EXPLAINABILITY AI

CALIFORNIAN HOUSING

ABOUT THE DATASET

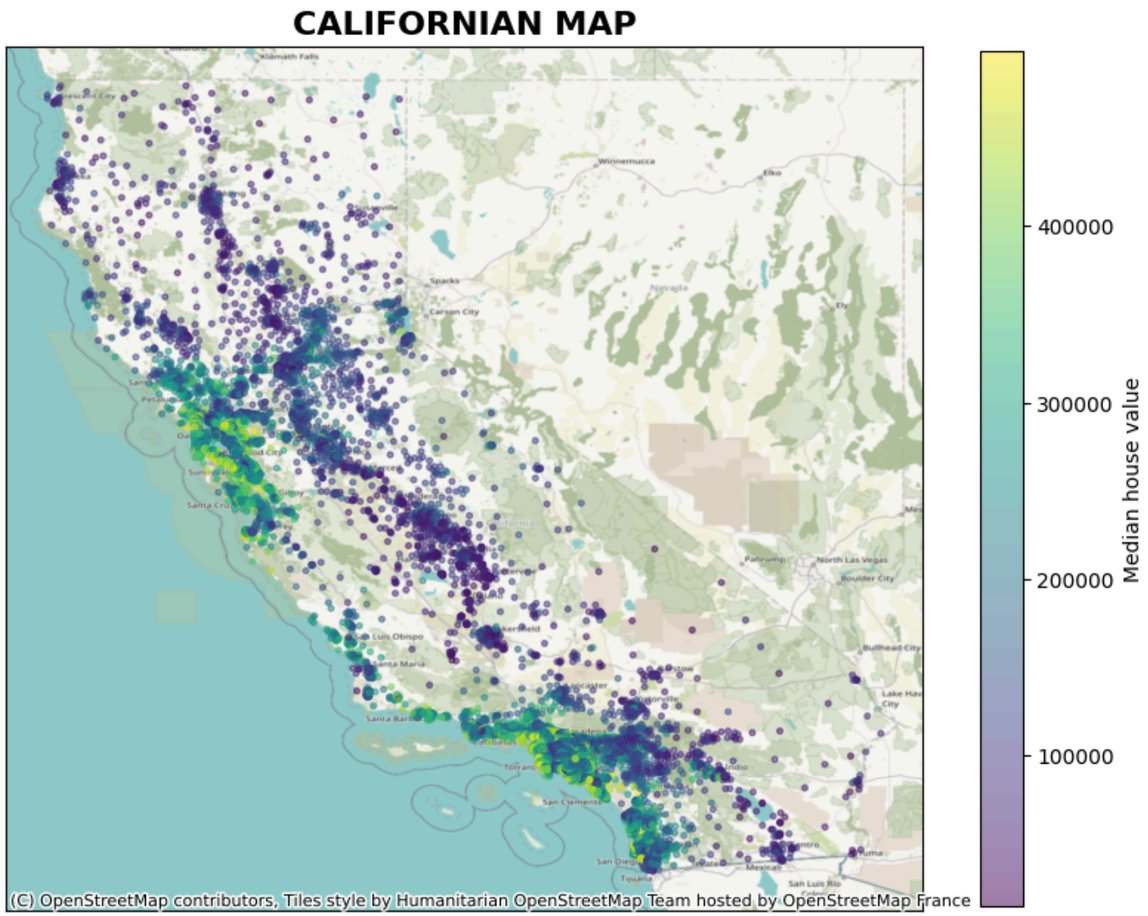
The data contains data found in a given California district and some summary stats about them based on the 1990 census data. The columns are as follows: longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, ocean proximity.

