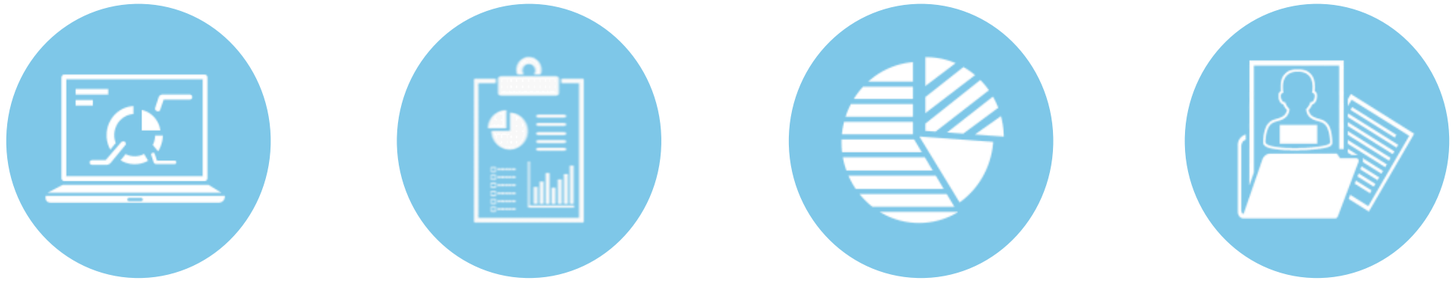
Data Science Capstone Project

Project 1: Real Estate



Problem Statement

A banking institution requires actionable insights into mortgage-backed securities, geographic business investment, and real estate analysis.

The mortgage bank would like to identify potential monthly mortgage expenses for each region based on monthly family income and rental of the real estate.

A statistical model needs to be created to predict the potential demand for amount of loan in dollars 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.

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 described here do not limit the dashboard to these few.

Dataset Description

Variables	Description
Second mortgage	Households with a second mortgage statistics
Home equity	Households with a home equity loan statistics
Debt	Households with any type of debt statistics
Mortgage Costs	Statistics regarding mortgage payments, home equity loans, utilities, and property taxes
Home Owner Costs	Sum of utilities, and property taxes statistics
Gross Rent	Contract rent plus the estimated average monthly cost of utility features
High school Graduation	High school graduation statistics
Population Demographics	Population demographics statistics
Age Demographics	Age demographic statistics
Household Income	Total income of people residing in the household
Family Income	Total income of people related to the householder

Project Task: Week 1

Data Import and Preparation:

1. Import data.

2. Figure out the primary key and look for the requirement of indexing.

3. Gauge the fill rate of the variables and devise plans for missing value treatment. Please explain explicitly the reason for the treatment chosen for each variable.

Project Task: Week 1

Exploratory Data Analysis (EDA):

4.. Perform debt analysis. You may take the following steps:

a) Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10 percent. Visualize using geo-map. You may keep the upper limit for the percent of households with a second mortgage to 50 percent

b) Use the following bad debt equation:

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

Bad Debt = second_mortgage + home_equity - home_equity_second_mortgage

c) Create pie charts to show overall debt and bad debt

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

e) Create a collated income distribution chart for family income, house hold income, and remaining income

Project Task: Week 2

Exploratory Data Analysis (EDA):

1. Perform EDA and come out with insights into population density and age. You may have to derive new fields (make sure to weight averages for accurate measurements):
   a) Use pop and ALand variables to create a new field called population density
   b) Use male_age_median, female_age_median, male_pop, and female_pop to create a new field called median age
   c) Visualize the findings using appropriate chart type
2. Create bins for population into a new variable by selecting appropriate class interval so that the number of categories don't exceed 5 for the ease of analysis.
   a) Analyze the married, separated, and divorced population for these population brackets
   b) Visualize using appropriate chart type
3. Please detail your observations for rent as a percentage of income at an overall level, and for different states.
4. Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.

Project Task: Week 3

Data Pre-processing:

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.

2. Each variable is assumed to be dependent upon a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as “specific variance” because it is specific to one variable. Obtain the common factors and then plot the loadings. Use factor analysis to find latent variables in our dataset and gain insight into the linear relationships in the data.

   Following are the list of latent variables:

   - Highschool graduation rates
   - Median population age
   - Second mortgage statistics
   - Percent own
   - Bad debt expense

Project Task: Week 4

Data Modeling :

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

   Please refer ‘deplotment_RE.xlsx’. Column hc_mortgage_mean is predicted variable. This is the mean monthly mortgage and owner costs of specified geographical location.

   Note: Exclude loans from prediction model which have NaN (Not a Number) values for hc_mortgage_mean.

   a) Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step.

   b) Run another model at State level. There are 52 states in USA.

   c) Keep below considerations while building a linear regression model.

   - Variables should have significant impact on predicting Monthly mortgage and owner costs
   - Utilize all predictor variable to start with initial hypothesis
   - R square of 60 percent and above should be achieved
   - Ensure Multi-collinearity does not exist in dependent variables
   - Test if predicted variable is normally distributed

Project Task: Week 4

Data Reporting:

2. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
   a) Box plot of distribution of average rent by type of place (village, urban, town, etc.).
   b) Pie charts to show overall debt and bad debt.
   c) Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10 percent. Visualize using geo-map.
   d) Heat map for correlation matrix.
   e) Pie chart to show the population distribution across different types of places (village, urban, town etc.).

Thank You