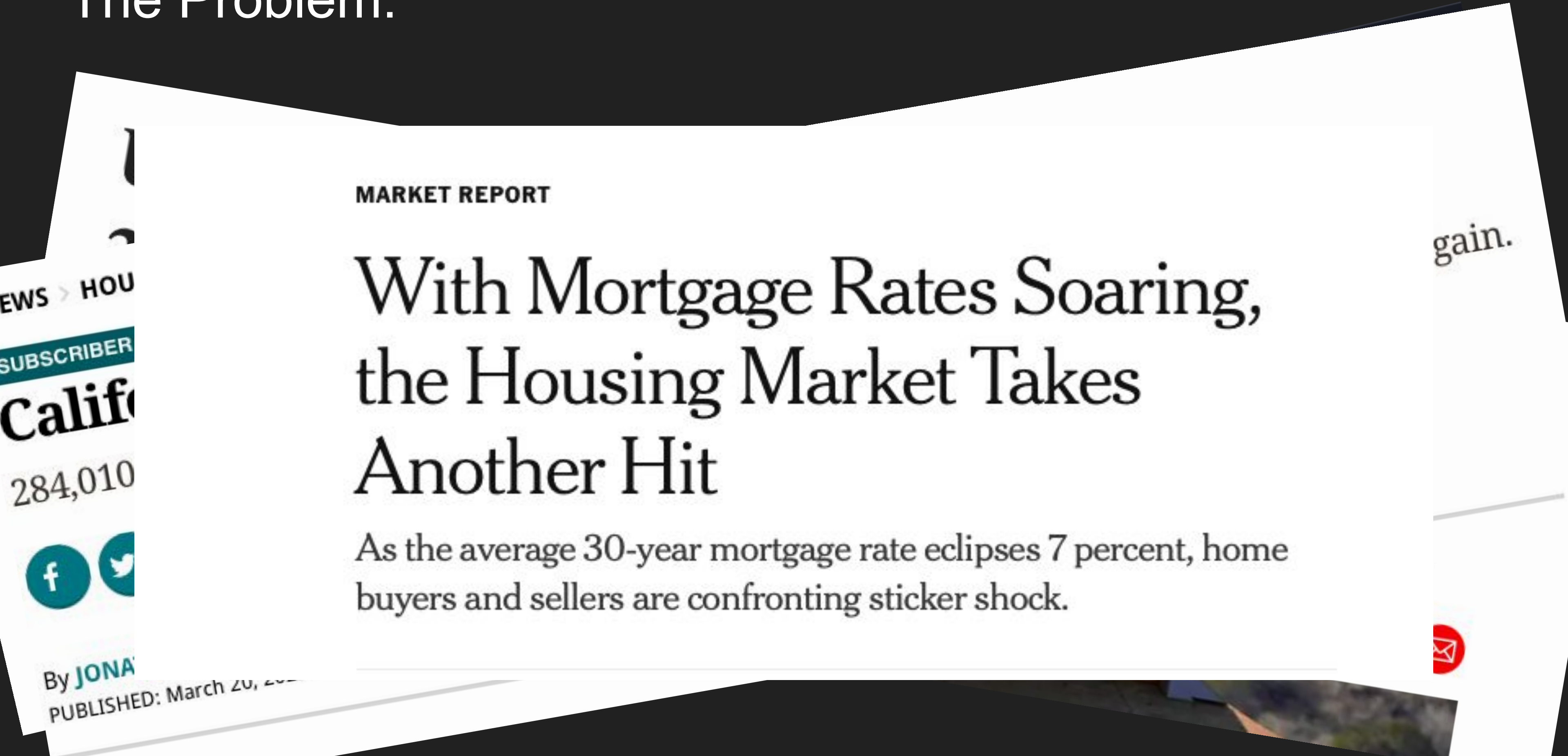
Predicting California Home Values

What determines a home’s value?

The Problem:



EWS › HOU

SUBSCRIBER

Calif

284,010

By JONA

PUBLISHED: March 20, 20

gain.

MARKET REPORT

With Mortgage Rates Soaring, the Housing Market Takes Another Hit

As the average 30-year mortgage rate eclipses 7 percent, home buyers and sellers are confronting sticker shock.

