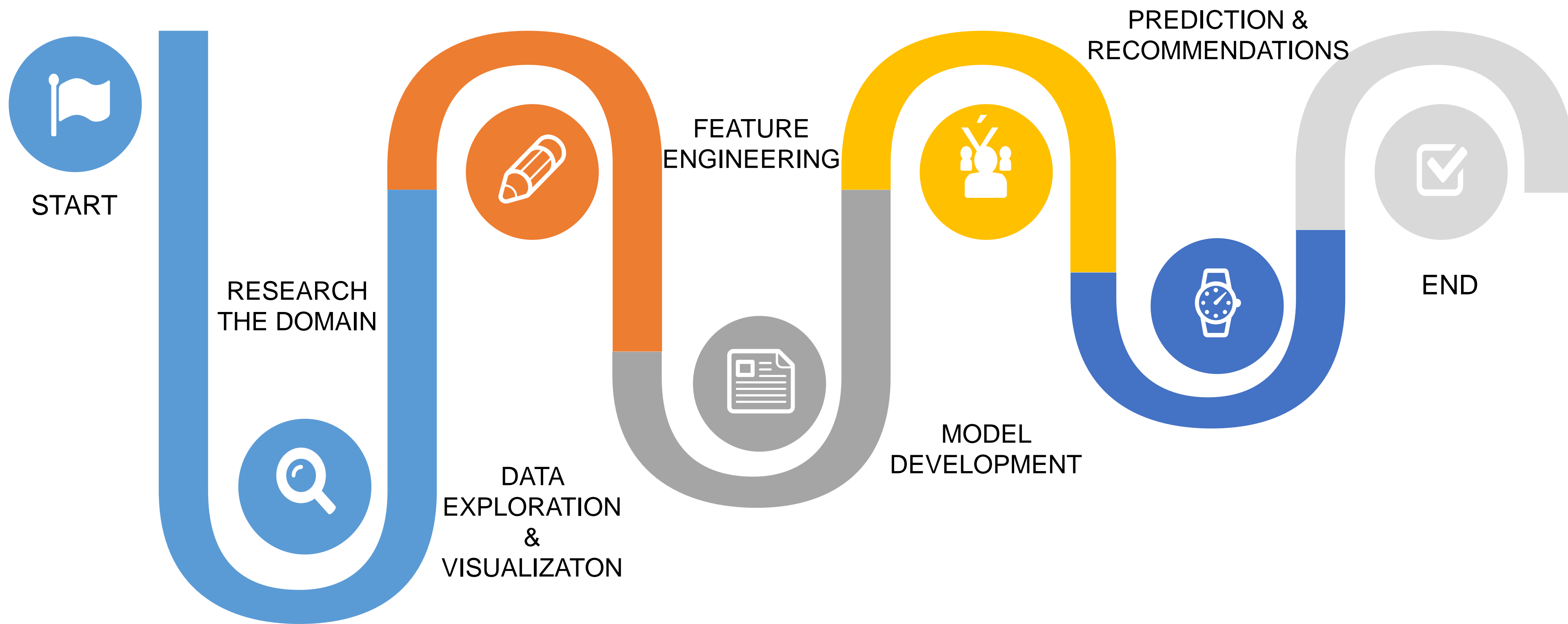


AUTO INSURANCE

USA



ROAD MAP



INTRODUCTION

✓ Auto Insurance in US

Motor Insurance's primary use is to provide financial protection against physical damage or bodily injury resulting from traffic collisions and against liability that could also arise from incidents in a vehicle.

✓ Duration of Policy

Policies are generally issued for six-month or one-year timeframes and are renewable.

✓ Policy Features

Basic policies in US mandatorily cover Property damage, Liability damage and Medical charges of the customers.

✓ Types of Policy

Most insurance coverage are of the following types:

- Collision reimbursement
- Comprehensive
- Glass Coverage





UNDERSTANDING THE DATASET

- Each unique customer id contain demographic data of customers who's insurance is expiring in the month of Jan & Feb 2011.
- The dataset contains 9134 records and 24 features including the target feature.
- There are 16 categorical & 6 numerical features.

Feature	Explanation
Customer	Unique Customer ID
State	US Province to where the customer belongs to
Response	Refers to whether customers have responded to marketing calls or not
Coverage	Nature of Insurance coverage
Education	Education level of customer
Effective To Date	Expiry date of policy
Gender	Gender of the customer
Employment Status	Current Employment status of the customer
Income	Customer annual income in USD
Location Code	Type of location where customer lives
Marital Status	Marital status of the customer
Vehicle Size	Size of vehicle
Vehicle Class	Type of vehicle
Sales Channel	Channel of sales
Renew Offer Type	Offer given during renewal
Total Claim Amount	Amount claimed till date
Monthly Premium Auto	Monthly premium for auto insurance
Months Since Last Claim	No. of months before which the last claim was made
Months Since Policy Inception	No. of months before which the policy commenced
Number of Open Complaints	No. of unresolved complaints from the customer
Number of Policies	No. of policies with the current customer
Policy Type	Type of policy
Policy	Policy sub category
Customer Life Time Value	CLV of the customer for the auto insurance company

PROBLEM STATEMENT



Target

- This is a Regression problem with Customer Lifetime Value as our Target Variable.

Scope

- There is scope to understand the Auto insurance policies and predict Customer Lifetime value.
- Understand the trend and up sell policies to the customers whose policies are expiring in Jan & Feb 2011.

Limitations

- No clarity on the offer types and policy types given in the data as to what they entail.
- There is no past data to understand the trend of customers buying behaviour or competition analysis.
- Source of the data and company is unknown.
- The company's presence is only in 5 states in the West of USA – California, Nevada, Arizona, Oregon and Washington.

A close-up photograph of the front right corner of a bright blue car. The car's headlight, side mirror, and front bumper are visible. The background is a blurred green landscape with trees. A semi-transparent dark blue circle is overlaid on the left side of the image, containing white text.

OBJECTIVE:
*Understand the demographics
to predict the customer lifetime
value for up-selling the new
products*



