Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра ИУ5. Курс «Методы машинного обучения»

Отчет по лабораторной работе №1

«Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5-23Б
Белоусов Евгений
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю. Е.
Подпись и дата:

Москва, 2022 г.

Цель лабораторной работы:

изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание:

Построение графиков, помогающих понять структуру данных, и их интерпретация.

Основой лабораторной работы является методология визуализации данных data-to-viz

In [ ]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib
%matplotlib inline
```

In [ ]:

```
from os import path
data = pd.read_csv(path.join('archive', 'winemag-data_first150k.csv'), sep=',')
data.head()
```

Out[ ]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	NaN	Tinta de Toro	Bodega Carmen Rodríguez
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	NaN	Provence red blend	Domaine de la Bégude

In [ ]:

```
data = data.drop(columns=['region_2'], axis=1)
```

In [ ]:

```
data.head()
```

Out[ ]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	variety	winery
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Cabernet Sauvignon	Heitz
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	Tinta de Toro	Bodega Carmen Rodríguez
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sauvignon Blanc	Macauley
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Pinot Noir	Ponzi
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	Provence red blend	Domaine de la Bégude

In [ ]:

```
data = data.dropna()
```

In [ ]:

```
data.dtypes
```

Out[ ]:

```
Unnamed: 0       int64
country         object
description     object
designation     object
points           int64
price          float64
province        object
region_1        object
variety         object
winery          object
dtype: object
```

In [ ]:

```
data.describe()
```

Out[ ]:

	Unnamed: 0	points	price
count	77284.000000	77284.000000	77284.000000
mean	74493.686261	88.231678	37.584817
std	43712.830627	3.303169	36.403885
min	0.000000	80.000000	4.000000
25%	36034.750000	86.000000	18.000000
50%	74450.000000	88.000000	29.000000
75%	112735.750000	91.000000	45.000000
max	150928.000000	100.000000	2013.000000

Самое дорогое вино из списка

In [ ]:

```
max_price = data['price'].max()
most_expensive_wine = data.loc[(data['price'] == max_price)]
most_expensive_wine
```

Out[ ]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	variety	winery
13318	13318	US	The nose on this single-vineyard wine from a s...	Roger Rose Vineyard	91	2013.0	California	Arroyo Seco	Chardonnay	Blair

In [ ]:

```
# define colors
GRAY1, GRAY2, GRAY3 = '#231F20', '#414040', '#555655'
GRAY4, GRAY5, GRAY6 = '#646369', '#76787B', '#828282'
GRAY7, GRAY8, GRAY9 = '#929497', '#A6A6A5', '#BFBEBE'
BLUE1, BLUE2, BLUE3, BLUE4 = '#174A7E', '#4A81BF', '#94B2D7', '#94AFC5'
RED1, RED2 = '#C3514E', '#E6BAB7'
GREEN1, GREEN2 = '#0C8040', '#9ABB59'
ORANGE1 = '#F79747'

# configure plot font family to Arial
plt.rcParams['font.family'] = 'Arial'
# configure mathtext bold and italic font family to Arial
matplotlib.rcParams['mathtext.fontset'] = 'custom'
matplotlib.rcParams['mathtext.bf'] = 'Arial:bold'
matplotlib.rcParams['mathtext.it'] = 'Arial:italic'
```

In [ ]:

```
# create new figure
plt.figure(figsize=(7.45, 4.9), # width, height in inches
           dpi=110)             # resolution of the figure

# remove chart border
for spine in plt.gca().spines.values():
    spine.set_visible(False)
```

```
# change the appearance of ticks, tick labels, and gridlines
plt.tick_params(bottom='off', left='off', labelleft='off', labelbottom='off')

plt.axis('off')

# titile the plot
plt.text(-0.15, 1.03,
         'The most expensive wine:'+' '*27,
         fontsize=26,
         color='white',
         # put a rectangular box around the text to set a background color
         bbox={'facecolor': GRAY7, 'pad': 10, 'edgecolor': 'none'})

# add note to the plot
plt.text(-0.15, 0.81,
         '{} · {} · {}'.format(most_expensive_wine['winery'].values[0], most_expensive_wine['variety'].values
         fontsize=19,
         color=BLUE2)

# add note to the plot
plt.text(-0.15, 0.41,
         '{}$'.format(round(most_expensive_wine['price'].values[0])), # use mathtext \\bf for bold text
         fontsize=122,
         color=BLUE2)

description_len = len(most_expensive_wine['description'].values[0])

