گزارش کار پروژه پایتون

تحلیل اطلاعات پرواز

آرین احدی‌نیا

رکسانا خباززاده مقدم

فربد عصاره

استاد: استاد علیرضا کدیور

راهنما: آقای امیرحسن فتحی

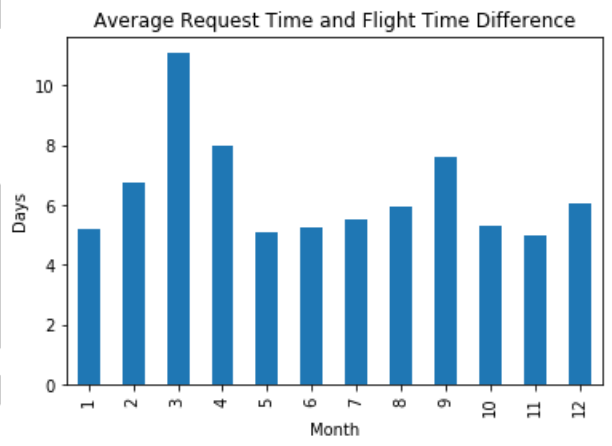
نیم‌سال اول ۹۸-۹۹

فهرست مطالب

مقدمه

در این گزارش به بررسی روند انجام پروژه از ابتدا تا انتها و نتایج به دست آمده از آن پرداخته می‌شود. در بخش اول، روش تقسیم کار بیان می‌شود. سپس جدول داده تجمیع شده، ستون‌های آن، نوع داده‌ها در هر قسمت جدول و کارکردشان تشریح می‌شود. در ادامه درباره‌ی کدهای هر خواسته، ورودی و خروجی آن‌ها و کاری که روی جدول انجام می‌دهند به طور مختصر توضیح داده می‌شود. بعد، پویایی و قابلیت به‌روز رسانی کد و تدابیر اتخاذ شده برای سهولت در ایجاد تغییرات دلخواه کاربر بررسی می‌شود. بخش بعدی گزارش به بررسی مختصری از داده تجمیع‌شده و مشخصات آن مانند تعداد رکوردها، بازه زمانی داده‌های گزارش، مجموع درآمدها و... اختصاص دارد. بعد از آن، درباره‌ی اهمیت، معیارها و کارهایی که برای کیفیت‌سنجی داده‌ها انجام شده، بحث می‌شود. در قسمت گزارش تحلیلی، خروجی‌ها و تحلیل‌ها با استفاده از نمودار و شکل با دقت بیش‌تری و به طرز ملموس‌تری بیان می‌شوند. نهایتا در قسمت آخر، به پرسش جدیدی که در مسیر انجام پروژه مطرح شد، پاسخ داده می‌شود.

در دنیای امروز، داده‌ها ارزشی کم‌تر از گنج ندارند. تحلیل داده برای هر شرکت یا سازمان، صرف نظر از کالا یا خدماتی که ارائه می‌کنند امری بسیار حیاتی است و می‌تواند عامل موفقیت یا شکست باشد. برای مثال، برای یک شرکت هواپیمایی بسیار مهم است که بداند فروشش به چه عواملی بستگی دارد؟ آیا فروش به ماه، روز هفته و یا حتی ساعت روز ارتباطی دارد؟ یا شاید هم این مبدا یا مقصد است که تعیین کننده است؟ آیا سرویس تامین‌کننده‌ای که مسئول فروش بلیط‌ها و پل ارتباطی شرکت و مشتریانش است، قابل اعتماد است؟ یا این که شاید بهتر باشد سرویس دیگری برای این کار انتخاب شود؟ اصلا شاید ترکیبی از همه‌ی این عوامل باشد، مگر نه؟



این نمودار اختلاف زمانی میان تاریخ خرید بلیط و پرواز را نشان می‌دهد که می‌تواند معیار خوبی از سطح اعتماد مشتریان به سرویس تامین‌کننده باشد

گزارش تقسیم کار

جدول زیر، تقسیم کار میان اعضای گروه را نشان می‌دهد:

کیفیت‌سنجی	رکسانا خباززاده مقدم-آرین احدی‌نیا
خواسته ۱	فربد عصاره
خواسته ۲	فربد عصاره
خواسته ۳	آرین احدی‌نیا- فربد عصاره
خواسته ۴	رکسانا خباززاده مقدم
خواسته ۵	آرین احدی‌نیا
خواسته ۶	رکسانا خباززاده مقدم
ماژولار کردن کدها	آرین احدی‌نیا
تهیه گزارش	رکسانا خباززاده مقدم
تهیه پاورپوینت	فربد عصاره-آرین احدی‌نیا

توصیف داده

نام ستون	type	نام ستون	type
id	int	destination	int
user_id	int	price	float
request_date_id	int	request_date	datetime.date
request_time	datetime.time	departure_date	int
departure_date_id	int	difference	int
departure_time	datetime.time	departure_month	int
company	int	request_month	int
source	int	departure_weekday	int
request_weekday	int	departure_hour	int
request_hour	int	reward	int

توضیح کدهای پایتون

فایل داده‌های دریافتی در ابتدا کیفیت‌سنجی می‌شود و از خروجی کد کیفیت‌سنجی(که شامل داده‌ها به شکل مطلوب و اصلاح شده است) که یک فایل csv. است، به عنوان ورودی تمامی بخش‌های دیگر استفاده می‌شود. در تمامی ماژول‌ها توابعی وجود دارند که کارشان رسم نمودارهای مربوطه است که در ادامه‌ی گزارش آمده‌اند.

