Can Data Science Help Guide Me in My Journey to Purchasing a House

Create a Classification Data Science model (C.50, Neural Network, CART, random forests) that generates me a decision tree for my real world quest: What attributes should I look for when trying to purchase a House?

# Things I’d be focusing on: Using Beautifulsoup

## Web Scrapping

[Web scrape](https://www.edureka.co/blog/web-scraping-with-python/) Data from Zillow using python.

## Develop Databases (DB): SQL vs NoSQL

Export web scraped data from Python and import into DB’s: MongoDB/MySQL

Create tables and relationships (if applicable)

Query/Play around in databse… see efficiency from both.

## Extract (from DB), Manipulate and create a new dataset using python: **Is this a data pipeline??? Or is there a way to transform this into a data pipeline (automate it)?**

Export and import data from DP to python

Use Queries to get what I want on Python and create 1 clean dataset: ETL if needed

🡪 If Null Values, take care of them

🡪 Dropping or creating new columns

## EDA of project:

* [Simple countplots of Numeric and Categorical Variables](https://towardsdatascience.com/how-to-perform-exploratory-data-analysis-with-seaborn-97e3413e841d)
* Make a Correlation plot to see which Numeric variables have high correlations
* [Make a Correlation plot to see which Categorical variables have high correlations](https://medium.com/@ritesh.110587/correlation-between-categorical-variables-63f6bd9bf2f7#:~:text=Chi%2DSquare%20test%20of%20independence,categorical%20variables%20is%20easily%20available.)
* Plot Numeric variables that have high correlation
* [See if any Numeric and Categorical variables are significant (P-values… Anova)](https://thinkingneuron.com/how-to-measure-the-correlation-between-a-numeric-and-a-categorical-variable-in-python/)

## Choose which variables to choose for model (Go’s hand and hand with EDA)

## Create Models

* Since this more of a classification type of Machine Learning Model Learn about those types of models:
  + Decision Trees:
    - C.50 (REFRESH)
    - CART MODELS (REFRESH)
    - Random Forest (NEW)
    - Boosting (????)
    - Logit Reg

## Optimizing Results

Look at how to optimize Model ROCR/AUC Curve

## (Maybe) Include an interactive ARCGIS MAP that shows all the houses, prices etc.….:

Would be cool to implement this in my EDA

# Notes/Brainstorm Ideas

**Pipelines**

Script to scrape data once a day

Run analysis once a night

CSC 480: Independetn Study Objectives:

**\*A pipeline is just the automation of a process\***

Web Scrape Using Pandas Beautiful Soup

Extract The Webscraping data, Import it into a DB (MongoDB/MySQL) and Create a DB

Develop a mongo Database and work with python to analyze data from that database.

Extract the information from the DB using Python

Do Data Manipulation, Cleaning, and other things on Python

Do an EDA of project DATA:

* **univariate non-graphical**
* **multivariate non- graphical**
* **univariate graphical**
* **multivariate graphical**

**ADHD:**

* Can my model predict people who have ADHD based on certain predictor variables?
* Any data set revolving around ADHD

**WHERE ATTRIBUTES ARE ASSOCIATED WITH BEING AN ECSU STUDENTS FROM 2017-2022:**