HIGH LEVEL DESIGN

WINE DATA ANALYSIS

WRITTEN BY	AKASH SINGH

The industry of wine is a very marketing driven business. Wine is now produced all over the world and the products are shipped globally. It is a natural product and the outcome depends heavily on the often on purpose indivualized production process. So as a result the products even from the same producer with the same name change every year. On the one hand the costumer is gaining more choice but on the other hand he is loosing the clear view and it is making the decision troublesome. So even sticking to one producers product can not guarantee you a standard if this season the grape has low quality. That is why in this opaque jungle reviews are so important. Nearly all buyers more or less rely on reviews because these are the only information they can get beside the metadata of the wine like name, origin etc. So in the past decades wine review itself became an industry. It produced so much data which can be mined for deductions and predictions which will be done here to make better decision when buying wine.

Data Under Standing

country(origin of wine in text)

description(description of the characteristics of the wine in text)

designation(vineyard of the graps in text)

points (rating in points from 80-100 ascending in quality as integer)

price(in US-Dollar as floating number)

province(province or state where wine is produced as text)

region_1(area in the province as text)

region_2(more specific area of the region_1 as text)

taster_name(name of the taster as text)

taster_twitter_handle(twitter handle of taster as text)

title(name of the wine as text)

variety(variety of the wine as text)

winery(name of the winery as text)

```
import pyforest

import plotly.express as px

sns.set_style('darkgrid')

matplotlib.rcParams['font.size'] = 15

matplotlib.rcParams['figure.figsize'] = (16,14)

matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

Uploading The Data and naming it as df

```
df=pd.read_csv("C:/Users/Akash/Desktop/Internship/wine reviews_small.csv")
```

Checking the number of rows & column in your dataset

```
print('The number of rows = {} and the number of columns = {}'.format(df.shape[0], df.shape[1]))
```

The number of rows = 29665 and the number of columns = 14

Checking The Types Of Data in our columns

```
df.info()
```

```
0   Unnamed: 0            29665 non-null  int64
 1   country               29650 non-null  object
 2   description           29665 non-null  object
 3   designation           21216 non-null  object
 4   points                29665 non-null  int64
 5   price                 27540 non-null  float64
 6   province              29650 non-null  object
 7   region_1              24699 non-null  object
 8   region_2              11411 non-null  object
 9   taster_name           23544 non-null  object
 10  taster_twitter_handle  22428 non-null  object
 11  title                 29665 non-null  object
 12  variety               29665 non-null  object
 13  winery                29665 non-null  object
```

```
df.head()
```

Dropping the unecessary columns in our data set

```
df.drop(columns=['Unnamed: 0'], inplace=True)
```

Checking For Missing Values In Our Set

