**Exercise 6.1**

1. Create a “Data Source” section in your project document and provide the following information:
   * A summary of your data source. We recommend you revisit [Exercise 1.4: Sourcing the Right Data](https://careerfoundry.com/en/steps/sourcing-the-right-data) for a recap on what to include in your summary.
   * An explanation for why you’ve chosen this data set.

**Data Source**

The data set that is being used for this project is open data provided by Zillow. These listings were scraped directly from Zillow by connecting API calls to examine average housing prices across Utah. The data is trustworthy as it was pulled from Zillow on 4/25/2024.

The data set contains nearly 4000 rows of information with several columns that detail homes. It includes things such as the price of the home, what kind of home it is, where it is, how many bathrooms and bedrooms it has and the living area. Since the data set has information like the city as well as longitude and latitude, we’ll be able to find the average prices of homes in each city to compare pricing across the state of Utah.

I have personally chosen this data because I have interest in the Utah housing market. Utah is currently ranked within the top 10 states that have the highest average housing pricing. This data set will be a great project since it will allow me to do an analysis on multiple locations within Utah to determine if high house pricing is within certain parts of Utah.

**Data Profile**

1. **Clean your data.** Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.
2. **Understand your data.** Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis. You might want to make a data profile like what you did in Achievement 1.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Attaching the Jupyter notebook for reference for step 6.

A table of text with black text

Description automatically generated with medium confidence

There wasn’t much change regarding the column names. We typically just changed all the uppercase to lowercase and added “\_” to separate words but most columns remained the same.

1. **Consider limitations and ethics.** Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected.

One such limitation on this data is the size. These are only the homes that are on sale as of April 25th. Since we only have homes that are currently on sale, we can consider this as a sample set. We have chosen to opt out of manufactured homes (mobile homes), apartments and listings for land.

The data provided was submitted to Zillow by the owners of the home or by a real estate agent that is working with the owners.

1. **Define questions to explore.** In a third section of your project document, define a list of questions to explore with your analysis. As mentioned in the Exercise, you may want to revisit [Exercise 1.2: Starting with Requirements](https://careerfoundry.com/en/course/data-immersion/exercise/starting-with-requirements) for a recap on writing good questions.

Which cities in Utah Have the highest and lowest average priced homes?

Are most homes that are on sale new or old?

Which cities have the highest and lowest homes for sale?

“Zestimate” is the Zillow estimate of the home. Are most homes lower or higher than the Zestimate amount?

Why do some cities have a higher average?