ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ГОРОДА МОСКВЫ
ДОПОЛНИТЕЛЬНОГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
ЦЕНТР ПРОФЕССИОНАЛЬНЫХ КВАЛИФИКАЦИЙ И СОДЕЙСТВИЯ ТРУДОУСТРОЙСТВУ
«ПРОФЕССИОНАЛ»

ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА

на тему
«Анализ данных с использованием Python»
(на примере анализа данных исследуемого продукта)
слушателя Спиридоновой Марины Валерьевны
группы № 143
по программе профессиональной переподготовки
«Аналитик данных»

Москва, 2023

Цель исследования:

Необходимо выявить определяющие популярность марки вина закономерности и попытаться выяснить, что можно предложить покупателям вина при выборе вина. Это позволит сделать ставку на потенциально популярный продукт и спланировать например рекламную кампанию для интернет-магазинов, осуществляющих продажи вина.

Выполнение задачи предполагает:

Цель этого проекта — выявить, какие признаки больше всего влияют на рейтинг вина. Для анализа используется набор данных из Kaggle, крупнейшего в мире сообщества специалистов по данным и машинному обучению. Набор данных состоит из 13 признаков (2 числовых признака и 11 категориальных признаков).

Столбцы данных

- Страна - страна происхождения вина.
- Описание — описание вкусового профиля вина.
- Обозначение - виноградник, откуда берется виноград для вина.
- Баллы - количество баллов на которое критик журнала Wine Enthusiast оценил вино по шкале от 1 до 100.
- Цена - стоимость одной бутылки вина.
- Провинция — провинция или штат, из которого произведено вино.
- Регион 1 — зона виноделия в провинции или штате (например, долина Напа в Калифорнии).
- Регион 2 — (не обязательно) более конкретный регион в винодельческой области (например, Резерфорд в долине Напа).
- Разновидность — сорт винограда, из которого делают вино (например, Пино Нуар).
- Винодельня — винодельня, производящая вино.

Шаг 1. Открытие файла с данными и изучение общей информации

Шаг 2. Подготовка данных

- Заменить названия столбцов (привести к нижнему регистру).
- Преобразовать данные в нужные типы. Описать, в каких столбцах заменили тип данных и почему.
- Обработать пропуски при необходимости.
- Объяснить, почему заполнили пропуски определённым образом или почему не стали это делать.
- Описать причины, которые могли привести к пропускам.
- Посчитать среднии цены для каждой страны.
- Внести новый столбец "Континенты" country_to_continent = {
  'Italy':'Europe',
  'Portugal':'Europe',
  'US':'North America',
  'Spain':'Europe',
  'France':'Europe',
  'Germany':'Europe',
  'Argentina':'Latin America',
  'Chile':'Latin America',
  'Australia': 'Oceania',
  'Austria': 'Europe',
  'South Africa': 'Africa',
  'New Zealand': 'Oceania',
  'Israel': 'Asia',
  'Hungary':'Europe',
  'Greece':'Europe',
  'Romania':'Europe',
  'Mexico':'Latin America',
  'Canada':'North America',
  'Turkey': 'Asia',
  'Czech Republic': 'Europe',
  'Slovenia': 'Europe',
  'Luxembourg': 'Europe',
  'Croatia': 'Europe',
  'Georgia':'Europe',

'Uruguay': 'Latin America',
'England': 'Europe',
'Lebanon': 'Asia',
'Serbia': 'Europe',
'Brazil': 'Latin America',
'Moldova': 'Europe',
'Morocco':'Africa',
'Peru':'Latin America',
'India':'Asia',
'Bulgaria':'Europe',
'Cyprus': 'Europe',
'Armenia':'Asia',
'Switzerland':'Europe',
'Bosnia and Herzegovina':'Europe',
'Ukraine':'Europe',
'Slovakia':'Europe',
'Macedonia':'Europe',
'China':'Asia',
'Egypt':'Africa'
}

Шаг 3. Провести исследовательский анализ данных

- Определить, какие сорта лидируют по рейтингам. Найти популярные сорта по региону.
- Выбрать сорта с наибольшими ценами. Для каждого региона найдите среднюю цену вина.
- Определить, популярные сорта вина в бюджетном сегменте.
- Определить, какие сорта вина лидируют по рейтингам.
- Построить график «ящик с усами» по рейтингам в разбивке по странам, по сортам вина.
- Выявить закономерность влияния на цену цвета и рейтинга. Построить диаграмму рассеяния и посчитать корреляцию.

Шаг 4. Составить портрет потребителя каждого региона

Определить для пользователя каждого континента :

- Самые популярные сорта (топ-5).
- Влияет ли рейтинг на цены по регионам?

Шаг 5. Провести исследование статистических показателей

- Выполнить подсчитать среднего количества, дисперсии и стандартного отклонения для цен на продукт различных регионов. Построить гистограммы. Описать распределения.
- Построить линейную регрессию зависимости между ценой продукта и его рейтингом.

Шаг 6. Проверка гипотез

- H0: Средние пользовательские рейтинги красного и белого вина одинаковые.
- H1:Средние пользовательские рейтинги красного и белого вина разные.
- H0:Средние цены двух популярных сортов вина одинаковые.
- H1:Средние цены двух популярных сортов вина разные.

Задать самостоятельно пороговое значение alpha.

Вывод

1.Предобработка данных

Импортируем необходимые библиотеки

In [132]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.cm as cm
%matplotlib inline
import scipy.stats as st
# импорт библиотеки warnings
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Загрузка данных

In [133]:

```
df = pd.read_csv('wine_reviews.csv')
df.head()
```

Out[133]:

	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	US	With a delicate, silky mouthfeel and bright ac...	NaN	86	23.0	California	Central Coast	Central Coast	Pinot Noir	MacMurray Ranch
1	Italy	D'Alceo is a drop dead gorgeous wine that ooze...	D'Alceo	96	275.0	Tuscany	Toscana	NaN	Red Blend	Castello dei Rampolla
2	France	The great dominance of Cabernet Sauvignon in t...	NaN	91	40.0	Bordeaux	Haut-Médoc	NaN	Bordeaux-style Red Blend	Château Bernadotte
3	Italy	The modest cherry, dark berry and black tea no...	NaN	81	15.0	Tuscany	Chianti Classico	NaN	Sangiovese	Valiano
4	US	Exceedingly light in color, scent and flavor, ...	NaN	83	25.0	Oregon	Rogue Valley	Southern Oregon	Pinot Noir	Deer Creek

Заменять названия столбцов (привести к нижнему регистру) нет необходимости, поскольку названия столбцов в нижнем регистре.

In [134]:

```
df.columns
```

Out[134]:

```
Index(['country', 'description', 'designation', 'points', 'price', 'province',
       'region_1', 'region_2', 'variety', 'winery'],
      dtype='object')
```

Анализируем размерность, ищем пропуски

In [135]:

```
df.shape
```

Out[135]:

```
(20000, 10)
```

In [136]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20000 entries, 0 to 19999
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   country      20000 non-null  object 
 1   description  20000 non-null  object 
 2   designation  13999 non-null  object 
 3   points       20000 non-null  int64  
 4   price        18198 non-null  float64
 5   province     20000 non-null  object 
 6   region_1     16543 non-null  object 
 7   region_2     8058 non-null   object 
 8   variety      20000 non-null  object 
 9   winery       20000 non-null  object 
dtypes: float64(1), int64(1), object(8)
memory usage: 1.5+ MB
```

Количество значений в столбцах различается. Это говорит о том, что в данных есть пустые значения. Признак points и price числовые. С помощью библиотеки seaborn построим тепловую карту для визуализации данных.

In [137]:

```
colours = ['#993366', '#FFFF00']
sns.heatmap(df.isnull(), cmap=sns.color_palette(colours))
# Decorations
plt.title('Матрица пропущенных значений набора данных', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.figtext(0.1, -0.2, "Рисунок 1. - Матрица пропущенных значений набора данных")
plt.show()
```

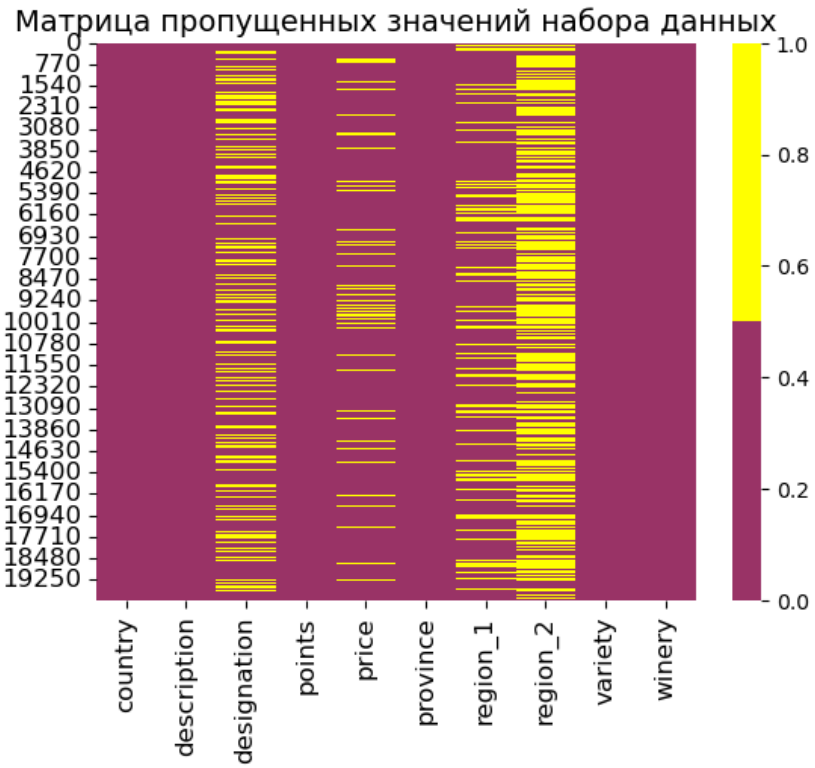


Рисунок 1. - Матрица пропущенных значений набора данных

Обрабатываем пропуски

In [138]:

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

Out[138]:

```
country            0
description        0
designation     6001
points             0
price           1802
province           0
region_1        3457
region_2       11942
variety            0
winery             0
dtype: int64
```

Посмотрим поближе на пропуски в цене.

In [139]:

```
df.query('price.isnull() == True')
```

Out[139]:

	country	description	designation	points	price	province	region_1	region_2	variety	winery
55	Italy	This dark Recioto dessert wine opens with an i...	Acinatico 500ml	88	NaN	Veneto	Recioto della Valpolicella Classico	NaN	Corvina, Rondinella, Molinara	Stefano Accordini
63	Italy	Black fruit, ash, cola, cherry, mesquite and s...	Le Sassine	91	NaN	Veneto	Valpolicella Classico Superiore Ripasso	NaN	Corvina, Rondinella, Molinara	Le Ragose
74	France	This is an earthy, traditional approach to Cro...	Laurus	86	NaN	Rhône Valley	Crozes-Hermitage	NaN	Syrah	Gabriel Meffre
92	Portugal	Spicy and firm in character, although the colo...	Late Bottled Vintage	86	NaN	Port	NaN	NaN	Port	Quinta Santa Eufemia
101	France	90-92 Barrel sample. Dense, smoky tannins over...	Barrel sample	91	NaN	Bordeaux	Margaux	NaN	Bordeaux-style Red Blend	Clos Magdelaine
...	...	...	...	...	...	...	...	...	...	...
19966	Italy	The beautiful Brunate cru is known for heavier...	Brunate	93	NaN	Piedmont	Barolo	NaN	Nebbiolo	Damilano
19970	France	92–94 A delicious, sweet wine, which has great...	Barrel Sample	93	NaN	Bordeaux	Sauternes	NaN	Bordeaux-style White Blend	Château Doisy-Daëne
19973	Chile	Creamy and herbal on the nose, with mild veget...	Ecos de Rulo	86	NaN	Colchagua Valley	NaN	NaN	Cabernet Sauvignon	Viña Bisquertt
19992	France	This wine is driven by its ripe fruit, black c...	Tradition	88	NaN	Southwest France	Cahors	NaN	Malbec-Merlot	Château la Caminade
19998	Italy	Organically farmed Cannonau grapes deliver sma...	Le Sabbie	87	NaN	Sicily & Sardinia	Cannonau di Sardegna	NaN	Cannonau	Meloni

1802 rows × 10 columns

In [140]:

```
df.isnull().sum().price / 20000
```

Out[140]:

```
0.0901
```

Видим, что распределение пропусков не зависит от страны, региона, сорта винограда, оценки. Возможно отсутствие цены связано с неправильностью заполнения данных, или данные могли быть утеряны.

Процент пустых значений цены составляет 9%. Есть вариант удаления пропусков, однако в связи с тем, что процент пропусков достаточно велик предлагаем сохранить соответствующие строки.

В этом исследовании предлагаем заполнить пропуски медианными значениями по каждой провинции.

In [141]:

```
df['price'] = df.groupby('province')['price'].apply(lambda x: x.fillna(x.median()))
```

In [142]:

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

Out[142]:

```
8
```

Незаполненных значений в столбце price осталось 8, удалим их.

In [143]:

```
df = df.dropna(subset = 'price')
df.isnull().sum()
```

Out[143]:

```
country            0
description        0
designation     6000
points             0
price              0
province           0
region_1        3449
region_2       11934
variety            0
winery             0
dtype: int64
```

Мы обработали пустые значения столбца price. Пропуски в остальных признаках не так принципиальны, при наличии провинции, регион 1 и регион 2 могли быть не заполнены или могут не иметь конкретизации. Эти признаки не относятся к исследуемым, поэтому предлагаем их сохранить в базе данных.

Преобразуем столбец price в int, поскольку все значения цены в базе данных целые

In [144]:

```
print(df.price.to_list())
```

```
7.0, 12.0, 26.0, 28.0, 80.0, 25.0, 89.0, 13.0, 12.0, 53.0, 60.0, 18.0, 18.0, 40.0, 22.0, 18.0, 13.0, 60.
0, 10.0, 12.0, 75.0, 12.0, 19.0, 16.0, 165.0, 60.0, 16.0, 10.0, 20.0, 14.0, 23.0, 13.0, 19.0, 27.0, 35.
0, 50.0, 38.0, 36.0, 21.0, 42.0, 15.0, 23.0, 25.0, 13.0, 50.0, 45.0, 23.0, 72.0, 21.0, 45.0, 48.0, 42.0,
20.0, 20.0, 25.0, 30.0, 65.0, 12.0, 25.0, 12.0, 20.0, 13.0, 9.0, 40.0, 75.0, 11.5, 10.0, 21.0, 10.0, 9.
0, 22.0, 65.0, 50.0, 30.0, 22.0, 25.0, 16.0, 30.0, 8.0, 19.0, 30.0, 40.0, 26.5, 25.0, 14.0, 17.0, 120.0,
36.0, 50.0, 47.0, 48.0, 35.0, 40.0, 13.0, 20.0, 29.0, 12.0, 14.0, 24.0, 15.0, 30.0, 13.0, 20.0, 28.0, 2
7.0, 44.0, 20.0, 39.0, 15.0, 35.0, 20.0, 45.0, 70.0, 15.0, 13.0, 19.0, 40.0, 25.0, 35.0, 20.0, 29.0, 11.
0, 85.0, 26.0, 11.0, 60.0, 12.0, 28.0, 44.0, 202.0, 10.0, 35.0, 16.0, 40.0, 14.0, 40.0, 29.0, 42.0, 17.
0, 10.0, 70.0, 38.0, 85.0, 38.0, 20.0, 18.0, 60.0, 32.0, 150.0, 11.0, 90.0, 42.0, 255.0, 25.0, 49.0, 16.
0, 9.0, 19.0, 23.0, 10.0, 20.0, 36.0, 38.0, 15.0, 15.5, 20.0, 48.0, 8.0, 18.0, 6.0, 17.0, 60.0, 11.0, 2
0.0, 30.0, 20.0, 32.0, 35.0, 48.0, 18.0, 15.0, 48.0, 16.0, 30.0, 35.0, 125.0, 30.0, 70.0, 12.0, 20.0, 12.
0, 10.0, 45.0, 18.0, 25.0, 17.0, 30.0, 19.0, 200.0, 20.0, 44.0, 8.0, 28.0, 11.0, 14.0, 36.0, 9.0, 49.0, 1
6.0, 8.0, 20.0, 13.0, 40.0, 25.0, 46.0, 35.0, 12.0, 50.0, 23.0, 50.0, 35.0, 32.0, 25.0, 15.0, 38.0, 22.0,
40.0, 15.0, 22.0, 16.0, 16.0, 14.0, 13.0, 50.0, 31.0, 26.0, 28.0, 24.0, 40.0, 19.0, 18.0, 20.0, 24.0, 80.
0, 27.0, 45.0, 17.0, 8.0, 20.0, 20.0, 30.0, 22.0, 29.0, 30.0, 8.0, 28.0, 50.0, 22.0, 11.0, 20.0, 45.0, 3
4.0, 27.0, 19.0, 28.0, 20.0, 42.0, 20.0, 55.0, 75.0, 10.0, 42.0, 65.0, 8.0, 50.0, 30.0, 10.0, 35.0, 16.0,
10.0, 75.0, 10.0, 45.0, 20.0, 48.0, 18.0, 12.0, 35.0, 23.0, 30.0, 30.0, 65.0, 22.0, 15.0, 18.0, 18.0, 32.
0, 34.0, 17.0, 20.0, 38.0, 60.0, 11.0, 14.0, 15.0, 33.0, 38.0, 115.0, 90.0, 14.0, 20.0, 50.0, 38.0, 45.0,
16.0, 26.5, 10.0, 18.0, 32.0, 39.0, 9.0, 90.0, 23.0, 20.0, 19.0, 12.0, 16.0, 36.0, 30.0, 22.0, 19.0, 32.
0, 23.0, 16.0, 25.0, 13.0, 9.0, 46.0, 15.0, 55.0, 24.0, 38.0, 20.0, 12.0, 35.0, 23.0, 20.0, 42.0, 22.0, 1
```

In [145]:

```
df.price = df.price.astype('int32')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19992 entries, 0 to 19999
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   country      19992 non-null  object
 1   description  19992 non-null  object
 2   designation  13992 non-null  object
 3   points       19992 non-null  int64
 4   price        19992 non-null  int32
 5   province     19992 non-null  object
 6   region_1     16543 non-null  object
 7   region_2     8058 non-null   object
 8   variety      19992 non-null  object
 9   winery       19992 non-null  object
dtypes: int32(1), int64(1), object(8)
memory usage: 1.6+ MB
```

Посчитаем среднюю цену для каждой страны

In [146]:

```
df.groupby('country')['price'].mean()
```

Out[146]:

```
country
Argentina                 22.441417
Australia                 30.557878
Austria                   30.365155
Bosnia and Herzegovina    12.000000
Brazil                    24.666667
Bulgaria                  10.600000
Canada                    44.538462
Chile                     19.597436
China                     27.000000
Croatia                   21.238095
Cyprus                    16.666667
France                    40.993820
Georgia                   16.000000
Germany                   35.423398
Greece                    20.812500
Hungary                   60.911765
Israel                    32.160920
Italy                     35.375485
Lebanon                   32.500000
Luxembourg                36.000000
Mexico                    25.571429
Moldova                   15.062500
Montenegro                10.000000
New Zealand               24.510112
Portugal                  27.141711
Romania                   12.500000
Serbia                    16.500000
Slovenia                  25.928571
South Africa              21.289474
South Korea               11.000000
Spain                     27.788073
Switzerland               19.000000
Turkey                    28.166667
US                        33.514126
US-France                 50.000000
Ukraine                   13.000000
Uruguay                   18.428571
Name: price, dtype: float64
```

Введем новый столбец "Континенты" country_to_continent

In [147]:

```
continent = {
'Italy':'Europe',
'Portugal':'Europe',
'US':'North America',
'Spain':'Europe',
'France':'Europe',
'Germany':'Europe',
'Argentina':'Latin America',
'Chile':'Latin America',
'Australia': 'Oceania',
'Austria': 'Europe',
'South Africa': 'Africa',
'New Zealand': 'Oceania',
'Israel': 'Asia',
'Hungary':'Europe',
'Greece':'Europe',
'Romania':'Europe',
'Mexico':'Latin America',
'Canada':'North America',
'Turkey': 'Asia',
'Czech Republic': 'Europe',
'Slovenia': 'Europe',
'Luxembourg': 'Europe',
'Croatia': 'Europe',
'Georgia':'Europe',
'Uruguay': 'Latin America',
'England': 'Europe',
'Lebanon': 'Asia',
'Serbia': 'Europe',
'Brazil': 'Latin America',
'Moldova': 'Europe',
'Morocco':'Africa',
'Peru':'Latin America',
'India':'Asia',
'Bulgaria':'Europe',
'Cyprus': 'Europe',
'Armenia':'Asia',
'Switzerland':'Europe',
'Bosnia and Herzegovina':'Europe',
'Ukraine':'Europe',
'Slovakia':'Europe',
'Macedonia':'Europe',
'China':'Asia',
'Egypt':'Africa'
}
df['country_to_continent'] = df['country'].map(continent)
df.head()
```

Out[147]:

	country	description	designation	points	price	province	region_1	region_2	variety	winery	country_to_continent
0	US	With a delicate, silky mouthfeel and bright ac...	NaN	86	23	California	Central Coast	Central Coast	Pinot Noir	MacMurray Ranch	North America
1	Italy	D'Alceo is a drop dead gorgeous wine that ooze...	D'Alceo	96	275	Tuscany	Toscana	NaN	Red Blend	Castello dei Rampolla	Europe
2	France	The great dominance of Cabernet Sauvignon in t...	NaN	91	40	Bordeaux	Haut-Médoc	NaN	Bordeaux-style Red Blend	Château Bernadotte	Europe
3	Italy	The modest cherry, dark berry and black tea no...	NaN	81	15	Tuscany	Chianti Classico	NaN	Sangiovese	Valiano	Europe
4	US	Exceedingly light in color, scent and flavor, ...	NaN	83	25	Oregon	Rogue Valley	Southern Oregon	Pinot Noir	Deer Creek	North America

Выводы по итогам раздела 1:

- Мы подгрузили данные, оценили пропуски. Заметили, что распределение пропусков цены не зависит от страны, региона, сорта винограда, рейтинга. Возможно отсутствие цены связано с неправильностью заполнения данных, или данные могли быть утеряны.
- Далее, обработали пропуски столбца price - заменили на медианные значения по каждой провинции.

- Выявили также пропуски в designation, region_1, region_2. При указании провинции, регион 1 и регион 2 могли быть не заполнены или могут не иметь конкретизации (нет соответствующего административно-территориального деления). Мы их не стали обрабатывать, поскольку эти признаки не относятся к исследуемым в рассматриваемой работе.
- Преобразовали столбец price в int, поскольку все значения цены в базе данных целые. Это позволяет разумнее использовать память.
- Посчитали среднюю цену бутылки вина для каждой страны.
- Ввели новый столбец "Континенты" country_to_continent

2. Исследовательский анализ данных

Изучим распределение значений рейтинга

In [148]:

```
plt.figure(figsize = (5,4))
df.points.hist(color = 'purple')
plt.xlabel('Оценка / рейтинг')
plt.ylabel('Количество')
plt.title('Распределение уровней рейтинга')
plt.figtext(0.1, -0.1, "Рисунок 2. - Распределение уровней рейтинга")
```

Out[148]:

```
Text(0.1, -0.1, 'Рисунок 2. - Распределение уровней рейтинга')
```

Рисунок 2. - Распределение уровней рейтинга

In [149]:

```
df.points.describe()
```

Out[149]:

```
count    19992.000000
mean        87.899960
std          3.242735
min         80.000000
25%         86.000000
50%         88.000000
75%         90.000000
max        100.000000
Name: points, dtype: float64
```

Видим, что максимальное значение рейтинга - 100, минимальное 80. Медианное значение близко к среднему, что также видно на графике. С высокой вероятностью, распределение нормальное.

Определим, какие сорта лидируют по рейтингам.

In [150]:

```
df.sort_values(by = 'points', ascending = False)[['variety', 'points']]
```

Out[150]:

	variety	points
323	Chardonnay	100
17967	Cabernet Blend	100
5955	Merlot	100
7306	Syrah	99
13188	Bordeaux-style Red Blend	99
...	...	...
14068	Lagrein	80
6888	Red Blend	80
12274	Tempranillo Blend	80
18391	Sauvignon Blanc	80
10320	Merlot	80

19992 rows × 2 columns

Сорта Шардоне (Chardonnay), Каберне бленд (Cabernet Blend), Мерло (Merlot), а также Шираз (Syrah), Красный бленд Бордо (Bordeaux-style Red Blend) получают самые высокие оценки экспертов.

Найдем популярные сорта по региону.

In [151]:

```
new = df.groupby(['province'])[['variety', 'points']].max().reset_index()
new[new.province.isin(['Tuscany', 'California','Bordeaux'])]
```

Out[151]:

	province	variety	points
27	Bordeaux	Sémillon	99
38	California	Zinfandel	100
270	Tuscany	White Blend	100

Нашли сорт винограда с самым высоким рейтином по каждой провинции. Например, в Bordeaux - это Sémillon, в Калифорнии - это Zinfandel, а в Тоскане - это White Blend.

Изучим распределение значений цены

In [152]:

```
plt.figure(figsize = (10,4))
df.price.hist(color = 'purple', bins = 500)
plt.xlabel('Цена за бутылку')
plt.xlim(0, 200)
plt.ylabel('Количество проданных')
plt.title('Распределение уровня цен')
plt.figtext(0.1, -0.1, "Рисунок 3. - Распределение уровня цен")
```

Out[152]:

```
Text(0.1, -0.1, 'Рисунок 3. - Распределение уровня цен')
```

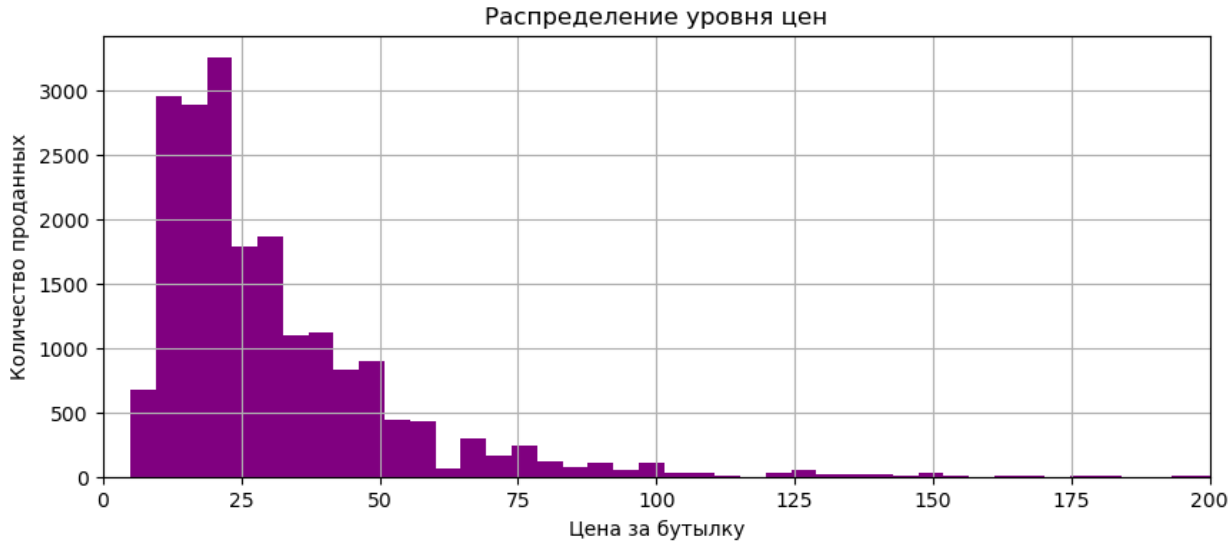


Рисунок 3. - Распределение уровня цен

In [153]:

```
df.price.describe()
```

Out[153]:

```
count    19992.000000
mean        32.709234
std         38.072288
min          5.000000
25%         16.000000
50%         24.000000
75%         39.000000
max       2300.000000
Name: price, dtype: float64
```

Видим, что максимальное значение цены - 2300 у.е., минимальное 5 у.е.. Медианное значение - 24, среднее - 32.7. Распределение имеет хвостик справа, что и отображается на графике. Действительно, имеется группа вин с очень высокими ценами. Это могут быть и выбросы, но, вероятнее всего, это коллекционные вина, имеющие длительный срок выдержки. На данном этапе, мы не будем убирать этот хвостик, однако, будем держать в голове при последующем статистическом анализе.

Выберем сорта с наибольшими ценами.

In [154]:

```
df.sort_values(by = 'price', ascending = False)[['variety', 'price']]
```

Out[154]:

	variety	price
13188	Bordeaux-style Red Blend	2300
323	Chardonnay	1400
4324	Grüner Veltliner	1100
19501	Bordeaux-style Red Blend	850
8493	Furmint	764
...	...	...
27	Cabernet Sauvignon	5
6381	Rosé	5
4321	Chardonnay	5
14547	Tempranillo	5
3583	Syrah	5

19992 rows × 2 columns

Сорта с наибольшими ценами - Bordeaux-style Red Blend, Chardonnay, Grüner Veltliner, Furmint. Заметим, что топ сортов по ценам лишь частично пересекается с топ сортов по рейтингам. Совпадение в сортах - Красный бленд Бордо (Bordeaux-style Red Blend), Шардоне (Chardonnay).

Для каждого региона найдем среднюю цену вина.

In [155]:

```
df.groupby('province')['price'].mean().sort_values(ascending = False)
```

Out[155]:

```
province
Tokaji                133.100000
Santa Cruz             95.000000
Champagne              92.701657
Israel                 70.000000
Burgundy               65.239362
                          ...
Table wine              8.000000
Requinoa                8.000000
Felso-Magyarország      7.000000
Recas                   7.000000
Primorska               7.000000
Name: price, Length: 307, dtype: float64
```

Интересно, что самая высокая средняя цена бутылки вина - это цена токайского вина (Венгрия).

Определим, популярные сорта вина в бюджетном сегменте. Возьмем цену за бутылку до 10 евро.

In [156]:

```
df.query('price <= 10').groupby('variety')['price'].count().sort_values(ascending = False).head(15)
```

Out[156]:

```
variety
Chardonnay            148
Cabernet Sauvignon    116
Sauvignon Blanc       108
Merlot                 89
Rosé                   86
Red Blend              61
White Blend            59
Malbec                 57
Portuguese Red         56
Pinot Grigio           50
Portuguese White       47
Tempranillo            44
Shiraz                 36
Riesling               34
Garnacha               32
Name: price, dtype: int64
```

Больше всего вин в сегменте до 10 евро продается из следующих сортов: Шардоне, Каберне Совиньон, Совиньон блан, Мерло и Розе. С результатом согласна. Указанные сорта являются достаточно неприхотливыми и получили распространение и в более суровых погодных условиях, соответствующие виноградники есть также на территории РФ (Крым, Таманский полуостров, Абрау-Дюрсо)

Построить график «ящик с усами» по рейтингам в разбивке по странам, по сортам вина.

In [157]:

```
df.sample(5)
```

Out[157]:

	country	description	designation	points	price	province	region_1	region_2	variety	winery	country_to_contine
4850	US	Tasted just at its official release, this very...	NaN	85	19	Washington	Columbia Valley (WA)	Columbia Valley	Syrah	DaMa	North Amer
19906	France	While demi-sec is the description, this almost...	Demi-Sec	89	46	Champagne	Champagne	NaN	Champagne Blend	Mailly Grand Cru	Euro
17776	Italy	Subdued aromas include white spring flowers, p...	NaN	86	15	Northeastern Italy	Alto Adige	NaN	Pinot Grigio	Kellerei Kaltern Caldaro	Euro
8136	US	Jammy and fat. All primary fruit (cherries) an...	Sanford & Benedict Vineyard	92	42	California	Sta. Rita Hills	Central Coast	Pinot Noir	Bonaccorsi	North Amer
16144	US	Malbec is relatively new and rare in Washingto...	NaN	86	38	Washington	Columbia Valley (WA)	Columbia Valley	Malbec	Michael Florentino Cellars	North Amer

In [158]:

```
df.boxplot(column='points', by='country', figsize = (30, 8))
plt.figtext(0.1, -0.2, "Рисунок 4. - График «ящик с усами» по рейтингам в разбивке по странам", fontsize = 30)
plt.xticks(rotation = 45, fontsize = 15) # повернуть и отформатировать надписи по оси x
```

Out[158]:

```
(array([ 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, 30, 31, 32, 33, 34,
        35, 36, 37]),
 [Text(1, 0, 'Argentina'),
  Text(2, 0, 'Australia'),
  Text(3, 0, 'Austria'),
  Text(4, 0, 'Bosnia and Herzegovina'),
  Text(5, 0, 'Brazil'),
  Text(6, 0, 'Bulgaria'),
  Text(7, 0, 'Canada'),
  Text(8, 0, 'Chile'),
  Text(9, 0, 'China'),
  Text(10, 0, 'Croatia'),
  Text(11, 0, 'Cyprus'),
  Text(12, 0, 'France'),
  Text(13, 0, 'Georgia'),
  Text(14, 0, 'Germany'),
  Text(15, 0, 'Greece'),
  Text(16, 0, 'Hungary'),
  Text(17, 0, 'Israel'),
  Text(18, 0, 'Italy'),
  Text(19, 0, 'Lebanon'),
  Text(20, 0, 'Luxembourg'),
  Text(21, 0, 'Mexico'),
  Text(22, 0, 'Moldova'),
  Text(23, 0, 'Montenegro'),
  Text(24, 0, 'New Zealand'),
  Text(25, 0, 'Portugal'),
  Text(26, 0, 'Romania'),
  Text(27, 0, 'Serbia'),
  Text(28, 0, 'Slovenia'),
  Text(29, 0, 'South Africa'),
  Text(30, 0, 'South Korea'),
  Text(31, 0, 'Spain'),
  Text(32, 0, 'Switzerland'),
  Text(33, 0, 'Turkey'),
  Text(34, 0, 'US'),
  Text(35, 0, 'US-France'),
  Text(36, 0, 'Ukraine'),
  Text(37, 0, 'Uruguay')])
```

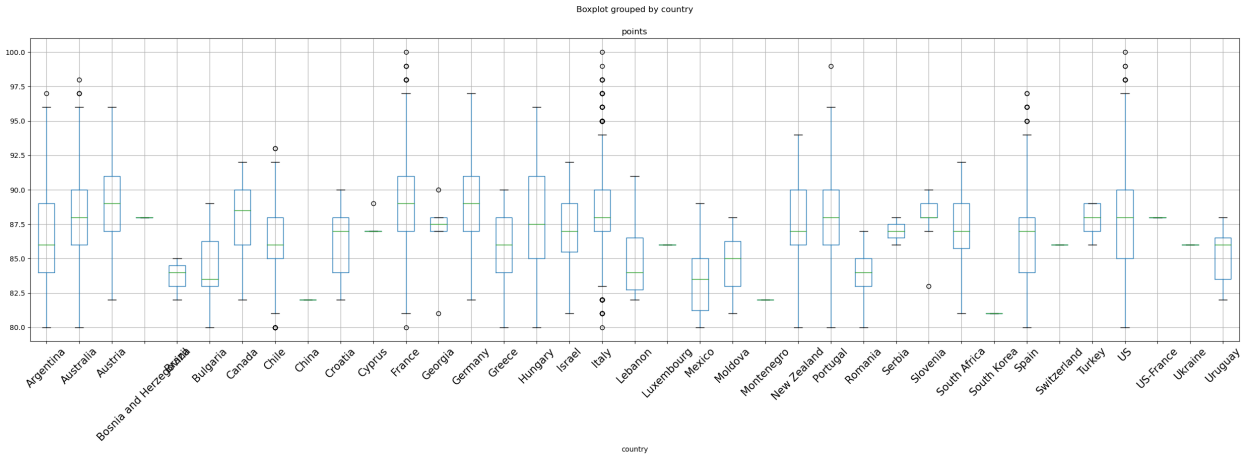


Рисунок 4. - График «ящик с усами» по рейтингам в разбивке по странам

In [159]:

```
#нашли топ 5 популярных сортов
df_variety_5top_list = df.variety.value_counts().nlargest(5).index
df_variety_5top = df[df.variety.isin(df_variety_5top_list)]
```

In [160]:

```
df_variety_5top.boxplot(column='points', by='variety', figsize = (11, 6))
plt.figtext(0.1, -0.2, "Рисунок 5. - График «ящик с усами» по рейтингам в разбивке по сорту вина на примере 5ти популярных
plt.xticks(rotation = 30, fontsize = 12) # повернуть и отформатировать надписи по оси x
```

Out[160]:

```
(array([1, 2, 3, 4, 5]),
 [Text(1, 0, 'Bordeaux-style Red Blend'),
  Text(2, 0, 'Cabernet Sauvignon'),
  Text(3, 0, 'Chardonnay'),
  Text(4, 0, 'Pinot Noir'),
  Text(5, 0, 'Red Blend')])
```

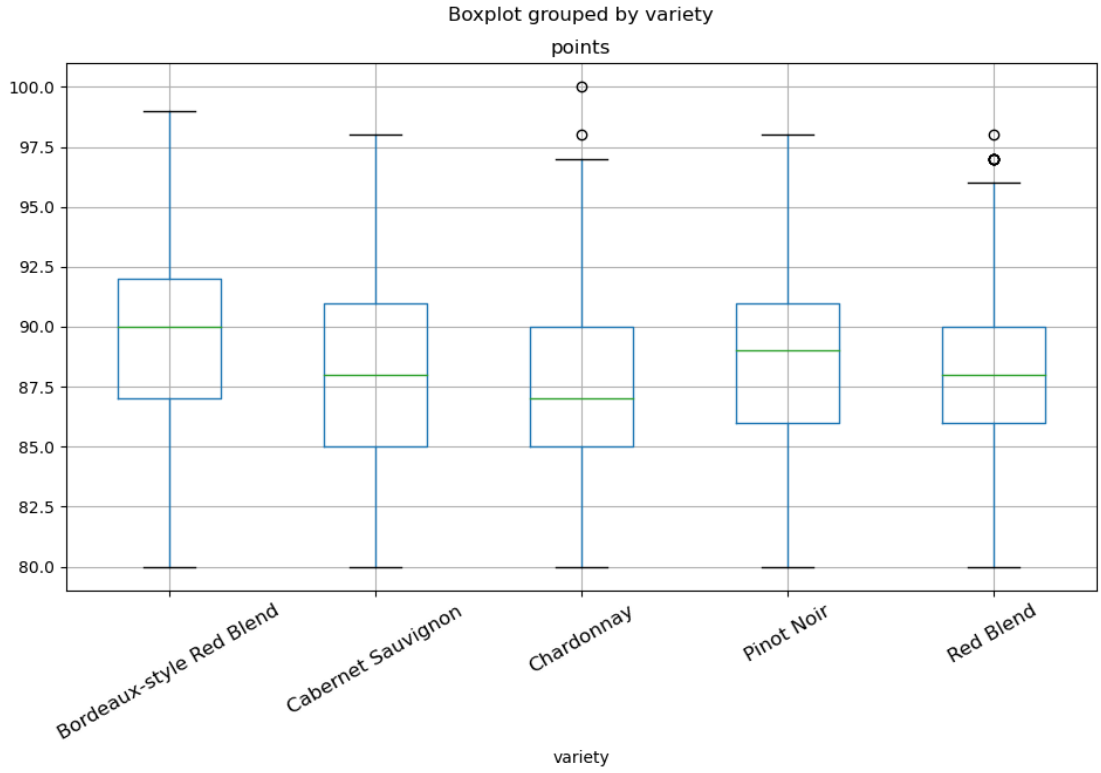


Рисунок 5. - График «ящик с усами» по рейтингам в разбивке по сорту вина на примере 5ти популярных сортов

Выявим закономерность влияния на цену цвета и рейтинга. Построим диаграмму рассеивания и посчитать корреляцию.

Создадим столбец с цветом вина

In [161]:

```
color = {
    "Chardonnay": "white",
    "Pinot Noir": "red",
    "Cabernet Sauvignon": "red",
    "Red Blend": "red",
    "Bordeaux-style Red Blend": "red",
    "Sauvignon Blanc": "white",
    "Syrah": "red",
    "Riesling": "red",
    "Merlot": "red",
    "Zinfandel": "red",
    "Sangiovese": "red",
    "Malbec": "red",
    "White Blend": "white",
    "Rosé": "other",
    "Tempranillo": "red",
    "Nebbiolo": "red",
    "Portuguese Red": "red",
    "Sparkling Blend": "other",
    "Shiraz": "red",
    "Corvina, Rondinella, Molinara": "red",
    "Rhône-style Red Blend": "red",
    "Barbera": "red",
    "Pinot Gris": "white",
    "Viognier": "white",
    "Bordeaux-style White Blend": "white",
    "Champagne Blend": "other",
    "Port": "red",
    "Grüner Veltliner": "white",
    "Gewürztraminer": "white",
    "Portuguese White": "white",
    "Petite Sirah": "red",
    "Carmenère": "red"
}
df['wine_color'] = df['variety'].map(color)
df.head()
```

Out[161]:

	country	description	designation	points	price	province	region_1	region_2	variety	winery	country_to_continent	wine_c
0	US	With a delicate, silky mouthfeel and bright ac...	NaN	86	23	California	Central Coast	Central Coast	Pinot Noir	MacMurray Ranch	North America	
1	Italy	D'Alceo is a drop dead gorgeous wine that ooze...	D'Alceo	96	275	Tuscany	Toscana	NaN	Red Blend	Castello dei Rampolla	Europe	
2	France	The great dominance of Cabernet Sauvignon in t...	NaN	91	40	Bordeaux	Haut-Médoc	NaN	Bordeaux-style Red Blend	Château Bernadotte	Europe	
3	Italy	The modest cherry, dark berry and black tea no...	NaN	81	15	Tuscany	Chianti Classico	NaN	Sangiovese	Valiano	Europe	
4	US	Exceedingly light in color, scent and flavor, ...	NaN	83	25	Oregon	Rogue Valley	Southern Oregon	Pinot Noir	Deer Creek	North America	

Проверяем количество пропущенных значений в столбце color. Это сорта вина, которых нет в книжке.

In [162]:

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

Out[162]:

```
3497
```

Для того чтобы не терять важные данные, заменим пустые значения на 'tbd' (to be defined - подлежит определению). Вероятно, мы сможем обсудить с экспертами и присвоить необходимый цвет позднее.

In [163]:

```
df.wine_color = df.wine_color.fillna('tbd')
df.wine_color.isnull().sum()
```

Out[163]:

```
0
```

Проверили - пустых значений в столбце 'wine_color' больше нет.

Посмотрим, какие сорта винограда попадают в 'other'. Это, прежде всего,

In [164]:

```
df[df.wine_color == 'other'].variety.unique()
```

Out[164]:

```
array(['Rosé', 'Champagne Blend', 'Sparkling Blend'], dtype=object)
```

Посчитаем корреляцию цены и рейтинга

In [165]:

```
df[['price', 'points']].corr()
```

Out[165]:

	price	points
price	1.000000	0.404637
points	0.404637	1.000000

Построим диаграмму рассеивания

In [166]:

```
sns.lmplot(x = 'points',
           y = 'price', data = df,
           height = 5, # height and width of the plot
           scatter_kws = {'color': 'purple'}, # color for the points
           line_kws = {'color': 'midnightblue'}) # color for the regression line
plt.xlim([79, 101])
plt.ylim([0, 1000])
plt.xlabel("Оценка / рейтинг вина")
plt.ylabel("Цена за бутылку")
plt.title("Зависимость цены от рейтинга")
plt.figtext(0.1, -0.2, "Рисунок 6. - Зависимость цены от рейтинга")
```

Out[166]:

```
Text(0.1, -0.2, 'Рисунок 6. - Зависимость цены от рейтинга')
```

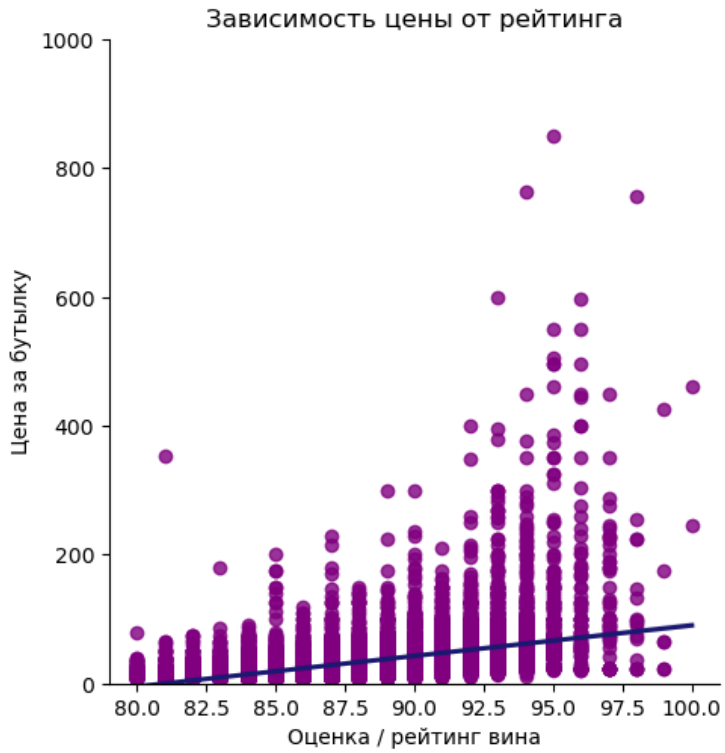


Рисунок 6. - Зависимость цены от рейтинга

Видим, что есть некоторая зависимость между ценой и рейтингом.

Выявим закономерность влияния на цену цвета

In [167]:

```
df.groupby('wine_color')['price'].agg(('mean','median'))
```

Out[167]:

wine_color	mean	median
other	31.202517	20.0
red	36.647191	28.0
tbd	26.810695	20.0
white	26.837878	20.0

Медианное и среднее значение цены красного вина выше. При этом медианное значение категории "другое" равно категории белого, а среднее выше среднего белого.

In [168]:

```
df['wine_color'].value_counts(normalize = True)
```

Out[168]:

```
red      0.579582
white    0.201781
tbd      0.174920
other    0.043717
Name: wine_color, dtype: float64
```

Видим, что красного вина в выборке представлено больше всего (58%), белого тоже немало (20%). Вместе с тем, вина с пометкой 'other'(напомним, что это шампанское, розе и игристое) в выборке немного, и, соответственно, делать какие-либо выводы по указанной категории мы можем лишь с осторожностью. Отметим, что категория 'tbd' занимает достаточно большой процент выборки (17,5%).

In [169]:

```
plt.scatter(df.wine_color, df.price, lw=0, alpha=.08, color='purple')
plt.xlabel("Вино")
plt.ylabel("Цена за бутылку")
plt.ylim(0, 600)
plt.figtext(0.1, -0.2, "Рисунок 7. - Зависимость цены от цвета вина")
```

Out[169]:

```
Text(0.1, -0.2, 'Рисунок 7. - Зависимость цены от цвета вина')
```

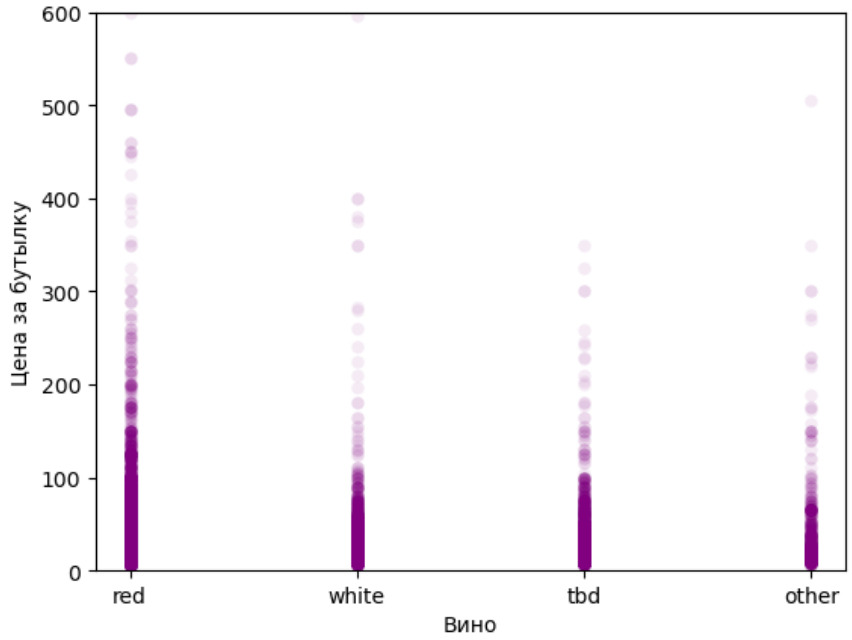


Рисунок 7. - Зависимость цены от цвета вина

Диаграмма также показывает, что красное вино в среднем выше по стоимости, чем белое вино.

In [170]:

```
sns.lmplot(x = 'price',
           y = 'points', data = df,
           height = 5, # height and width of the plot
           scatter_kws = {'color': 'purple'}, # color for the points
           line_kws = {'color': 'midnightblue'}) # color for the regression line
plt.xlim([0, 300])
plt.ylim([70, 101])
plt.xlabel("Цена за бутылку")
plt.ylabel("Оценка / рейтинг вина")
plt.title("Зависимость оценки / рейтинга вина от стоимости")
plt.figtext(0.1, -0.2, "Рисунок 8. - Зависимость оценки / рейтинга вина от стоимости")
```

Out[170]:

```
Text(0.1, -0.2, 'Рисунок 8. - Зависимость оценки / рейтинга вина от стоимости')
```

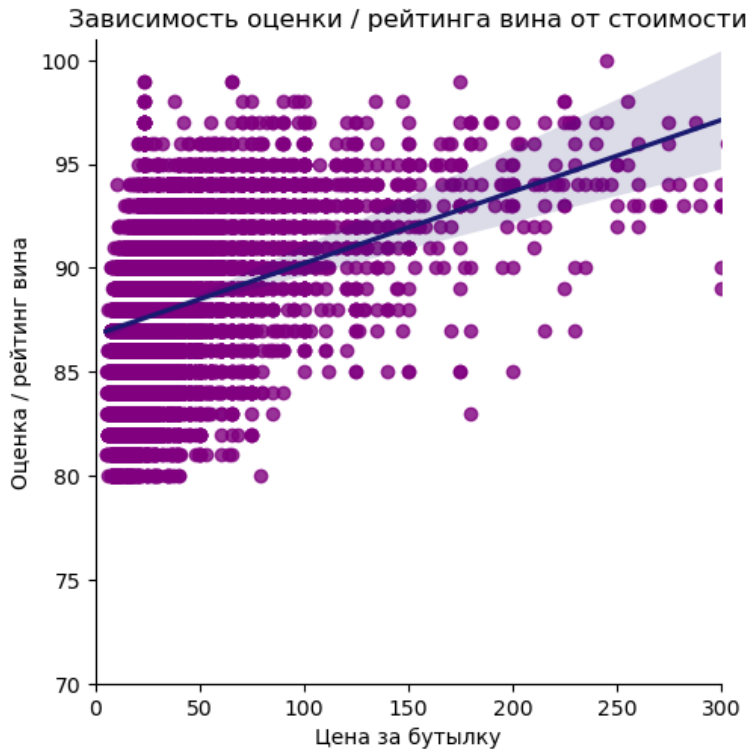


Рисунок 8. - Зависимость оценки / рейтинга вина от стоимости

Вывод: Мы оценили зависимость качества (вкусовых качеств) вина от цены. На основании графика можно предположить, что при цене до 80 - 100 у.е. с ростом цены, как правило, растет и качество вина. Тем не менее, после 100 евро эта зависимость сглаживается, и дальнейшее увеличение цены не приводит к увеличению качества вина.

Пейте качественное вино и не переплачивайте! :)

Выводы по итогам раздела 2:

Мы провели исследовательский анализ данных и выявили следующее:

- Определелили, что сорта Шардоне (Chardonnay), Каберне бленд (Cabernet Blend), Мерло (Merlot), а также Шираз (Syrah), Красный бленд Бордо (Bordeaux-style Red Blend) получают самые высокие оценки экспертов.
- Нашли сорт винограда с самым высоким рейтином по каждой провинции. Например, в Bordeaux - это Sémillon, в Калифорнии - \to Zinfandel, а в Тоскане - это White Blend.
- Выявили группу вин с очень высокими ценами. Предполагаем, что это коллекционные вина, имеющие длительный срок выдержки и высоко ценящиеся на рынке вин.
- Выбрали сорта с наибольшими ценами - Bordeaux-style Red Blend, Chardonnay, Grüner Veltliner, Furmint. Заметим, что топ сортов по ценам лишь частично пересекается с топ сортов по рейтингам. Совпадение в сортах - Красный бленд Бордо (Bordeaux-style Red Blend), Шардоне (Chardonnay).
- Для каждого региона нашли среднюю цену вина. Интересно, что самая высокая средняя цена бутылки вина - это цена токайского вина (Венгрия).
- Определили, популярные сорта вина в бюджетном сегменте (цену за бутылку до 10 евро). Выявили, что больше всего вин в сегменте до 10 евро продается из следующих сортов: Шардоне, Каберне Совиньон, Совиньон блан, Мерло и Розе.
  Указанные сорта являются достаточно неприхотливыми и получили распространение и в более суровых погодных условиях, соответствующие виноградники есть также на территории РФ (Крым, Таманский полуостров, Абрау-Дюрсо)
- Выявили закономерность влияния цвета на цену:
- Медианное и среднее значение цены красного вина выше. Предполагаем, что красное вино, в среднем, дороже белого.
- Вина с пометкой 'other'(напомним, что это шампанское, розе и игристое) в среднем продаются по цене как белые, однако по в выборке немного, и, соответственно, делать какие-либо выводы по указанной категории мы можем лишь с осторожностью.
- Выявили закономерность влияния рейтинга на цену:
- Закономерность некоторая есть - чем выше рейтинг, тем выше цена (коэффициент корреляции - 0,4).
- Вместе с тем, можно сделать и обратный вывод - чем выше цена, тем выше рейтинг Требуется дополнительное исследование

3. Портрета пользователя

Определим для пользователя каждого континента самые популярные сорта (топ-5).

In [171]:

```
df.groupby('country_to_continent')['variety'].describe()
```

Out[171]:

country_to_continent	count	unique	top	freq
Africa	304	28	Sauvignon Blanc	44
Asia	105	28	Cabernet Sauvignon	25
Europe	8702	332	Red Blend	694
Latin America	1538	60	Malbec	303
North America	8273	136	Pinot Noir	1408
Oceania	1067	51	Shiraz	173

для Африки:

In [172]:

```
top5 = df.groupby('country_to_continent')['variety'].get_group('Africa')
top5.value_counts().head(5)
```

Out[172]:

```
Sauvignon Blanc       44
Shiraz                44
Chardonnay            43
Cabernet Sauvignon    27
Pinotage              26
Name: variety, dtype: int64
```

для Азии:

In [173]:

```
df.groupby('country_to_continent')['variety'].get_group('Asia').value_counts().head(5)
```

Out[173]:

```
Cabernet Sauvignon          25
Chardonnay                  16
Red Blend                   12
Bordeaux-style Red Blend     9
Merlot                       5
Name: variety, dtype: int64
```

для Европы:

In [174]:

```
df.groupby('country_to_continent')['variety'].get_group('Europe').value_counts().head(5)
```

Out[174]:

```
Red Blend                   694
Bordeaux-style Red Blend    625
Chardonnay                  462
Riesling                    460
Sangiovese                  362
Name: variety, dtype: int64
```

для Латинской Америки:

In [175]:

```
df.groupby('country_to_continent')['variety'].get_group('Latin America').value_counts().head(5)
```

Out[175]:

```
Malbec                303
Cabernet Sauvignon    281
Chardonnay            142
Sauvignon Blanc       116
Red Blend             115
Name: variety, dtype: int64
```

для Северной Америки:

In [176]:

```
df.groupby('country_to_continent')['variety'].get_group('North America').value_counts().head(5)
```

Out[176]:

```
Pinot Noir            1408
Cabernet Sauvignon    1171
Chardonnay            1077
Syrah                  551
Zinfandel              515
Name: variety, dtype: int64
```

для Океании:

In [177]:

```
df.groupby('country_to_continent')['variety'].get_group('Oceania').value_counts().head(5)
```

Out[177]:

```
Shiraz             173
Sauvignon Blanc    164
Chardonnay         153
Pinot Noir         139
Riesling            80
Name: variety, dtype: int64
```

- В Африке самые популярные сорта винограда: Савиньон Блан, Шираз и Шардоне
- В Азии Каберне савиьон, Шардоне и Ред бленд
- В Латинской Америке Мальбек, Каберне савиьон, Шардоне
- В Северной Америке: Пино Нуар, Каберне савиьон, Шардоне
- В Океании:Шираз, Савиньон Блан, Шардоне и Пино Нуар

Теперь для пользователя каждого континента также определим, влияет ли рейтинг на цены по регионам.

Посчитаем корреляцию и отобразим тепловую матрицу

In [178]:

```
df.groupby('country_to_continent')['price','points'].corr()
```

Out[178]:

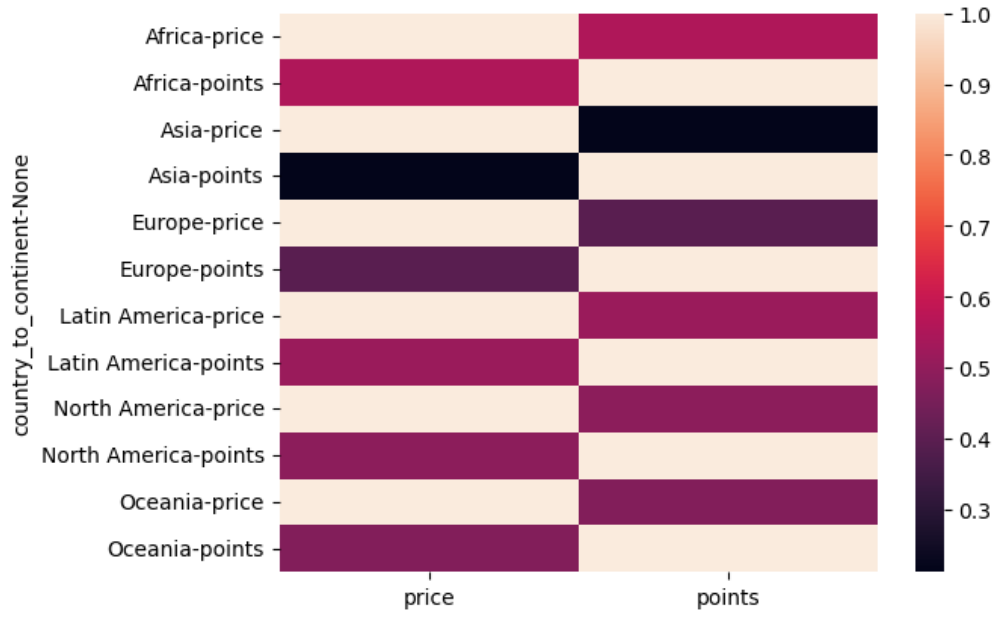
		price	points
country_to_continent			
Africa	**price**	1.000000	0.551393
	points	0.551393	1.000000
Asia	**price**	1.000000	0.211848
	points	0.211848	1.000000
Europe	**price**	1.000000	0.395288
	points	0.395288	1.000000
Latin America	**price**	1.000000	0.515544
	points	0.515544	1.000000
North America	**price**	1.000000	0.490214
	points	0.490214	1.000000
Oceania	**price**	1.000000	0.472855
	points	0.472855	1.000000

In [179]:

```
sns.heatmap(df.groupby('country_to_continent')['price','points'].corr())
```

Out[179]:

```
<Axes: ylabel='country_to_continent-None'>
```



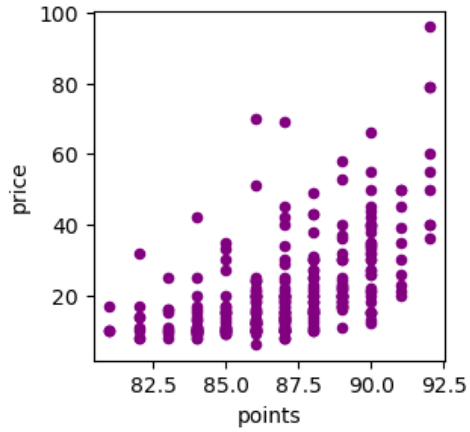
На основании таблицы корреляции по регионам, видим, что прослеживается некоторая связь между ценой и рейтингом в Африке, Латинской Америке, Северной Америке и Океании. В Европе и Азии связь цены и рейтинга гораздо ниже. Это может быть связано отчасти с тем, что Европа имеет длинную историю виноделия, и, вероятно, самое большое количество коллекционных вин, которые стоят выше рынка. Это отрывает цену и рейтинг друг от друга.

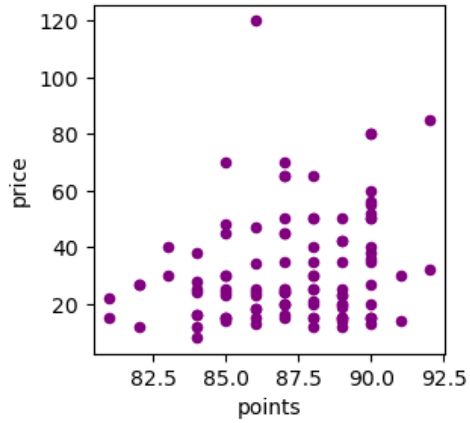
In [180]:

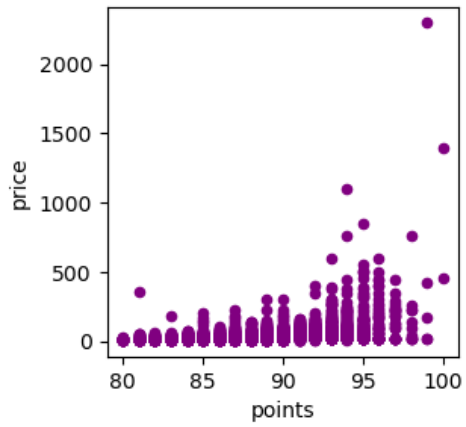
```
df.groupby('country_to_continent')['price','points'].plot(kind = 'scatter', x ='points', y = 'price', figsize=(3, 3), col
plt.figtext(0.1, -0.2, "Рисунок 9. - Корреляция между ценой и рейтингом по странам")
```

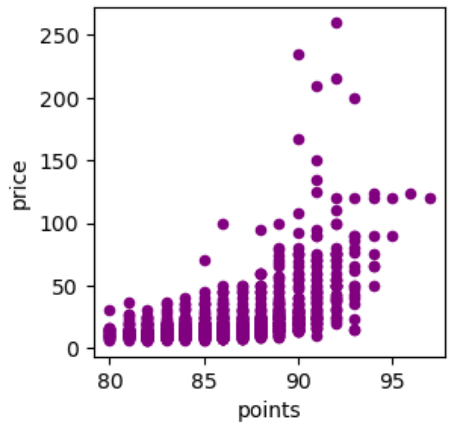
Out[180]:

```
Text(0.1, -0.2, 'Рисунок 9. - Корреляция между ценой и рейтингом по странам')
```









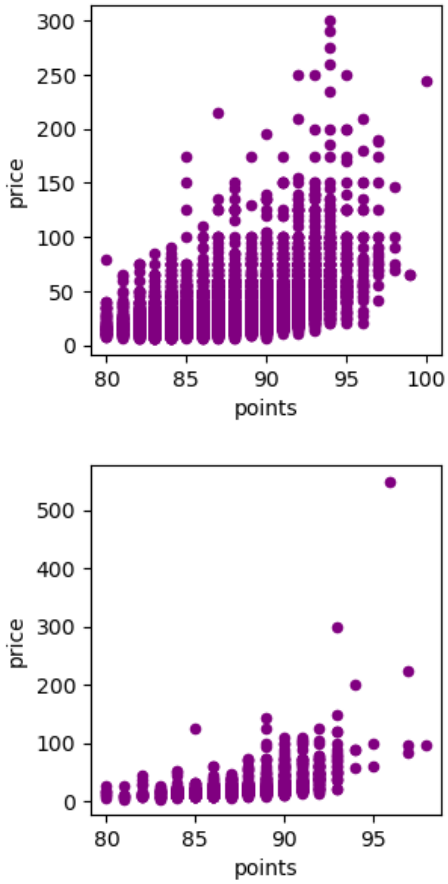


Рисунок 9. - Корреляция между ценой и рейтингом по странам

На графике также можно проследить некоторую зависимость цены от рейтинга

Выводы по итогам раздела 3:

Мы определили для пользователя каждого континента самые популярные сорта (топ-5).

- В Африке самые популярные сорта винограда: Савиньон Блан, Шираз и Шардоне
- В Азии Каберне савиьон, Шардоне и Ред бленд
- В Латинской Америке Мальбек, Каберне савиьон, Шардоне
- В Северной Америке: Пино Нуар, Каберне савиьон, Шардоне
- В Океании:Шираз, Савиньон Блан, Шардоне и Пино Нуар
- В Европе: Ред бленд, Бордо Ред бленд, Шардоне и Рислинг

Для пользователя каждого континента также определили, влияет ли рейтинг на цены по регионам:

- На основании таблицы корреляции и графиков по регионам, видим, что прослеживается некоторая связь между ценой и рейтингом в Африке, Латинской Америке, Северной Америке и Океании. В Европе и Азии связь цены и рейтинга гораздо ниже.
- Это может быть связано отчасти с тем, что Европа имеет длинную историю виноделия, и, вероятно, самое большое количество коллекционных вин, которые стоят выше рынка. Это отрывает цену и рейтинг друг от друга.

4. Исследование статистических показателей.

Подсчитаем среднее количество, дисперсию и стандартное отклонение для цен на продукт различных регионов.

In [181]:

```
df.groupby('country_to_continent')['price'].describe()
```

Out[181]:

country_to_continent	count	mean	std	min	25%	50%	75%	max
Africa	304.0	21.289474	13.629531	6.0	11.75	17.0	25.0	96.0
Asia	105.0	31.619048	19.297958	8.0	16.00	25.0	40.0	120.0
Europe	8702.0	34.966444	50.941065	5.0	16.00	23.0	39.0	2300.0
Latin America	1538.0	21.013654	21.078913	6.0	11.00	15.0	21.0	260.0
North America	8273.0	33.548773	23.218583	6.0	19.00	28.0	41.0	300.0
Oceania	1067.0	28.035614	27.997733	5.0	15.00	20.0	30.0	550.0

Построим гистограммы. Опишем распределения.

In [182]:

```
df[['country_to_continent', 'price']].hist(column = 'price', by ='country_to_continent', bins = 50, figsize = (11,9), leg
plt.xlim(0, 150)
plt.figtext(0, 0, "Рисунок 9. - Распределение уровня цен по континентам")
```

Out[182]:

```
Text(0, 0, 'Рисунок 9. - Распределение уровня цен по континентам')
```

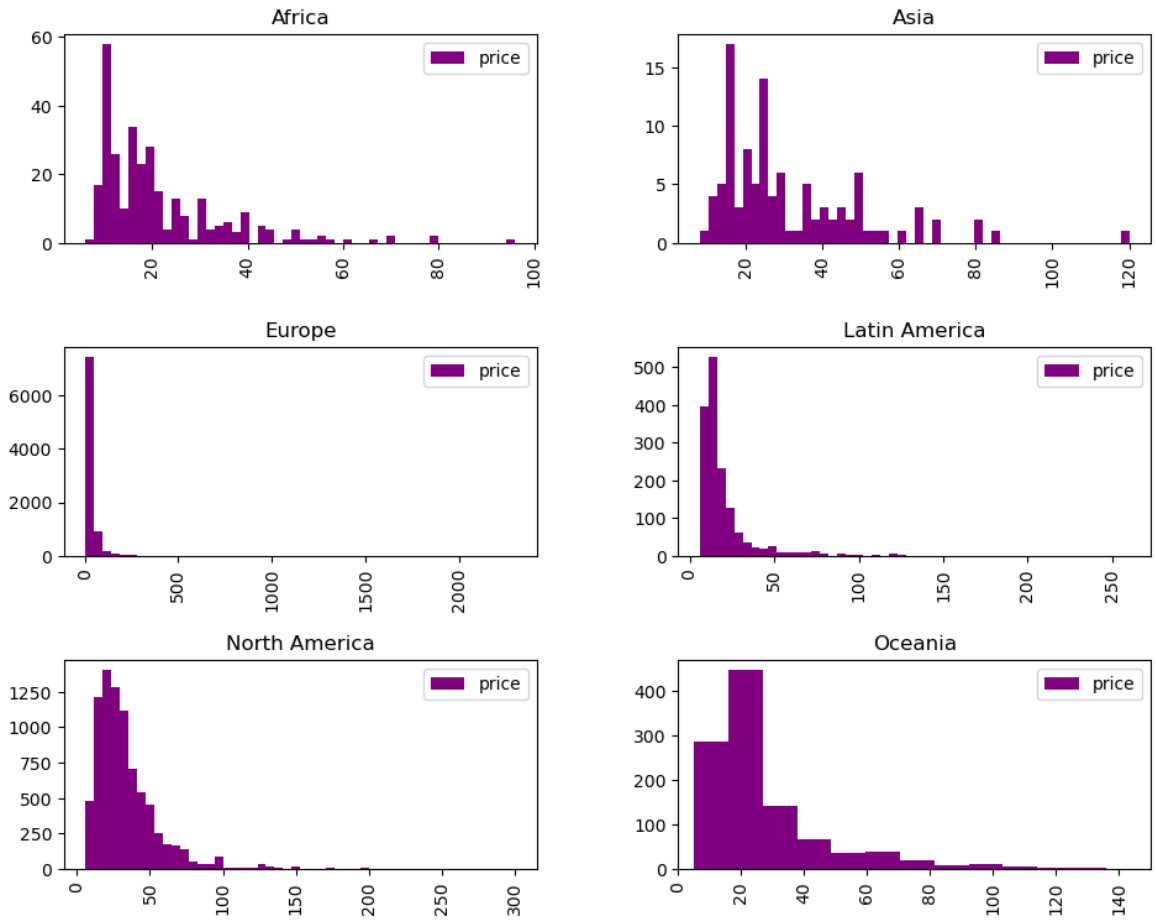


Рисунок 9. - Распределение уровня цен по континентам

In [183]:

```
df.price.hist(bins = 500, color = 'purple')
plt.xlim(0, 150)
plt.xlabel('Цена')
plt.ylabel('Количество проданных')
plt.title('Распределение уровня цен')
plt.figtext(0.1, -0.1, "Рисунок 10. - Распределение уровня цен")
```

Out[183]:

```
Text(0.1, -0.1, 'Рисунок 10. - Распределение уровня цен')
```

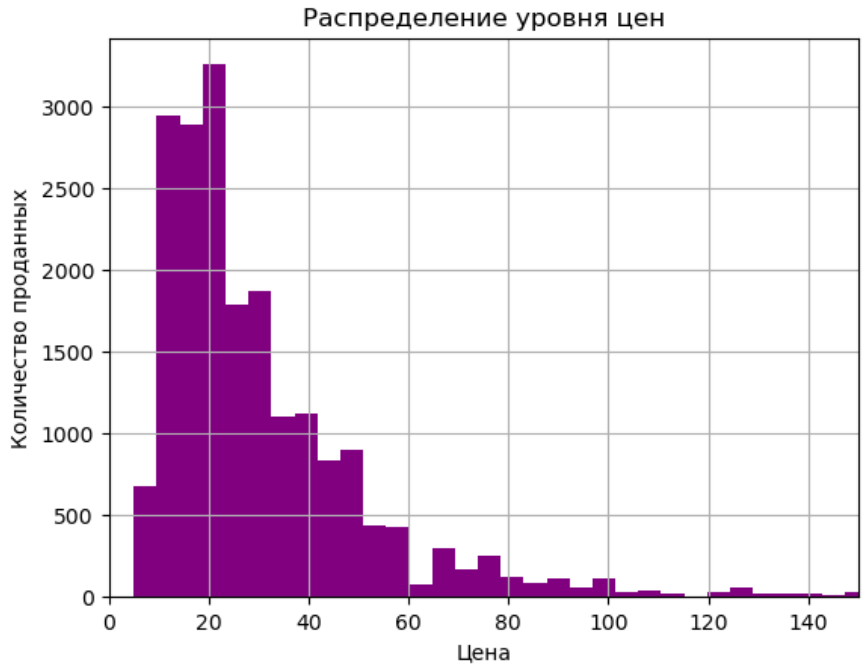


Рисунок 10. - Распределение уровня цен

Распределения цены по регионам похожи на нормальные. Есть аномалии у Европы, поскольку есть группа очень дорогих вин, но за исключением этих вин распределение нормальное.

Построим линейную регрессию зависимости между ценой продукта и его рейтингом:

In [184]:

```
feature_cols = ['price']
X = df[feature_cols]
y = df.points
```

In [185]:

```
from sklearn.linear_model import LinearRegression
linreg = LinearRegression()
linreg.fit(X, y)
```

Out[185]:

```
LinearRegression
LinearRegression()
```

In [186]:

```
print (linreg.intercept_)
print (linreg.coef_)
```

```
86.77266295060727
[0.03446418]
```

Мы построили линейную регрессию зависимости цены от рейтинга. При нулевом значении рейтинга - цена составит 86,77 у.е. Каждая новая добавленная единица в цене, прибавляет к рейтингу 0.03.

Выводы по итогам раздела 4:

Распределение цены по регионам выглядит как нормальное. Есть аномально высокие значения цены у вин Европы, поскольку нами ранее была выявлена группа очень дорогих вин. За исключением этих вин распределение нормальное.

Мы построили линейную регрессию зависимости цены от рейтинга. При нулевом значении рейтинга - цена составит 86,77 у.е. Каждая новая добавленная единица в цене, прибавляет к рейтингу 0.03.

5. Проверка гипотез

Рассмотрим взаимосвязь рейтингов красного и белого вина. Выделим 2 гипотезы:

H0: Средние пользовательские рейтинги красного и белого вина одинаковые.

H1: Средние пользовательские рейтинги красного и белого вина разные.

In [187]:

```
red = df[df.wine_color == 'red'].points
white = df[df.wine_color == 'white'].points
```

Подготовили выборки

In [188]:

```
print(red.mean())
print(white.mean())
# Средние отличаются
```

```
88.19504617243463
87.56445215666832
```

In [189]:

```
value,p = st.normaltest(red)
if p <0.05:
    print('red распределено нормально')
else:
     print('red распределено не нормально')
```

```
red распределено нормально
```

In [190]:

```
value,p = st.normaltest(white)
if p <0.05:
    print('white распределено нормально')
else:
     print('white распределено не нормально')
```

```
white распределено нормально
```

В качестве уровня значимости выберем альфа = 0.05.

In [191]:

```
Ho = "Средние пользовательские рейтинги красного и белого вина одинаковые"
H1 = "Средние пользовательские рейтинги красного и белого вина разные"

t,p_value  = st.ttest_ind(red, white, axis = 0)

if p_value < 0.05:
    print(f'{H1}, поскольку p_value {p_value.round(3)} < 0.05')
else:
    print(f'{Ho}, поскольку p_value {p_value.round(3)} > 0.05')
```

```
Средние пользовательские рейтинги красного и белого вина разные, поскольку p_value 0.0 < 0.05
```

In [192]:

```
sns.histplot(red, color='purple')
sns.histplot(white, color='midnightblue')
plt.figtext(0.1, -0.1, "Рисунок 11. - Распределение пользовательских рейтингов красного и белого вина")
```

Out[192]:

```
Text(0.1, -0.1, 'Рисунок 11. - Распределение пользовательских рейтингов красного и белого вина')
```

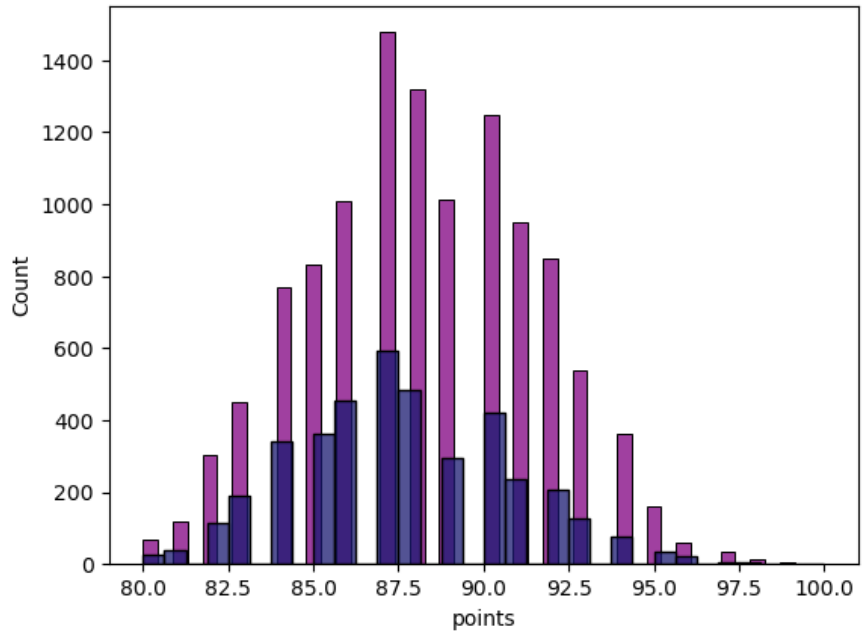


Рисунок 11. - Распределение пользовательских рейтингов красного и белого вина

Нулевая гипотеза не подтвердилась. Можем сделать вывод, что средние пользовательские рейтинги красного и белого вина отличаются. Наложение гистограмм также подтверждает отличие средних пользовательских рейтинговкрасного и белого вина.

Рассмотрим взаимосвязь цен двух популярных сортов вина. Выделим 2 гипотезы:

H0: Средние цены двух популярных сортов вина одинаковые.

H1: Средние цены двух популярных сортов вина разные.

In [193]:

```
df.variety.value_counts()
```

Out[193]:

```
Pinot Noir                  1945
Chardonnay                  1893
Cabernet Sauvignon          1636
Red Blend                   1328
Bordeaux-style Red Blend     952
                            ...
Schiava                        1
Merlot-Shiraz                  1
White Port                     1
Meoru                          1
Mansois                        1
Name: variety, Length: 420, dtype: int64
```

Вспомним, что 2 самых популярных сорта винограда - Pinot Noir и Chardonnay.

In [194]:

```
pinot_noir = df[df.variety == 'Pinot Noir'].price
chardonnay = df[df.variety == 'Chardonnay'].price
```

In [195]:

```
print(pinot_noir.mean())
chardonnay.mean()
# Средние отличаются
```

```
43.883290488431875
```

Out[195]:

```
32.770206022187004
```

In [196]:

```
value,p = st.normaltest(pinot_noir)
if p <0.05:
    print('pinot_noir распределено нормально')
else:
     print('pinot_noir распределено не нормально')
```

```
pinot_noir распределено нормально
```

In [197]:

```
value,p = st.normaltest(chardonnay)
if p <0.05:
    print('chardonnay распределено нормально')
else:
     print('chardonnay распределено не нормально')
```

```
chardonnay распределено нормально
```

В качестве уровня значимости выберем альфа = 0.05.

In [198]:

```
Ho = "Средние цены двух популярных сортов вина одинаковые"
H1 = "Средние цены двух популярных сортов вина разные"

t,p_value  = st.ttest_ind(pinot_noir, chardonnay, axis = 0)

if p_value < 0.05:
    print(f'{H1}, поскольку p_value {p_value.round(3)} < 0.05')
else:
    print(f'{Ho}, поскольку p_value {p_value.round(3)} > 0.05')
```

```
Средние цены двух популярных сортов вина разные, поскольку p_value 0.0 < 0.05
```

In [199]:

```
sns.histplot(pinot_noir, color='b')
sns.histplot(chardonnay, color='y')
plt.figtext(0.1, -0.1, "Рисунок 11. - Распределение средних цен двух популярных сортов вина Пино Нуар и Шардоне")
plt.xlim(0, 250)
plt.figure(figsize = (10,4))
```

Out[199]:

```
<Figure size 1000x400 with 0 Axes>
```

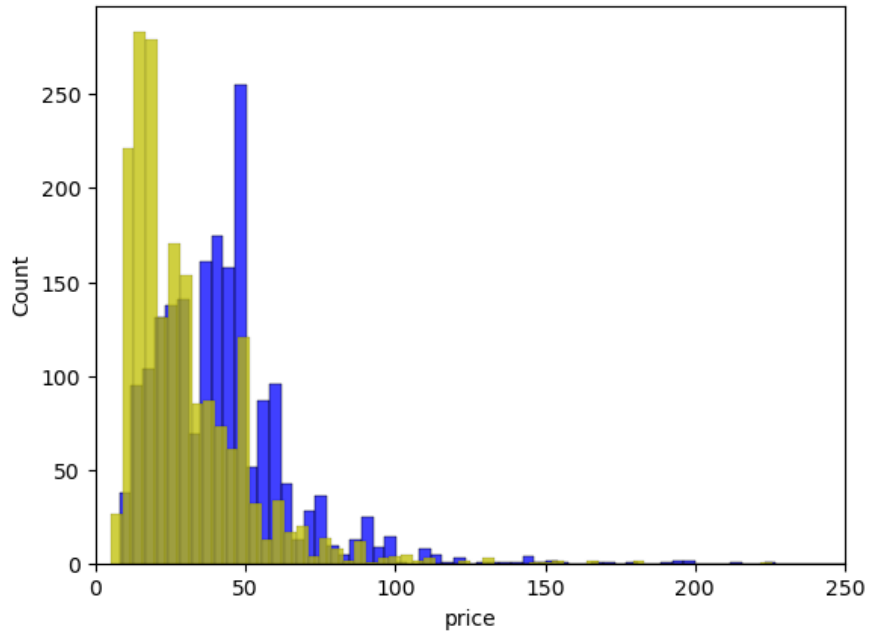


Рисунок 11. - Распределение средних цен двух популярных сортов вина Пино Нуар и Шардоне

```
<Figure size 1000x400 with 0 Axes>
```

Нулевая гипотеза не подтвердилась. Можем сделать вывод, что средние цены двух популярных сортов вина отличаются. Наложение гистограмм также подтверждает отличие средних цен двух популярных сортов вина.

Выводы по итогам раздела 5:

Нулевая гипотеза не подтвердилась. Можем сделать вывод, что средние пользовательские рейтинги красного и белого вина отличаются. Наложение гистограмм также подтверждает отличие средних пользовательских рейтинговкрасного и белого вина.

Нулевая гипотеза не подтвердилась. Можем сделать вывод, что средние цены двух популярных сортов вина отличаются. Наложение гистограмм также подтверждает отличие средних цен двух популярных сортов вина.

6. Выводы

В ходе работы проанализировано 20 000 записей. Мы обработали пропуски столбца price - заменили на медианные значения по каждой провинции. Пропуски в designation, region_1, region_2 мы не заполняли, поскольку эти признаки мы не будем использовать в исследовании. Преобразовали тип данных столбца price из float в int, поскольку все значения цены в базе данных целые. Это позволяет разумнее использовать память.

В ходе анализа мы:

- Выявили группу вин с очень высокими ценами. Предполагаем, что это коллекционные вина, имеющие длительный срок выдержки и высоко ценящиеся на рынке вин.
- Выяснили, что сорта Красный бленд Бордо (Bordeaux-style Red Blend), Шардоне (Chardonnay) занимают высшие места в топе как по рейтингу, так и по цене.
- Определили, популярные сорта вина в бюджетном сегменте (цену за бутылку до 10 евро). Выявили, что больше всего вин в сегменте до 10 евро создается из следующих сортов: Шардоне, Каберне Совиньон, Совиньон блан, Мерло и Розе. Указанные сорта являются достаточно неприхотливыми и получили распространение и в более суровых погодных условиях, соответствующие виноградники есть также на территории РФ (Крым, Таманский полуостров, Абрау-Дюрсо)

Самые популярные сорта:

- В Европе: Ред бленд, Бордо Ред бленд, Шардоне и Рислинг

- В Азии: Каберне савиьон, Шардоне и Ред бленд
- В Северной Америке: Пино Нуар, Каберне савиьон, Шардоне
- В Африке: самые популярные сорта винограда: Савиньон Блан, Шираз и Шардоне
- В Латинской Америке: Мальбек, Каберне савиьон, Шардоне
- В Океании: Шираз, Савиньон Блан, Шардоне и Пино Нуар

Ответили на вопрос: влияет ли рейтинг на цены в зависимости от континента?

- На основании таблицы корреляции и графиков по континентам, видим, что прослеживается некоторая связь между ценой и рейтингом в Африке, Латинской Америке, Северной Америке и Океании.
- В Европе и Азии связь цены и рейтинга гораздо ниже. Это может быть связано отчасти с тем, что Европа имеет длинную историю виноделия, и, вероятно, самое большое количество коллекционных вин, которые стоят выше рынка. Это отрывает цену и рейтинг друг от друга.

Мы построили линейную регрессию зависимости цены от рейтинга. При нулевом значении рейтинга - цена составит 86,77 у.е. Каждая новая добавленная единица в цене, прибавляет к рейтингу 0.03.

По итогам проверки гипотез:

- Средние пользовательские рейтинги красного и белого вина отличаются.
- Средние цены двух популярных сортов вина отличаются.

Результаты:

Самые большие рынки вина - это Европа и Северная Америка. Покупатели в этих регионах ценят качество, и готовы приобретать коллекционное вино. Коллекционные вина, вероятно, будут иметь спрос в этих регионах и дальше, на что стоит обратить внимание ритейлерам. В остальных регионах коллекционное вино следует продавать ограничено.

В Европе наиболее популярные сорта, на которые стоит обратить внимание: Ред бленд, Бордо Ред бленд, Шардоне и Рислинг

В Северной Америке - это Пино Нуар, Каберне савиньон, Шардоне

Всеми любимый и "понимаемый" лидер во всех регионах сорт винограда - это Шардоне. Он популярен как в бюджетном сегменте, так и в премиальном. Магазинам следует всегда иметь запас такого вина.

Также красное вино, как правило, продается дороже и чаще белого, поэтому стоит подготовить запасы вина соответствующим образом.

Список литературы

1. Андерсон, К, Аналитическая культура: от сбора данных до бизнес-результатов / Карл Андерсон. - Москва : Манн, Иванов и Фербер, 2017. - 324 с.
2. Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони, Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — СПб.: Питер, 2019.
3. Мэтиз Э., Изучаем Python. Программирование игр, визуализация данных, веб-приложения. — СПб.: Питер, 2017.
4. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018.
5. Рашка С., Рашка С. Р28 Python и машинное обучение / пер. с англ. А. В. Логунова. - М.: ДМК Пресс, 2017.
6. Шарден Б., Массарон Л., Боскетти А., Крупномасштабное машинное обучение вместе с Python. Пер. с анг. А. В. Логунова. – М.: ДМК Пресс, 2018.