Probability Theory and Statistics

ALY6010, WINTER 2022

Module 2 Final Project – Milestone 1

Week-2

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Tom Breur

Date: 03/07/2022

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.

```
wine_data                1000 obs. of 14 variables
  $ ï..                  : int [1:1000] 0 1 2 3 4 5 6 7 8 9 ...
  $ country              : chr [1:1000] "Italy" "Portugal" "US" "US" ...
  $ description          : chr [1:1000] "Aromas include tropical fruit, broom, brimstone and …
  $ designation          : chr [1:1000] "Vulkâ\210šâ\200  Bianco" "Avidagos" "" "Reserve Late…
  $ points               : int [1:1000] 87 87 87 87 87 87 87 87 87 87 ...
  $ price                : int [1:1000] NA 15 14 13 65 15 16 24 12 27 ...
  $ province             : chr [1:1000] "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
  $ region_1             : chr [1:1000] "Etna" "" "Willamette Valley" "Lake Michigan Shore" …
  $ region_2             : chr [1:1000] "" "" "Willamette Valley" "" ...
  $ taster_name          : chr [1:1000] "Kerin Oâ\200šÃ„Ã´Keefe" "Roger Voss" "Paul Gregutt" …
  $ taster_twitter_handle: chr [1:1000] "@kerinokeefe" "@vossroger" "@paulgwineÂ¬â\200 " "" …
  $ title                : chr [1:1000] "Nicosia 2013 Vulkâ\210šâ\200  Bianco  (Etna)" "Quint…
  $ variety              : chr [1:1000] "White Blend" "Portuguese Red" "Pinot Gris" "Riesling…
  $ winery               : chr [1:1000] "Nicosia" "Quinta dos Avidagos" "Rainstorm" "St. Juli…
```

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,

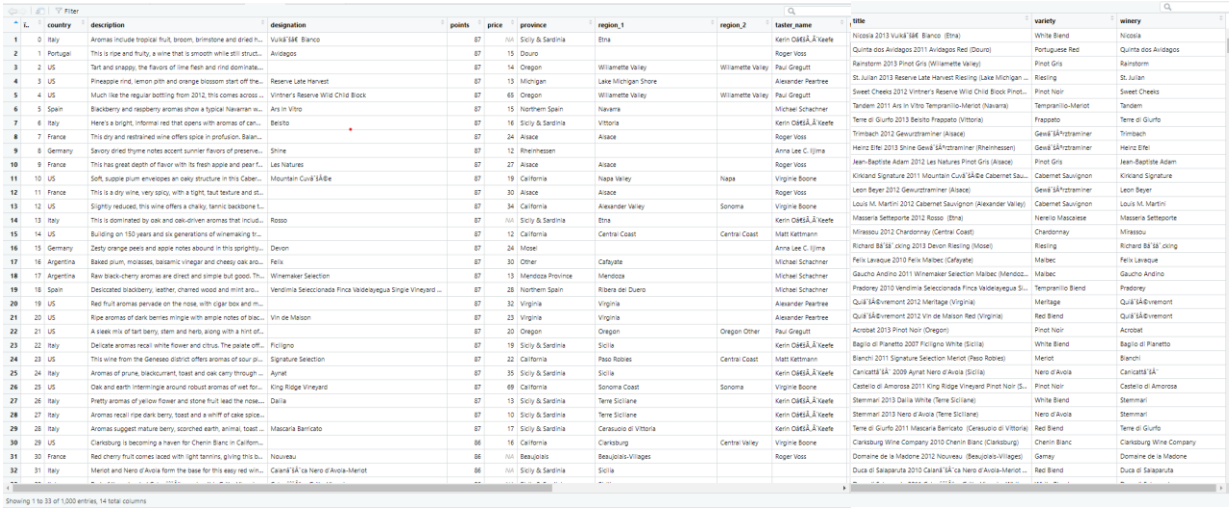
region (reigion_1 and region_2) the wine is available depending on the consumption of the wine in a particular country. Apart from this dataset gives insights about the taste of the wine with taster_name and information about the social media site where they handle the wine taste follower on Twitter. 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.



