Data sources

Training data Kaggle <https://www.kaggle.com/anthonypino/melbourne-housing-market?select=Melbourne_housing_FULL.csv>

Testing data Domain.com – scraping https://www.domain.com.au/

Victoria crime

Steps:

1. Obtain training dataset
   1. Data cleaning (drop non relevant columns, null values, merging dataframes)
2. Scraping test data from domain.com
   1. Retrieve page with chrome driver manager and get URL
   2. Parse URL into beautiful soup to scrape data
   3. Initialise browser / return house features
3. Setting up model
   1. multiple linear regression model using sklearn
      1. read in csvs
      2. correlation matrix
      3. assign data to x and y values
      4. convert categorial data (suburbs) to numbers through one-hot encoding/get\_dummies
      5. scaling and normalisation
      6. Fit the Model to the scaled training data and make predictions using the scaled test data
4. Create Flask app to create routes to the relevant data points used
   1. Creating routes for index.html, tables/visualisations (if any?)
      1. Route to render index.html
      2. Route to trigger the scrape function ( parse in Domain URL and scrape info from website)
5. Prepare web page
   1. Page with predictive data
   2. Page with data sources
6. Heroku deployment

Caveats

* Dropped columns that had nulls/NAs.
* Bedrooms 2 came from other source to ignored and used number of rooms instead
* What date to scrape and pass in to the model for the prediction. Not all listings are auctions so we don’t know the date that they will sell. Using the date the prediction is made instead.
* Use bootstrap templates / include reference/credit on webpage