گزارش تمرین ۱.۱ داده‌کاوی

علیرضا آزادبخت ۹۹۴۲۲۰۱۹

مجموعه داده New York City Airbnb Open Data:

مقدمه:

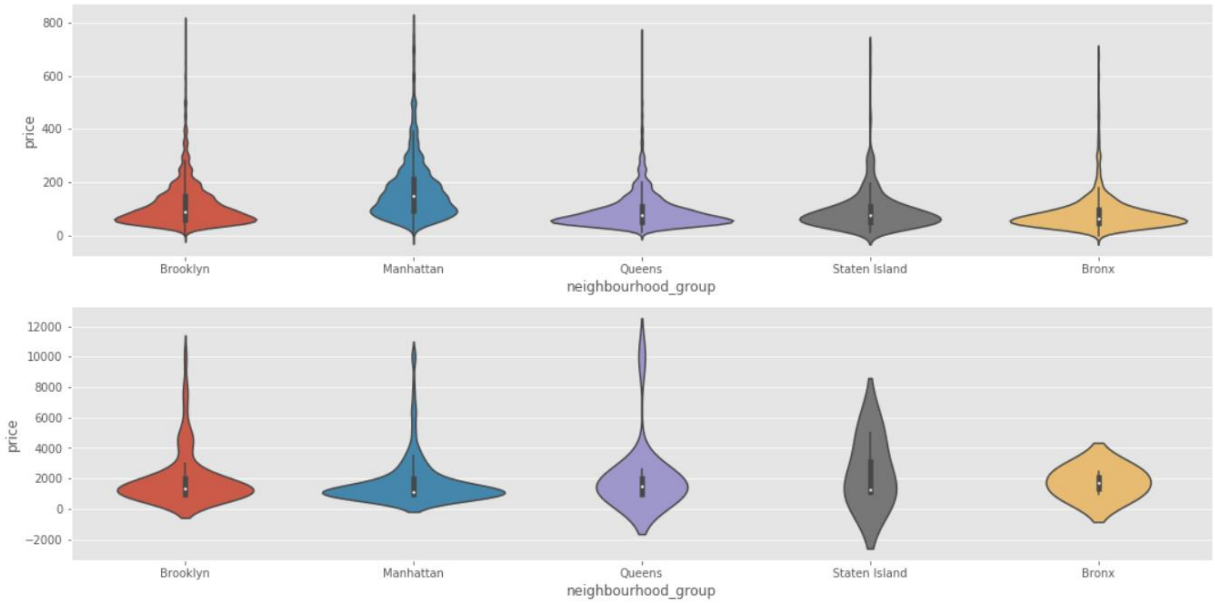
دیتای اولیه شرکت ایر بی اند بی شامل دیتای افرادی در محله های مختلف نیویورک است که خانه های خود را به افراد بومی نیویورک یا مسافران یا توریست ها یا افرادی که برای کار به نیویورک می‌آیند و کار بازه ای و فصلی انجام می‌دهند، برای اقامت خانه این افراد را اجاره میکنند ابتدا مقداری تحلیل های معمولی و بیسیک رو بررسی میکنیم سپس بدنبال ارتباط های پنهان م مخفی توی دیتا ها و روابط مختلف اون ها میپردازیم و نتایجی که تونستیم از دیتا ها بگیریم رو نشون میدیم

این پروژه یکی از پروژه های درس دیگری که در دوره کارشناسی داشته ام می‌باشد و نوتبوک پروژه شامل تحلیل های بیشتری با داکیومنت درون نوتبوک می‌باشد اما در این گزارش تنها در حد مورد نیاز تحلیل ها آورده شده اند.

سوالات:

- **What can we learn about different hosts and areas?**

در مرحله اول که شروع به بررسی مقادیر فیچر ها کردیم متوجه چند اسم میزبان نا مشخص شدیم، ایده این بود که شاید یک هاست توی یک رکورد اسمش ثبت نشده باشد و بتوانیم از سطر های دیگر این مقادیر رو پر کنیم ولی وقتی چک کردیم متوجه شدیم که اینطوری نیست و ساختار دیتا بیسشون احتمال زیاد رابطه ای هست و اگر مقداری نال ثبت شده باشه کلا توی هیچ تیبلی نمیشه پیداش کرد و حدس می‌زنیم که این افراد یا یک شرکت کوچک هستند که در کار اجاره خانه قبل از شروع ایر بی اند بی بوده اند و اسم رسمی نداشته اند و موقعی که به سیستم اضافه شدن اسمی ثبت نکردن یا شاید از اولین هاست ها بودن و سیستم ولیدیشن دیتا نداشته اون موقع و ما نمیتونیم از هویت حقیقی آن ها باخبر شویم.

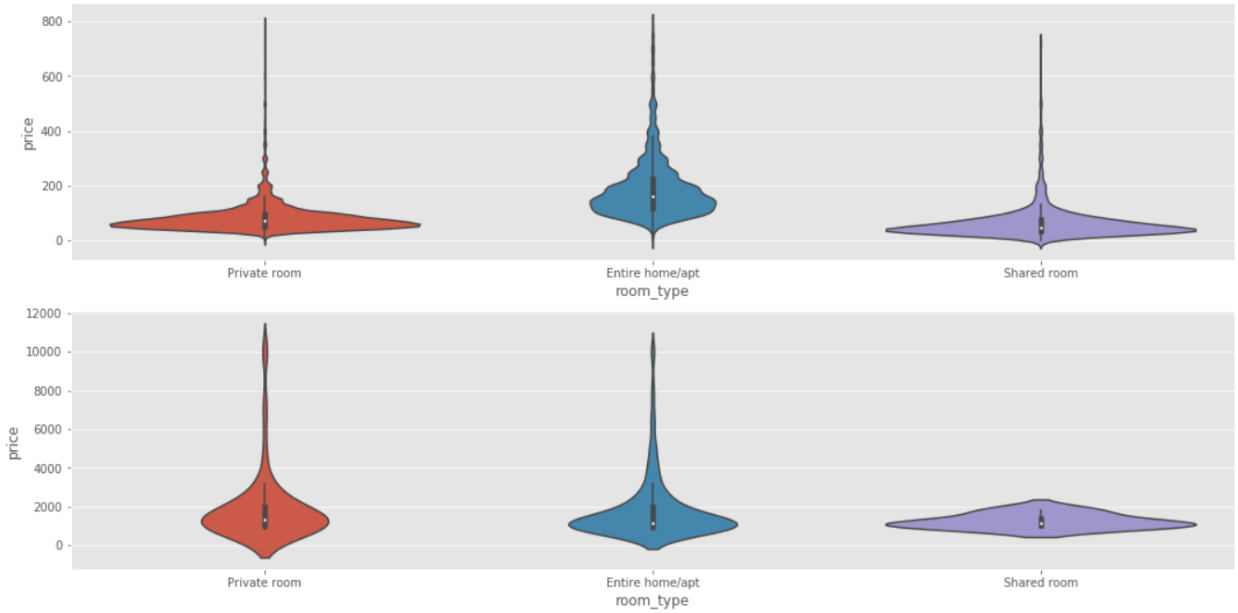


همین طور که از نمودار ها میشه فهمید نمودار بالا برای خونه های زیر ۸۰۰ دلار هست و نمودار پایینی برای نمودار های بالای ۸۰۰ دلار و میشه از از توزیع قیمتا نتیجه گرفت که منطقه کویین ارزون تر هست و منهتن گرون ترین منطقه هست ولی نکته جالب اینجاست درسته منهتن توی کل گرون تر هست از بقیه ولی استاتن آیلند خونه های گرون بیشتری داره ولی به طور کلی خیلی منطقه گرونی نیست ولی رویال تر است و بقیه مناطق رابطه نزدیک به هم دارن.

