دانشگاه شهید بهشتی
دانشکده علوم ریاضی
گروه علوم کامپیوتر

تمرین های سری اول
درس داده‌کاوی

والا خسروی
۹۹۴۲۲۰۶۸

جناب آقای دکتر فراهانی
جناب آقای دکتر خردپیشه

اسفند ۱۳۹۹

مقدمه

کلیه شکل‌ها، جداول، محاسبات و آزمون‌ها بر روی داده‌ها با استفاده از نرم افزار R به دست آماده است و دستورها به ترتیب در فایل script.r قرار دارد و خروجی‌ها و تحلیل‌های لازم در همین فایل قرار گرفته است.

تمرین اول

معرفی

مجموعه داده اول مربوط به داده‌های جمع آوری شده از خانه‌های اجارهای برای اقامت کوتاه مدت (Airbnb) در شهر نیویورک آمریکا است و در آن اطلاعاتی در مورد میزبان‌ها، میهمان‌ها، مکان اقامتگاه‌ها، زمان و مدت اجاره، قیمت اجاره و ... وجود دارد. که این داده‌ها شامل: قیمت، محله، طول و عرض جغرافیایی و اطلاعات دیگر از موارد اجاره شده از سال ۲۰۰۸ الی ۲۰۱۹ می باشد.

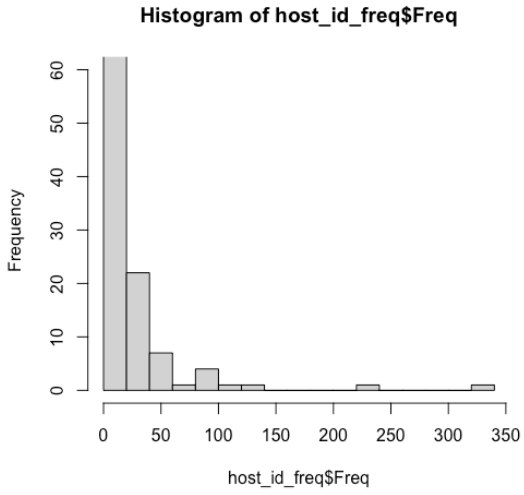
سوال اول

تعداد داده‌های موجود ۴۸۸۹۵ ردیف می‌باشد که مختص به ۳۷۴۵۷ میزبان منحصر به فرد می‌باشد. این موضوع نشانه دهنده این است که از سال ۲۰۰۸ الی ۲۰۱۹ مواردی بیش از یکبار خانه خود را اجاره دادند. حال به محاسبه تعداد دفعاتی که هر میزبان خانه خود را اجاره داده است می‌پردازیم.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	1.000	1.305	1.000	327.000

جدول ۱.۱

و هیستوگرام آن را رسم می‌کنیم



شکل ۱.۱.۱

با توجه به جدول ۱.۱ و شکل ۱.۱ می توان متوجه شد عده زیادی فقط یکبار خانه خود را برای اجاره گذاشته اند. حال این موضوع را آزمون می‌کنیم

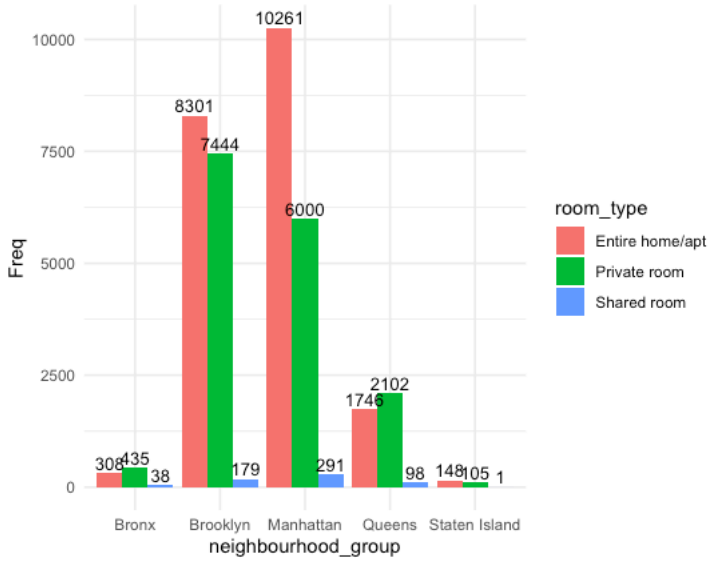
$$\begin{cases} H_0 : \mu > 2 \\ H_1 : \mu < 2 \end{cases}$$

```
> t.test(host_id_freq$Freq, mu=2, alternative="less")

        One Sample t-test

data:  host_id_freq$Freq
t = -48.696, df = 37456, p-value < 2.2e-16
alternative hypothesis: true mean is less than 2
95 percent confidence interval:
     -Inf 1.328827
sample estimates:
mean of x
 1.305363
```

آزمون ۱.۱.۱

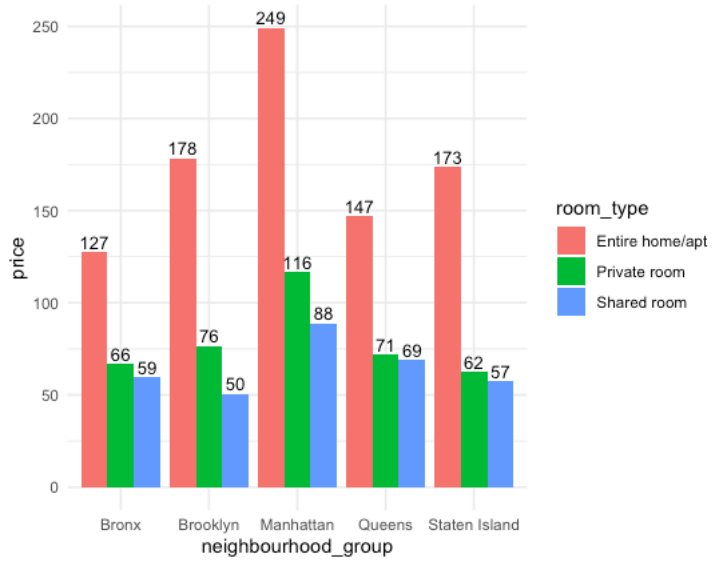
نتیجه این آزمون نشان دهنده این است که با احتمال قوی فرض صفر رد می‌شود که به این معنی است میانگین تعداد دفعات اجاره هر خانه از ۲ بار کمتر است.

در ادامه ابتدا از مجموعه داده host_id های تکراری را خذف کرده و تعداد خانه های موجود در هر منطقه محاسبه می کنیم و آن‌ها را بر اساس نوع خانه (خانه کامل، اتاق اختصاصی و اتاق اشتراکی) و محله دسته‌بندی می‌کنیم و نمودار میله‌ای تعداد خانه‌ها را رسم می‌کنیم.



شکل ۲.۱.۱

همچنین می‌توانیم این داده را بر اساس نوع خانه و محله دسته‌بندی کرده و نمودار میله‌ای قیمت خانه‌ها را رسم می‌کنیم.

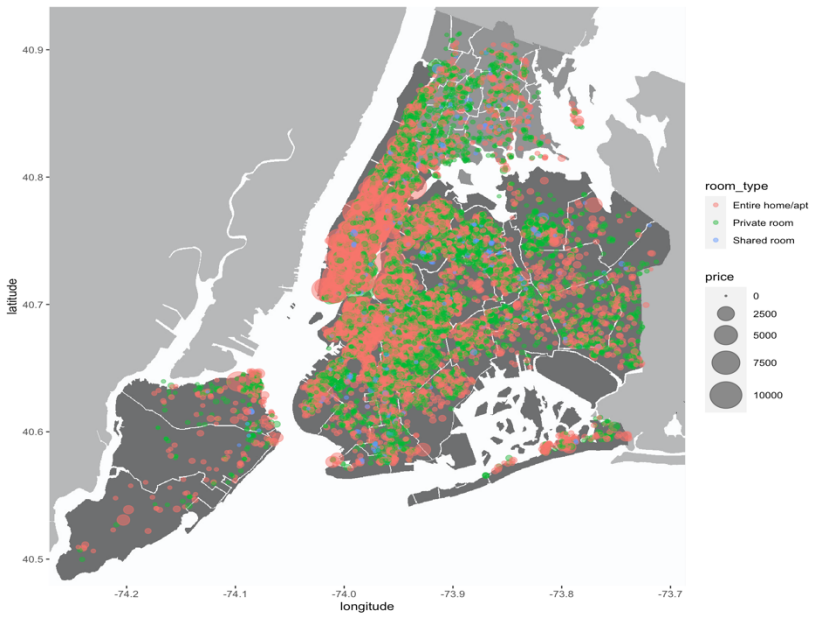


شکل ۳.۱.۱

با توجه به این دو نمودار نتیجه می‌گیریم که گران‌ترین محله Manhattan است و تعداد خانه‌های برای اجاره در آن بیشتر از بقیه محله‌ها است.

در نمودار بعدی با استفاده از bubblechart خانه‌ها را روی نقشه نمایش می‌دهیم اندازه حباب‌ها نشان دهنده قیمت و رنگ آن‌ها نشان دهنده نوع خانه (خانه کامل، اتاق اختصاصی و اتاق اشتراکی) است.

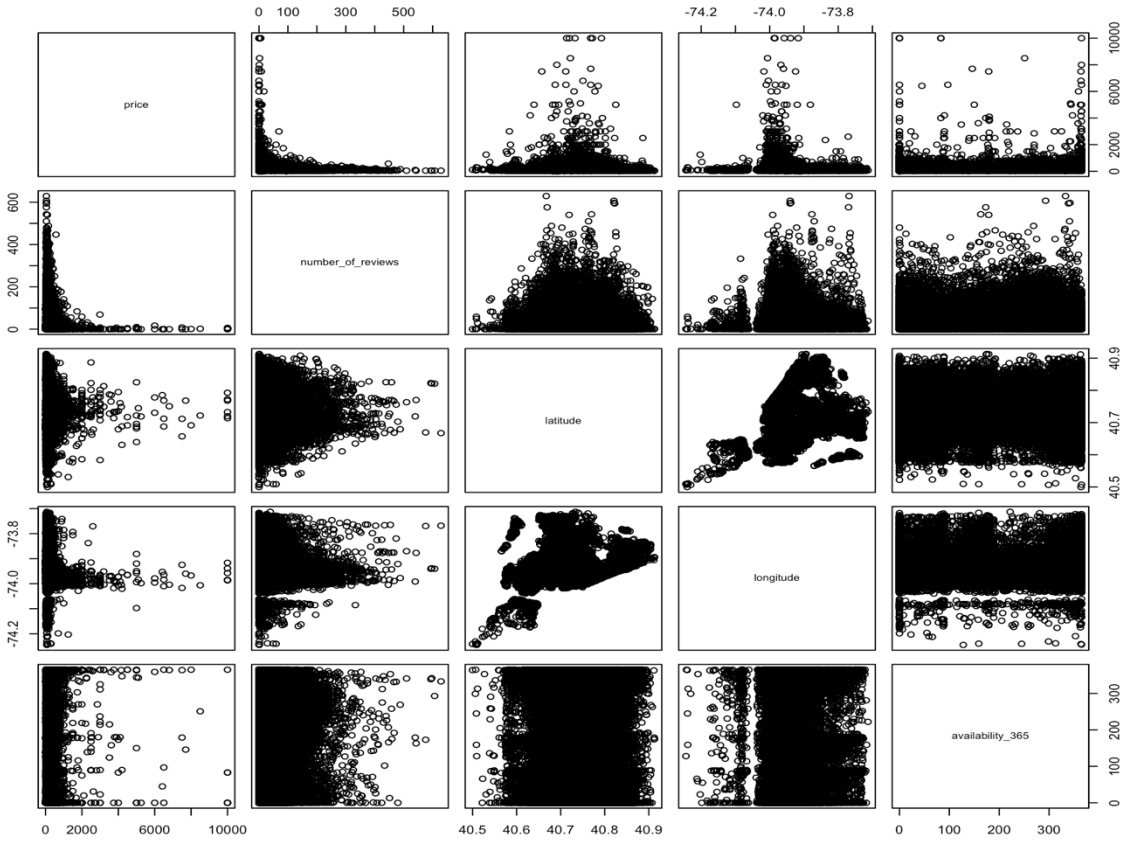
همچنین چون این نمودار روی نقشه رسم شده است به صورت تقریبی محله هر خانه را می‌توانیم تشخیص بدهیم.



شکل ۴.۱.۱

سوال دوم

در این سوال پنج عامل قیمت، تعداد نظرات، طول جغرافیایی، عرض جغرافیایی و دسترسی در ۳۶۵ روز را دو به دو بررسی می‌کنیم و نمودار نقطه‌ای آن را رسم می کنیم.



شکل ۱.۲.۱

متغییر پاسخ را قیمت در نظر می‌گیریم، حال به کمک آزمون ANOVA تاثیر این عوامل را بر قیمت بررسی می‌کنیم.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
number_of_reviews	1	6.485e+06	6484696	116.73	< 2e-16	***
latitude	1	3.109e+06	3109091	55.97	7.5e-14	***
longitude	1	6.417e+07	64170058	1155.08	< 2e-16	***
availability_365	1	3.009e+07	30090076	541.63	< 2e-16	***
Residuals	48890	2.716e+09	55555			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

جدول ۱.۲.۱

با توجه به مقادیر کم p-value هر چهار عامل بر روی قیمت تاثیرگذار اند پس مدل ما به شکل زیر است.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

حال ضرایب بتا را محاسبه می‌کنیم

```
Call:
lm(formula = price ~ number_of_reviews + latitude + longitude +
    availability_365, data = data)

Coefficients:
      (Intercept)  number_of_reviews           latitude          longitude
       -6.979e+04         -3.013e-01          2.103e+02         -8.297e+02
 availability_365
        1.919e-01
```

جدول ۲.۲.۱

سوال سوم

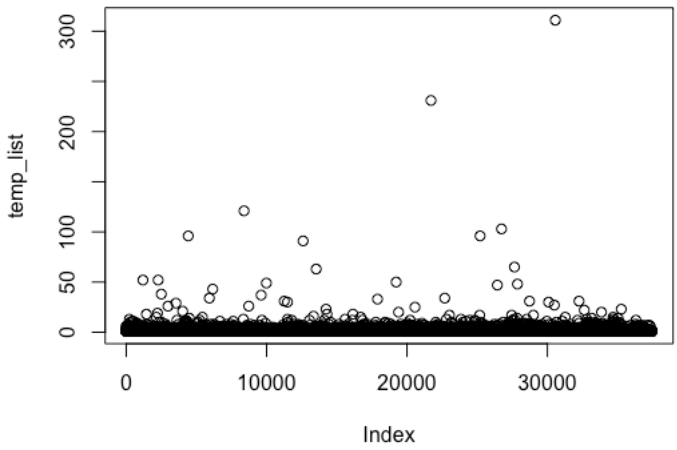
برای پیدا کردن مشغول‌ترین میزبان کافی است تعداد دفعاتی که هر میزبان خانه خود را اجاره داده را محاسبه کنیم و آن را بر اساس فراوانی مرتب می کنیم

```
> host_id_freq <- host_id_freq[order(-host_id_freq$Freq),]
> host_id_freq
           Var1 Freq
34647 219517861  327
29408 107434423  232
19575  30283594  121
31080 137358866  103
12807  12243051   96
14437  16098958   96
```

جدول ۱.۳.۱

با توجه جدول بالا میزبان با شناسه 219517861 بیشترین تعداد اجاره را از سال ۲۰۰۸ الی ۲۰۱۹ داشته است. حال تمام اطلاعات میزبان با این شناسه را جدا می‌کنیم و به تحلیل آن می پردازیم.

این فرد با نام soner ، ۳۱۱ واحد خانه متمایز برای اجاره در ثبت کرده است. این رقم به نظر رقم بالایی می‌آید پس آن را تعداد خانه‌هایی که هر میزبان دیگر اجاره داده است مقایسه می‌کنیم.



شکل ۱.۳.۱

با توجه به شکل بالا عدد ۳۱۱ با اختلاف از بیشتر داده‌ها فاصله دارد و میانگین خانه‌های هر میزبان ۱.۲۹ است. پس می‌تواند این موضوع را دلیل مشغول بودن این میزبان دانست.
همچنین مساله قیمت را می‌توان بررسی کرد که در ادامه به آن می‌پردازیم.

شکل ۲.۳.۱

همان طور که در شکل مشاهده می‌کنیم قیمت‌های خانه‌هایی که soner اجاره داده است بیشتر از باقی خانه‌ها است. این موضوع را مورد آزمون قرار می‌دهیم.

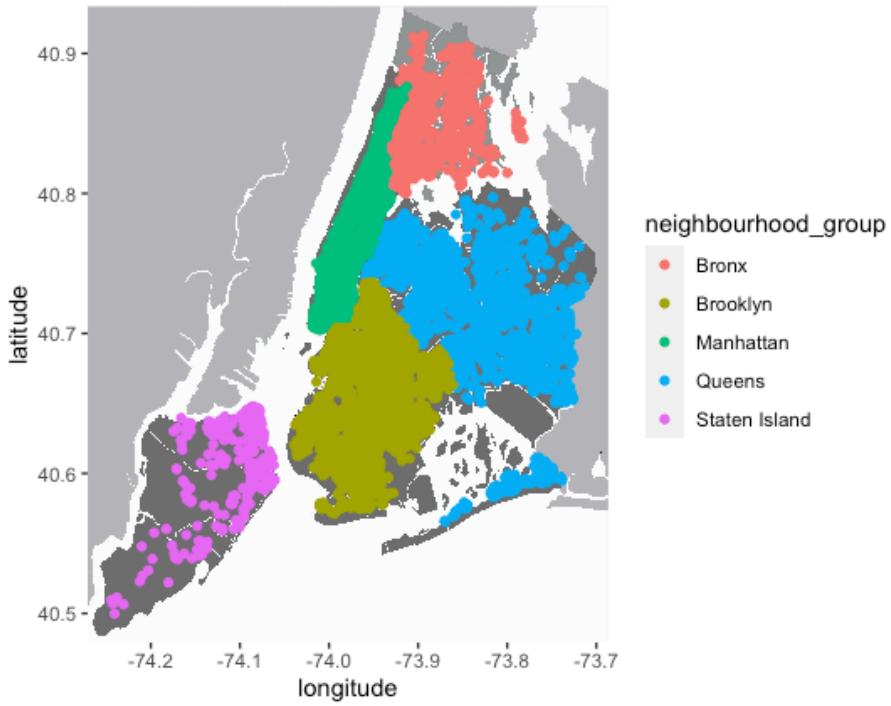
```
        Welch Two Sample t-test

data:  most_frequent_host$price and others$price
t = 15.091, df = 344.01, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 90.09679      Inf
sample estimates:
mean of x mean of y
 253.1957  152.0442
```

جدول ۲.۳.۱

با توجه به مقدار p-value متوجه می‌شویم قیمت‌های خانه‌هایی که soner اجاره داده است بیشتر از باقی خانه‌ها است. پس قیمت را نمی‌توان به عنوان یک عامل مثبت در نظر گرفت.

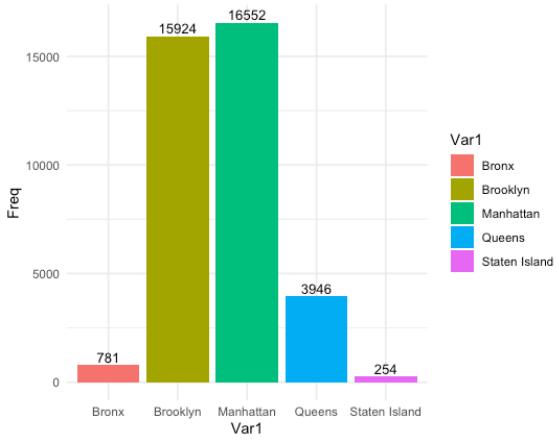
سوال چهارم

برای مقایسه محله‌های مختلف از نظر ترافیک (به منظور میزان اجاره شدن خانه‌ها) ابتدا تعداد خانه‌های اجاره شده در هر محله را بدست می‌آوریم.



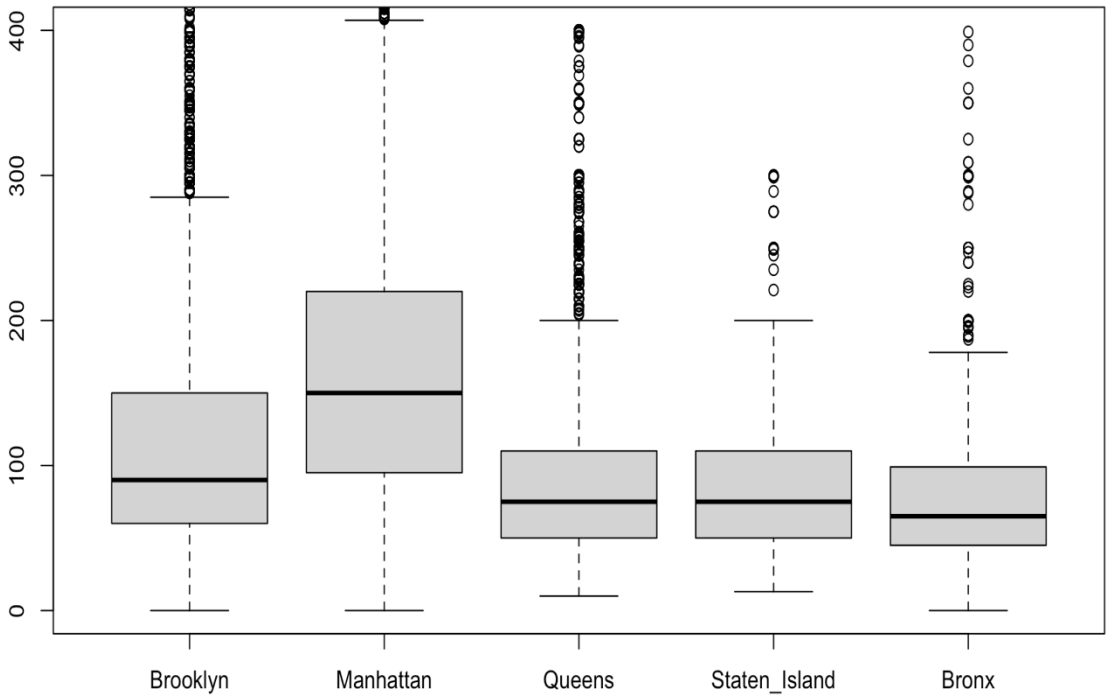
شکل ۱.۴.۱

با توجه به شکل بالا محله Manhattan و Brooklyn بالا ترین میزان ترافیک را دارند که در ادامه نمودار میله‌ای آن را رسم می‌کنیم.



شکل ۲.۴.۱

نمودار جعبه‌ای هر قیمت در هر محله را رسم می‌کنیم



شکل ۳.۴.۱

همانطور که مشخص است میانگین قیمت در محله‌ها با هم تفاوت دارد. برای این مجموعه داده برابری واریانس‌ها را به کمک آزمون بارتلت بررسی می‌کنیم.

```
        Bartlett test of homogeneity of variances

data:  price by neighbourhood_group
Bartlett's K-squared = 6081.3, df = 4, p-value < 2.2e-16
```

با توجه به مقدار p-value فرض صفر که برابری واریانس‌ها است رد می‌شود.

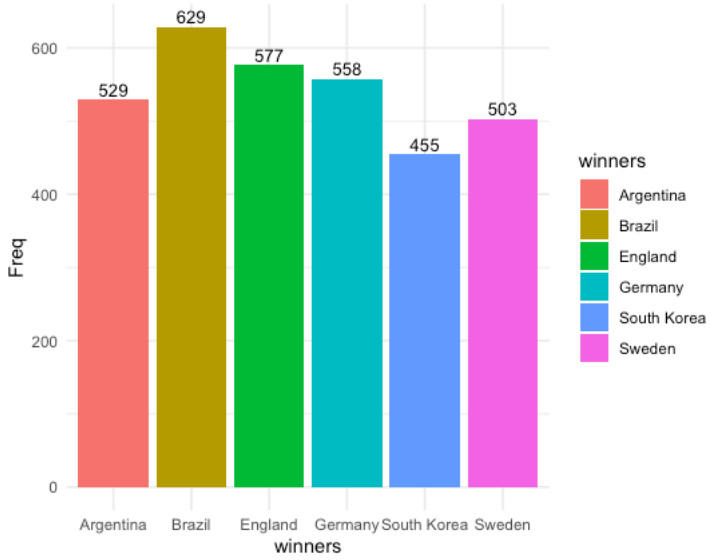
سوال دوم

معرفی

این مجموعه داده مربوط به مسابقات فوتبال بین‌المللی است که در قابل مسابقات جام‌جهانی، جام‌های قاره‌ای، تورنومنت‌ها، بازی‌های دوستانه و ... انجام شده است .به این منظور اطلاعات مختلفی از قبیل، نام تیم‌ها، محل انجام مسابقه، میزبان و میهمان، زمان بازی، تعداد گل‌ها، نتیجه بازی و ... ذخیره شده‌اند.

سوال اول

اگر تنها تعداد بردها را برای انتخاب بهترین تیم در نظر بگیریم تیم **برزیل** با ۶۲۹ برد بهترین تیم است



شکل ۱.۱.۲

سوال دوم

برای پاسخ به این سوال نیاز به پیش پردازش داده‌ها می باشد. از ستون تاریخ، سال را جدا کرده و در ستون جدید ذخیره کردیم. سپس در تمام مسابقات برگزار شده برنده را تشخیص داده و آن را در ستون جدیدی ذخیره کردیم. سپس برای تشخیص اسطوره هر دوره بازه های ۱۰ ساله را جدا کرده و تعداد بردهای هر تیم

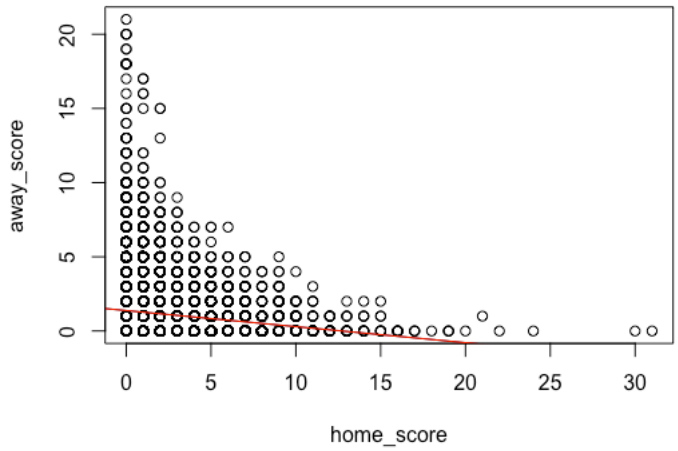
را حساب کردیم، تیمی که بیشترین تعداد برد را آن ۱۰ سال داشت را به عنوان اسطوره بر می گزینیم. حاصل به صورت جدول زیر بدست می آید. البته این جدول با احتساب بازی های دوستانه نیز می باشد.

1	Scotland	8	1870
2	Scotland	23	1880
3	England	24	1890
4	England	25	1900
5	Argentina	33	1910
6	Sweden	41	1920
7	Germany	55	1930
8	Argentina	43	1940
9	Hungary	60	1950
10	Brazil	77	1960
11	South Korea	102	1970
12	South Korea	73	1980
13	Brazil	105	1990
14	Saudi Arabia	105	2000
15	Mexico	97	2010
16	France	6	2020

جدول ۱.۲.۲

سوال سوم

در سوال رابطه بین گل‌های تیم میزبان و تیم مهمان را بررسی کردیم. در شکل زیر نمودار نقطه‌ای و خط رگررسیونی آن را رسم می‌کردیم.



جدول ۱.۲.۳

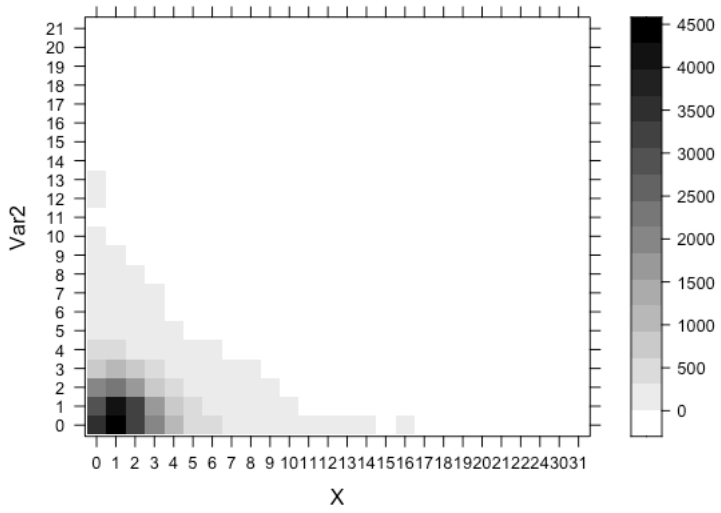
با آزمون ANOVA وجود رابطه بین این دو متغییر را بررسی می‌کنیم.

```
              Df Sum Sq Mean Sq F value Pr(>F)
home_score     1   1516  1515.7   784.4 <2e-16 ***
Residuals  41874  80918     1.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

مقدار p-value نشان دهنده آن است که رابطه وجود دارد و رابطه خطی آن به صورت زیر است.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.375893   0.009585  143.54   <2e-16
home_score  -0.108577   0.003877  -28.01   <2e-16
```

همچنین این مقادیر به این موضوع اشاره دارد که شانس تعداد گل بیشتر (یعنی برد) برای میزبان بیشتر است. heatmap داده‌ها به شکل زیر است.



شکل ۲.۲.۲

سوال چهارم

تعداد دفعاتی که همه تیم‌ها باهم بازی کردند را محاسبه می‌کنیم بر اساس بیشترین تعداد بازی آن را مرتب می‌کنیم ۱۱ تیم ابتدا آن را جدا می‌کنیم.

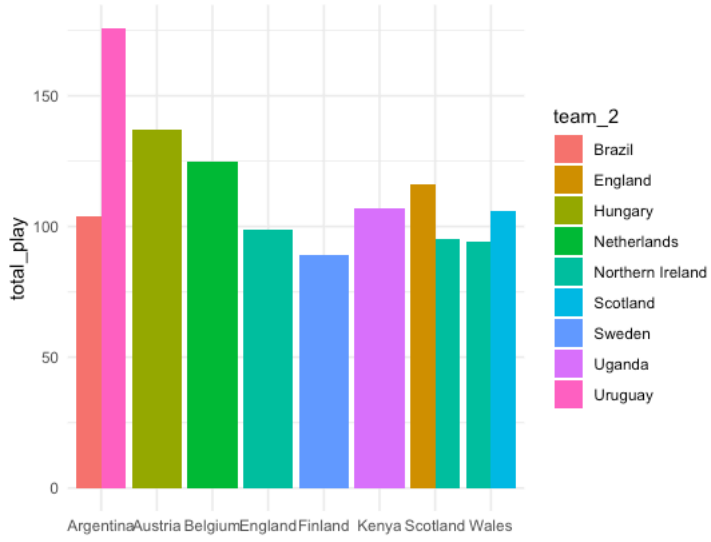
چالش اصلی برای حل این سوال ارائه اگوریتمی بود که بتواند تیم‌ها را به صورت زوجی از دیتافریم جدا کند چون داده‌ها به صورت تیم میزبان و تیم مهمان ذخیره شده اند. در جدول زیر مشکل مطرح شده واضح است.

87794	Argentina	Uruguay	96
4293	Uruguay	Argentina	80

که بعد از اعمال الگورتیم به شکل زیر تبدیل می‌شود

	team_1	team_2	total_play
1	Argentina	Uruguay	176

نمودار داده‌های بدست آماده به شکل زیر است.

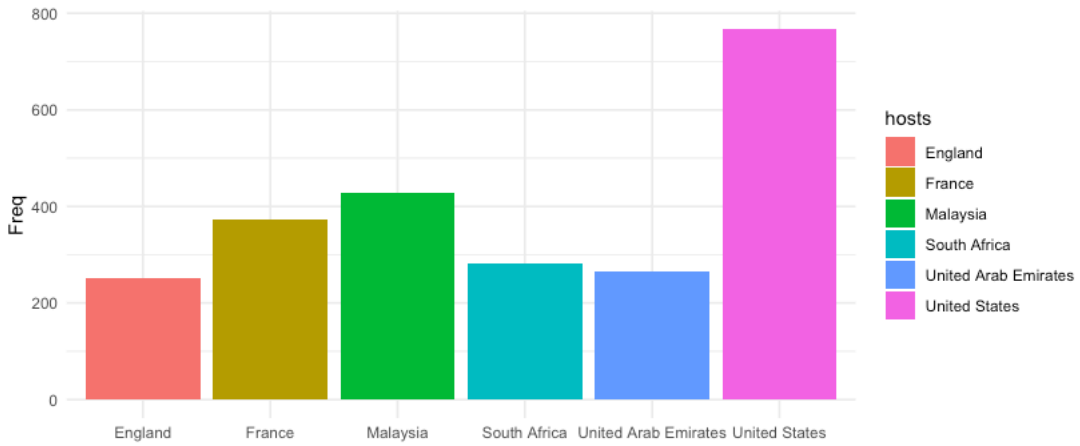


شکل ۱.۳.۲

این شکل نشان دهنده این است که بازی‌های پر تکرار درون قاره‌ای رخ می‌دهد.

سوال پنجم

نمودار کشورهایی که بیشترین میزبانی مسابقاتی را داشتند که در آن شرکت نداشتند به شکل زیر است.



جدول ۱.۵.۲

سوال ششم

مسابقات جام جهانی را از دیتافریم جدا می‌کنیم احتمال برد تیم‌های میزبان را محاسبه می‌کنیم و جدول آن به شکل زیر است.

	host	count	total	prob
1	Uruguay	4	4	1.0000000
2	Italy	10	10	1.0000000
3	France	7	8	0.8750000
4	Brazil	7	10	0.7000000
5	Switzerland	2	4	0.5000000
6	Sweden	4	5	0.8000000
7	Chile	4	6	0.6666667
8	England	5	5	1.0000000
9	Mexico	5	6	0.8333333
10	Germany	11	13	0.8461538
11	Argentina	5	6	0.8333333
12	Spain	1	3	0.3333333
13	United States	1	3	0.3333333
14	South Korea	3	5	0.6000000
15	Japan	2	3	0.6666667
16	South Africa	1	2	0.5000000
17	Russia	2	3	0.6666667

جدول ۱.۶.۲

با توجه به این جدول و تحلیل‌های قبلی میزبان بودن در تورنومنت‌ها شانس برد را افزایش می‌دهد. این موضوع را مورد آزمون قرار می‌دهیم.

```
> t.test(probs$prob, mu=.5, alternative="greater")

        One Sample t-test

data:  probs$prob
t = 4.1611, df = 16, p-value = 0.000368
alternative hypothesis: true mean is greater than 0.5
95 percent confidence interval:
 0.6247742       Inf
sample estimates:
mean of x
0.7149698
```

نتیجه این آزمون این فرض را تایید می‌کند.

سوال هفتم

بازی‌های دوستانه را جدا می‌کنیم و تعداد تکرار تیم‌ها را می‌شماریم.

91	Germany	296
84	France	293
148	Mexico	291
107	Hungary	276
15	Austria	254
218	Sweden	252
24	Belgium	245
182	Poland	245
159	Netherlands	243
219	Switzerland	240
114	Italy	232
32	Brazil	228
236	United States	217

۱۰ تیم اول این لیست عضو تیم‌های قدرتمند فوتبال هستتند پس تعداد بازی‌های دوستانه تاثیرگذار است.