All about Wine

March 12, 2021

1 All about Wine

Purpose To find out correlation of prices and points for each wine, and look for which country has the best wine of all. Also, calculate mean, min, and max points for countries Data can be found here: Kaggle

Key Performance Indicators: - Relationship between price and points - Highest wine points originated country - World choropleth

[ ]:

```

```

[1]:

```
import numpy as np
import pandas as pd
```

[2]:

```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import chart_studio.plotly as py
import plotly.graph_objects as go
from plotly import __version__
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot

init_notebook_mode(connected = True)
```

[3]:

```
import cufflinks as cf
```

[4]:

```
wine1 = pd.read_csv("winemag-data_first150k.csv")
wine2 = pd.read_csv("winemag-data-130k-v2.csv")
```

[5]:

```
wine = [wine1, wine2]
```

[6]:

```
winemag = pd.concat(wine)
```

[38]:

```
winemag.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 280901 entries, 0 to 129970
Data columns (total 14 columns):
 #   Column                 Non-Null Count   Dtype
```

```
---  ------                 --------------  -----
 0   Unnamed: 0             280901 non-null  int64
 1   country                280833 non-null  object
 2   description            280901 non-null  object
 3   designation            197701 non-null  object
 4   points                 280901 non-null  int64
 5   price                  258210 non-null  float64
 6   province               280833 non-null  object
 7   region_1               234594 non-null  object
 8   region_2               111464 non-null  object
 9   variety                280900 non-null  object
 10  winery                 280901 non-null  object
 11  taster_name            103727 non-null  object
 12  taster_twitter_handle  98758 non-null   object
 13  title                  129971 non-null  object
dtypes: float64(1), int64(2), object(11)
memory usage: 32.1+ MB
```

[39]:

```
winemag.head()
```

[39]:

```
   Unnamed: 0 country                                         description  \
0           0      US  This tremendous 100% varietal wine hails from …
1           1   Spain  Ripe aromas of fig, blackberry and cassis are …
2           2      US  Mac Watson honors the memory of a wine once ma…
3           3      US  This spent 20 months in 30% new French oak, an…
4           4  France  This is the top wine from La Bégude, named aft…

                            designation  points  price        province  \
0                     Martha's Vineyard      96  235.0      California
1  Carodorum Selección Especial Reserva      96  110.0  Northern Spain
2         Special Selected Late Harvest      96   90.0      California
3                               Reserve      96   65.0          Oregon
4                            La Brûlade      95   66.0        Provence

            region_1           region_2             variety  \
0        Napa Valley               Napa  Cabernet Sauvignon
1               Toro                NaN       Tinta de Toro
2     Knights Valley             Sonoma     Sauvignon Blanc
3  Willamette Valley  Willamette Valley          Pinot Noir
4             Bandol                NaN  Provence red blend

                    winery taster_name taster_twitter_handle title
0                    Heitz         NaN                   NaN   NaN
1  Bodega Carmen Rodríguez         NaN                   NaN   NaN
2                 Macauley         NaN                   NaN   NaN
3                    Ponzi         NaN                   NaN   NaN
4     Domaine de la Bégude         NaN                   NaN   NaN
```

Filter only for relevant columns and exclude non-numerical or non-categorical columns

[7]:
```
winemag = winemag[['country', 'points', 'price', 'province', 'region_1',␣
 ↪'region_2', 'variety']]
```

[65]:
```
#winemag['numbers'] = 1
```

[66]:
```
winemag.head()
```

[66]:
```
  country  points  price         province            region_1  \
0      US      96  235.0       California         Napa Valley
1   Spain      96  110.0  Northern Spain                Toro
2      US      96   90.0       California     Knights Valley
3      US      96   65.0           Oregon  Willamette Valley
4  France      95   66.0         Provence              Bandol

            region_2             variety  numbers
0               Napa  Cabernet Sauvignon        1
1                NaN       Tinta de Toro        1
2             Sonoma     Sauvignon Blanc        1
3  Willamette Valley          Pinot Noir        1
4                NaN  Provence red blend        1
```