Relevance

Home Buyers

- Narrow your search to find more affordable housing
- Avoid overpaying for homes less likely to increase in value

Investors

- Identify homes that are most likely to increase in value

Developers

- Identify locations where homes will sell for the most

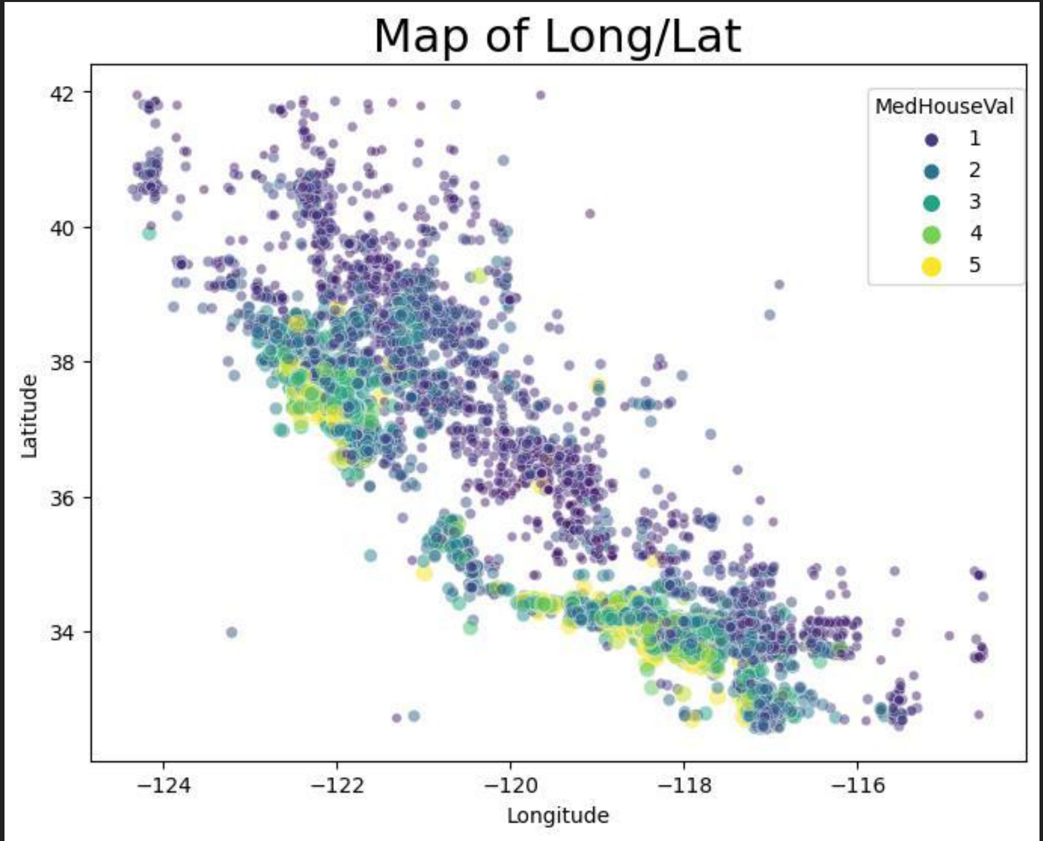
California Housing Data Set

- Available on scikit learn
- Data from 1990 Census
  - Contains home information about blocks of housing in California
  - 37,000 rows
  - 8 Features
- Target Variable: Median Home Value (in each housing block)

What Predicts Home Value?