ماژول برنامه پرپرواز(frequent_flyer_program):

reward

این تابع امتیاز هر بلیط را محاسبه می‌کند و به عنوان یک ستون به دیتافریم اصلی اضافه می‌کند.

customer_reward

این تابع امتیاز هر مشتری را با توجه به مجموع امتیازهای دریافت شده محاسبه می‌کند.

customer_club

این تابع مشتریان را در چهار سطح مختلف باشگاه مشتریان قرار می‌دهد. ورودی این تابع دیتافریم اصلی و سه حد نصاب برای سطوح مختلف باشگاه است.

chart

این تابع توزیع تعداد مشتریان را در هر سطح نمایش می‌دهد.

purchase_user_id

میزان خرید هر مشتری را محاسبه می‌کند.

count_user_id

تعداد خریدهای هر مشتری را محاسبه می‌کند.

normal_distribution

نمودار توزیع نرمال خریدهای مشتری‌ها را رسم می‌کند.

club_chart

این تابع نمودار تعداد مشتریان در هر چهار سطح باشگاه را نشان می‌دهد.

ماژول محبوبیت(popularity):

route

این تابع تعداد پروازهای انجام شده در هر دوتایی مبدا-مقصد را محاسبه می‌کند و تحت یک دیتاسری برمی‌گرداند.

source

این تابع تعداد پروازهای انجام شده از هر مبدا را در یک دیتاسری برمی‌گرداند.

destination

این تابع تعداد پروازهای انجام شده به هر مقصد را در یک دیتاسری برمی‌گرداند.

company

این تابع سهم پروازهای هر شرکت از کل را به صورت درصدی محاسبه می‌کند و به عنوان یک دیتاسری برمی‌گرداند.

month

تعداد پروازهای برخاسته در هر ماه را تحت یک دیتاسری برمی‌گرداند.

company_by_month

تعداد پروازهای هر کمپانی در هر ماه را برمی‌گرداند.

company_by_source

تعداد پروازهای هر کمپانی از هر مبدا را برمی‌گرداند.

company_by_destination

تعداد پروازهای هر کمپانی به هر مقصد را برمی‌گرداند.

airport_plot

این تابع نمودار تعداد پروازهای ورودی و خروجی از هر فرودگاه را رسم می‌کند.

ماژول اقتصاد(economics):

highest_grossing_company

مجموع درآمدهای هر شرکت را محاسبه می‌کند.

Plot_pie_data_highest_grossing_company

نمودار دایره‌ای درآمدها کل ایرلاین‌ها را رسم می‌کند.

Plot_bar_data_highest_grossing_company

نمودار میله‌ای درآمد کل ایرلاین‌ها را رسم می‌کند.

plot_data_highest_sharing_company

نمودار سهم شرکت‌ها از کل پروازها را رسم می‌کند.

ماژول وابستگی(dependence_analysis):

weekday_sell

تعداد درخواست خرید در روزهای مختلف را در یک دیناسری ارائه می‌دهد.

weekday_departure

تعداد پروازهای خروجی در روزهای مختلف را در یک دیناسری ارائه می‌دهد.

hour_sell

تعداد درخواست خرید در ساعات مختلف روز را در یک دیناسری ارائه می‌دهد.

hour_departure

تعداد پروازهای خروجی در ساعات مختلف روز را در یک دیتاسری ارائه می‌دهد.

month_sell

تعداد درخواست خرید در ماه‌های مختلف سال را در یک دیناسری ارائه می‌دهد.

ماژول اختلاف(difference analysis):

month

این تابع مقدار میانگین اختلافات در هر ماه را بدست میآورد.

company

در این تابع مقدار میانگین اختلافات برای هر کمپانی محاسبه می‌شود و نمودار میانگین اختلافات را بر حسب هر کمپانی رسم می‌کند.

plot_company

این تابع نمودار میانگین اختلافات را در هر ماه رسم می‌کند.

قابلیت به‌روزرسانی

پویایی کد، عامل بسیار مهمی در کیفیت و کارایی آن است.

یکی از عواملی که پویایی کد کمک شایانی می‌کند، استفاده از ماژول‌ها و توابع است. این شیوه‌ی نوشتن منسجم‌تر و مرتب‌تر است که باعث می‌شود کد برای کاربر قابل‌فهم‌تر شود، با آن احساس راحتی بیش‌تری کند و سردرگم نشود. به همین دلیل؛ تمامی کدها به صورت ماژولار و توابع نوشته‌شده و سعی‌شده حتی‌المکان کد هر بخش مختصر و خوانا باشد. برای مثال؛ اسم توابع و متغیرها با توجه به کاربردشان انتخاب شده‌است تا کاربر صرفا با دیدن نام‌ها بتواند بفهمد هر مورد چه کاری را انجام می‌دهد. علاوه بر این، با کامنت‌گذاری، خطوطی که اعمال مشابهی دارند توضیح داده‌شده‌اند تا کاربر بتواند به راحتی مطابق میل و نیازش آن را تغییر دهد.