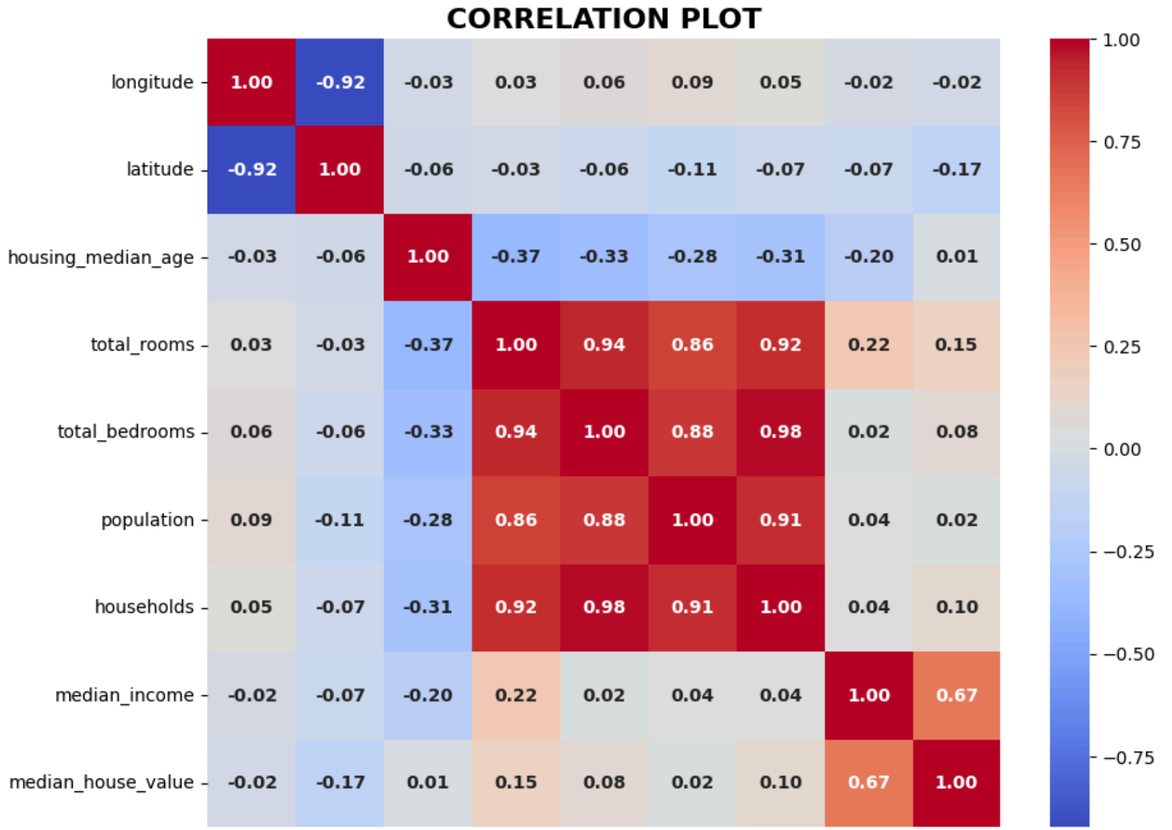
We cleaned the dataset by removing the outliers and removing the points that are located on island.