[6]:
```
#winemag = winemag.rename(columns = {'Unnamed: 0': 'num'})
```

[42]:
```
winemag.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 280901 entries, 0 to 129970
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   country   280833 non-null  object
 1   points    280901 non-null  int64
 2   price     258210 non-null  float64
 3   province  280833 non-null  object
 4   region_1  234594 non-null  object
 5   region_2  111464 non-null  object
 6   variety   280900 non-null  object
dtypes: float64(1), int64(1), object(5)
memory usage: 17.1+ MB
```

[43]:
```
#How many unique countries on data
winemag['country'].nunique()
```

[43]:
```
50
```

[44]:

```
#unique countries array
winemag['country'].unique()
```

[44]:

```
array(['US', 'Spain', 'France', 'Italy', 'New Zealand', 'Bulgaria',
       'Argentina', 'Australia', 'Portugal', 'Israel', 'South Africa',
       'Greece', 'Chile', 'Morocco', 'Romania', 'Germany', 'Canada',
       'Moldova', 'Hungary', 'Austria', 'Croatia', 'Slovenia', nan,
       'India', 'Turkey', 'Macedonia', 'Lebanon', 'Serbia', 'Uruguay',
       'Switzerland', 'Albania', 'Bosnia and Herzegovina', 'Brazil',
       'Cyprus', 'Lithuania', 'Japan', 'China', 'South Korea', 'Ukraine',
       'England', 'Mexico', 'Georgia', 'Montenegro', 'Luxembourg',
       'Slovakia', 'Czech Republic', 'Egypt', 'Tunisia', 'US-France',
       'Peru', 'Armenia'], dtype=object)
```

[45]:

```
winemag['country'].value_counts()
```

[45]:

```
US                        116901
France                     43191
Italy                      43018
Spain                      14913
Portugal                   11013
Chile                      10288
Argentina                   9431
Australia                   7286
Austria                     6402
New Zealand                 4739
Germany                     4617
South Africa                3659
Greece                      1350
Israel                      1135
Canada                       453
Hungary                      377
Romania                      259
Bulgaria                     218
Uruguay                      201
Slovenia                     181
Croatia                      162
Turkey                       142
Mexico                       133
Moldova                      130
Georgia                      129
England                       83
Brazil                        77
Lebanon                       72
Cyprus                        42
Morocco                       40
Macedonia                     28
```

```
Serbia                        26
Ukraine                       19
Czech Republic                18
India                         17
Peru                          16
Luxembourg                    15
Switzerland                   11
Lithuania                      8
Bosnia and Herzegovina         6
Egypt                          4
South Korea                    4
China                          4
Slovakia                       4
Armenia                        2
Tunisia                        2
Albania                        2
Montenegro                     2
Japan                          2
US-France                      1
Name: country, dtype: int64
```

[8]:

```
count = winemag['country'].value_counts()
count = count.to_frame().reset_index()
count.rename(columns={'index': 'country', 'country': 'value'}, inplace=True)
count
```

[8]:

```
                   country   value
0                       US  116901
1                   France   43191
2                    Italy   43018
3                    Spain   14913
4                 Portugal   11013
5                    Chile   10288
6                Argentina    9431
7                Australia    7286
8                  Austria    6402
9              New Zealand    4739
10                 Germany    4617
11            South Africa    3659
12                  Greece    1350
13                  Israel    1135
14                  Canada     453
15                 Hungary     377
16                 Romania     259
17                Bulgaria     218
18                 Uruguay     201
19                Slovenia     181
```

```
20                 Croatia    162
21                  Turkey    142
22                  Mexico    133
23                 Moldova    130
24                 Georgia    129
25                 England     83
26                  Brazil     77
27                 Lebanon     72
28                  Cyprus     42
29                 Morocco     40
30               Macedonia     28
31                  Serbia     26
32                 Ukraine     19
33          Czech Republic     18
34                   India     17
35                    Peru     16
36              Luxembourg     15
37             Switzerland     11
38               Lithuania      8
39  Bosnia and Herzegovina      6
40                   Egypt      4
41             South Korea      4
42                Slovakia      4
43                   China      4
44                   Japan      2
45                 Armenia      2
46              Montenegro      2
47                 Tunisia      2
48                 Albania      2
49               US-France      1
```

