DSCI510 Final Report

Project Description

This project used wine data collected from Target.com and Kaggle to analyze and compare the relationships between wine characteristics and ratings. The name of this project is how wine characteristics are related to wine ratings. Wines have many attributes such as ingredients, region, and flavor, and people would rate the wines based on these attributes. Therefore, I collected multiple wine attributes and ratings from two data sources. The data contains descriptive data of wines such as price, alcohol percentage, wine variant, etc., given this information, I am interested in figuring out what types of wine attributes contribute to a high or low wine rating. This project is a data analysis project with Python that incorporates data collection, data analysis, and data visualization.

Installation

The dependency and version are recorded in the requirements.txt file. To install the Python libraries needed for this project, enter the following command in terminal:

- pip install -r requirements.txt

Running the Project

The code is stored in Jupyter notebook called final_project.ipynb. The code can be run using the Cell->Run All.

The system will prompt two times to ask to enter the username and user key to download the wine_reviews dataset from Kaggle. Enter the username for the first prompt and enter the user key for the second prompt.

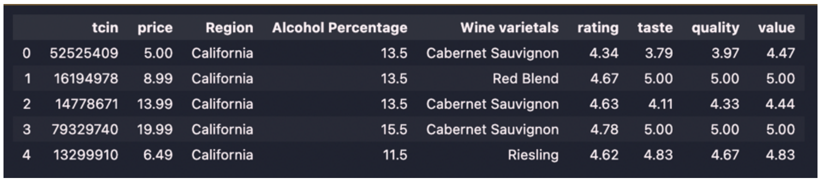
- Username: ruitingma
- User key: 27c4355da483b349890f5213ed6d1c5z

The link to my GitHub repository is https://github.com/ruitingm/WineReviews.

Data Source

For this project, I collected wine data from two different sources with Python.

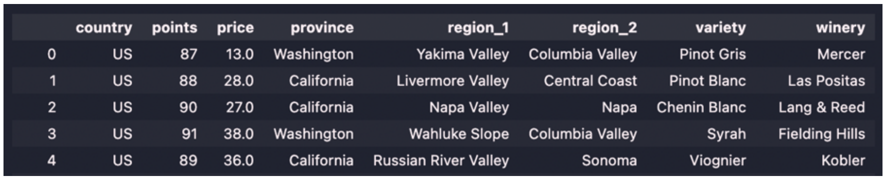
The first dataset is called target_wine where I fetched Target.com's API and collected 1200 wine samples. The target_wine data contains information on wines that Target.com has. The webpage link is https://www.target.com/s?searchTerm=wine&sortBy=relevance&moveTo=product-list-grid&facetedValue=fwtfrZpahdwZxhfwzZal25lff2zppZb6ubZjkr5nZocuu8Z6brr6Zli9hbZvwqqq. Target.com is an online department store that sells various kinds of groceries, clothing, and many other kinds of stuff. Because Target.com does not offer public APIs, I looked at the cURL under the network tab of the inspect function and converted the cURL to a Python request to retrieve the wine data from Target.com. Target listed much information about each wine; however, I only specified 9 of the most common wine attributes. The target_wine data is stored in a CSV file with 9 columns and 1200 rows, the sample target_wine data is stored in the data file. The 9 columns are product number, price, region, alcohol percentage, type, rating, quality, value, and taste. Below is a screenshot of the sample target_wine data.



	tcin	price	Region	Alcohol Percentage	Wine varietals	rating	taste	quality	value
0	52525409	5.00	California	13.5	Cabernet Sauvignon	4.34	3.79	3.97	4.47
1	16194978	8.99	California	13.5	Red Blend	4.67	5.00	5.00	5.00
2	14778671	13.99	California	13.5	Cabernet Sauvignon	4.63	4.11	4.33	4.44
3	79329740	19.99	California	15.5	Cabernet Sauvignon	4.78	5.00	5.00	5.00
4	13299910	6.49	California	11.5	Riesling	4.62	4.83	4.67	4.83

The wine data collected from Target.com has 300 samples less than I planned originally. In the beginning, I planned to collect 1500 samples. To avoid the denial of access to the website, I reduced my samples to 1200 so that I could collect all wine data in a short period. The challenge of retrieving wine data from Target.com is that although I could get wine attributes from the wine browsing page, there are only 24 wines displayed on one page. Since I plan to collect 1200 samples, I had to create a method to loop over each page to get 1200 samples. The method that I created is to use a while loop to loop over each page, for every loop, I change the URL and the page count to submit a new request. I have never done such a thing in the past and I was finally able to fetch all the wine data through trial and error.