همون نمودار بالارو برای هر محله هم انجام دادیم ولی خب این دیتا بدرد ما نمیخوره چون ما شناختی با مناطق کوچیک نیویورک نداریم که بخوایم ازش مفهوم بکشیم بیرون ولی میتونیم این دیتا رو نگه داریم و اگر خواستیم اپلای کنیم و خونه بخریم ازش استفاده کنیم (((: ولی حالا میتونیم همینطوری تحلیل کلی داشته باشیم ک وقتی منطقه کوچیک میشه نوع بافت اون منطقه یکسان و یکدست تر میشه ینی اگر ما بگیم که یک منطقه فقیر نشین هست خیلی فرق داره تا بگیم یک محله فقیر نشین هست بخاطر همین اگر من بگم چون میانگین قیمت خونه های اجاره ای توی منهتن مثلا از بقیه بالاتره یا بگم یک محله کوچیک توی استاتن ایلند گرون هست جامعیت کمتری داره ینی به زبان دیگه اگر نتیجه گیری برای سطح مادی و فرهنگی بخوایم بگیریم اگر در مورد محله ها صحبت کنیم صحبتمون جامعیت بیشتری خواهد داشت پس ینی اگر توی نمودارای بالا بگیم فلان منطقه منطقه فقیر نشینی هست میتونیم نتیجه مون رو جامع بدونیم ایده ای که اینجا داشتیم این بود که بیایم محله های فقیر نشین رو با درصد جرم که توی اون محله اتفاق میافته مقایسه کنیم و اثر فقر روی جرم جنایت رو بررسی کنیم ولی به دو دلیل این کارو نکردیم یک اینکه دیتا رو از کجا بیاریم؟ (((: دو اینکه اکثر محله هایی ک خونه اجاره میدن مناطق توریستی و کاری هستن و بنظرم میاد که اگر خونه ای توی ی منطقه

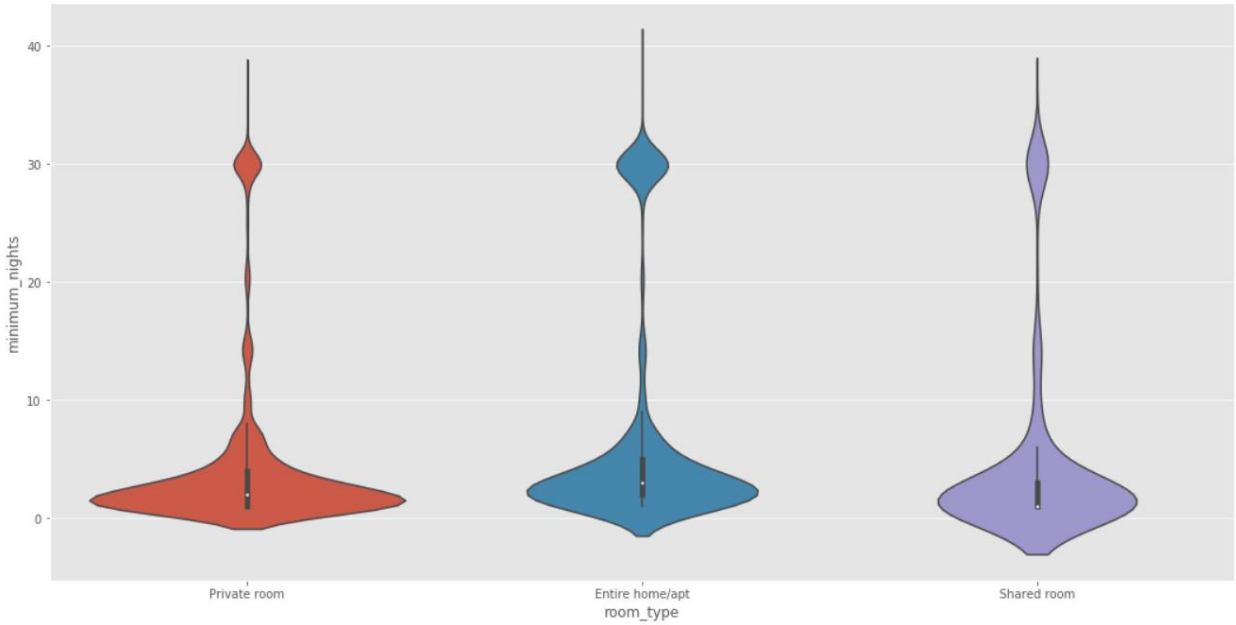
فقیر نشین باشه اگر برای اجاره قرار داده بشه کسی طرفش نمیره و کم کم توی سایت نمایشش کمتر میشه (ریویو کمتر میگیره و از این حرفا) پس چون اجاره ای هستن قیمتامون اگر نتیجه ای از روی این دیتا بگیریم اشتباه هست و جامعیت کافی رو نداره.

وقتی قیمت رو مورد بررسی قرار دادیم نکته جالبی پیدا کردیم اینکه قیمتا توی بازه زیر ۸۰۰ دلار به طور میانگین توی اجاره کل خونه بالاتر از ۲ نوع دیگه هست ولی وقتی وارد بازه لاکچری و بالا ۸۰۰ دلار میشیم میبینیم که قیمت ی اتاق تنها بالا تر قرار میگیره حالا چرا ؟ تعداد خونه های پنت هوس طور و گرون قیمت نسبتا کمتر هست ولی یه اتاق لاکچری تر تعدادشون بیشتره و بعد از مطالعه میدانی متوجه شدیم بعضی از هتل های کمتر معروف ولی لاکچری طور اتاقاشون روتوی ایر بی اند بی به فروش میذارن بخاطر همین هست که اتاق های تکی توی قیمت بالا از نظر میانگین بالاتر از اجاره کل خونه لاکچری قرار میگیرن



همونطور که دیده میشه خونه هایی که اتاق اشتراکی هستن اکثرا یک روزه هستن و منطقی هم هست مثلا فرض کنیم یک نفر فقط یک کار اداری یک روزه دارد و فقط یجا برای وسایلش میخواد تا به کاراش برسه پس میره یه اتاق اشتراکی میگیره که ارزون تر باشه بخاطر همین توی میانگین پایین تر هستن ولی همون طوری که دیده میشه اقامت های بیشتر کل خونه رو اجاره میکنن و نکته ای که دیده میشه اون بازه ۳۰ روزه هست که ما بهش میگیم اجاره ماهانه این مدل قرارداد که طرف یک ماه تمام خونه رو اجاره کنه برای بیزینس‌من ها و افرادی هست که در حال حرکت کار میکنن و مجبورن بازه ای مثلا ۳۰ روزه رو توی نیویورک برای کار یا تفریح باشن و این نمودار نشون میده ک ترجیه اونا برای این است که کل خونه رو اجاره کنن پس اگر مثلا کنارمناطق

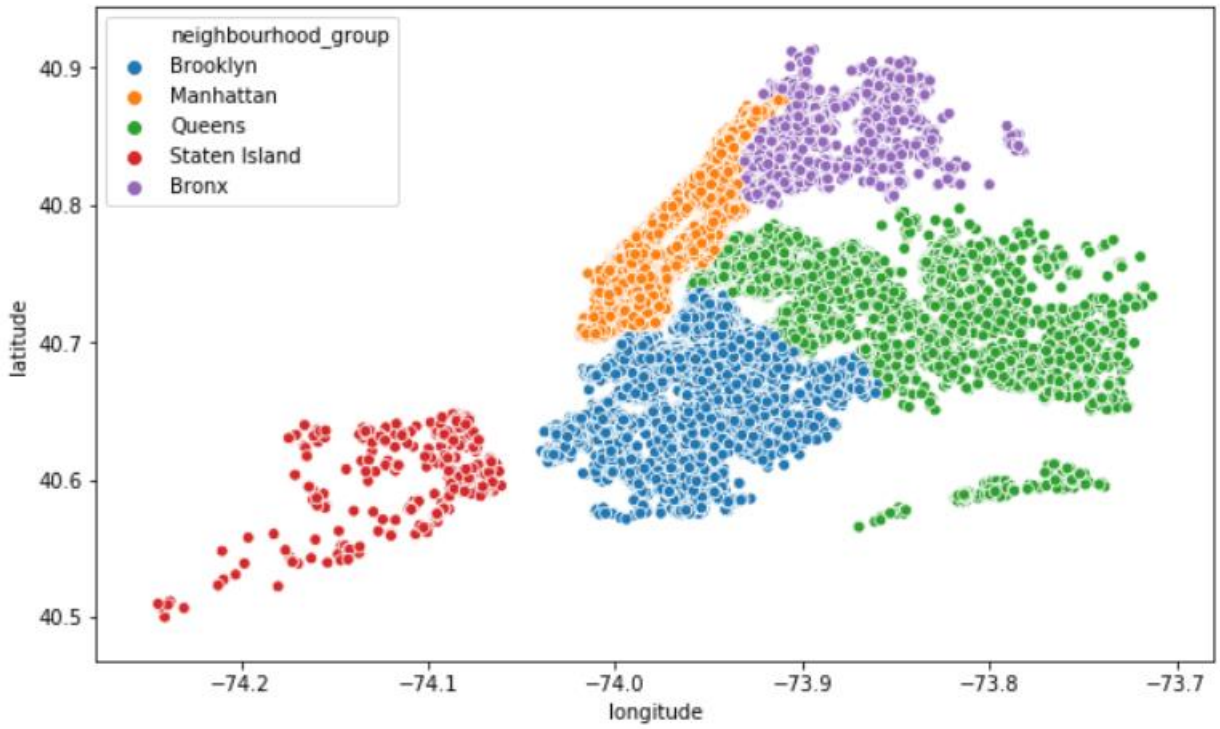
اداری هستین بهتره کل خونتون رو ۳۰ روزه اجاره بدین