Kaggle Link

VISUALISATION

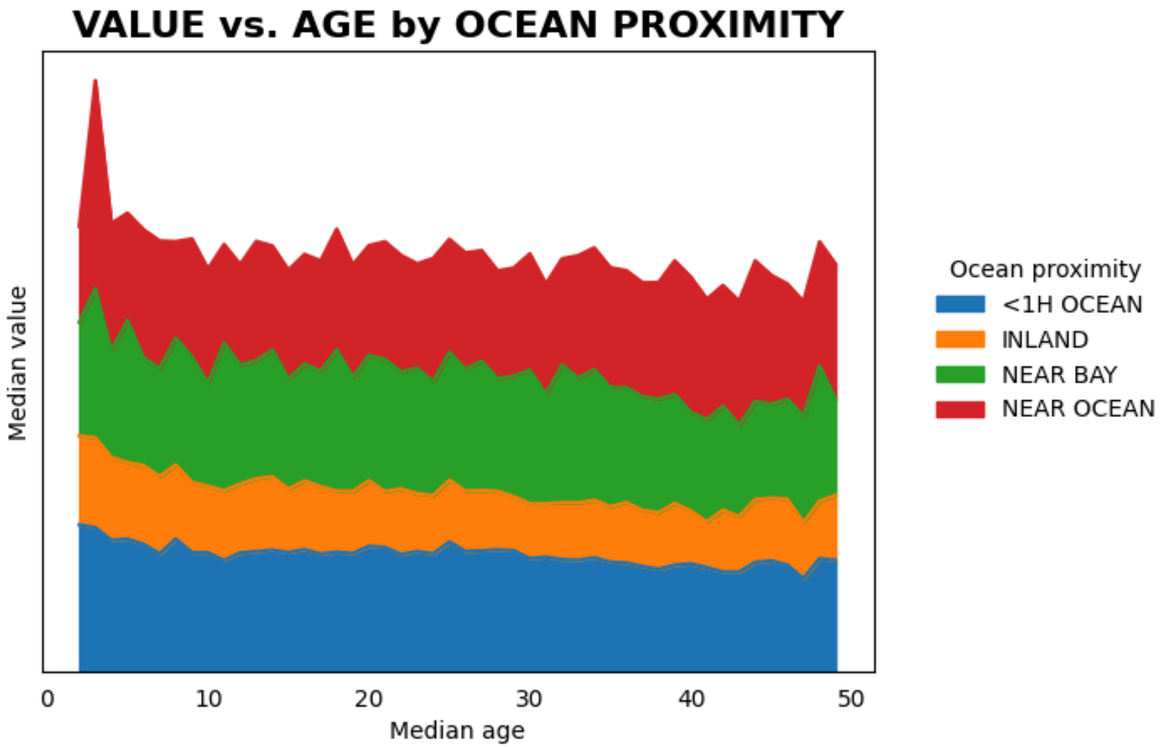
Map of the locations
and their median value





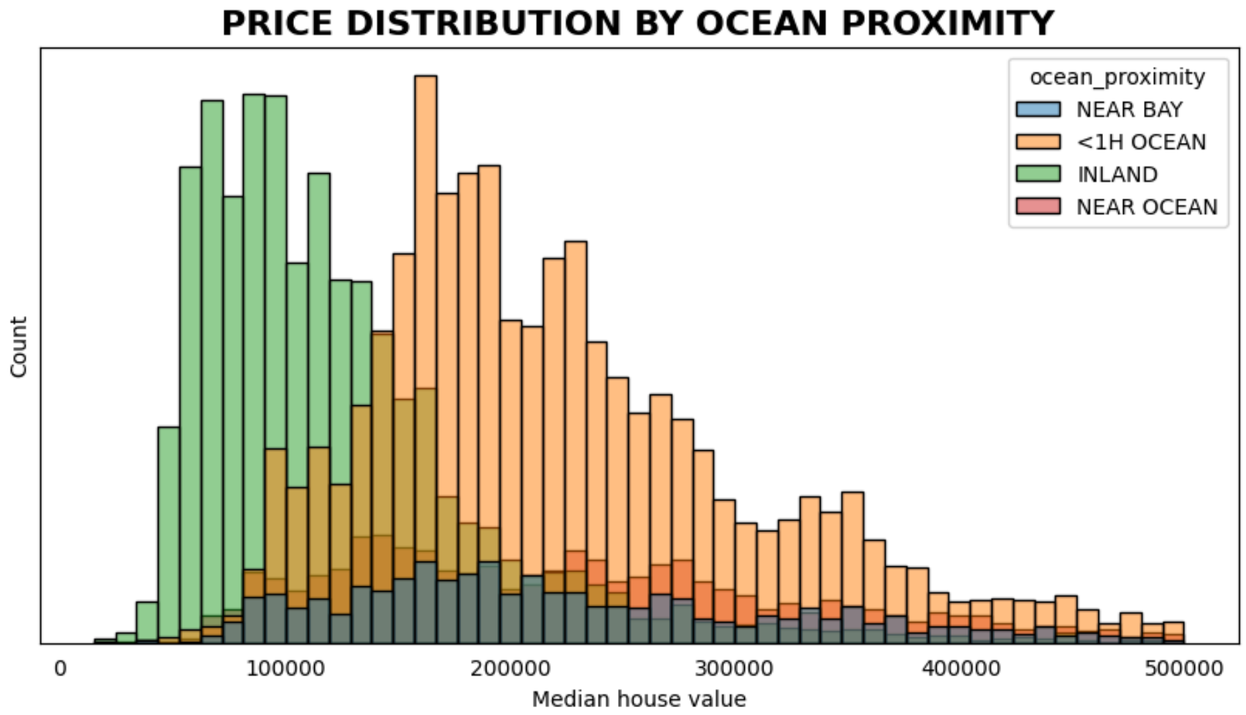
Correlation matrix of the dataset

VISUALISATION



Skew on the median age and value repartition

Importance of the ocean proximity



We tuned a Random Forest and an XGBoost model and the most performant turns out to be **Random Forest**.