The second dataset is called wine_reviews which contains information on wine reviews by wine masters. The original data source is Kaggle and the link is https://www.kaggle.com/datasets/zynicide/wine-reviews?q=wine&select=winemag-data_first150k.csv. The wine_reviews data on Kaggle is in both CSV and JSON format and can be downloaded directly from Kaggle. The original data has about 150k samples with 14 wine review attributes. Many wine review samples contain missing values; therefore, I removed all the samples with missing values and randomly kept 1500 samples. I also removed some columns that I think are not relevant to this project and kept only 8 attributes. The 8 columns are country, points, price, province, region 1, region 2, variety, and winery. Below is a screenshot of the sample wine_reviews data; the wine_reviews data is stored in the data file as a CSV file.



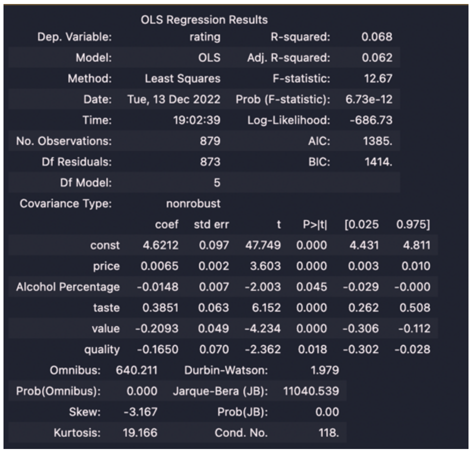
	country	points	price	province	region_1	region_2	variety	winery
0	US	87	13.0	Washington	Yakima Valley	Columbia Valley	Pinot Gris	Mercer
1	US	88	28.0	California	Livermore Valley	Central Coast	Pinot Blanc	Las Positas
2	US	90	27.0	California	Napa Valley	Napa	Chenin Blanc	Lang & Reed
3	US	91	38.0	Washington	Wahluke Slope	Columbia Valley	Syrah	Fielding Hills
4	US	89	36.0	California	Russian River Valley	Sonoma	Viognier	Kobler

The wine_reviews data was retrieved as I planned in the beginning. The challenge regarding retrieving the wine_reviewsdata is that instead of downloading the CSV file from the Kaggle website, I spent a lot of time figuring out how I can use Python commands to automatically download the dataset and the solution was to use the opendatasets library to download the dataset given the link of the data source.

The two datasets are similar in the sense that they both contain several wine characteristics that describe the wines and their respective ratings. The differences are that the ratings are measured on different scales and are rated by different people. The wines in the target_wine data are rated by customers on a scale of 0-5, denoted by rating. The wines in the wine_reviews data are reviewed by wine masters on a scale of 0-100, denoted by points.

Analysis and Methodology

For this project, I conducted statistical analysis to investigate the relationship between various wine attributes and wine ratings. The two datasets that I have are similar in a way that there are some shared attributes (with different attribute names) and some unique attributes and a wine rating. The analysis approach is the same for similar attributes. In addition, I examined continuous and categorical variables separately. I first built a OLS Regression with 4 the numeric attributes in the target_wine data with regard to wine rating. The results show that wine price and taste have a positive relationship and alcohol percentage, value, and quality have negative relationship with wine rating; all four attributes have p-value below 0.05. The OLS results are shown below. In addition, the wine_reviews data indicates that wine ratings are positively associated with wine price and have a Pearson correlation equal roughly to 0.42. The wine_reviews data also suggests that price is statistically significant which indicates that the positive relationship unlikely due to chance.