[10]:

```
bywinery_multiple = winemag.groupby(['country']).agg({'points':['mean', 'min',␣
 ↪'max']})
bywinery_multiple.columns = ['points_mean', 'points_min', 'points_max']
bywinery_multiple = bywinery_multiple.reset_index()
bywinery_multiple
```

[10]:

```
                   country  points_mean  points_min  points_max
0                  Albania    88.000000          88          88
1                Argentina    86.283851          80          97
2                  Armenia    87.500000          87          88
3                Australia    88.112407          80         100
4                  Austria    89.707591          81          98
5   Bosnia and Herzegovina    85.333333          83          88
6                   Brazil    84.207792          80          89
7                 Bulgaria    87.064220          80          91
8                   Canada    88.880795          82          94
```

```
9                    Chile    86.382290           80          95
10                   China    83.750000           82          89
11                 Croatia    86.703704           81          91
12                  Cyprus    86.214286           80          89
13          Czech Republic    86.777778           84          89
14                   Egypt    83.750000           83          84
15                 England    91.722892           89          95
16                  France    88.884559           80         100
17                 Georgia    86.961240           81          92
18                 Germany    89.200996           80          98
19                  Greece    86.520000           80          93
20                 Hungary    88.050398           80          97
21                   India    89.000000           82          93
22                  Israel    87.752423           80          94
23                   Italy    88.481147           80         100
24                   Japan    85.000000           85          85
25                 Lebanon    86.666667           81          91
26               Lithuania    84.250000           84          85
27              Luxembourg    87.666667           86          90
28               Macedonia    85.678571           81          89
29                  Mexico    85.022556           80          92
30                 Moldova    85.846154           81          91
31              Montenegro    82.000000           82          82
32                 Morocco    88.450000           82          93
33             New Zealand    87.778434           80          95
34                    Peru    83.562500           80          86
35                Portugal    88.157178           80         100
36                 Romania    85.606178           80          92
37                  Serbia    87.615385           86          89
38                Slovakia    84.500000           82          87
39                Slovenia    88.154696           82          92
40            South Africa    87.543591           80          95
41             South Korea    81.500000           81          82
42                   Spain    86.932542           80          98
43             Switzerland    88.090909           83          90
44                 Tunisia    86.000000           85          87
45                  Turkey    88.091549           84          92
46                      US    88.166106           80         100
47               US-France    88.000000           88          88
48                 Ukraine    84.210526           82          88
49                 Uruguay    85.711443           80          92
```

[11]:

```
#Top 20 points by country
bywinery_multiple_20 = bywinery_multiple.nlargest(20, 'points_mean')
bywinery_multiple_20
```

[11]:

```
        country  points_mean  points_min  points_max
15      England    91.722892          89          95
4       Austria    89.707591          81          98
18      Germany    89.200996          80          98
21        India    89.000000          82          93
16       France    88.884559          80         100
8        Canada    88.880795          82          94
23        Italy    88.481147          80         100
32      Morocco    88.450000          82          93
46           US    88.166106          80         100
35     Portugal    88.157178          80         100
39     Slovenia    88.154696          82          92
3     Australia    88.112407          80         100
45       Turkey    88.091549          84          92
43  Switzerland    88.090909          83          90
20      Hungary    88.050398          80          97
0       Albania    88.000000          88          88
47    US-France    88.000000          88          88
33  New Zealand    87.778434          80          95
22       Israel    87.752423          80          94
27   Luxembourg    87.666667          86          90
```

[ ]:

```

```

Choropleth shows map figure but not identifying values – needs improvement and update

[12]:

```
data = dict(
        type = 'choropleth',
        locations = winemag['country'],
        z = winemag['price'],
        text = winemag['country'],
        colorbar = {'title' : 'winery spread world'},
      )
```

[13]:

```
layout = dict(
    title = 'Wine World Map',
    geo = dict(
        showframe = False,
        projection = {'type':'natural earth'}
    )
)
```

