**Probability Theory and Statistics**

Logo

Description automatically generated

ALY6010, WINTER 2022

Module 2 Final Project – Milestone 2

Week-5

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Tom Breur

Date: 03/28/2022

**Module 2**

**Week 2**

**Introduction: Final Project Milestone-1**

**Exploratory Data Analysis**

Introduction:

Wine tasting is as old as the industry itself, with a more systematic approach gradually emerging from the 14th century onwards. To define a wine's spectrum of perceived flavors, aromas, and overall attributes, professional wine tasters (such as sommeliers or retail consumers) use a constantly developing specialized lexicon. Similar terminology could be used in more casual, recreational evaluations, which often require a considerably less objective manner for a more general, portion of one's personality.

This project is intended to use exploratory data analysis (EDA) approaches to investigate correlations between 14 variables, as well as to find the solutions for visualizations, distributions, outliers in a selected wine taste data set.

There are 1000 observations and 14 variables in this dataset. This dataset, which includes the variables price, designation, description, province, region 1, region 2, flavor, taster, taster name, Twitter handle, title, variety, and winery, will help in offering insights into the wine taste. Here, will show performing statistical analysis while performing command in R.

Text

Description automatically generated

It is defined with the country where the wine has been tasted, with the description according to the taste of the wine like Tropical fruit, broom, brimstone, and dried herb are among the aromas. The tongue is understated, with unripened apple, citrus, and dry sage accented by crisp acidity. Moreover, the designation as the name of the wine like Levele, Cignale, Rosado, with we can see points which help in defining the rating of the wine as per the rating price of the wine will be available in the dataset. We can see the availability of wine in various countries. This dataset will give an understanding of which province, region (reigion\_1 and region\_2) the wine is available depending on the consumption of the wine in a particular country. Apart from this dataset gives insights about the taste of the wine with taster\_name and information about the social media site where they handle the wine taste follower on Twitter. Likewise, this dataset defines the variety of wines available in various regions and provinces around different parts of the country. However, the most interesting thing we can retrieve the production details of the wine with winery details available in the dataset. Like sparkling blend is available in Unites States (US) which is available in Iron Horse winery.

Wine\_ Tasting data has various 1000 rows and 14 columns which contain ‘id’ as an integer (int), there are many character variables (chr) like, description, designation, province, region\_1, region\_2, taster\_name, taster\_twitter\_handle, country title, variety, and winery are defined as a character. Whereas Point and Price is a continuous variable is defined as integer(int).

**Purpose of Dataset:**

To get an understanding of the wine price in different countries with the specific brand and availability as per the rating (points) of the dataset. This Exploratory data analysis helps in visualizing the statistics through, bar scatterplot, boxplot, histogram, chart, and analysis of the dataset in the tabular form.

**Task 1: Dataset Imported in R:** In this task wine \_tasting contains 14 variables and 1000 observations which help in analyzing and visualizing the data in meaningful data available in the dataset.

Graphical user interface, application

Description automatically generatedGraphical user interface

Description automatically generated with low confidence

**Task 2: Data cleaning**: Create Duplicate Dataset to fill missing value with NA to make the data in an evaluation formate.

A picture containing letter

Description automatically generated

A picture containing text, indoor

Description automatically generated

A picture containing text

Description automatically generated

**Task 3**: In this task change the name of the data variable available in dataset

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Task 4:** In this task dropped variable to do our analysis precise Text

Description automatically generatedGraphical user interface, application

Description automatically generated

**Task 4**: In this task structure of the dataset after dropping a few variables.

Text, letter

Description automatically generated

**Task 5**: In this task as the point has min. value 80 and max. value 100 with the mean and median 88, 88.58 respectively. Moreover, the minimum price of the wine is 7, and the maximum price is 775

A picture containing calendar

Description automatically generated

**Task 6:** In this task before analyzing the facts about the dataset- glimpse()

Scatter chart

Description automatically generated

**Task 7**: In this task to create the scatterplot for understanding the check the correlation between the variables of the wine\_tasting dataset. We can see the correlation between price and points. Let’s discuss this

A picture containing text, crossword puzzle, receipt

Description automatically generated

**Task 8:** In this task, the boxplot shows the outlier of the price variable, which shows around 84 to 94 points the price of the wine is below 100, and above the price, boxplot shows the outlier which above price 400.

A picture containing box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

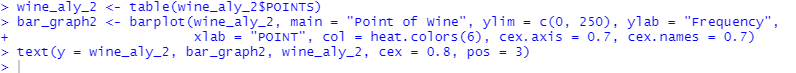
**Task 9**: In this task shows the price of the wine with frequency

Text

Description automatically generated

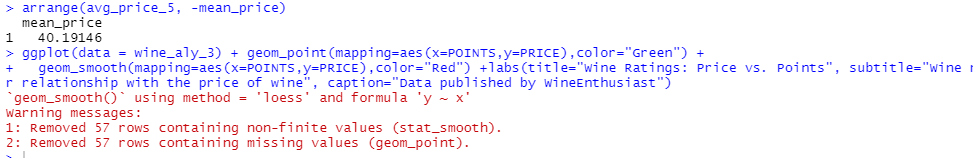
A screenshot of a computer

Description automatically generated with low confidence

**Task 10**: In this task points if the wine shows at point 87 maximum number of wine are available i.e. 221 ****Chart, bar chart