```
df.isnull().sum()
```

```
country                15
description             0
designation          8449
```

points 0

price 2125

province 15

region_1 4966

region_2 18254

taster_name 6121

taster_twitter_handle 7237

title 0

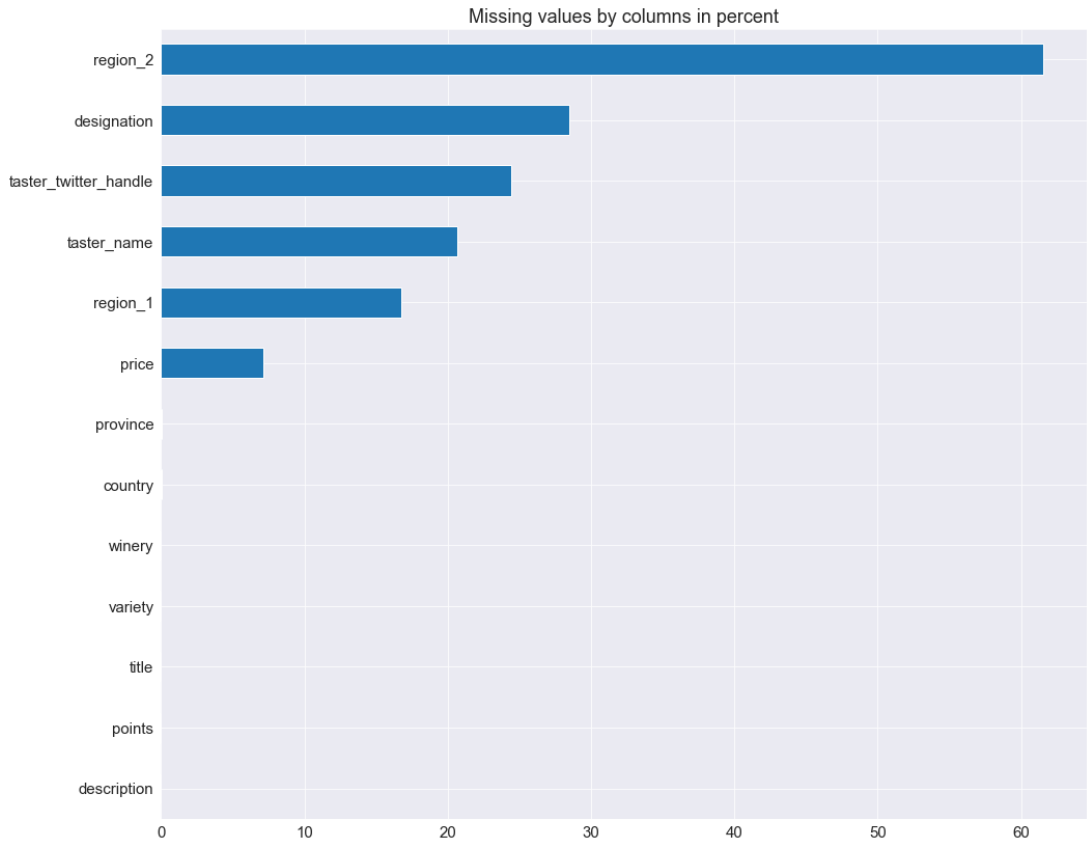
variety 0

winery 0

dtype: int64

Percentage of missing values in our columns

```
(df.isnull().sum()/len(df) * 100).sort_values().plot.barh(title = 'Missing values by columns in percent')
plt.show()
```



Data Preparation

In the beginning before answering the business questions the columns with more than 10 % missing values(region_1, region_2, taster_name, taster_twitter_handle, designation) are removed. These columns are text columns and have very individual and distinctive information for every review, so they cannot easily be imputed by for example the most common value. The price column contains too important information for statistics and modeling, so only the rows where the price is missing are deleted. For the columns country, province and variety the same is done but the number of these missing values are like I said before negligible.

#drop columns with more than 10 % of missing values

```
missing_percent = df.isnull().sum()/len(df) * 100

drop_list = missing_percent[missing_percent > 10].index.tolist()

data_columns_dropped = df.drop(drop_list, axis=1)

#dropping rows in missing values

df1= data_columns_dropped.dropna(axis=0)
```

data structure after cleaning

```
df1.info()
```

<class 'pandas.core.frame.DataFrame'> Int64Index: 27528 entries, 1 to 29664 Data columns (total 8 columns): # Column Non-Null Count Dtype --- ------ -------------- ----- 0 country 27528 non-null object 1 description 27528 non-null object 2 points 27528 non-null int64 3 price 27528 non-null float64 4 province 27528 non-null object 5 title 27528 non-null object 6 variety 27528 non-null object 7 winery 27528 non-null object dtypes: float64(1), int64(1), object(6) memory usage: 1.9+ MB

Exploration about the data.

Number Of Wine Makes By Countries

