Exercise 6.1

5. 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 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

6. **Clean your data.** Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.
7. **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.

Out[8]:

	zpid	zipcode	latitude	longitude	price	bathrooms	bedrooms	living_area	zestimate	tax_assessed_value	price_change
count	2.773000e+03	2773.000000	2773.000000	2773.000000	2.773000e+03	2771.000000	2768.000000	2772.000000	2.384000e+03	1.795000e+03	1.297000e+03
mean	5.718922e+08	84251.085828	40.031484	-112.111501	1.271843e+06	3.231685	3.852601	2969.922078	1.057951e+06	7.466256e+05	-2.104654e+04
std	8.111100e+08	286.370139	1.372271	0.714666	2.426021e+06	1.555618	1.488404	1915.794970	1.712701e+06	1.157739e+06	2.060740e+05
min	1.188538e+07	84003.000000	37.000530	-113.649120	1.500000e+05	0.000000	0.000000	1.000000	1.633000e+05	6.900000e+03	-2.000000e+06
25%	6.322383e+07	84060.000000	40.315407	-112.025750	4.700000e+05	2.000000	3.000000	1705.500000	4.630500e+05	3.492000e+05	-3.000000e+04
50%	1.247783e+08	84098.000000	40.549900	-111.894750	6.550000e+05	3.000000	4.000000	2462.000000	6.345000e+05	4.820000e+05	-1.000000e+04
75%	3.479064e+08	84404.000000	40.699500	-111.779106	1.100000e+06	4.000000	5.000000	3804.250000	9.852250e+05	7.645500e+05	-3.100000e+03
max	2.142959e+09	84790.000000	41.764180	-109.519150	5.200000e+07	14.000000	14.000000	17661.000000	2.869490e+07	2.792344e+07	3.275000e+06

Attaching the Jupyter notebook for reference for step 6.

Variables List	Time Variant or Time Invariant	Structured or Unstructured	Qualitative / Quantitative	Data Type: (If Qualitative [Nominal/Ordinal]) (If Quantative [Discrete/Continuous])
zpid	Time Invariant	Structured	Quantitative	Discrete
streetName	Time Invariant	Structured	Qualitative	Nominal
city	Time Invariant	Structured	Qualitative	Nominal
state	Time Invariant	Structured	Qualitative	Nominal
zipcode	Time Invariant	Structured	Quantitative	Discrete
latitude	Time Invariant	Structured	Quantitative	Continuous
longitude	Time Invariant	Structured	Quantitative	Continuous
price	Time Invariant	Structured	Quantitative	Discrete
bathrooms	Time Invariant	Structured	Quantitative	Discrete
bedrooms	Time Invariant	Structured	Quantitative	Discrete
livingArea	Time Invariant	Structured	Quantitative	Discrete
homeType	Time Invariant	Structured	Qualitative	Ordinal
homeStatus	Time Invariant	Structured	Qualitative	Nominal
daysOnZillow	Time Variant	Structured	Quantitative	Discrete
zestimate	Time Invariant	Structured	Quantitative	Discrete
rentZestimate	Time Invariant	Structured	Quantitative	Discrete
isPreforeclosureAuction	Time Invariant	Structured	Qualitative	Binary
isZillowOwned	Time Invariant	Structured	Qualitative	Binary
currency	Time Invariant	Structured	Qualitative	Nominal
country	Time Invariant	Structured	Qualitative	Nominal
taxAssessedValue	Time Invariant	Structured	Quantitative	Discrete
lotAreaValue	Time Invariant	Structured	Quantitative	Discrete
lotAreaUnit	Time Invariant	Structured	Qualitative	Nominal
is_newHome	Time Variant	Structured	Qualitative	Nominal
newConstructionType	Time Invariant	Structured	Qualitative	Nominal
datePriceChanged	Time Invariant	Structured	Quantitative	Discrete
priceReduction	Time Invariant	Structured	Quantitative	Discrete
priceChange	Time Invariant	Structured	Quantitative	Discrete
is_openHouse	Time Invariant	Structured	Qualitative	Nominal
openHouse	Time Invariant	Structured	Qualitative	Binary
open_house_showing	Time Invariant	Structured	Qualitative	Nominal
imgSrc	Time Invariant	Unstructured	Qualitative	Nominal
homeDetailUrl	Time Invariant	Structured	Qualitative	Nominal
price_to_rent_ratio	Time Invariant	Structured	Quantitative	Discrete

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.

---

8. **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.

---

10. **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 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?