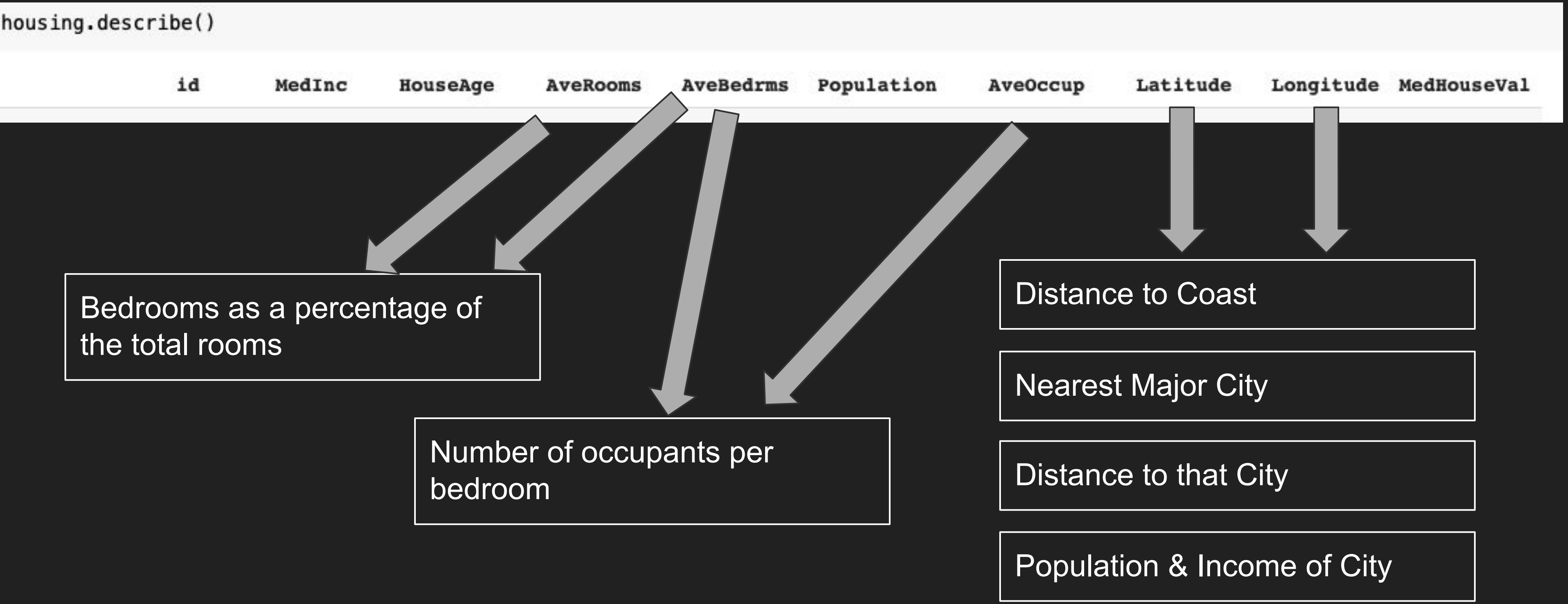
- Owner Income
- Number of Rooms
- Number of Bedrooms
- Age of the House
- Occupancy or how crowded the neighborhood is
- Location (Coordinates)

Exploring the Data: Location, Location, Location!



NEVADA
Sacramento
San Francisco
San Jose
CALIFORNIA
Las Vegas
Los Angeles
San Diego

More Features



```
housing.describe()
```

```
        id        MedInc     HouseAge     AveRooms    AveBedrms   Population      AveOccup     Latitude     Longitude  MedHouseVal
```