	i..	country	description	designation	points	price	province	region_1	region_2	taster_name
1	0	Italy	Aromas include tropical fruit, broom, brimstone and dried h...	Vulkà Bianco	87	NA	Sicily & Sardinia	Etna		Kerin Oâ€Â Keefe
2	1	Portugal	This is ripe and fruity, a wine that is smooth while still struct...	Avidagos	87	15	Douro			Roger Voss
3	2	US	Tart and snappy, the flavors of lime flesh and rind dominate....		87	14	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt
4	3	US	Pineapple rind, lemon pith and orange blossom start off the...	Reserve Late Harvest	87	13	Michigan	Lake Michigan Shore		Alexander Peartree
5	4	US	Much like the regular bottling from 2012, this comes across ...	Vintner's Reserve Wild Child Block	87	65	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt
6	5	Spain	Blackberry and raspberry aromas show a typical Navarran w...	Ars In Vitro	87	15	Northern Spain	Navarra		Michael Schachner
7	6	Italy	Here's a bright, informal red that opens with aromas of can...	Belsito	87	16	Sicily & Sardinia	Vittoria		Kerin Oâ€Â Keefe
8	7	France	This dry and restrained wine offers spice in profusion. Balan...		87	24	Alsace	Alsace		Roger Voss
9	8	Germany	Savory dried thyme notes accent sunnier flavors of preserve...	Shine	87	12	Rheinhessen			Anne Lee C. Iijima
10	9	France	This has great depth of flavor with its fresh apple and pear f...	Les Natures	87	27	Alsace	Alsace		Roger Voss
11	10	US	Soft, supple plum envelopes an oaky structure in this Caber...	Mountain Cuvã Âe	87	19	California	Napa Valley	Napa	Virginie Boone
12	11	France	This is a dry wine, very spicy, with a tight, taut texture and st...		87	30	Alsace	Alsace		Roger Voss
13	12	US	Slightly reduced, this wine offers a chalky, tannic backbone t...		87	34	California	Alexander Valley	Sonoma	Virginie Boone
14	13	Italy	This is dominated by oak and oak-driven aromas that includ...	Rosso	87	NA	Sicily & Sardinia	Etna		Kerin Oâ€Â Keefe
15	14	US	Building on 150 years and six generations of winemaking tr...		87	12	California	Central Coast	Central Coast	Matt Kettmann
16	15	Germany	Zesty orange peels and apple notes abound in this sprightly...	Devon	87	24	Mosel			Anne Lee C. Iijima
17	16	Argentina	Baked plum, molasses, balsamic vinegar and cheesy oak aro...	Felix	87	30	Other	Cafayate		Michael Schachner
18	17	Argentina	Raw black-cherry aromas are direct and simple but good. Th...	Winemaker Selection	87	13	Mendoza Province	Mendoza		Michael Schachner
19	18	Spain	Desiccated blackberry, leather, charred wood and mint aro...	Vendimia Seleccionada Finca Valdelayegua Single Vineyard ...	87	28	Northern Spain	Ribera del Duero		Michael Schachner
20	19	US	Red fruit aromas pervade on the nose, with cigar box and m...		87	32	Virginia	Virginia		Alexander Peartree
21	20	US	Ripe aromas of dark berries mingle with ample notes of blac...	Vin de Maison	87	23	Virginia	Virginia		Alexander Peartree
22	21	US	A week mix of tart berry, stem and herb, along with a hint of...		87	20	Oregon	Oregon	Oregon Other	Paul Gregutt
23	22	Italy	Delicate aromas recall white flower and citrus. The palate off...	Ficligno	87	19	Sicily & Sardinia	Sicilia		Kerin Oâ€Â Keefe
24	23	US	This wine from the Geneseo district offers aromas of sour pl...	Signature Selection	87	22	California	Paso Robles	Central Coast	Matt Kettmann
25	24	Italy	Aromas of prune, blackcurrant, toast and oak carry through ...	Aynat	87	35	Sicily & Sardinia	Sicilia		Kerin Oâ€Â Keefe
26	25	US	Oak and earth intermingle around robust aromas of wet foo...	King Ridge Vineyard	87	69	California	Sonoma Coast	Sonoma	Virginie Boone
27	26	Italy	Pretty aromas of yellow flower and stone fruit lead the nose....	Dalila	87	13	Sicily & Sardinia	Terre Siciliane		Kerin Oâ€Â Keefe
28	27	Italy	Aromas recall ripe dark berry, toast and a whiff of cake spice...		87	10	Sicily & Sardinia	Terre Siciliane		Kerin Oâ€Â Keefe
29	28	Italy	Aromas suggest mature berry, scorched earth, animal, toast...	Mascarla Barricato	87	17	Sicily & Sardinia	Cerasuolo di Vittoria		Kerin Oâ€Â Keefe
30	29	US	Clarksburg is becoming a haven for Chenin Blanc in Californ...		86	16	California	Clarksburg	Central Valley	Virginie Boone
31	30	France	Red cherry fruit comes laced with light tannins, giving this b...	Nouveau	86	NA	Beaujolais	Beaujolais-Villages		Roger Voss
32	31	Italy	Merlot and Nero d'Avola form the base for this easy red win...	Calanã ca Nero d'Avola-Merlot	86	NA	Sicily & Sardinia	Sicilia		

Showing 1 to 33 of 1,000 entries, 14 total columns

title	variety	winery
Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan ...	Riesling	St. Julian
Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot...	Pinot Noir	Sweet Cheeks
Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)	Tempranillo-Merlot	Tandem
Terre di Giurfo 2013 Belsito Frappato (Vittoria)	Frappato	Terre di Giurfo
Trimbach 2012 Gewurztraminer (Alsace)	Gewã Âºrztraminer	Trimbach
Heinz Eifel 2013 Shine Gewã Âºrztraminer (Rheinhessen)	Gewã Âºrztraminer	Heinz Eifel
Jean-Baptiste Adam 2012 Les Natures Pinot Gris (Alsace)	Pinot Gris	Jean-Baptiste Adam
Kirkland Signature 2011 Mountain Cuvã Âe Cabernet Sau...	Cabernet Sauvignon	Kirkland Signature
Leon Beyer 2012 Gewurztraminer (Alsace)	Gewã Âºrztraminer	Leon Beyer
Louis M. Martini 2012 Cabernet Sauvignon (Alexander Valley)	Cabernet Sauvignon	Louis M. Martini
Masseria Setteporte 2012 Rosso (Etna)	Nerello Mascalese	Masseria Setteporte
Mirassou 2012 Chardonnay (Central Coast)	Chardonnay	Mirassou
Richard Bã ¶cking 2013 Devon Riesling (Mosel)	Riesling	Richard Bã ¶cking
Felix Lavaque 2010 Felix Malbec (Cafayate)	Malbec	Felix Lavaque
Gaucho Andino 2011 Winemaker Selection Malbec (Mendoz...	Malbec	Gaucho Andino
Predorey 2010 Vendimia Seleccionada Finca Valdelayegua Si...	Tempranillo Blend	Predorey
Quilã Âvremont 2012 Meritage (Virginia)	Meritage	Quilã Âvremont
Quilã Âvremont 2012 Vin de Maison Red (Virginia)	Red Blend	Quilã Âvremont
Acrobat 2013 Pinot Noir (Oregon)	Pinot Noir	Acrobat
Baglio di Pianetto 2007 Ficligno White (Sicilia)	White Blend	Baglio di Pianetto
Bianchi 2011 Signature Selection Merlot (Paso Robles)	Merlot	Bianchi
Canicattã Â 2009 Aynat Nero d'Avola (Sicilia)	Nero d'Avola	Canicattã Â
Castello di Amorosa 2011 King Ridge Vineyard Pinot Noir (S...	Pinot Noir	Castello di Amorosa
Stemmari 2013 Dalila White (Terre Siciliane)	White Blend	Stemmari
Stemmari 2013 Nero d'Avola (Terre Siciliane)	Nero d'Avola	Stemmari
Terre di Giurfo 2011 Mascaria Barricato (Cerasuolo di Vittoria)	Red Blend	Terre di Giurfo
Clarksburg Wine Company 2010 Chenin Blanc (Clarksburg)	Chenin Blanc	Clarksburg Wine Company
Domaine de la Madone 2012 Nouveau (Beaujolais-Villages)	Gamay	Domaine de la Madone
Duca di Salaparuta 2010 Calanã ca Nero d'Avola-Merlot ...	Red Blend	Duca di Salaparuta

