CALIFORNIA HOUSING PRICE PREDICTION

BY JAYA RAGHAVENDRA

AGENDA

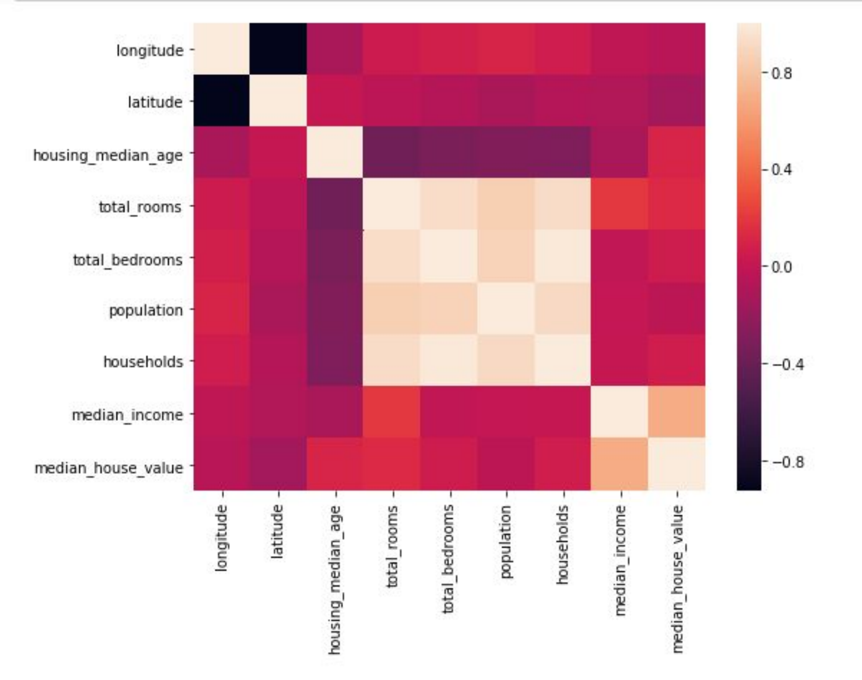
- Objective
- Data Visualization
- Data Preprocessing and cleaning data
- ML Model built and validation
- Test set Prediction and measuring accuracy with metric RMSE [Root Mean Squared Error]

OBJECTIVE

- The purpose of the project is to predict median house values in Californian districts, given many features from these districts.
- The project also aims at building a model of housing prices in California using the California census data. The data has metrics such as the population, median income, median housing price, and so on for each block group in California. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

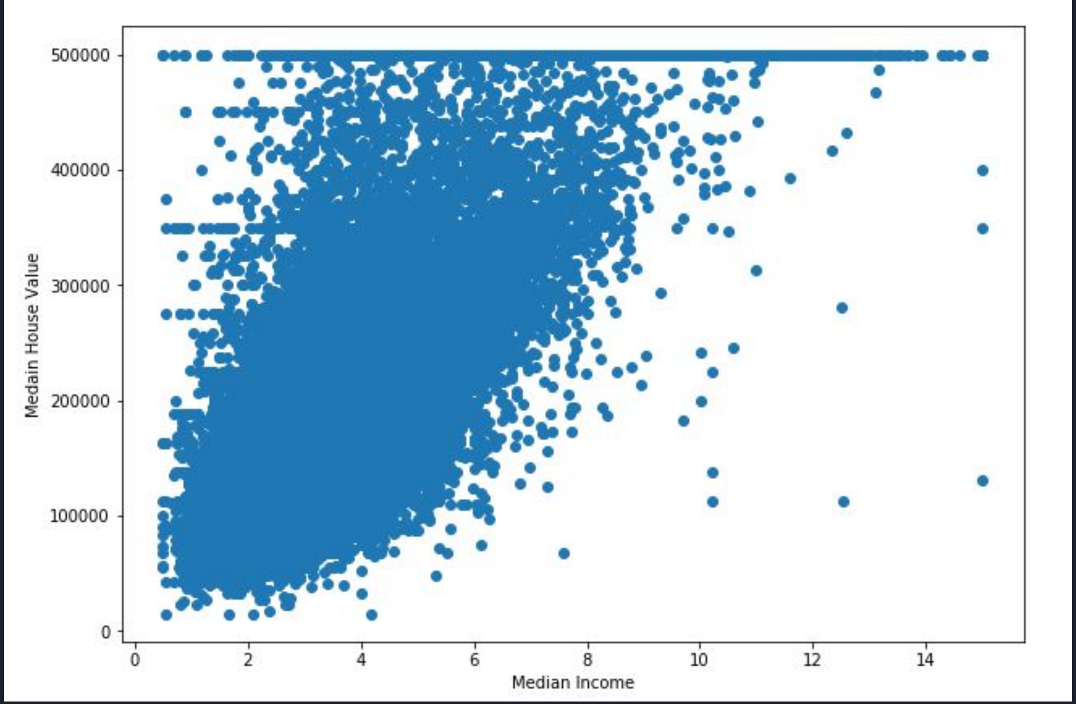
DATA VISUALIZATION

- **Dataset has 20640 records and 10 features.**
- **Correlation heatmap shows relationship among all features.**



