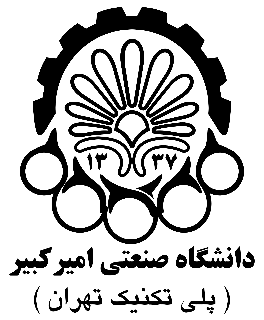
به نام خداوند بخشنده و مهربان



**درس داده کاوی**

پروژه پایانی درس

استاد درس: دکتر ناظرفرد

نام دانشجو:

روزبه قاسمی 9531424

تیر 1399

فهرست مطالب

[مقدمه 3](#_Toc46253422)

[مرحله اول؛ ارزیابی داده‌ها 3](#_Toc46253423)

[مرحله دوم؛ استفاده از روش‌های هوش مصنوعی برای پیش‌بینی 16](#_Toc46253424)

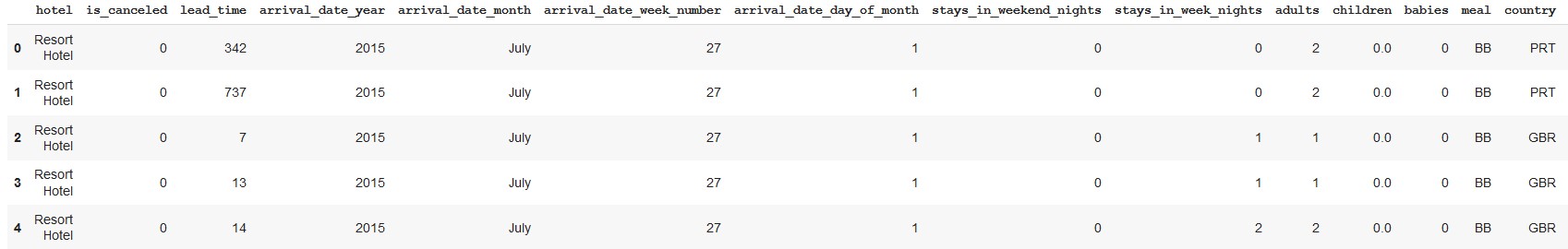
[**دسته‌بندی بیز ساده 17**](#_Toc46253425)

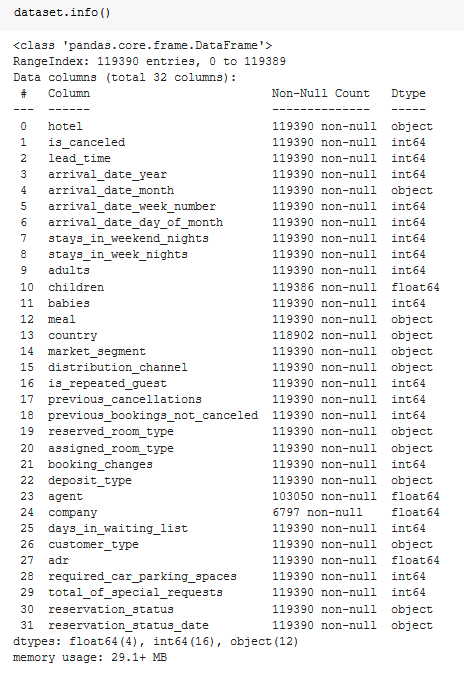
[**دسته‌بندی جنگل تصادفی 19**](#_Toc46253426)

# مقدمه

هدف از انجام این پروژه آشنایی با مفهوم تجزیه و تحلیل داده‌های اکتشافی[[1]](#footnote-1) آشنا شویم. تجزیه و تحلیل داده‌های اکتشافی به مجموعه‌ای از تکنیک‌ها اشاره دارد که در اصل توسط جان توکی[[2]](#footnote-2) استفاده شده‌است تا داده‌ها را به گونه‌ای نمایش دهد که ویژگی‌های جالب آشکار شوند. بر خلاف روش‌های کلاسیک که معمولا با یک مدل فرض شده برای داده‌ها شروع می‌شوند، از تکنیک‌های EDA برای تشویق داده‌ها برای پیشنهاد مدل‌هایی استفاده می‌شود که ممکن است مناسب باشند. این مفهوم به ما کمک می‌کنند تا بتوانیم زمانی که با یک مجموعه‌داده مواجه می‌شویم، چگونه آنرا بررسی و تحلیل کنیم. تحلیل داده‌ها به ما کمک می‌کند تا بتوانیم از هزاران داده و ویژگی موجود در مجموعه‌داده اطلاعات مفیدی را بدست آوریم. در مرحله بعدی ما باید بتوانیم ویژگی‌هایی را از میان ویژگی‌های موجود انتخاب کرده و آن‌ها را جدا کنیم و در نهایت برای پیش‌بینی این که راهی برای پیش‌بینی احتمال لغو رزرو هتل به طور اختصاصی پیدا کنیم.

