**This dataset contains house sale prices for XYZ. It includes homes sold between May 2014 and May 2015.**

**Perform the following Task Mentioned below :**

**Exploratory Analysis**

**1 Ques**.Plot a distribution and box plot for the price variable. Find out if there are any outliers and list them.

**2 Ques**.Find out which variables are highly correlated with 'price'.

**3 Ques**.Find out which zip codes have the highest median house 'price'. Do a bar plot to depict the top 10.

**4 Ques**.Find out if there are any missing values in the dataset and your strategy for imputing those.

**5 Ques**.Do some variables need feature engineering or transformation, if they need to be used to predict prices. Explain what transformations you will apply.

**Model Building**

**6 Ques**.Build a regression model to predict the prices.

**Evaluation**

**7 Ques**.Build the linear model with L1 and L2 regularization parameters. Do a Grid Search for to find optimal values for hyper parameters.

--Lasso

--Ridge

**8 Ques.**Draw a diagram to depict RMSE values for different hyper parameters and show the lowest RMSE at optimal value for L1 and L2 parameters.

**Models Comparison**

**9 Ques**.Build the following models and find the best performing model with lowest RMSE value.

1.Ridge

2.Lasso

3.Elastic Net (With Lasso and Ridge regularizations)

4.Decision Tree

5.Random Forest

**Tools to Deploy : pyspark , Reading Data through MySQL**