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

```
wine_new                  1000 obs. of 14 variables
  $ ï..                 : int  0 1 2 3 4 5 6 7 8 9 ...
  $ country             : chr  "Italy" "Portugal" "US" "US" ...
  $ description         : chr  "Aromas include tropical fruit, broom, brimstone and dried herb. The pa
  $ designation         : chr  "Vulkâ\210šâ\200  Bianco" "Avidagos" NA "Reserve Late Harvest" ...
  $ points              : int  87 87 87 87 87 87 87 87 87 87 ...
  $ price               : int  NA 15 14 13 65 15 16 24 12 27 ...
  $ province            : chr  "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
  $ region_1            : chr  "Etna" NA "Willamette Valley" "Lake Michigan Shore" ...
  $ region_2            : chr  NA NA "Willamette Valley" NA ...
  $ taster_name         : chr  "Kerin Oâ\200šÃ„Á´Keefe" "Roger Voss" "Paul Gregutt" "Alexander Peartre
  $ taster_twitter_handle: chr  "@kerinokeefe" "@vossroger" "@paulgwineÂ¬â\200 " NA ...
  $ title               : chr  "Nicosia 2013 Vulkâ\210šâ\200  Bianco  (Etna)" "Quinta dos Avidagos 201
  $ variety             : chr  "White Blend" "Portuguese Red" "Pinot Gris" "Riesling" ...
  $ winery              : chr  "Nicosia" "Quinta dos Avidagos" "Rainstorm" "St. Julian" ...
```

	ï..	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle
1	0	Italy	Aromas include tropical fruit, broom, brimstone and dried h...	Vulkâ€šÃ„Â Bianco	87	NA	Sicily & Sardinia	Etna	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
2	1	Portugal	This is ripe and fruity, a wine that is smooth while still struct...	Avidagos	87	15	Douro	NA	NA	Roger Voss	@vossroger
3	2	US	Tart and snappy, the flavors of lime flesh and rind dominate...	NA	87	14	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwineÂ¬â€
4	3	US	Pineapple rind, lemon pith and orange blossom start off the...	Reserve Late Harvest	87	13	Michigan	Lake Michigan Shore	NA	Alexander Peartree	NA
5	4	US	Much like the regular bottling from 2012, this comes across ...	Vintner's Reserve Wild Child Block	87	65	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwineÂ¬â€
6	5	Spain	Blackberry and raspberry aromas show a typical Navarran w...	Ars In Vitro	87	15	Northern Spain	Navarra	NA	Michael Schachner	@wineschach
7	6	Italy	Here's a bright, informal red that opens with aromas of can...	Belsito	87	16	Sicily & Sardinia	Vittoria	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
8	7	France	This dry and restrained wine offers spice in profusion. Balan...	NA	87	24	Alsace	Alsace	NA	Roger Voss	@vossroger
9	8	Germany	Savory dried thyme notes accent sunnier flavors of preserve...	Shine	87	12	Rheinhessen	NA	NA	Anna Lee C. Iijima	NA
10	9	France	This has great depth of flavor with its fresh apple and pear f...	Les Natures	87	27	Alsace	Alsace	NA	Roger Voss	@vossroger
11	10	US	Soft, supple plum envelopes an oaky structure in this Caber...	Mountain Cuvâ´šÃ©e	87	19	California	Napa Valley	Napa	Virginie Boone	@vboone
12	11	France	This is a dry wine, very spicy, with a tight, taut texture and st...	NA	87	30	Alsace	Alsace	NA	Roger Voss	@vossroger
13	12	US	Slightly reduced, this wine offers a chalky, tannic backbone t...	NA	87	34	California	Alexander Valley	Sonoma	Virginie Boone	@vboone
14	13	Italy	This is dominated by oak and oak-driven aromas that includ...	Rosso	87	NA	Sicily & Sardinia	Etna	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
15	14	US	Building on 150 years and six generations of winemaking tr...	NA	87	12	California	Central Coast	Central Coast	Matt Kettmann	@mattkettmann
16	15	Germany	Zesty orange peels and apple notes abound in this sprightly...	Devon	87	24	Mosel	NA	NA	Anna Lee C. Iijima	NA
17	16	Argentina	Baked plum, molasses, balsamic vinegar and cheesy oak aro...	Felix	87	30	Other	Cafayate	NA	Michael Schachner	@wineschach
18	17	Argentina	Raw black-cherry aromas are direct and simple but good. Th...	Winemaker Selection	87	13	Mendoza Province	Mendoza	NA	Michael Schachner	@wineschach
19	18	Spain	Desiccated blackberry, leather, charred wood and mint aro...	Vendimia Seleccionada Finca Valdelayegua Single Vineyard ...	87	28	Northern Spain	Ribera del Duero	NA	Michael Schachner	@wineschach
20	19	US	Red fruit aromas pervade on the nose, with cigar box and m...	NA	87	32	Virginia	Virginia	NA	Alexander Peartree	NA
21	20	US	Ripe aromas of dark berries mingle with ample notes of blac...	Vin de Maison	87	23	Virginia	Virginia	NA	Alexander Peartree	NA
22	21	US	A sleek mix of tart berry, stem and herb, along with a hint of...	NA	87	20	Oregon	Oregon	Oregon Other	Paul Gregutt	@paulgwineÂ¬â€
23	22	Italy	Delicate aromas recall white flower and citrus. The palate off...	Ficiligno	87	19	Sicily & Sardinia	Sicilia	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
24	23	US	This wine from the Geneseo district offers aromas of sour pl...	Signature Selection	87	22	California	Paso Robles	Central Coast	Matt Kettmann	@mattkettmann
25	24	Italy	Aromas of prune, blackcurrant, toast and oak carry through ...	Aynat	87	35	Sicily & Sardinia	Sicilia	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
26	25	US	Oak and earth intermingle around robust aromas of wet for...	King Ridge Vineyard	87	69	California	Sonoma Coast	Sonoma	Virginie Boone	@vboone
27	26	Italy	Pretty aromas of yellow flower and stone fruit lead the nose....	Dalila	87	13	Sicily & Sardinia	Terre Siciliane	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
28	27	Italy	Aromas recall ripe dark berry, toast and a whiff of cake spice...	NA	87	10	Sicily & Sardinia	Terre Siciliane	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
29	28	Italy	Aromas suggest mature berry, scorched earth, animal, toast ...	Mascarà Barricato	87	17	Sicily & Sardinia	Cerasuolo di Vittoria	NA	Kerin Oâ€šÃ„Á´Keefe	@kerinokeefe
30	29	US	Clarksburg is becoming a haven for Chenin Blanc in Californ...	NA	86	16	California	Clarksburg	Central Valley	Virginie Boone	@vboone
31	30	France	Red cherry fruit comes laced with light tannins, giving this b...	Nouveau	86	NA	Beaujolais	Beaujolais-Villages	NA	Roger Voss	@vossroger
32	31	Italy	Merlot and Nero d'Avola form the base for this easy red win...	Calanâ´šÃ¨ca Nero d'Avola-Merlot	86	NA	Sicily & Sardinia	Sicilia	NA	NA	NA

```
> wine_new <- wine_data
> wine_new[wine_new == ""] <- NA
> wine_new
   ï..    country
1    0      Italy
2    1   Portugal
3    2         US
4    3         US
5    4         US
6    5      Spain
7    6      Italy
8    7     France
9    8    Germany
10   9     France
11  10         US
12  11     France
```