Random forest

```
# Perform hyperparameter tuning using GridSearchCV
param_grid = {
    'max_depth': [None, 5, 6],          # Maximum depth of a tree
    'min_child_weight': [5, 10, 15],    # Minimum sum of instance weight needed in a child
    'gamma': [0, 0.1],                  # Minimum loss reduction required to make a further partition on a leaf node of the tree
}
```

```
{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
```

XGBoost

```
# Define the parameter grid to search
param_grid = {
    'n_estimators': [50, 100, 200],  # Number of trees in the forest
    'max_depth': [None, 10, 20, 30],  # Maximum depth of the tree
    'min_samples_split': [2, 5, 10],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 2, 4]  # Minimum number of samples required to be at a leaf node
}
```

```
{'gamma': 0, 'max_depth': 5, 'min_child_weight': 15}
```

R-SQUARED & RMSE

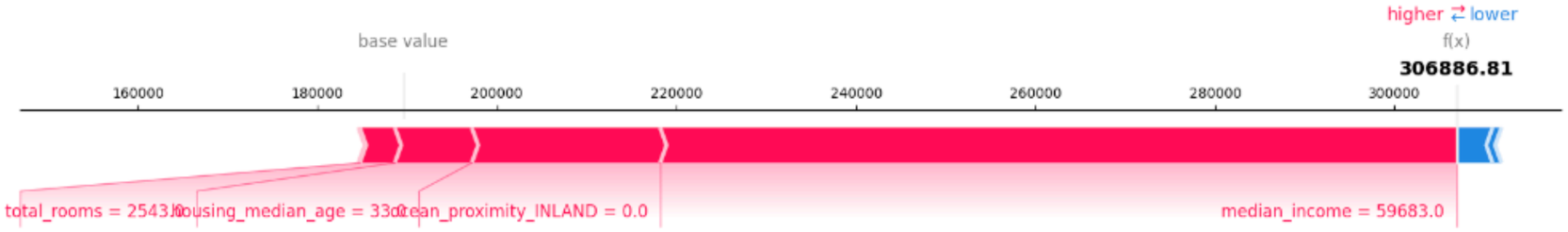
```
============  RANDOM FOREST  ==============
Root Mean Squared Error: 59771.948
R-squared: 0.598
=========== LINEAR REGRESSION  =============
Root Mean Squared Error: 61372.484
R-squared: 0.576
===============  XGBoost  ================
Root Mean Squared Error: 67570.904
R-squared: 0.492
```