همچنین، تلاش شده تا هر تابع صرفا یک کار مشخص را انجام دهد. این باعث می‌شود خروجی هر کد قابل فهم و پیش‌بینی باشد . بنابراین اگر مشکلی پیش بیاید، راحت‌تر می‌توان

علاوه بر موارد بالا، ماژولار بودن کدها مزایای دیگری هم دارد؛ برای مثال اگر نیاز جدیدی پیش بیاید که بخشی از آن با کدهای کنونی قابل حل باشد، می‌توان ماژول مربوط به آن بخش را صدا و از آن استفاده کرد. به این صورت از دوباره‌کاری و اشغال بی‌مورد حافظه پرهیز می‌شود و امکان پاسخ(هر چند در حد نسبی) به برخی پرسش‌های جدید با استفاده از ماژول‌های قبلی بدون نیاز به کدنویسی دوباره و یا استخدام برنامه‌نویس پاسخ داد!

گزارش توصیفی از اطلاعات تجمیع کلان

	داده خام	داده پس از کیفیت‌سنجی
تعداد رکوردها	۲۱۱۷۷۶	۱۹۸۸۹۲
بازه زمانی	۲۰۱۶/۳/۲۱ تا ۲۰۱۸/۳/۲۰	۲۰۱۶/۳/۲۱ تا ۲۰۱۸/۳/۲۰
کل فروش(واحد پولی)	۱۶۵۶۷۳۱۷۲,۵	۱۵۷۲۵۲۰۵۴,۶
تعداد فرودگاه‌ها	۲۸	۲۸
تعداد شرکت‌ها	۱۳	۱۳
کل آی‌دی‌ها	۴۳۲۶۰	۴۱۸۱۲
حداقل قیمت	۶۵,۹۳	۶۵,۹۳
حداکثر قیمت	۹۴۴۳	۹۴۴۳
حداقل اختلاف زمانی	-۴۹	۰
حداکثر اختلاف زمانی	۲۱۲	۲۱۲

کیفیت‌سنجی

ماژول استانداردسازی داده‌ها(data_standardisation):

کیفیت‌سنجی داده به بررسی حالات کیفی و کمی اطلاعات می‌پردازد. تعاریف زیادی برای کیفیت داده وجود دارد اما به طور کلی می‌توان گفت یک مجموعه داده وقتی از کیفیت بالایی برخوردار است که برای استفاده‌های موردنظر، تصمیم‌گیری‌ها و برنامه‌ریزی مناسب باشد. علاوه بر این، یک کد زمانی باکیفیت است که بتواند مسئله را در دنیای واقعی به درستی توصیف و پیش‌بینی کند. جدا از این‌ها، زمانی که تعداد منابع داده زیاد می‌شود، مسئله‌ی جدیدی (با صرف نظر از مناسب بودن داده برای هر گونه استفاده و تحلیل) مطرح می‌شود: آیا داده با ثبات است؟ آیا از انسجام درونی برخوردار است؟

اگر چه کیفیت داده در نظر افراد مختلف می‌تواند تعاریف گوناگونی داشته باشد، اما به طور کلی می‌توان معیارهای زیر را برای کیفیت داده در نظر گرفت:

دقت انسجام درونی به‌روز بودن کامل بودن مرتبط بودن به نیاز

برآورده شدن برخی از این معیارها (برای مثال در این پروژه) مانند به‌روز بودن و انسجام درونی، وابسته به سازمان گردآورنده‌ی این داده‌هاست و نمی‌توان آن‌ها را با کد زدن کنترل کرد.

کد کیفیت‌سنجی از سه بخش تشکیل شده‌است:

۱- کنترل کامل بودن: در این بخش تمامی ستونها بررسی می‌شوند و اگر در ستونی صفر یا null باشد، کل آن ردیف حذف می‌شود. پس از اتمام این بخش، ردیف‌های دیتافریم دوباره شماره‌گذاری(indexing) می‌شوند.

۲- بررسی و در صورت نیاز، اصلاح نوع ورودی ستون‌ها، یکسان کردن فرمت ساعت‌های درخواست و پرواز و تاریخ‌ها.

۳- انجام محاسبات و افزودن ستون‌هایی که در پاسخ دادن به خواسته‌ها به آن‌ها نیاز خواهد شد:

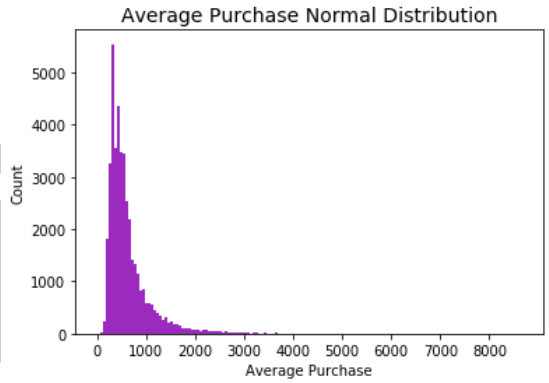
ماه(درخواست و خرید) روز هفته(درخواست و خرید) ساعت(درخواست و خرید)

اختلاف زمانی بین تاریخ‌های درخواست و خرید بر اساس روز

پس از اعمال تمامی این تغییرات، دیتافریم اصلاح شده با دو فرمت .csv و .xlsx ذخیره می‌شوند.