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

```
> wine_aly_rename1 <- rename_with(wine_new, toupper)
> 
```

```
   Ï..   COUNTRY
1    0     Italy
2    1  Portugal
3    2        US
4    3        US
5    4        US
6    5     Spain
7    6     Italy
8    7    France
9    8   Germany
10   9    France
11  10        US
12  11    France
13  12        US
14  13     Italy
15  14        US
16  15   Germany
17  16 Argentina
18  17 Argentina
```

	Ï..	COUNTRY	DESCRIPTION	DESIGNATION	POINTS	PRICE	PROVINCE	REGION_
1	0	Italy	Aromas include tropical fruit, broom, brimstone and dried h...	Vulkâ´šâ€ Bianco	87	NA	Sicily & Sardinia	Etna
2	1	Portugal	This is ripe and fruity, a wine that is smooth while still struct...	Avidagos	87	15	Douro	NA
3	2	US	Tart and snappy, the flavors of lime flesh and rind dominate....	NA	87	14	Oregon	Willamett
4	3	US	Pineapple rind, lemon pith and orange blossom start off the...	Reserve Late Harvest	87	13	Michigan	Lake Mic
5	4	US	Much like the regular bottling from 2012, this comes across ...	Vintner's Reserve Wild Child Block	87	65	Oregon	Willamett
6	5	Spain	Blackberry and raspberry aromas show a typical Navarran w...	Ars In Vitro	87	15	Northern Spain	Navarra
7	6	Italy	Here's a bright, informal red that opens with aromas of can...	Belsito	87	16	Sicily & Sardinia	Vittoria
8	7	France	This dry and restrained wine offers spice in profusion. Balan...	NA	87	24	Alsace	Alsace
9	8	Germany	Savory dried thyme notes accent sunnier flavors of preserve...	Shine	87	12	Rheinhessen	NA
10	9	France	This has great depth of flavor with its fresh apple and pear f...	Les Natures	87	27	Alsace	Alsace
11	10	US	Soft, supple plum envelopes an oaky structure in this Caber...	Mountain Cuvâ´šÂ€e	87	19	California	Napa Vall
12	11	France	This is a dry wine, very spicy, with a tight, taut texture and st...	NA	87	30	Alsace	Alsace

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

```
> #Drop
> wine_aly_drop <- subset(wine_aly_rename1, select = -c (DESCRIPTION, TITLE, TASTER_NAME, TASTER_TWITTER_HANDLE, WINERY, REGION_1, REGION_2))
> 
```

```
    Ï..      COUNTRY                                 DESIGNATION POINTS PRICE            PROVINCE
1     0        Italy                          Vulkâ€šâ€  Bianco     87    NA      Sicily & Sardinia
2     1     Portugal                                   Avidagos     87    15               Douro
3     2           US                                       <NA>     87    14              Oregon
4     3           US                        Reserve Late Harvest     87    13            Michigan
5     4           US                 Vintner's Reserve Wild Child Block     87    65              Oregon
6     5        Spain                                 Ars In Vitro     87    15      Northern Spain
7     6        Italy                                    Belsito     87    16      Sicily & Sardinia
8     7       France                                       <NA>     87    24             Alsace
9     8      Germany                                      Shine     87    12          Rheinhessen
```