Bedrooms as a percentage of the total rooms

Number of occupants per bedroom

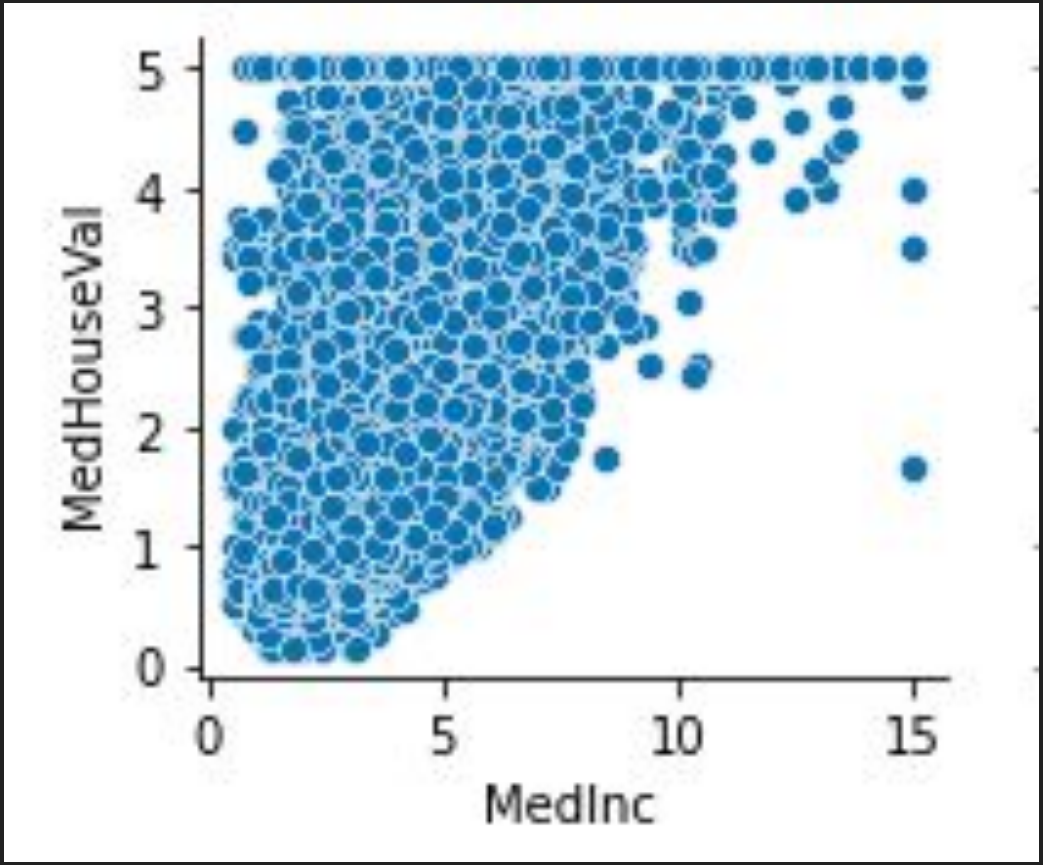
Distance to Coast

Nearest Major City

Distance to that City

Population & Income of City

More Exploratory Analysis



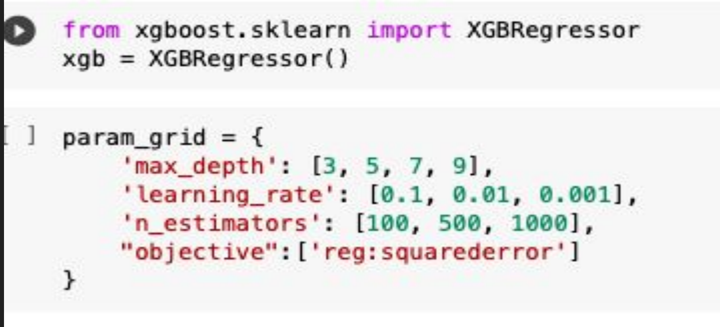
Modelling

- Linear Regression
- Random Forest
- XGBoost

Scaling and Preprocessing

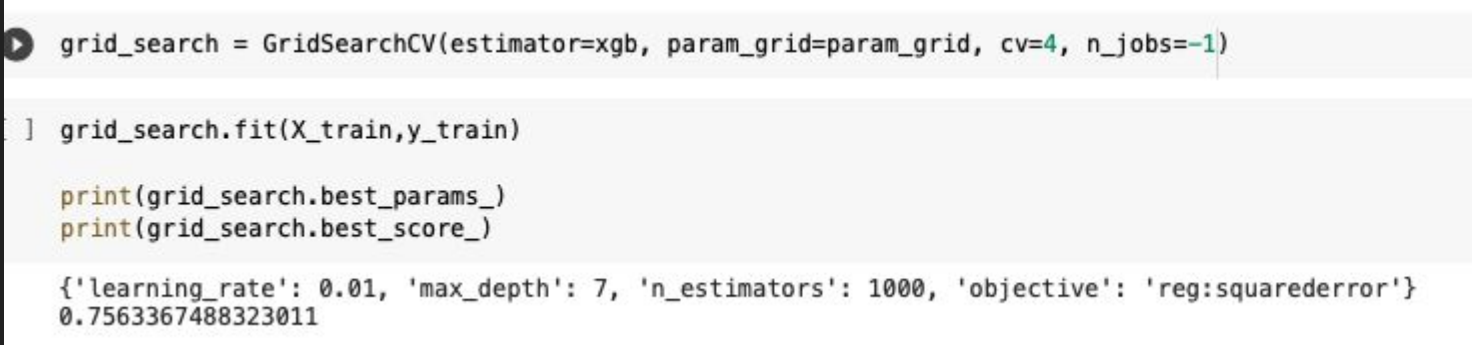
- Linear Regression Requirements
    - Scaling-Used Standard Scaler
        - Mean now equals 0
        - Standard Deviation now equals 1
    - One-Hot-Encoding of Categorical Variables

