به نام خدا

گزارش تمرین سری اول

درس داده کاوی

جناب آقای دکتر سعیدرضا خرد پیشه و هادی فراهانی

دانشجو: رضا جمشیدکیانی (۹۸۴۲۲۰۴۴)

فهرست مطالب

مجموعه داده اول: اماکن اجاره ای در نیویورک AB NYC 2019

۱- فراخوانی پکیج ها و داده:

در ابتدا لازم است پکیج های لازم فراخوانی شوند:

```
In [32]:
from sklearn.model_selection import cross_val_score
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as ss
```

سپس فایل دیتاست که روی گوگل درایو ذخیره شده را در برنامه لود می کنیم تا پس از خواندن در متغیر rb ذخیره شود:

```
from google.colab import drive
drive.mount('/content/gdrive');
! ls '/content/gdrive/MyDrive/Data/Data Mining/'
root = '/content/gdrive/MyDrive/Data/Data Mining/'
rb = pd.read_csv(root+"AB_NYC_2019.csv")
rb.head()

rb.isnull().sum()
rb.duplicated().value_counts()
rb.columns
rb.drop(['id', 'name', 'host_id', 'host_name',
         'last_review'], inplace=True, axis=1)
```

۲- همبستگی (Correlation) بین قیمت مکان و حداقل شب های اقامت چگونه است؟ آیا همبستگی بین قیمت و حداقل شب‌های اقامت وجود دارد یا خیر؟

باتوجه به اینکه اطلاعی در مورد نرمال بودن نمونه ها نداریم و ممکن است داده های ما پرت باشند، Spearman correlation می زنیم:

```
scipy.stats.spearmanr(rb['price'], rb['minimum_nights'])
```

خروجی شامل مقدار correlation و pvalue به شکل زیر خواهد بود:

```
SpearmanrResult(correlation=0.10128900445001728,
pvalue=1.1582087858911544e-111)
```

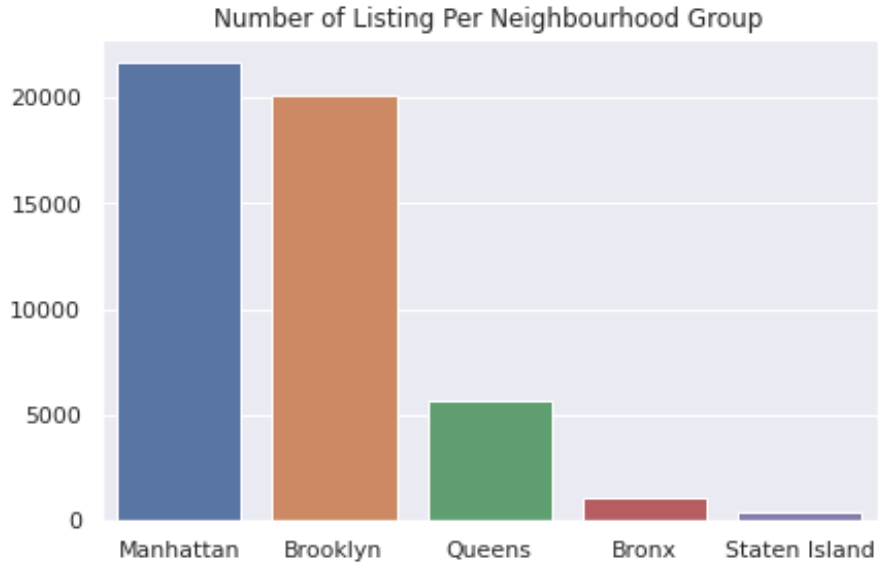
از آنجایی که pvalue بیش از 0/05 است پس فرض H0 رد می شود و correlation بین این دو ستون وجود ندارد. در واقع همبستگی بین قیمت مکان و حداقل شب های اقامت مشاهده نمی کنیم.

۳- میانگین قیمت روزانه هر گروه از مناطق اجاره چگونه است؟

- ابتدا ستون جدیدی با نام price_per_day برای محاسبه میانگین قیمت هر روز اجاره، لحاظ می کنیم. نحوه محاسبه مقدار این ستون، تقسیم قیمت بر حداقل شب های اجاره است.

```
• rb["price_per_day"] = rb["price"]/rb["minimum_nights"]
```

- تعداد اماکن (لیست های) هر منطقه:



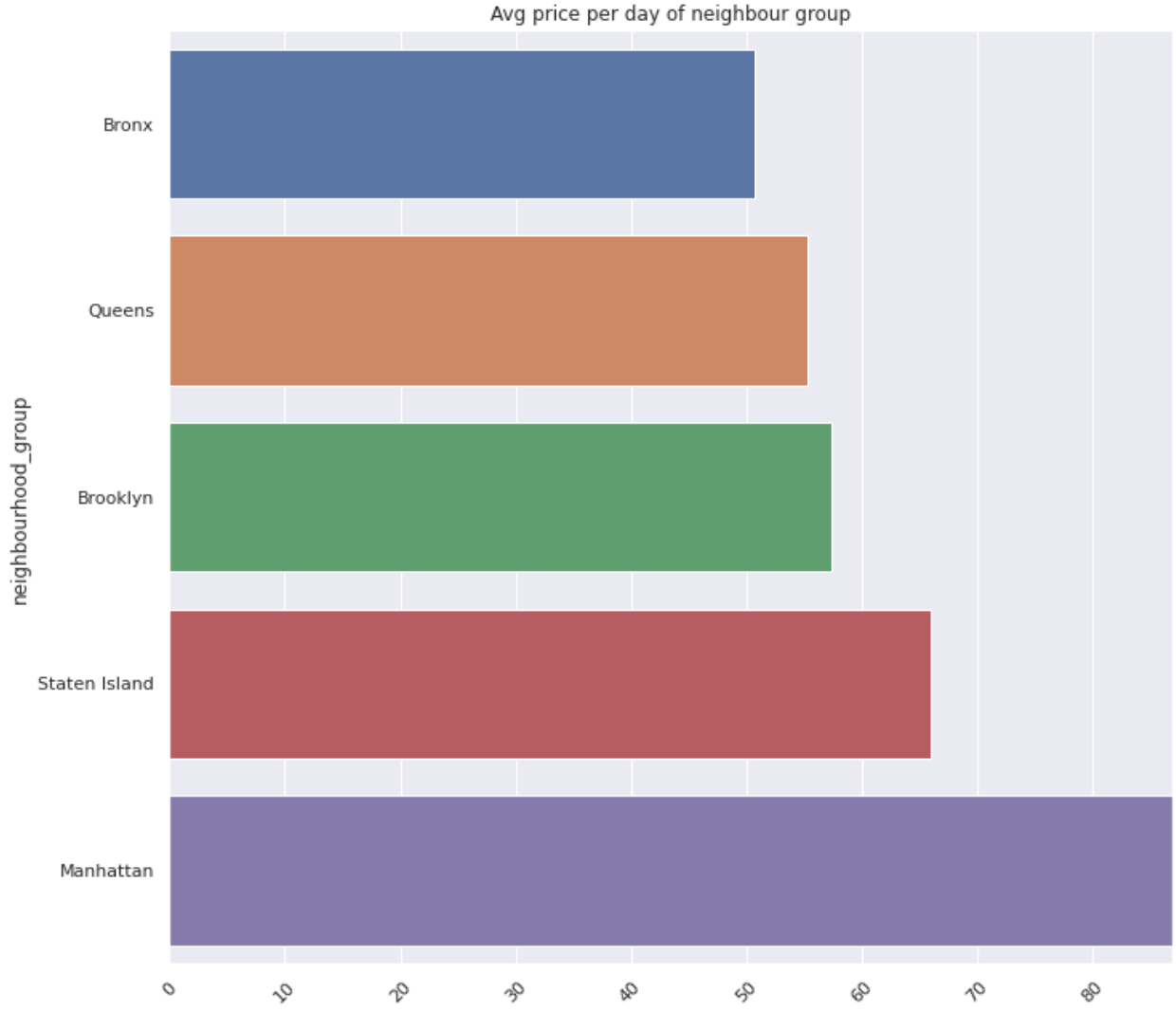
- سپس بر اساس میانگین قیمت روزانه و تعداد خانه های هر گروه که در مرحله قبل بدست آمد، مجموع قیمت روزانه هر گروه منطقه بدست می آید.

```
Avg Price of Neighbourhood Group Brooklyn: 57.42877835024624
Avg Price of Neighbourhood Group Manhattan: 86.94598085173197
Avg Price of Neighbourhood Group Queens: 55.30723215856462
Avg Price of Neighbourhood Group Staten Island: 65.94196326984459
Avg Price of Neighbourhood Group Bronx: 50.70361009604071
```

```
rb["price_per_day"] = rb["price"]/rb["minimum_nights"]
rb["Neighbour"] = rb["neighbourhood"].astype(str)+"_"+rb["neighbourhood_group"].astype(str)
sns.barplot(rb["neighbourhood_group"].value_counts().index, rb["neighbourhood_group"].value_counts().values)
plt.title("Number of Listing Per Neighbourhood Group")
plt.show()
for ng in rb["neighbourhood_group"].unique():
  print(f'Avg Price of Neighbourhood Group {ng}: {rb[rb["neighbourhood_group"]==ng]["price_per_day"].sum()/len(rb[r

grp_neighbour = rb.groupby("neighbourhood_group")["price_per_day"].mean().sort_values()
plt.figure(figsize=(10, 10))
plt.xticks(rotation=45)
plt.xlim(0,max(grp_neighbour.values))
sns.barplot(grp_neighbour.values,grp_neighbour.index)
plt.title("Avg price per day of neighbour group")
plt.show()
```