	Ï..	COUNTRY	DESIGNATION	POINTS	PRICE	PROVINCE	VARIETY
1	0	Italy	Vulkâ€šâ€ Bianco	87	*NA*	Sicily & Sardinia	White Blend
2	1	Portugal	Avidagos	87	15	Douro	Portuguese Red
3	2	US	*NA*	87	14	Oregon	Pinot Gris
4	3	US	Reserve Late Harvest	87	13	Michigan	Riesling
5	4	US	Vintner's Reserve Wild Child Block	87	65	Oregon	Pinot Noir
6	5	Spain	Ars In Vitro	87	15	Northern Spain	Tempranillo-Merlot
7	6	Italy	Belsito	87	16	Sicily & Sardinia	Frappato
8	7	France	*NA*	87	24	Alsace	GewâšÂºrztraminer
9	8	Germany	Shine	87	12	Rheinhessen	GewâšÂºrztraminer
10	9	France	Les Natures	87	27	Alsace	Pinot Gris
11	10	US	Mountain Cuvâ€šÂ©e	87	19	California	Cabernet Sauvignon
12	11	France	*NA*	87	30	Alsace	GewâšÂºrztraminer
13	12	US	*NA*	87	34	California	Cabernet Sauvignon
14	13	Italy	Rosso	87	*NA*	Sicily & Sardinia	Nerello Mascalese
15	14	US	*NA*	87	12	California	Chardonnay
16	15	Germany	Devon	87	24	Mosel	Riesling
17	16	Argentina	Felix	87	30	Other	Malbec
18	17	Argentina	Winemaker Selection	87	13	Mendoza Province	Malbec
19	18	Spain	Vendimia Seleccionada Finca Valdelayegua Single Vineyard ...	87	28	Northern Spain	Tempranillo Blend
20	19	US	*NA*	87	32	Virginia	Meritage

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

```
> str(wine_aly_1)
'data.frame':   1000 obs. of  7 variables:
 $ Ï..        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ COUNTRY    : chr  "Italy" "Portugal" "US" "US" ...
 $ DESIGNATION: chr  "Vulkâ\210šâ\200  Bianco" "Avidagos" NA "Reserve Late Harvest" ...
 $ POINTS     : int  87 87 87 87 87 87 87 87 87 87 ...
 $ PRICE      : int  NA 15 14 13 65 15 16 24 12 27 ...
 $ PROVINCE   : chr  "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
 $ VARIETY    : chr  "White Blend" "Portuguese Red" "Pinot Gris" "Riesling" ...
> 
```

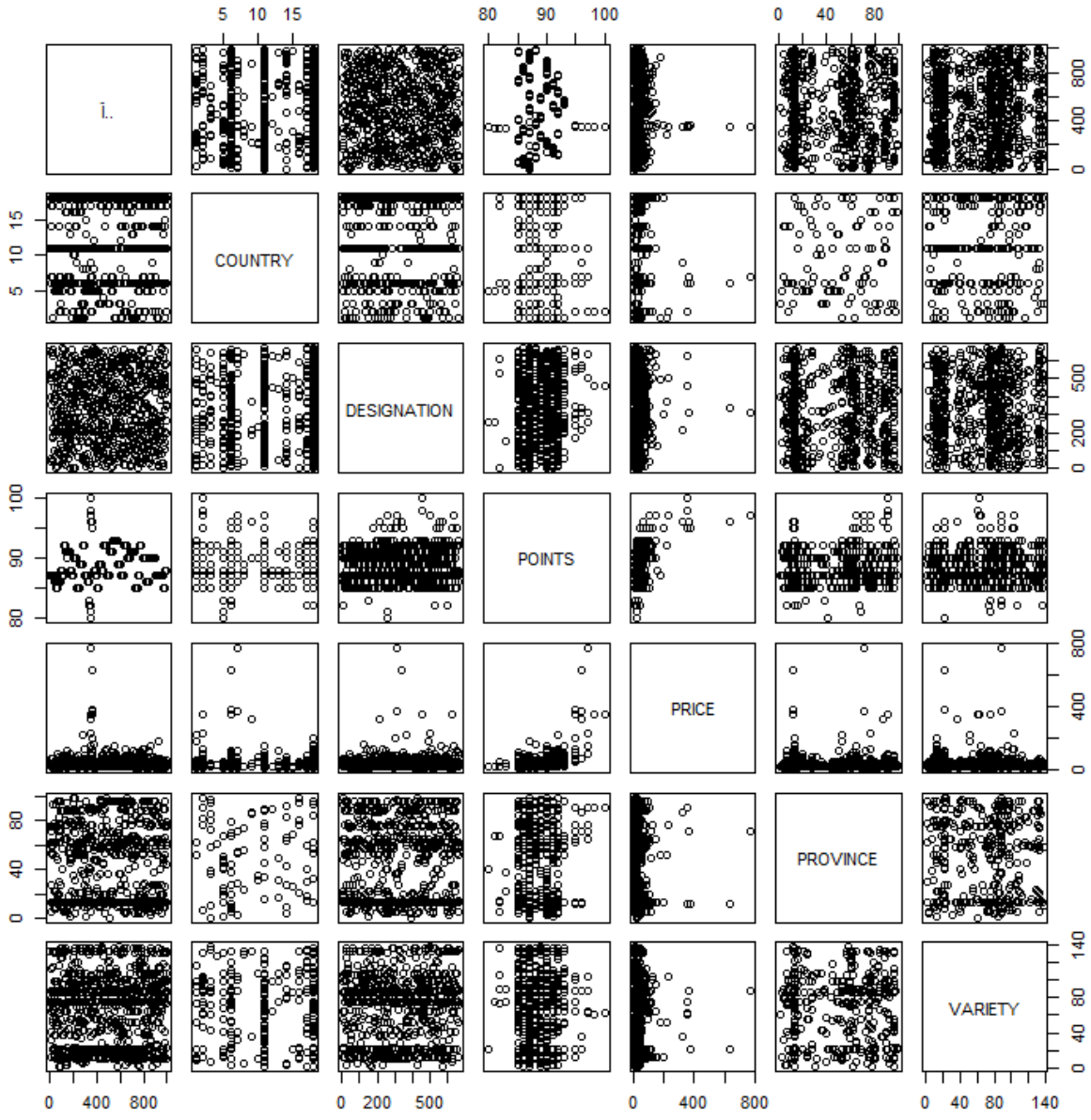
Task 5: In this task as the point has min. value 80 and max. value 100 with the mean and median 88, 88.58 respectively. Moreover, the minimum price of the wine is 7, and the maximum price is 775

```
> summary(wine_aly_1)
      Ï..          COUNTRY          DESIGNATION           POINTS          PRICE          PROVINCE           VARIETY         
 Min.   :  0.0   Length:1000        Length:1000        Min.   : 80.00   Min.   :  7.00   Length:1000        Length:1000       
 1st Qu.:249.8   Class :character   Class :character   1st Qu.: 87.00   1st Qu.: 17.00   Class :character   Class :character  
 Median :499.5   Mode  :character   Mode  :character   Median : 88.00   Median : 27.00   Mode  :character   Mode  :character  
 Mean   :499.5                                         Mean   : 88.58   Mean   : 37.35                                        
 3rd Qu.:749.2                                         3rd Qu.: 90.00   3rd Qu.: 43.50                                        
 Max.   :999.0                                         Max.   :100.00   Max.   :775.00                                        
                                                                        NA's   :57                                            
> 
```