UNIVARIATE & BIVARIATE ANALYSIS

VARIABLE IDENTIFICATION

NUMERICAL VARIABLES

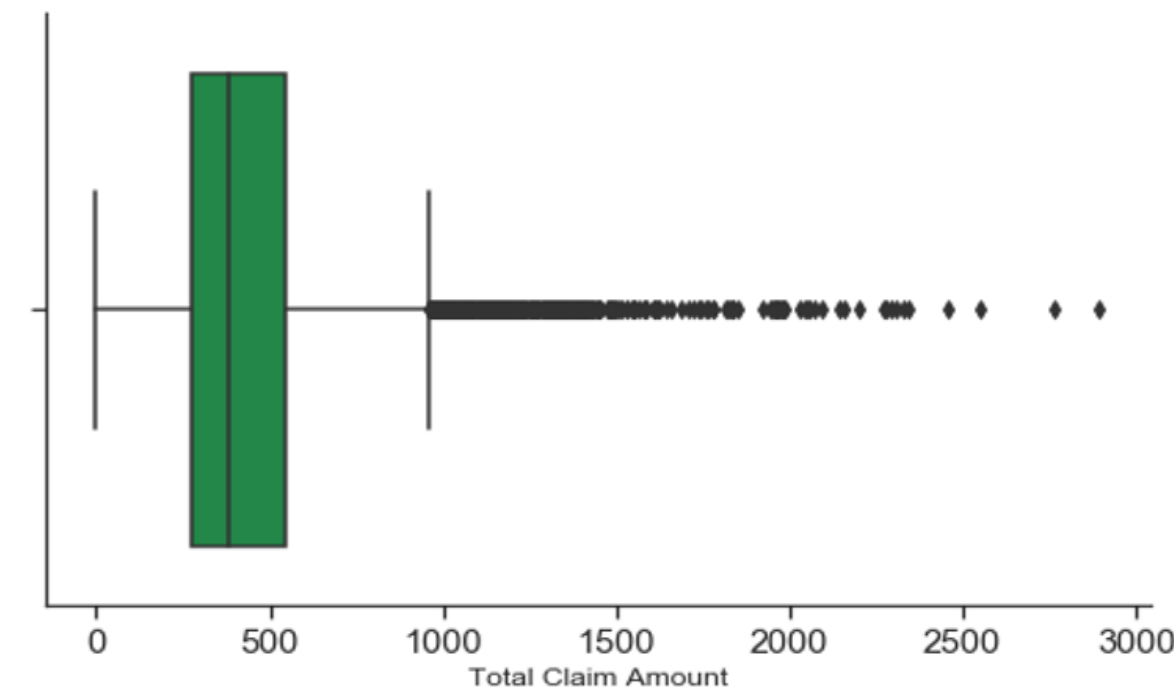
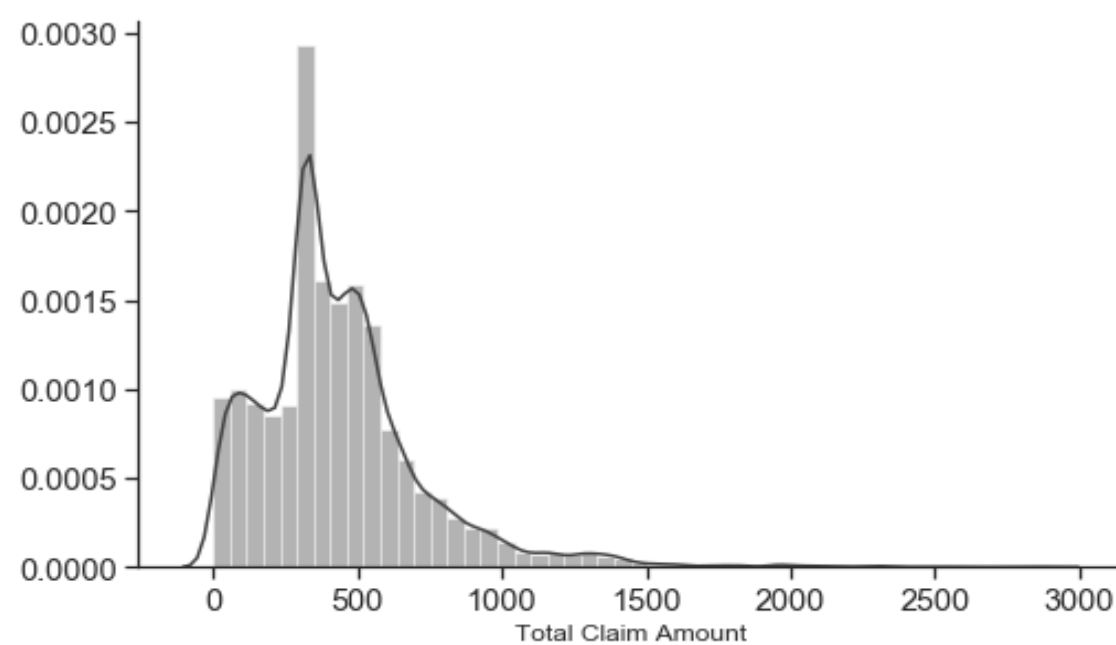
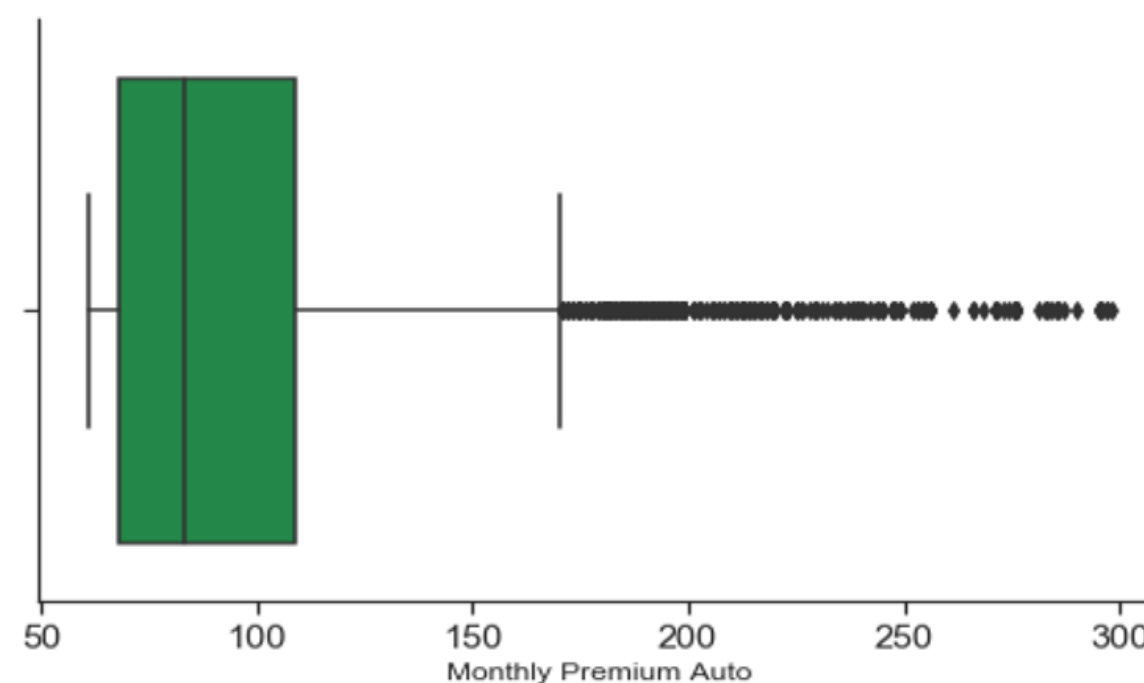
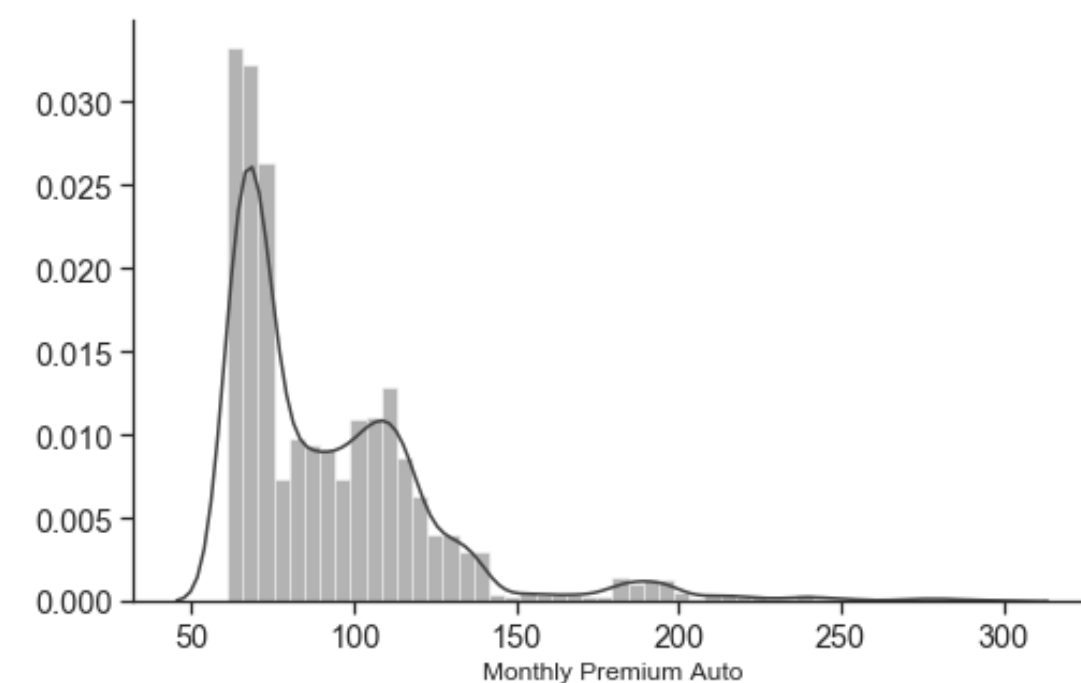
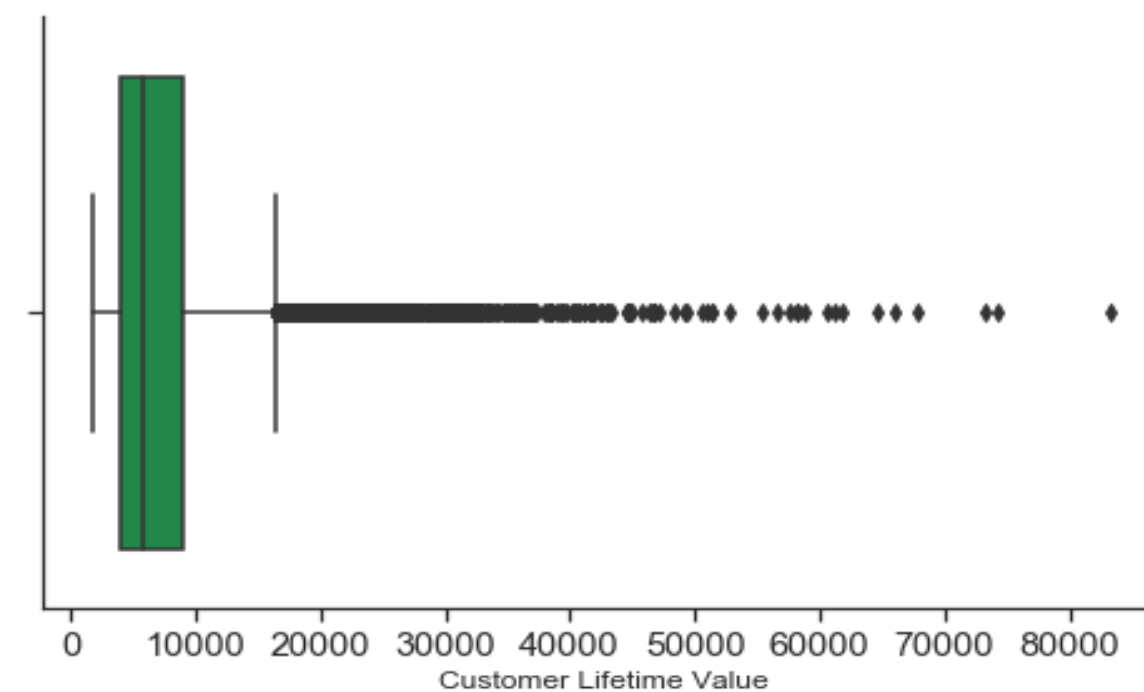
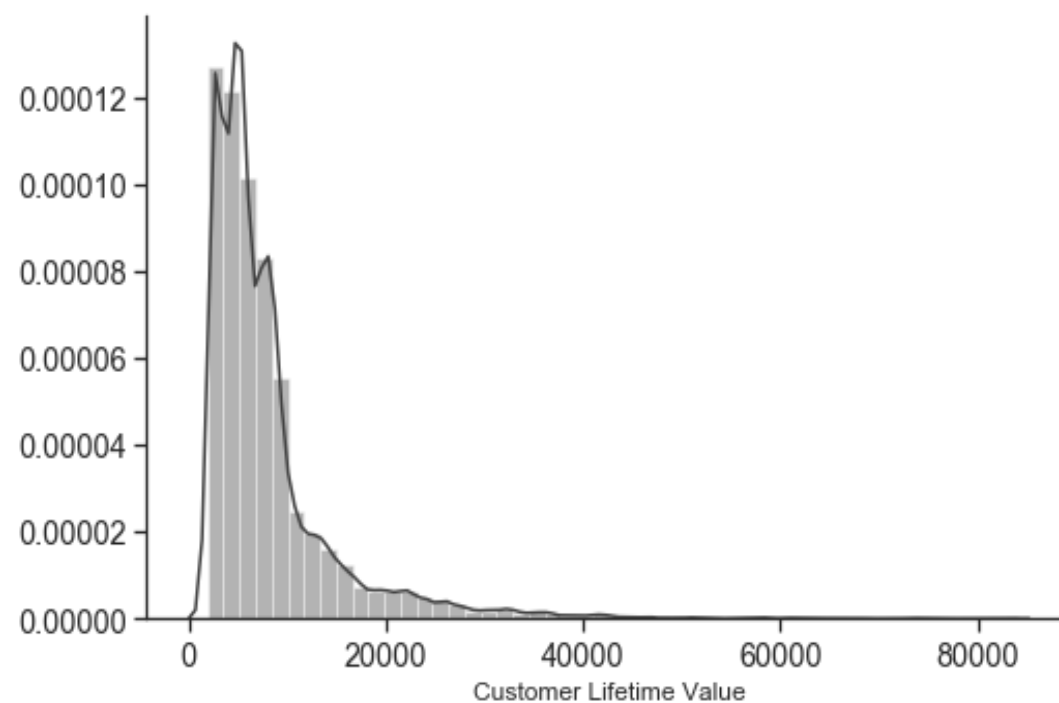
	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000
mean	8004.940475	37657.380009	93.219291	15.097000	48.064594	0.384388	2.966170	434.088794
std	6870.967608	30379.904734	34.407967	10.073257	27.905991	0.910384	2.390182	290.500092
min	1898.007675	0.000000	61.000000	0.000000	0.000000	0.000000	1.000000	0.099007
25%	3994.251794	0.000000	68.000000	6.000000	24.000000	0.000000	1.000000	272.258244
50%	5780.182197	33889.500000	83.000000	14.000000	48.000000	0.000000	2.000000	383.945434
75%	8962.167041	62320.000000	109.000000	23.000000	71.000000	0.000000	4.000000	547.514839
max	83325.381190	99981.000000	298.000000	35.000000	99.000000	5.000000	9.000000	2893.239678

- We can see the 25% (1st quartile) income being 0 which can be said that those people were the unemployed customers
- With just an overview, we can say that our target column, the CLV along with Income, Total Claim Amt, and the Monthly premium is highly right skewed.
- The max no. of Months since Policy Inception is 99.

Numerical Variables

Our target variable is Customer Lifetime Value. The data is highly skewed towards the right and a large number of extreme values can be seen.

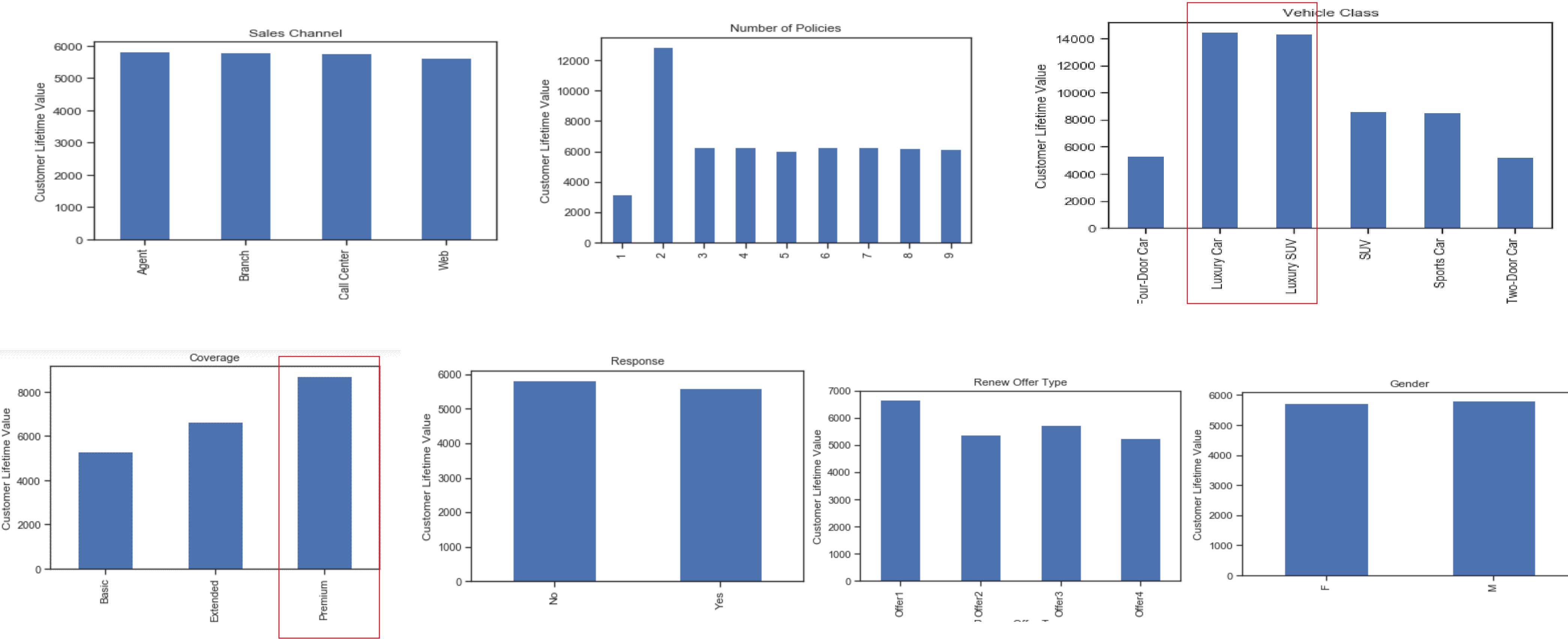
Also the Monthly Premium Auto and the Total Claim amount show similar features.