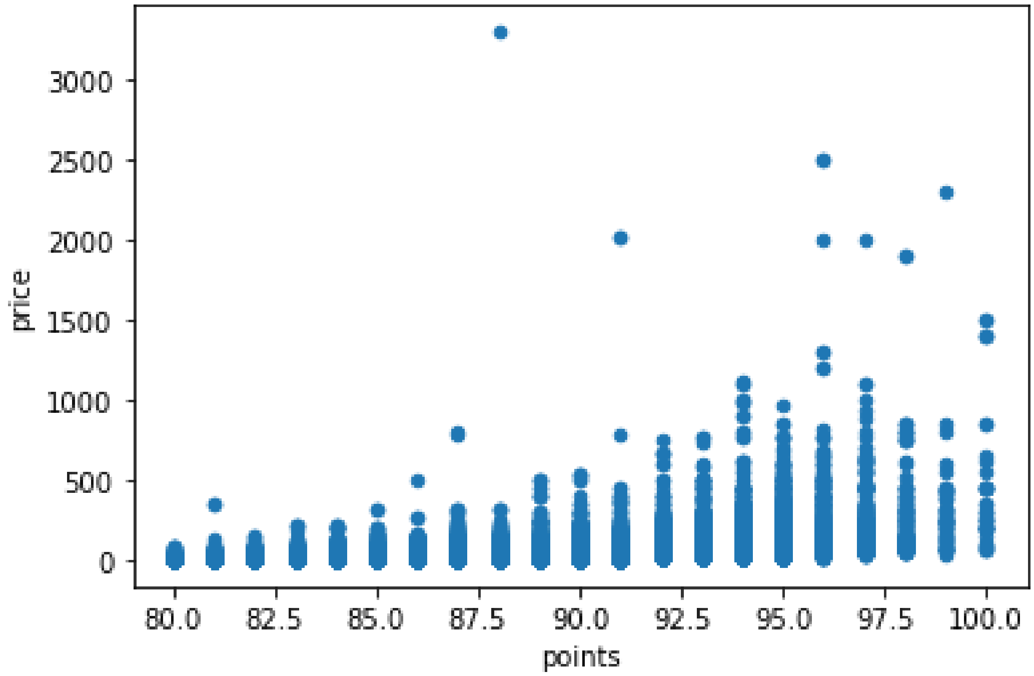
[14]:

```
choromap = go.Figure(data = [data],layout = layout)
iplot(choromap)
```

[19]:

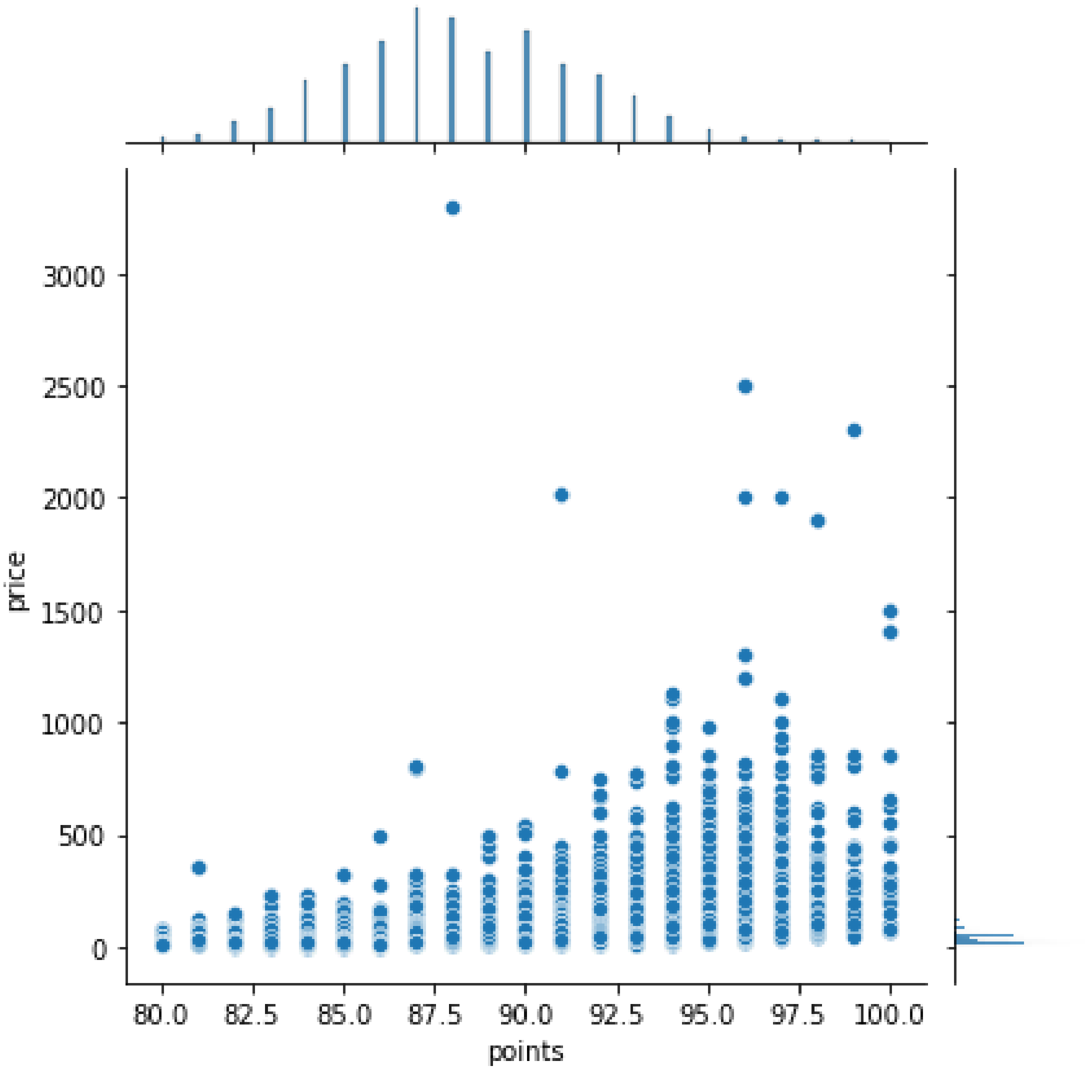
```
winemag.plot.scatter(x='points',y='price')
```

```
[19]: <AxesSubplot:xlabel='points', ylabel='price'>
```



```
[20]: sns.jointplot(x = 'points', y = 'price', data = winemag)
```

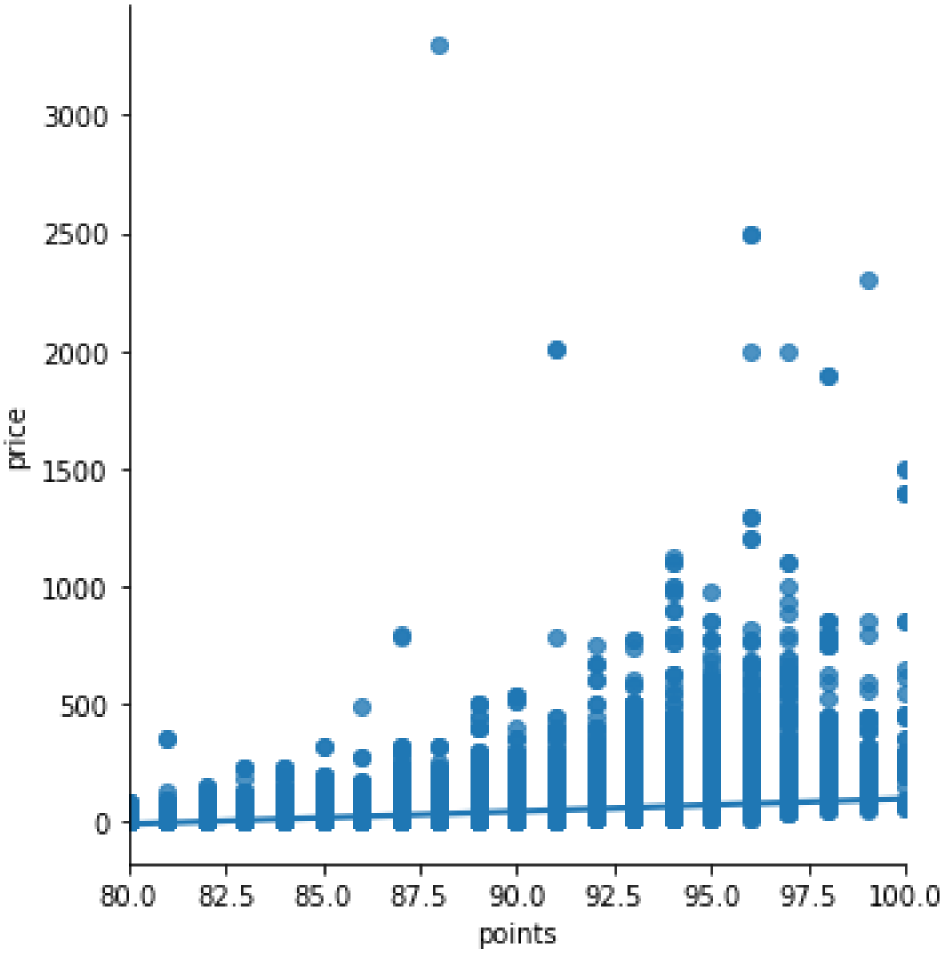
```
[20]: <seaborn.axisgrid.JointGrid at 0x7fa63d129460>
```



```
[26]: #sns.heatmap(winemag)
```

```
[22]: sns.lmplot(x='points',y='price',data=winemag)
```