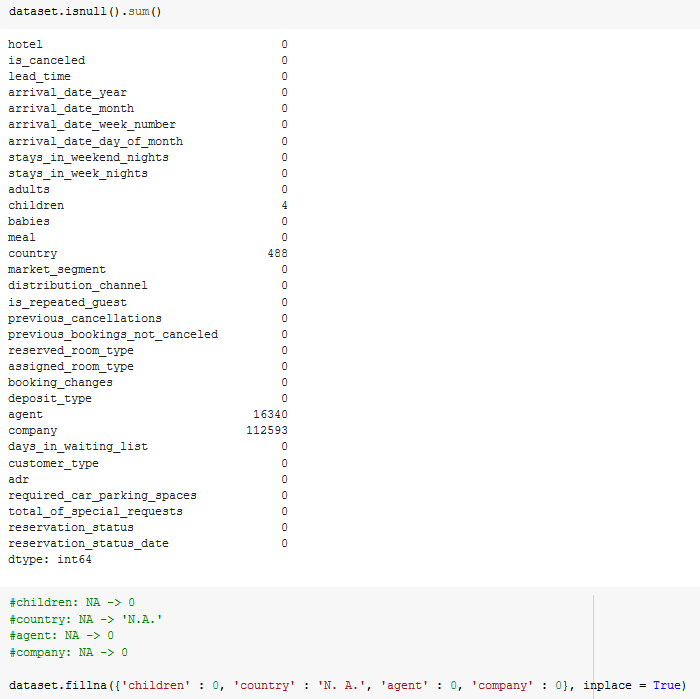
# مرحله اول؛ ارزیابی داده‌ها

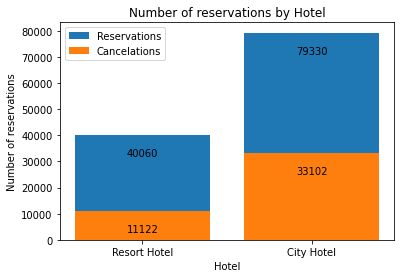
در ابتدا پس از آدرس‌دهی کردن محل ذخیره مجموعه دادن و خواندن آن به کمک کتابخانه Pandas، با دستور Head، 5 رکورد اول آنرا نشان می‌دهیم.

سپس اطلاعات دیتاست را بدست می‌آوریم:

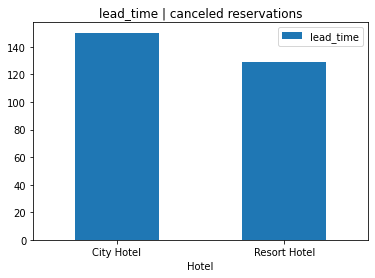
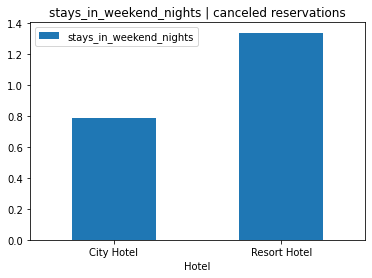
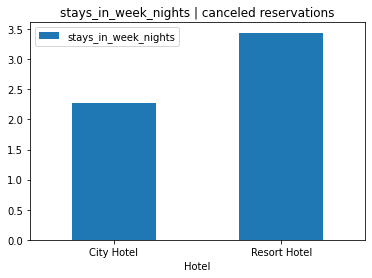
از تابع اطلاعات برای چاپ یک خلاصه مختصر از یک DataFrame استفاده می‌شود. این روش اطلاعات را در مورد یک DataFrame از جمله the شاخص و dtypes ستون، مقادیر غیر صفر و کاربرد حافظه را چاپ می‌کند.

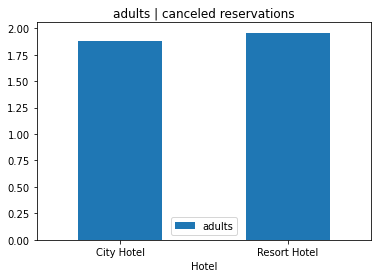
سپس در ادامه، آن ویژگی‌هایی از مجموعه‌داده که نیاز است اصلاح شوند، را اصلاح می‌کنیم.



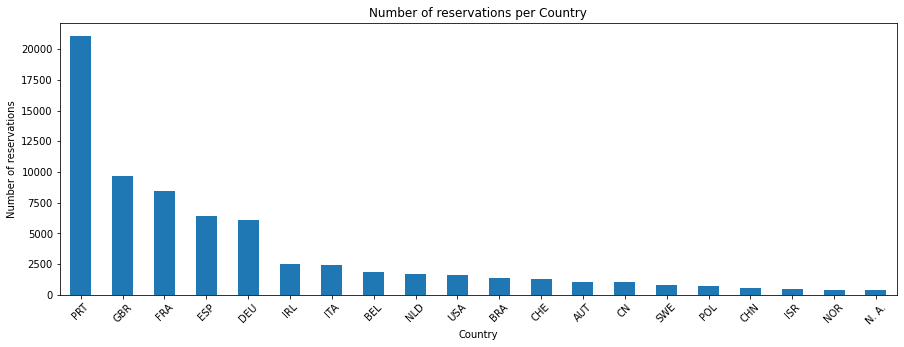
سپس در ادامه، به ترتیب اطلاعاتی را بدست می‌آوریم که برای جلوگیری از طولانی شدن گزارش صرفاً نتایج بدست‌آمده را به صورت خلاصه تحلیل می‌کنیم و کد را دیگر در گزارش نمی‌آوریم.

