6.1: Sourcing Open Data
Wine Reviews

Data Sourcing
The Wine Reviews dataset that I am choosing to work with was sourced from an external data source:
https://www.kaggle.com/datasets/zynicide/wine-reviews

Data Collection
The dataset was acquired through web scraping methods. It was scraped from WineEnthusiast during the week of June 15th, 2017.

Data Limitations and Ethics
The dataset on Kaggle titled "Wine Reviews" contains information about various wines, including reviews, ratings, and descriptions. While it provides valuable insights into the world of wines, it may also have certain limitations. Some possible data limitations of this dataset include:

1. **Sampling Bias**: The dataset may contain a biased sample of wines, as it relies on reviews from WineEnthusiast. Wines reviewed by this source may not be representative of all wines available in the market, leading to potential biases in the data.
2. **Limited Coverage**: The dataset may not include information about all wines globally. It may be limited to wines reviewed by WineEnthusiast or those available in specific regions or markets, which could limit its generalizability.
3. **Incomplete Information**: Certain variables or fields within the dataset may contain missing or incomplete information. For example, some wines may have missing descriptions or ratings, which could affect the analysis and interpretation of the data.
4. **Subjectivity of Reviews**: Wine reviews and ratings are subjective and may vary depending on individual preferences and biases of reviewers. As a result, the dataset may contain subjective information that may not be consistent across reviewers.
5. **Quality of Reviews**: The quality and reliability of reviews may vary, as they are based on the opinions and experiences of reviewers. Some reviews may be more detailed and informative than others, potentially affecting the accuracy and usefulness of the data.
6. **Limited Variables**: While the dataset contains several variables such as wine variety, region, and price, it may lack certain variables that could provide additional context or insights into the wines, such as production methods, vineyard characteristics, or aging process.
7. **Temporal Bias**: The dataset may be skewed towards more recent wines, as older wines may be less likely to be reviewed or included in the dataset. This could introduce a temporal bias, especially if analyzing trends or changes in wine characteristics over time.

Data Contents
The Wine Reviews dataset consists of 150,930 rows and 10 columns (excluding the indexing column). It contains reviews of various wines with their area of origin, as well as price.

Why I chose this dataset
I first chose this dataset because it fulfills the project brief requirements. Additionally, I chose this dataset because wine tasting has always been a bit of a hobby of mine, so I have personal interest in this particular type of data.

Data Profile

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
country	time-invariant	structured	qualitative	nominal
description	time-variant	unstructured	qualitative	nominal
designation	time-invariant	structured	qualitative	ordinal
points	time-variant	structured	quantitative	continuous
price	time-invariant	structured	quantitative	continuous
province	time-invariant	structured	qualitative	nominal
reion_1	time-invariant	structured	qualitative	nominal
region_2	time-invariant	structured	qualitative	nominal
variety	time-invariant	structured	qualitative	nominal
winery	time-invariant	structured	qualitative	nominal

Data Consistency Checks

Missing Values	Missing Value Treatment
'country' column contains 5 missing values	.00003% missing values in this column. I have decided to remove these missing values as it is an extremely insignificant number.
'designation' column contains 45,735 missing values	This column is not important to my analysis, so it will be dropped.
'price' column contains 13,695 missing values	I will be creating a new variable 'price_missing' to flag these missing values.
'province' column contains 5 missing values	.00003% missing values in this column. I have decided to remove these missing values as it is an extremely insignificant number.
'region_1' column contains 25,060 missing values	I already have my 2 geographical components required for this project within the 'country' and 'province' columns. I will be dropping this column.
'region_2' column contains 89,977 missing values	I already have my 2 geographical components required for this project within the 'country' and 'province' columns. I will be dropping this column.

There are no duplicates present in this dataset.
There is no mixed-type data present in this dataset.

Data Wrangling

Column Dropped	Column Type Changed	Column Added	Comments
description			This column contains data that is too unstructured for my analysis and not needed.
designation			This column is unneeded for my analysis.
	points		I've changed the 'points' column's data type from integer to string because the numbers do not represent real numeric values, but rather a rating scale.
region_1			This column contains several thousand missing values. I already have all the location information I need with other columns present in this dataset.
region_2			This column contains several thousand missing values. I already have all the location information I need with other columns present in this dataset.
winery			This column is unneeded for my analysis.
		price_missing	Approximately 9% of the data is missing in the 'price' column. I decided to create a new variable to flag these missing values. Entries with a missing price value will be flagged as 'True'

Dropped the five missing values present in the 'country' and 'province' columns.

Questions to Explore

Where are the highest-rated wines located?

What varieties of wine are consistently rated highest?

Is there a relationship present between the price of wine and a high rating?

Is there a relationship present between the location of wine and price?

What countries produce the most wine?

- Are there specific provinces within those countries that produce higher-rated wines?

What varieties of wine are produced most commonly?