6.1: Sourcing Open Data

DATA SET SUMMARY

Data Source	The original data set has been taken from publicly available source of realtor.com. Data source link: https://www.realtor.com/research/data/ (updated on a monthly basis)
Data Collection	The data comes from open source, is administrative and reliable.
Data Contents	The data contains core metrics on residential real estate listings by US state from July 2016 through January 2023.
Data Relevance	The data is recent, refers to the period from July 2016 through January 2023.

Why I have chosen this data set:

I have chosen the data set, since it fits the criteria specified in the Achievement 6 brief: it has geographical and time components, it contains continuous variables and categorical variables can be easily created. The data is also big enough: 4,029 entries. Plus, real estate is an area of interest for me.

DATA PROFILE

Cleaning	• 25 unnecessary columns with missing vales have been removed (in Jupiter notebook). • Last row that did not contain applicable data was removed (in Excel) • 27 missing values for the state of Alaska in each of 'pending-listing-count' and 'pending_ratio' have been kept AS IS to keep the remaining of the data for Alaska available for analysis. • No duplicates have been detected. • No mixed-type columns have been detected. • No outliers have been detected.

Understanding

```
core_met.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4029 entries, 0 to 4028
Data columns (total 15 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   month_date_yyyymm                     4029 non-null   int64
 1   state                                 4029 non-null   object
 2   state_id                              4029 non-null   object
 3   median_listing_price                  4029 non-null   int64
 4   active_listing_count                  4029 non-null   int64
 5   median_days_on_market                 4029 non-null   int64
 6   new_listing_count                     4029 non-null   int64
 7   price_increased_count                 4029 non-null   int64
 8   price_reduced_count                   4029 non-null   int64
 9   pending_listing_count                 4002 non-null   float64
 10  median_listing_price_per_square_foot  4029 non-null   int64
 11  median_square_feet                    4029 non-null   int64
 12  average_listing_price                 4029 non-null   int64
 13  total_listing_count                   4029 non-null   int64
 14  pending_ratio                         4002 non-null   float64
dtypes: float64(2), int64(11), object(2)
memory usage: 472.3+ KB
```

```
core_met.dtypes
```

```
month_date_yyyymm                       int64
state                                  object
state_id                               object
median_listing_price                    int64
active_listing_count                    int64
median_days_on_market                   int64
new_listing_count                       int64
price_increased_count                   int64
price_reduced_count                     int64
pending_listing_count                 float64
median_listing_price_per_square_foot    int64
median_square_feet                      int64
average_listing_price                   int64
total_listing_count                     int64
pending_ratio                         float64
dtype: object
```

	month_date_ yyyymm	*median_listing_ price*	*active_listing_ count*	*median_days_ on_market*	*new_listing_ count*	*price_increased_ count*	*price_reduced_ count*
minimum	201607	129,913.00	727.00	14.00	296.00	-	72.00
maximum	202301	880,000.00	155,274.00	211.00	56,416.00	9,460.00	56,908.00
mean	N/A	329,270.41	18,508.47	66.77	9,216.82	680.45	5,172.68

	pending_listing_ count	*median_listing_ price_per_ square_foot*	*median_square_ feet*	*average_listing_ price*	*total_listing_ count*	*pending_ratio*
minimum	-	79.00	988.00	206,299.00	793.00	-
maximum	82,101.00	697.00	2,798.00	1,719,561.00	208,019.00	3.13
mean	8,782.63	177.66	1,917.17	520,856.60	27,158.08	0.58

Limitations

- 27 missing values for the state of Alaska in each of 'pending-listing-count' and 'pending_ratio' have been kept AS IS to keep the remaining

	of the data for Alaska available for analysis. Please, keep in mind this limitation when involving 'pending-listing-count' and 'pending_ratio' variables for Alaska in your analysis.
Ethics	• The data was sourced from Realtor.com® Economic Research on their official site. The data is open-source and administrative, based on the most comprehensive and accurate database of MLS-listed for-sale homes in the industry. The data set was designed specifically for research purpose in mind and is rather reliable. • The PII was removed from the data set by the data set owner - Realtor.com® Economic Research and is safe for public use.

QUESTIONS TO EXPLORE

1. What are the trends in the last few years on the US real estate market? What is the seasonality?
2. What is the ranking of the US states from the most to the least affordable?
3. Was there any noticeable change in the real estate market during COVID-19?
4. What are the hottest markets in the US?
5. What are the demand trends, and which states have the least demand?