Feature	P-Value (Shapiro test)
Customer Lifetime Value	0
Income	0
Monthly Premium Auto	0
Months Since Last Claim	0
Months Since Policy Inception	7.704e-44
Total Claim Amount	0

BIVARIATE RELATIONSHIP WITH CLV

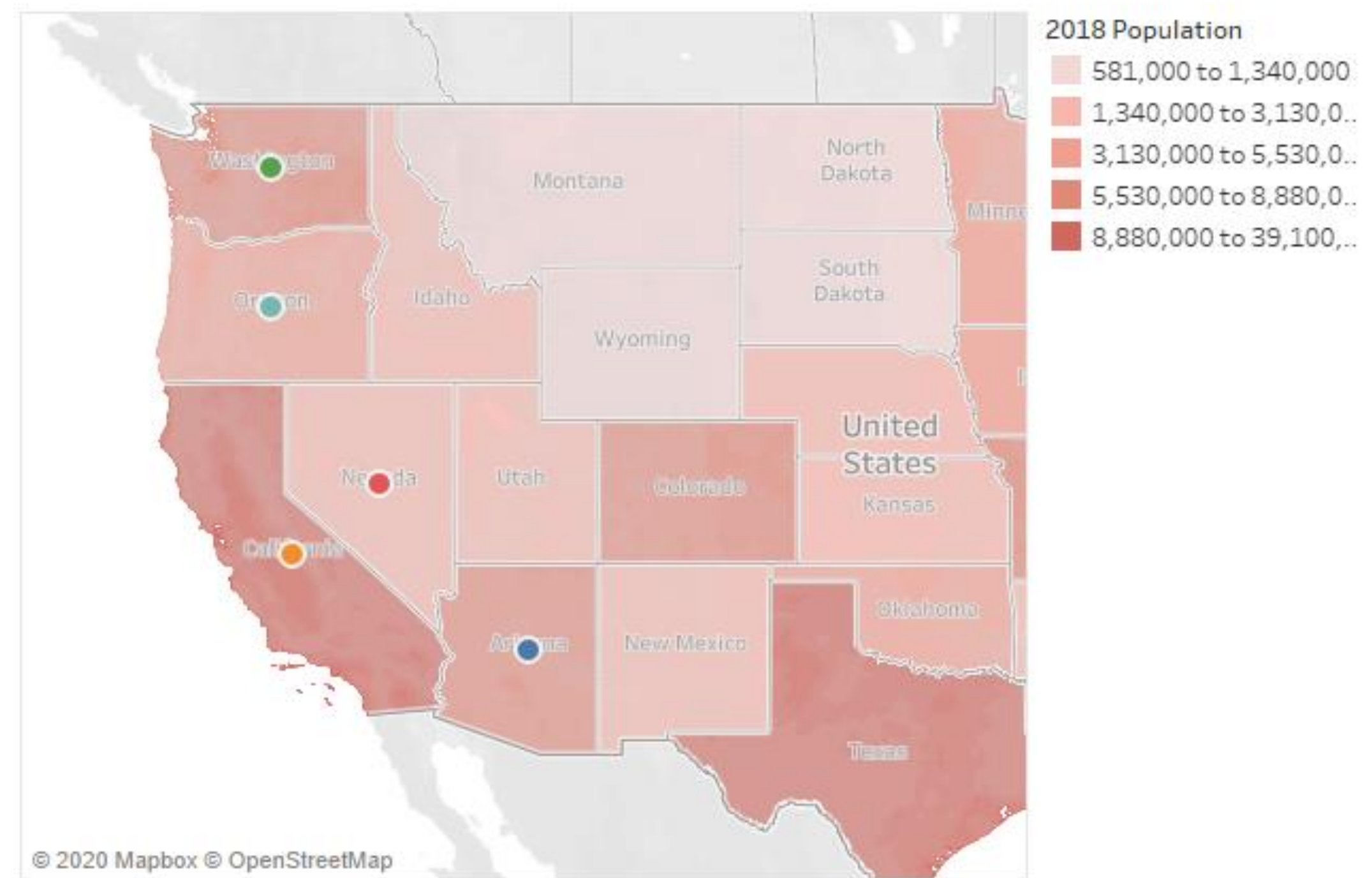
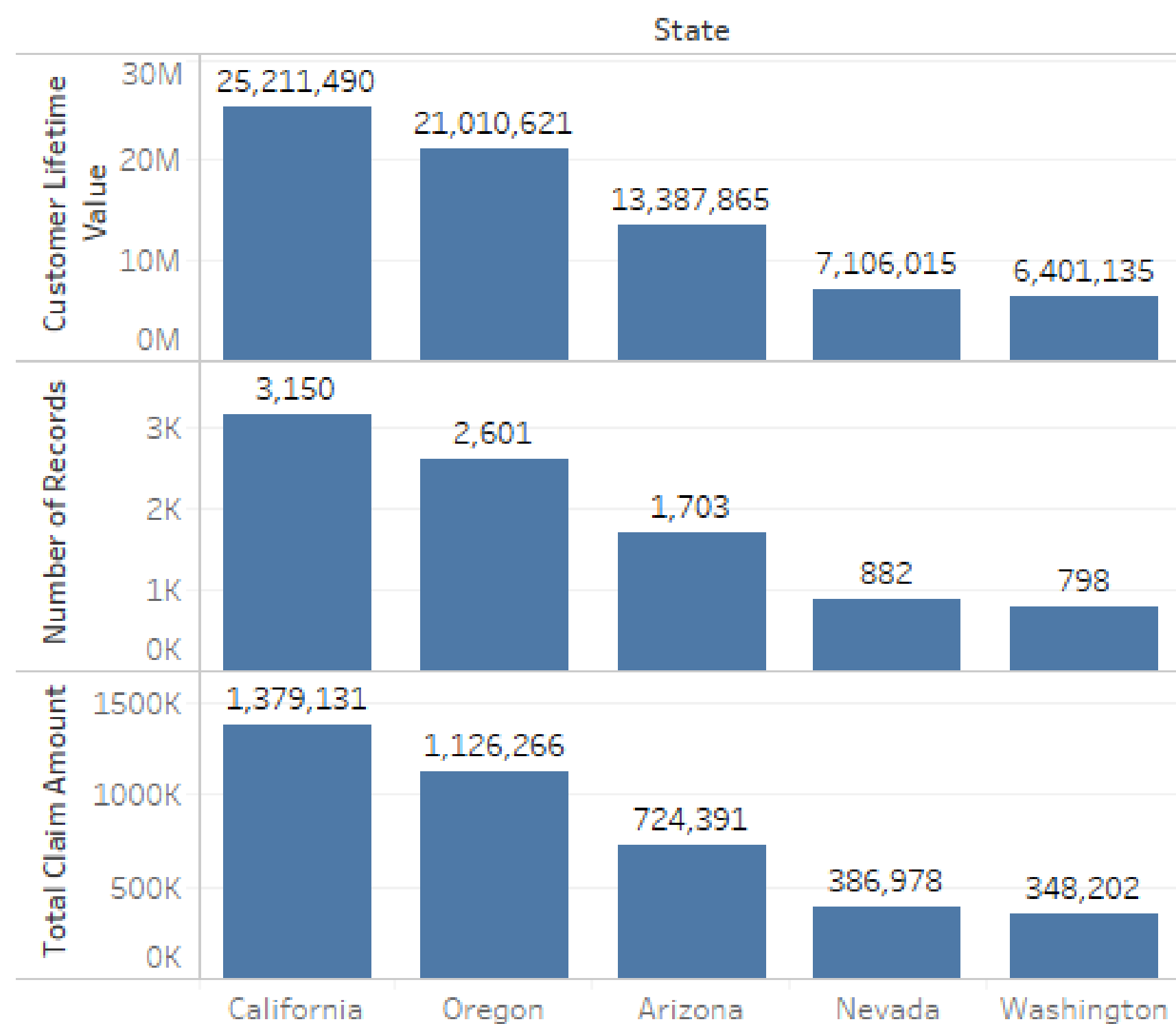
CATEGORICAL VARIABLES