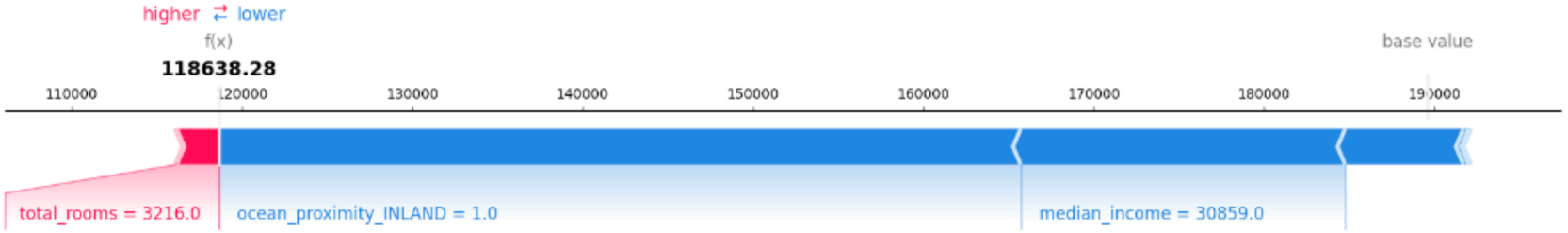
RANDOM FOREST
LINEAR REGRESSION

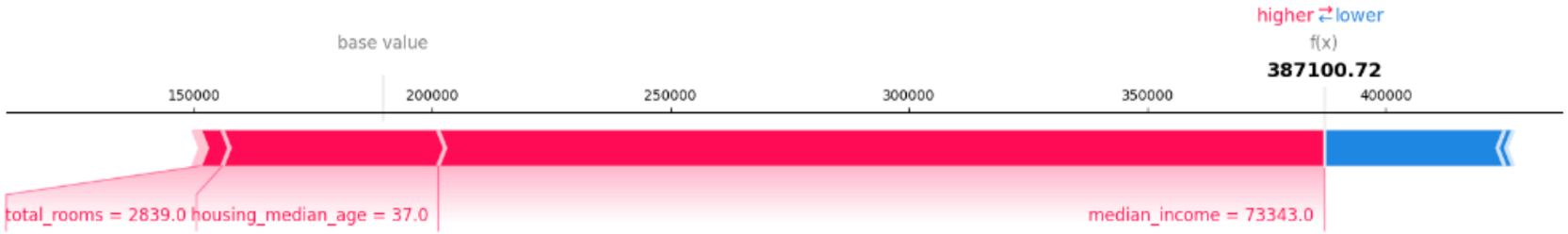
