

**BANA 6760 - Data Visualization**

Final Report



By

SAI lokesh siddanathi

**Introduction:**

It is Important for a business to track the existing customers and Increase/decrease in the count. While adding new customers is important, it is more important to track the satisfaction of the existing customers. While there is already a system to collect the feedback, there is no mechanism for the Management to view these numbers. Hence this Analysis is critical for the business to identify the CSAT and other KPI’s like Customer Behavior, Loans, Revenue and Assets.

**Solution Proposed:**

There are multiple channels and programs created to collect the feedback. Statistical Analysis is done with R and then Visualizations are created with it. With the visuals, As an Analyst I can derive some recommendations for the business to action immediately for improving the business.

**Proposed Methodology:**

This Analysis is called Exploratory Data Analysis, which helps us understand the datapoints and identify some patterns. Post Identification of these, I shall conduct Regression Analysis and Interpret the results. Then Based on the above results, I shall conclude the suggestions for the business.

Yes, It is definitely an interesting analysis for any Business. As Everyone plans for the improvement and increase the revenue, it is very important for a business so that Management can see the insights and draw actionable items from this. Here customer database is used and the analysis is based on the feedbacks received from the existing Customers.

**Dataset :**

Dataset is taken from the below links. I have referred Mockaroo, where I can create the required Data by assuming that data is currently available and start the analysis based on the available data.

<https://www.iguazio.com/blog/best-13-free-financial-datasets-for-machine-learning/>

<https://www.mockaroo.com/>

Details like Customer Name, Location, ID, Home Branch Details, Education, Employment, Gender, Income, Loans Taken, Coverage…etc. are the known values from the Customer in this dataset.

**Plan of Approach:**

**1. Data Preprocessing:**

* Data Cleaning
* Identifying Missing Values
* Matching the Datatypes / Converting into Proper Datatypes
* Removing Duplicates

**2. EDA – Exploratory Data Analysis**

Visualizing the Data by checking the Metrics that are not clear in initially but clear after preprocessing. Ex: Histogram, Bar plot etc.

Identifying the outliers by ddetermining which of our features are likely to have a connection to our target and take a closer look at those

**3. Train Model**

Using a Supervised Model to do this project. I shall target a column like customer Feedback and shall work on training the model based on this Column. Shall split the features/ Target Split and then plan accordingly. Linear Regression and Random Forecast are two models in my thoughts as of now.

**4. Test Model**

Model can be checked and Validated against the test set for AUC – ROC and Accuracy. Tuning shall be performed accordingly and model is validated.

**5. Draw Conclusions / Insights**

Based on the above steps the model is validated and the visuals are drawn based on the model. The data can be downloaded and then prepared after EDA has been completed. Final Model created is expected to be around AUC – ROC 0.75

**Tools Used in this Process:**

Pre-Processing – MS Excel/ MS Power BI (Power Query Editor)

EDA – R Studio

Visualization – MS Power BI

**Analysis with R:**

R is used for the statistical Analysis for finding the data and the derivations. R helped in understanding to find the insights of data with the help of Regression Analysis.