Hyperparameter Tuning



```
from xgboost.sklearn import XGBRegressor
xgb = XGBRegressor()
```

```
param_grid = {
     'max_depth': [3, 5, 7, 9],
     'learning_rate': [0.1, 0.01, 0.001],
     'n_estimators': [100, 500, 1000],
     "objective":['reg:squarederror']
}
```



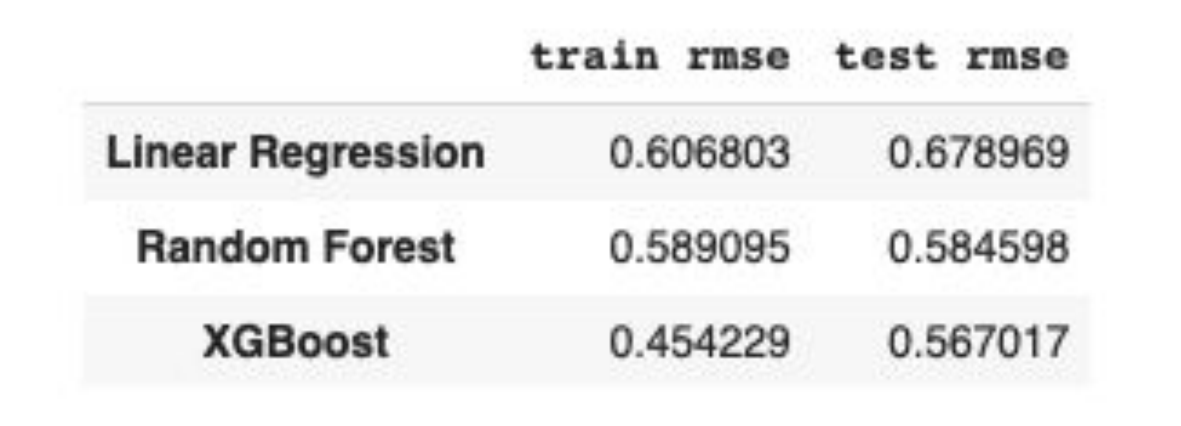
```
grid_search = GridSearchCV(estimator=xgb, param_grid=param_grid, cv=4, n_jobs=-1)
```

```
grid_search.fit(X_train,y_train)

print(grid_search.best_params_)
print(grid_search.best_score_)
```