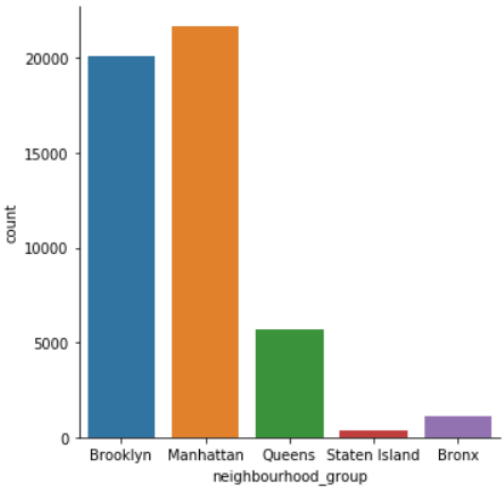


طبق تجربه ای که توی ایدی دادن دیتا بیسا توی کارای مختلفمون داشتیم خواستیم یک بررسی بکنیم که ببینیم نکته ای از اون قبیل میتونیم در بیاریم. سیستم ایدی دهی دیتا بیسا نکته خاصی نداره فقط ایدی هایی که اول وارد سیستم میشن معمولا ایدی پایین تری دارن ینی از روی ایدی میشه فهمید که اولین کاربرا چه نوع افرادی بودن اولین کاربران اکثرا توی طول جغرافیایی بیشتر بودن ینی طرفای منهتن اول همه گیر شده و با توجه به نکته هایی که در مورد نوع منهتن در اوریم خیلی منطقیه که اول جایی که نیاز سیستم بیشتر بوده کاربرا بیشتر بیان دردسترس بودن در طول سال هم رابطه نسبی داره پس بازم میتونیم بگیم حرفی که زدیم مبنی بر اینکه اول وقتی سیستم اماده به کار شده رفته با افرادی که توی این کار بودن و توی اجاره دادن خونه بودن قرارداد بسته حالا ۲ تا حالت داره یا سیستم رو وقتی زدن اون افراد که توی این کار بودن رو جذب کردن یا بنیان گذارای این سیستم خودشون توی این کار بودن و میخواستن یه سیستم بسازن که کل بازار کارشونو بگیرند و گسترده ترش کنند بخاطر همین که مشتریای ثابت داشتن این افراد ریویو هایی که میگیرفتن هم بالا تر بوده.

- **Is there any noticeable difference of traffic among different areas and what could be the reason for it?**

ابتدا داده ها رو براساس موقعیت جغرافیایشون رسم کردیم که تقریبا شبیه نقشه شهر نیورک هم میشه و منطقی هم هست ولی چیزی که از نموار پیداست اینه که تعداد داده هامون یعنی تعداد میزبان ها در رنگ در آبی و سبز بیشتر است یعنی در شهر Brooklynو شهر Queensبیشتر است ولی این لزوما نمیتونه درست باشه شاید در محله دیگه تعداد اون نسبت به این دو محله بیشتر باشن ولی متراکم تر باشن به خاطر به همین در نمودار زیر متوجه میشم که آیا حدسمون درست بوده یا نه





مردم توی بروکلین و منهتن تعداد بیشتری خونه هاشونو اجاره میدن ولی اگر با اون نمودار طول و عرض جغرافیایی کشیدیم مقایسه کنیم متوجه میشیم منهتن با توجه به اینکه مساحت کمتری داره و بیشتر ساختمونای اصلی و بزرگ و هولدینگ های شرکت های بیزینسی و کاری اونجاست ملت تعداد خونه بیشتری برای اقامت نیاز دارن بخاطر همین میشه گفت منهتن پر ترافیک ترین منطقه نیویورک میتونه باشه ولی چون بروکلین مساحت نسبتا بیشتری داره نمیتونیم به طور قطع بگیم رنک شماره یک و از نظر ترافیک بروکلین هست

- **هاست ها از نظر جنسیت چه ویژگی هایی دارند؟**

اولین ایده ای به ذهن یک فردی که از ایران این دیتاهارو میبینه این است که ایا جنسیت میزبان روی قیمت یا مینیمم شب رزرو تاثیر داره یا نه. بخاطر همین تصمیم گرفتیم که از یه مدل استفاده کنیم که از روی اسم میزبان هامون جنسیت اونارو حدس بزنیم وقتی ک فیچر جنسیت رو اضافه کردیم به کارمون چندتا نمودار کشیدیم بنظر ی سری رابطه پیدا کردیم و بعد از اون ی ازمون فرض تی تست زدیم و نتیجه گیری کردیم

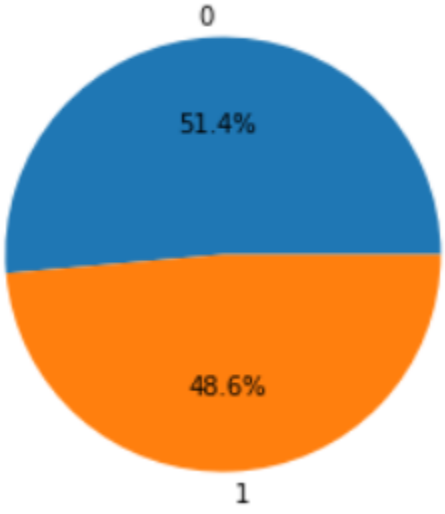
دیتا هامون کورلیشن خاصی با فیچر هامون ندارن ولی چون دیتای جنسیت باینری هست منطقیه که خیلی رابطه ای ب چشم نیاد مخصوصا چون ما رکورد هایی که جنسیتشون مجهول بود رو از دیتامون حذف کردیم ولی نکته ای که وجود داره اینه که وقتی تحلیلای توزیع هارو تست کردیم نتایج خوبی گرفتیم

قسمت اول رابطه جنسیت و قیمت خونه هارو بررسی کردیم طبق شهود نموداری بنظر یکسان میان ولی طبق نتیجه تی تست یعنی یا سمپلینگ ما اشتباه هست یا هر دو جامعه در میانگین با هم اختلاف دارند و با فرض گرفتن درست بودن سمپلینگ نتیجه میگیریم که میانگین ها اشتباه هستند پس جنسیت زدگی توی نیویورک وجود داره و قیمت یه خونه بخاطر جنسیت دارنده اون زیاد میشه قسمت دوم دنبال این بودیم که نکنه مردم به

میزبان های خانوم ریویو بیشتری میدن یا نه؟ با توجه به حرفایی که توی قسمت قبل زدیم نتیجه این شد که بله جنسیت روی تعداد ریویو تاثیر داره ولی جالب تر اینه که اقایون ریویو بیشتری میگیرن نه خانوم ها در قسمت سوم تاثیر جنسیت را روی ملاک شمارش لیست میزبان بررسی کردیم و مشابه قبل متوجه ارتباط جنسیت شدیم در قسمت سوم بررسی کردیم که ایا مینیمم شب رزرو روی جنسیت تاثیر داشته یا ن و طبق نتیجه تی تست متوجه شدیم که خیر جنسیت روی مینیمم شب رزرو تاثیر نداره

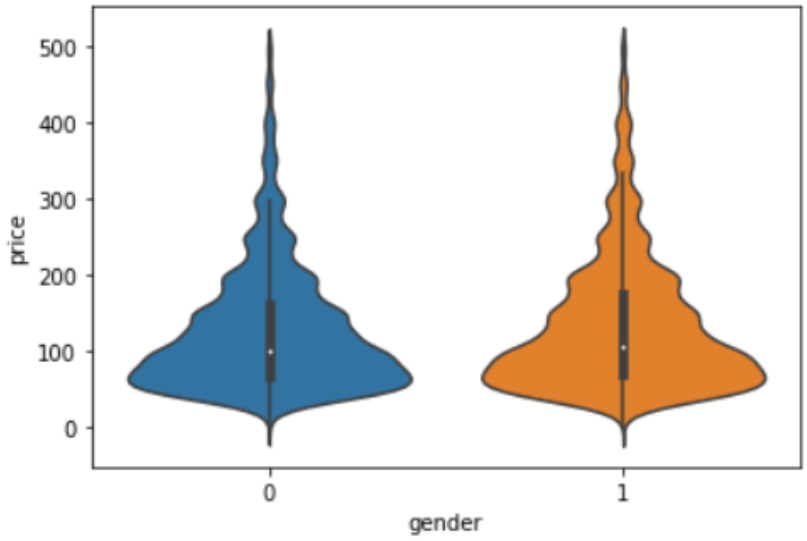
نتیجه کلی کار این شد که جنسیت تقریبا روی همه جوانب کار تاثیر دار ولی اینطوری نیست که مثلا فقط خانوم ها میانگین بالارو داشته باشن و در بعضی از موارد کلا تاثیری نداشت جنسیت پس میشه گفت جنسیت زدگی خیلی توی نیویورک خیلی مطرح نیست ونتیجه گیری قبلی صرفا میانگین های متفاوت رو نشون میداد.