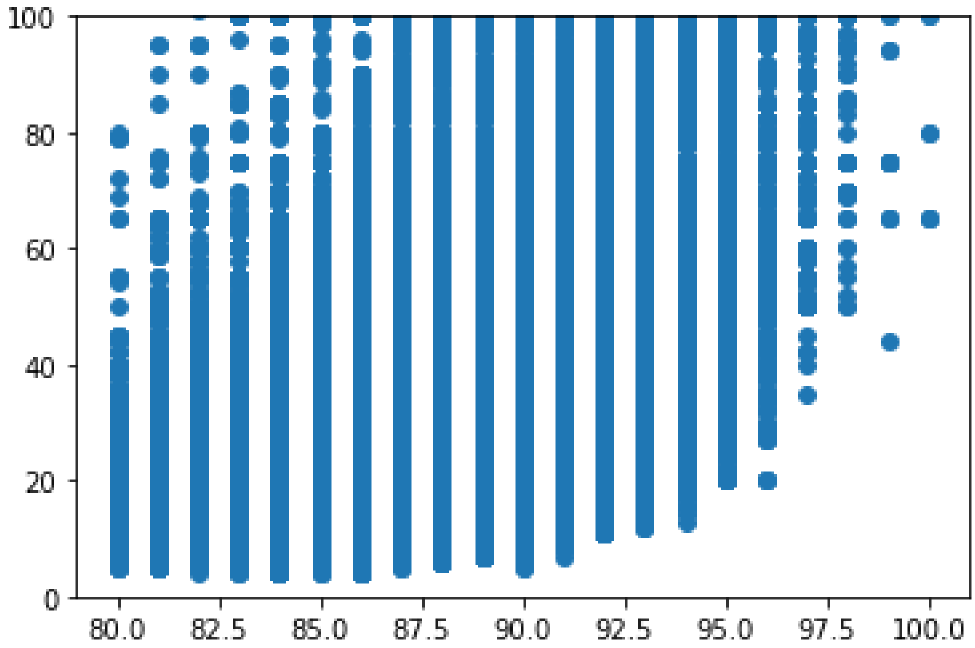
```
[22]: <seaborn.axisgrid.FacetGrid at 0x7fa6250c2d30>
```



```
[24]: fig = plt.figure()

      plt.scatter(x = 'points', y = 'price', data = winemag)
      plt.ylim([0,100])
```

```
[24]: (0.0, 100.0)
```