# add note to the plot
plt.text(-0.15, 0.25,
         # use mathtext \\bf for bold text
         '{}'.format(most_expensive_wine['description'].values[0][:description_len//3]),
         fontsize=14,
         color=GRAY9)

plt.text(-0.15, 0.17,
         # use mathtext \\bf for bold text
         '{}'.format(most_expensive_wine['description'].values[0][description_len//3:2*description_len//3]),
         fontsize=14,
         color=GRAY9)

plt.text(-0.15, 0.09,
         # use mathtext \\bf for bold text
         '{}'.format(most_expensive_wine['description'].values[0][2*description_len//3:]),
         fontsize=14,
         color=GRAY9)

# add note to the plot
plt.text(-0.15, 0.02,
         '{} · {} · {}'.format(most_expensive_wine['country'].values[0], most_expensive_wine['province'].valu
         fontsize=19,
         color=GRAY7)
```

Out[ ]:

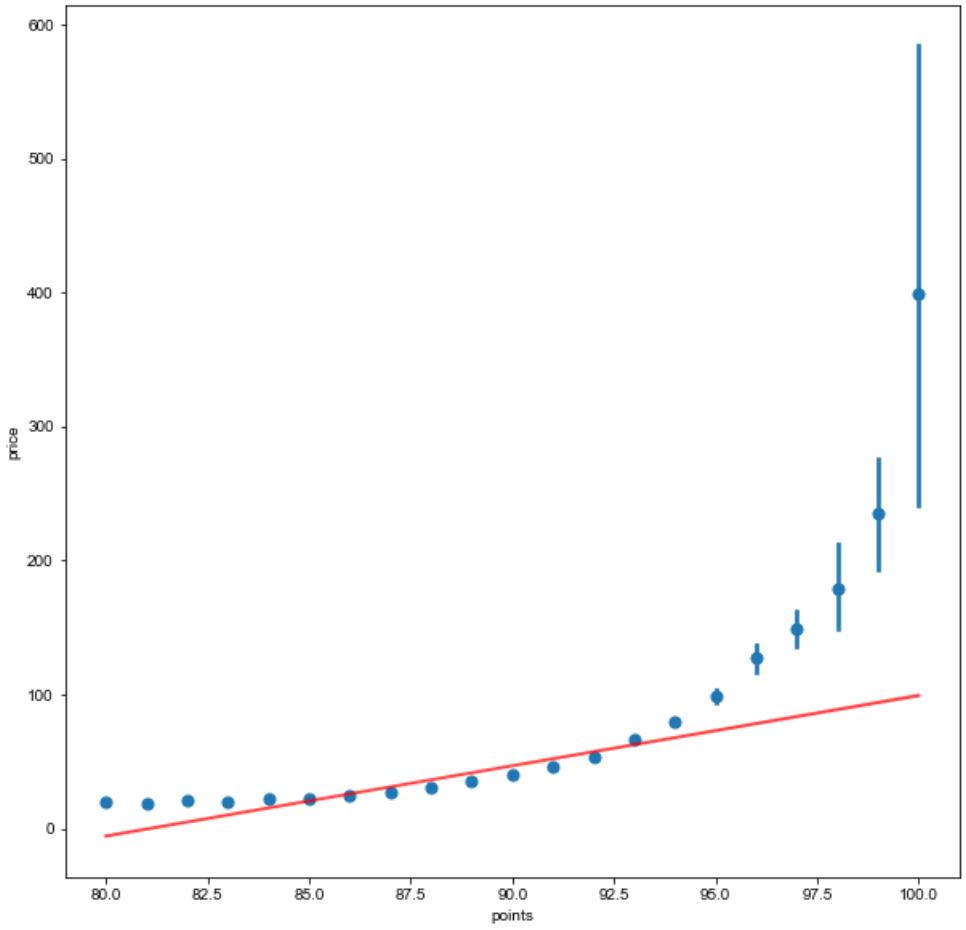
```
Text(-0.15, 0.02, 'US · California · Arroyo Seco')
```

Самым дорогим вином в датасете является Blair Chardonnay и стоит 2013 баксов.

Зависимость стоимости вина от его оценки

In [ ]:

```
g = sns.regplot(x=data['points'], y=data['price'], x_estimator=np.mean, line_kws={"color":"r","alpha":0.7,"lw"
g.get_figure().set_size_inches(10, 10)
plt.show()
```



Стобальные вина в среднем стоят заметно дороже остальных (хотя это может быть вызвано несколькими особенно дорогими винами). При этом в датасете присутствуют 90 бальные вина примерно за 20 долларов.

Посмотрим какие страны получают более высокие оценки

In [ ]:

```
data['country'].unique()
```

Out[ ]:

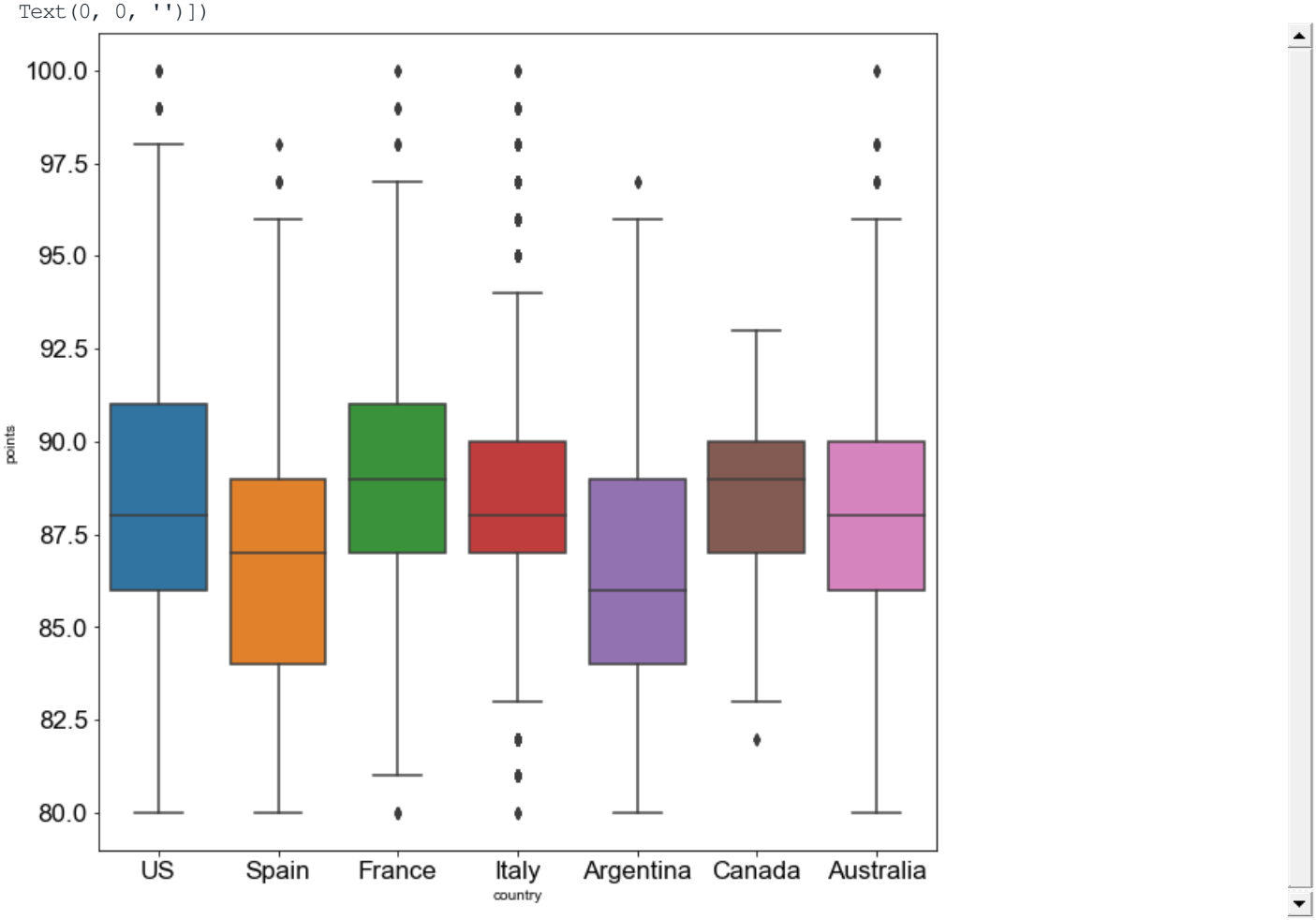
```
array(['US', 'Spain', 'France', 'Italy', 'Argentina', 'Canada',
       'Australia'], dtype=object)
```

In [ ]:

```
g = sns.boxplot( y=data['points'], x=data['country'] )
g.get_figure().set_size_inches(10, 10)
plt.xticks(fontsize='17')
plt.yticks(fontsize='17')
```

Out[ ]:

```
(array([ 77.5,  80. ,  82.5,  85. ,  87.5,  90. ,  92.5,  95. ,  97.5,
        100. , 102.5]),
 [Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, ''),
  Text(0, 0, '')])
```



Франция!!

Более высокие оценки эксперты отдают винам из Франции

Посмотрим какие сорта стоят дороже

In [ ]:

```
data['variety'].unique().shape
```

Out[ ]:

```
(427,)
```

In [ ]:

```
data['variety'].unique()
```

Out[ ]:

```
array(['Cabernet Sauvignon', 'Tinta de Toro', 'Sauvignon Blanc',
       'Pinot Noir', 'Provence red blend', 'Friulano', 'Tannat',
       'Chardonnay', 'Tempranillo', 'Malbec', 'Rosé', 'Tempranillo Blend',
       'Syrah', 'Sparkling Blend', 'Sangiovese', 'Red Blend', 'Mencía',
       'Palomino', 'Riesling', 'Cabernet Sauvignon-Syrah', 'Nebbiolo',
       'Meritage', 'Glera', 'Malbec-Merlot', 'Merlot-Malbec',
       'Ugni Blanc-Colombard', 'Cabernet Sauvignon-Cabernet Franc',
       'Moscato', 'Pinot Grigio', 'Zinfandel', 'White Blend', 'Barbera',
       'Grenache', 'Rhône-style Red Blend', 'Albariño',
       'Bordeaux-style Red Blend', 'Viognier', 'Picpoul', 'Godello',
       'Cabernet Franc', 'G-S-M', 'Mourvèdre', 'Petit Verdot',
       'Rhône-style White Blend', 'Muscat', 'Cabernet Sauvignon-Merlot',
       'Pinot Bianco', 'Garganega', 'Sauvignon', 'Tannat-Cabernet',
       'Ugni Blanc', 'Grüner Veltliner', 'Sylvaner', 'Chasselas',
       'Alsace white blend', 'Merlot', 'Vermentino', 'Sherry',
       'Primitivo', 'Grenache-Syrah', 'Pinot Blanc', 'Pinot Gris',
       'Gewürztraminer', 'Torrontés', 'Malbec-Cabernet Sauvignon',
       'Gros Manseng', 'Nerello Mascalese', 'Shiraz', 'Champagne Blend',
       'Romorantin', 'Pinot Nero', 'Tannat-Merlot', 'Duras', 'Garnacha',
       'Bordeaux-style White Blend', 'Gamay', 'Turbiana', 'Monastrell',
       'Roussanne', 'Touriga Nacional', 'Pinot Auxerrois', 'Port',
       'Cabernet Blend', 'Colombard-Sauvignon Blanc', 'Moscatel',
       'Marsanne', 'Blaufränkisch', 'Garnacha Blanca',
       'Merlot-Cabernet Sauvignon', 'Melon', 'Carricante',
       'Sangiovese-Syrah', 'Cabernet Franc-Merlot',
       'Sauvignon Blanc-Semillon', 'Chenin Blanc', 'Macabeo',
       'Grenache Blanc', 'Ciliegiolo', 'Petite Sirah', 'Auxerrois',
       'Alicante Bouschet', 'Aglianico', 'Negrette', 'Rosado',
       'Carignane', 'Grillo', 'Charbono', 'Trepat', 'Trebbiano',
       'Pinot Noir-Gamay', 'Chardonnay-Viognier', 'Syrah-Mourvèdre',