```
df2=df1.country.value_counts()

df2
```

US 12234

France 4014

Italy 3863

Spain 1486

Portugal 1156

Chile 1071

Argentina 893

Austria	654
Australia	494
Germany	482
New Zealand	341
South Africa	301
Israel	108
Greece	92
Canada	63
Uruguay	34
Romania	30
Bulgaria	29
Croatia	26
Hungary	23
Mexico	18
Georgia	18
Brazil	17
Turkey	16
England	11
Slovenia	10
Moldova	9
Lebanon	7
Peru	5
Czech Republic	4
Serbia	3
Morocco	3
Cyprus	3
India	2
Switzerland	2
Luxembourg	1

Armenia 1

Bosnia and Herzegovina 1

Ukraine 1

Slovakia 1

Macedonia 1

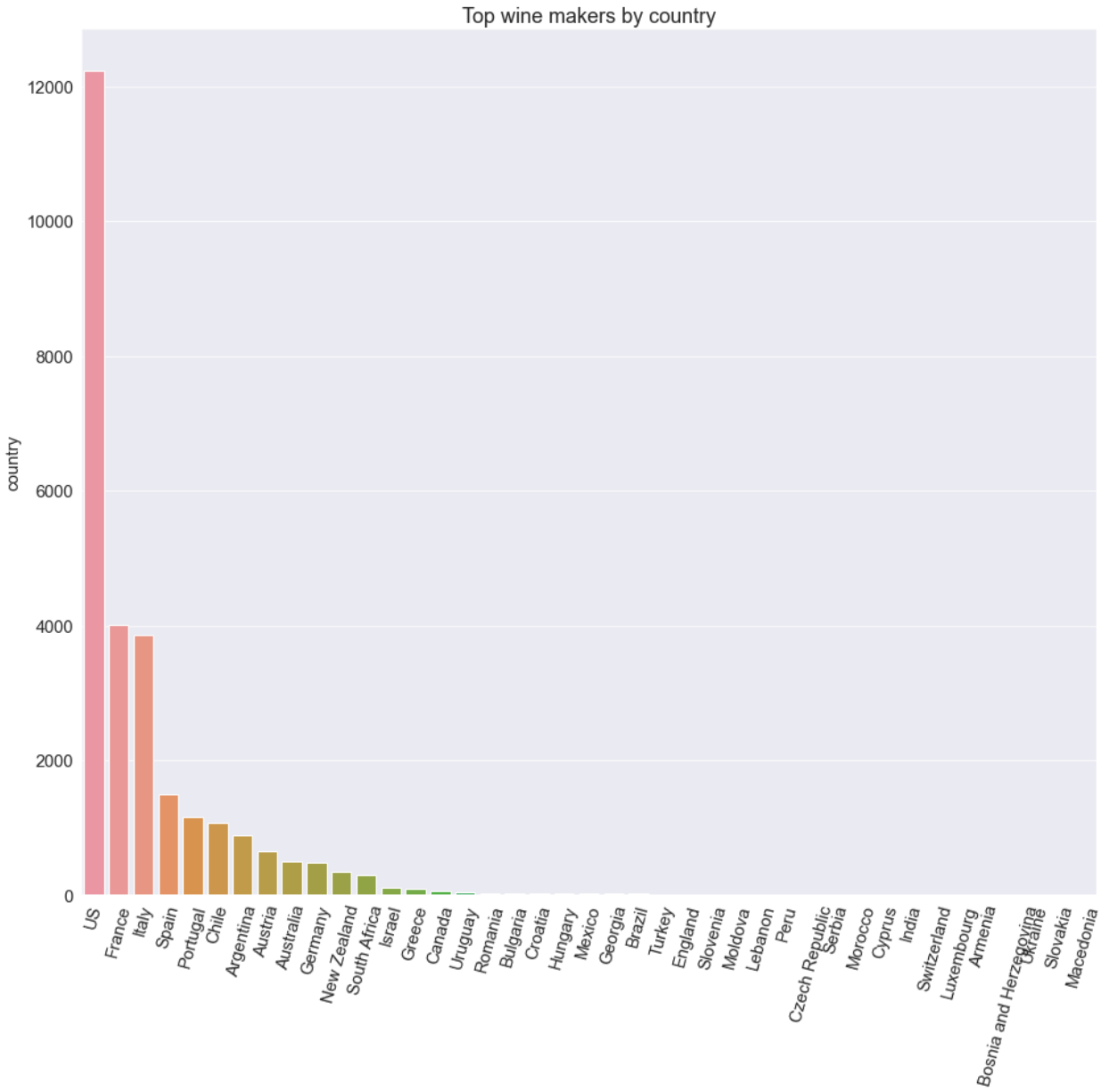
Name: country, dtype: int64