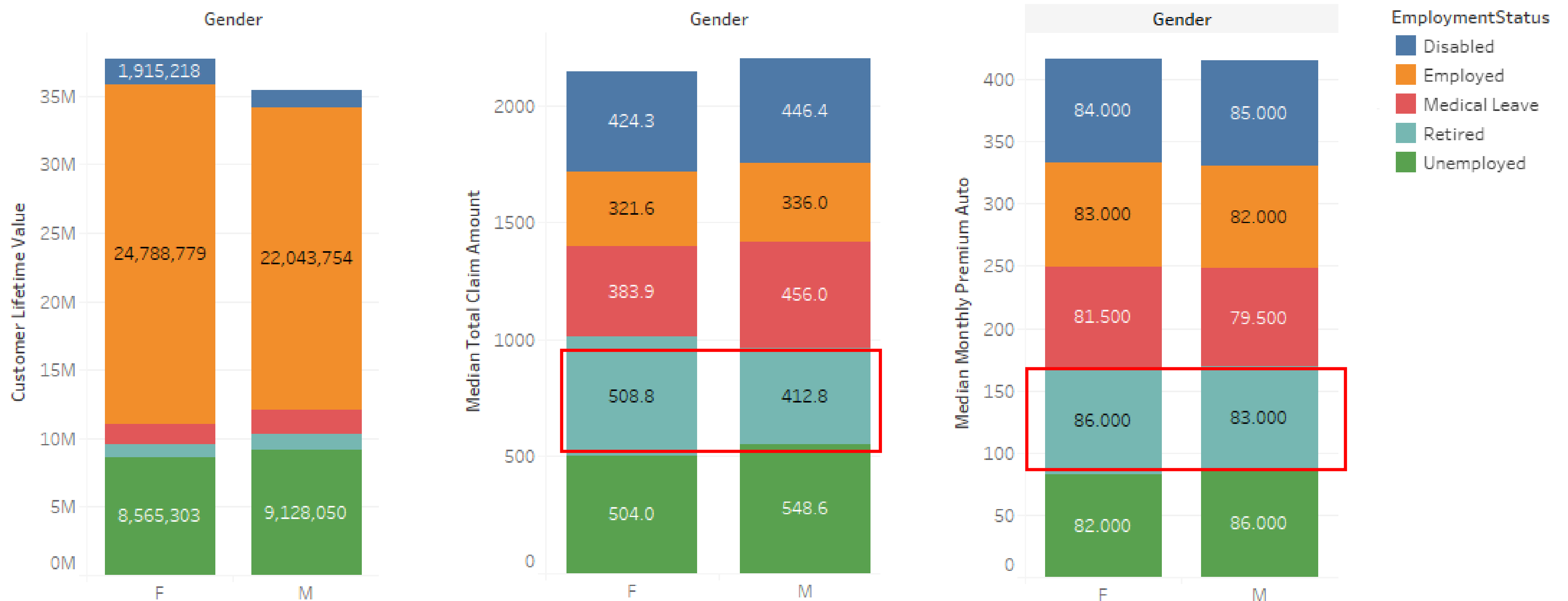
TABLEAU

STATE WISE PRESENCE



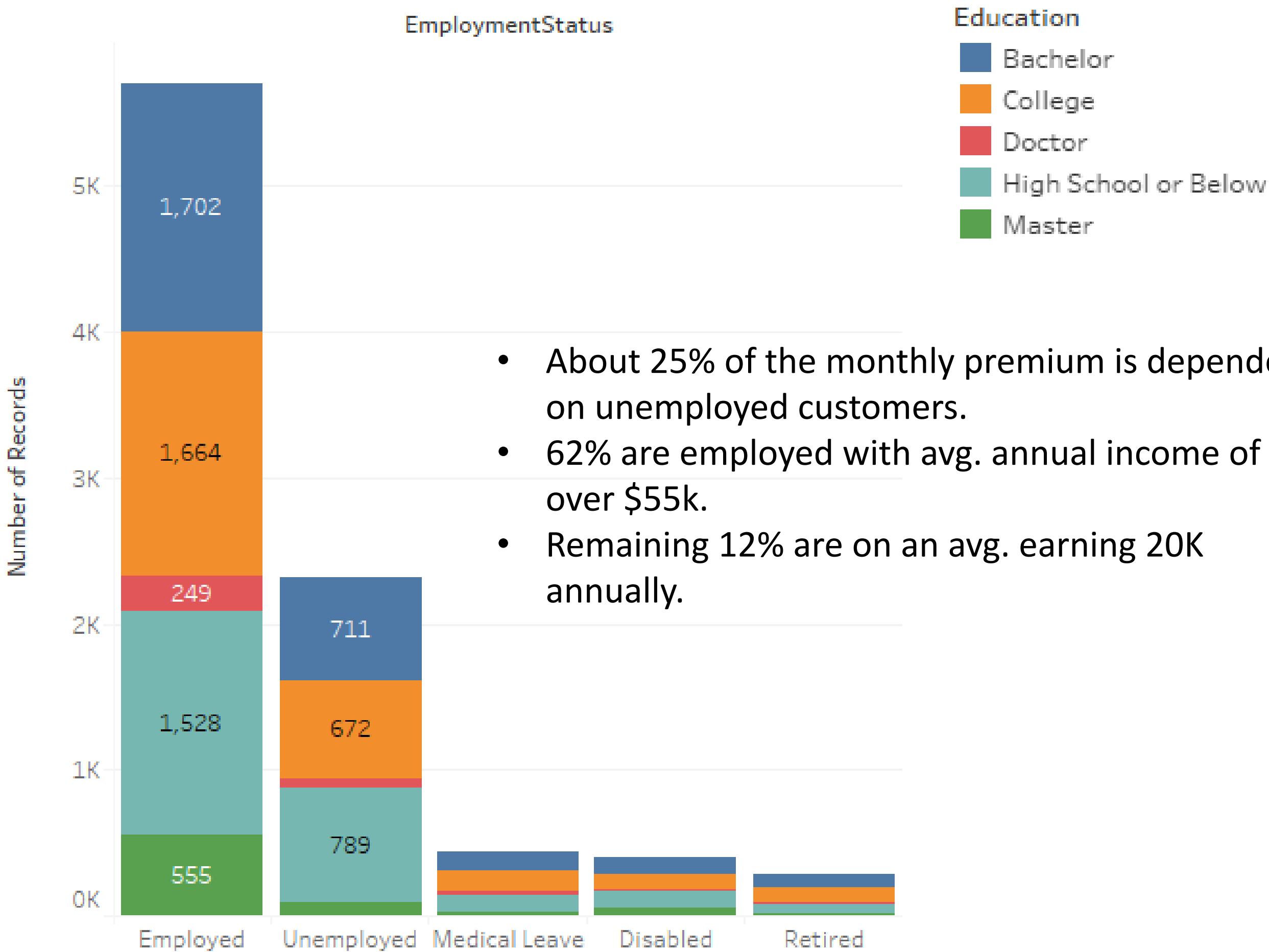
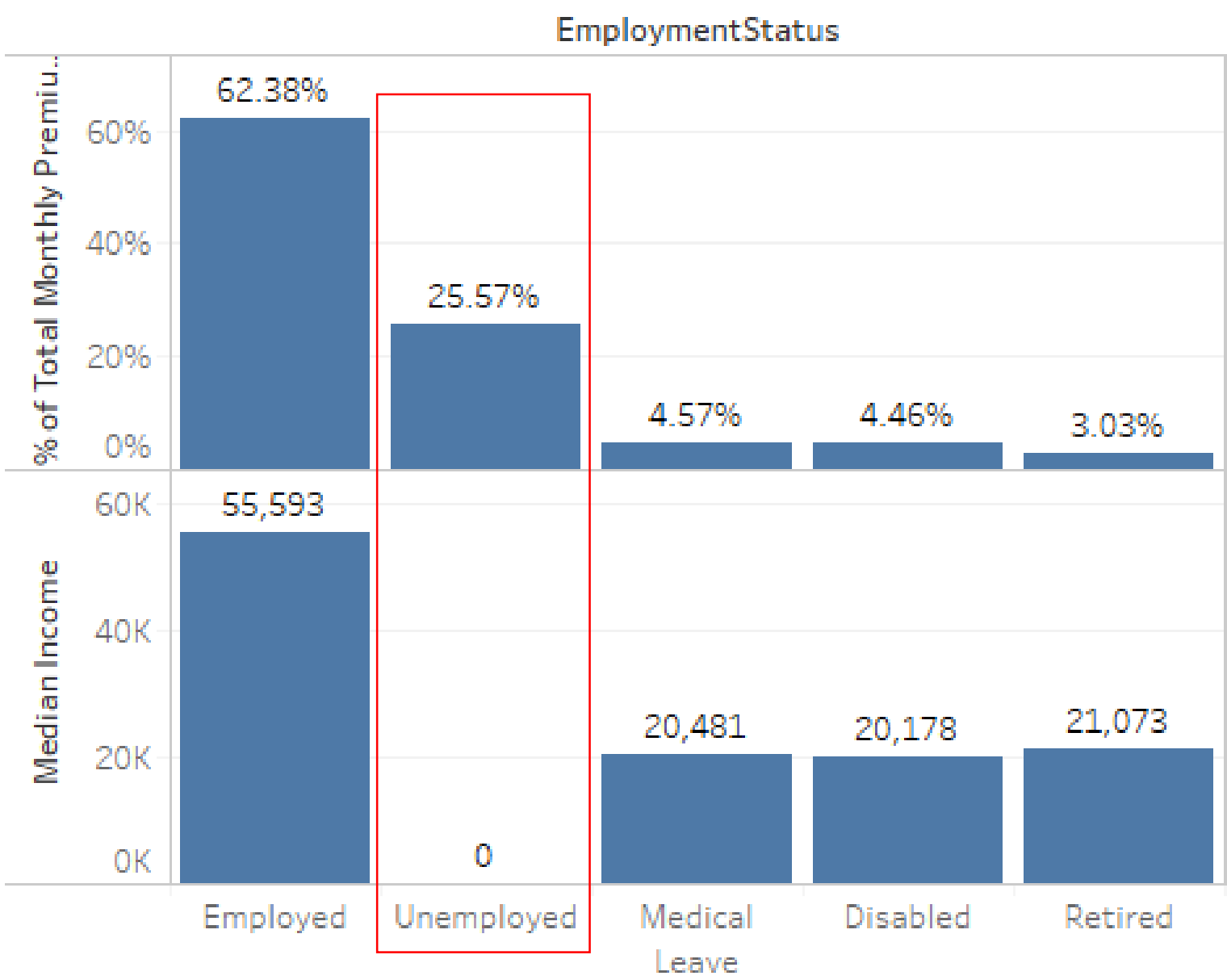
- California is the most populated amongst the 5 states, correspondingly the Auto insurance has the maximum presence here.
- Comparatively the Auto insurance is doing well in Oregon even though the population is lesser.
- Washington State is highly population as well but the penetration by the insurance company is the least.
- This forms a scope to analyze the market potential for expansion.

GENDER VARIATIONS

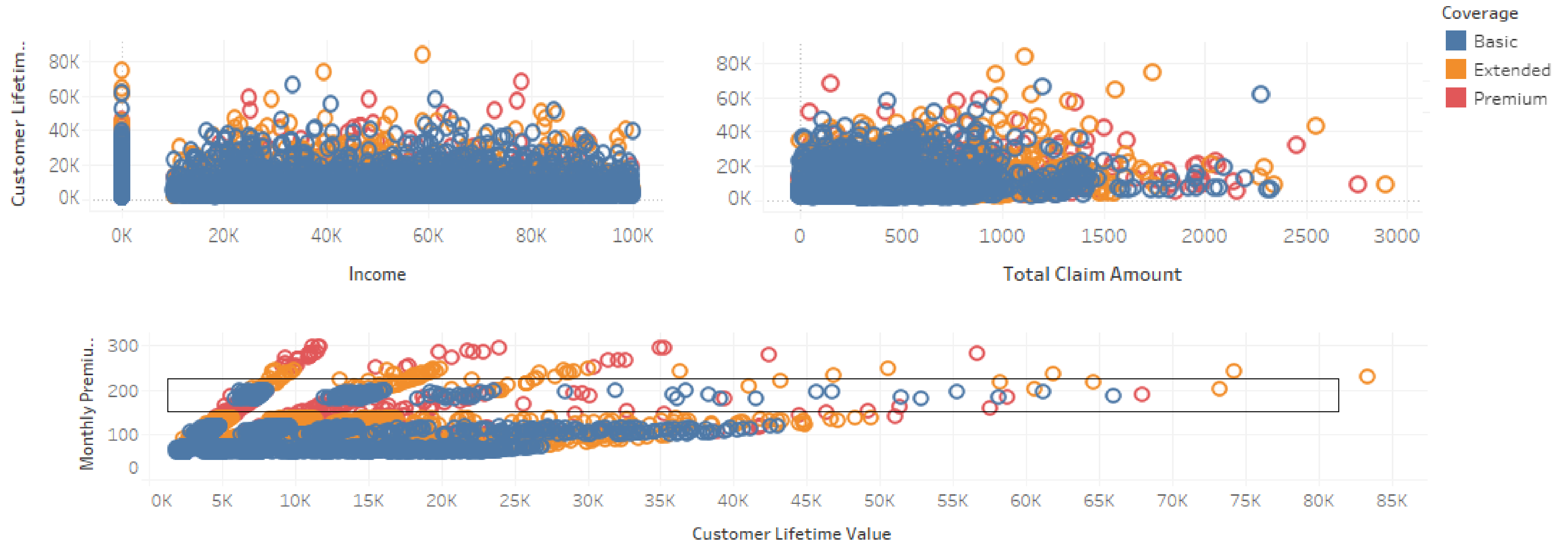


- While there is no discrepancy in genders, but Male tend to claim more amount on an average than females.
- Except for Retired females who are claiming more amount, which is also reflected in the monthly premium avg. being \$3 expensive.
- Males on Medical leave are paying monthly premium lesser than the females, however the average claim amount is more than the females by ~\$70

