Presented by Becca|Nalicia|Will|Chris

PREDICTIVE HOUSING VALUE
BOSTON, MA

An overview of our journey working through the final project of the UofR Data Analytics Bootcamp

PROJECT DEFINITION

Work together as a team to create a machine learning model

OUR TEAM GOALS

Goal #1

Choose a useable Dataset

Goal #2

Incorporate interactive visual components

Goal #3

Design a machine learning model that produced accurate predictive results

Goal #4

Work successfully in a virtual environment

COMMUNCATION CHANNEL FOR VIRTUAL TEAM WORK

Scheduled Class Time

main collaborative work sessions

Slack Group Chat

-Ideas

-Links/Documents

- Team Meeting Summaries

Github

-Technical Work

If needed, additional zoom meetings for group work

ROADMAP

Find a workable dataset

Determine Features to build Machine Learning Model on

Test and Train a Linear Regression Model

- User to choose preferences
- Model provides median price of home
- Model provides a heatmap for full range of house values

PICKING A DATA SET

Our First Attempt

01 NOAA Storm Events Dataset

- Question: *Safe Places to live*
- Machine Learning Model?
  Time-Series Analysis

Ultimately, not enough consistent data to apply a reliable time-series analysis model.

Let's pivot!

PICKING A DATA SET

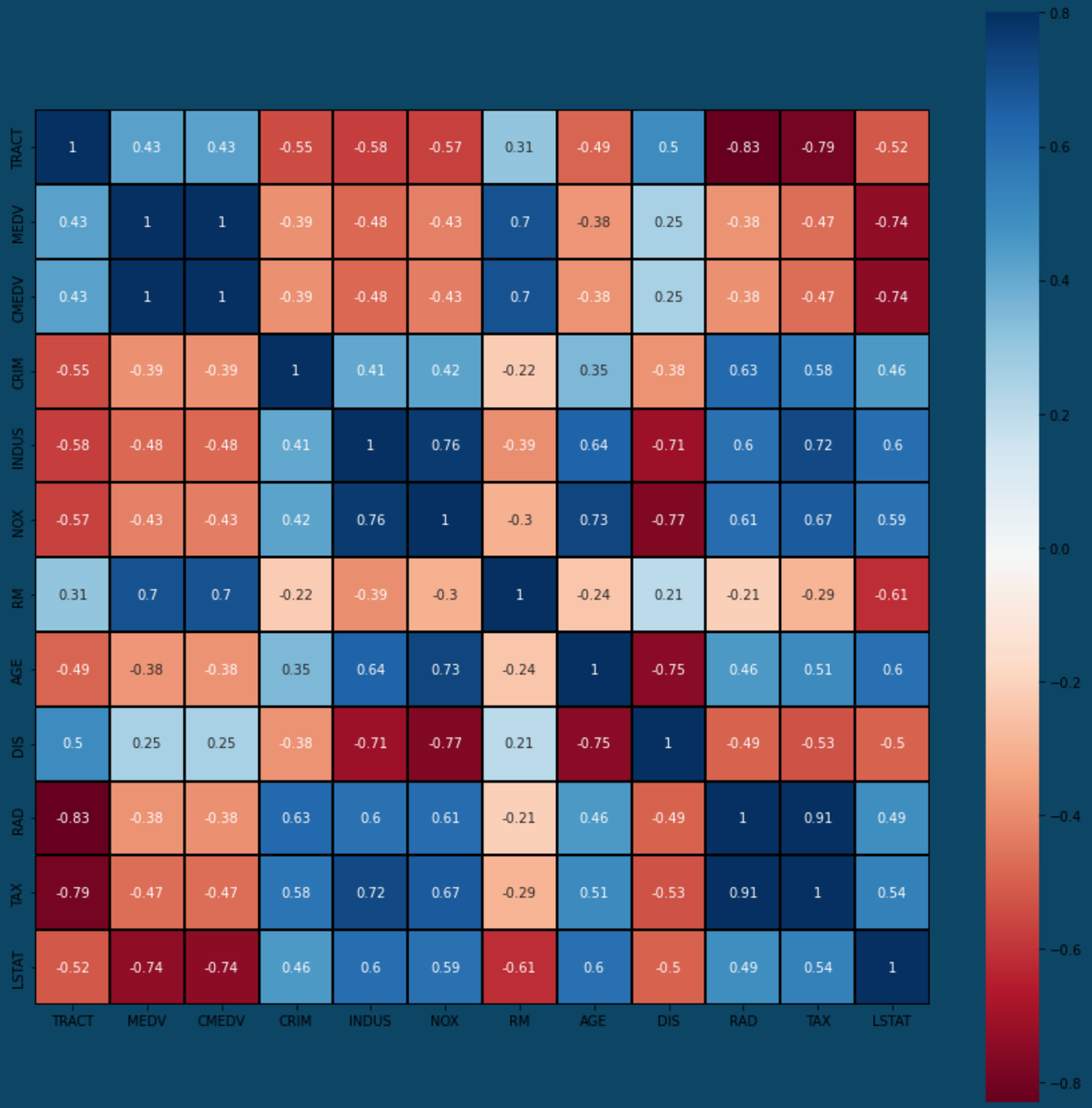
Our Second Attempt

02 Boston Housing Dataset

- Previously proven dataset
- Clear Defined Features
- Machine Learning Model?
  Linear Regression