```
len(df2)
```

41

```
plt.xticks(rotation=75)

plt.title('Top wine makers by country')

sns.barplot(df2.index, df2);
```

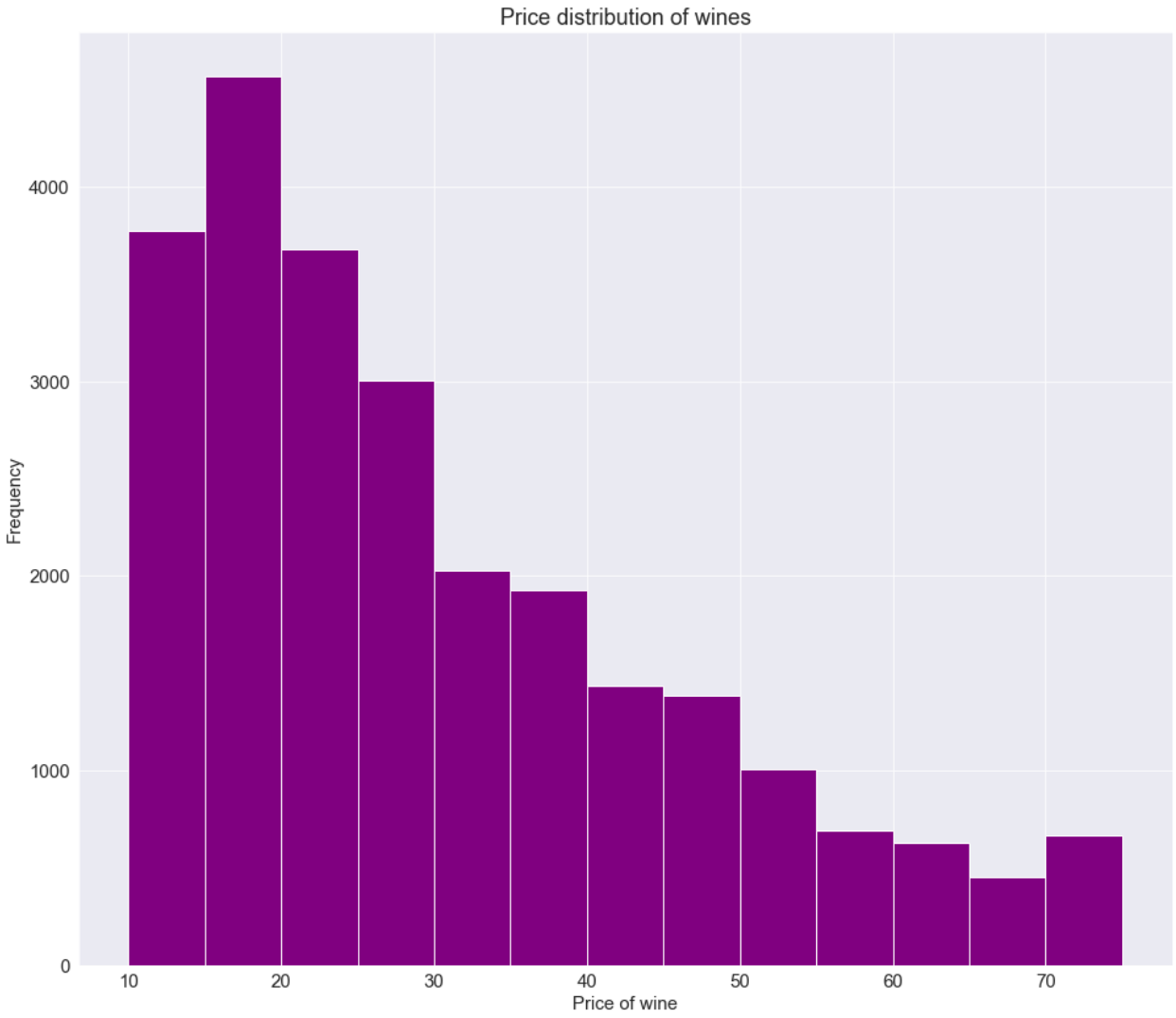
C:\Users\Akash\AppData\Local\Programs\Python\Python310\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation. warnings.warn(