در شکل بالا، رزروهای صورت گرفته بر اساس نوع هتل‌ها مشخص شده است. در این شکل مشخص می‌کند که چه تعداد از رزرواسیون‌ها کنسل شده است و در کل برای هر نوع هتل، چند رزرواسیون صورت گرفته است.

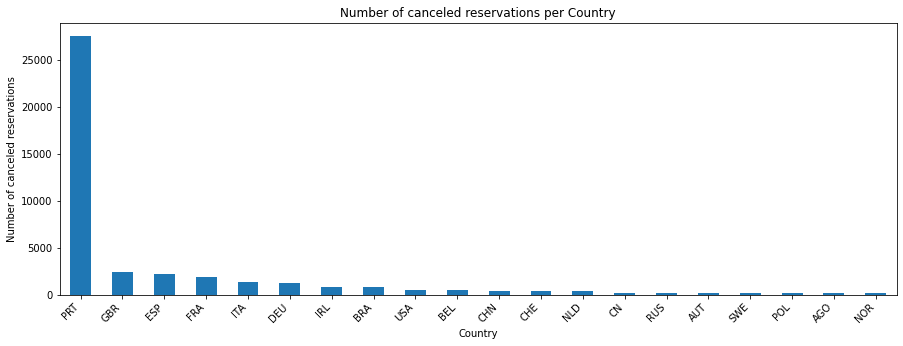
در ادامه، برای پیدا کردن ویژگی‌هایی که بیشترین تاثییر در کنسل کردن یک رزرو هتل دارند را بررسی می‌کنیم.



در ادامه، بر اساس ارزش هر کدام از ویژگی‌ها و بر اساس نوع هتل آن‌ها را لیست می‌کنیم.

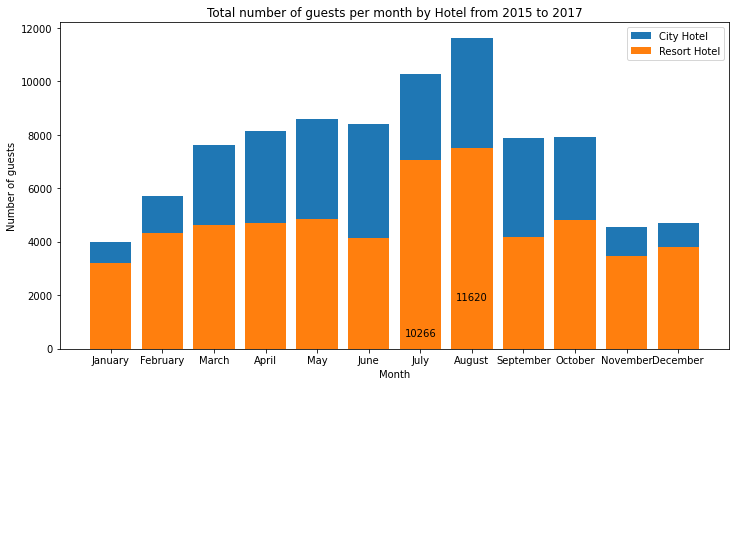
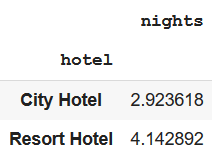
در ادامه، رزروهای صورت گرفته بر اساس کشورها در شکل زیر نمایش داده می‌شود:

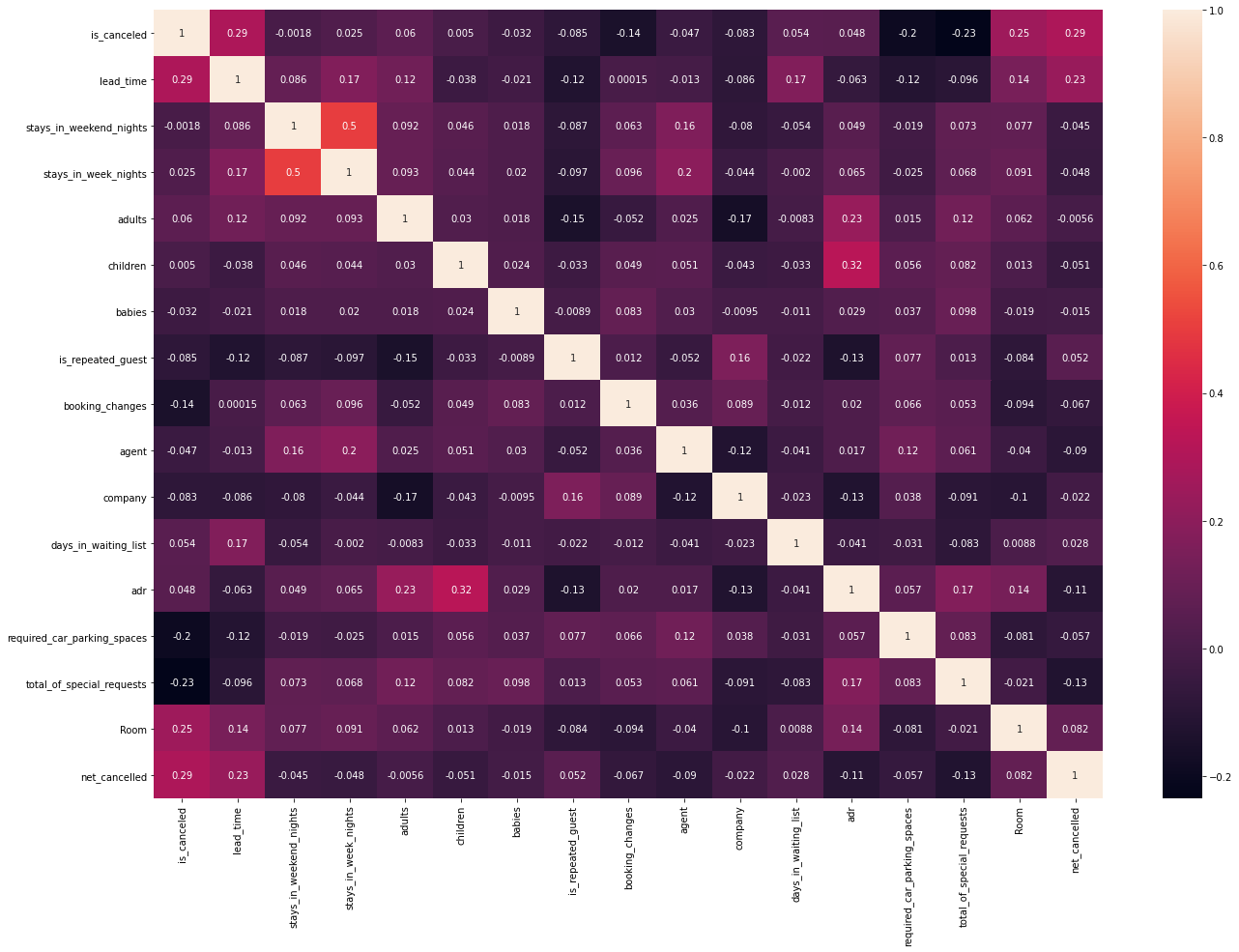
سپس، رزروهای کنسل شده بر اساس کشورها را در شکل زیر مشاهده می‌کنیم:

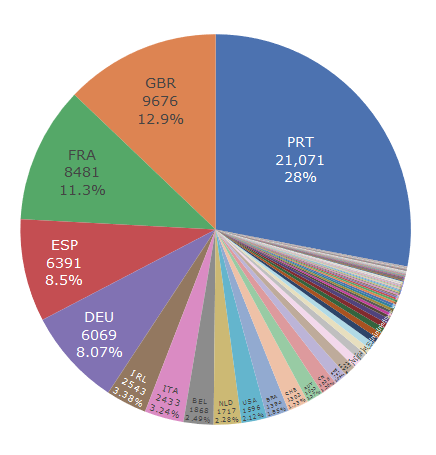


همانطور که در دو شکل بالا مشخص است، کشور پرتغال بیشترین نرخ کنسل کردن رزرواسیون‌ها را دارد. همچنین پرتغال دارای بالاترین نرخ رزرواسیون نیز می‌باشد!

در شکل زیر، تعداد مهمان‌های هر هتل از ابتدای سال 2015 تا ابتدای سال 2017 یعنی تا دسامبر 2016 آورده شده است.

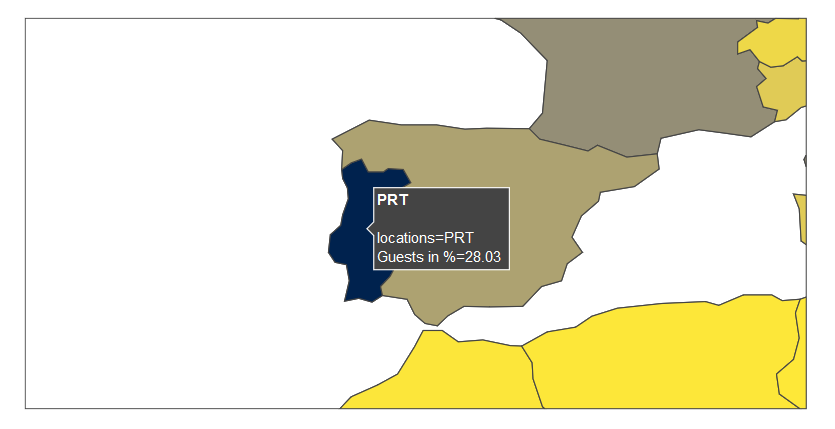
سپس در ادامه نشان دادم که به طور متوسط هر مهمان در هر نوع از هتل‌ها چند شب مهمان بوده است:

سپس در شکل زیر یک heatmap از correlation میان ویژگی‌های این مجموعه‌داده آورده شده است:

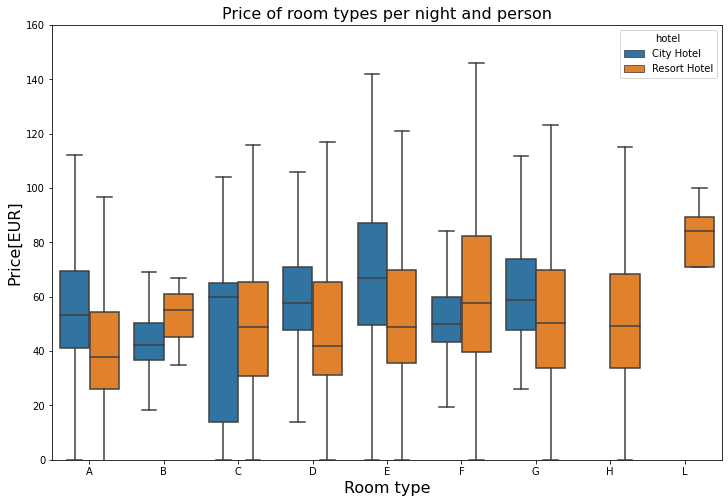
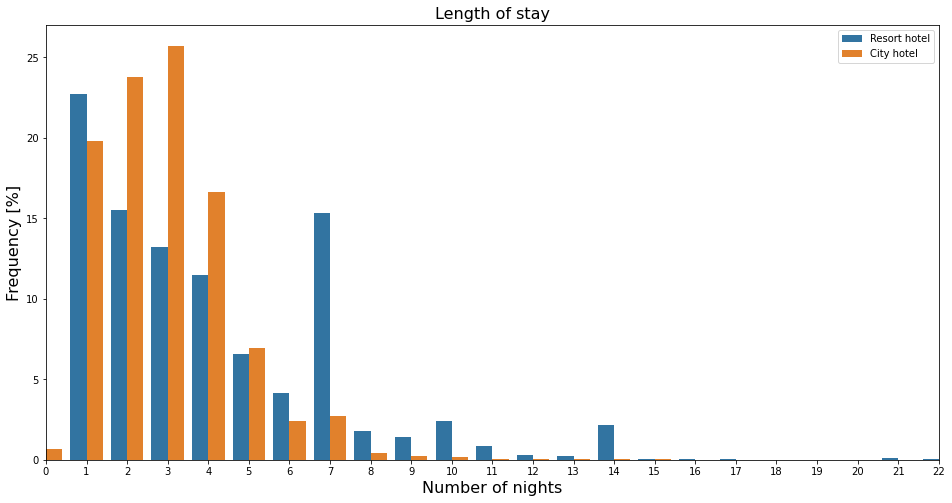
 سپس در شکل زیر، توزیع مهمان‌های هتل‌ها بر اساس کشوری که از آنجا می‌آید آورده شده است:

سپس مشابه شکل بالا، بر روی نقشه جهان نشان می‌دهیم:

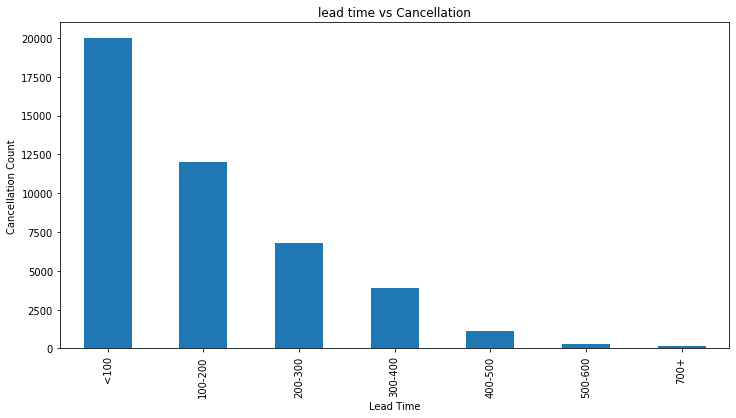
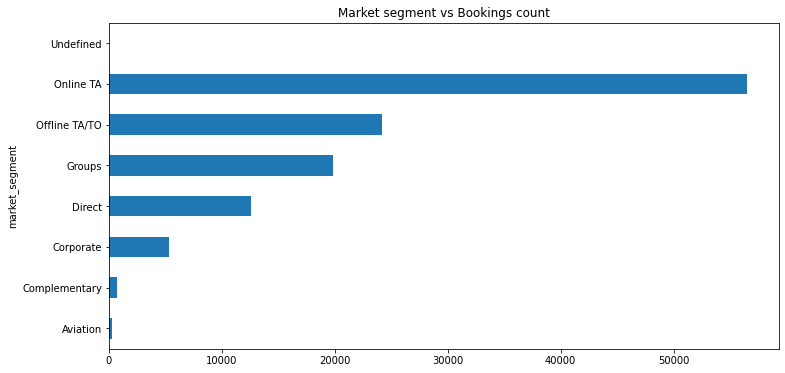


به این صورت که هر چه تعداد مهمان‌ها از آن کشورها بیشتر باشد، رنگ تیره‌‌تری دارد. همانطور که در قبل اشاره شد، کشور پرتغال دارای بیشترین متقاضی رزرواسیون هتل هاست که در شکل زیر نشان داده شده است:

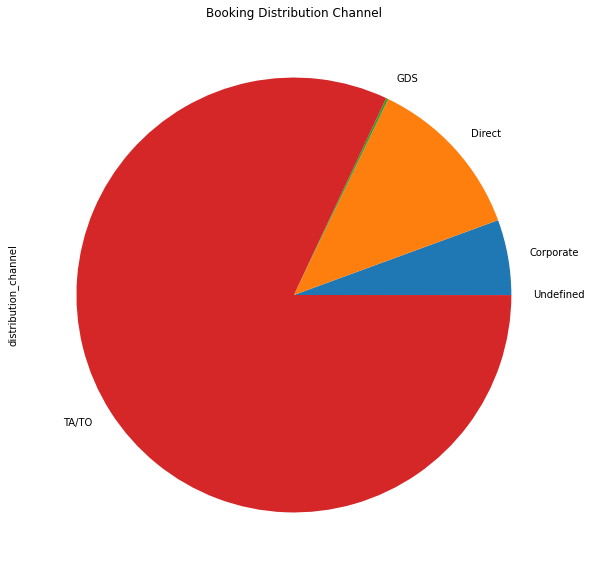
سپس در قسمت زیر، قیمت نوع اتاق‌ها را بر اساس نوع هتل‌ها به یورو نشان می‌دهم:

سپس در شکل زیر، فرکانس طول اقامت مهمان‌ها نشان داده شده است:

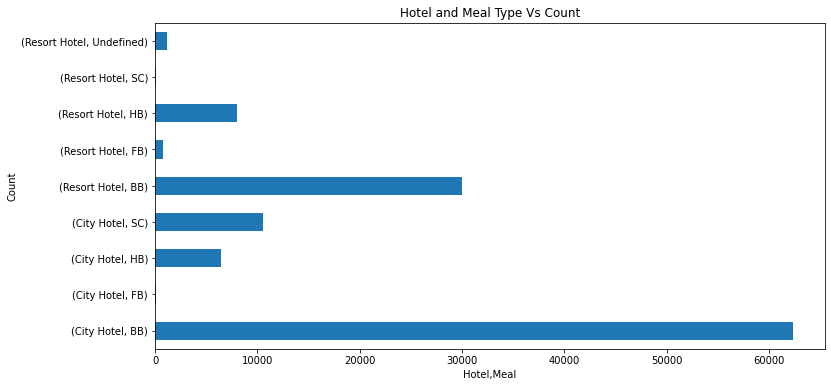
سپس بررسی می‌کنیم که آیا لغو رزرو تحت تأثیر عوامل دیگری است؟

در ادامه، بررسی می‌کنیم که حداکثر تعداد رزروهای انجام شده در کدام بخش بازار است؟

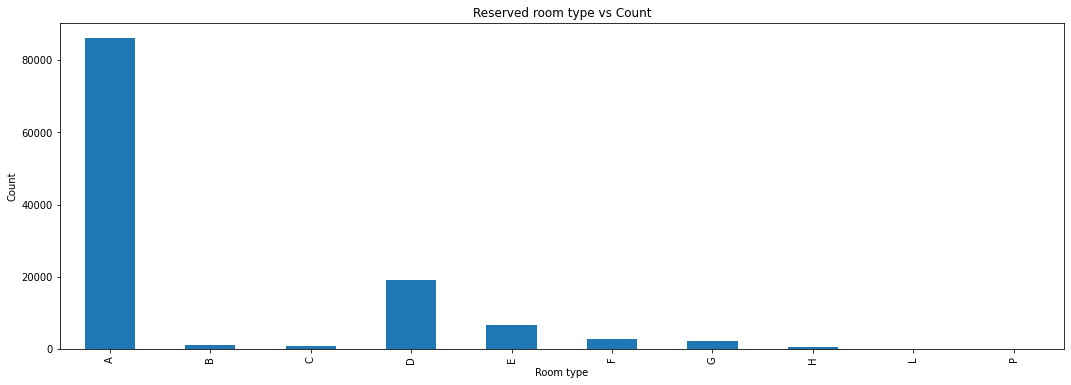
همانطور که مشخص است، بیشترین رزرو از طریق آژانس مسافرتی آنلاین انجام می شود.

در ادامه بررسی می‌کنم که بیشتر از کدام روش برای رزرو هتل استفاده می‌شود؟

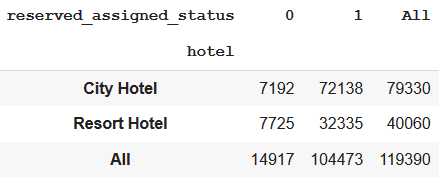
همانطور که مشخص است، بیشتر رزروها از طریق آژانس مسافرتی / اپراتور تور انجام می‌شود.

در ادامه بررسی می‌کنیم که چه نوع غذای combo توسط مردم به شدت ترجیح داده شد؟

اکثر مردم "BB" را در هر دو نوع هتل ترجیح می‌دهند. همچنین افراد کمی ترجیح می‌دهند "بسته غذا" را در هتل Resort ترجیح دهند. بنابراین اغلب آن‌ها غذا را در هتل Resort ترجیح می‌دهند.

همچنین بررسی می‌کنم که در زمان رزرو کدام نوع اتاق را افراد ترجیح می دهند؟

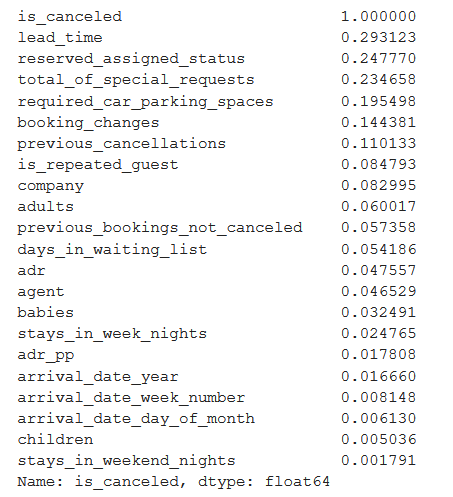
همانطور که مشخص است نوع اتاق "A" بیشترین متقاضی را دارد.

در آخر بررسی می‌کنم که آیا در زمان ثبت نام در اتاق، به همه اتاقهای مشابه اتاق انتخاب شده توسط آنها اختصاص داده شده است؟

همانطور که از نتایج مشخص است، در City Hotel، ۹ % از bookings در زمان چک کردن به همان نوع اتاق نمی‌رسیدند. در Resort Hotel ، ۱۹ درصد از bookings در زمان چک کردن به همان نوع اتاق نمی‌رسیدند. این هتل‌ها نیاز به برداشتن گام‌های لازم برای تخصیص نوع اتاق ارجح به مشتریان دارند چون از پریشانی غیر ضروری در بین مشتریان اجتناب می‌کنند و همچنین به آن‌ها کمک می‌کند تا اعتماد مشتری را به دست آورند.

# مرحله دوم؛ استفاده از روش‌های هوش مصنوعی برای پیش‌بینی

همانطور که در مرحله اول بدست آوردیم، نرخ لغو برای رزرو هتل در صنعت رزرواسیون آنلاین هتل‌ها بسیار بالا است. هنگامی که رزرو لغو شد، تقریبا هیچ کاری نمی‌شود کرد. این باعث ناراحتی بسیاری از هتل‌ها می‌شود و اشتیاق به انجام اقدامات احتیاطی را ایجاد می‌کند. بنابراین، پیش‌بینی reservations که می‌تواند لغو شود و جلوگیری از این حذف لغو شود یک مقدار اضافی برای هتل‌ها ایجاد خواهد کرد. لغو رزرو کردن تاثیر قابل‌توجهی بر تصمیمات مدیریت تقاضا در صنعت مهمان‌نوازی دارد. برای کاهش تاثیر حذف، هتل‌ها سیاست‌های لغو سخت و تاکتیک‌های overbooking را اجرا می‌کنند که به نوبه خود می‌تواند تاثیر منفی بر درآمد و شهرت هتل داشته باشد. برای کاهش این اثر، یک نمونه اولیه سیستم مبتنی بر یادگیری ماشین توسعه یافت. این سیستم از داده‌های سیستم‌های مدیریت املاک هتل استفاده می‌کند و هر روز یک مدل طبقه‌بندی را آموزش می‌دهد تا پیش‌بینی کند که رزرو شده "به احتمال زیاد" و با محاسبه تقاضای خالص است. این نمونه اولیه که در محیط تولید در دو هتل مستقر شده‌است، با اجرای آزمایش A/B، اندازه‌گیری تاثیر اقدامات انجام‌شده بر روی bookings را به احتمال زیاد لغو می‌کند. نتایج نشان می‌دهند که عملکرد نمونه اولیه خوب است و نشانه‌های مهمی برای پیشرفت تحقیقات ارایه می‌دهد در حالی که این است که bookings با هتل‌ها تماس گرفته تا با هواپیما تماس گرفته نشده تا با آن‌ها تماس گرفته نشده باشد.

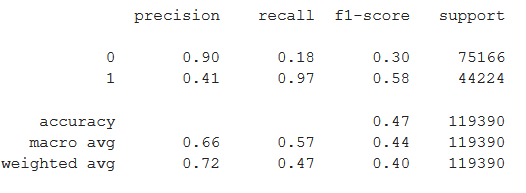
حال برای انجام این مرحله، ابتدا باید از جدول Heatmap داده شده، ویژگی‌هایی که تاثییر بیشتری داشت را جدا می‌کنیم. (5 ویژگی برتر)

سپس برای این پروژه من از دو روش استفاده کردم:

## دسته‌بندی بیز ساده

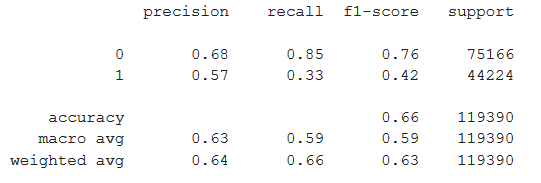
در حوزه یادگیری ماشین تکنیک و روش دسته‌ بند بیز ساده، با بکارگیری قضیه بیز و فرض استقلال بین متغیرها، به عنوان عضوی از خانواده دسته‌بندهای برمبنای احتمال (Probabilistic Classifiers) قرار می‌گیرد. در سال‌های ۱۹۶۰ تحقیق و بررسی‌های زیادی پیرامون بیز ساده بخصوص در زمینه بازیابی متن صورت گرفت و حتی امروز هم به عنوان ابزاری برای دسته‌بندی متن (Text Categorization) برای حل مسائلی مانند تشخیص هرزنامه‌ها (Spam Mails) به کار می‌رود. معمولا این کار به کمک برآورد تابع احتمال و از طریق فراوانی یا فراوانی نسبی کلمات در اسناد متنی صورت می‌گیرد. به این ترتیب به کمک حداکثرسازی تابع درست‌نمایی (Likelihood maximization)برآورد پارامترهای مدل میسر می‌شود. در حوزه آمار و دانش رایانه، مدل بیز ساده با نام‌های دیگری نظیر بیز ساده (Simple Bayes) و بیز مستقل نیز شناخته می‌شود که در بسیاری از حوزه‌های دیگر نیز کاربرد دارد. به منظور آشنایی خوانندگان با این زمینه کاربردی بخصوص برای افرادی که به تازگی خواستار ورود به گروه متخصصان علم داده‌ هستند، متن حاضر را تهیه کرده‌ایم.

برای تقسیم‌بندی داده‌های train و تست از k-fold cross validation استفاده کردم و k = 10 را انتخاب کردم به این صورت که هر بار 9 بخش را به عنوان داده train و 1 بخش را به عنوان داده test در نظر می‌گیریم. سپس error میانگین را به دست آورده و accuracy را بدست می‌آوریم.

دقت بدست‌آمده برابر است با:

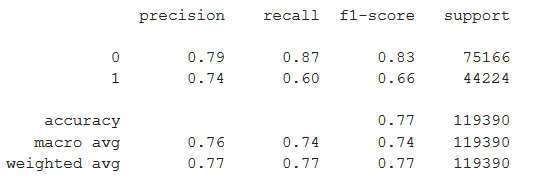
دقت در این مرحله بدست آمده برابر با 47 درصد است.

حال در همین روش، همانطور که در توضیح پروژه خواسته شده بود، از الگورتیم PCA برای پیدا کردن ویژگی‌هایی که بیشترین تاثییر دارد را به صورت یک الگوریتم feature extraction based انتخاب شده است و یک سری ویژگی جدید در واقع تولید می‌شود.

نتایج الگوریتم بیزساده بر اساس استفاده از PCA به شکل زیر است:

همانطور که مشخص است، دقت بهبود پیدا کرد! این نشان از عملکرد قوی الگوریتم‌های کاهش ابعاد مثل PCA هست.

## دسته‌بندی جنگل تصادفی

الگوریتم جنگل تصادفی (Random Forest) یک الگوریتم یادگیری ماشین با قابلیت استفاده آسان است که اغلب اوقات نتایج بسیار خوبی را حتی بدون تنظیم فراپارامترهای آن، فراهم می‌کند. این الگوریتم به دلیل سادگی و قابلیت استفاده، هم برای دسته‌بندی (Classification) و هم رگرسیون، یکی از پر کاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌شود. در این مطلب، چگونگی عملکرد جنگل تصادفی و دیگر مباحث مهم پیرامون آن مورد بررسی قرار خواهند گرفت. برای درک چگونگی عملکرد جنگل تصادفی، ابتدا باید الگوریتم درخت تصمیم (Decision Tree) که بلوک سازنده جنگل تصادفی است را آموخت. انسان‌ها همه روزه از درخت تصمیم برای تصمیم‌گیری‌ها و انتخاب‌های خود استفاده می‌کنند، حتی اگر ندانند آنچه که از آن بهره می‌برند نوعی الگوریتم یادگیری ماشین است.

همانطور که در شکل بالا مشخص است، دقت بیشتری نسبت به بیز ساده بدست آورد که بر اساس قواعد یادگیری ماشین این نتیجه قابل انتظار بود.

پایان

1. Exploratory Data Analysis [↑](#footnote-ref-1)
2. John Tukey [↑](#footnote-ref-2)