گزارش تحلیلی

- خواسته ۱

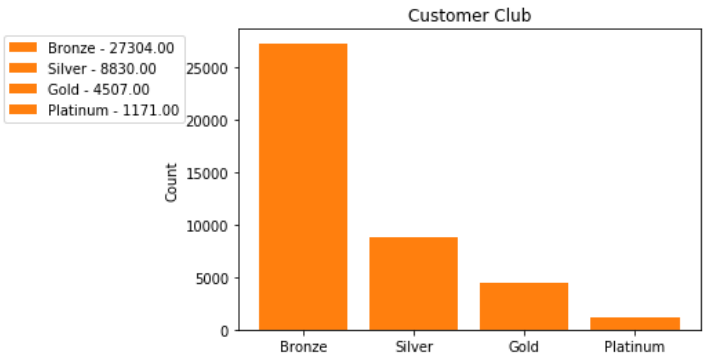


این نمودار، توزیع نرمال میانگین خریدهای مشتریان را نشان می‌دهد. با توجه به این نمودار، غالبا مشتریان حدود ۳۰۰-۴۰۰ خرید می‌کنند. علاوه بر این، به طور میانگین هر نفر ۳۷۶۰ خرج کرده‌است. دلیل این اختلاف بالا تعدادی داده است که با وجود کم بودن تعدادشان، مقدارشان چنان زیاد است که تاثیر زیادی در میانگین اعمال می‌کنند. بنابراین بررسی توزیع خریدها می‌تواند ابزار سنجش مناسب‌تری باشد.

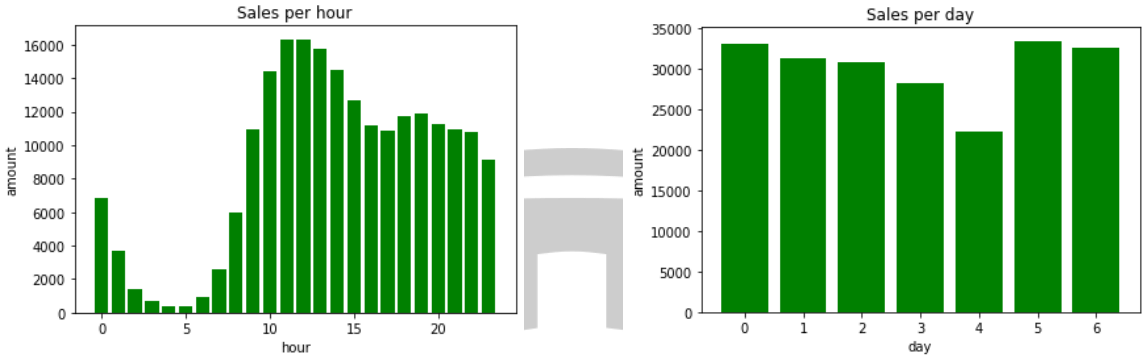
- خواسته ۲

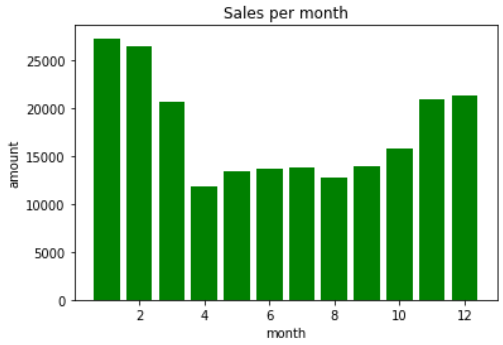
برای امتیازدهی به مشتریان، یک رابطه‌ی خطی در نظر گرفته شد: خارج قسمت تقسیم مجموع خریدها بر ۱۰۰ در ۵۰۰ ضرب شد و در ستونی جدید در دیتافریم ذخیره شد.

سپس با توجه به میانگین، ماکسیمم و مینیمم امتیازات مشتریان، سه عدد ۱۰۰۰۰ و ۳۰۰۰۰ و ۱۰۰۰۰۰ به عنوان معیار دسته‌بندی مشتریان در سه دسته‌ی `Silver, Gold & Platinum` در نظر گرفته شدند.