USA is top of the list among the wine making countries at over 12000 units, followed by ITALY & FRANCE slighty bellow 4000 by volumn .

Wine Price Distribution

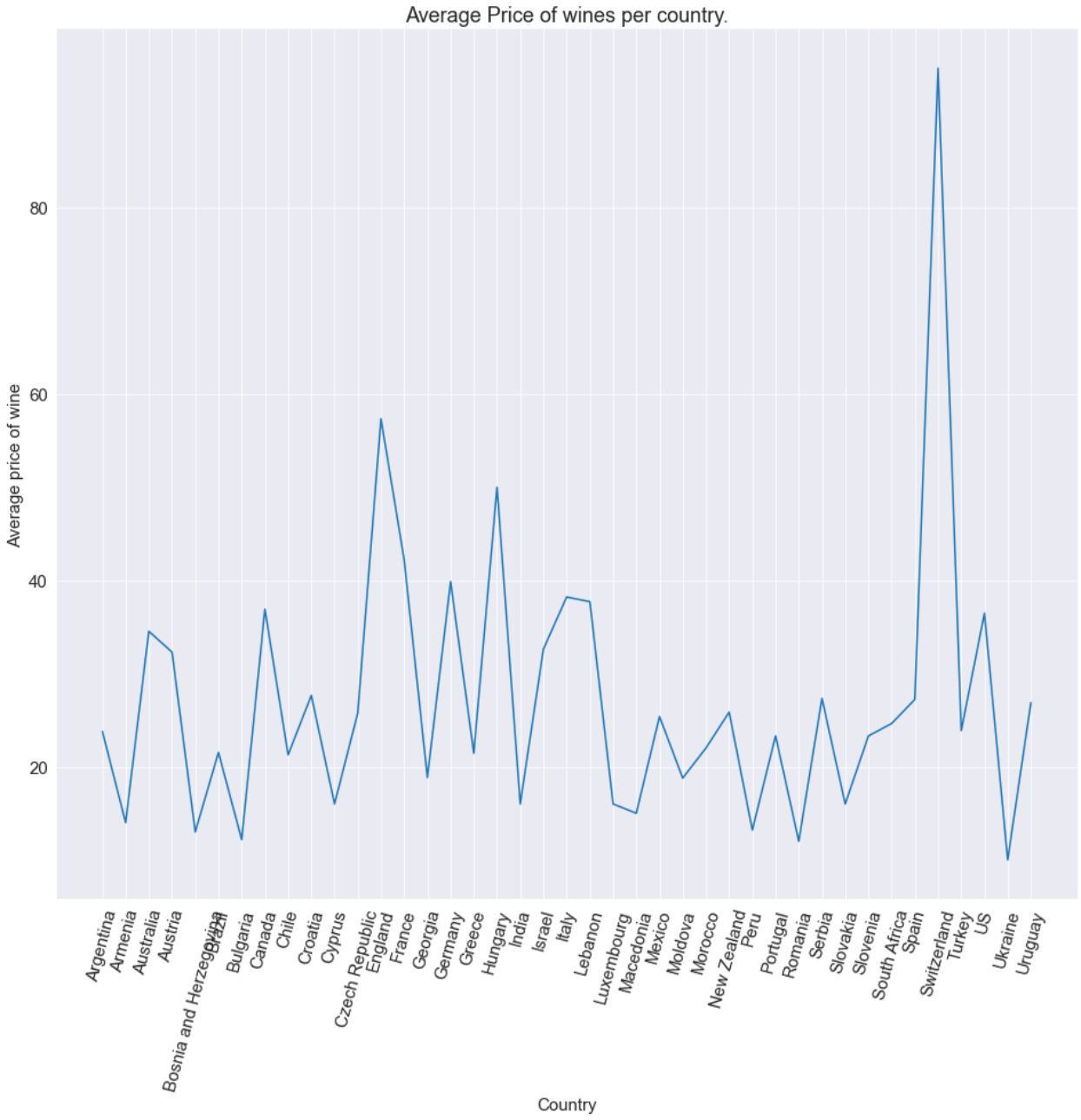
```
plt.title('Price distribution of wines')