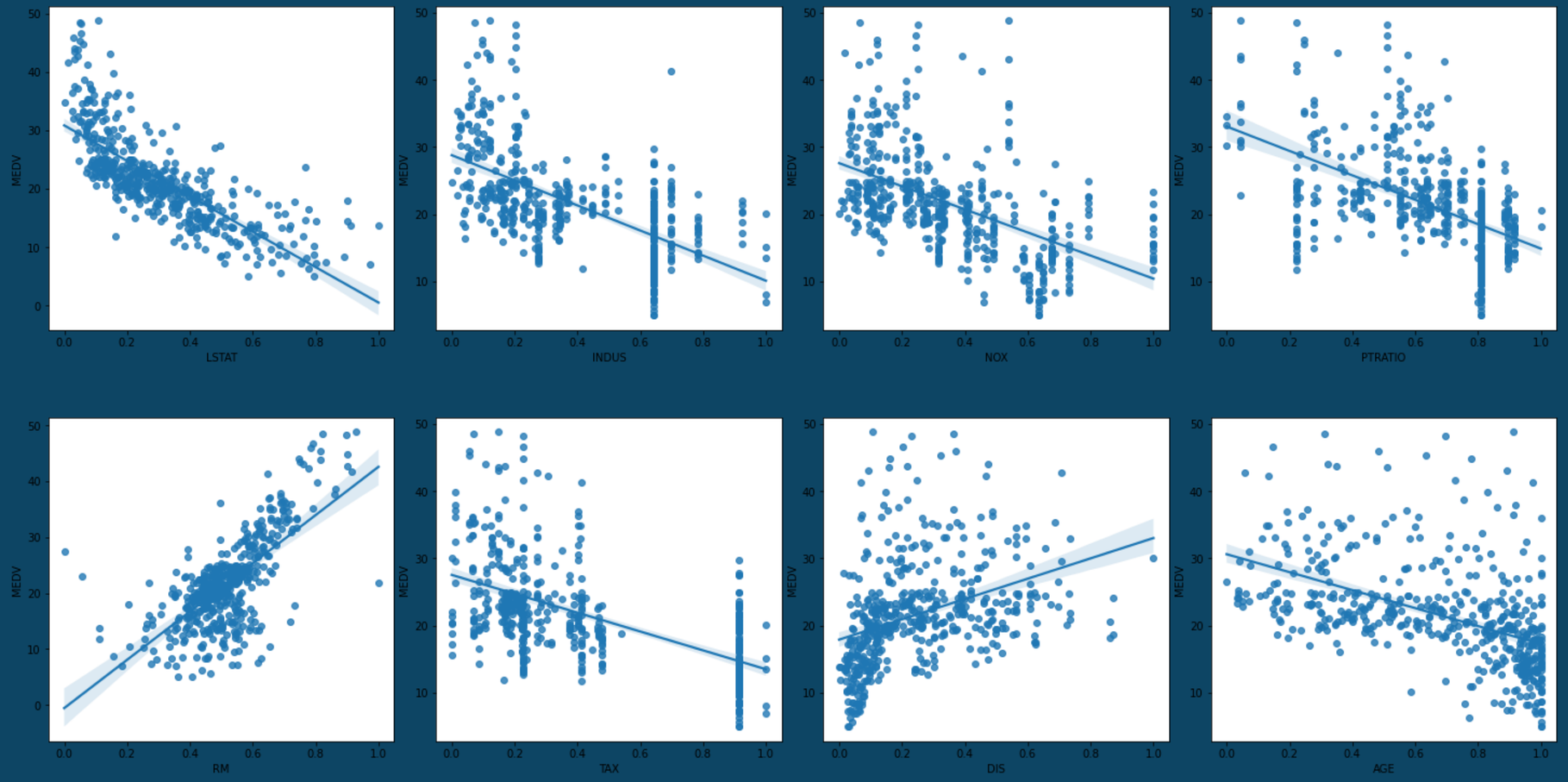
DATA EXPLORATION

- Clean Data (nulls, outliers, etc)
- Pearson Standard Correlation Heatmap



DATA EXPLORATION

