**Project Report (Fidelity International – Data Analytics Training)**

**Author: Amit Gupta (Emp ID: A380501)**

**GitHub URL**

[amit0877/UCDPA\_amitgupta (github.com)](https://github.com/amit0877/UCDPA_amitgupta)

**Abstract**

The project is divided into 2 parts

1st part focuses on building a Machine Learning model for predicting house values in a particular region using Python and its various concepts in a structured approach comprising of Data loading, Data Cleansing & Analysis, Model Building & Analysis

2nd part focuses on using a complex relational database to illustrate the concepts of downloading data from a Relational database and merging dataframes

**Introduction (Usecase 1)**

The objective of project (Use case 1) is to build model to predict median value of houses in the state of California, US based on different metrics like house type, its location, income of people, etc. The dataset would be used to build a supervised learning model and help build understanding of different regression models and hyper parameter tuning techniques. The process of building model would involve use of various data analysis techniques and visualizations. The modem would be build using Python and help demonstrate various concepts as learnt during the training program.

**Dataset**

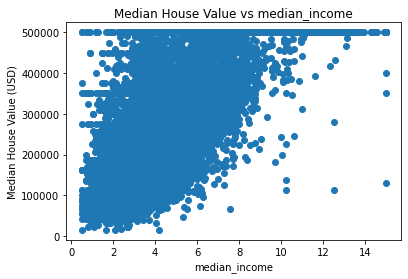
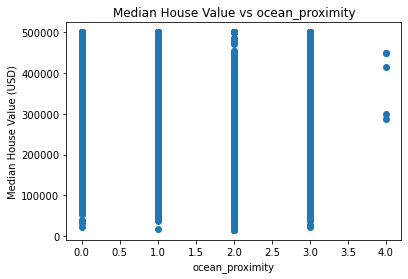
Dataset source ([California House Price | Kaggle](https://www.kaggle.com/datasets/shibumohapatra/house-price))

The above dataset has been taken from Kaggle. This is a California Census Data published by the US Census Bureau and has 10 types of metrics as listed below

1. **Longitude** (float): Longitude value for the block in California, USA
2. **Latitude** (float): Latitude value for the block in California, USA
3. **Housing Median Age** (int): Median age of the house in the block
4. **Total Rooms** (int): Count of the total number of rooms (excluding bedrooms) in all houses in the block
5. **Total bed rooms** (float): Count of the total number of bedrooms in all houses in the block
6. **Population** (int): Count of the total number of populations in the block
7. **Households** (int): Count of the total number of households in the block
8. **Median Income** (float): Median of the total household income of all the houses in the block
9. **Ocean Proximity** (categorical): Type of the landscape of the block [ Unique Values: 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND’]
10. **Median House Value** (int): Median of the household prices of all the houses in the block

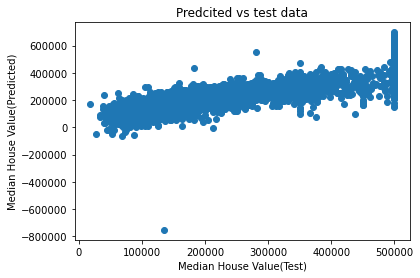
The dataset is a mix of integer, float and categorical variables and would provide a good understanding of how different datatypes contribute to building a Regression Model

**Implementation Process**

1. Data Loading
   1. Dataset (CSV file) was downloaded at the data location from the above mentioned Kaggle URL
   2. Import pandas package
   3. The csv file imported in Pandas DataFrame df using read\_csv function
   4. Verify all columns have been imported as DataFrame columns using df.head()
   5. Print few rows of the data to see how the data looks like (e.g., datatypes, etc)
   6. Check for number of rows(data records)
2. Data Cleansing & Analysis
   1. Check for any missing values. It was found the column total\_bedrooms has some missing values(~1%). Those were replaced by the mean value of the column
   2. Also check for any duplicate values- No duplicate records were found
   3. An alphanumeric categorical numeral (Ocean\_proximity) was converted to numeric to help with model building
   4. Scatter Plot charts were created between each independent variable and the dependent variable (Median Housing Value)
      1. In order to build charts, various Python concepts and libraries were used, namely
         1. List of columns in a dataset to be used for loop iterators
         2. Loop Iterator in order to build the charts between independent variables and the dependent variable
         3. Scatter plot Charts build using matplotlib library

A total of 9 charts created using an iterator. A couple of charts have been added above. From charts, its observed that there is maximum correlation between Median Income the Median House value and the one with Ocean Proximity has least correlation

* 1. A correlation matrix also created which confirms above inference

1. Machine Learning Model Building (using sklearn)
   1. Data arrays created for the Independent variables and the Dependent variable using List
   2. The overall dataset was divided into Training and Test datasets in 4:1 ratio
   3. Linear Regression Model built using LinearRegression Library; KFold library used with n\_splits=6
      1. The model built with a 63% accuracy.
         1. Median House Value= 206855
         2. RMSE = 69489
   4. Scatter plot created between predicted and test values of the dependent variable using Matplotlib
   5. The chart has a low slope indicating a low accuracy of the model.
   6. The data set was then used to model using ither Regression algorithms, namely
      1. Lasso Regression
      2. Ridge Regression
   7. But it’s observed that the model accuracy remains at 63% and no improvement is seen
      1. Hyper Parameter Tuning algorithm (GridSearchCV)was used for Lasso & Ridge Regression models for diff alpha values (0.0001, 1, 10); but no improvement in seen in model accuracy
   8. The probable reason, HyperTuning and other complex regression algorithms haven’t worked is that only one independent variable has a high correlation with the dependent variable and others have very low correlation leaving very little scope of tuning to the models
2. Insights
   1. There is a high correlation between Median Income of households and Median House Value
   2. The Linear Regression model has an accuracy of 63% in predicting the Median House value in State of California
   3. Except for Median Income, all other variables considered have a very correlation with the Median House value

**Introduction (Usecase 2)**

The objective of project (Use case 2) is to illustrate following concepts

* Connect to a Relational database and download data in dataframes
* Merge Dataframes

**Dataset**

Dataset ([World Development Indicators 2022 | Kaggle](https://www.kaggle.com/datasets/psycon/world-development-indicators))

The World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially-recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates.

Name of dataset (**Indicators.sqlite**)

This is an SQLite data base having 6 tables

1. Country
2. CountryNotes
3. Footnotes
4. Indicators
5. Series
6. SeriesNotes

**Implementation Process**

1. Data Loading & create dataframes
   1. Loaded pandas and sqlite3 packages
   2. Connection made with the database Indicators.sqlite and all 6 tables fetched
      1. Country
      2. CountryNotes
      3. Footnotes
      4. Indicators
      5. Series
      6. SeriesNotes
   3. Data from all 6 tables loaded into 6 individual dataframes
2. Merge Dataframes
   1. Understand the column names in different tables to be merged and identify the join conditions on the business keys
   2. Data in dataframes merged based on common business keys and using outer join