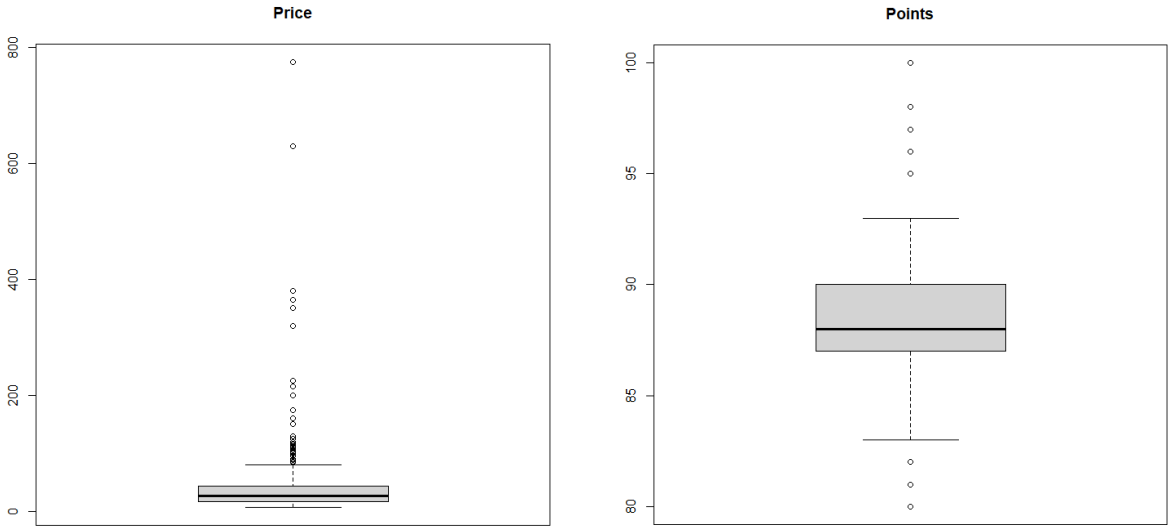
Task 6: In this task before analyzing the facts about the dataset- glimpse()

```
> glimpse(wine_aly_1)
Rows: 1,000
Columns: 7
$ Ï..        <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, ~
$ COUNTRY    <chr> "Italy", "Portugal", "US", "US", "US", "Spain", "Italy", "France", "Germany", "France", "US", "France", "US", ~
$ DESIGNATION <chr> "Vulkâ\210šâ\200  Bianco", "Avidagos", NA, "Reserve Late Harvest", "Vintner's Reserve Wild Child Block", "Ars ~
$ POINTS     <int> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87~
$ PRICE      <int> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, 19, 30, 34, NA, 12, 24, 30, 13, 28, 32, 23, 20, 19, 22, 35, 69, 13, 10~
$ PROVINCE   <chr> "Sicily & Sardinia", "Douro", "Oregon", "Michigan", "Oregon", "Northern Spain", "Sicily & Sardinia", "Alsace",~
$ VARIETY    <chr> "White Blend", "Portuguese Red", "Pinot Gris", "Riesling", "Pinot Noir", "Tempranillo-Merlot", "Frappato", "Ge~
> 
```