- خواسته ۳





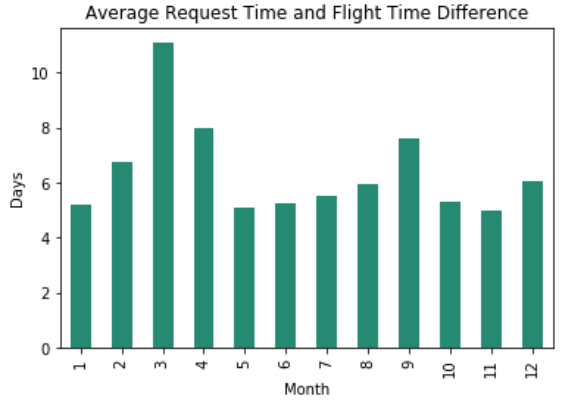
با توجه به این سه نمودار، پرطرفدارترین ماه(میلادی)، روز و ساعت برای خرید بلیط به ترتیب ژانویه، شنبه و ساعت ۱۱ و ۱۲ هستند.

فروش در ساعات مختلف روز با سبک زندگی غالب بر جامعه ارتباط تنگاتنگی دارد. برای مثال از حدود ساعت ۱ بامداد تا ۷ صبح که با ساعت خواب اکثریت جامعه هماهنگ است، فروش کمی دیده می‌شود. بعد، از حدود ساعت ۹ که فعالیت‌های روز آغاز می‌شوند، فروش به شدت افزایش می‌یابد. فروش در ساعات بعد از ظهر کاهش می‌یابد و در ساعات عصر و شب میزان فروش کمابیش ثابت می‌ماند.

از شنبه تا جمعه، فروش روزانه کاهش می‌یابد. برای اطمینان از تطبیق نمودار فروش روزانه باید اطلاعاتی درباره‌ی محل جمع‌آوری این داده‌ها داده‌شود؛ در غیر این صورت نمی‌توان نظری داد.

درباره‌ی ماه هم می‌توان گفت که به احتمال زیاد مناسبتی مهم در ماه‌های سه و چهار رخ می‌دهد که باعث می‌شود مردم از پیش بلیط‌های خود خریداری کنند.

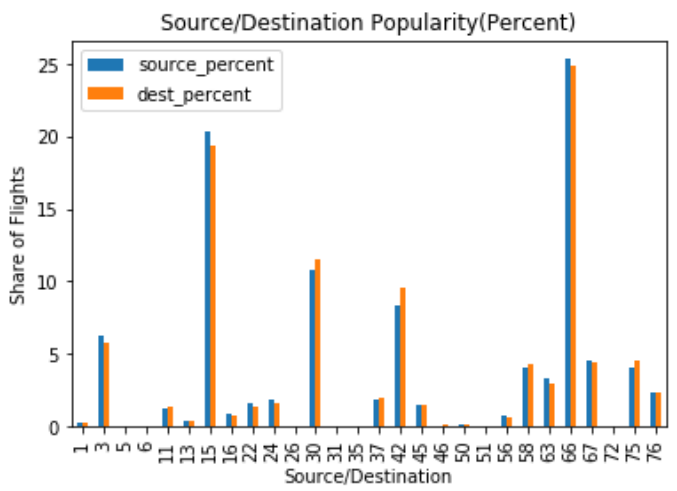
- خواسته ۴



بررسی اختلاف زمانی میان خرید بلیط و تاریخ پرواز می‌تواند معیار خوبی برای بررسی میزان اعتماد مشتریان به سرویس تامین‌کننده باشد. با توجه به این که به طور میانگین، خریداران ۶ روز قبل از پرواز بلیط‌هایشان را تهیه می‌کنند که مقدار معقولی است(اگر چه باید در مقایسه با تامین‌کننده‌های دیگر اظهار نظر کرد)، می‌توانیم بگوییم که خریداران اعتماد خوبی به این سرویس دارند و ادامه دادن همکاری میان این سرویس و شرکت‌های هواپیمایی سودمند است.

- خواسته ۵

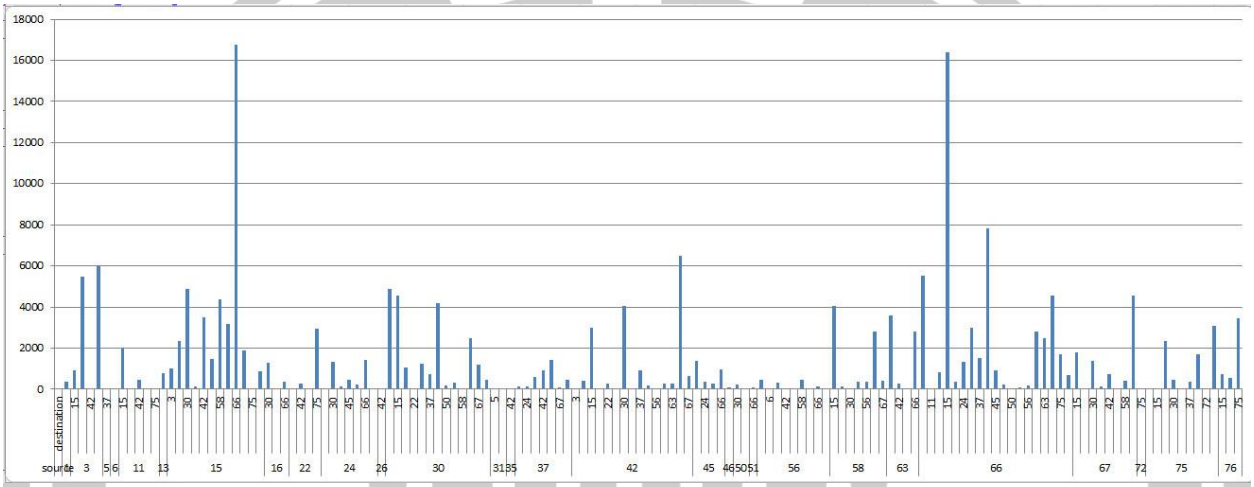
در نمودار زیر محور افقی نام فرودگاه ها و محور عمودی درصد پرواز هایی است که از آن فرودگاه به عنوان مبداء یا مقصد انجام شده است. منطقا باید تعداد پرواز های ورودی و خروجی از فرودگاه ها برابر باشد. چرا که با هر هواپیمای ورودی، پروازی خروجی انجام خواهد شد. همان طور که مشاهده می شود طول ستون های آبی و نارنجی باید برابر باشد. توجه کنید که اختلاف اندکی که مشاهده می شود به دلیل ناقص بودن تعداد اطلاعات است.