- و در آخر میانگین قیمت روزانه هر گروه منطقه به همراه نمودار آن:

میانگین قیمت روزانه هر گروه منطقه محاسبه و مرتب شده اند. که بر اساس آن Manhattan بیشترین میانگین قیمت روزانه و Bronx کمترین میانگین قیمت روزانه را در گروه های مختلف دارد.



```
grp_neighbour = rb.groupby("neighbourhood_group")["price_per_day"].mean().sort_values()
plt.figure(figsize=(10, 10))
plt.xticks(rotation=45)
plt.xlim(0,max(grp_neighbour.values))
sns.barplot(grp_neighbour.values,grp_neighbour.index)
plt.title("Avg price per day of neighbour group")
plt.show()
```

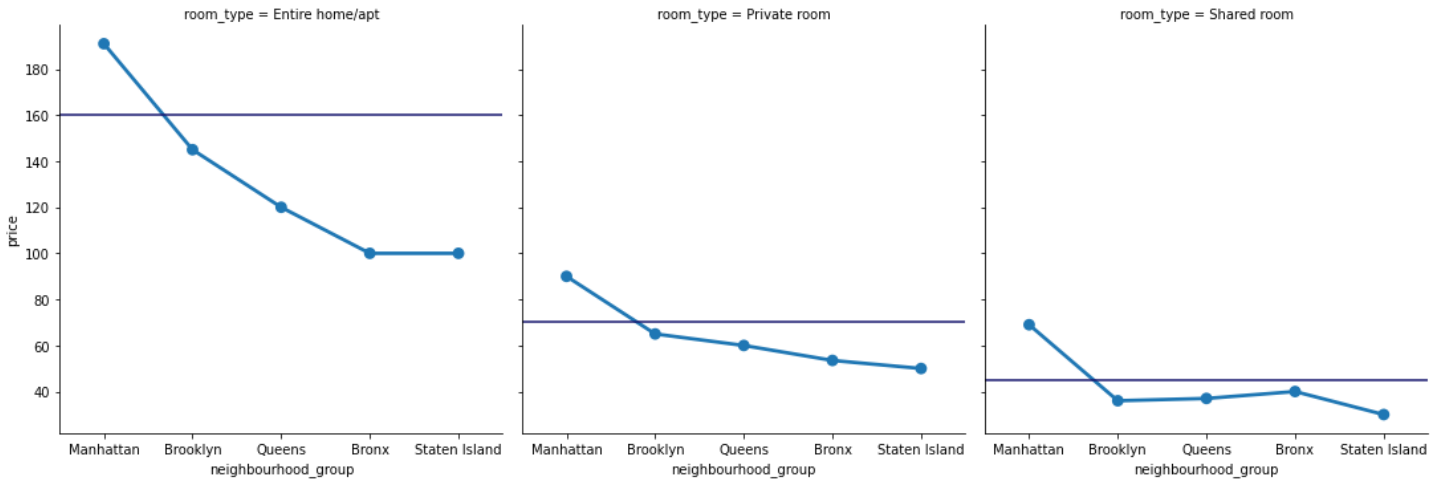
۴- اجاره انواع اتاق ها در هر ناحیه چگونه است؟

برای مقایسه اجاره انواع اتاق در ناحیه های مختلف، از میانه (Median) قیمت ها استفاده می کنیم.

بدین منظور، انواع اتاق ها را بر اساس میانه آن، گروه بندی می کنیم و در متغیر آرایه ای جدیدی ذخیره می کنیم.

در واقع میانه انواع اتاق ها را محاسبه می کنیم. سپس بر اساس مناطق مختلف، گروه بندی می کنیم.

- خانه های دربست/آپارتمان در منطقه منهتن بیشترین قیمت را دارند و در محله Stalen island کمترین قیمت را این خانه ها دارند. (با درنظرگرفتن میانه قیمت این خانه ها در مناطق مختلف)
- اتاق های شخصی در منطقه منهتن بیشترین قیمت و در محله Stalen island کمترین قیمت را دارند.
- اتاق های اشتراکی در منطقه منهتن، بیشترین قیمت و در محله Stalen island کمترین قیمت را دارند.

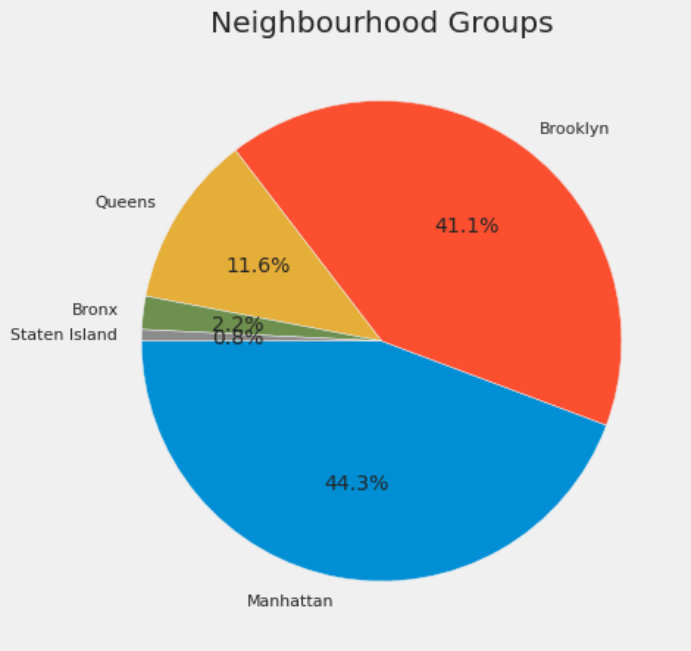



```
# A
avg_roomtype_cost = rb.groupby('room_type').price.median()
top_price = rb.groupby(['neighbourhood_group', 'room_type']).median(
).sort_values(by='price', ascending=False).reset_index()
g = sns.catplot(x='neighbourhood_group', y='price', data=top_price,
                ci=False, estimator=np.median, kind='point', col='room_type')
for i in range(len(avg_roomtype_cost)):
    g.axes[0][i].axhline(avg_roomtype_cost[i], color='midnightblue')
plt.show()
```