SHAPLEY VALUES

FORCEPLOT EXAMPLES

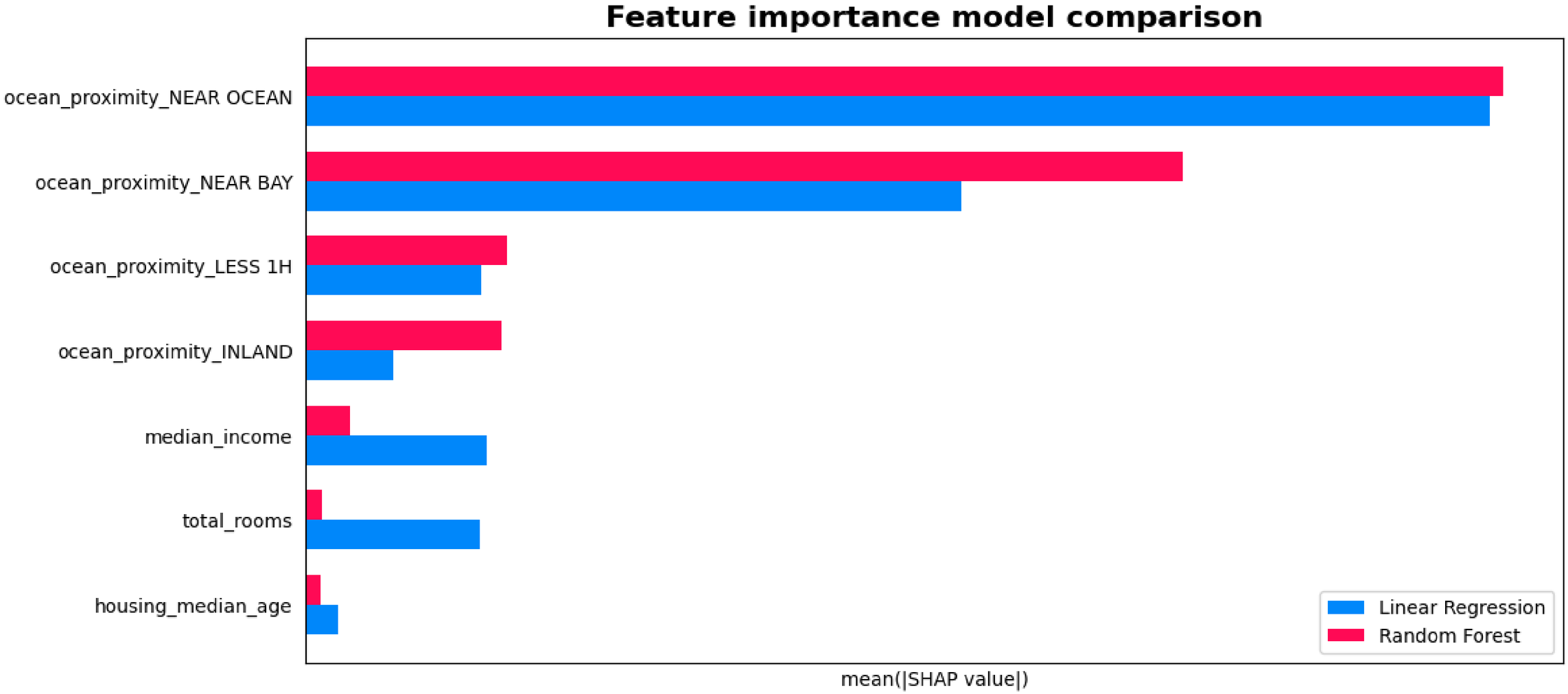
RANDOM FOREST



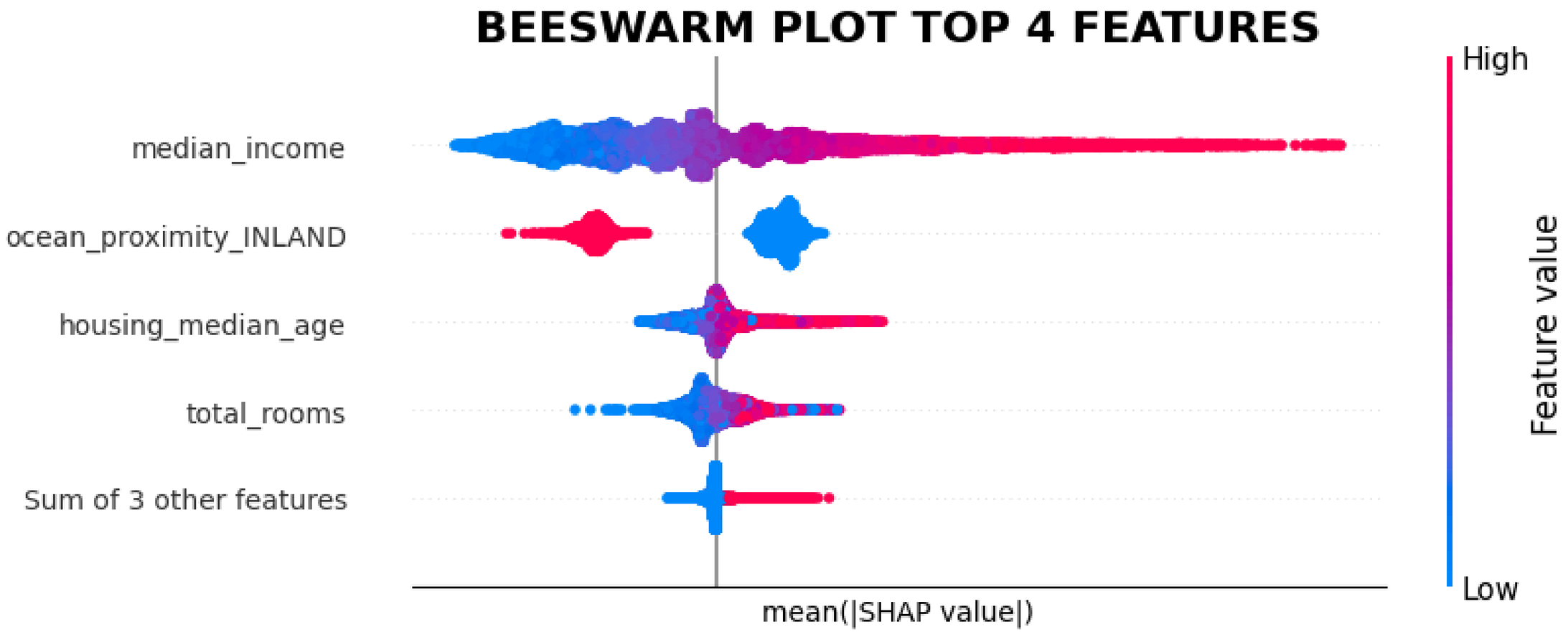