DATA VISUALIZATION

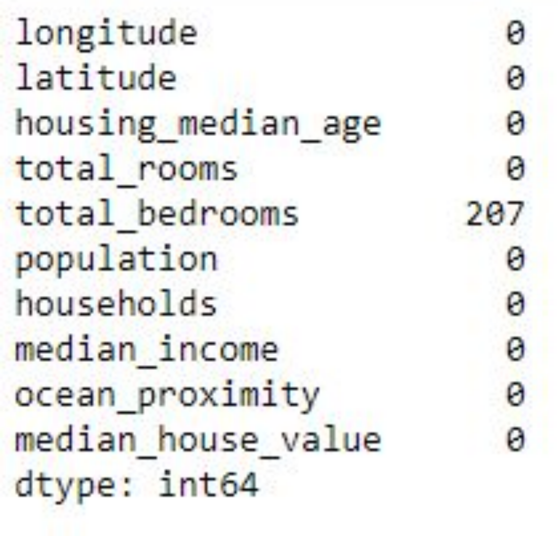
- **Scatter plot of Median Income vs Median house value.**



DATA PREPROCESSING AND CLEANING DATA

Missing values count:

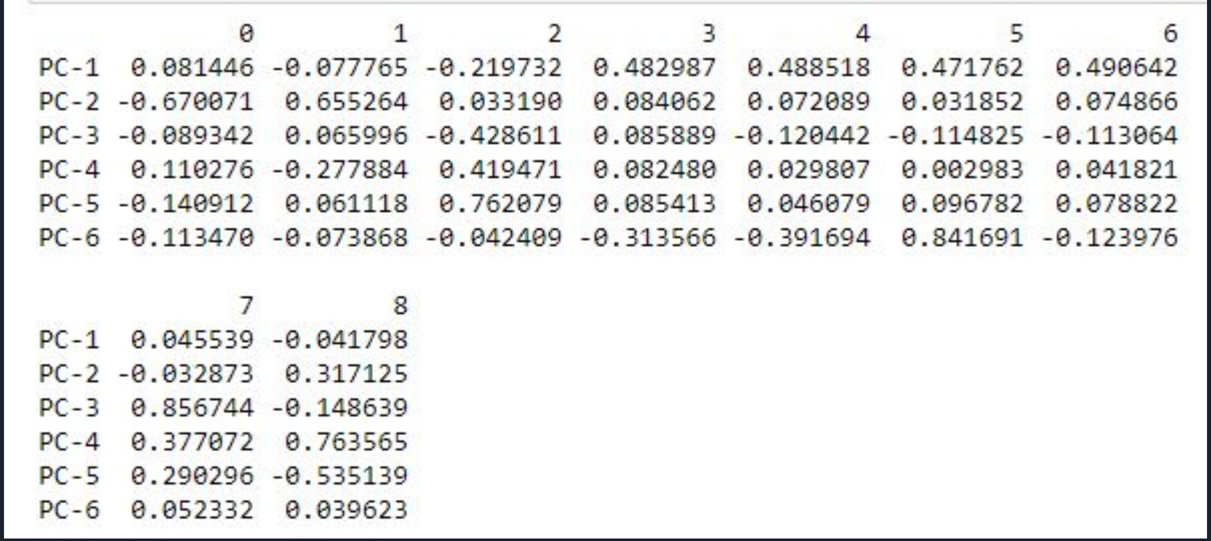
Imputed total bedrooms missing data with mean.



```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
ocean_proximity         0
median_house_value      0
dtype: int64
```

DATA PREPROCESSING AND CLEANING DATA

PCA: Dump components relations with features: This gives us the picture of how features are related to components



```
             0         1         2         3         4         5         6
PC-1  0.081446 -0.077765 -0.219732  0.482987  0.488518  0.471762  0.490642
PC-2 -0.670071  0.655264  0.033190  0.084062  0.072089  0.031852  0.074866
PC-3 -0.089342  0.065996 -0.428611  0.085889 -0.120442 -0.114825 -0.113064
PC-4  0.110276 -0.277884  0.419471  0.082480  0.029807  0.002983  0.041821
PC-5 -0.140912  0.061118  0.762079  0.085413  0.046079  0.096782  0.078822
PC-6 -0.113470 -0.073868 -0.042409 -0.313566 -0.391694  0.841691 -0.123976

             7         8
PC-1  0.045539 -0.041798
PC-2 -0.032873  0.317125
PC-3  0.856744 -0.148639
PC-4  0.377072  0.763565
PC-5  0.290296 -0.535139
PC-6  0.052332  0.039623
```

MODEL BUILDING AND EVALUATION

ML Model	Test Data Set Accuracy	RMSE
Linear regression	53.5%	79072
Decision Trees	63.4%	70080
Random Forest	73%	59251

THANK YOU