برای اینکه نشون بدیم چقدر از جامعه مرد و چقدر زن هستند از نمودار زیر استفاده کردیم و متوجه شدیم که تعداد خانومایی که خونشونو اجاره میدن بیشتر هستن حالا چرا؟ اگر ما در نظر نگیریم که خطای مدلمون برای حدس جنسیت زیاد بوده و بیخیال دیتاهایی که نتونستیم چک کنیم جنسیتشون چیه و با در نظر گرفتن اینکه مدل مشابه با مدلی که برای بررسی جنسیت وجود داره برای سن افراد هم وجود داره ولی چون وب اپ بود نتونستیم توی نوت بوک نشون بدیم ولی سن افراد از روی اسمشون تقریبا بالای ۳۰ ۴۰ بود با توجه ای این نکته فک کردیم شاید منطقی باشه اینطوری فکر کنیم که خانوم های بازنشسته که منبع درامدی دیگه ای ندارن سرویسی که ایر بی اند بی در اختیارشون قرار میده که میتونن خونشون رو اجاره بدن با کمترین دردسر و درصد سایت کم میتونن به طور ماهیانه یه درامد داشته باشن و دیگه مجبور نباشن توی سن های بالا کار خاصی انجام بدن و درامدشون امن بشه



- **آیا جنسیت هاست بر روی قیمت موثر است؟**

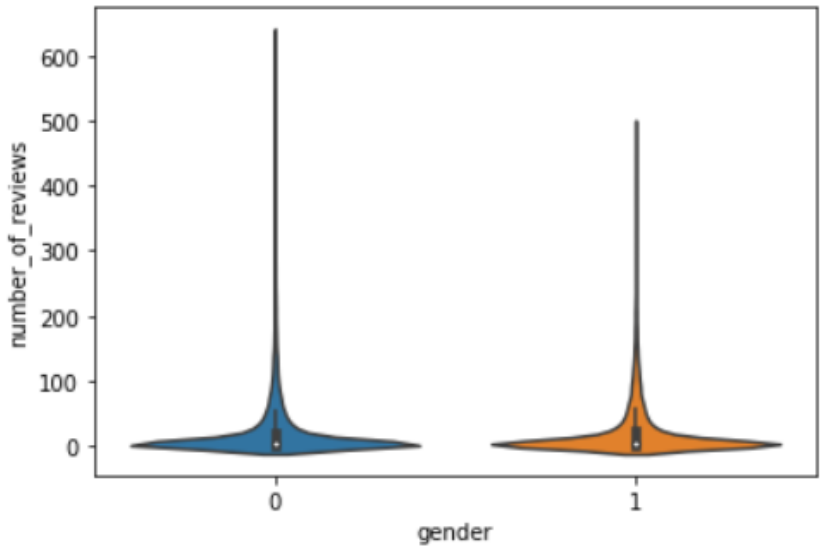
```
Ttest_indResult(statistic=1.7918969601450074, pvalue=0.07315755802542322)
```



نمیتوانیم فرض صفر را رد کنیم چون مقدار pvalue بیشتر از ۰.۰۵ است ولی اختلاف چندان زیادی ندارد

- **آیا جنسیت هاست بر روی تعداد ریویر موثر است؟**

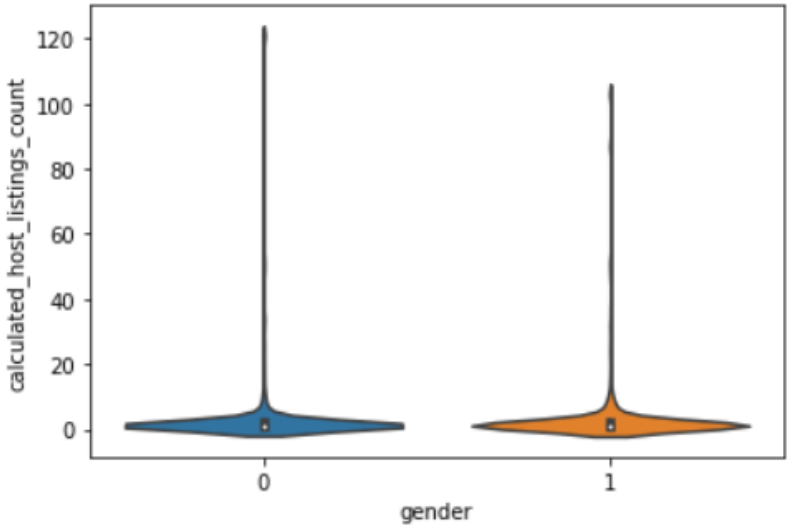
```
Ttest_indResult(statistic=2.043856646336926, pvalue=0.04097470461432081)
```



فرض صفر رد می‌شود و میانگین دو جمعیت متفاوت است

- **آیا جنسیت هاست بر روی calculated_host_listings_count موثر است؟**

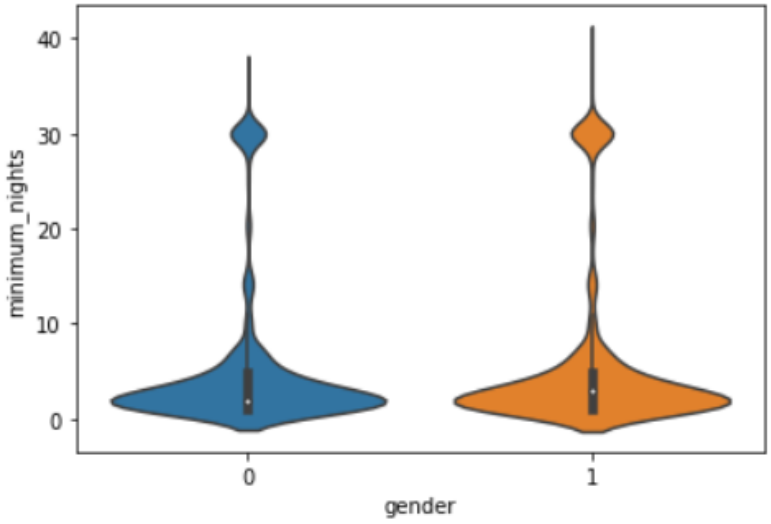
```
Ttest_indResult(statistic=5.610587116141991, pvalue=2.030656769150408e-08)
```



فرض صفر رد میشود و میانگین دو جامعه متفاوت است

- **آیا جنسیت هاست بر روی حداقل شب اجاره موثر است؟**

```
Ttest_indResult(statistic=-0.06396568177457355, pvalue=0.9489978948270685)
```



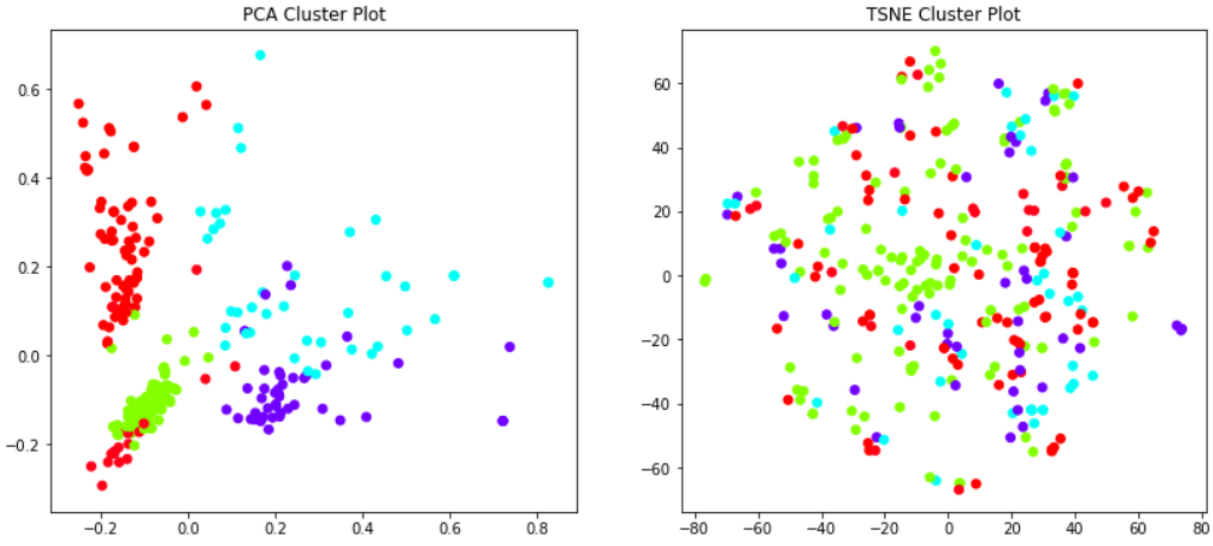
نمیتوانیم فرض صفر را رد کنیم اما به طور شهودی تفاوت اندکی دیده میشود.