دسته بندی بر اساس مبداء و مقصد انجام شده است و تعداد پرواز های انجام شده در هر مسیر که دوتایی مبدا و مقصد است، محاسبه شده است. بر اساس این داده و داده ای که بر اساس دسته بندی پرواز ها بر مبنای مبداء و مقصد صورت گرفته است، تحلیل انجام شده است.

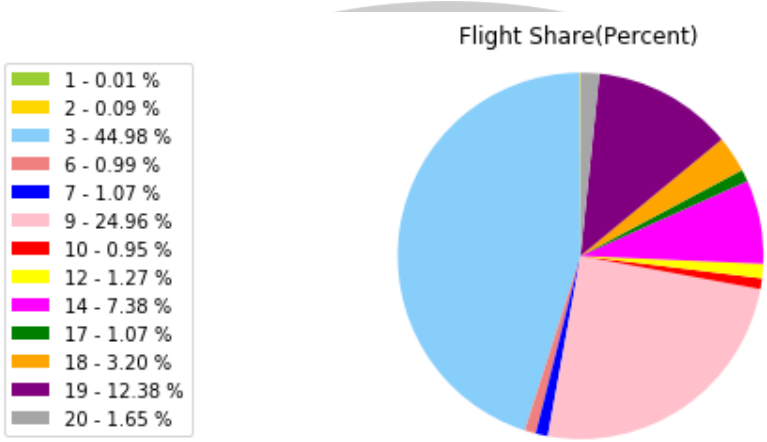
تعداد پرواز های ورودی و خروجی از فرودگاه های ۶۶ و ۱۵ با انحراف قابل توجهی، بیشتر از سایر فرودگاه‌ها است. بنابرین قابل توجیه است که پرواز های رفت و برگشت از این دو فرودگاه تعداد بیشتری نسبت به سایر پرواز ها داشته باشد.

همچنین بین ۵ فرودگاه پرتردد، ۲۰ مسیر وجود دارد که ۱۰ تا از این مسیر ها دقیقا ۱۰ مسیر پرتردد می باشند.