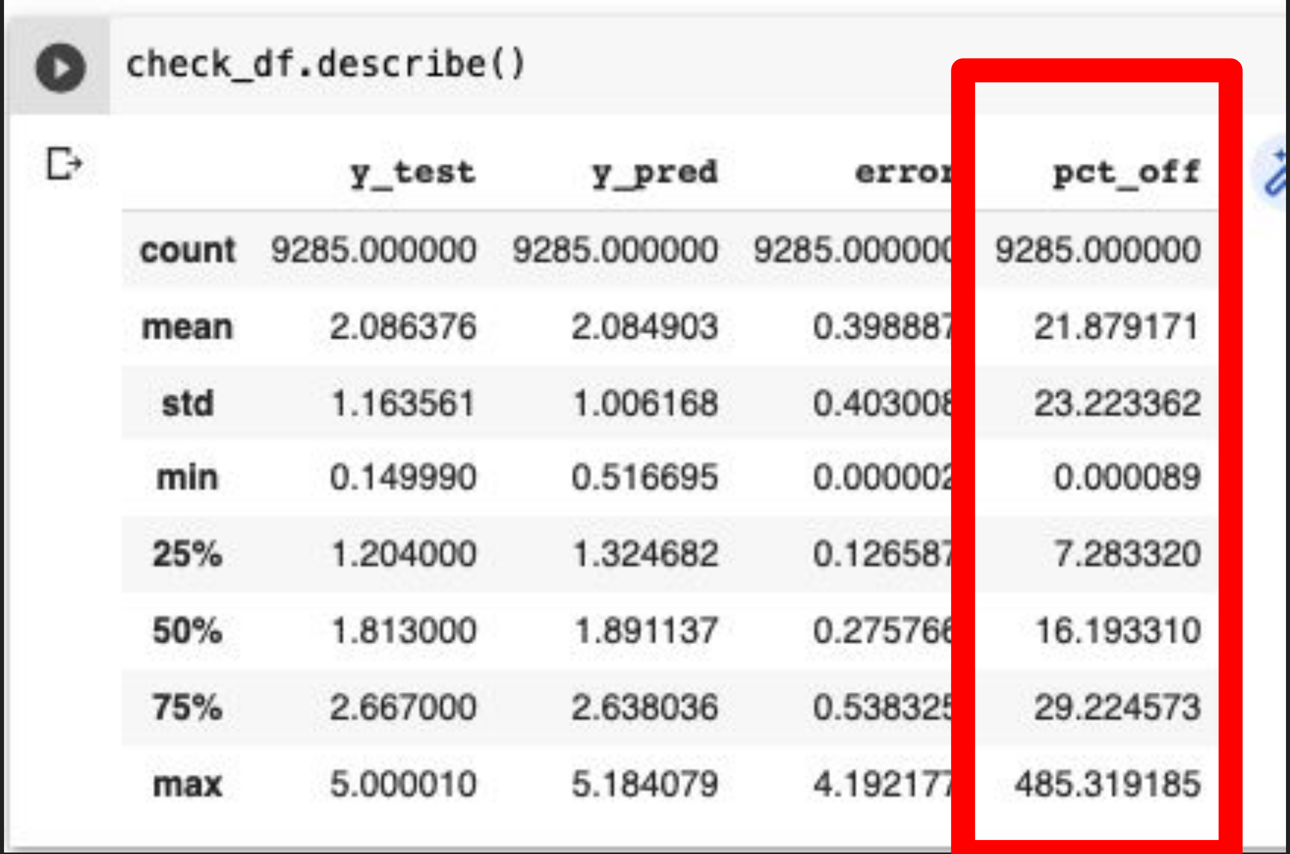
```
{'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 1000, 'objective': 'reg:squarederror'}
0.7563367488323011
```

Results-RMSE



	train rmse	test rmse
Linear Regression	0.606803	0.678969
Random Forest	0.589095	0.584598
XGBoost	0.454229	0.567017

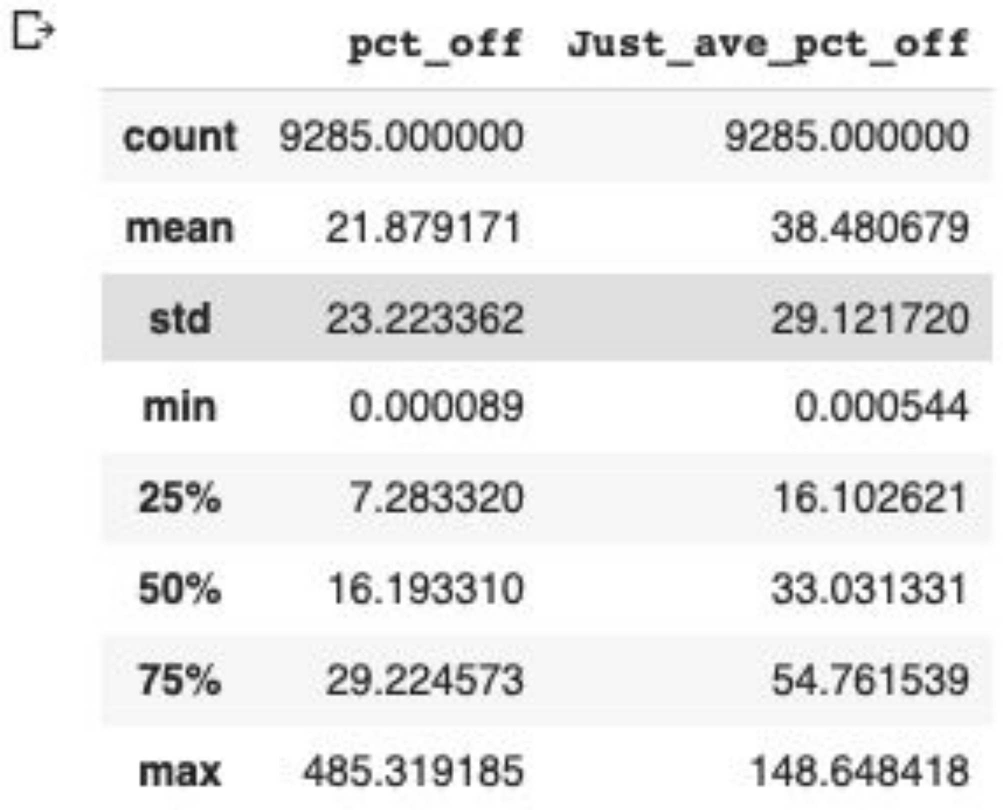
Results-XGBoost



```
check_df.describe()
```

