به نام خدا

گزارش تمرین 1 داده کاوی

1 – تحلیل داده ها و پیدا کردن جنسیت میزبان ها

در ابتدا پکیجی برای حدس زدن جنسیت نام ها نصب می کنیم و در مرحله بعد مواردی که نیاز داریم import میکنیم. سپس گوگل درایو را به نوت بوک خود اضافه می کنیم و فایل csv داده ها را می خوانیم.

با تایپ ny که نام دیتافریم ما می باشد چک می کنیم که فایل csv را خوانده ایم.

حالا شروع به تمیز کردن داده ها می کنیم. در مرحله اول تعداد سطر های خالی هر ستون را می بینیم. همچنین تعداد شماره هاست های تکراری را محاسبه می کنیم و می بینیم 11438 تبلیغ کننده تکراری وجود دارد و با چک کردن سطر های تکراری می بینیم که تبلیغ تکراری وجود ندارد.

برای ستون name در هر جای خالی عبارت “no-description” را وارد می کنیم و با چک کردن دوباره می بینیم که 16 جای خالی پر شده اند.

از آن جایی که می خواهیم از جنسیت نام ها در تحلیل ها استفاده کنیم سطر هایی که نامی ندارند حذف می کنیم و همچنین ستون تعداد نظرات و میانگن نظرات را بدلیل تعداد بالای سطرهای خالی حذف می کنیم. مشاهده می کنیم که دیگر هیچ ستونی سطر خالی ندارد.

حال برای آشنایی بیشتر با داده ها مقادیر پرتکرار ستون های object و Boolean را محاسبه کنیم. همچنین بدلیل عدم اهمیت ستون id این ستون هم حذف می کنیم.

حالا به پیش پردازش داده ها می رسیم. می بینیم که حداقل و حداکثر قیمت فاصله زیادی با میانگین و انحراف از معیار داده ها دارند و منطقی نیستند به همین دلیل 2 درصد پایین و بالای داده ها را حذف می کنیم.(در ابتدا سعی شد که 3 برابر انحراف از معیار از میانگین حساب شد ولی هنوز قیمت کف بیش از اندازه پایین بود.)

در مرحله بعد تعداد تکرار اسامی همسایگی ها را می شماریم و همچنین تعداد تکرار انواع اتاق ها را می شماریم. مشاهده می شود که تعداد همسایگی هایی که تنها یکبار آمده اند زیاد هستند و از آنجایی که می خواهیم از نمودار برای دیدن نتایج استفاده کنیم آن هایی که تعداد تکرار کمتر از 300 عدد دارند را بهother تغییر نام می دهیم و در غیر این صورت نام آن ها را نگه می داریم.

همچنین همان طور که گفتیم می خواهیم از جنسیت برای تحلیل داده ها استفاده کنیم پس تابع get\_gender را روی ستون host\_name اجرا کرده و آن را در ستون Gender می ریزیم.(همچنین برای جلوگیری از مشکل در خواندن نام ها را به حروف بزرگ تبدیل می کنیم).

با بررسی نتیجه در مرحله بعد می بینیم که تعدادی نام زن هستند تعدادی مرد و تعدادی هم به طور مساوی استفاده می شوند با نوشتن تابعی به نام gdrc() این گروه ها را به سه گروه زن مرد و ناشناخته تبدیل می کنیم.

حالا برای حل مشکل احتمالی آگهی تکراری ستون های host\_name و طول و عرض جغرافیایی را با هم ترکیب می کنیم و در ستونی به نام combined می ریزیم. از آن جایی که فقط دو مورد آگهی با خانه تکراری وجود دارد تکراری ها را حذف می کنیم. بنابراین تعداد آگهی که باقی مانده برای خانه های مختلف توسط افراد است.

2- تحلیل داده ها و دیداری سازی

برای شروع تحلیل تعدادی از شماره میزبان هایی که بیشترین نظرات را داشته اند دریافت کرده سپس سطر هایی که این میزبان ها دارند دریافت می کنیم. برای توضیح این که چرا این افراد بیش از بقیه مورد نظر قرار گرفته اند می توان گفت که JJ و WANDA و DONNA هزینه کمتری نسبت به میانگین محله خود دریافت می کنند اما شاید دلیل پرطرفداری CAROL نحوه آگهی دادن او باشد البته ممکن است این موضوع بدلیل نزدیک بودن به مترو یا مکان دیدنی هم باشد.

در ادامه می توانیم میانگین تعداد نظرات برای مردان و زنان را دید که با توجه به نمودار این میانگین برای مردان بالاتر است همچنین میزان قیمت را برای زنان و مردان دید که در این نمودار هم مردان میانگین بالاتری دارند.

در قیمت بعد نمودار دیگری داریم که نشان می دهد تعداد آگهی ها در هر محله چه میزان می باشد که نمودار این موضوع را به خوبی نشان می دهد.

نمودار بعدی تعداد آگهی ها را در هر گروه آگهی نشان می دهد که همان طور که می توان دید منهتن و بروکلین بیشترین تعداد آگهی را دارند.

نمودار بعدی قیمت را در هر محله نشان می دهد که می تواند در شناخت محله های ارزان تر و گران تر به ما کمک کند.

دو نمودار بعد گروه های اصلی هر محله با میانگین قیمت و میانگین تعداد نظرات نشان می دهد و شاید با توجه به این دو نمودار بتوان گفت که بین قیمت ها و میزان نظرات رابطه ای عکس وجود دارد. برای چک کردن این موضوع می توان قیمت و تعداد نظرات را با همسایگی ها مقایسه کرد – با کشیدن نمودار می توان دید که ظاهرا چنین رابطه ای در بیش تر محلات وجود دارد و قیمت با تعداد نظرات رابطه ای معکوس دارد.

نمودار بعدی تلفیق گروه های محله با نوع خانه ها و قیمت ها را نشان می دهد که می تواند در تحلیل داده ها به ما کمک کند نکته ای که این نمودار دارد این است که قیمت اتاق های اشتراکی استاتن ایلند از همه محله ها بیشتر است که بنظر منطقی نمی آید چون انواع دیگر خانه ها قیمت کمتری دارند. با بررسی بیشتر و نمودار های قبلی می بینیم که این موضوع می تواند به دلیل کم بودن تعداد آگهی های این نوع خانه باشد.

نمودار بعدی هم تعداد روزهای در دسترس را برای محله های مختلف نشان می دهد که همان طور که می بینیم نه تنها میانه بلکه کوارتایل اول و سوم هم در بروکلین و منهتن از بقیه مکان ها کمتر است دلیل این موضوع می تواند توریستی بودن این مکان ها باشد و میانه ها و کوارتایل های اول و سوم هم در استاتن ایلند و برانکس از همه جا است یعنی خانه ها در زمان های بیش تری قابل دسترس هستند.

نمودار بعدی میزان حداقل شب رزرو را نشان می دهد که دوباره در بروکلین و منهتن از همه جا بیش تر است.

در نمودار بعدی می توان پرتکرار ترین کلمات را که در توضیحات آگهی ها آمده اند دید.

حالا میزان تعداد آگهی های هر فرد را بررسی می کنیم. تعداد آگهی ها به طور میانگین 1.3 است و میانه و مد تعداد آگهی برای هر فرد 1 است. همچنین می توان دید که کوارتایل های اول و سوم هر دو یک هستند که با توجه به میانگین پایین این موضوع منطقی است. در بخش بعد با نمایش بیشترین افرادی که آگهی داده اند می توان دید مثلا نفر اول 317 آگهی داده نفر دوم 232 آگهی داده و الی آخر. موضوع دیگری که باید اشاره شود این است که با توجه با حذف نام ها و مکان هایی که تکراری بودند آگهی هایی که اکنون وجود دارند همگی منحصر بفرد هستند و هیچ فردی یک خانه را دوبار اجاره نداده است پس میانگین تعداد خانه های هر فرد برابر 1.3 است و بقیه نکاتی که برای تعداد آگهی هر فرد گفتیم در واقع برای تعداد خانه های هر فرد هم برقرار است پس میانه و مد تعداد آگهی برای هر فرد 1 است. همچنین کوارتایل های اول و سوم هر دو یک هستند.

3- تست های فرض

دو تست اول تست های فرض برای چک کردن نرمال بودن یا نبودن داده ها است. تست اول تست شاپیرو است که روی قیمت داده هایی که چنسیت میزبانشان مرد بوده انجام شده است و همان طور که نتیجه را می توان دید نتیجه این تست منفی است تست دوم تست نرمال تست است که بر روی قیمت داده هایی که جنسیت میزبانشان زن بوده انجام شده است و نتیجه این تست هم این است که این داده ها نرمال نیستند. ابتدا می خواهیم ببینیم که آیا تفاوت واقعی بین میزان قیمت خانه های میزبان های زن یا میزبان های مرد وجود دارد یا نه. برای این کار از آزمون های فرض غیر پارامتری Kruskal و mannwhitneyu استفاده می کنیم. آزمون اول متفاوت بودن توزیع های زن و مرد را می سنجد و آزمون دوم علاوه بر این دو متفاوت بودن آن هایی که جنسیتشان مشخص نیست هم می سنجد که نتیجه هر دو با اطمینان 95 درصد این است که این توزیع ها متفاوت هستند. برای نرمال کردن این داده ها میانگین را از هرنقطه داده کم کرده و بر انحراف از معیار تقسیم می کنیم و حالا می خواهیم ببینیم که آیا تفاوت واقعی بین میزان قیمت خانه های میزبان های زن یا میزبان های مرد وجود دارد یا نه. با استفاده از آزمون ttest این موضوع را تست می کنیم و می بینیم که با اطمینان 95 درصد تفاوتی بین این دو توزیع وجود ندارد. بنابراین نتیجه آزمون غیرپارامتری با آزمون پارامتری متفاوت است.

حالا دو متغیر دیگر تعریف می کنیم برای تعداد نظرات هر جنسیت و دوباره قبل از بررسی متفاوت بودن آن ها نرمال بودن آن ها را چک می کنیم. پس از این موضوع دوباره متغیر های تعریف شده را نرمال می کنیم و اینبار میخواهیم ببینیم که آیا در حالت کلی تفاوتی بین تعداد نظرات در بین مرد ها زن ها و افرادی که جنسیتشان مشخص نبود وجود دارد یا نه که برای این کار از تست یک طرفه ANOVA استفاده می کنیم. پس از اجرای این تست می بینیم که با احتمال 95 درصد تفاوتی بین ای سه توزیع وجود نداشته و بر اساس آزمون های پارامتری جنسیت تاثیری در میزان نظرات ندارد.

5- تلاش در ساخت مدل برای پیش بینی قیمت

ابتدا ماتریس کوریلیشن را می سازیم و در ادامه هنگام انتخاب فیچر ها می توانیم آن ها را با ماتریس مقایسه کنیم.

سپس رگرسیون خطی را روی مدل خود با تمام ستون ها اجرا می کنیم. می بینیم که امتیاز رگرسیون 0.3 است در مرحله بعد دو ستون طول و عرض جغرافیایی را حذف می کنیم و رگرسیون را روی آن اجرا می کنیم. امتیاز رگرسیون 0.05 کمتر می شود که به نظر نسبتا زیاد است با توجه به ماتریس می توان گفت عرض جغرافیایی دلیل این موضوع است چون ارتباط نزدیکتری با هزینه دارد که از لحاظ منطقی هم باید مکانی که خانه در آن قرار گرفته شده با هزینه ارتباط داشته باشد و این موضوع منطقی است.

حالا رگرسیون لسو و ریج هم روی این دیتا فریم اجرا می کنیم که متوجه می شویم امتیاز آن ها چندان تفاوتی با رگرسیون خطی ندارد.

در مرحله بعد host\_listings\_count و availabilty365 را حذف می کنیم و می بینیم که تاثیر خیلی زیادی روی امتیاز محاسبه شده ندارد و البته انتظار نمی رود قابل استفاده بودن در کل سال و تعداد آگهی های هر فرد ارتباط زیادی با قیمت داشته باشد. با این وجود باز هم تاثیری وجود دارد برای چک کردن این که کدام یک از این دو ستون این تاثیر را گذاشته اند یکبار availabilty365 را حذف می کنیم و یکبار host\_listings\_count را حذف می کنیم و می بینیم که availabilty365 تاثیر بیشتری دارد.

حالا دو مورد دیگر با نام های minimum\_nights و number\_of\_reviews را حذف می کنیم و می بینیم که تاثیر محسوسی روی امتیاز مدل ما ندارد که هم از لحاظ منطقی این موضوع واضح است و هم می توان از روی ماتریس این را بررسی کرد.

حالا ستون نوع اتاق را حذف می کنیم و نتیجه را بررسی می کنیم و می بینیم که نوع اتاق تاثیر محسوسی روی امتیاز محاسبه شده دارد که البته واضح است که نوع اتاق باید بر قیمت تاثیر بگذارد و این موضوع را هم می توان در ماتریس مشاهده کرد.

در ادامه دو ویژگی که کمترین ارتباط را با قیمت دارند یعنی تعداد نظرات و همسایگی را حذف می کنیم و امتیاز را محاسبه می کنیم(البته انتظار می رود که همسایگی ارتباط خوبی با قیمت داشته باشد دلیل این موضوع می تواند پرت بودن و کم بودن تعداد بعضی از همسایگی ها باشد که باعث کم شدن رابطه این دو شده است.) پس از این کار از دو روش Descissiontreeregressor و Randomforrestregressor استفاده می کنیم و مشاهده می کنیم که روش دوم امتیاز بالاتری نسبت به بقیه روش ها دارد.