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

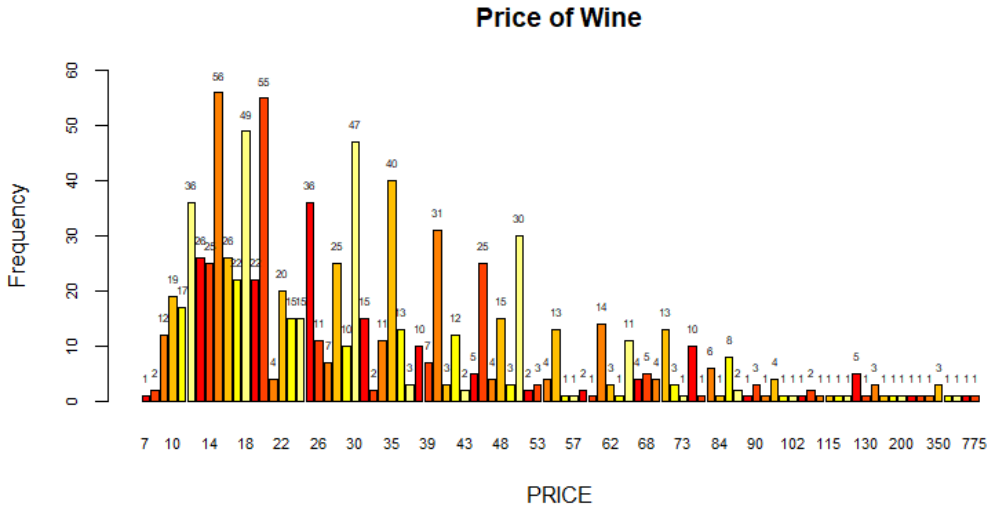


Task 8: In this task, the boxplot shows the outlier of the price variable, which shows around 84 to 94 points the price of the wine is below 100, and above the price, boxplot shows the outlier which above price 400.



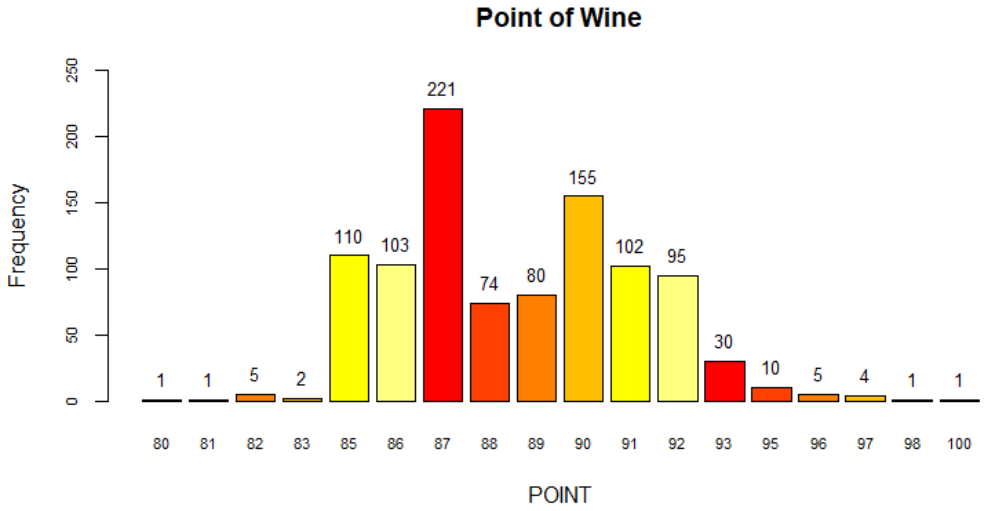
Task 9: In this task shows the price of the wine with frequency

```
> # BOX PLOT
> boxplot(wine_aly_1$PRICE)
> #Plot a frequency barchart of a price of wine
> wine_aly_1 <- table(wine_aly_1$PRICE)
> bar_graph1 <- barplot(wine_aly_1, main = "Price of Wine", ylim = c(0, 60), ylab = "Frequency",
+                       xlab = "PRICE", col = heat.colors(6), cex.axis = 0.7, cex.names = 0.7)
> text(y = wine_aly_1, bar_graph1, wine_aly_1, cex = 0.5, pos = 3)
>
```



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

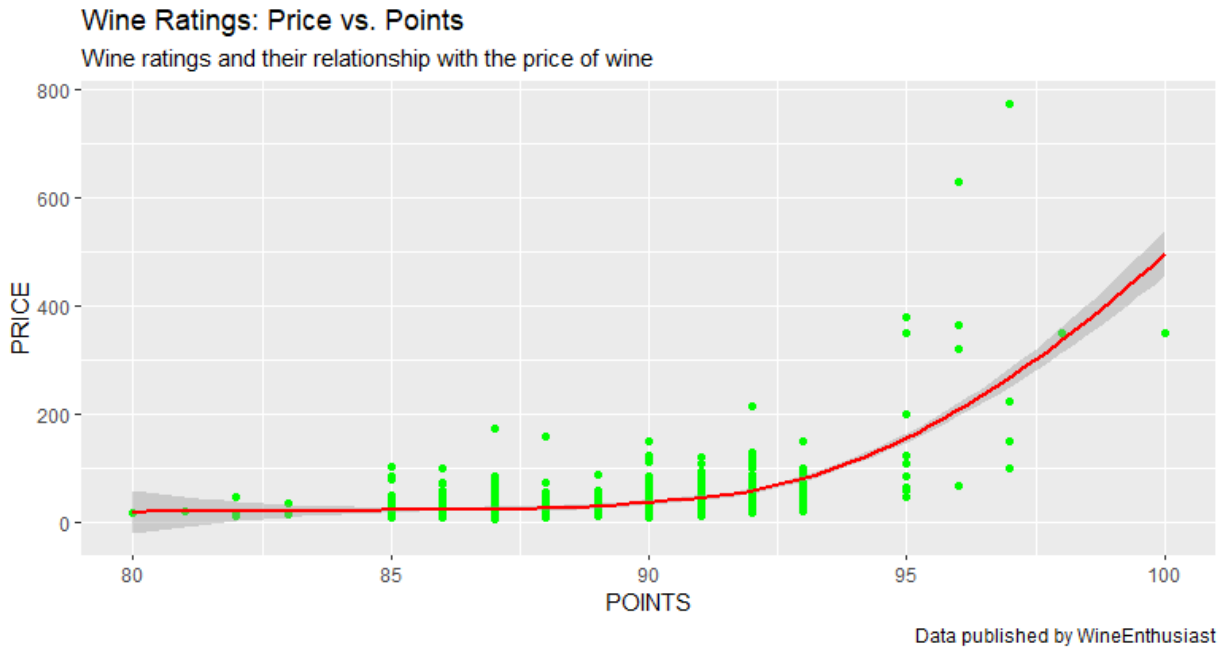
```
> wine_aly_2 <- table(wine_aly_2$POINTS)
> bar_graph2 <- barplot(wine_aly_2, main = "Point of Wine", ylim = c(0, 250), ylab = "Frequency",
+                       xlab = "POINT", col = heat.colors(6), cex.axis = 0.7, cex.names = 0.7)
> text(y = wine_aly_2, bar_graph2, wine_aly_2, cex = 0.8, pos = 3)
>
```