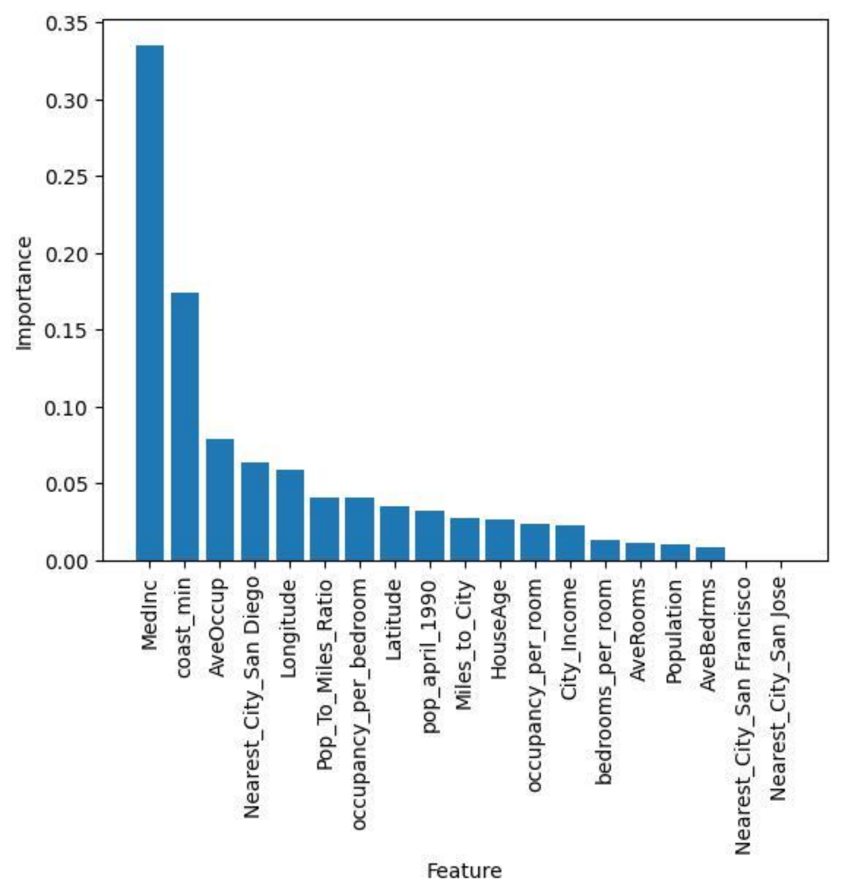
```
             y_test       y_pred        error      pct_off
count   9285.000000  9285.000000  9285.000000  9285.000000
mean       2.086376     2.084903     0.398887    21.879171
std        1.163561     1.006168     0.403008    23.223362
min        0.149990     0.516695     0.000002     0.000089
25%        1.204000     1.324682     0.126587     7.283320
50%        1.813000     1.891137     0.275766    16.193310
75%        2.667000     2.638036     0.538325    29.224573
max        5.000010     5.184079     4.192177   485.319185
```

Results-XGBoost



	pct_off	Just_ave_pct_off
count	9285.000000	9285.000000
mean	21.879171	38.480679
std	23.223362	29.121720
min	0.000089	0.000544
25%	7.283320	16.102621
50%	16.193310	33.031331
75%	29.224573	54.761539
max	485.319185	148.648418

Best Features For XGBoost



Conclusion & Next Steps

- Test model on more recent data
  - Review Outliers to improve model
- Do a similar study with rental prices. Are the same or similar features important to the price of rent?
- Does real estate follow similar patterns or are there other features that are important?