MONTHLY PREMIUM

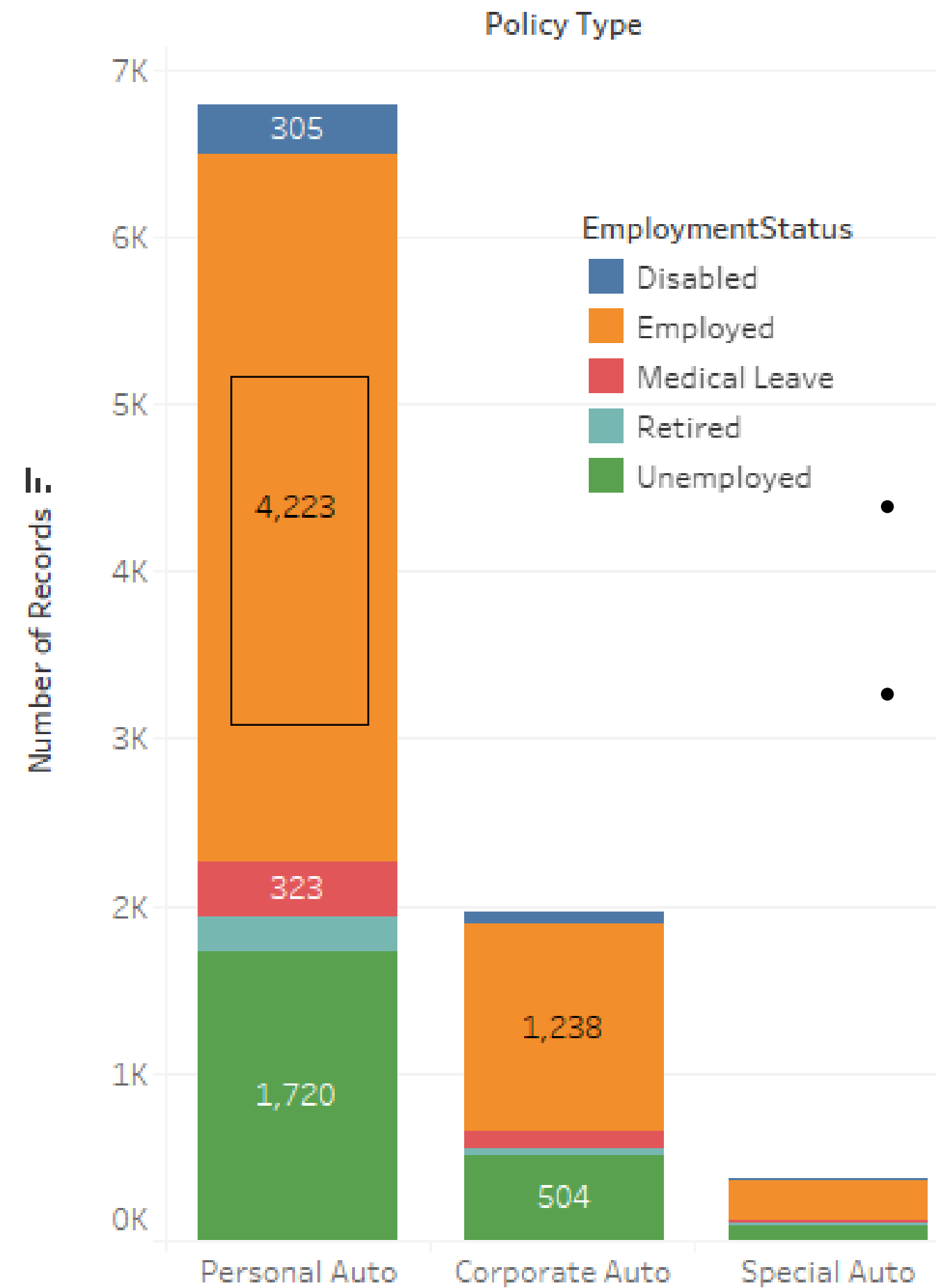
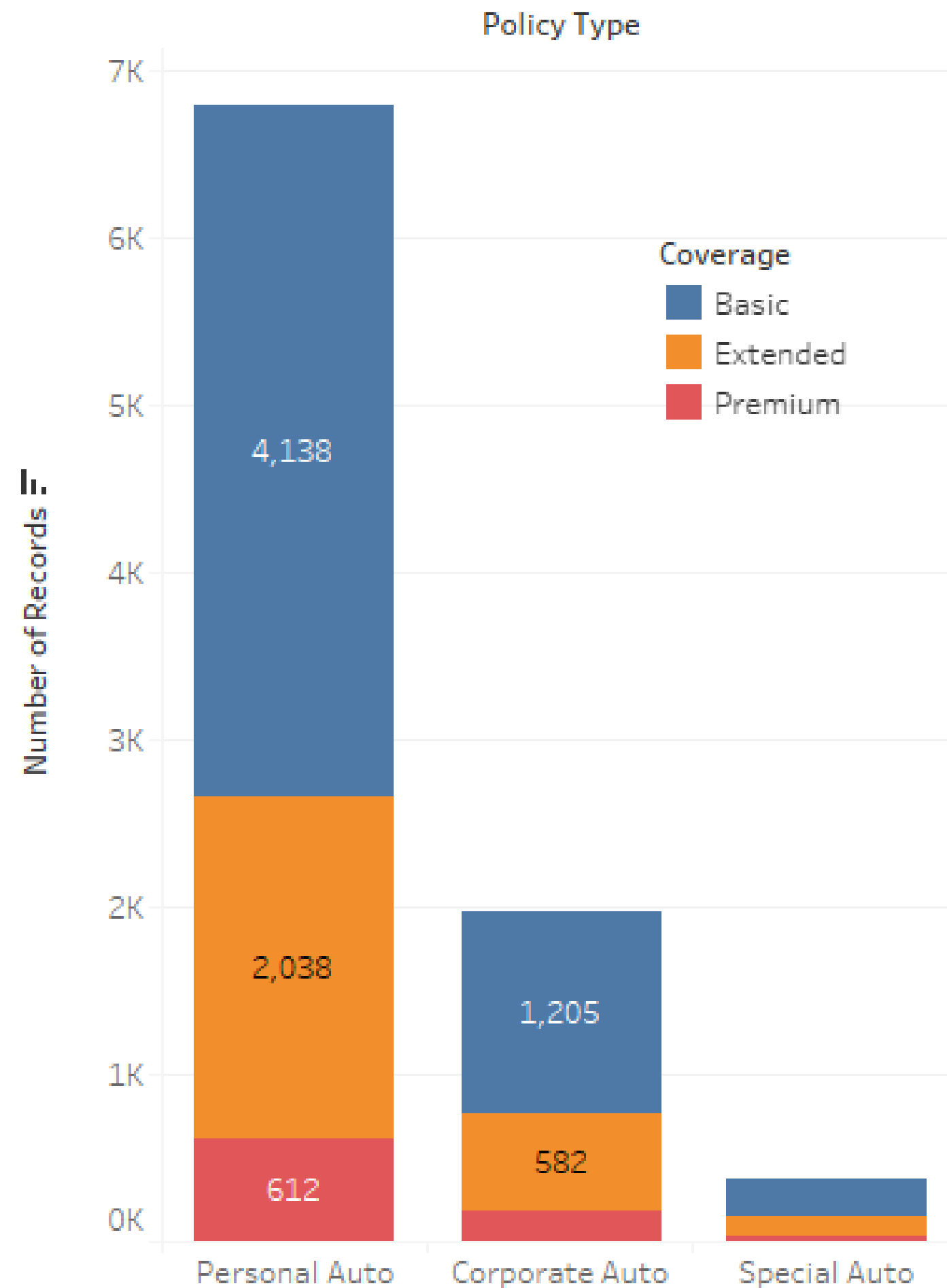


COVERAGE ANALYSIS



- There is no linear relationship between CLV and Income/ Total Claim amount / Premium.
- It is clear that there are more people under the basic coverage since its cheapest followed by extended and Premium.
- There are many customers on a basic plan but their monthly premium is at par with the extended or premium coverage. These are the potential customers to up-sell the coverage.

POLICY TYPE



- The largest bracket of employed customers have opted for a Personal Auto policy.
- Special policy is opted by only 4% customers.



FEATURE ENGINEERING

CATEGORICAL DATA

Target Variable – CLV is not normally distributed, hence we need to go for non-parametric tests.

Mann-Whitney U Test: 2 Categories

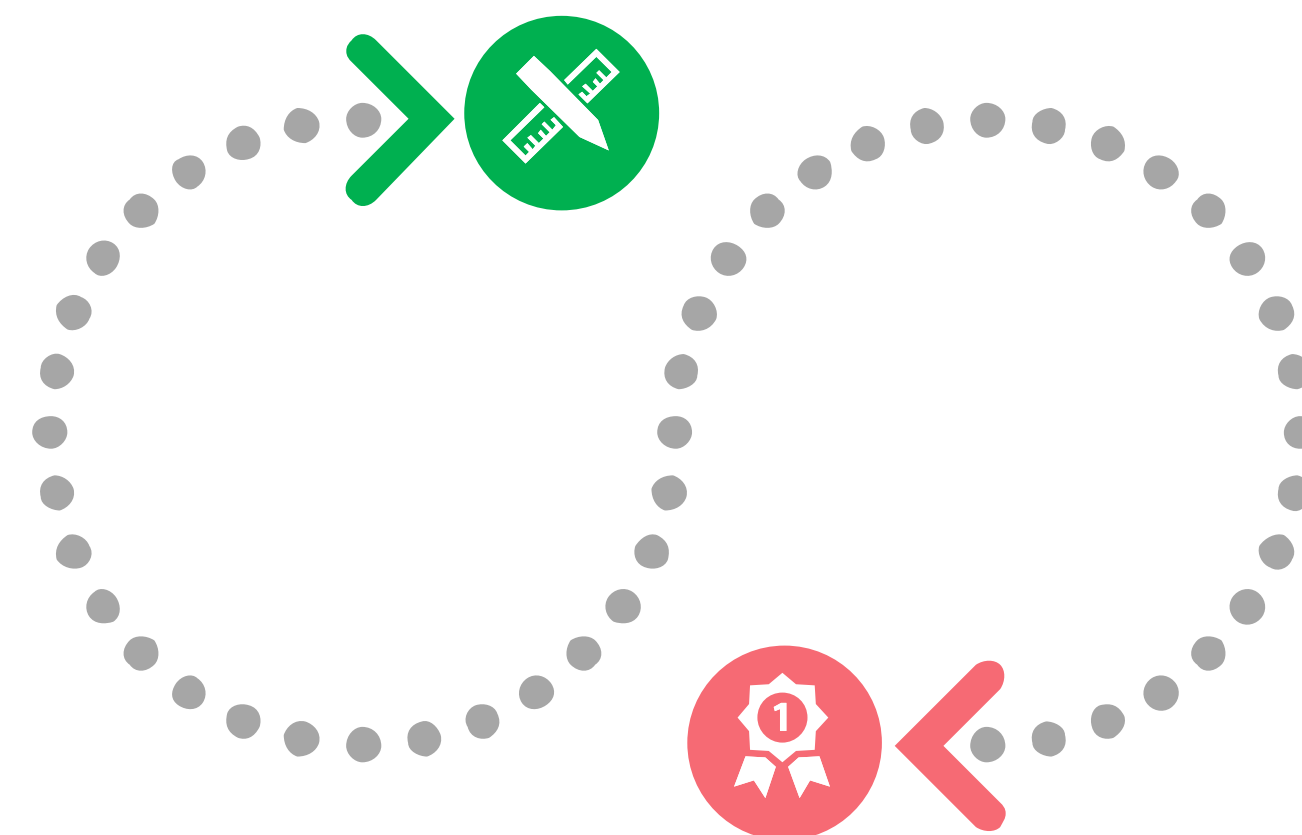
- ✓ Your dependent variable should be measured at the ordinal or continuous level.
- ✓ Your independent variable should consist of two categorical, independent groups.
- ✓ You should have independence of observations.

Features: 'Response' & 'Gender'

Kruskal-Wallis Test : 2+ Categories

- ✓ Your dependent variable should be measured at the ordinal or continuous level.
- ✓ Your independent variable should consist of two or more categorical, independent groups.
- ✓ You should have independence of observations.

Features: 'State', 'Coverage', 'Education', 'EmploymentStatus', 'Location Code', 'Marital Status', 'Policy Type', 'Policy', 'Renew Offer Type', 'Sales Channel', 'Vehicle Class', 'Vehicle Size', 'Number of Open Complaints', and 'Number of Policies'.




CATEGORICAL DATA


All categorical columns are tested against the assumptions and have been considered with a P-value benchmark of 5%