       'Graciano', 'Syrah-Cabernet Sauvignon', 'Fiano', 'Falanghina',
       'Carignan', 'Cabernet-Shiraz', 'Verdelho', 'Pedro Ximénez',
       'Malbec Blend', 'Catarratto', 'Greco',
       'Chardonnay-Sauvignon Blanc', 'Vidal', 'Chenin Blanc-Chardonnay',
       'Tempranillo-Cabernet Sauvignon', 'Petite Verdot',
       'Pinot Noir-Syrah', 'Gamay Noir', 'Cannonau', 'Mauzac',
       'Gros and Petit Manseng', 'Lambrusco di Sorbara', 'Lemberger',
       'Cinsault', 'Teroldego', 'Frappato', 'Malbec-Petit Verdot',
       'Veltliner', 'Rosato', 'Lambrusco', 'Cabernet Sauvignon-Shiraz',
       'Tocai Friulano', 'Verdejo', 'Fer Servadou', 'Nerello Cappuccio',
       "Nero d'Avola", 'Dolcetto', 'Malbec-Tannat', 'Hondarrabi Zuri',
       'Syrah-Merlot', 'Tinto Fino', 'Montepulciano', 'Prié Blanc',
       'Chardonnay-Semillon', 'Cabernet Sauvignon-Sangiovese', 'Viura',
       'Garnacha-Syrah', 'Zibibbo', 'Xarel-lo', 'Inzolia',
       'Cabernet-Syrah', 'Lambrusco Grasparossa', 'Syrah-Grenache',
       'Cabernet Franc-Malbec', 'Tempranillo-Shiraz', 'Tinta Fina',
       'Zweigelt', 'Colombard-Ugni Blanc', 'Müller-Thurgau',
       'Grenache-Carignan', 'Orange Muscat', 'Vignoles',
       'Carignan-Grenache', 'Fumé Blanc', 'Bobal', 'Norton', 'Rkatsiteli',
       'Roussanne-Viognier', 'Shiraz-Viognier', 'Bonarda', 'Traminette',
       'Semillon-Sauvignon Blanc', 'Grenache Blend', 'Jaen', 'Mondeuse',
       'Carmenère', 'Teroldego Rotaliano',
       'Sangiovese-Cabernet Sauvignon', 'Syrah-Petite Sirah', 'Jacquère',
       'Sémillon', 'Tinto del Pais', 'Mission', 'Carineña',
       'Garnacha-Tempranillo', 'Pecorino', 'Negroamaro', 'Cococciola',
       'Passerina', 'Chambourcin', 'Gaglioppo', 'Garnacha Tintorera',
       'Viognier-Chardonnay', 'Tempranillo Blanco', 'Aligoté',
       "Cesanese d'Affile", 'Muscat Canelli', 'Malvasia Nera', 'Premsal',
       'Sauvignon Blanc-Chardonnay', 'Petit Manseng', 'Verdicchio',
       'Sagrantino', 'Counoise', 'Mantonico', 'Grecanico',
       'Cariñena-Garnacha', 'Malvasia', 'Cabernet Sauvignon-Malbec',
       'Shiraz-Grenache', 'Claret', 'Syrah-Tempranillo',
```