- Linear Collerations



DATA EXPLORATION

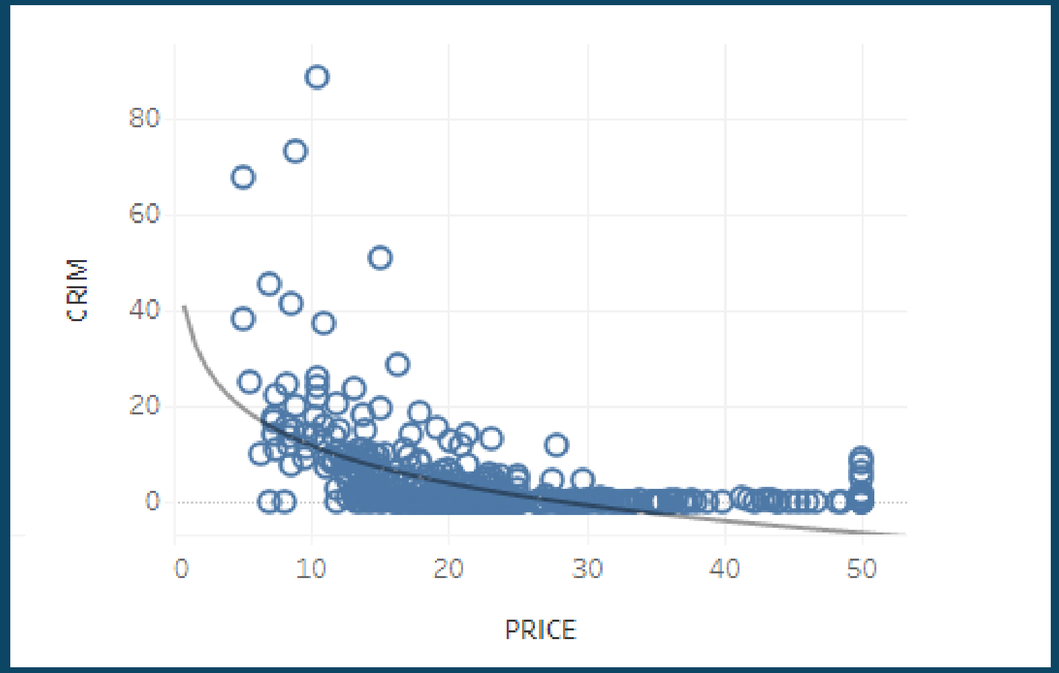
Impactful features to add to our Machine Learning Model