- **آیا در مناطق و محله های مختلف نیویورک ادبیات و زیر ساخت های فرهنگی متفاوتی در نوع ادبیات موجود در متن آگهی ها وجود دارد؟**

نکته جالب بعدی که ما دنبالش بودیم این بود ایا توی نیویورک منطقه ای که هاست توی اون حضور داره روی نوع حرف زدنش توی متن اگهی اون تاثیری میذاره یعنی مثلا هر کدوم از منطقه های نیویورک با یک نوع حرف زدن خاص توی اگهی هاشون حرف بزنن و با لحن خاصی متن بنویسن به زبان دیگه مثلا بافت فرهنگی گفتاری متفاوتی توی ۵ منطقه اصلیه نیویورک یا بین مناطق کوچیک تر اون وجود داره یا نه

بخاطر اینکه ببینیم ایا همچین رابطه ای وجود داره متن اگهی های رو کلاستر کردیم و رند ایندکس لیبل های کلاستر هارو با اسم مناطق بررسی کردیم.

برای این کار اول اومدیم یه word to vec روی متن اگهی ها زدیم و اونارو به ۵ کلاستر تبدیل کردیم و همونطور که از نمودار و نتایج کلمات توی هر کلاستر میبینیم متوجه میشیم که خیلی ارتباطی روی متن اگهی ها وجود نداره و دقت رند ایندکس هم بسیار کم است پس کلا نمیتوانیم نتیجه بگیریم که حداقل توی ۵ منطقه اصلیه نیویورک تفاوت فرهنگی گفتاری ای بین افراد وجود داره



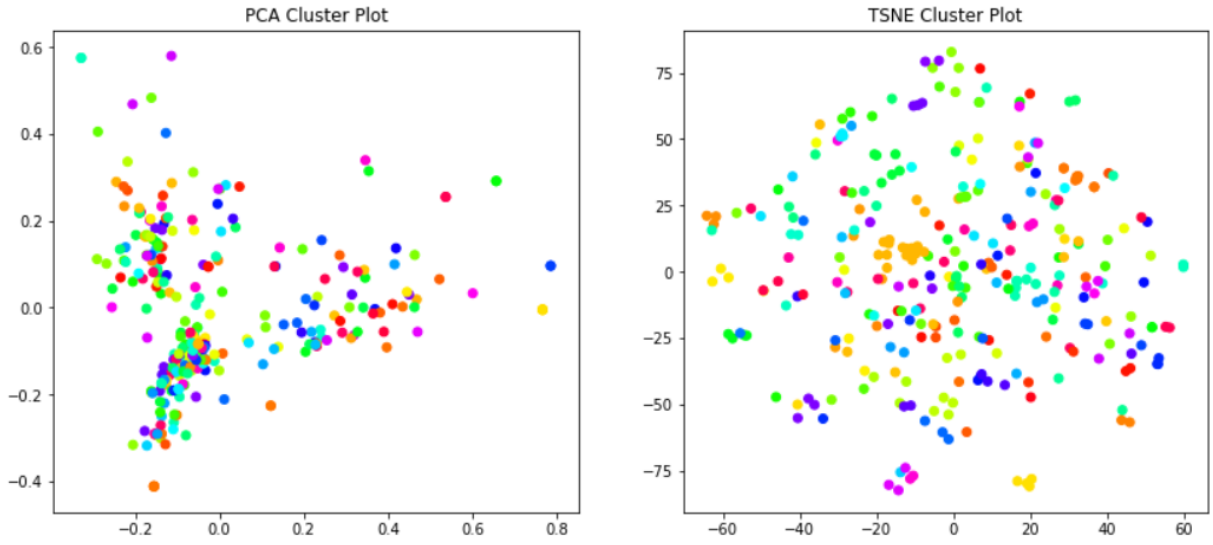
اما در نمودار سمت چپ شاهد ۴ کلاستر مختلف دیگر هستیم درست است که همهی اعضای این کلاستر ها متعلق به یک منطقه خاص نیستند اما متوجه میشویم که دست کم ۴ نوع ادبیات گوناگوندر اگهی ها وجود دارد.

مهمترین کلماتی که در ۵ کلاستر پیدا کردیم به صورت زیر بود

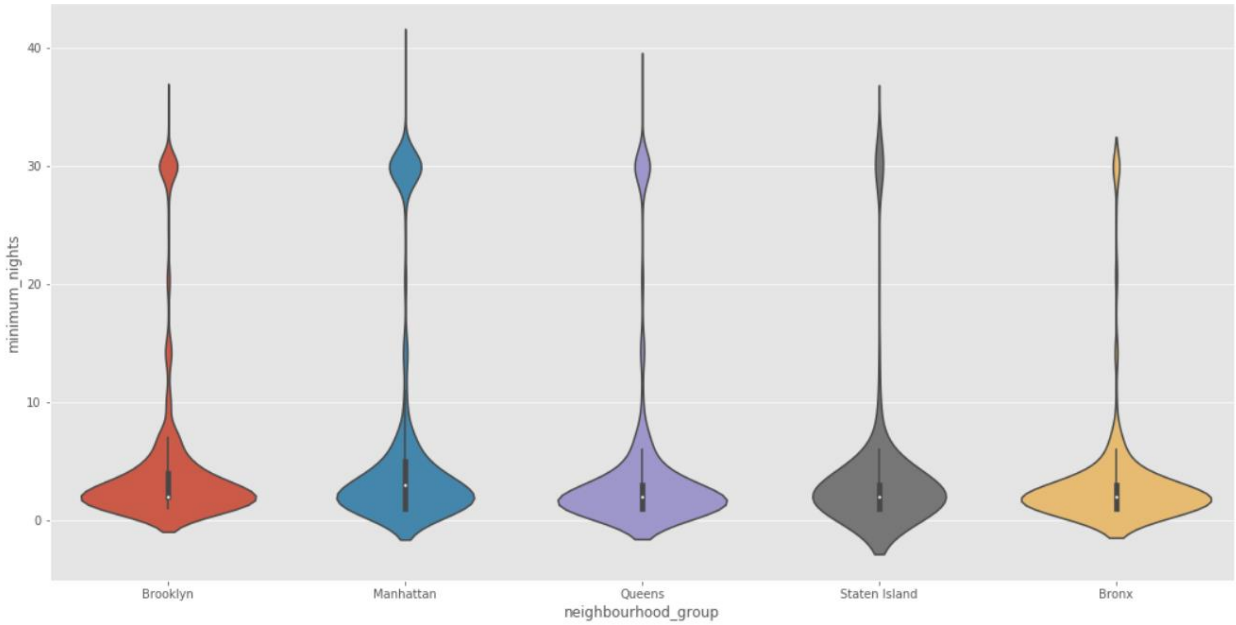
```
Cluster 0
apt,williamsburg,east,beautiful,sunny,brooklyn,spacious
,cozy,apartment,bedroom

Cluster 1
nyc,williamsburg,cozy,park,manhattan,loft,1br,home,broo
klyn,apt

Cluster 2
apt,williamsburg,brooklyn,manhattan,bath,bathroom,cozy,
bedroom,room,private

Cluster 3
beautiful,near,large,williamsburg,brooklyn,manhattan,su
nny,spacious,cozy,room

Cluster 4
west,private,manhattan,spacious,park,village,apartment,
east,cozy,studio
```

