# Abstract:

For my data science project, I decided I would get a head start on life and start looking for houses. The way that I will be conducting this experiment is: I will: use [Zillow as my primary source to scrape web information](https://scrapfly.io/blog/how-to-scrape-zillow/).

**Hopefully,** I will create a script using the ‘mac terminal crontab’ so that every day at XX:00, my laptop automatically conducts an automated web scrape. (Pipeline)

I will then export my scraped data from python and important it into 2 databases MySQL (SQL) & MongoDB (NoSQL). The database part of my project will serve as another research question for my Computer Science Senior Research class: Which type of database (SQL VS NoSQL) is more efficient?

I will learn a lot about databases: Constructing, maintaining, creating relationships, queries, etc.…

Once I have constructed this database, I will try to see if I could learn how to script how to automate the exportation of my datasets from python and importation to my database.

Moving on, I will then use python to extract information from my databases to conduct an EDA.

When I finish doing an EDA, I’ll see which variables (numeric) and Exploratory variables have the most correlation. From there I will create a classification model - most likely a decision tree (CART, C.50, Logistic Reg, Random Forest, etc…) – to see what attributes are most associated with my price range when looking to buy a house. I would be playing with the Complexity and Accuracy of the models.

# Attributes for houses:

To not look at thousands upon thousands of homes across Connecticut, I will be looking at all the **houses listed and recently sold on Zillow with a 15-mile radius of Hartford.** This is a crucial filter for me because I want to make sure I do not live more than 15 miles away from the city where I will be working.

To do this I’m going to have to do a Geospatial Query of Connecticut. It will be a \_\_\_\_ for the city of Hartford.

I will grab as many attributes as possible. Whatever Attributes Zillow throws out at me, I want to make them as consistent as possible across my dataset.