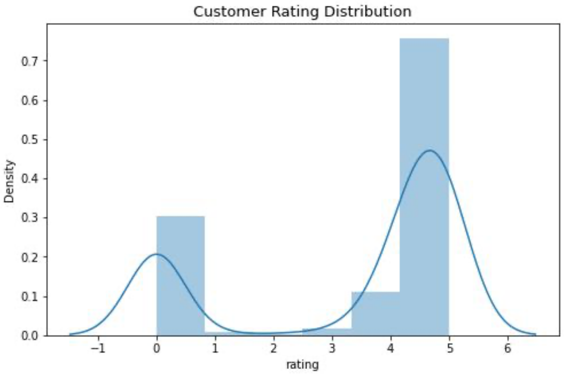


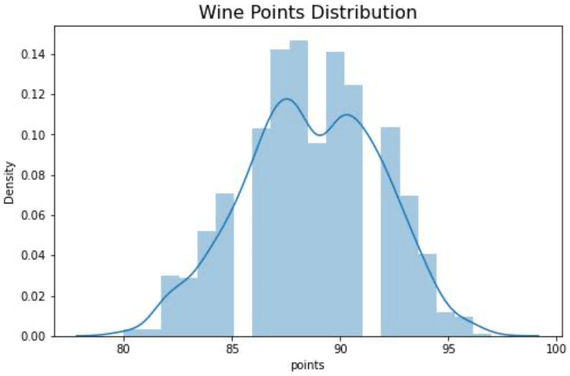
```
                     OLS Regression Results
    Dep. Variable:            rating         R-squared:      0.068
           Model:                OLS     Adj. R-squared:     0.062
          Method:      Least Squares        F-statistic:     12.67
            Date:   Tue, 13 Dec 2022  Prob (F-statistic):  6.73e-12
            Time:           19:02:39     Log-Likelihood:   -686.73
No. Observations:                879                AIC:     1385.
    Df Residuals:                873                BIC:     1414.
        Df Model:                  5
 Covariance Type:          nonrobust
                         coef   std err        t   P>|t|   [0.025   0.975]
              const    4.6212     0.097   47.749   0.000    4.431    4.811
              price    0.0065     0.002    3.603   0.000    0.003    0.010
 Alcohol Percentage   -0.0148     0.007   -2.003   0.045   -0.029   -0.000
              taste    0.3851     0.063    6.152   0.000    0.262    0.508
              value   -0.2093     0.049   -4.234   0.000   -0.306   -0.112
            quality   -0.1650     0.070   -2.362   0.018   -0.302   -0.028
       Omnibus:  640.211      Durbin-Watson:      1.979
Prob(Omnibus):     0.000   Jarque-Bera (JB):  11040.539
          Skew:   -3.167           Prob(JB):       0.00
      Kurtosis:   19.166          Cond. No.        118.
```

For categorical variables in both datasets, I visualized the results with different plots using Seaborn and Matplotlib. The analysis approach that I took is to assess the distribution of the categorical variables with respect to wine ratings. Detailed visualization and explanation are introduced in the next section.

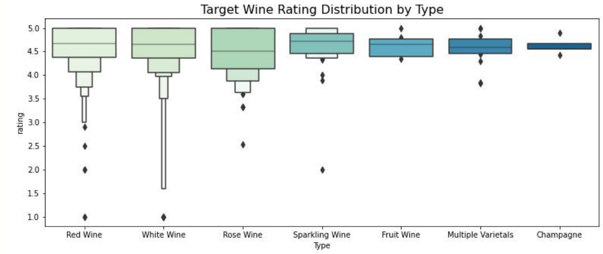
Visualization

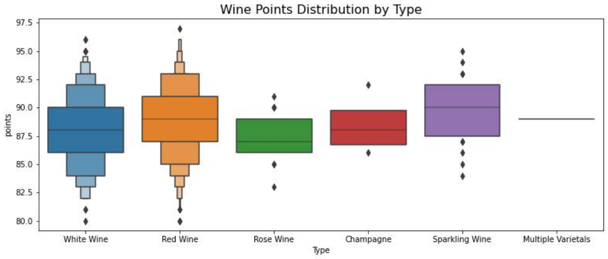
For ratings in both datasets, I plotted two histogram charts to better understand the distribution of the ratings/points (I will use them interchangeably as they share the same definition). For the target_wine data, the majority of the ratings are between 4 and 5. On the other hand, for the wine_reviews data, most points centered around 87 to 89 and 90 to 92. The results show that most wines receive good to high ratings.