همین کار هارو برای ۲۲۱ منطقه کوچیک تر نیویورک انجام دادیم و باز هم نتیجه مشابه قبل شد و نتونستیم نتیجه ای بگیریم از متن اگهی ها

نکته جالب تر این هست که یک سری از کلمات توی کلاستر های مختلفمون پر تکرار بودن که همین کمک میکنه که نتیجه بگیریم تفاوت فرهنگی گفتاری خاصی وجود نداره و در یک قالب متن اگهی ها نوشته میشه اگر بخاطر قوانین محدود کننده برای متن اگهی ها نباشه و مردم وقتی به تازگی عضو سایت میشن سعی میکنن مشابه بقیه اگهی های موجود توی سایت متن بزنن پس اینطوری اون نوع اگهی های یوزر های اولیه تعمیم پیدا میکنه به نوع حرف زدن یوزر های جدیدمون توی بازه زمان و باعث یکسانی فرهنگ و نوع گفتار افراد مناطق مختلف در نیویورک اگر موارد بالا نباشند که خیلی محتمل هستند میشه نتیجه گرفت یک فرهنگ قالب گفتاری نوشتاری توی کل مناطق نیویورک برقرار است و نتیجه گیری بهتری با این دیتاها نمیشه گرفت.



- **آیا در مناطق مختلف اجاره مدت زمان ماهانه تفاوت است؟**



نکته قشنگ این نمودار این هست ک اجاره ماهانه توی منهتن بیشتر از بقیه است پس نتیجه ای که در بخش اول گرفتیم درست محسوب میشود.

- **What can we learn from predictions?**

اگر بتوانیم زمان آگهی ها را هم داشته باشیم یا حداقل فصلی که آن ها جمع آوری شده اند میتوانیم فصل های ارزان قیمت را پیشبینی کنیم و برای سفر در آن فصل ها به نیویورک برویم.

میتوانیم برای سرمایه گزاری خانه هایی که در محله هایی با اجاره خانه به صورت تک اتاق بیشتر است را خریداری کنیم و اتاق های آن را در فصل های گران قیمت به صورت اتاقی اجاره دهیم و در فصل های عادی به صورت کل خانه ای اجاره دهیم، یعنی یک سیستم طراحی کنیم که پیشنهاد دهد که در هر بازه از زمان خانه به چه صورتی و چند روزه برای اجاره قرار گیرد در بازه طولانی مدت سود هاست ماکزیمم میشود.

برای ساختن هتل در مناطقی که تقاضا زیاد است را میتوانیم بدست آوریم.

اگر باز هم زمان را داشته باشیم میتوانیم بررسی کنیم که تقاض در چه محل هایی روبه افزایش است و در آن محل ها هتل احداث کنیم و یا خانه هارا بخریم و اجاره دهیم.

- **Which hosts are the busiest and why?**

جواب این سوال را در بخش های قبلی تا حد خوبی پاسخ دادیم. به طور کلی هاست های بروکلین به دلایل بالا که گفتیم مشغول ترین ها هستند و حتی هاست هایی که ایدی کمتری دارند به دلیلی که اشاره کردیم جزو فعال ترین ها محسوب میشوند چون شغل اصلی آن ها همین است. ولی به طوری کلی تر میتوانیم به گوییم مناطقی که به صورت روزانه بیشتر خانه های خود را اجاره میدادندشلوغ تر محسوب میشوند اما این نتیجه گیری به این دلیل است که صرفا رکورد های بیشتری از آن ها در دیتا بیست وجود دارد ولی طبق تعریف شلوغ محسوب میشوند.

مجموعه داده International Football Results:

مقدمه:

مجموعه داده در این بخش شامل تاریخچه همه بازی های رسمی فوتبال از سال ۱۸۷۲ تاکنون است. و اولین بازی صورت گرفته بین دو تیم انگلستان و اسکاتلند بوده که نتیجه بازی مساوی بوده است.
مراحل کلینینگ و بعضی مطالعاتEDA بر روی دیتا ست انجام شده ولی چون هدف از تمرین ۱.۱ تنها بررسی آماری و تحلیل دیتا ها میباشد آن ها را ذکر نمیکنیم و مستقیما به پاسخ دادن به سوالات و مراحل طی شده برای رسیدن به آن ها را ذکر میکنیم. توضیحات تکمیلی این بخش ها در نوتبوک قابل مشاهده است.

سوالات:

- **Who is the best team of all time?**

برای رده بندی کشورها سیستمی که نوشتیم این است در هر برشی از تاریخ که بخواهیم میتوانیم رده بندی تا آن تاریخ دلخواه انجام دهیم. قوانین رده بندی که پیاده سازی کردیم بدین صورت است:

۱. هر تیمی که برنده مسابقه شود۶ امتیاز مثبت می‌گیرد اما اگر تیمی که امتیاز پایین تری نسبت به تیم دیگر داشته باشد و برنده شود بدین صورت عمل می‌کنیم:

امتیاز تیم برنده = اختلاف امتیاز دو تیم × (ضریبی بین ۰ و ۱) + ۶

مثلا فرض کنیم تیم a در رده بندی دارای ۱۵۰۰ امتیاز و تیم b دارای ۱۲۰۰ امتیاز باشد و تیم a برنده شود امتیاز تیم a بدین صورت محاسبه میشود:

به امتیاز تیم a (که در رده پایین تری قرار داشت

(۱۵۰۰-۱۲۰۰)*c+۶ اضافه میشود که در آن c یک عدد بین صفر و یک است

اگر تیم b برنده شود به آن ۶ امتیاز اضافه می‌گردد

۲. تیمی که بازنده مسابقه شود ۱ امتیاز منفی می‌گیرد

۳. مسابقه ای که به تساوی بیانجامد به هر تیم ۳ امتیاز مثبت داده میشود اما اگر تیم که امتیاز پایین تری نسبت به تیم دیگری داشته باشد و مساوی کند امتیاز آن دو تیم بدین صورت محاسبه می‌گردد که میانگین اختلاف امتیاز دو تیم رو محاسبه می‌کنیم و به تیمی که رده پایین تری قرار دارد این مقدار به امتیازات آن

اضافه و از امتیازات تیمی که در رده بالاتری قرار دارد کم میشود مثلا فرض کنیم تیم a در رده بندی دارای ۱۵۰۰ امتیاز و تیم b دارای ۱۰۰۰ امتیاز است و این دو تیم مساوی میکنند امتیاز این دو تیم بدین صورت محاسبه میشود

امتیاز تیم a و b برار است با (۲۵۰) + ۳ که عدد ۲۵۰ میانگین اختلاف دو تیم است.

نکته : اگر تیمی که تاکنون موجود نبوده اضافه شود برای محاسبه امتیازات آن بدین صورت عمل میکنیم به ازای هر باخت برای آن -۱ امتیاز در نظر میگیریم تا این که این تیم اولین مسابقه خود را برنده و شود و امتیاز آن مانند تیم برنده که در رده بندی پایین تر است که در قسمت (۱) توضیح داده شد محاسبه میگردد.

به طور کلی برای اینکه بتوانیم تحلیل درستی روی دیتا داشته باشیم و بتوانیم به سوالات مسئله پاسخ بدیم باید سیستمی برای رنک کرن تیم ها ارائه می‌دادیم این کار با چالش های بسیاری روبرو بود مثلا خیلی از تیم ها رفته رفته وارد مسابقات میشدند و دیتایی برای اون ها نداشتیم و اگر نمیخواستیم این مورد را بررسی کنیم انگلستان به عنوان کهن ترین تیم همیشه رنک شماره یک را به خود اختصاص می‌داد و هیچ روند مناسبی برای رنکینگ در طول تاریخ بدست نمی‌امد هدف کلی ما در این بخش این بود که بتوانیم رنکینیگی در هر برش از تاریخ فوتبال اراعه دهیم برای این کار از سیستم کلی شرط بندی الهام گرفتیم و روند فوق را ارائه دادیم و بعد از تعریف کلی این روند ها تنها چیزی که لازم بود تیون کردن پرامترهای الگوریتممون بود

به طور کلی اگر مسابقه ی فوتبالی برگزار شود هر دو تیم امتیازی که تاکنون در رنکین دارند رو وارد بازی می‌کنند اگر نتیجه ی هر دو تیم برابر میانگین هر دو امتیاز میشود اینطوری اگر یک تیم خیلی قوی با یک تیم خیلی ضعیف مسابقه دهد یعنی افت داشته و باید رنکش کم شود و تیم ضعیف عملکرد خوبی نسبت به کاری که انجام داده داشته و امتیاز کسب میکند حال اگر مسابقه مساوی نشود اگر تیم ضعیف برنده شده باشد امتیازش برابر تیم قوی تر می‌شود و درصدی از اختلاف امتیاز هر ۲ تیم را میگیرد و تیم بازنده یک امتیاز از دست می‌دهد اگر تیم قوی تر برنده باشد امتیاز تیم قوی ۶ واحد اضافه میشود و تیم بازنده یک واحد کم می‌شود.

این سیستم این مزیت را دارید که اگر تیمی ضعیف تیم قوی ای را ببرد که اختلاف امتیاز بالایی دارد این شانس را به تیم های ته جدول می‌دهد که بتوانند به صدر صعود کنند و تیم های صدر جدول همیشه یکه تاز نیستند.

حال اگر تیمی برای اولین بار وارد مسابقات شود هنگامی که اولین مسابقه که برنده شود طبق فرمول بالا امتیاز بگیرد و جایگاهش را در جدول کسب می‌کند روند فوق را برای هر رکورد اجرا کردیم و اینگونه رنکینگی در هر بازه تاریخ ارائه دادیم.

نتایج بدست آمده بر روی بازه از اول تاریخ دیتا تا اخرین دیتای ثبت شده به صورت زیر می‌باشد:

کشور	امتیاز
Belgium	1121.279795
Netherlands	1069.773044
Germany	1051.457570
Italy	1050.822424
Mexico	1037.490505

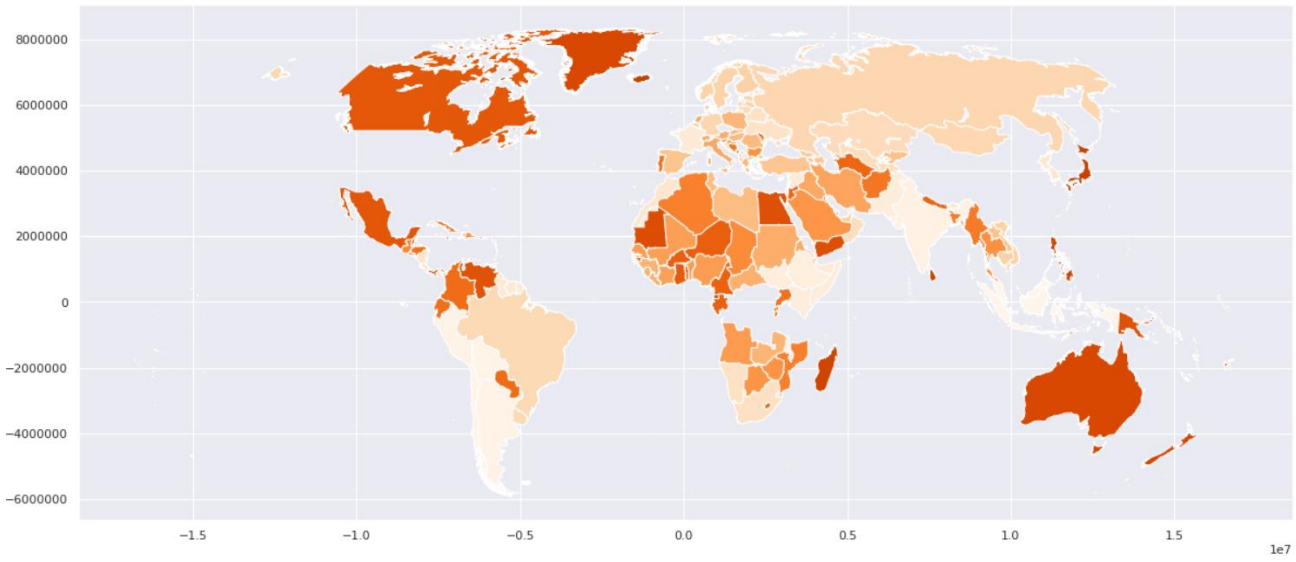
با در نظر گرفتن متد استفاده شده تیم بلژیک بهترین تیم دنیا می‌باشد.

- **Which teams dominated different eras of football?**

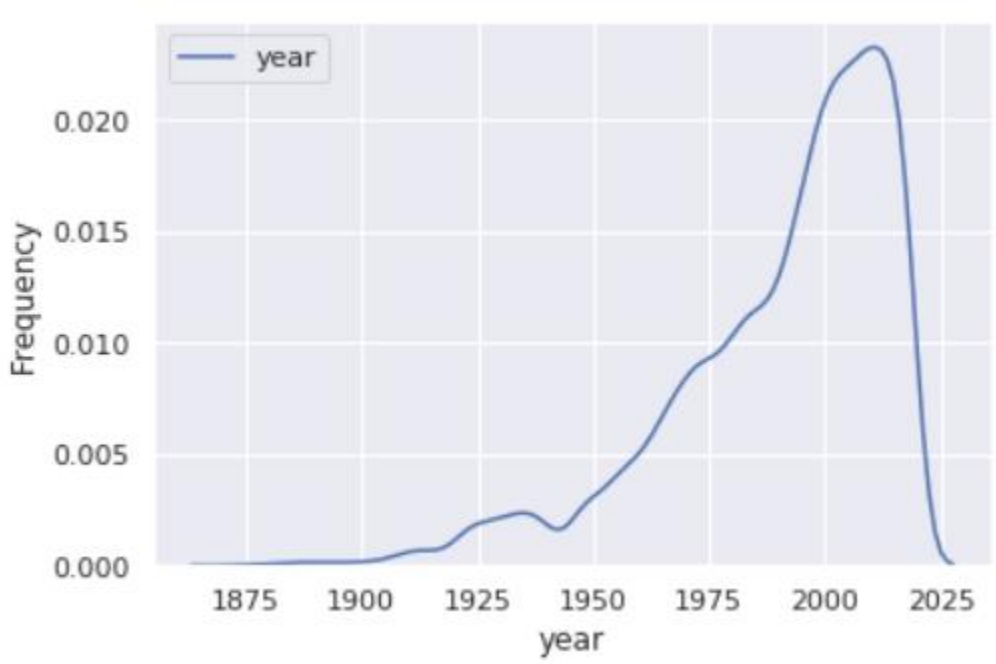
با این سیستم که ما ارائه دادیم اولا در حال حاضر نتایج سیستم رنکینگ ما با رنکینگ داغون فیفا برابر هست و تاریخ پیدا میکنیم وقتی تیمی برای ۵ سال همیشه رنگ یک رو به خودش اختصاص میده یعنی در این ۵ سال دامینیتور بود و میتونیم عصر های فوتبال رو دنبال کنیم مثلا دوره ای که پلله همچنان بازی می‌کرد همیشه ارژانتین بالاست.

- **Can we say anything about geopolitics from football fixtures - how has the number of countries changed, which teams like to play each other?**

در کل جهان ۲۵۵ کشور داریم در حالی که در لیست مسابقاتی که برگزار شده با بررسی هایی که انجام دادیم تعداد ۳۱۴ کشور موجود است که نشان دهنده این است که تعدادی از کشور های قبلا وجود داشته اند و بنا به دلایل مختلفی از بین رفتند یا تجزیه شدند یا به کشورهای دیگری ملحق شدند.



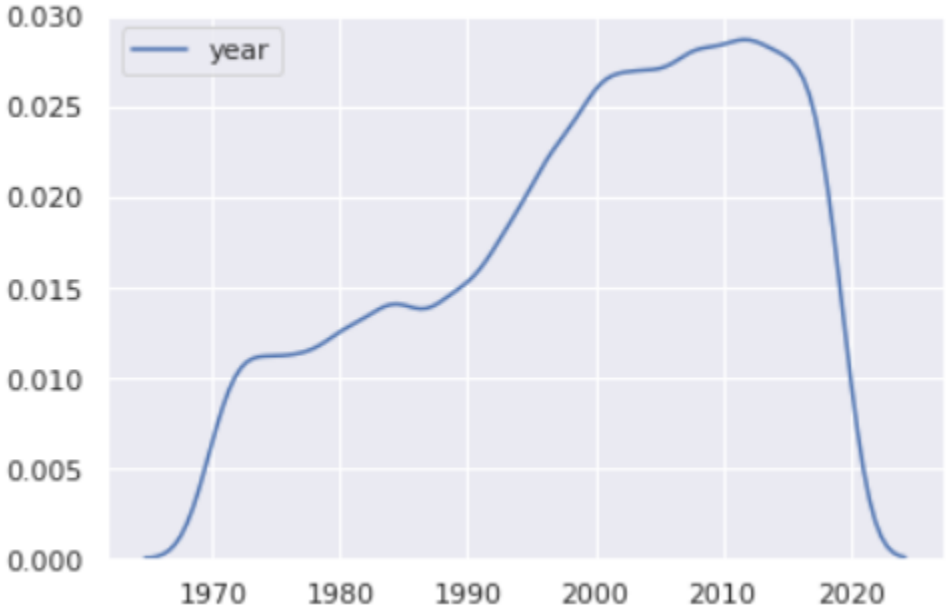
این نمودار هم هیت مپ کشورها رو براساس تاریخی که به فوتبال پیوستند نشون میده



این نمودار هم روند برگزاری مسابقات فوتبال در سطح ملی در سال های مختلف رو نشون میده که روند که از سال ۱۸۷۶ سالی که اولین کشور ها به فوتبال پیوستند هواره سیر صعودی داشته فقط در یک بازه سیر نزولی داشته اون قبل از سال ۱۹۵۰ بوده که که علت اون رو هم میتوان جنگ جهانی دوم دانست بین سال های

۱۹۳۹ تا ۱۹۴۴ رخ داده و دقیقا در این بازه بوده که سیر بازی فوتبال نزولی بوده و از آن سال به بعد یعنی از سال ۱۹۴۵ سیر صعودی فوتبال نیز افزایش بیشتری پیدا کرد

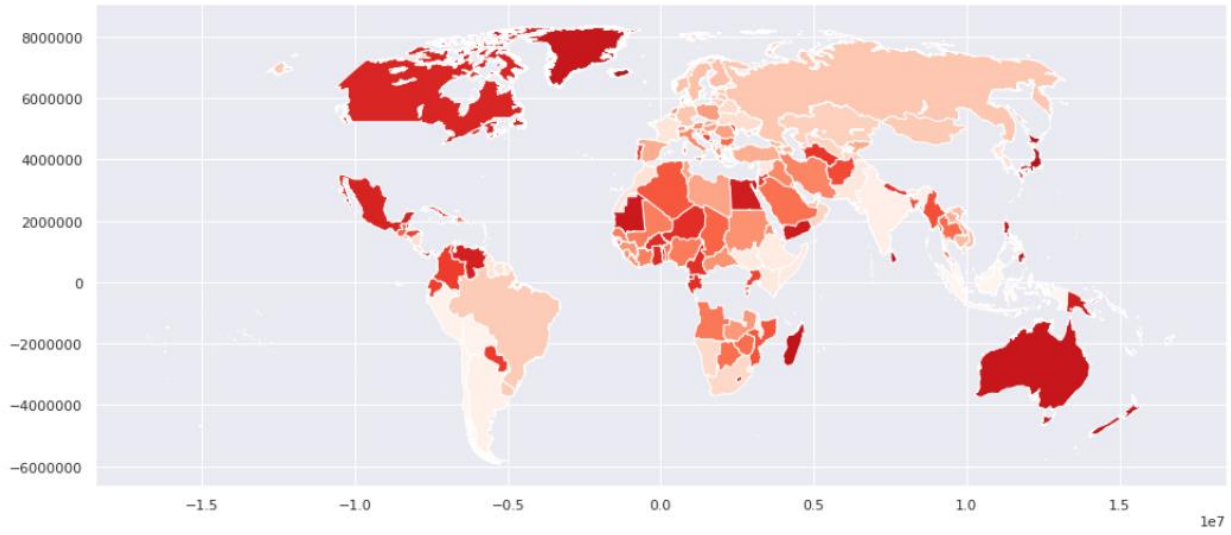
در نمودار بعد هم همین نمودار فقط برای مودش رو نشون میدهد که طبق اون روند برگزاری مسابقات در بین سال های ۱۹۷۰ تا ۲۰۱۲ بیشترین سیر صعودی برگزاری مسابقات در سطح ملی رو دارد.



- **Which countries host the most matches where they themselves are not participating in?**

کشور	تعداد بازی دیگر تیم ها در آن کشور
United States	761
Malaysia	428
France	374
South Africa	283

United Arab Emirates	257
England	250
Qatar	227
Spain	214
Thailand	207
Brazil	193
Sweden	191
Germany	171
Soviet Union	164
South Korea	155
Egypt	150
Singapore	138
Kuwait	134
Mexico	128
India	126
Tanzania	123



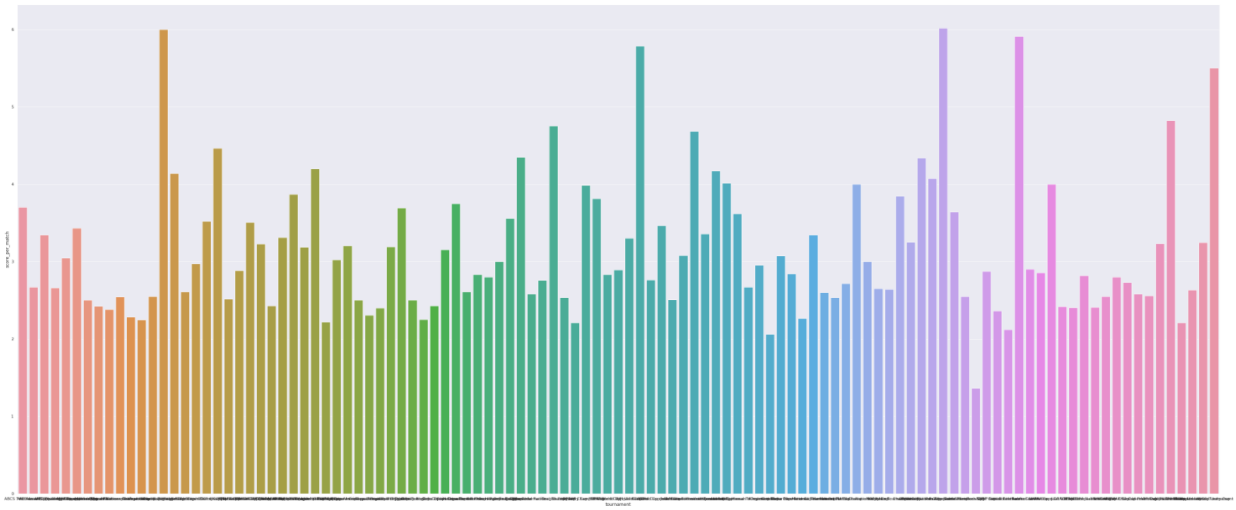
در همین قسمت میزبان ها, اومدیم کشور هایی که میزبان مسبقات دیگر کشور ها بودند که مثلا یا میزبان تورنمتی بودند یا کشور های دیگه برای مسابقات دوستانه خود این کشور ها رو انتخاب کردند رو مورد بررسی قرار دادیم.

نکته جالب توجه این جاست که در بین کشور هایی که میزبان مسابقات سایر کشور ها بودند نیز امریکا با ۷۶۱ میزبانی بیشترین میزبانی را دارد و بعد از آن مالزی و فرانسه قرار دارد.

نمودار بالا هم هیت مپ همین قسمت است

- **کدام تورنومنت ها جذابیت بیشتری دارند؟**

یکی از معیار هایی که در هر تورنمنت در نظر میگیرن مقدار گل زده در هر مسابقه در اون تورنمنت هست که می تواند نشان دهنده هیجان و جذابیت مسابقات در ان تورنمنت باشد چون هر چه تعداد گل زده در هر مسابقه بیشتر باشد معمولا آن مسابقه جذاب تر ودیدینی تر است . به همین ما برای هر تورنمنت این معیار رو حساب کردیم و نمودار ان را کشیدیم



tournament	Score per match
Pacific Games	6.019608
Atlantic Heritage Cup	6.000000
South Pacific Games	5.912195
GaNEFo	5.785714
World Unity Cup	5.500000

- **آیا روند ها استعماری کشور های اروپایی از روی فوتبال و بازی ها قابل برداشت است؟**

بله همانطور که درنمودار سال ورود کشور ها به بازی های فوتبال قابل مشاهده است کشور های هند و همسایگاه انگلستان و کشور های آفریقایی کشور هایی هستند که زود تر از باقی کشور ها وارد بازی ها شدند که میتوان نتیجه گرفت که به این دلیل میباشد که در زمان استعمار آن ها فرهنگ کشور استعمار گر در قالب بازی های فوتبال وارد آن کشور ها میشدند.