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

```
> arrange(avg_price_5, -mean_price)
  mean_price
1   40.19146
> ggplot(data = wine_aly_3) + geom_point(mapping=aes(x=POINTS,y=PRICE),color="Green") +
+   geom_smooth(mapping=aes(x=POINTS,y=PRICE),color="Red") +labs(title="Wine Ratings: Price vs. Points", subtitle="Wine r
r relationship with the price of wine", caption="Data published by WineEnthusiast")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
Warning messages:
1: Removed 57 rows containing non-finite values (stat_smooth).
2: Removed 57 rows containing missing values (geom_point).
>
```

```
> ggplot(data = wine_aly_3) + geom_point(mapping=aes(x=POINTS,y=PRICE),color="Green") +
+   geom_smooth(mapping=aes(x=POINTS,y=PRICE),color="Red") +labs(title="Wine Ratings: Price vs. Points", subtitle="Wine
r relationship with the price of wine", caption="Data published by WineEnthusiast")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
Warning messages:
1: Removed 57 rows containing non-finite values (stat_smooth).
2: Removed 57 rows containing missing values (geom_point).
>
```


Wine Ratings: Price vs. Points
Wine ratings and their relationship with the price of wine
800
600
400
200
0
PRICE
80
85
90
95
100
POINTS
Data published by WineEnthusiast

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

Appendix: R code

```
install.packages("tidyverse")
install.packages("dplyr")
install.packages("plyr")
install.packages("epiDisplay")
install.packages("gmodels")
install.packages("ggplot")

library(datasets)
library(tidyverse)
library(dplyr)
library(plyr)
library(psych)
library(epiDisplay)
library(skimr)
library(gmodels)
library(ggplot2)



wine_data <- read.csv("C:\\Abhinav _ NEU BOSTON\\ALY6010 Probability Theory & Intro Statistics\\Module 1 Submission\\Wine_tasting_20220222.csv")
wine_data
wine_new <- wine_data
wine_new[wine_new == ""] <- NA
wine_new


# Rename
wine_aly_rename1 <- rename_with(wine_new, toupper)
wine_aly_rename1

#Drop
wine_aly_drop <- subset(wine_aly_rename1, select = -c (DESCRIPTION, TITLE, TASTER_NAME, TASTER_TWITTER_HANDLE, WINERY, REGION_1, REGION_2))
wine_aly_drop

wine_aly_1 <- wine_aly_drop

#Structure of the dataset
str(wine_aly_1)

#Summary of the dataset
summary(wine_aly_1)

#Glimpse of the dataset
glimpse(wine_aly_1)

# BOX PLOT
boxplot(wine_aly_1$PRICE)

#Plot a frequency barchart of a price of wine
wine_aly_1 <- table(wine_aly_1$PRICE)
bar_graph1 <- barplot(wine_aly_1, main = "Price of Wine", ylim = c(0, 60), ylab = "Frequency",
                      xlab = "PRICE", col = heat.colors(6), cex.axis = 0.7, cex.names = 0.7)
text(y = wine_aly_1, bar_graph1, wine_aly_1, cex = 0.5, pos = 3)
```

```
#Plot a frequency barchart for points of wine
wine_aly_2 <- wine_aly_drop
wine_aly_2
wine_aly_2 <- table(wine_aly_2$POINTS)
bar_graph2 <- barplot(wine_aly_2, main = "Point of Wine", ylim = c(0, 250), ylab = "Frequency",
                      xlab = "POINT", col = heat.colors(6), cex.axis = 0.7, cex.names = 0.7)
text(y = wine_aly_2, bar_graph2, wine_aly_2, cex = 0.8, pos = 3)

# Through ggplot find highest price of provices
wine_aly_3 <- wine_aly_drop[!(is.na(wine_aly_drop$PROVINCE) | wine_aly_drop$PROVINCE==""), ]
wine_aly_3

avg_price <-  wine_aly_3 %>% group_by(PROVINCE) %>% drop_na() %>% summarize(mean_price = mean(PRICE))
arrange(avg_price, -mean_price)

province_reviews_5 <- wine_aly_3 %>% group_by(PROVINCE) %>% filter(n() >= 5)

avg_price_5 <- province_reviews_5 %>% group_by(PROVINCE) %>% drop_na() %>% summarize(mean_price = mean(PRICE))

arrange(avg_price_5, -mean_price)

ggplot(data = wine_aly_3) + geom_point(mapping=aes(x=POINTS,y=PRICE),color="Green") +
  geom_smooth(mapping=aes(x=POINTS,y=PRICE),color="Red") +
  labs(title="Wine Ratings: Price vs. Points",
       subtitle="Wine ratings and their relationship with the price of wine", caption="Data published by WineEnthusiast")
```