MANN-WHITNEY U TEST

KRUSKAL-WALLIS TEST

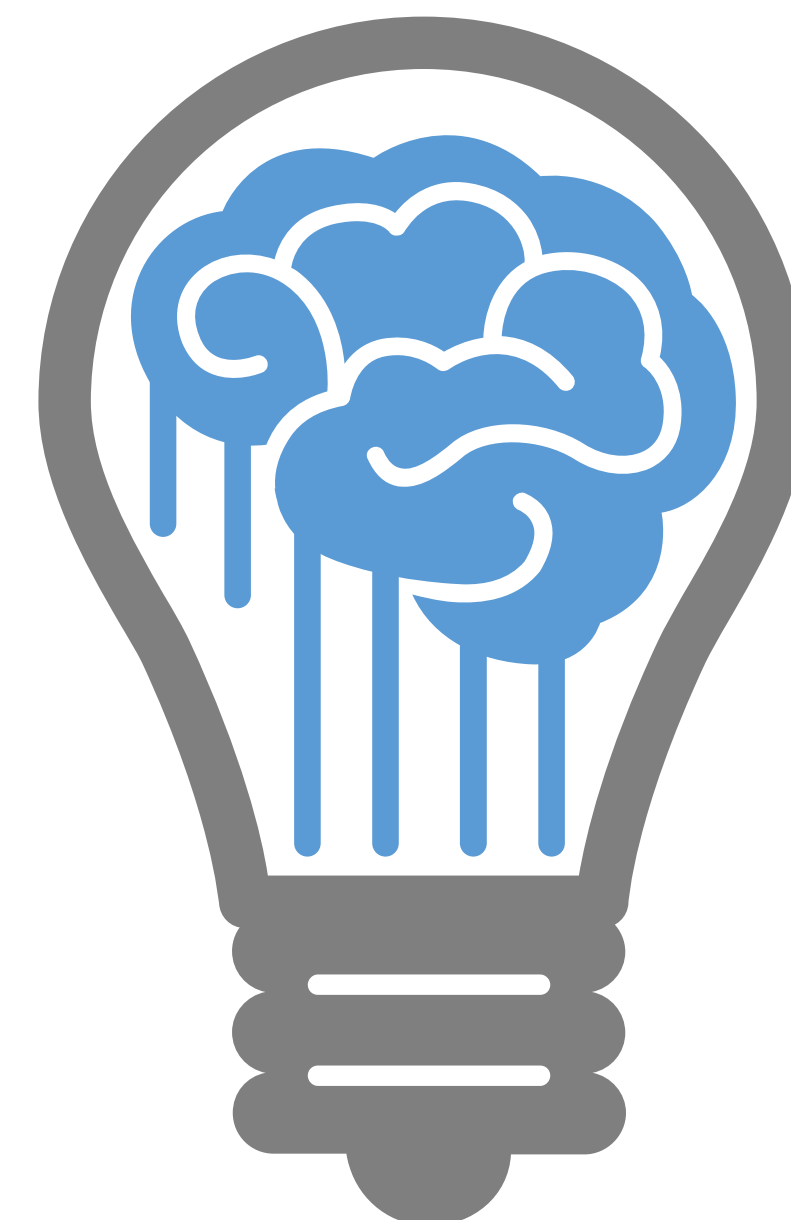
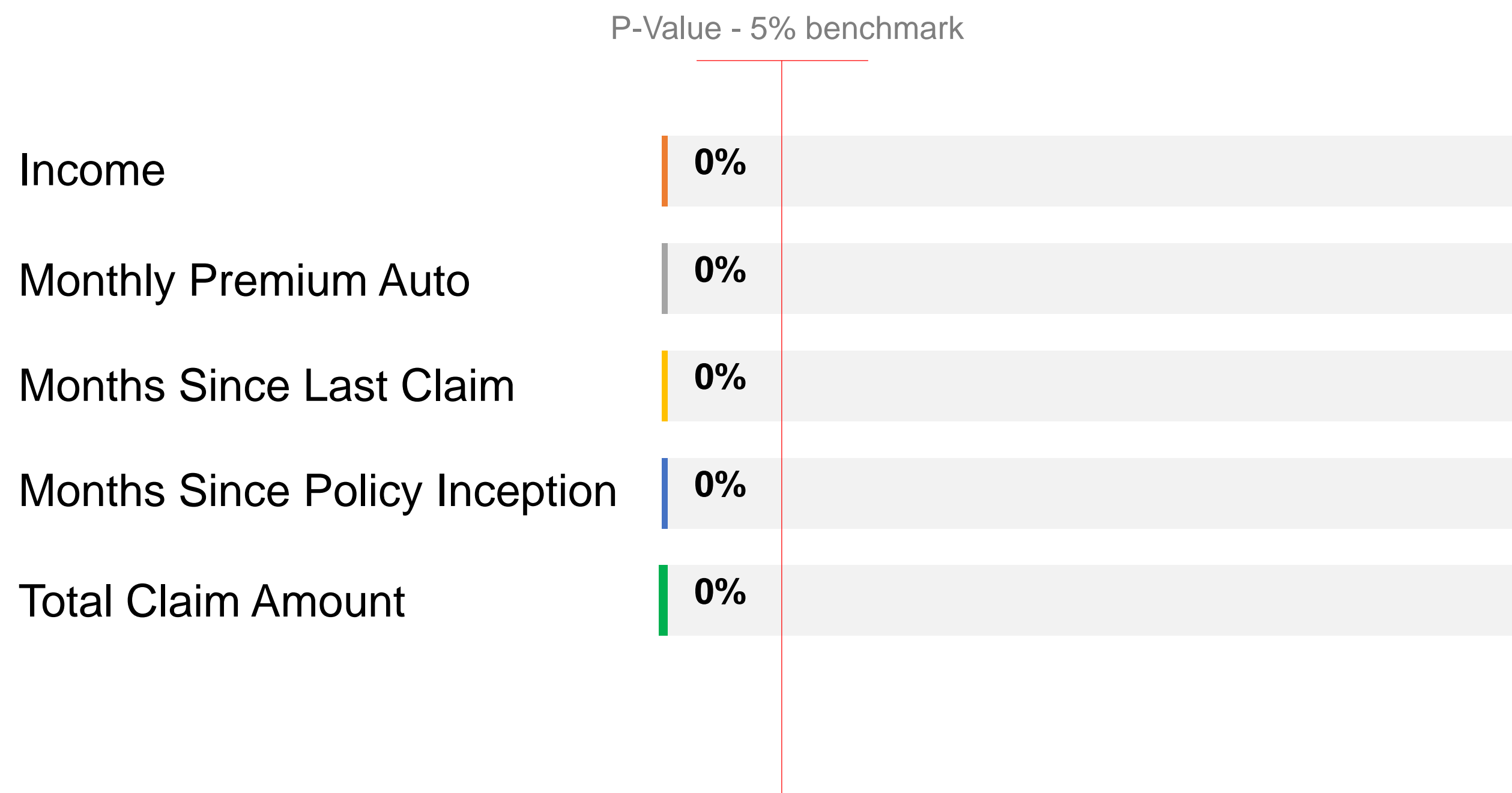
FEATURES	P-VALUE (%)
Gender	24
Response	25
Coverage	0
Education	1.5
State	27
Employment Status	0
Location Code	29
Marital Status	0
Vehicle Size	0.8
Vehicle Class	0
Sales Channel	21
Renew Offer Type	0
Number of Open Complaints	0
Number of Policies	0
Policy Type	9
Policy	43

 P-Value more than 5%

 P-Value less than 5%

NUMERICAL DATA

As we have already proved that our numerical data is non-normal, hence *Levene* test to check whether the variables have same variance.



Feature list

The categorical variables have been ordinally converted into numeric to support the Machine Learning models.
Also to treat the skewness in the data, BoxCox transformation has been applied

Education	EmploymentStatus	Income	Marital Status	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Renew Offer Type	Vehicle Class	Vehicle Size	Customer Lifetime Value_boxcox	Monthly Premium Auto_boxcox	Total Claim Amount_boxcox
2	4	56274	0	32	5	0	1	3	1	1	2.703839	0.685347	33.035752
2	1	0	1	13	42	0	8	2	0	1	2.754926	0.685871	56.357614
2	4	48767	0	18	38	0	2	3	1	1	2.780772	0.686039	40.071789
2	1	0	0	18	65	0	7	3	2	1	2.759125	0.686018	38.763252
2	4	43836	1	12	44	0	1	3	0	1	2.704995	0.685461	19.560154

Final features to be used for model building

['Coverage', 'Education', 'EmploymentStatus', 'Marital Status', 'Renew Offer Type', 'Vehicle Class', 'Vehicle Size', 'Number of Open Complaints', 'Number of Policies', 'Income', 'Monthly Premium Auto', 'Months Since Last Claim', 'Months Since Policy Inception', 'Total Claim Amount']



MODEL BUILDING



MODEL RESULTS

After applying wrapper & embedded methods on the base OLS models, the results revolve around 29.5% as the data is not linear.

BASE Model

29.4%

Backward Elimination

29.5%

Forward Selection

29.5%

Recursive Feature Elimination

29.4%

Lasso

29.5%

Ridge

29.3%

Elastic Net

16.2%



DATASET VARIATIONS

On applying the Machine Learning models on various versions of the dataset, highlighted in green are the versions that showed significant results.

BACKWARD
ELIMINATION
(8+1)

FORWARD
SELECTION
(10+1)

POLYNOMIAL 2 &
POLYNOMIAL 3

STATISTICALLY
PROVEN VARIABLES
(14+1)

LABEL ENCODER

SCALAR
TRANSFORMATION

STATS+OLS
(6+1)

ONE HOT ENCODING

BOXCOX
TRANSFORMATION

SQUARE ROOT
TRANSFORMATION

LOG
TRANSFORMATION

FEATURE SELECTION

List of variables in each dataset

Statistically proven Variables

['Coverage', 'Education', 'EmploymentStatus', 'Marital Status', 'Renew Offer Type', 'Vehicle Class', 'Vehicle Size', 'Number of Open Complaints', 'Number of Policies', 'Income', 'Monthly Premium Auto', 'Months Since Last Claim', 'Months Since Policy Inception', 'Total Claim Amount']

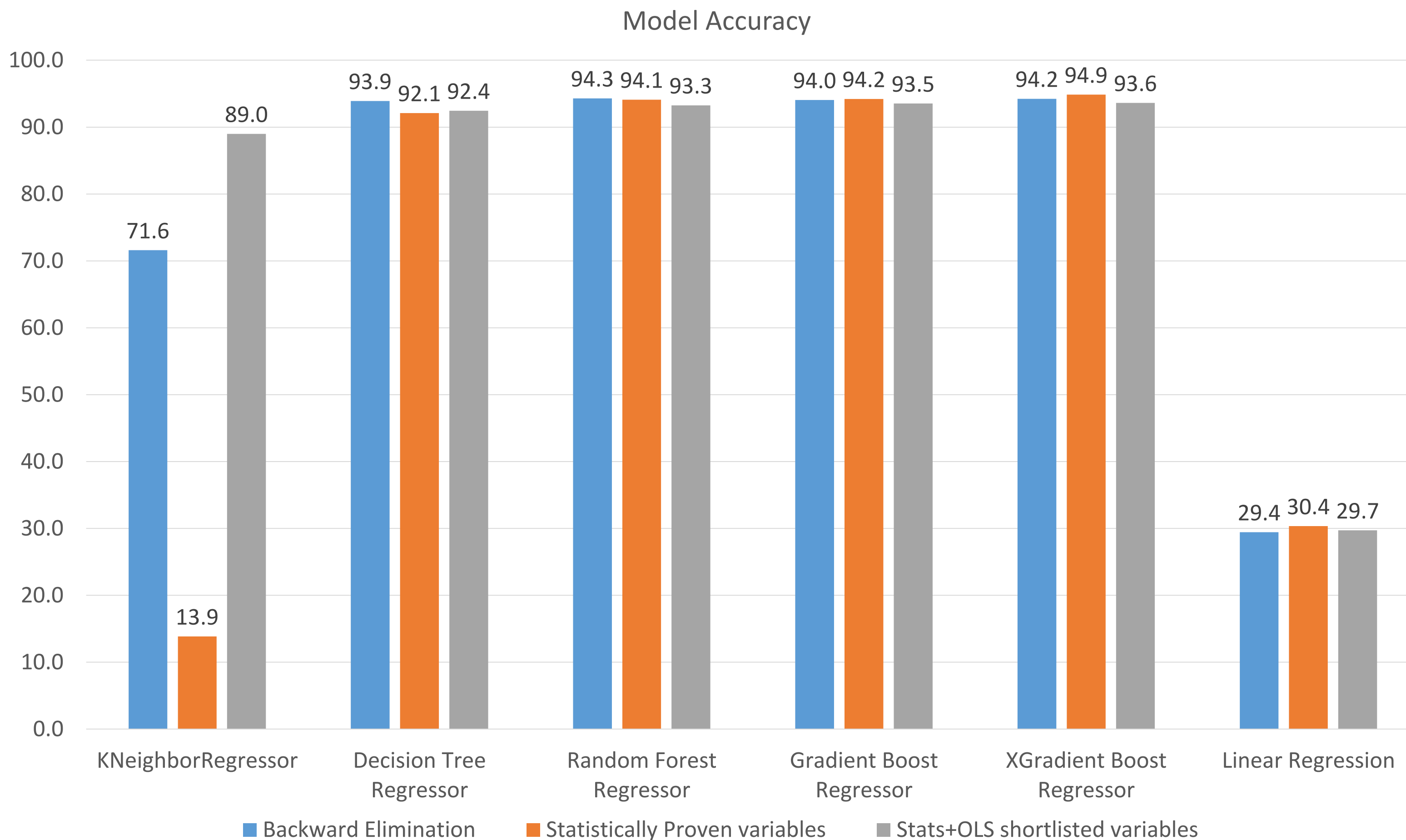
Backward Elimination and Recursive

['Response', 'EmploymentStatus', 'Number of Open Complaints', 'Number of Policies', 'Policy Type', 'Renew Offer Type', 'Vehicle Class', 'Monthly Premium Auto_boxcox']