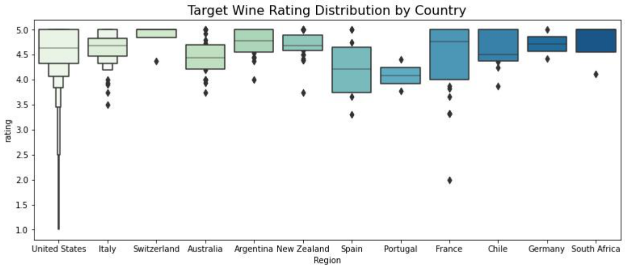


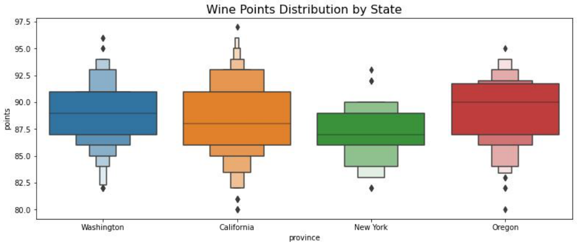
Both wine datasets contain a variable that describes the wine varietals. Because the wine varietals are too specific in the data and hard to visualize, I generalized the types of wines by their varieties. For example, wine has a variety named Cabernet Sauvignon, I categorized it as red wine. I created two boxplots to examine the rating distributions of wines by type. The vertical axes of the boxplot denote the ratings/points of the wine, and the horizontal axes denote the general wine type such as whether the wine is a red wine or a white wine. The dots on the graphs represent outliers. The boxplot displays the 75% and 25% percentile and average of the ratings of each type of wine. The left graph displays the ratings of target_wine and the right graph shows wine_reviews. The results of both graphs show that sparkling wine has the highest average rating, followed by red wine. In addition, rose wine seems to have the lowest rating.





In addition to rating distributions by wine type, I also visualized the rating distributions by region. The target_wine data contains wines from different countries while the wine_reviews data contains only US wines. This allows me to explore the relationships between wine rating and the country as well as the relationships between wine rating and US states. The vertical axes represent the ratings while the horizontal axes represent the countries or the US states. The graph on the left shows the Target wine rating distribution by country. The result suggests that France wine has the highest average rating, but the span of the rating is wide. US wine has an average rating and has a smaller lower whisker which indicates that the wine with the lowest rating is a US wine. The graph on the right shows the wine points distribution by US states. This graph gives a specific emphasis on US wine. The result shows that the state with the highest average rating is Oregon, followed by Washington and California.





The visualization went smoothly as I originally planned. I have plotted other graphs but did not end up using them because of repetition. The challenge that I encountered with my visualizations is that there are too many (more than 60) wine varieties in both datasets. For people that are not familiar with wine, plotting each wine based on its specific variety is not intuitive as plotting the type of wine. Despite many people are not familiar with wines, they may know that wine has types of red wine, white wine, etc. Therefore, I spent a lot of time classifying each wine by its type and visualized the types of wine instead of their varietals.

Future Work

For this project, I explored the relationships between numeric wine characteristics and wine ratings and the relationships between categorical wine characteristics separately. Given more time to improve the project, I would utilize machine learning technologies such as Decision Tree to predict the wine ratings based on a given set of wine attributes as it takes in both continuous and categorical variables. Although looking at statistical results of numeric attributes and categorical attributes separately still provides insights into how wine characteristics are related to wine ratings, the relationship may be different if combine all attributes are examined. In this project, I only explored the rating distributions by different regions, and I believe more visualizations can be done with regions such as geospatial plots. In conclusion, this project could be improved by incorporating machine learning techniques.