```
'Malvasia Bianca', 'Viognier-Marsanne',
'Pinot Grigio-Sauvignon Blanc', 'Pallagrello Nero',
'Chardonnay-Albariño', 'Savagnin', 'Nero di Troia',
'Ribolla Gialla', 'Pinotage', 'Carignano', 'Vidal Blanc',
'Vernaccia', 'Corvina, Rondinella, Molinara', 'Pinot Meunier',
'Garnacha-Monastrell', 'Cabernet Merlot', 'Malbec-Tempranillo',
'Uva di Troia', 'Verdeca', 'Insolia', 'Garnacha-Cabernet',
'Sangiovese Grosso', 'Merlot-Cabernet Franc', 'Maturana', 'Malvar',
'Airen', 'Monica', 'Lagrein', 'Shiraz-Cabernet Sauvignon',
'Picolit', 'Prosecco', 'White Riesling',
'Cabernet Sauvignon-Carmenère', 'Tempranillo-Garnacha',
'Perricone', 'Vidadillo', 'Syrah-Cabernet', 'Traminer', 'Arneis',
'Cortese', 'Moscato Giallo', 'Torbato', 'Debit', 'Bovale',
'Shiraz-Tempranillo', 'Mansois', 'Merlot-Cabernet', 'Black Muscat',
'Kerner', 'Pallagrello', 'Muscat Blanc', 'Schiava',
'Monastrell-Syrah', 'Trebbiano di Lugana', 'Raboso', 'Colombard',
'Tannat-Cabernet Franc', 'Greco Bianco', 'Tokay', 'Muscadel',
'Scheurebe', 'Tintilia ', 'Piedirosso', 'Segalin', 'Lacrima',
'Cayuga', 'Prieto Picudo', 'Corvina', 'Macabeo-Moscatel',
'Moscadello', 'Albana', 'Viognier-Roussanne', 'Prugnolo Gentile',
'Verduzzo', 'Albarín', 'Syrah-Viognier', 'Aleatico',
'Morio Muskat', 'Alicante', 'Marsanne-Roussanne',
'Gewürztraminer-Riesling', 'Casavecchia', 'Malvasia-Viura',
'Nosiola', 'Incrocio Manzoni', 'Cabernet Sauvignon-Tempranillo',
'Viura-Verdejo', 'Dornfelder', 'Erbaluce', 'Pansa Blanca',
'Catalanesca', 'Cabernet', 'Verdejo-Viura', 'Cabernet Pfeffer',
'Syrah-Cabernet Franc', 'Valdiguié', 'Mazuelo', 'Brachetto',
'Jacquez', 'Chardonnay-Sauvignon', 'Madeleine Angevine', 'Ruché',
'Moscatel de Alejandría', 'Doña Blanca',
'Roussanne-Grenache Blanc', 'Muscadelle', 'Malbec-Syrah',
'Picapoll', 'Roussanne-Marsanne', 'Pugnitello',
'Provence white blend', 'Carignan-Syrah', 'Albarossa',
'Chenin Blanc-Viognier', 'Baco Noir', 'Sauvignon Blanc-Verdejo',
"Loin de l'Oeil", 'Rolle', 'Verdejo-Sauvignon Blanc',
'Grenache-Mourvèdre', 'Braucol', 'Tocai Rosso', 'Pinot-Chardonnay',
'Pigato', 'Bombino Bianco', 'Trebbiano-Malvasia', 'Magliocco',
'Verduzzo Friulano ', 'Vespaiolo', 'Seyval Blanc', 'Marzemino',
'Tempranillo-Malbec', 'Viura-Chardonnay', 'Crespiello',
'Cabernet Franc-Tempranillo', 'Tempranillo-Merlot',
'Shiraz-Mourvèdre', 'Roviello', 'Caprettone', 'Garnacha-Graciano',
'Mataro', 'Symphony', 'Nasco', 'Coda di Volpe',
'Pallagrello Bianco', 'Grenache-Shiraz', 'Pelaverga Piccolo',
'Touriga Franca', 'Nuragus', 'Alvarelhão', 'Durif', 'Angevine',
'Pinot Blanc-Pinot Noir', 'Manzoni', 'Johannisberg Riesling',
'Silvaner', 'Malvasia Istriana', 'Susumaniello',
'Macabeo-Chardonnay', 'Shiraz-Malbec', 'Pignoletto',
'Cabernet Franc-Cabernet Sauvignon', 'Freisa', 'Petite Syrah',
'Pinot Blanc-Chardonnay', 'Roscetto', 'Malbec-Bonarda', 'Grolleau',
'Gragnano', 'Ansonica', 'Sangiovese Cabernet', 'Syrah-Bonarda',
'Durello', 'Marsanne-Viognier', 'Malbec-Cabernet Franc', 'Rufete',
'St. Vincent', 'Groppello', 'Saperavi', 'Muscat of Alexandria',
'Muscat Blanc à Petit Grain', 'Merlot-Grenache', 'Grechetto',
'Macabeo-Gewürztraminer', 'Grenache Gris', 'Muscat Hamburg',
"Muscat d'Alexandrie", 'Merlot-Syrah', 'Semillon-Chardonnay',
'Chardonnay-Pinot Gris', 'Pardina', 'Apple', 'Clairette',
'Refosco', 'Sauvignon Musqué', 'Cabernet Sauvignon Grenache',
'Shiraz-Merlot', 'Chardonelle', 'Muscadet',
'Viura-Sauvignon Blanc', 'Tocai', 'Tokay Pinot Gris',
'Chardonnay-Pinot Grigio'], dtype=object)
```

Так как в датасете представлено 427 сортов винограда, а человек способен адекватно воспринимать только около 7, у нас не получится представить их всех адекватно. Будем рассматривать только красные вина из 4 сортов, которые вызывают у меня наибольший интерес.

In [ ]:

```
data_variety = data.loc[(data['variety'].isin(['Cabernet Sauvignon', 'Pinot Noir', 'Barbera', 'Merlot']))]
data_variety.shape
```

Out[ ]:

```
(17631, 10)
```

In [ ]:

```
data_variety.head()
```

Out[ ]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	variety	winery
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Cabernet Sauvignon	Heitz
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Pinot Noir	Ponzi
8	8	US	This re-named vineyard was formerly bottled as...	Silice	95	65.0	Oregon	Chehalem Mountains	Pinot Noir	Bergström
9	9	US	The producer sources from two blocks of the vi...	Gap's Crown Vineyard	95	60.0	California	Sonoma Coast	Pinot Noir	Blue Farm
11	11	US	From 18-year-old vines, this supple well-balan...	Estate Vineyard Wadensvil Block	95	48.0	Oregon	Ribbon Ridge	Pinot Noir	Patricia Green Cellars

In [ ]:

```
print(data_variety.loc[(data['variety']=='Cabernet Sauvignon')].shape)
print(data_variety.loc[(data['variety']=='Pinot Noir')].shape)
print(data_variety.loc[(data['variety']=='Barbera')].shape)
print(data_variety.loc[(data['variety']=='Merlot')].shape)
```

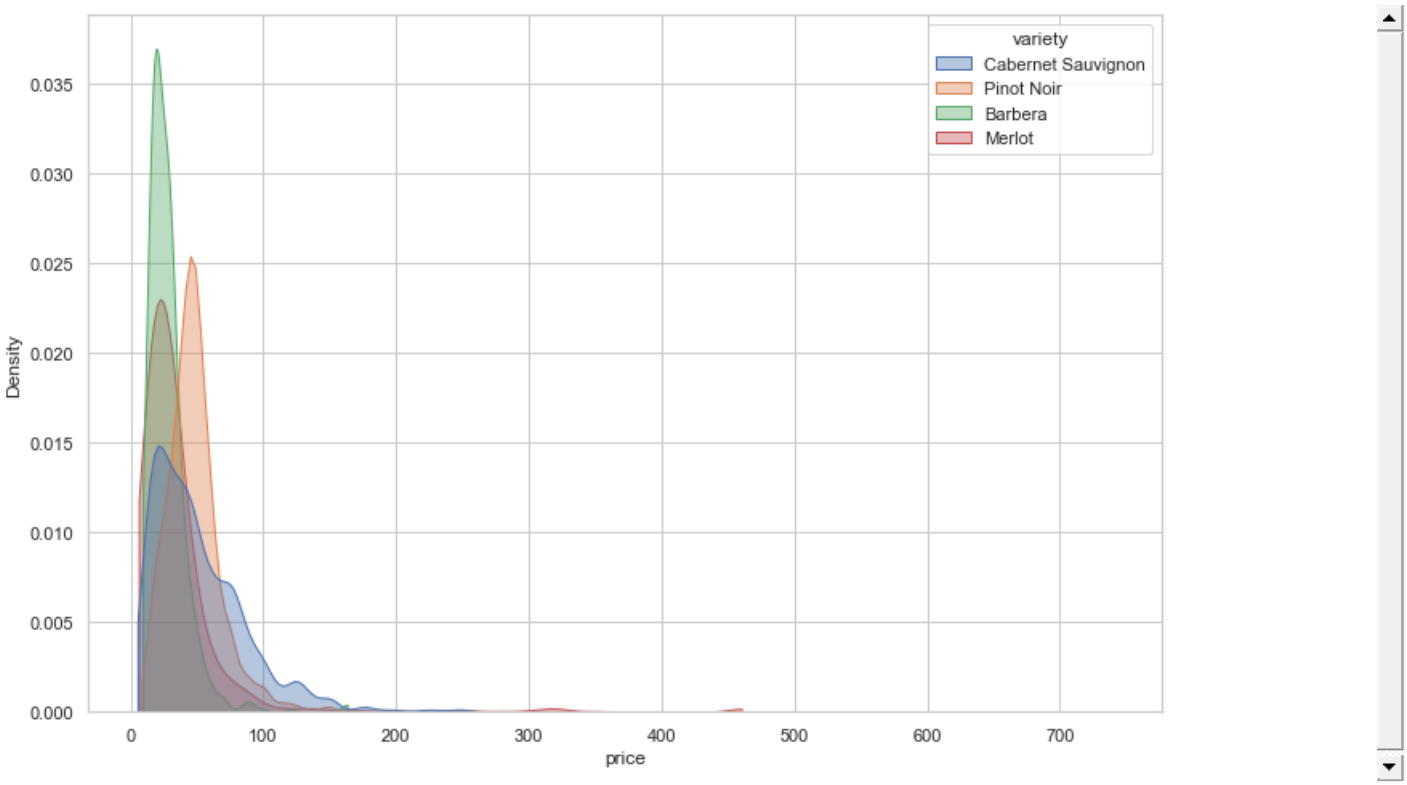
```
(6181, 10)
(8813, 10)
(675, 10)
(1962, 10)
```

In [ ]:

```
# Set figure size for the notebook
plt.rcParams["figure.figsize"]=12,8

# set seaborn whitegrid theme
sns.set(style="whitegrid")

# Without transparency
sns.kdeplot(data=data_variety, x="price", hue="variety", cut=0, fill=True, common_norm=False, alpha=0.4)
plt.show()
```



Пик плотности стоимостей Пино-Нуар находится дальше всего от нуля (где-то около 50$ за бутылку) и не совпадает с пиком, общим для остальных вин. Из чего можно сделать вывод, что Бургундское вино в среднем стоит дороже чем Каберне-Совиньон, Барбера и Мерло.

Посмотрим как связаны между собой страны и сорта винограда

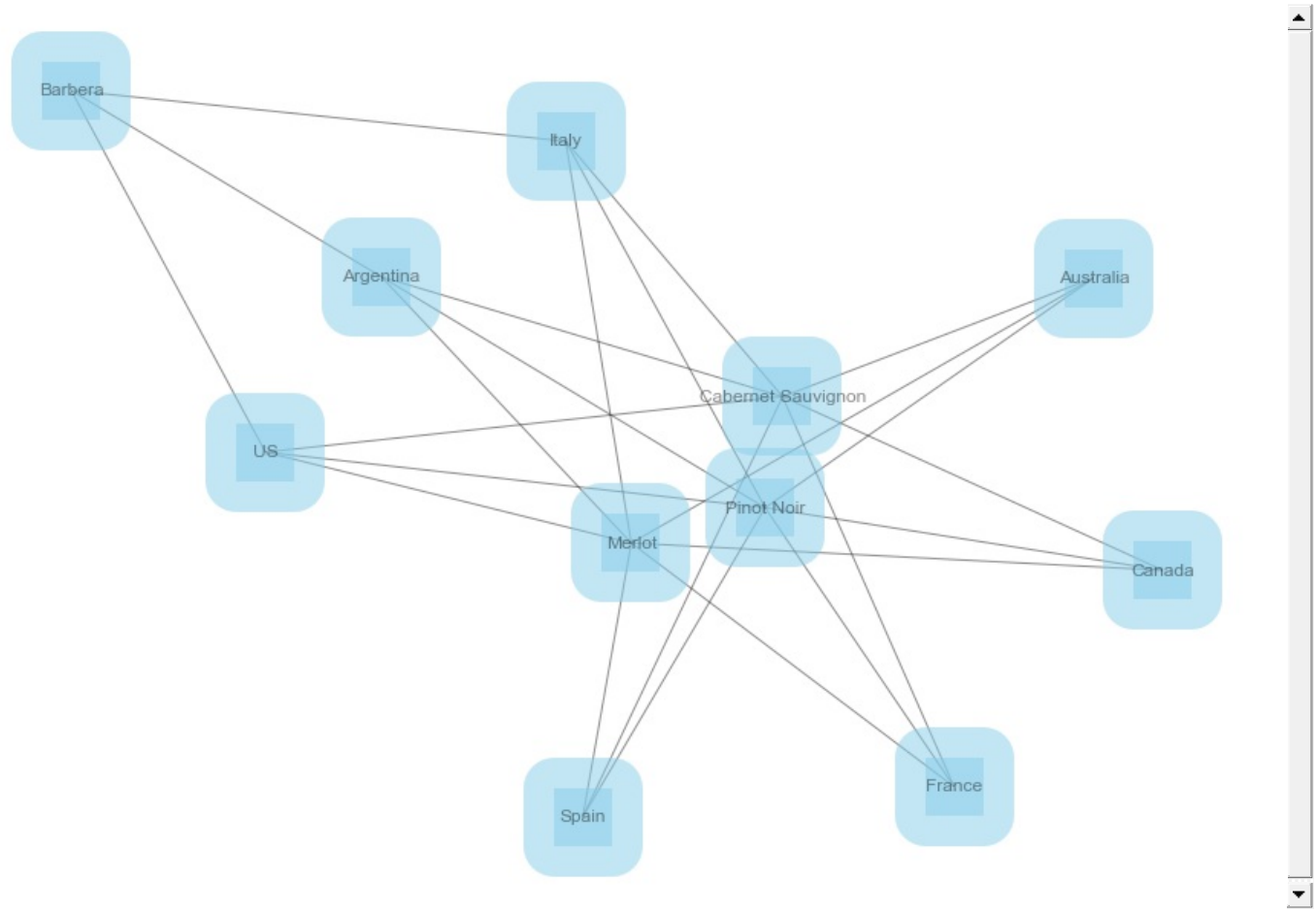
(Не могу же я обойтись без графа в конце-концов)

In [ ]:

```
import networkx as nx

G = nx.from_pandas_edgelist(data_variety, 'country', 'variety')

nx.draw(G, with_labels=True, node_size=1500, node_color="skyblue", node_shape="s", alpha=0.5, linewidths=40)
plt.show()
```



Из графа видно, что Барбера произрастает только в Италии, Аргентине и Штатах, а остальные сорта являются настолько популярными, что произрастают во всех странах, представленных в датасете.

Итог

1. Некоторые вина могут стоить весьма внушительные суммы денег.
2. Их стоимость резко поднимается после получения высокой оценки (выше 95 по Паркеру) экспертов.
3. В среднем французкие вина немного дороже остальных, что обуславливается винными традициями этой страны.
4. Знакомство с Пино-Нуар будет стоить несколько дороже, чем знакомство с Мерло, Каберне-Совиньон или Барберой.
5. Однако, сейчас виноградники распространяются почти по всем винным странам, поэтому не стоит удивляться Пино-Нуар не из Бургундии и Каберне-Совиньон не из Бордо.