- Crime
- Adjacent to River
- Poverty Level
- Close to employment centers
- Single Family Homes
- Accessibility to highways
- Pupil-Teacher Ratio
- Rooms per Dwelling
- NO2 Concentrations

CRIME AS THE DEPENDENT VARIABLE

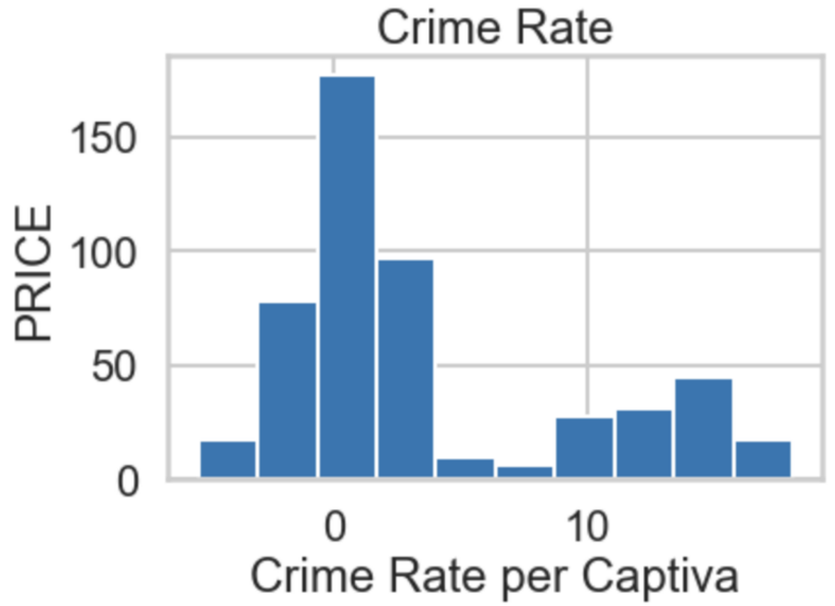
SOME OF US ARE PARENTS AND FROM PAST EXPERIENCE A MAIN FACTOR FOR CHOOSING WHERE TO LIVE IS THE CRIME RATE.

THERE IS WILLINGNESS TO PAY MORE FOR A HOUSE AND SACRIFICE ELSE WHERE, IF THAT MEANS LIVING IN A SAFER NEIGHBORHOOD. WITH THIS MINDSET WE TOOK A CLOSER LOOK INTO THE CRIME RATE.



Using Linear Regression models we can see that the higher the price of the house the lower the crime rate is.

```
plt.hist(lm.predict(X))
plt.title("Crime Rate")
plt.xlabel("Crime Rate per Captiva")
plt.ylabel("PRICE")
plt.show()
```

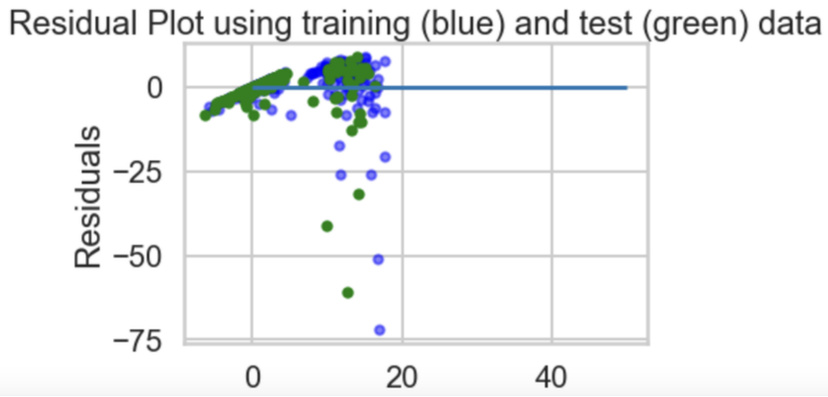



```
Mean squared error with test data: 50.42
Mean squared error with train data: 37.90
```

```
# Residual plots

plt.scatter(regr2.predict(X_train), regr2.predict(X_train) - Y_train, c='b', s=40, alpha=0.5)
plt.scatter(regr2.predict(X_test), regr2.predict(X_test) - Y_test, c='g', s=40)
plt.hlines(y = 0, xmin=0, xmax = 50)
plt.title('Residual Plot using training (blue) and test (green) data')
plt.ylabel('Residuals')
```

```
Text(0, 0.5, 'Residuals')
```



CHOOSING FEATURES TO BASE ML MODEL ON

01. Linear Regression

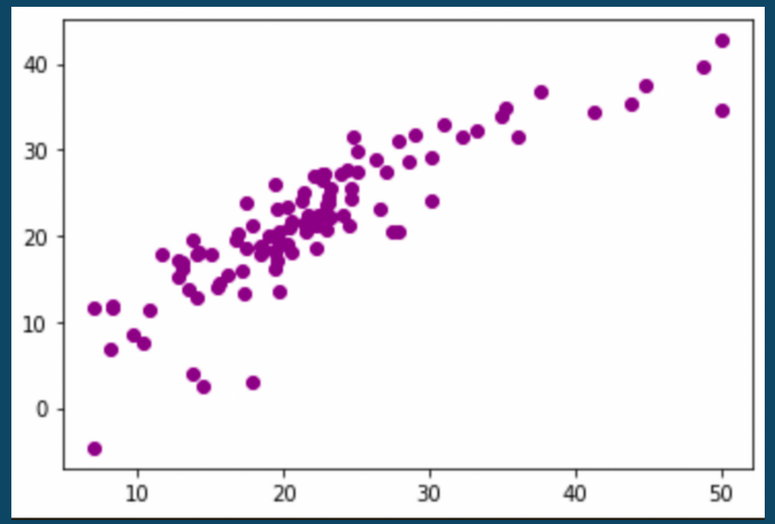
Linear regression is a simple machine learning technique that helps us predict.

02. R scores

The R2 score tells us how much influence the independent variables have on the dependent variable.

EXPLORING VARIABALES

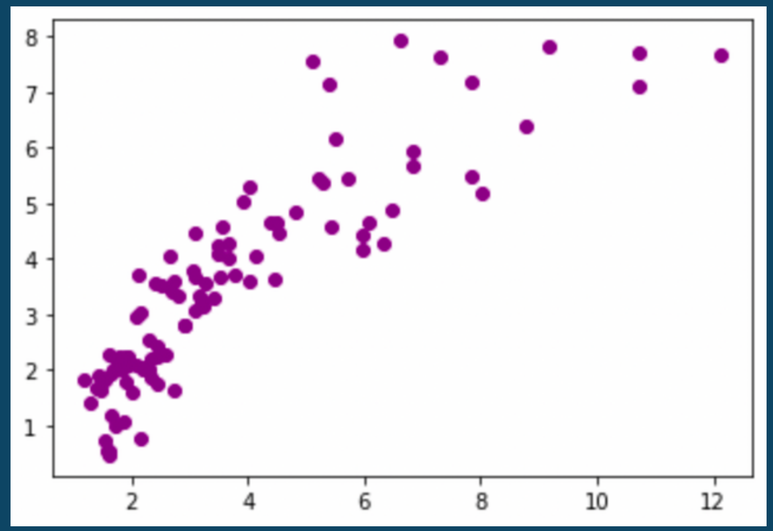
MEDV




```
Performance for training set
R2 score = [ 0.7205589620756452 ]
RMSE = [ 4.899434613036799 ]


Performance for testing set
R2 score = [ 0.7457528698651539 ]
RMSE score = [ 4.4616130967660155 ]
```

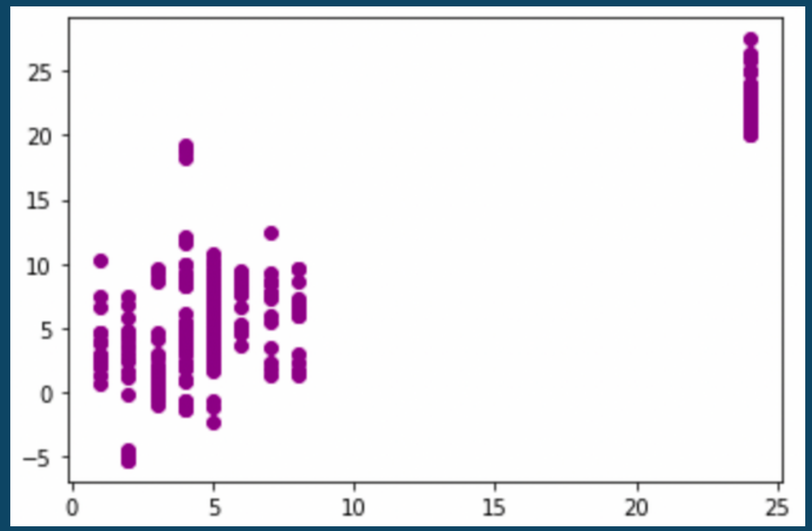
DIS




```
Performance for training set
R2 score = [ 0.7718772068094518 ]
RMSE = [ 0.9808881221513807 ]


Performance for testing set
R2 score = [ 0.7682787527628805 ]
RMSE score = [ 1.099679325800086 ]
```

RM




```
Performance for training set
R2 score = [ 0.8981116636289589 ]
RMSE = [ 2.803635857692753 ]


Performance for testing set
R2 score = [ 0.8596982364477532 ]
RMSE score = [ 3.249148486798341 ]
```

MACHINE LEARNING MODEL

Reasons for design

01 Model Choice: Linear Regression

02 Reason for data split at .8 for train and. 2 for test

03 Results, including final accuracy score

DATA PRESENTATION

Idea #1

Allow user to generate new scenario by exposing feature variables as parameter values, real-time response

Idea #2

Visualize as mapped data at the census tract level

DATA INTEGRATION

TabPy Analytics Extension - allows user to execute Python scripts. Regression model was deployed to local TabPy server.

```
: #fold regression model into python function
def SuggestHomeValue(CRIM, ZN, CHAS, NOX, RM,
                     DIS, RAD, PTRATIO, LSTAT):
    X = np.column_stack([CRIM, ZN, CHAS, NOX, RM,
                         DIS, RAD, PTRATIO, LSTAT])
    #X = scaler.transform(X)
    pred= LR.predict(X)
    pred=pred[0]

    return pred
```

```
# Connect to TabPy server using the client library
from tabpy.tabpy_tools.client import Client
connection = Client('http://localhost:9004/')
connection
```

```
# Publish the SuggestHomeValue function to TabPy server so it can be used from Tableau
# Using the name DiagnosticsDemo and a short description of what it does
connection.deploy('HomeValue',
                  SuggestHomeValue,
                  'Returns est median home value trained on the boston housing dataset', override = True)
```

DATA INTEGRATION

TabPy Analytics Extension - use Tableau calculated fields and script functions to query endpoint on tabpy server

Deployed Models:


```
{
    "HomeValue": {
        "description": "Returns est median home value trained on the boston housing dataset",
        "type": "model",
        "version": 1,
        "dependencies": [],
        "target": null,
        "creation_time": 1650651043,
        "last_modified_time": 1650651043,
        "schema": null,
        "docstring": "-- no docstring found in query function --"
    }
}
```

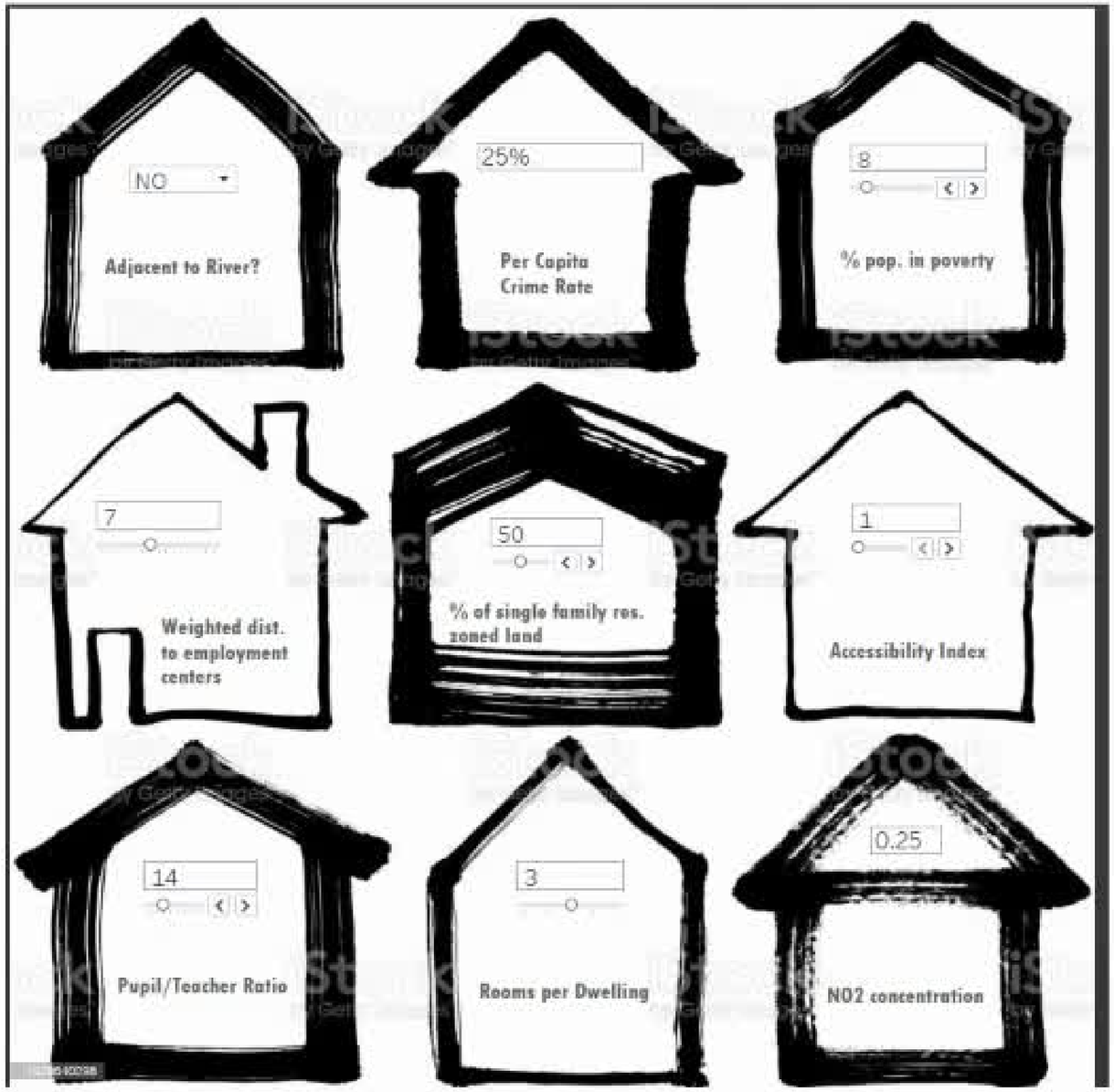

Estimate | bost_data_for_R

Results are computed along Table (across).

```
SCRIPT_REAL("return tabpy.query('HomeValue', arg1, arg2, arg3, arg4, arg5, arg6, arg7, arg8, arg9)['response']",
[pCRIM], [pZN], [pCR], [pNOX], [pRM],[pDIST], [pRAD], [pPTRATIO], [LSTAT])
```

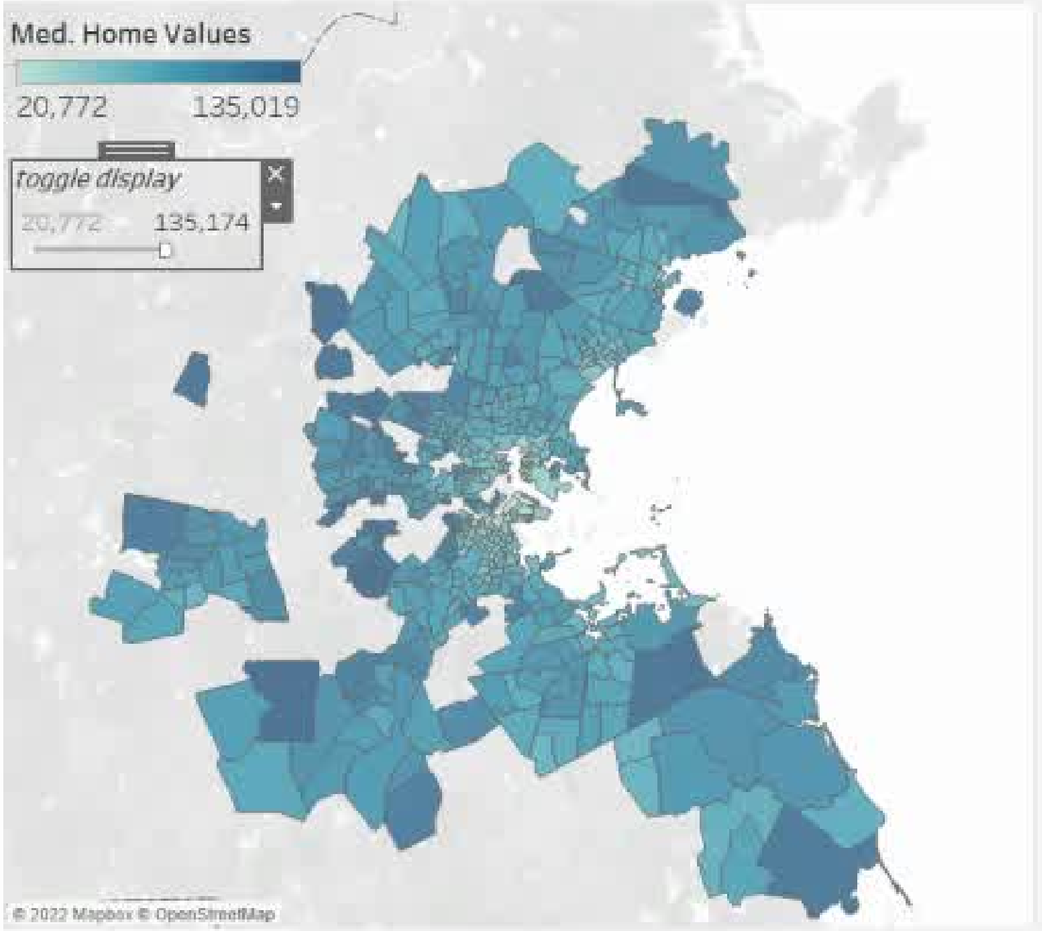
DATA INTEGRATION

Tableau Data Blending - allows for querying and aggregation of data on the fly without storing new data structure. Results are presented visually on a sheet.



Home Value Est.

$ 66,471



TECHNOLOGY USED OR WORKED WITH

-Postgres
-pdAdmin
-AWS DB
-Orcale

-Tableau
-Github
-Canva
-Zoom

-Python
-Pandas
-Jupyter Notebook
- Pearson Standard Correlation Coefficent
- R

-Linear Regression
-Time-Series Analysis

MEET OUR TEAM

REBECCA HINKLE

WILL WILSON

CHRIS NEWELL

NALICIA TILLMAN