plt.xlabel('Price of wine')

plt.ylabel('Frequency')

plt.hist(df1.price, bins=np.arange(10,80,5), color='purple');
```



Most of the wines are on lowerside of the price range

Average Price By Country

```
df3 = df1.groupby("country").mean()

df3

plt.title('Average Price of wines per country.')

plt.plot(df3 .index, df3.price)

plt.xticks(rotation=75)

plt.xlabel('Country')

plt.ylabel('Average price of wine');
```



Stzerland has The highest wine price which is just below 100 per unit while Armenia ,Bosnia and Herzegovina , Bulgaria , Peru , Romania have the lower price as compare to others which is just above 10 per unit and Ukrain has the lowest wine price below 10 per unit . Ther is no other countires which have unit price more than 60 except US. Hungary & England has uint price just below 60.Maximum Number Of coutries have the have there unit price between 20 to 40 .

Price Distribution of top 10 countries

```
df4=df1.groupby('country').price.agg(['count', 'min', 'max', 'mean']).reset_index().sort_values('count', ascending=False).head(10)
```

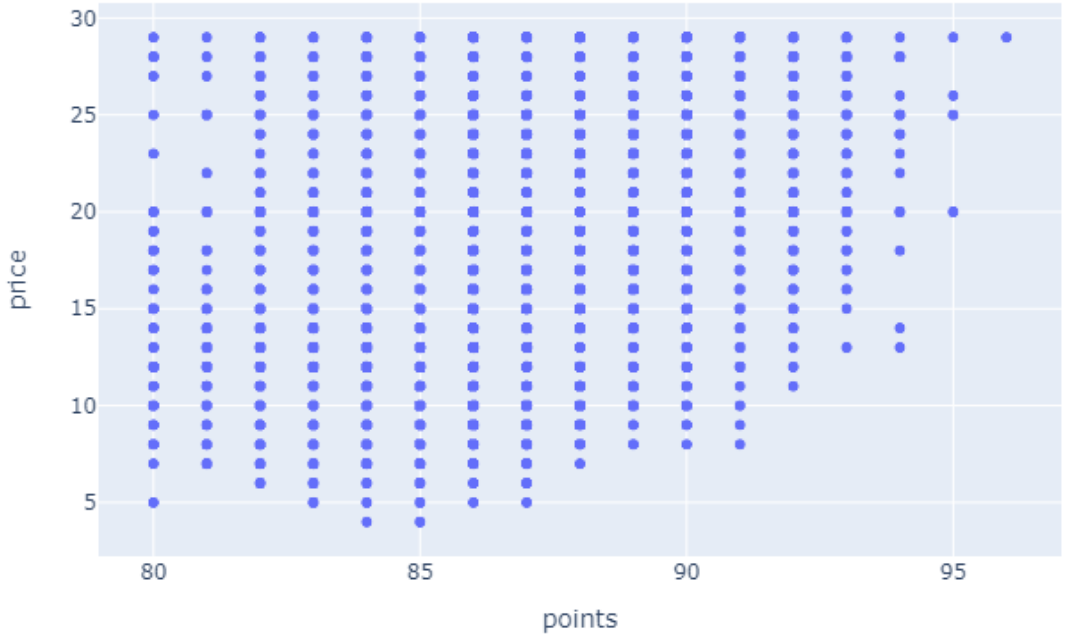
```
df4
```

As you can see while all the wines unit despite their origins start around 5 ,only French wines have an actual unit of more than 2,000

The higher the ratings points of the wine, the higher the price seems to be .

Correlation between Price & Point