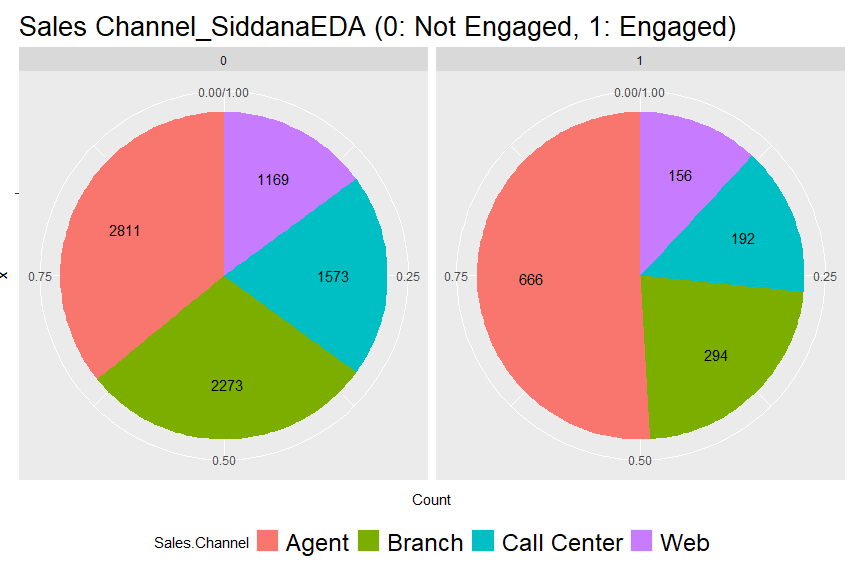
**Charts Used:**

Bar Chart

Pie Chart

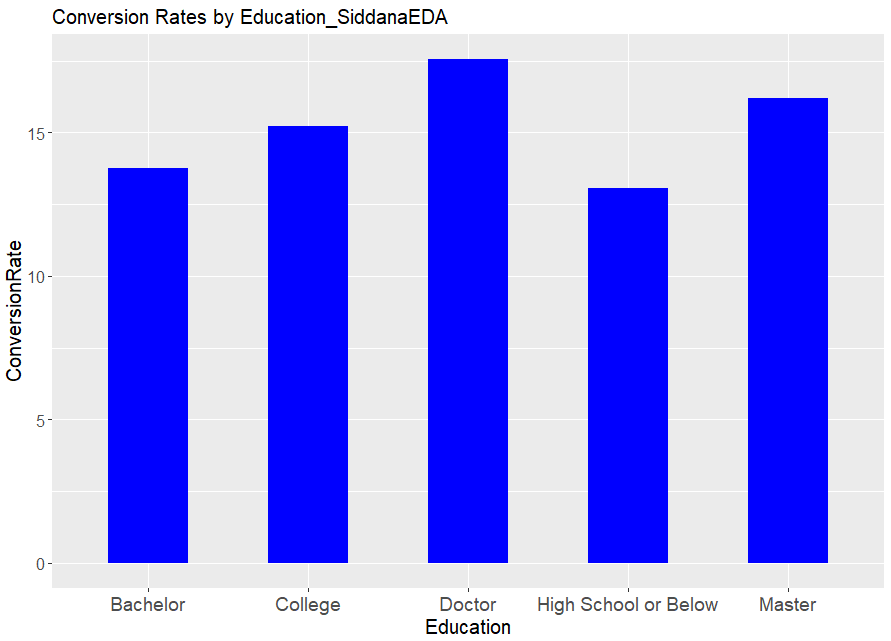
**1. Conversion Rate in Sales Channel:**

This Analyses the Engagement and the summary gives the enagagement summary.



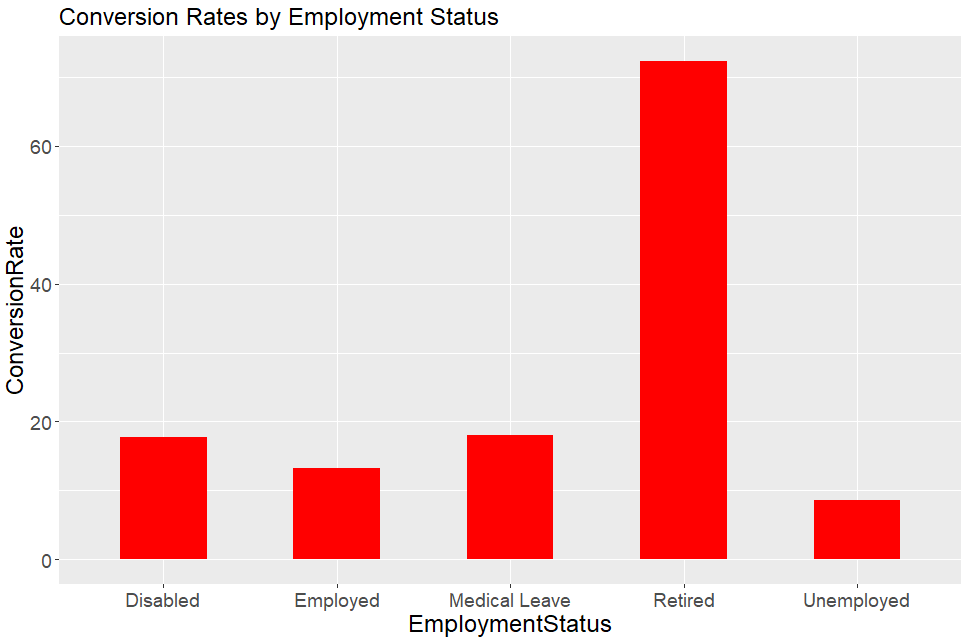
2. **Conversion Rate in State:**

This gives the conversion rate according to each state given in the dataset. Below graph shows that Doctor and Masters students has the high conversion rates.

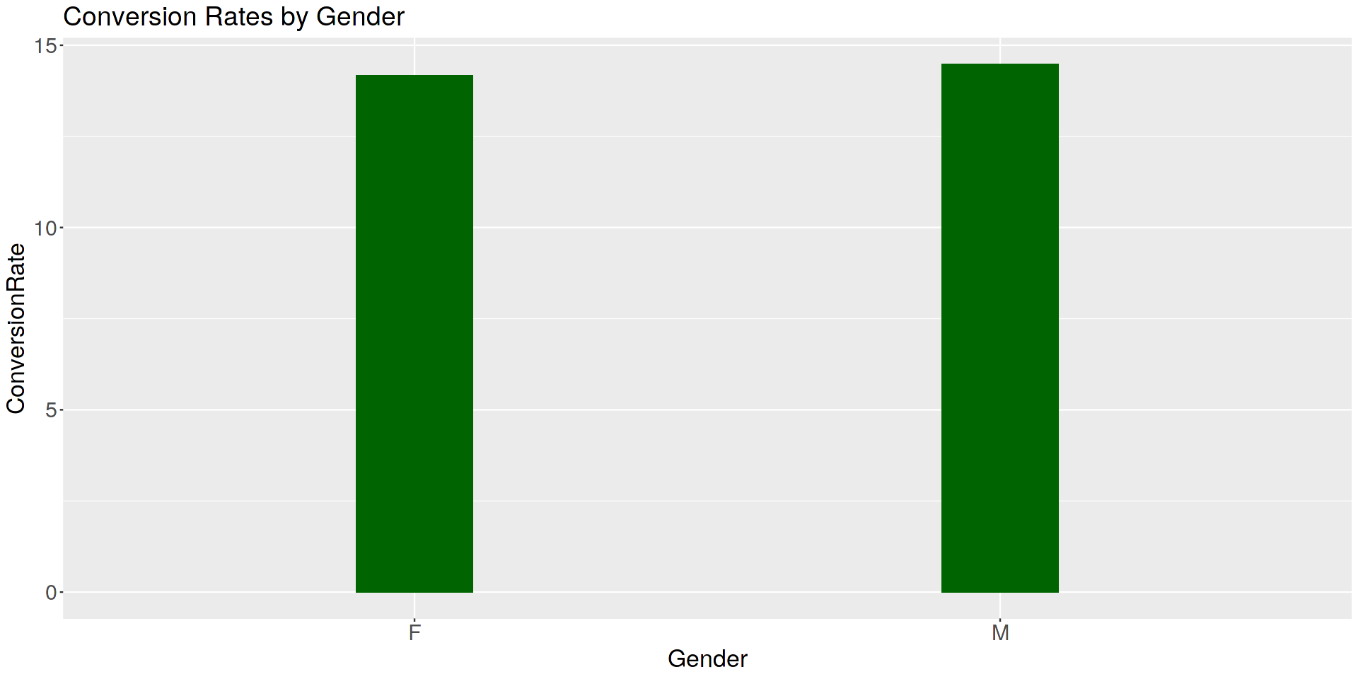


**3. Conversion Rate by Employment Status:**

Retired Employees are having the high conversion Rate. Below is the image for the reference.

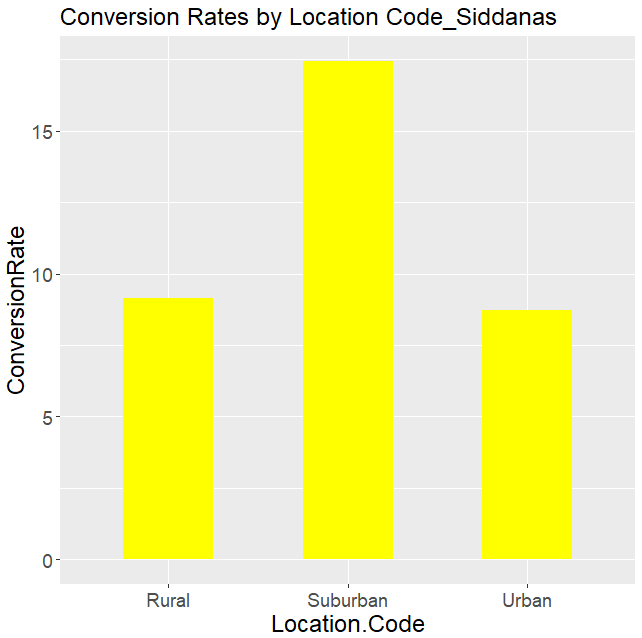


4. **Conversion Rate by Gender:**



**5. Location Wise Conversion Analysis:**

This is the Conversion Rate across the locations. Suburban has the high rate compared to the rural and Urban areas.



**Data Visualization’s with Power BI**

To Analyse the Customer Engagement and calculate the engagement summary with the helpof Data Visualisation Tool such as Power BI. Here we have used Power BI for analysing the data and the metrics for deriving from Analysis we had from R.

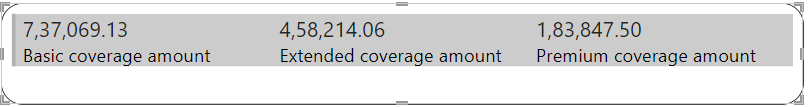
**Charts Used:**

* Bar Chart
* Pie Chart
* Column Chart
* Card Visual
* Slicers
* Table Visual

**Analysis detailed on the below components:**

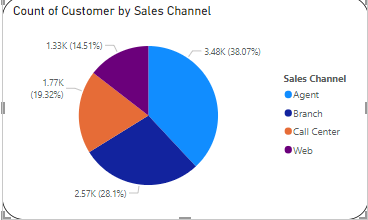
**Amounts Acquired:**

There are different amounts that are generated from the customers. Here is the quick analysis on the amounts collected through different channels. Here is the card visual that displays all the information in a single visual



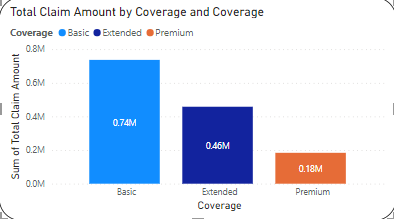
**Customers Through Sales Channel:**

There were different channels for acquiring customers in this sales. Here we have used a pie chart to visualize the number of customers acquired through different channels. Analysis says that ﻿Agent had the highest Count of Customer at 3,477, followed by Branch, Call Center, and Web.﻿﻿ ﻿﻿ ﻿﻿Agent accounted for 38.07% of Count of Customer.﻿﻿ ﻿﻿ ﻿



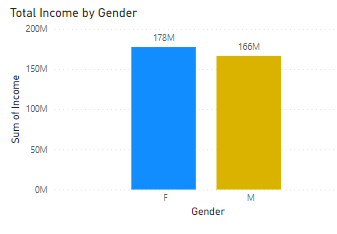
**Total Amounts by Coverage:**

There are different amounts that are covered for the customers. Below is the analysis for the coverage and the customers as it says ﻿At 21,10,474.19, Basic had the highest Sum of Total Claim Amount and was 293.20% higher than Premium, which had the lowest Sum of Total Claim Amount at 5,36,745.56.﻿﻿ ﻿﻿ ﻿﻿Basic had the highest Sum of Total Claim Amount at 21,10,474.19, followed by Extended at 13,17,747.30 and Premium at 5,36,745.56.﻿﻿ ﻿﻿ ﻿﻿Basic accounted for 53.23% of Sum of Total Claim Amount.﻿﻿ ﻿﻿ ﻿﻿Basic had 21,10,474.19 Sum of Total Claim Amount, Extended had 13,17,747.30, and Premium had 5,36,745.56.﻿﻿ ﻿﻿ ﻿



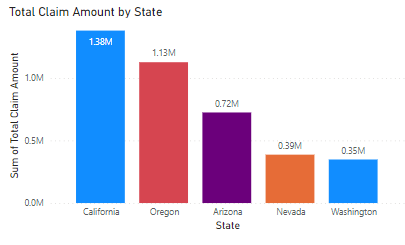
**Total Income by Gender:**

The income is analyzed overall and the Gender is the analysis that is explained for the income generated.﻿﻿﻿Sum of Income for F (177507886) was higher than M (166454623).﻿﻿ ﻿﻿ ﻿﻿F accounted for 51.61% of Sum of Income.﻿﻿ ﻿﻿ ﻿﻿F had 177507886 Sum of Income and M had 166454623.﻿﻿ ﻿﻿ ﻿ F stands for Female and M stands for Male.



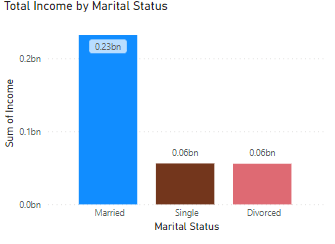
**Total Claims by State:**

This Analysis states the Number of claims made for every state. Here we have used a Column chart to visualize the amounts that are claimed. It is understood that ﻿At 13,79,130.69, California had the highest Sum of Total Claim Amount and was 296.07% higher than Washington, which had the lowest Sum of Total Claim Amount at 3,48,202.19.﻿﻿ ﻿﻿﻿﻿ ﻿﻿California accounted for 34.78% of Sum of Total Claim Amount.﻿﻿ ﻿﻿ ﻿﻿Across all 5 State, Sum of Total Claim Amount ranged from 3,48,202.19 to 13,79,130.69.﻿﻿ ﻿﻿ ﻿



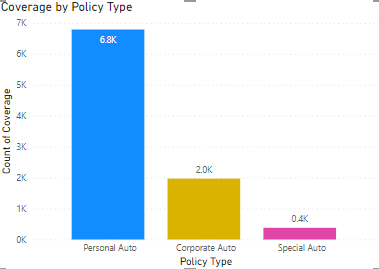
**Total Income by Marital Status:**

Income by marital status plays an important role in understanding the income statistics. It is understood that ﻿At 232128823, Married had the highest Sum of Income and was 316.67% higher than Divorced, which had the lowest Sum of Income at 55710897.﻿﻿ ﻿﻿ ﻿﻿Married had the highest Sum of Income at 232128823, followed by Single at 56122789 and Divorced at 55710897.﻿﻿ ﻿﻿ ﻿﻿Married accounted for 67.49% of Sum of Income.﻿﻿ ﻿﻿ ﻿﻿Married had 232128823 Sum of Income, Single had 56122789, and Divorced had 55710897.﻿﻿ ﻿﻿ ﻿



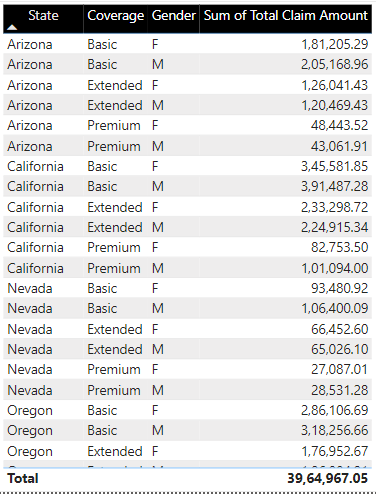
**Coverage By Policy Type:**

﻿At 6,788, Personal Auto had the highest Count of Coverage and was 1,695.77% higher than Special Auto, which had the lowest Count of Coverage at 378.﻿﻿ ﻿﻿ ﻿﻿Personal Auto had the highest Count of Coverage at 6,788, followed by Corporate Auto at 1,968 and Special Auto at 378.﻿﻿ ﻿﻿ ﻿﻿Personal Auto accounted for 74.32% of Count of Coverage.﻿﻿ ﻿﻿ ﻿﻿Corporate Auto had 1,968 Count of Coverage, Personal Auto had 6,788, and Special Auto had 378.﻿﻿ ﻿﻿ ﻿



**Over all Analysis:**

﻿Basic Coverage had the highest total Sum of Total Claim Amount at 21,10,474.19, followed by Extended at 13,17,747.30 and Premium at 5,36,745.56.﻿﻿ ﻿﻿ ﻿﻿California in Coverage made up 18.59% of Sum of Total Claim Amount.﻿﻿ ﻿﻿ ﻿﻿Basic had the highest average Sum of Total Claim Amount at 4,22,094.84, followed by Extended at 2,63,549.46 and Premium at 1,07,349.11.﻿﻿ ﻿﻿ ﻿﻿[]﻿﻿ ﻿﻿ ﻿



**Conclusion & Derivations:**

From the Regression Analysis, we can use it for finding the significance of Variables and only Income, Number. Of. Policies, and Total. Claim. Amount had significant correlations with the variable Engaged, according to the results of this regression study. Income and Total. Claim. Amount are positively related to Engaged (coefficients are positive). That means that the higher our clients' income and total claim amount, the more likely they are to interact. There are certain uncertainties working with this data because the range of the data available is very less to come to exact conclusions and give precise interpretations.

**References:**

<https://www.iguazio.com/blog/best-13-free-financial-datasets-for-machine-learning/>

<https://www.mockaroo.com/>