Shortlisted variables basis Stats + OLS

[Education, 'Number of Open Complaints', 'Renew Offer Type', 'Vehicle Class', 'Monthly Premium Auto_boxcox']

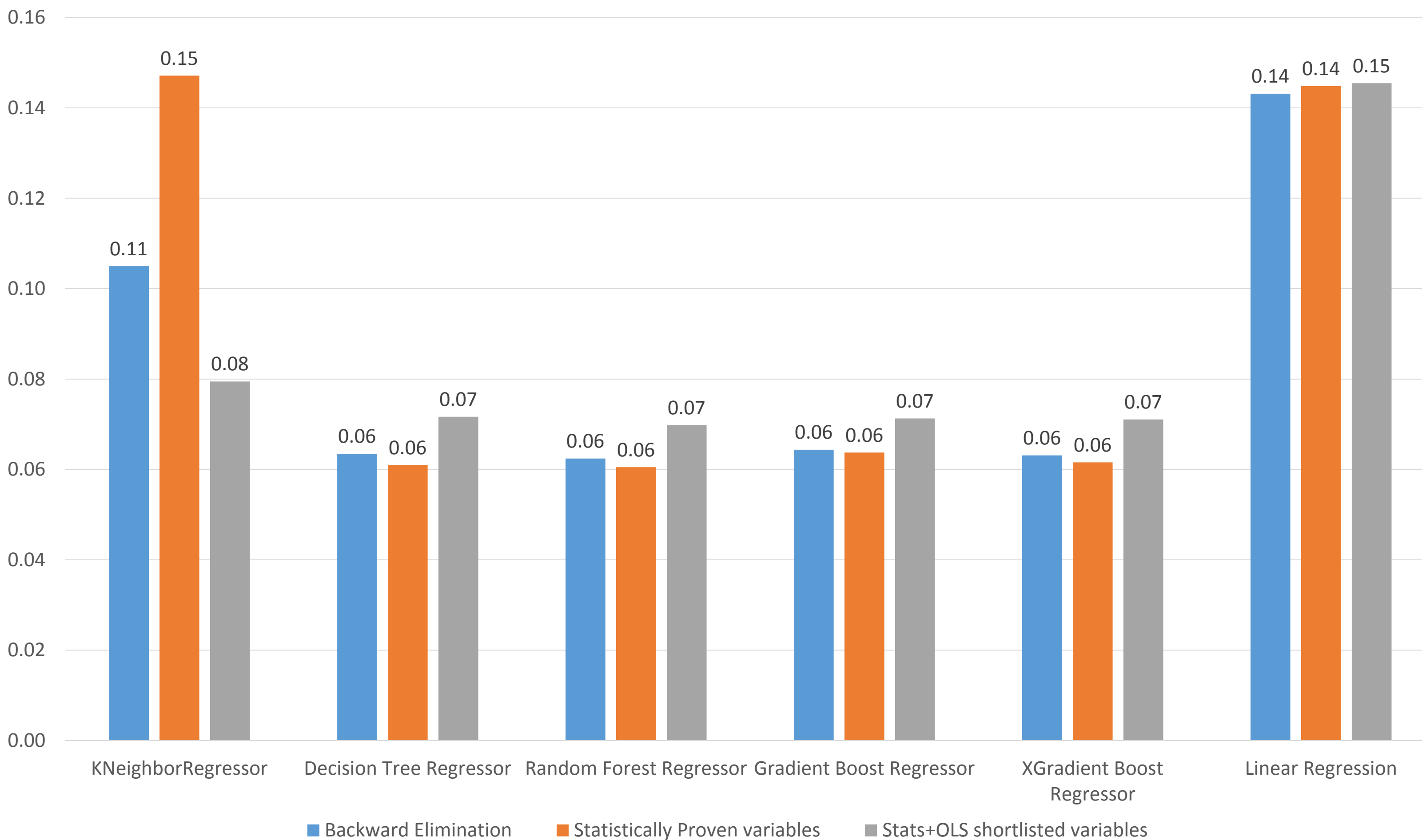
MACHINE LEARNING MODEL RESULTS



- KNN is clearly a very unpredictable model as the for few datasets it gives really poor results but for some it gives higher accuracy.
- Decision Tree, Random Forest and both the Gradient Booster models perform fairly the same with different datasets. With slightly higher error rate for the Stats+OLS model.
- Linear Regression model shows least accuracy and maximum errors.

MACHINE LEARNING MODEL RESULTS

Mean Absolute Errors



- KNN is clearly a very unpredictable model as the for few datasets it gives really poor results but for some it gives higher accuracy.
- Decision Tree, Random Forest and both the Gradient Booster models perform fairly the same with different datasets. With slightly higher error rate for the Stats+OLS model.
- Linear Regression model shows least accuracy and maximum errors.

SELECTING THE BEST MODEL



Model Interpretability

For Stats +OLS model is a complex dataset involving dual elimination methods to select the 6 favorable variables.



Model Accuracy

Basis the accuracy scores, we will not proceed with KNN and Linear Models.



Model Reliability

Simpler model of Backward Elimination is definitely more reliable. Basis on the company's budget the remaining DT, RF & GB models can be picked.

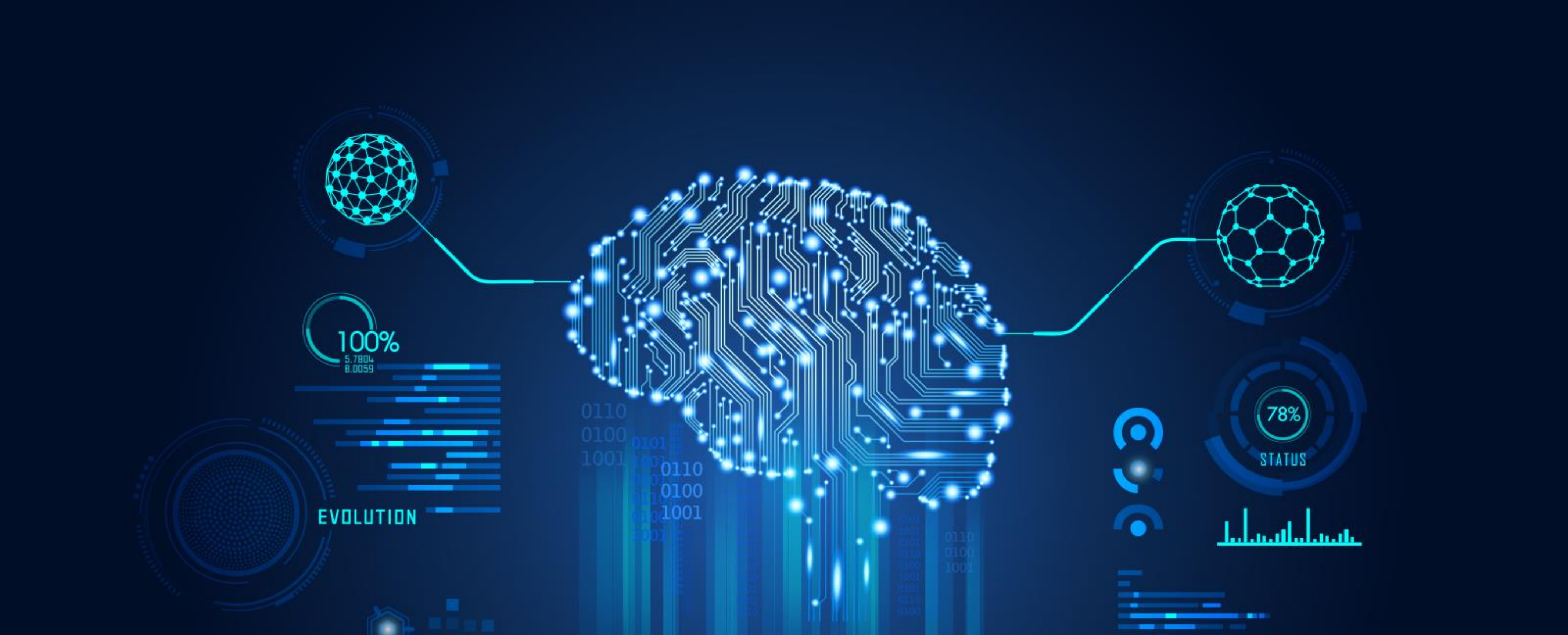


Model Simplicity

On the basis of simplicity, the final model should have least dimensions. Backward Elimination is a much simpler model.

Proceed with backward elimination dataset with 8+1 variables.

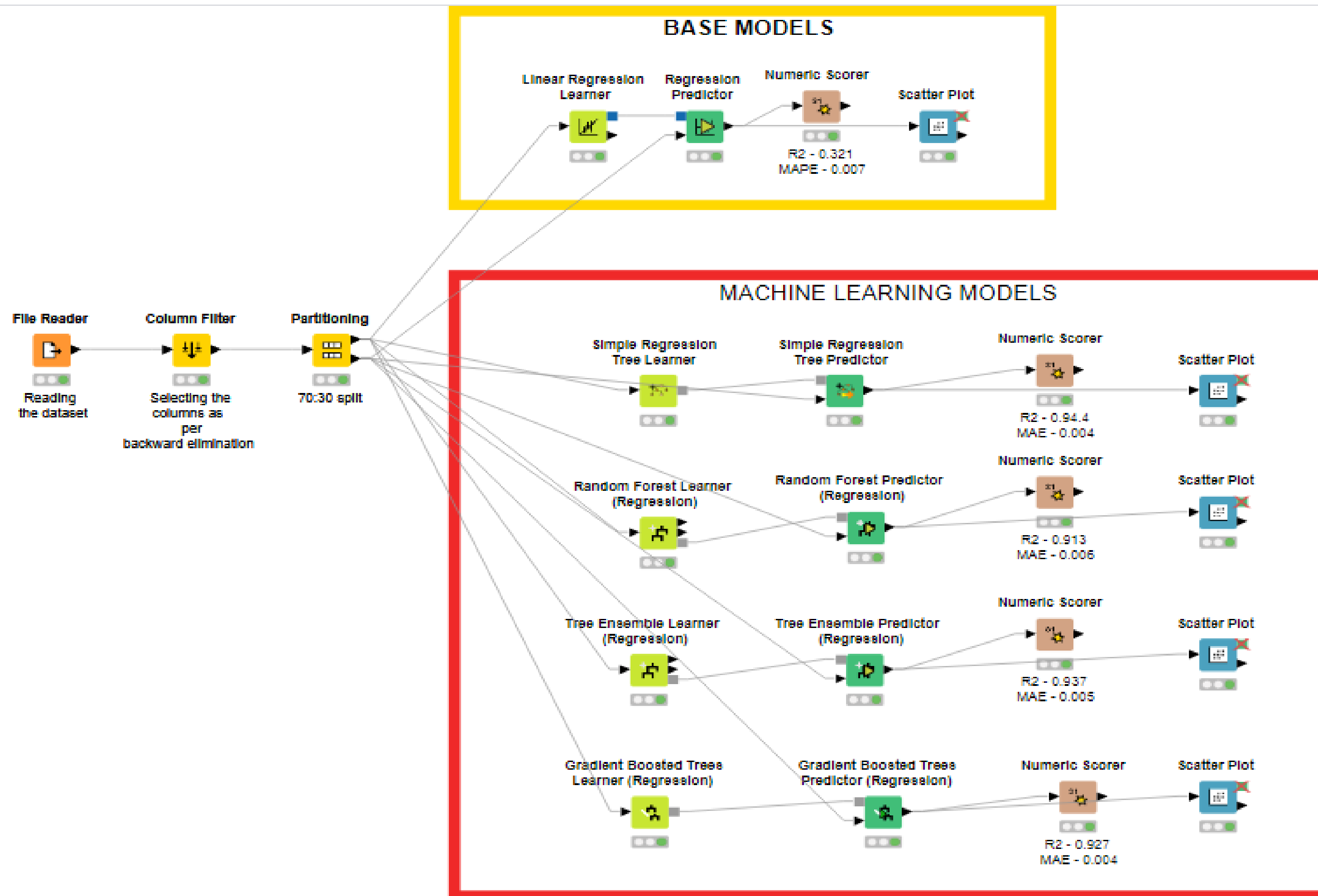
Best performing models are – Decision Tree Regressor, Random Forest Regressor, Gradient Boost Regressor and XGBoost Regressor.