```
px.scatter(df1[df1.price < 30], x = 'points', y = 'price')
```



Consider paying max 30/unit for a bottle which is quite acceptable, so here we can see the points begin at 80.

You can find a bottle at any price but to have a bottle at min 94 pts, you have to pay at least 15$

Q. Which countries have the top 5 highly rated wine?

```
df4=df1.sort_values('points',ascending=False).head(5)
```

```
df4
```

We can clearly see that Italy has the most rated wine in the world, followed closely by the USA and Australia.

Q.Which wine has the highest unit price per wine rating?

First I have to add a new column to hold the unit price per wine rating

```
df1['price_value_per_rating']=df1.points/df1.price
```

```
df1
```

```
C:\Users\Akash\AppData\Local\Temp\ipykernel_8936\889160114.py:1: SettingWithCopyWarning: A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row_indexer,col_indexer] = value instead See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
df5=df1.sort_values("price_value_per_rating",ascending=False)
```

```
df5
```

It's clear that the US wine is highly priced per unit rating. Spain,Argentina,Chile & Italyfollow in that order.

Q.Which 10 wine varieties are rated the least?

```
df6=df1.sort_values("points").head(10)
```

```
df6
```

The least rated wines all have a tie of 80 points each. Most of them are from Argentina.

Q. How many varities of wine are there in production all over the world ?

```
df7=df1.variety.value_counts()
```

```
print('There are {} different varieties of wines in the world.'.format(len(df7)))
```

There are 465 different varieties of wines in the world.

Q.Which wines are rated the highest in the world?

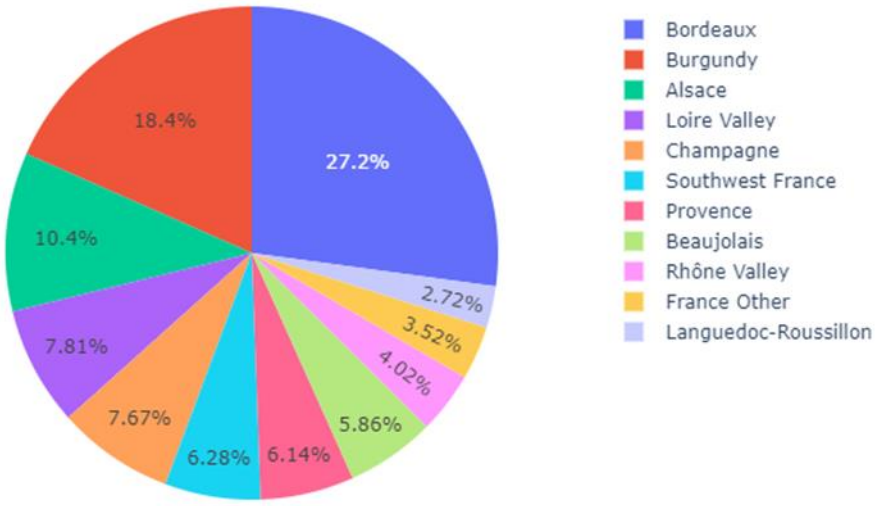
```
df8= df1.sort_values("points",ascending=False).head(4)df8
```

Italy's Prugnolo Gentile & Australia's Muscat are the top rated wine in the world.

Zoom in to French province

```
france = df[df.country == 'France'].groupby('province').size().reset_index(name='count').sort_values('count', ascending=False)
```

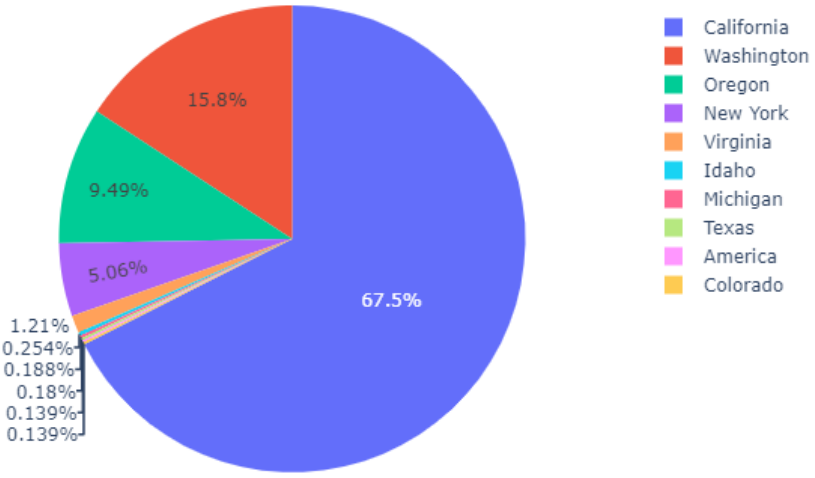
```
px.pie(france, names = 'province', values = 'count')
```



Although Bordeaux is the most popular province for wine, France has quite a list of good wines through the country.

US provinces

```
us = df[df.country == 'US'].groupby('province').size().reset_index(name='count').sort_values('count', ascending=False)

px.pie(us.head(10), names = 'province', values = 'count')
```



California is the place-to-be if you want to produce wine in US with over 67.5%, following with Washington and Oregon who cover already over 90% of the market.

Q. Which Variety of wine is mentioned here the most

```
df1.groupby('variety').size().reset_index(name='count').sort_values('count', ascending=False).head(5)
```

The variety Pinor Noir Wine Is Mentioned Here The most followed by Chardonnay wine .

INFERENCES & CONCLUSIONS

I've drawn many inferences from the data. Here's a summary of them .

About the dataset

The number of rows are 29665 and the number of columns are 14 .

Top wine making countries

USA is top of the list among the wine making countries at over 12000 units, followed by ITALY & FRANCE slighty bellow 4000 by volumn .

Wine Price Distribution

Most of the wines are on lowerside of the price range .

Stzerland has The highest wine price which is just below 100 per unit while Armenia ,Bosnia and Herzegovina , Bulgaria , Peru , Romania have the lower price as compare to others which is just above 10 per unit and Ukrain has the lowest wine price below 10 per unit . Ther is no other countires which have unit price more than 60 except US. Hungary & England has uint price just below 60.Maximum Number Of coutries have the have there unit price between 20 to 40 .

Top 10 countries as per wine price distribution

As you can see while all the wines unit despite their origins start around 5 ,only French wines have an actual unit of more than 2,000. The higher the ratings points of the wine, the higher the price seems to be .

Relation between the price & rating of a wine

After analyzing the dat we have consider paying max 30/unit for a bottle which is quite acceptable, so here we can see the points begin at 80. You can find a bottle at any price but to have a bottle at min 94 pts, you have to pay at least 15 unit .

Countries with top 5 highly rated wines

We can clearly see that Italy has the most rated wine in the world, followed closely by the USA and Australia .

Countries with highest unit price per wine rating

It's clear that the US wine is highly priced per unit rating. Spain,Argentina,Chile & Italy follow in that order.

About the 10 least rated wines

The least rated wines all have a tie of 80 points each. Most of them are from Argentina

total number of wines varities in the world

There are 465 different varieties of wines in the world.

About the world's highest rated wines

Italy's Prugnolo Gentile & Australia's Muscat are the top rated wine in the world.

Top Variety Wine mentioned in the most in the dataset

Italy's Prugnolo Gentile & Australia's Muscat are the top rated wine in the world.

About the French Province

Although Bordeaux is the most popular province for wine, France has quite a list of good wines through the country.

About The US Province

California is the place-to-be if you want to produce wine in US with over 67.5%, following with Washington and Oregon who cover already over 90% of the market.