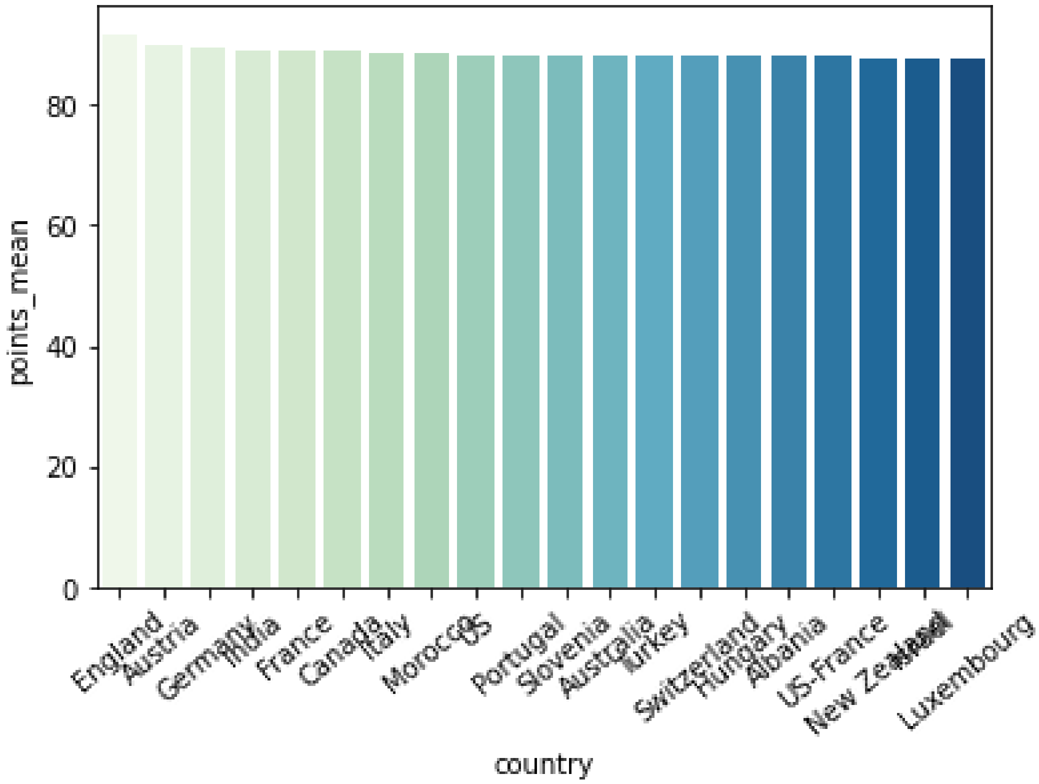


```
[15]: import plotly.express as px
      fig = px.bar(bywinery_multiple_20, x = 'country', y = 'points_mean')
      fig.show()
```

```
[16]: fig = px.scatter(winemag, x = 'price', y = 'points',
                      hover_name = 'country', log_x = True, size_max = 60)
      fig.show()
```

```
[18]: g = sns.barplot(x = 'country', y = 'points_mean', data = bywinery_multiple_20,␣
       ↪palette = 'GnBu')
      g.set_xticklabels(g.get_xticklabels(), rotation = 40)
```

```
[18]: [Text(0, 0, 'England'),
       Text(1, 0, 'Austria'),
       Text(2, 0, 'Germany'),
       Text(3, 0, 'India'),
       Text(4, 0, 'France'),
       Text(5, 0, 'Canada'),
       Text(6, 0, 'Italy'),
       Text(7, 0, 'Morocco'),
       Text(8, 0, 'US'),
       Text(9, 0, 'Portugal'),
       Text(10, 0, 'Slovenia'),
       Text(11, 0, 'Australia'),
       Text(12, 0, 'Turkey'),
       Text(13, 0, 'Switzerland'),
       Text(14, 0, 'Hungary'),
       Text(15, 0, 'Albania'),
       Text(16, 0, 'US-France'),
       Text(17, 0, 'New Zealand'),
       Text(18, 0, 'Israel'),
       Text(19, 0, 'Luxembourg')]
```



Conclusion

As you can see from Seaborn scatter plot, joint plot, lm (regression) plot, and Plotly scatter plot, wine price increases with its quality. High-quality wine costs a lot more than low-quality wine. Also, from separate data frame ‘bywinery_multiple_20’, it is shown top 20 points by country. Surprisingly, England takes first place while France and Italy which widely known as famous for wine sit 5th and 7th, respectively. However, when you look into data closely, England only has 83 unique data value counts (France: 43191, Italy: 43018 counts) and that lead to highest points (mean) overall. One thing I couldn’t solve was with choropleth map. I was able to show map figure and color scales, but couldn’t identify values. To do this, I added columns on dataframe to count, but didn’t work out. This part will be updated.

[ ]: