**Real Estate Project – Summary Report**

Problem Statement:

A banking institution requires actionable insights from the perspective of Mortgage-Backed Securities, Geographic Business Investment and Real Estate Analysis.

The objective is to identify white spaces/potential business in the mortgage loan. The mortgage bank would like to identify potential monthly mortgage expenses for each of region based on factors which are primarily monthly family income in a region and rented value of the real estate. Some of the regions are growing rapidly and Competitor banks are selling mortgage loans to subprime customers at a lower interest rate. The bank is strategizing for better market penetration and targeting new customers. A statistical model needs to be created to predict the potential demand in dollars amount of loan for each of the region in the USA. Also, there is a need to create a dashboard which would refresh periodically post data retrieval from the agencies. This would help to monitor the key metrics and trends.

The dashboard must demonstrate relationships and trends for the key metrics as follows: number of loans, average rental income, monthly mortgage and owner’s cost, family income vs mortgage cost comparison across different regions. The metrics are described not to limit the dashboard to these few only.

Solution:

1. **The training data was loaded in python and the column “UID” was set as the Index.**
2. **Data Pre-processing: It included the following:**
   1. **Removing duplicate records**

The training data set comprised of 27321 records, of which 160 were duplicate. The unique records were 27161.

* 1. **Filling missing values**

The missing values in any of the record was not more than 1.76%. However, the treatment was carried out based on each feature.

* The BlockID was completely blank across all the records. Hence the same was removed.
* Since “Primary” and “SumLevel” columns had the same value in all the records, the same were also removed.
* Out of all the records, there were 142 records where the population was Zero. In these records, all other columns had no data. Hence, these records were deleted.
* There were 65 records, where there was no data regarding the Family Income. Since these records also had missing information for other key features such as Rent, Mortgage cost, home owner cost, loan/debt details, hence these were also deleted.
* There were still some records where the rental information, household income, family income, monthly mortgage and owners cost was missing. ***Since these columns are related to the geographic location, the missing values were filled with average values for the respective "state" to which it belonged.***
* The remaining missing values were filled with the average of the respective columns.

1. **Perform debt analysis:**
2. **Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10%. Visualize using geo-map.**

EDA carried out in Tableau. Most of the cities in the eastern states were part of the Top 2500 locations.

1. **Bad Debt = P (Second Mortgage ∩ Home Equity Loan)**

**Bad Debt = Second Mortgage + Home Equity – Home Equity Second Mortgage**

**Good Debt = Debt – Bad Debt**

* Separate fields were created in Tableau.
* Out of the total records, about 50% cases had Bad Debt percentage of less than 10%, whereas, about 41% cases had Bad Debt percentage between 10-20%.

1. **Create pie charts (Venn diagram) to show overall debt (% bad and good debt) and bad debt (Second Mortgage and Home Equity loan).**

The average Good Debt percentage across all locations was 52.4%, and average Bad Debt percentage was 10.5%.

1. **Create Box and whisker plot and analyze the distribution for 2nd mortgage, home equity, good debt and bad debt for different cities.**

The analysis was carried out in Tableau for Top 10 cities. Based on the count of cases, the top 10 cities were: Baltimore, Bronx, Brooklyn, Chicago, Houston, Las Vegas, Los Angeles, Philadelphia, Phoenix and San Antonio.

* Second Mortgage:
  + Los Angeles, Brooklyn and Chicago had higher outliers.
  + One of the locations in Los Angeles had highest 2nd mortgage of 56%.
  + Highest median values for LA and Bronx of 4.5% and 4.6% respectively was observed.
* Home Equity:
  + Bronx, Brooklyn, Chicago and Los Angeles had higher outliers.
  + One of the locations in Bronx had highest home equity of 100%.
  + Highest median values for LA were observed as 12.1%.
* Good Debt:
  + Las Vegas had more outliers at the lower end i.e. less Good Debt %.
  + Highest median values for Las Vegas were observed as 61.6%.
* Bad Debt:
  + Bronx, Brooklyn, Chicago and Los Angeles had higher outliers.
  + One of the locations in Bronx and LA had highest Bad Debt of 100%.
  + Highest median values for LA were observed as 12.7%.

1. **Create a collated income distribution chart for family income, household income and remaining income.**

* Household Income:
  + Median value of the Household Income was $ 64k.
  + Around 50% of the records had mean Family Income between $40k to 70k.
  + Hence the distribution is right skewed.
* Family Income:
  + Median value of the Family Income was $ 79k.
  + Around 49% of the records had mean Family Income between $49k to 80k.
  + Hence the distribution is right skewed.
* Remaining Income: (Mean Family Income – Mean Household Income)
  + Median value of the Household Income was $ 7.5k.
  + There are few records with negative remaining Income.
  + Around 67% of the records had mean Family Income between $0k to 10k.

1. **Perform EDA and come out with insights into population density and age. Visualize the findings using appropriate chart type.**

* **Population density**
* **Median age**

EDA for population analysis and age analysis was carried out in Tableau.

* Population Analysis:
  + Population density computed as Population/Area of Land for each location.
  + The overall population density across all locations was 0.21%
  + About 54% of the records were having population density of <0.1%. And 38% records between 0.1% to 05%.
  + Around 73% of the locations were having average population between 2000 to 5000.
* Median age
  + Computed as: (Male Median Age \* Male Population + Female Median Age \* Female Population)/Total Population
  + The median age across all records was about 39 years. The Male median age was about 38 years and female median age was about 41 years.
  + About 70% of the locations were having median age between 30 to 45 years.

1. **Create bins for population. Analyze the married, separated and divorced population for these population brackets. Visualize using appropriate chart type.**

* The population bins were created for <2K, 2K-4K, 4K-6K, 6K-8K and >8K.
* The average married population ranged between 45% to 56%. It kept on increasing as the population bracket kept on increasing.
* Similarly, the average divorced population ranged between 8% to 11%. It kept on reducing as the population bracket kept on increasing.
* Similarly, the average separated population ranged between 1.6% to 2.4. It also kept on reducing as the population bracket kept on increasing.

1. **Please detail your observations for rent as a percentage of income at an overall level and for different states.**

* The overall rent to income ratio across all states was about 1.4%.
* The state-wise ratio ranged from 0.9% to 1.8%.
* The top 6 states with highest rent to income ratio were Hawaii, California, Florida, Puerto Rico, Nevada and Arizona. The ratio was above 1.6%.
* Similarly, states with lowest rent to income ratio (closer to 1%) were South Dakota, North Dakota, Iowa, West Virginia and Wyoming.

1. **Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.**

* Correlation analysis was carried out for the key parameters: COUNTYID, STATEID, pop, rent\_mean, hi\_mean, family\_mean, hc\_mortgage\_mean, hc\_mean, home\_equity\_second\_mortgage, second\_mortgage, home\_equity, debt, male\_age\_mean, female\_age\_mean, pct\_own, married, separated and divorced.
* Higher correlation (above 0.7) was observed between mean rent, mean household income, family income and house owner cost.
* Higher correlation of about 0.7 was observed ownership percentage and married population.
* Higher correlation of about 0.93 was observed between home equity second mortgage and second mortgage cases.

1. **The economic multivariate data has a significant number of measured variables. The goal is to find where the measured variables depend on a number of smaller unobserved common factors or latent variables. Obtain the common factors and then plot the loadings.**

* Exploratory factor analysis was carried out on 64 variables.
* Kaiser-Meyer-Olkin (KMO) test was performed to measure the suitability of data for factor analysis.
  + The test measures sampling adequacy for each variable in the model and for the complete model.
  + KMO values between 0.8 and 1 indicate the sampling is adequate.
  + LMO model score was 0.86 indicating that the data was suitable for carrying out factor analysis.
* FactorAnalyzer with initial no. of factors = 25 was considered. The eigenvalues and loadings were plotted.
* 10 factors were selected as the optimum number of factors, since their Eigenvalues was greater than 1.
* Factor Analyzer using ‘Varimax’ rotation and 10 factors was performed.
* 80% of cumulative variance was explained by 10 factors.
* Factor Analysis was also carried out for the latent factors such as: 'hs\_degree', 'hs\_degree\_male', 'hs\_degree\_female', 'male\_age\_median', 'female\_age\_median', 'home\_equity\_second\_mortgage', 'second\_mortgage', 'second\_mortgage\_cdf', 'pct\_own' and 'Bad\_Debt'.
  + For HS degree, factor 5 explains the maximum variation in these variables.
  + for Median Age, factor 4 explains the maximum variation in these variables.
  + For Second mortgage statistics, factor 6 explains the maximum variation in these variables.
  + For Ownership percentage, factor 3 explains the maximum variation.
  + For Bad Debt, factor 7 explains the maximum variation.

1. **Build a linear Regression model to predict the total monthly expenditure for home mortgages loan.**

* **Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step**
* **Run another model at State level. There are 52 states in USA.**
* The Sci-kit learn module was used to perform Linear regression model.
* The test data was also prepared on similar lines to the training data sets.
* The test results are as follows:
  + Mean Absolute Error (MAE): 174.9382
  + Mean Squared Error (MSE): 61938.9373
  + Root Mean Squared Error (RMSE): 248.8753
  + Mean Absolute Percentage Error (MAPE): 0.1117
  + **R2: 0.8452**
  + **Adjusted R2: 0.8449**
* **Since the co-efficient of determination R2 (0.8452) and adjusted R2 (0.8449) is closer to 1, the model is acceptable.**
* Hence, the state level analysis is not carried out.
* The predicted values were plotted in matplot library to ascertain the distribution of the predicted values.
  + The predicted values are positively skewed. Visually, they do not appear to be normally distributed.
* However, to check whether they are normally distributed using a statistical measure, Shapiro-Wilk test and Anderson-Darling tests are performed.
* **Shapiro-Wilk test**
  + The test statistic was 0.932 and the P-value was 0.0
  + Since the p-value was below 0.05, it can be concluded that the predicted values were not normally distributed.
* **Anderson-Darling test**
  + The critical value for α = 0.025 is 0.858.
  + Since the test statistic i.e. 226.127 is greater than all the critical values (for different levels of significance), the results are significant at all significance levels.
  + Hence, the null hypothesis can be rejected. i.e. there is sufficient evidence to say that the data is not normally distributed.

1. **Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business:**
2. **Box plot of distribution of average rent by type of place (Village, urban, town etc.)**
3. **Pie charts (Venn diagram) to show overall debt (% bad and good debt) and bad debt (second mortgage and home equity loan)**
4. **Explore the top 2,500 locations where percentage of households with a second mortgage is the highest and percent ownership is above 10.**
5. **Heat map for correlation matrix**
6. **Pie chart to show the population distribution across different types of places (Village, urban, town etc.)**

* Dashboards have been prepared in Tableau.
  + Dashboard 1: Top cities and Debt distribution
  + The median rent was higher in city as compared to other places.
  + Dashboard 2: Debt Analysis
  + Dashboard 3: Income Distribution
  + Dashboard 4: Age Analysis
  + Dashboard 5: Population Analysis
  + 57% of the population was in city. And 15% in towns.