Description automatically generated

**Task 13:** In this task we can see the relationship between the wine price and the rating of the wine, higher the prices in ascending order with very fewer reviews(points) **ggplot()**

****

**Text

Description automatically generated with medium confidence**

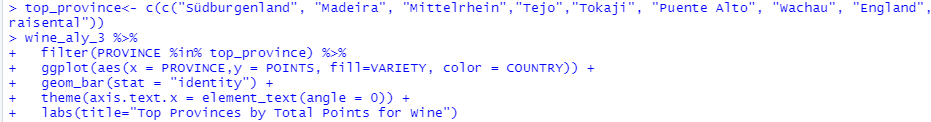
Chart

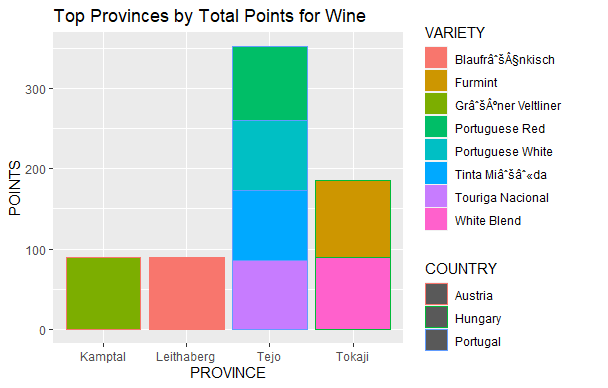
Description automatically generated

**Milestone 2**

**Question1: What is the top-rank variety of wine by province?’’**

Points are showing a variety of wines are used in the various countryin top provinces.

****



**Question2: What is the price of a variety of wines by province?**

**Province by Price:**

Prices are showing a variety of wines are used in various countriesin top provinces.

**Text

Description automatically generated**

Chart, bar chart

Description automatically generated

**Hypothesis Testing**

**Two-Sample t-test:**

* When comparing two vectors of numeric data, t-test() is commonly used. It is entirely dependent on how the items in the dataset have been configured.
* We conducted a two-sample test on the POINT and PRICE, with a t= 2.0526, 18 degrees of freedom, and a 0.054 p-value.
* The -0.021 and 1.812 are represented by the confidence interval. with a mean of 0.2416 for x and -0.6538 for y.
* The lesser result, 0.054, provides strong evidence in favor of the null hypothesis, hence it is accepted.

**Graphical user interface, text

Description automatically generated**

**One-Sample test:**

The one-sample t-test is used to test the population of the sample. A population dataset is used to run the t-test.

We used the Price variable in the one-sample test to understand the hypothesis degree of freedom is 9 and the p-value is larger than 0.05, indicating that alternative hypothesis testing is unnecessary.Graphical user interface, text, application, email

Description automatically generated

**Summary:**

In this wine\_tasting dataset, prices depend on the points of the wine higher the review higher the price. All the variables in this data provide analysis above the graphs. Explain what the tables and visualizations tell you about the data.

This research aims to study relationships between 14 variables, as well as find solutions for visualizations, distributions, and outliers in a selected wine taste data set, using exploratory data analysis (EDA) methodologies.

This dataset has 1000 observations and 14 variables. This dataset will aid in providing insights into the wine taste by including the variables designation, description, flavor, price, province, region 1, region 2, taster name, tastera twitter handle, title, variety, and winery. Here, we'll demonstrate how to perform the statistical analysis while using the R.

There are many character variables (chr) such as country, description, designation, province, region 1, region 2, taster name, taster twitter handle, title, variety, and winery that are defined as a character. Wine\_ Tasting data has various 1000 rows and 14 columns that contain 'id' as an integer (int). Point and Price, on the other hand, is a continuous variable that is defined as an integer (int).

It is described by the country in which the wine was tasted, as well as a description of the wine's flavor, such as tropical fruit, broom, brimstone, and dry herb, among other scents. Unripened apple, citrus, and dry sage are enhanced by crisp acidity on the tongue. Furthermore, the designation as the name of the wine, such as Levele, Cignale, or Rosado, will be available in the dataset, along with points that assist in determining the rating of the wine based on the rating price of the wine.

The availability of wine in many countries may be seen. This dataset will show which provinces and regions (reigion 1 and region 2) the wine is available in, based on the wine's consumption in a given country. Apart from that, this dataset contains information about the wine taster name and the social media site where they manage the wine taste follower on Twitter.

**Reference:**

[1]Create a new ggplot

<https://ggplot2.tidyverse.org/reference/ggplot.html>

[2]Exploring, cleaning, and analysing data in R - YouTube

A replay of a non-technical livestream that walked through how to explore, clean, and analyse data in R, using the 'starwars' dataset that is built into the ...

<https://www.youtube.com/watch?v=Ap1Q2fkqO_I>

[3]How To Make Frequency Table in R - Programming R Tutorials

<https://www.programmingr.com/statistics/frequency-table/>

[4]Wine tasting

<https://en.wikipedia.org/wiki/Wine_tasting#:~:text=There%20are%20five%20basic%20steps,expressiveness%2C%20complexity%2C%20and%20connectedness.>

[5]R in Action

<https://www.google.com/books/edition/R_in_Action/1TkzEAAAQBAJ?hl=en&gbpv=1&printsec=frontcover>

[6] Introduction to Hypothesis Testing in R - Learn every concept from Scratch!

<https://data-flair.training/blogs/hypothesis-testing-in-r/>

**Appendix: R code**

**Table

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**