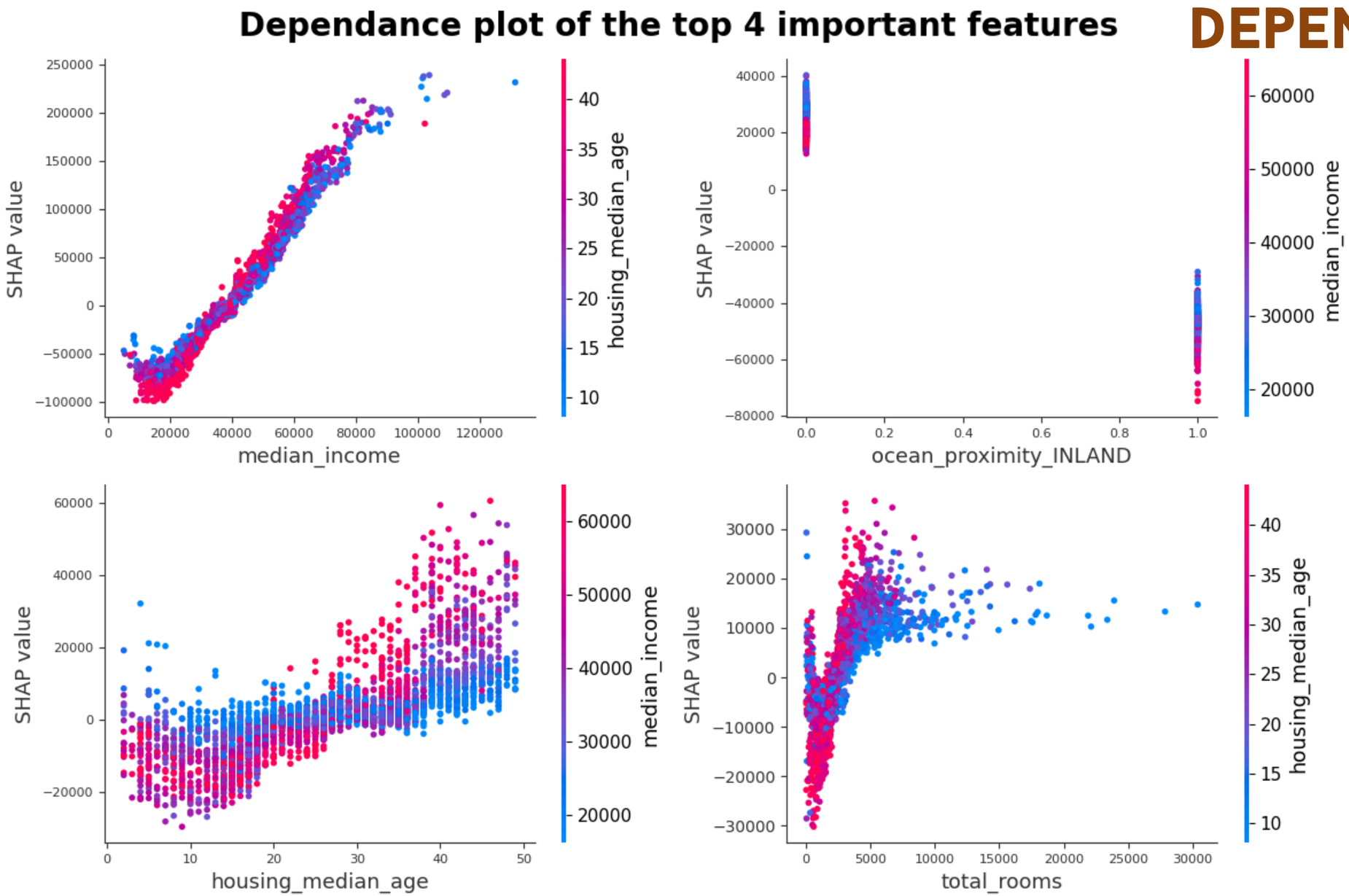
THE MOST IMPORTANT VARIABLES



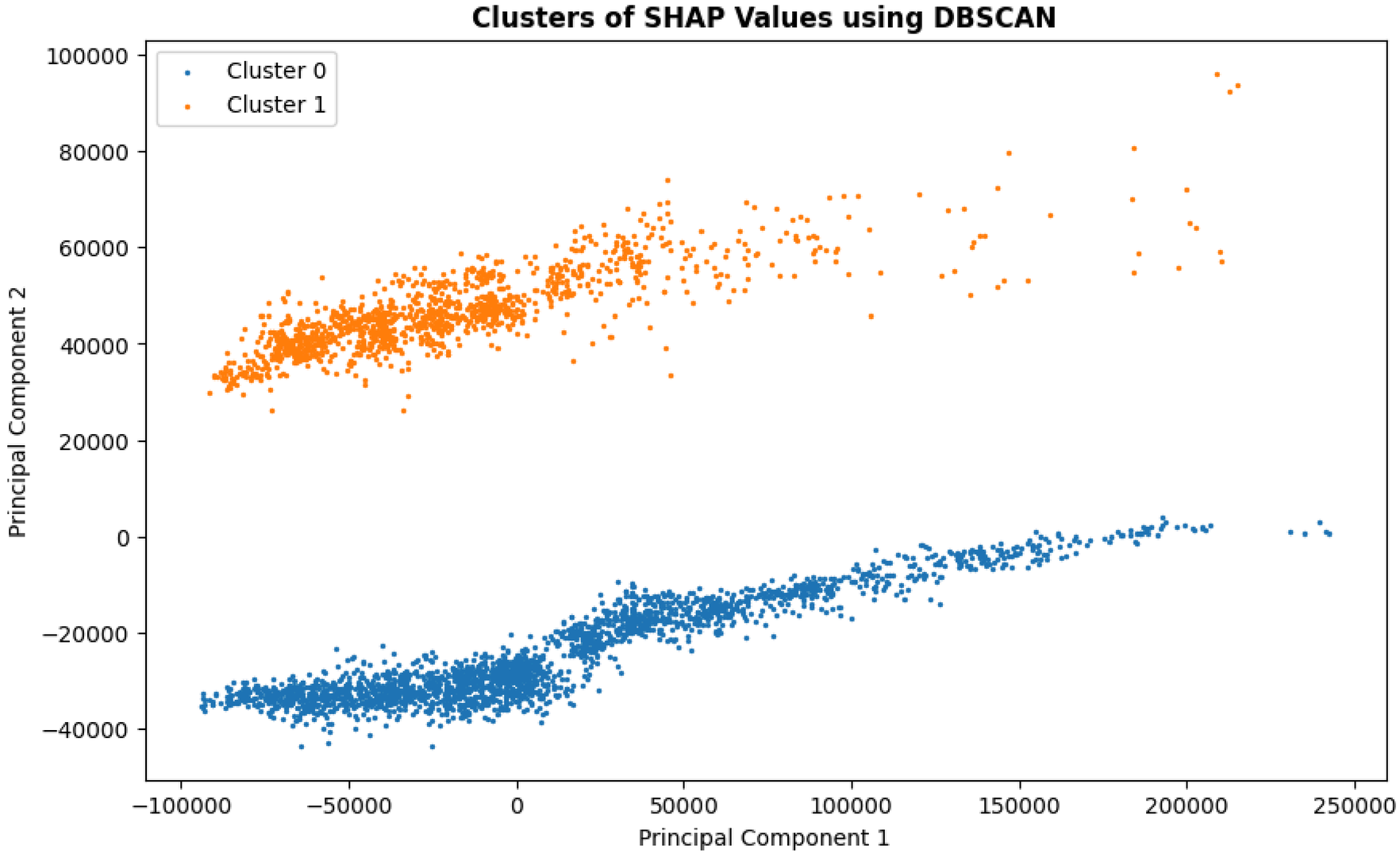
BEESWARM PLOT



DEPENDANCE PLOT



CLUSTERING SHAPE VALUES DBSCAN



THANK YOU

Further details on the GITHUB Repo