۵- پراکنش اماکن اجاره ای در محله ها چگونه است؟

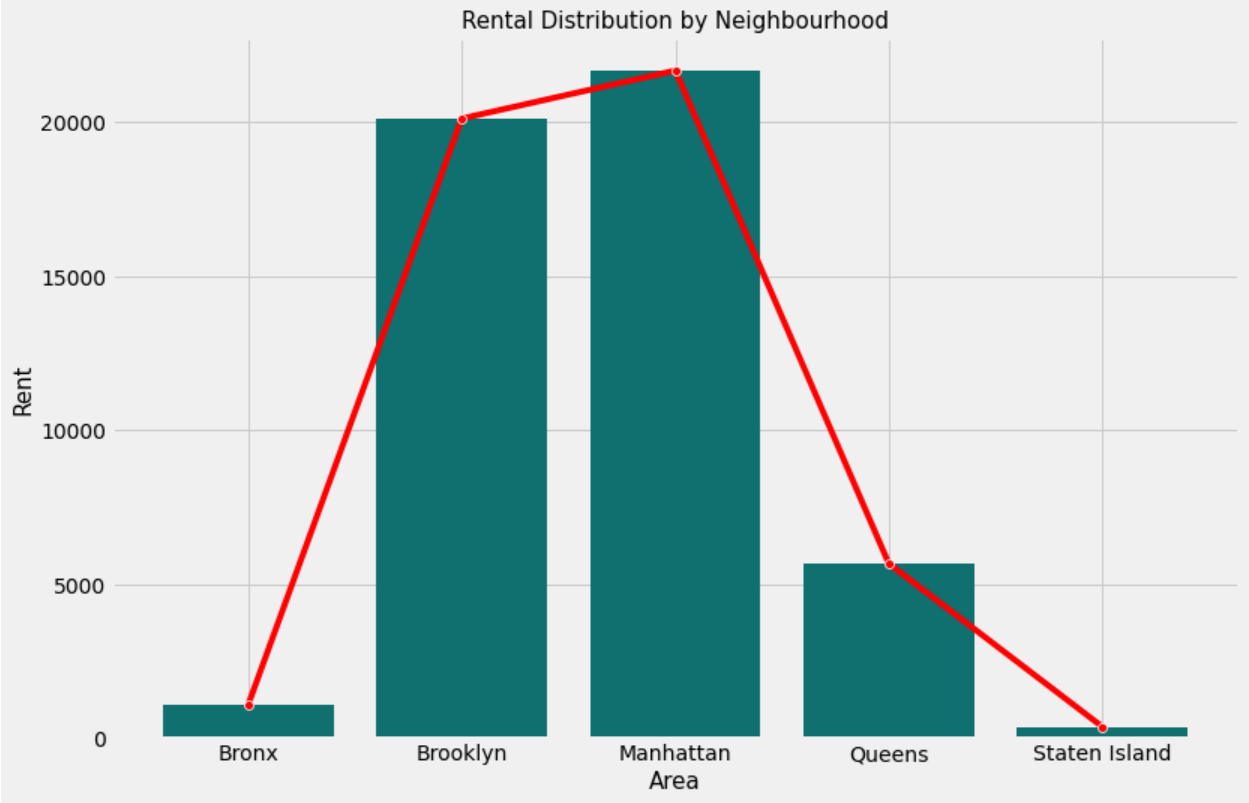
محله ها را بر اساس تعداد اماکن اجاره ای، گروه بندی می کنیم.

بر این اساس بسیاری از خانه های اجاره‌ای در Manhattan (۴۴.۳ درصد) و Brooklyn (۴۱.۱ درصد) هستند.



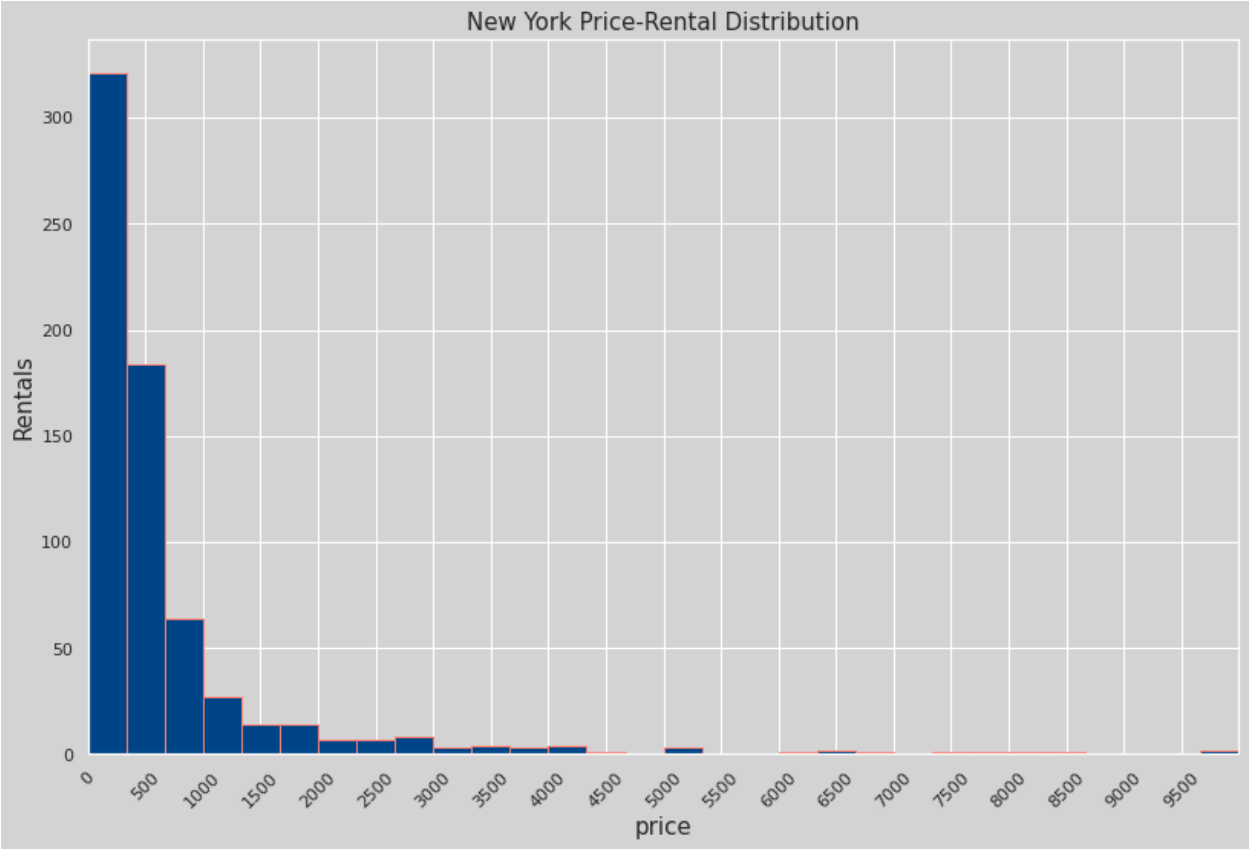

```
:
# B
plt.style.use('fivethirtyeight')
plt.figure(figsize=(13, 7))
plt.title("Neighbourhood Groups")
g = plt.pie(rb.neighbourhood_group.value_counts(
), labels=rb.neighbourhood_group.value_counts().index, autopct='%1.1f%%', startangle=180)
plt.show()
plt.close()
```

در نمودار زیر، توزیع اجاره ها در مناطق مختلف را مشاهده می کنیم. مشابه نمودار دایره ای قبل، منطقه منهتن بیشترین مکان های اجاره ای و Staten Island کمترین مکان ها را دارد.




```
neighbourhood = rb.groupby('neighbourhood_group')[
    'neighbourhood'].count().reset_index()
fig, ax = plt.subplots(figsize=(12, 8))
sns.barplot(x=neighbourhood[neighbourhood.columns[0]],
            y=neighbourhood[neighbourhood.columns[1]], color='teal', ax=ax)
sns.lineplot(x=neighbourhood[neighbourhood.columns[0]],
             y=neighbourhood[neighbourhood.columns[1]], color='r', marker='o', ax=ax)
plt.ylabel('Rent', fontsize='15')
plt.xlabel('Area', fontsize='15')
plt.title('Rental Distribution by Neighbourhood', fontsize='15')
plt.grid('x')
plt.show()
sns.set()
```

۶- توزیع اجاره ها در نیویورک چگونه است؟

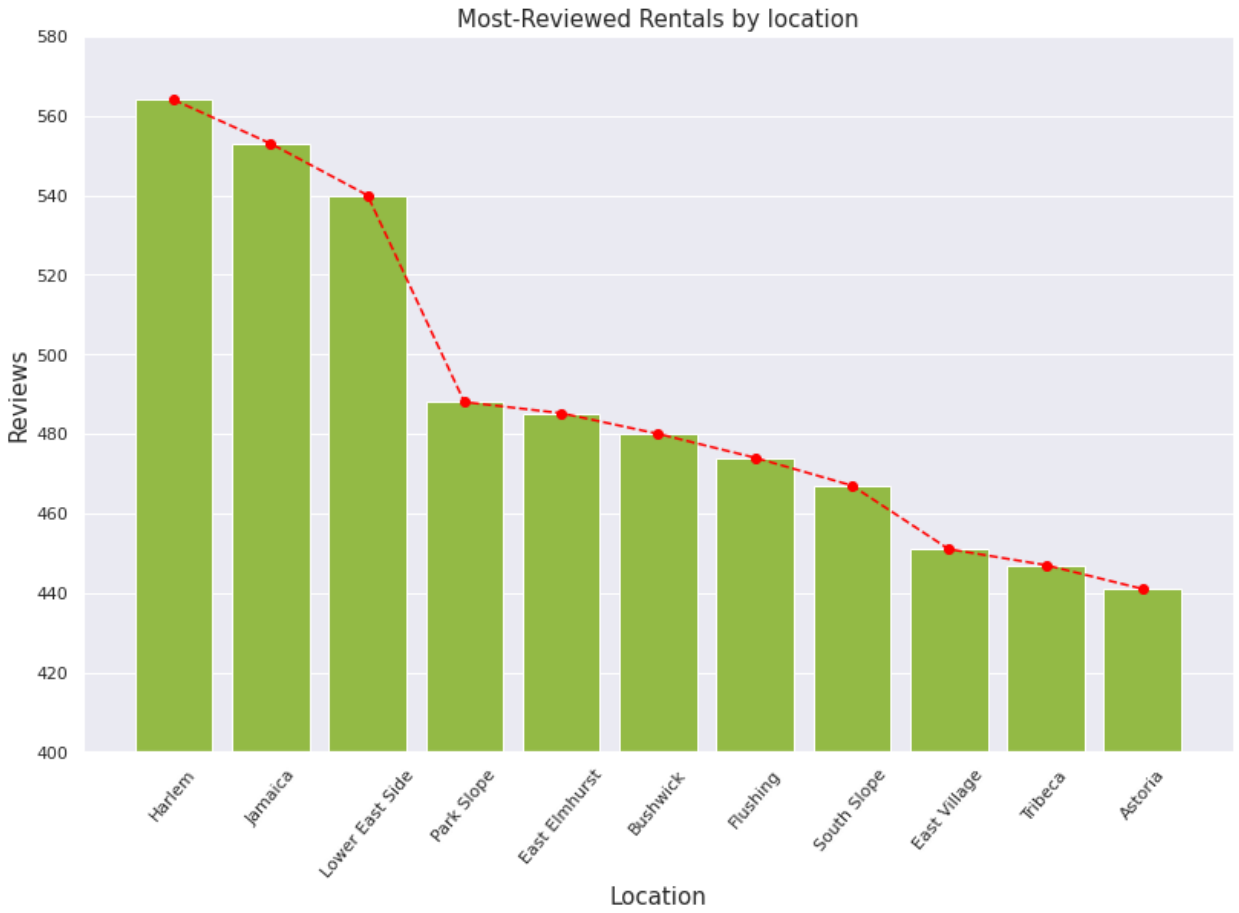


در اینجا فراوانی قیمت اجاره ها را می خواهیم بررسی کنیم که هر رنج قیمت، چه میزان فراوانی دارد. محور افقی را به فواصل ۵۰۰ تایی تقسیم می کنیم. ستون اول نشان می دهد که رنج قیمت ۰-۵۰۰ دلار، بیش از ۳۰۰ فراوانی دارد. در واقع بیش از ۳۰۰ واحد اجاره ای در رنج بین ۰-۵۰۰ دلار هستند. ستون آخر یعنی رنج قیمت ۹۵۰۰-۱۰۰۰۰ دلار، کمترین فراوانی واحد اجاره ای را دارد.

```
price = rb.loc[:, ['neighbourhood', 'price']].set_index('neighbourhood')
price_stats = rb['price'].describe().reset_index()
price_counts = price.price.value_counts().reset_index()
price_counts.rename(columns={'index': 'price', 'price': 'count'}, inplace=True)
fig2, ax = plt.subplots(figsize=(12, 8))
fig2.patch.set_facecolor('lightgray')
ax.set_facecolor('lightgray')
plt.hist(price_counts['price'], bins=30, color='#004488', edgecolor='salmon')
ax.set_xticks(range(0, 10000, 500))
for tick in ax.get_xticklabels():
    tick.set_rotation(45)
plt.xlabel('price', fontsize='15')
plt.ylabel('Rentals', fontsize='15')
plt.xlim((-0.5, 10000))
plt.title('New York Price-Rental Distribution', fontsize='15')
plt.show()
```

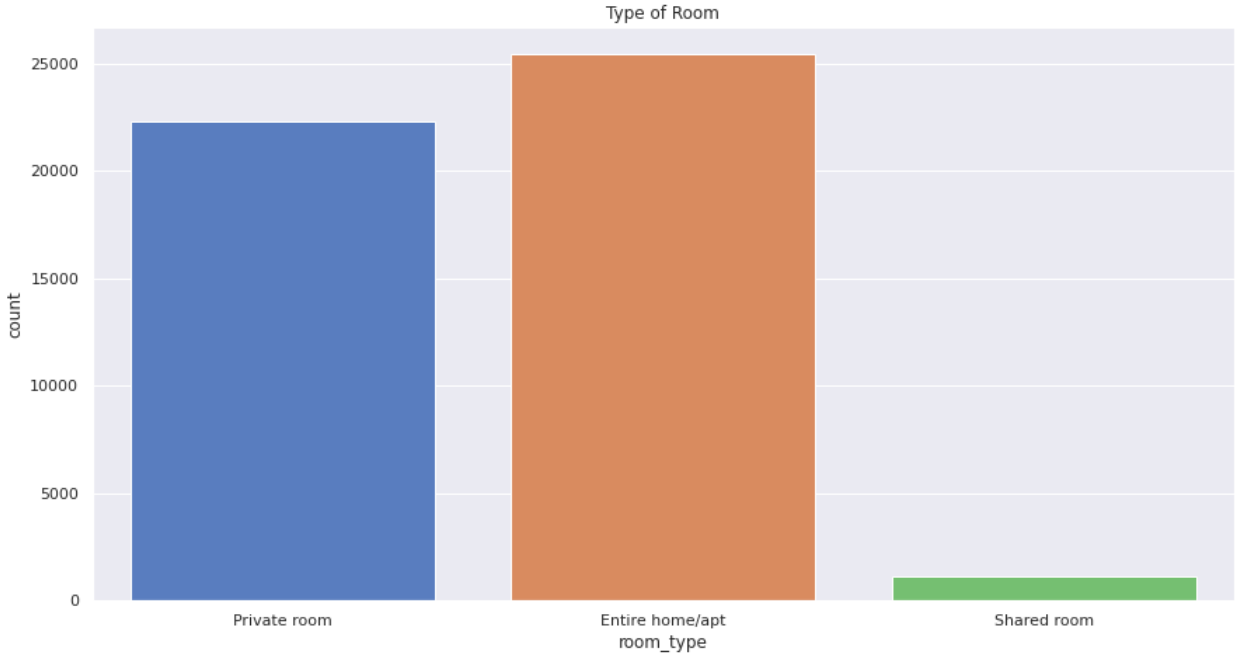
۷- پرطرفدارترین محله ها کدام هستند؟

بیشترین بازدیدهای اجاره ای بر حسب محله ها را بدست می آوریم. در اینجا تعداد بازدیدها بر اساس محله تفکیک شده است. بر این اساس محله Harlem بیشترین بازدید و Astoria کمترین بازدید را در بین سایر محله ها داشته است.




```
review = rb.sort_values('number_of_reviews', ascending=False)
top_reviewed = review.loc[:, ['neighbourhood', 'number_of_reviews']][:20]
top_reviewed = top_reviewed.groupby('neighbourhood').mean().sort_values(
    'number_of_reviews', ascending=False).reset_index()
fig4, ax3 = plt.subplots(figsize=(12, 8))
sns.barplot(x=top_reviewed['neighbourhood'],
            y=top_reviewed['number_of_reviews'].values, color='yellowgreen', ax=ax3)
plt.plot(top_reviewed['number_of_reviews'],
         marker='o', color='red', linestyle='--')
plt.ylabel('Reviews', fontsize='15')
plt.xlabel('Location', fontsize='15')
plt.ylim((400, 580))
for ax in ax3.get_xticklabels():
    ax.set_rotation(50)
plt.title('Most-Reviewed Rentals by location', fontsize='15')
plt.show()
sns.set()
```

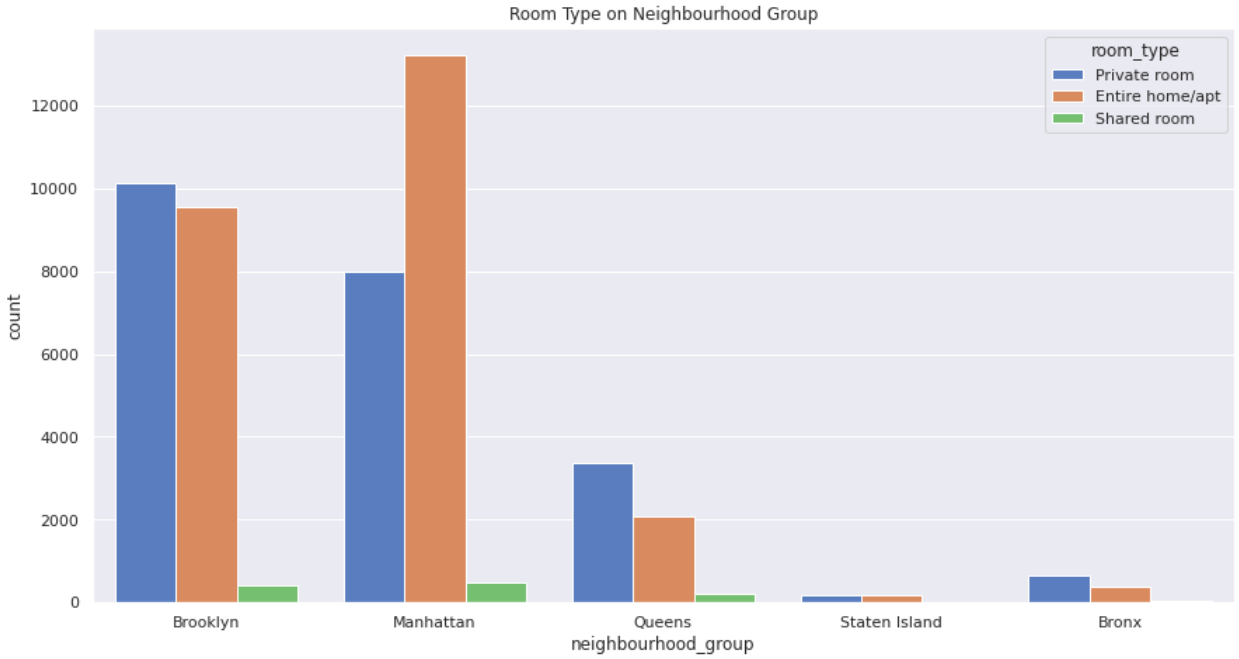
۸- تعداد اماکن اجاره‌ای را با هم مقایسه کنید.




```
plt.figure(figsize=(13,7))
plt.title("Type of rooms")
sns.countplot(rb.room_type,palette="muted")
fig=plt.gcf()
plt.show()
```

۹- فراوانی انواع اماکن اجاره ای در هر منطقه چگونه است؟

در اینجا فراوانی انواع اماکن اجاره ای (private room , Entire home/apt , Shared room) را در ۵ منطقه بررسی کرده ایم. بر این اساس در منطقه Brooklyn، مکان های Private room بیشترین اجاره و مکان های Shared room کمترین میزان اجاره را دارند. همچین این مقایسه برای سایر محله ها نیز انجام شده که در نمودار آمده است.




```
plt.figure(figsize=(13,7))
plt.title("Room Type on Neigbourhood Group")
sns.countplot(rb.neighbourhood_group,hue=rb.room_type,palette="muted")
plt.show()
```

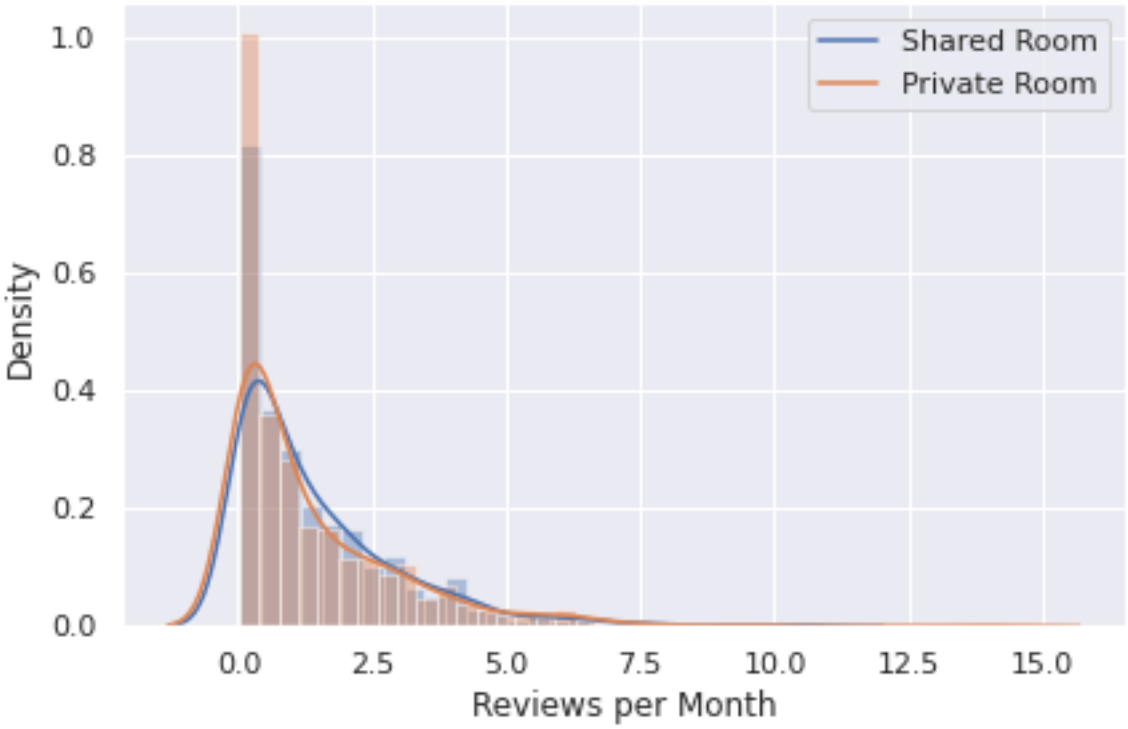
۱۰- آیا بین بازدیدهای ماهانه میزبان های "اتاقهای مشترک" و "اتاقهای خصوصی" تفاوت قابل توجهی وجود دارد؟

از آنجایی که فیلد "reviews_per_month " توزیع نرمال ندارد، از آزمون یو من ویتنی استفاده می کنیم که معادل غیر پارامتری (non parametric) آزمون 'Two Sample t test' است. بنابراین می توانیم بجای میانگین، بر روی میانه های (Median) تمرکز کنیم.

H0: تفاوت معناداری بین میانه های دو گروه وجود ندارد.

H1: بین میانه های دو گروه تفاوت معنی‌داری وجود دارد.

سطح معناداری: $\alpha = 0.05$




```
(607492.0, '0.00002580171171316850')
```

```
shared_rooms=rb.loc[rb.room_type ==
    'Shared room']['reviews_per_month'].values.tolist()
private_rooms=rb.loc[rb.room_type == 'Private room'].sample(len(
    shared_rooms), replace=False, random_state=1)['reviews_per_month'].values.tolist()

sns.distplot(shared_rooms)
sns.distplot(private_rooms)
plt.xlabel('Reviews per Month')
plt.legend(['Shared Room', 'Private Room'])
plt.show()
def mann_whitney_u_test(d1, d2):
    u_stat, p_val=ss.mannwhitneyu(d1, d2)
    return u_stat, f'{p_val:.20f}'
mann_whitney_u_test(shared_rooms, private_rooms)
```

مجموعه داده دوم: مجموعه داده مربوط به مسابقات فوتبال بین المللی

۱- فراخوانی پکیج ها و داده:

در ابتدا لازم است پکیج های لازم فراخوانی شوند. سپس فایل دیتاست که روی گوگل درایو ذخیره شده را در برنامه لود می کنیم تا پس از خواندن در متغیر rb ذخیره شود:

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

```
from google.colab import drive
drive.mount('/content/gdrive')
! ls '/content/gdrive/MyDrive/Data Mining/'
root='/content/gdrive/MyDrive/Data Mining/'
rb=pd.read_csv(root+"results.csv")
```

۲- آیا بین امتیاز تیم در زمین خانگی و تیم دیگر کوریلیشن یا همبستگی وجود دارد؟

```
rb.corr()
```

	home_score	away_score	neutral
home_score	1.000000	-0.135600	-0.031953
away_score	-0.135600	1.000000	0.084819
neutral	-0.031953	0.084819	1.000000

ضریب همبستگی (Correlation) عددی بین -۱ تا ۱ است. هر چه فاصله آن از صفر بیشتر باشد، روند دو پارامتر (هم جهت یا مخالف جهت بودن) نیز بیشتر می شود. مثبت و منفی بودن نیز جهت تغییرات را نشان می دهد. مطابق خروجی فوق، امتیاز یک تیم با امتیاز تیم دیگر دارای همبستگی ۰/۱۳۵- است. این بدین معناست که با افزایش امتیاز یک تیم، امتیاز تیم دیگر به میزان ۰/۱۳۵- کم می شود یا به عبارت دیگر افزایش تعداد گل های یک تیم، کاهش تعداد گل های تیم دیگر را دارد. از نظر شهودی نیز می توان این انتظار را داشت که هرچه تعداد گل های زده یک تیم بیشتر می شود، تیم دیگر معمولاً ضعیف تر است و تعداد گل های زده آن کمتر می شود.

۳- چه کشور (کشورهایی) در تمام زمان بهترین تیم بوده است؟

در اینجا انتخاب بهترین تیم ها بر این مبنا که آن کشور بیشترین برد را داشته است، صورت گرفته است.

برای این منظور، با توجه به نتیجه هر مسابقه، برنده آن را تعیین می کنیم. بر این اساس با شمردن تعداد برنده شدن هر کشور، می توانیم بهترین یا قویترین تیم را مشخص کنیم.

برای این کار، ستون جدیدی با عنوان winner اضافه می کنیم تا برنده هر بازی را بر اساس امتیاز (تعداد گل های آن بازی) تعیین می کنیم. بازی هایی که مساوی شده اند با عبارت non مشخص کرده ایم.

روش دیگری نیز برای این کار وجود دارد. می توان بعد از اضافه کردن ستون برنده، همین ستون گروه بندی کنیم. (بر اساس تعداد آن، بازی های برنده هر کشور مشخص می شود.)

```
def winner(row):
    if row['home_score'] > row['away_score']: return row['home_team']
    elif row['home_score'] < row['away_score']: return row['away_team']
    else: return 'non'
```

```
rb['winner'] = rb.apply(lambda row: winner(row), axis=1)
rb.head()
winners = pd.value_counts(rb.winner)
winners = winners.drop('non')
winners.head(20)
```

کد فوق، خروجی زیر را خواهد داشت:

```
Brazil          629
England         577
Germany         558
Argentina       529
Sweden          503
South Korea     455
Mexico          442
Hungary         439
Italy           428
France          423
Spain           409
Netherlands     402
Uruguay         384
Scotland        374
Denmark         361
Russia          358
Poland          355
Belgium         339
Austria         333
Zambia          330
```

کشور برزیل با ۶۲۹ بازی برده، انگلستان با ۵۷۷ بازی برده، آلمان با ۵۵۸ بازی برده و ... تیم های برتر هستند.

۴- چه کشور (کشورهایی) ضعیف ترین تیم بوده است؟

مشابه با روش سوال قبل، ستون جدیدی با عنوان losser یا بازنده ایجاد می کنیم و به همان ترتیب، بر حسب تعداد گل های کمتر هر بازی، بازنده آن را تعیین می کنیم.

```
def loser(row):
    if row['home_score'] < row['away_score']: return row['home_team']
    elif row['home_score'] > row['away_score']: return row['away_team']
    else: return 'non'

rb['loser'] = rb.apply(lambda row: loser(row), axis=1)
rb.head()
lossers = pd.value_counts(rb.loser)
lossers = lossers.drop('non')
lossers.head(20)
```

با اجرای دستورات فوق، خروجی زیر را خواهیم داشت:

```
Finland             403
```

```
Switzerland          348
Northern Ireland     336
Norway               330
Luxembourg           321
Chile                312
Wales                308
Sweden               299
Hungary              294
Austria              291
Uruguay              288
Malta                284
Singapore            280
Peru                 273
Paraguay             273
Belgium              272
Denmark              270
Poland               261
Bulgaria             259
Thailand             254
```

مطابق با این خروجی، کشور فنلاند با ۴۰۳ بازی باخته، سئیس با ۳۴۸ بازی باخته، ایرلند شمالی با ۳۳۶ بازی باخته و ... ضعیفترین تیم ها هستند.

۵- تیم های برنده چه تعداد گل زده داشته اند؟

ما باز هم از همان ستون برنده ای که برای سوالات قبل ایجاد کردیم، استفاده می کنیم. اینبار اندیس ردیف هایی که شامل تیم برنده هستند را در متغیری بنام goals قرار می دهیم. سپس در جدول اصلی، به ازای هر برنده، جمع تعداد گل های زده آن را جمع و حساب می کنیم.

```
goals = pd.Series(index=winners.index, dtype='int32')
for col in goals.index:
    goals[col] = rb[rb.home_team == col].home_score.sum() + rb[rb.away_team == col
].away_score.sum()
goals = goals.fillna(0).sort_values(ascending=False)
goals.head(20)
```

خروجی به صورت زیر خواهد بود که انگلستان با ۲۲۲۱ گل زده، برزیل با ۲۱۶۱ گل زده، آلمان با ۲۱۳۹ گل زده و ... بیشترین امتیاز را در بازی هایی که برنده بوده اند، کسب کرده اند.

```
England          2221
Brazil           2161
Germany          2139
Sweden           2025
Hungary          1905
Argentina        1836
Netherlands      1629
Mexico           1522
South Korea      1516
France           1512
Denmark          1430
Austria          1421
Spain            1414
```

```
Uruguay          1398
Italy            1387
Poland           1384
Belgium          1375
Scotland         1357
Norway           1222
Russia           1209
```

۶- بیشترین امتیاز (گل زده) در چه مسابقاتی بوده است؟

برای این منظور، جمع تعداد گل ها (امتیازات) همه مسابقات را محاسبه و در ستونی بنام total score ذخیره می کنیم.
در نهایت جدول را بر اساس همین ستون مرتب و چاپ می کنیم:

```
rb['total_score'] = rb.home_score + rb.away_score
re=rb.sort_values(by='total_score',ascending=False)
re.head(10)
```

خروجی این دستور به صورت زیر است که در آن تاریخ، نام تیم و .... بازی هایی که بیشترین امتیاز را داشته اند، مشاهده می شود.

date	home_team	away_team	home_score	away_score	tournament	city	country	neutral	year	loser	winner	total_score
2001-04-11	Australia	American Samoa	31	0	FIFA World Cup qualification	Coffs Harbour	Australia	False	2001	American Samoa	Australia	31
1971-09-13	Tahiti	Cook Islands	30	0	South Pacific Games	Papeete	French Polynesia	False	1971	Cook Islands	Tahiti	30
1979-08-30	Fiji	Kiribati	24	0	South Pacific Games	Nausori	Fiji	False	1979	Kiribati	Fiji	24
2006-11-24	Sápmi	Monaco	21	1	Viva World Cup	Hyères	France	True	2006	Monaco	Sápmi	22
2001-04-09	Australia	Tonga	22	0	FIFA World Cup qualification	Coffs Harbour	Australia	False	2001	Tonga	Australia	22
2005-03-11	Guam	North Korea	0	21	EAFF Championship	Taipei	Chinese Taipei	True	2005	Guam	North Korea	21
1987-12-15	American Samoa	Papua New Guinea	0	20	South Pacific Games	Nouméa	New Caledonia	True	1987	American Samoa	Papua New Guinea	20
2000-02-14	Kuwait	Bhutan	20	0	AFC Asian Cup qualification	Kuwait City	Kuwait	False	2000	Bhutan	Kuwait	20
2014-06-01	Darfur	Padania	0	20	CONIFA World Football Cup	Östersund	Sweden	True	2014	Darfur	Padania	20
2003-06-30	Sark	Isle of Wight	0	20	Island Games	St. Martin	Guernsey	True	2003	Sark	Isle of Wight	20

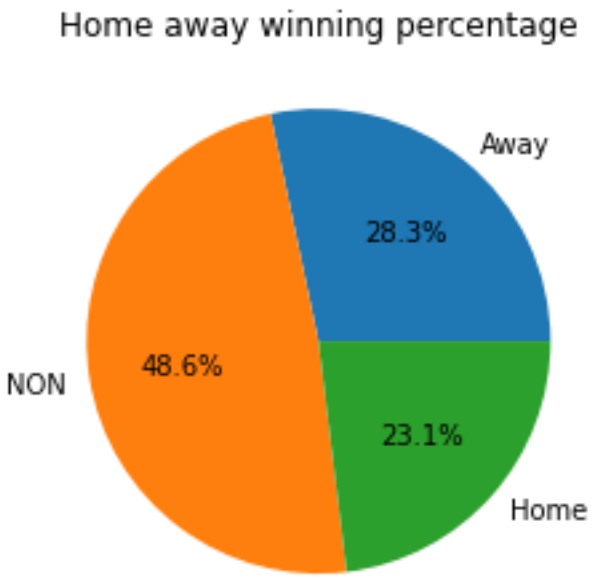
۷- میزبانی یک تورنمنت، چقدر در شانس یک کشور برای برنده شدن کمک می کند؟

برای بررسی این موضوع، سوابق برنده شدن تیم ها را در مسابقاتی که خود میزبان یا مهمان بوده اند، بررسی می کنیم.
وضعیت برنده هر مسابقه را از حیث اینکه میزبان بوده یا خیر ثبت می کنیم. در واقع تابعی تعریف می کنیم که مشخص می کند آیا برنده این بازی میزبان بوده (home) یا تیم دیگر (away) بوده و یا هیچ کدام (non) – برای بازی های مساوی یا بازی هایی که در زمین بی طرف انجام شده اند.
در ادامه درصد هرکدام از این وضعیت ها را away – home – non در نمودار دایره ای ترسیم می کنیم که نشاندهنده تعداد برنده هایی است که خود میزبان بوده اند یا خیر.

```
def find_homeaway(row):
    if row["home_team"] == row["winner"]:
        return "Home"
    elif row["away_team"] == row["winner"]:
        return "Away"
    else:
        return "non"
```

```
rb["winning_team2"] = rb.apply(find_homeaway,axis=1)
rb.head()
table = pd.pivot_table(rb,index="winning_team2",values="tournament",aggfunc="count
").reset_index()
P = table["tournament"].unique()
plt.title("Home away winning percentage")
plt.pie(P,
        explode = (0, 0, 0),
        labels=["Away","NON","Home"],
        autopct='%1.1f%%',
         )
plt.show()
```

خروجی به صورت شکل زیر است که نشان می دهد ۲۳/۱ درصد میزبانی ها، منجر به برنده شدن تیم شده است. ۲۸/۳ درصد تیم هایی که میزبان نبوده اند، در بازی ها برنده شده اند و ۴۸/۶ درصد بازی ها نیز یا مساوی شده اند یا برنده در زمین بی طرف بوده است. (برنده در زمین هیچ کدام از کشورهای دو طرف بازی نکرده است)



۸- در چه سال هایی، میانگین امتیازات بیشتر بوده است؟

ابتدا ستون تاریخ را تبدیل به سال می کنیم. (فقط بخش اول تاریخ که سال است را نگه می داریم)

سپس بر اساس تاریخ، گروه بندی می کنیم و همزمان میانگین ستون total_score را محاسبه می کنیم.

در پایان مرتب سازی را انجام داده و نتیجه را چاپ می کنیم:

```
rb['date'] = [x.split('-')[0] for x in rb['date'].tolist()]
goals_per_match = rb.groupby('date').total_score.mean()
goals_per_match=goals_per_match.sort_values(ascending=False)
```

```
goals_per_match.head(20)
```

دستورات فوق، خروجی زیر را دارد که در سال ۱۸۷۸ میانگین گل های زده شده ۹، سال ۱۸۸۲ میانگین گل های زده شده ۸، سال ۱۸۸۸ میانگین ۷.۱۴۲ گل زده شده و ....

```
1878    9.000000
1882    8.000000
1888    7.142857
1893    7.000000
1880    6.666667
1899    6.500000
1873    6.000000
1891    5.666667
1890    5.500000
1885    5.428571
1896    5.333333
1908    5.307692
1886    5.285714
1897    5.000000
1879    5.000000
1884    5.000000
1945    4.890625
1894    4.833333
1912    4.825000
1934    4.820755
```

۹- میانگین امتیازات در چه تورنمنت هایی بیشتر بوده است؟

برای این کار، بر اساس تورنمنت گروه بندی می کنیم و روی ستون total_score میانگین می گیریم و داده ها را مرتب می کنیم.

```
tour = pd.DataFrame(rb.groupby("tournament")["total_score"].mean()).reset_index().
sort_values(by="total_score",ascending=False).head(10)
tour
```

خروجی به صورت زیر خواهد بود که در تورنمنت Paciific Games با میانگین ۶/۰۹ امتیاز، بیشترین میانگین امتیاز را در بین تورنمنت های دیگر داشته است.

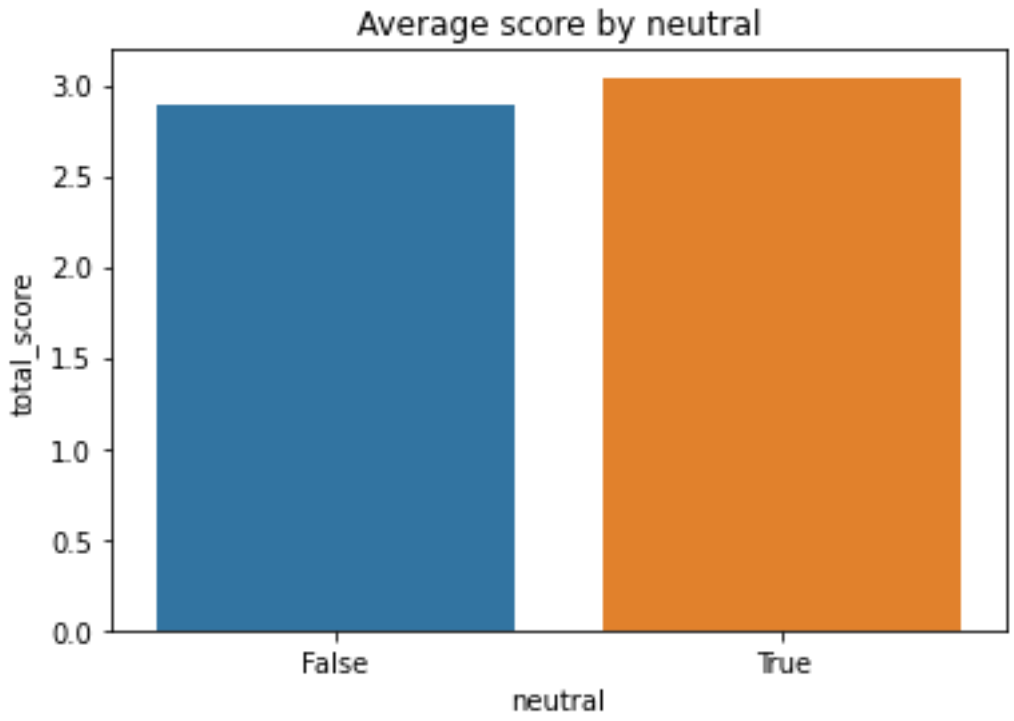
	tournament	total_score
86	Pacific Games	6.019608
13	Atlantic Heritage Cup	6.000000
93	South Pacific Games	5.912195
58	GaNEFo	5.785714
111	World Unity Cup	5.500000
107	Viva World Cup	4.821429
49	Dragon Cup	4.750000
63	Inter Games Football Tournament	4.684211
18	CCCF Championship	4.463415
46	Copa Roca	4.347826

۱۰- آیا امتیازات در مسابقاتی که در زمین های بیطرف برگزار می شود تفاوتی با سایر مسابقات دارد؟

برای بررسی این موضوع، داده ها را بر اساس ستون neutral گروه بندی می کنیم و ستون total_score آن ها را میانگین‌گیری می کنیم تا ببینیم میانگین امتیازات در مسابقات در زمین بیطرف چگونه بوده است.

```
n neutral = pd.DataFrame(rb.groupby("neutral")["total_score"].mean()).reset_index()

plt.title("Average score by neutral")
sns.barplot(data=neu,x="neutral",y="total_score")
plt.show()
```

خروجی به صورت نمودار زیر خواهد بود که نشان می دهد میانگین امتیاز در مسابقات در زمین بیطرف بیش از میانگین امتیاز در مسابقات بدون زمین بیطرف است.



۱۱- تحلیل آماری مسابقاتی که ایران در آن شرکت داشته و وضعیت میزبانی آن چگونه بوده است؟

در این تحلیل، مسابقاتی که ایران در آن شرکت داشته را بررسی می کنیم. در یک حالت، ایران هم میزبان بوده و هم شرکت کننده در مسابقه، و در حالت دیگر، ایران طرف دیگر بازی بوده است. در اینجا خواهیم دید که آیا مسابقاتی که ایران در آن شرکت داشته و میزبان بوده، با مسابقاتی که میزبان آن نبوده، چه تفاوتی داشته است.

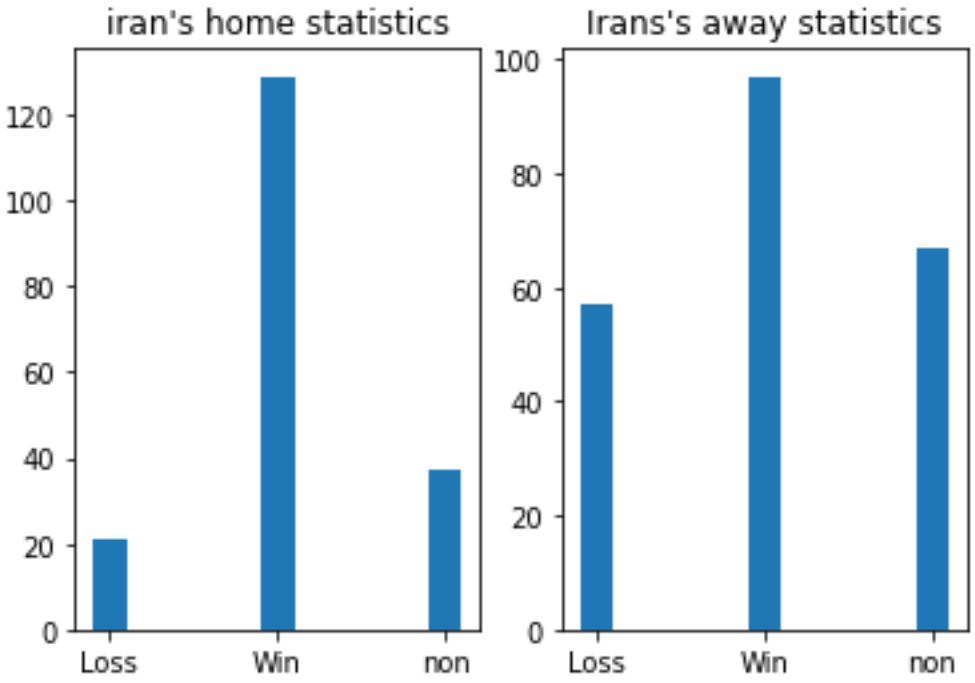
```
iranh=rb[(rb["home_team"]=="Iran") & (rb["country"]=="Iran")]
irana=rb[rb.away_team=='Iran']

irh_away=iranh[iranh.away_score>iranh.home_score]
irh_home=iranh[iranh.home_score>iranh.away_score]
irh_non=iranh[iranh.home_score==iranh.away_score]

ira_away=irana[irana.away_score>irana.home_score]
ira_home=irana[irana.home_score>irana.away_score]
ira_non=irana[irana.home_score==irana.away_score]

t1=np.array([irh_away.shape[0],irh_home.shape[0],irh_non.shape[0]])
t2=np.array([ira_home.shape[0],ira_away.shape[0],ira_non.shape[0]])

z=["Loss","Win","non"]
plt.subplot(1,2,1,)
plt.bar(z,t1, width=0.2)
plt.title("iran's home statistics")
plt.subplot(1,2,2)

plt.bar(z,t2,width=0.2)
```

```
plt.title("Irans's away statistics")
plt.show()
```

مطابق این آمار (نمودار سمت چپ)، در بازی هایی که ایران خودش میزبان بوده، در بیش از ۱۲۰ بازی برنده بوده نزدیک ۴۰ بازی مساوی و حدود ۲۰ بازی باخت داشته است.

اکثراً مسابقاتی که در زمین حریف باز کرده، (نزدیک ۱۰۰ بازی) را برده است. همچنین بیش از ۶۰ بازی در زمین حریف را مساوی و کمتر از ۶۰ بازی نیز باخته است.

برای ایران تفاوتی ندارد در چه زمینی بازی می کند و ایران چه در زمین حریف و چه در زمین خانگی با اختلاف بسیار زیاد برنده بوده و خواهد بود!



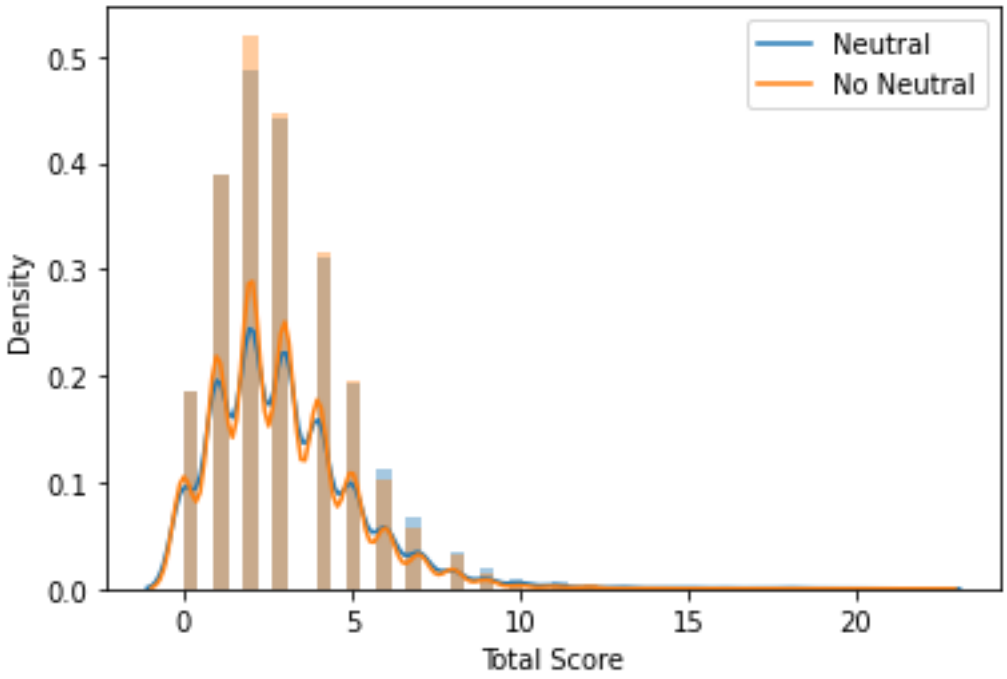
۱۲- آیا تعداد گل ها در مسابقاتی که در زمین بیطرف برگزار می شوند با بازی هایی که در زمین خانگی برگزار می شوند، یکسان است یا خیر؟

دو جامعه (مسابقاتی که در زمین بیطرف هستند و در زمین خانگی) متغیر مستقل هستند و مقدار آماره U را بر اساس این مشاهدات محاسبه می کنیم.

H0: تفاوت معناداری بین میانه های دو گروه وجود ندارد.

H1: بین میانه های دو گروه تفاوت معنی‌داری وجود دارد.

سطح معناداری: $\alpha = 0.05$




```
(52072638.0, '0.0028220727122658640 3')
```

از آنجایی که p-value کوچکتر از ۰/۰۵ است، پس فرض H0 رد و H1 پذیرفته می شود.

یعنی تفاوت معناداری بین دو گروه وجود دارد. به عبارت دیگر بازی هایی که در زمین خانگی برگزار می شوند امتیاز (گل های بیشتری) نسبت به بازی هایی که در زمین بی طرف برگزار می شوند، زده می شود و تفاوت معنا داری وجود دارد.