KNIME – AUTO ML TOOL

KNIME ARCHITECTURE

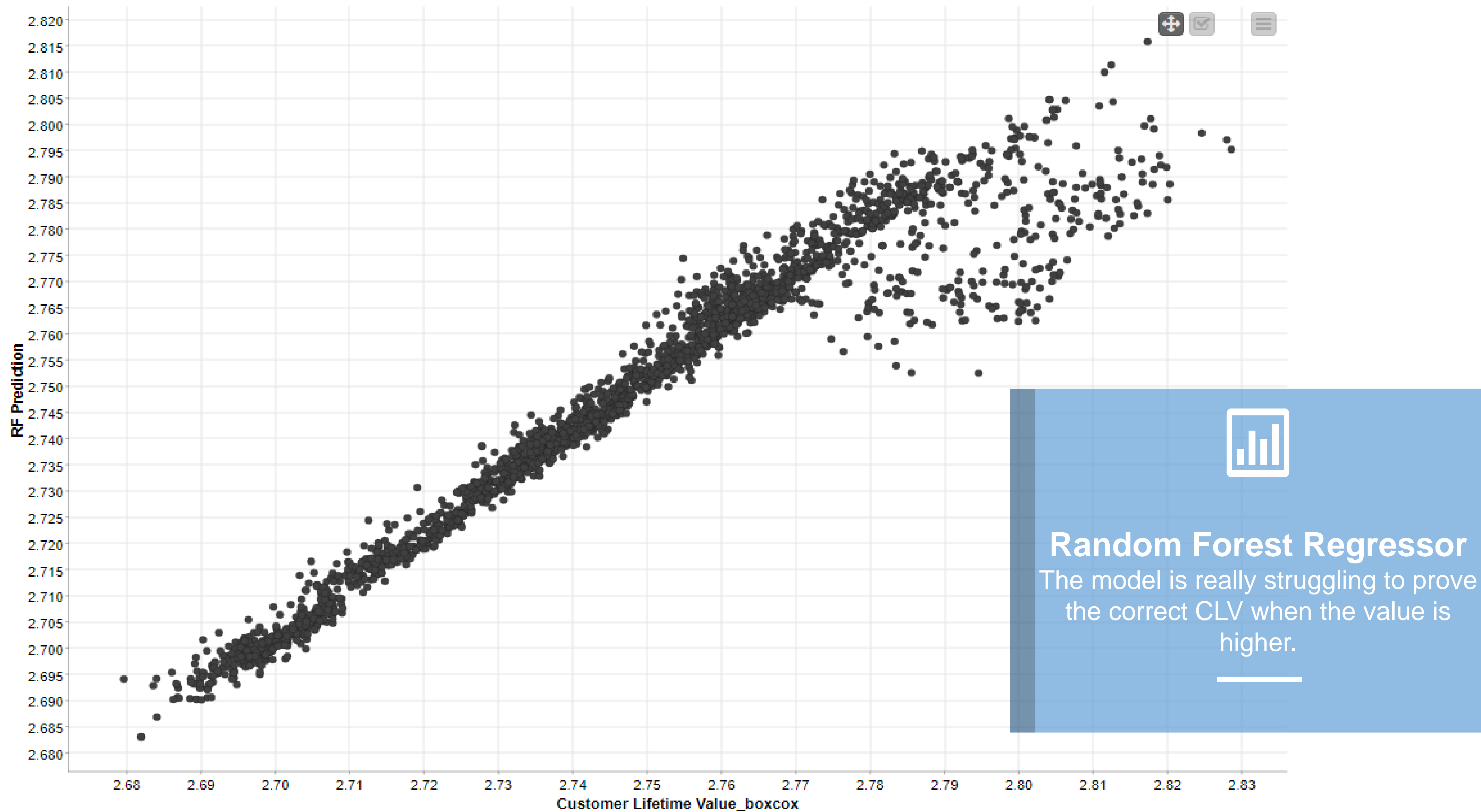
Model with 8+1 features shortlisted using Backward Elimination



KNIME RESULTS VS PYTHON

Data Type	Backward Elimination			
No. of X variables	8+1			
Platform	Knime		Python	
	R2	MAE	R2	MAE
	*100		*100	
Base Models				
Linear Regression	32.1	0.007	29.5	0.14
Machine Learning models				
Decision Tree	94.4	0.004	93.2	0.06
Random Forest	91.3	0.006	94.26	0.06
Gradient Booster	92.7	0.004	94.05	0.06

- DT in Knime has performed better than python by 1%, however the DT is still giving better scores with the best parameters from python.
- RF scores have fallen drastically as the n_estimators in RF is currently at 20, however if taken as 100 which is default, the R2 score reaches 94%.
- Also the mean absolute errors are lower than python, there is less room for errors in Auto ML models.





MODEL DEPLOYMENT USING FLASK

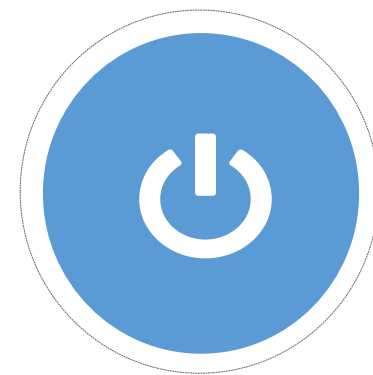
MODEL DEPLOYMENT - FLASK

Model deployment in Flask is a 3 step procedure

Auto Insurance - CLV prediction

Step1: HTML

Create an HTML form with all the required data from the user



1. Response

2. EmploymentStatus

3. Monthly Premium Auto:

4. Number of Open Complaints [0-5]:

5. Number of Policies [0-9]:

6. Policy Type

7. Renew Offer Type

8. Vehicle Class

Step 2: Model.py

Re-create the model in the form of model.py which can be read Flask.



```
RF = RandomForestRegressor(n_estimators = 20, max_depth=9, min_samples_leaf= 9, min_samples_s
RF.fit(X, y)

# Saving model to disk
pickle.dump(RF, open('model.pkl', 'wb'))

# Loading model to compare the results
model = pickle.load(open('model.pkl', 'rb'))
```

The model which is in the form of a python object into a character stream using pickling. The idea is that this character stream contains all the information necessary to reconstruct the to be read by app.py

Step 3: App.py

Create an API which receives details through GUI and computes the predicted lifetime value based on our model.



i 127.0.0.1:5000

For this we de-serialized the pickled model in the form of python object. We set the main page using index.html. On submitting the form values using POST request to /predict, we get the predicted lifetime value.

Final Model

The model returns the predicted value.

The original value to be predicted was \$3102.99



Auto Insurance - CLV prediction

1. Response

No

▼

2. EmploymentStatus

Employed

▼

3. Monthly Premium Auto:

78

4. Number of Open Complaints [0-5]:

0

5. Number of Policies [0-9]:

1

6. Policy Type

Personal Auto

▼

7. Renew Offer Type

Offer2

▼

8. Vehicle Class

Four-Door Car

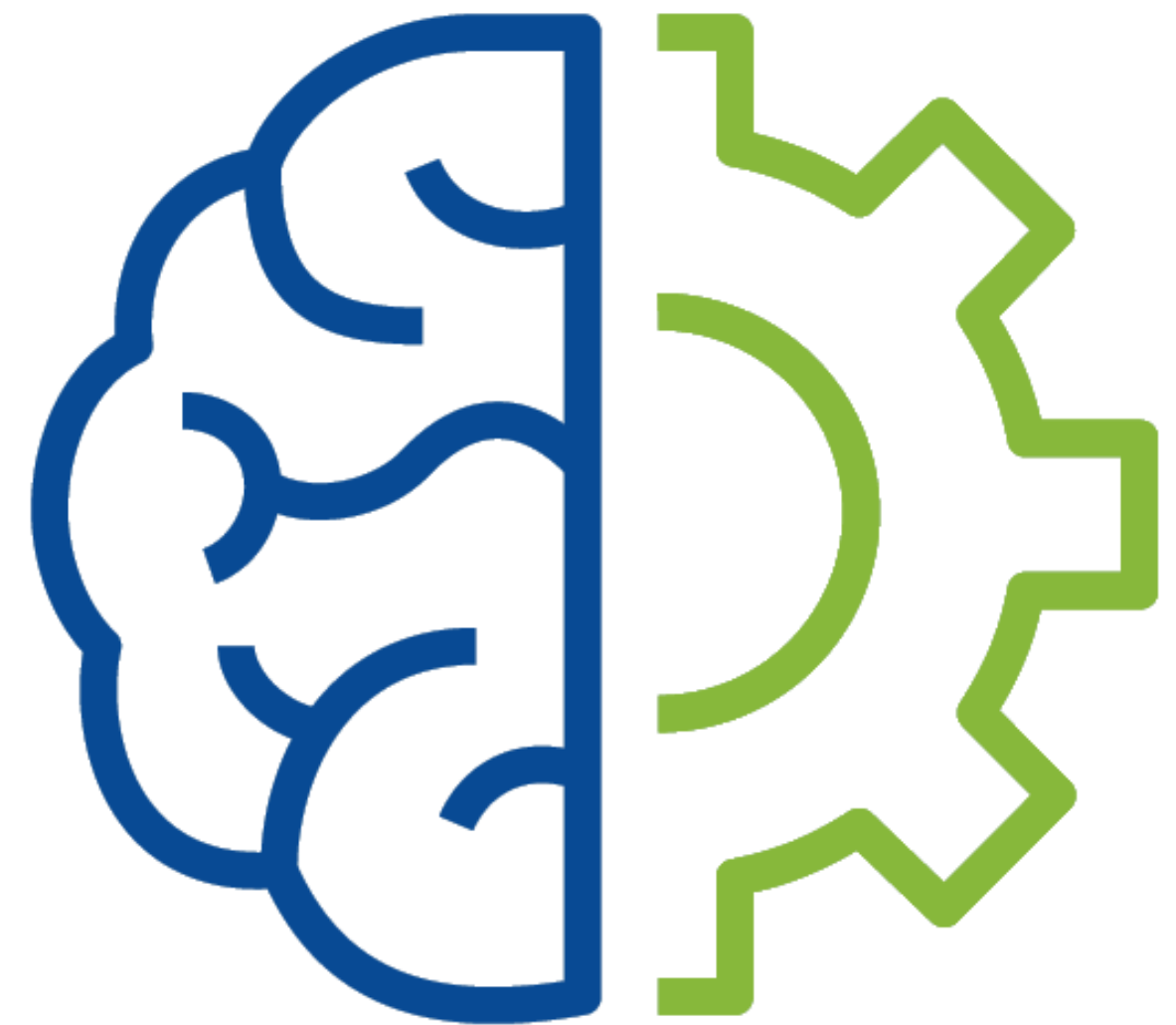
▼

Predict

Customer Lifetime Value \$ 3095.44

CHALLENGES

- Highly right skewed Target Variable and other Numerical data.
- Move to non-parametric tests for feature filtration which increases the complexity.
- The 2 clusters created are not distinct enough to base our conclusions on them.
- There is no strong correlation or any relationship amongst the numerical variables and the CLV.
- The models faced major issue of data leakage which was tweaked as the train & test split were transformed separately.



RECOMMENDATIONS

Depending on the clients requirements, we can suggest the following models:



Cost Effective
Model

Decision Tree
Regressor
R2 - 93.9%



High accuracy
expensive
Models

Gradient Booster
R2 - 94%
XGBooster
R2 – 94.4%



Cost effective
Reliable model

Random Forest
Regressor
R2 – 94.3%

UPSELLING & SCOPE

62%

There are only 5699 customers who are employed and potential candidates for upselling.

25% unemployed customers, most belonging to Cluster 1 with higher premium and no fixed income is a huge concern.

1854 customers are on the higher end of the Basic Coverage paying equivalent to Extended or Premium plans.

Customers upgrading to higher plans by just paying a little more and enjoying more benefits is always an easy sell.

21%

46%

Corporate plan needs to be pitched to 4223 employed customers.

It needs to be promoted enough or tweaked to offer enough perks that customers switch from Personal to Corporate.

Only 365 customers avail the Special Auto policy which needs to be re-considered.

Either the policy can be tweaked to offer better or it can be dropped from the bouquet of offers.

4%

Scope

Washington is highly populated, however the number of customers are lesser.

For certain categories the monthly premium amounts can be revised based on the Avg. amount claimed by the category.



THANK YOU