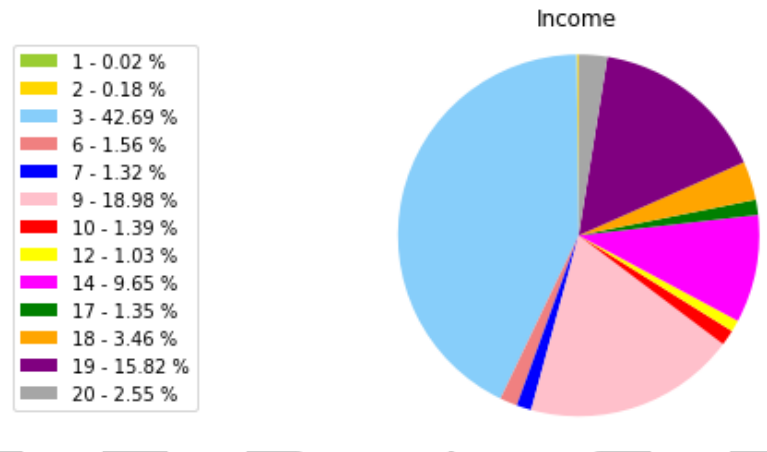


- خواسته ۶

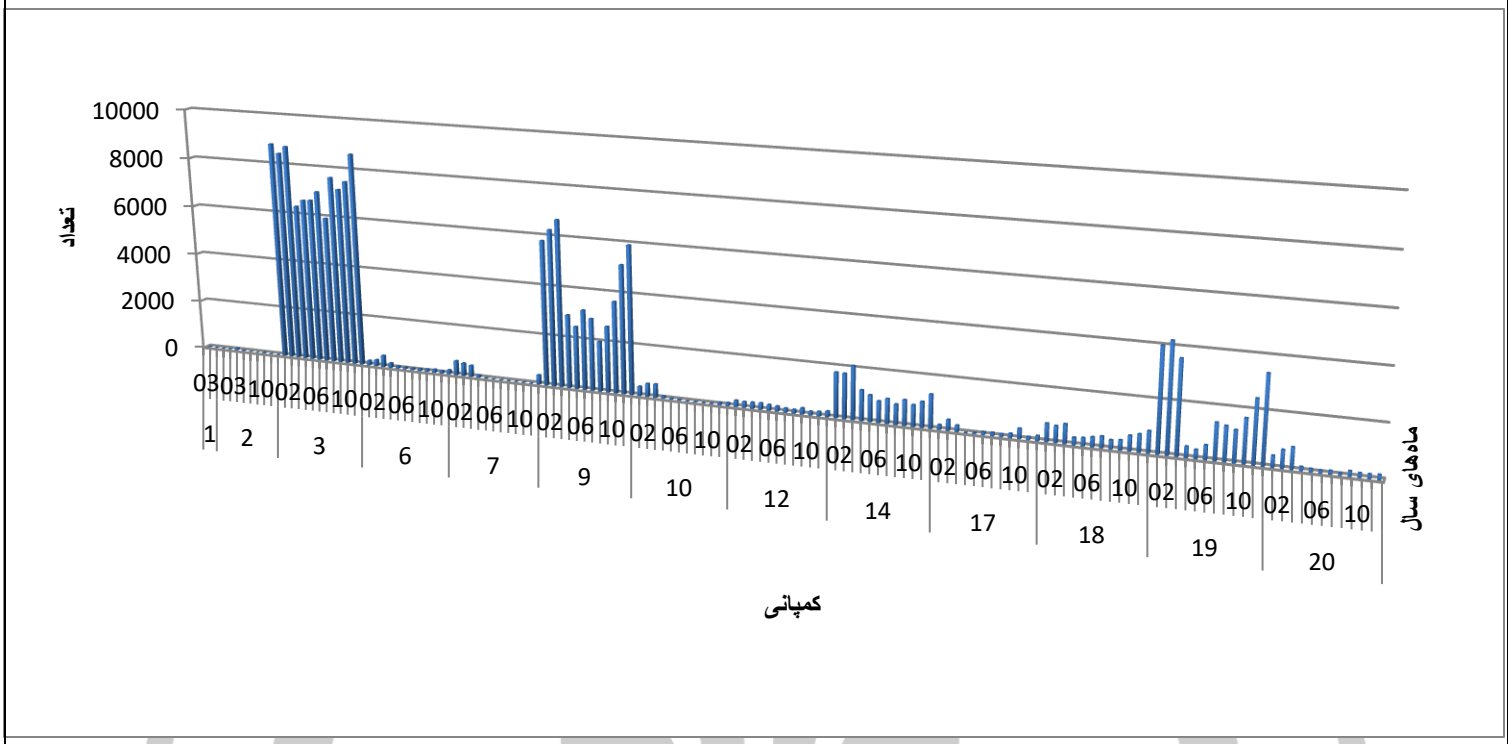
سهم هر کمپانی از کل پروازها(درصد):



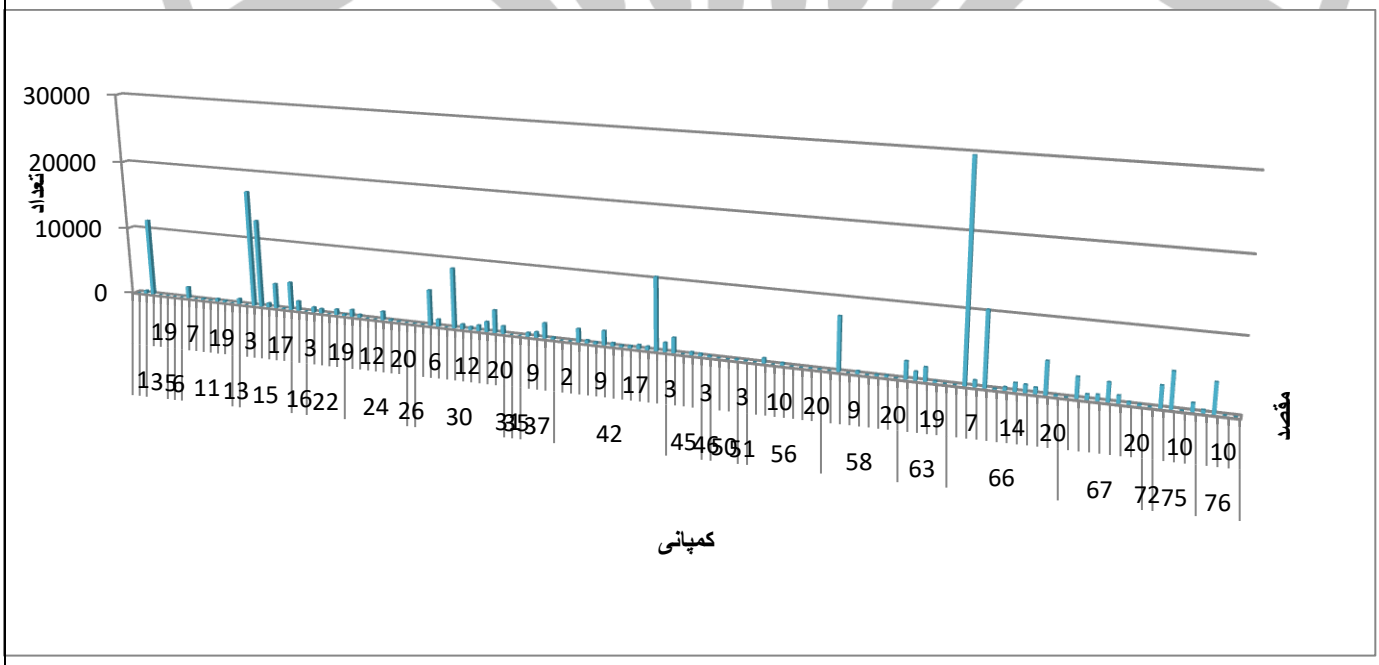
درآمد هر شرکت:

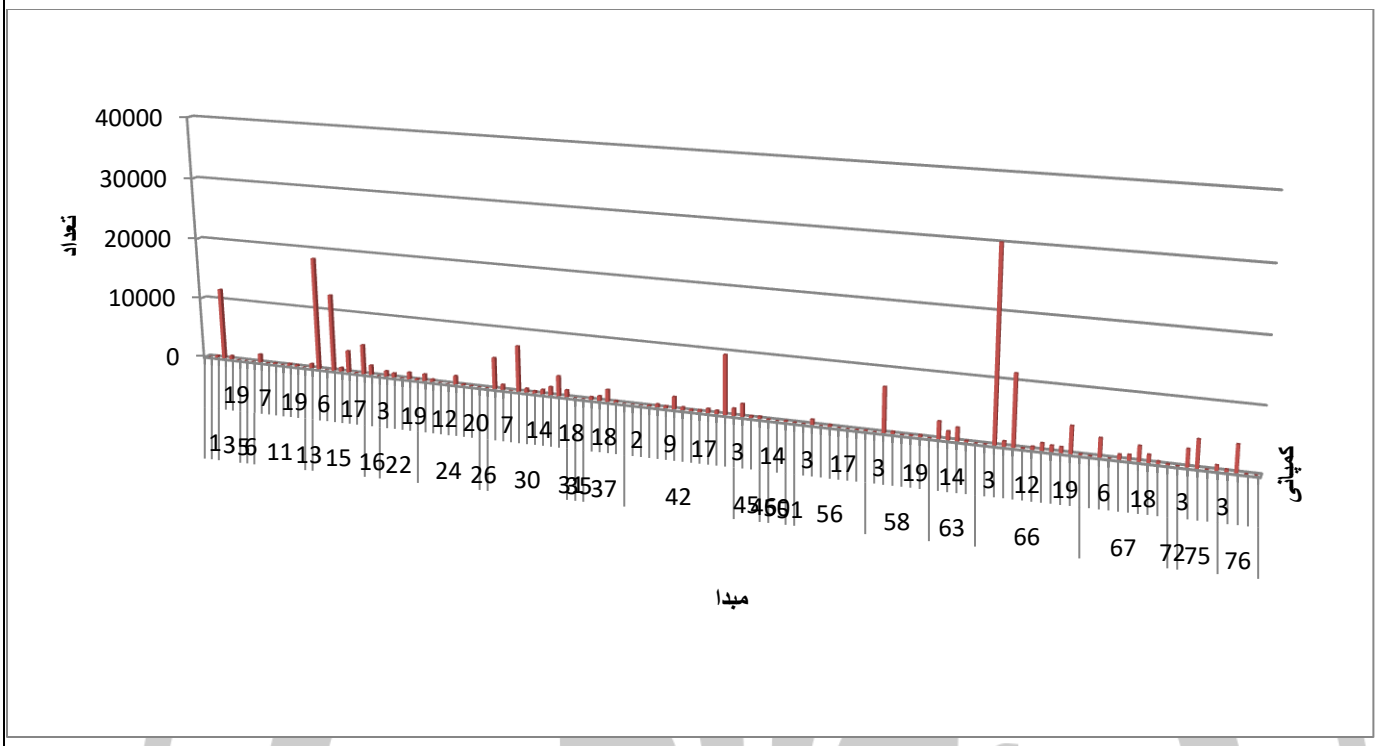


سهم هر کمپانی از کل پروازها با تعداد بلیط‌های فروخته شده ارتباط مستقیم دارد؛ زیرا سهم از درصد تعداد پروازهای هر شرکت نسبت به کل به دست آمده است. انتظار می‌رفت که با بیشتر شدن سهم یک شرکت، درآمد آن نیز افزایش یابد که با مقایسه‌ی دو نمودار بالا می‌بینیم که درصدها بسیار نزدیک هم هستند و اختلاف کم آن‌ها میتواند ناشی از عواملی مانند تفاوت میانگین قیمت بلیط‌های شرکت‌ها و ... باشد.



این نمودار تعداد فروش کمپانی‌های مختلف را در ماه‌های مختلف سال نمایش می‌دهد. مطابق این نمودار برخی از شرکت‌ها فقط در تعداد محدودی از ماه‌های سال پرواز دارند. اما با این وجود، باز هم می‌توان گفت در هر ماه، نسبت فروش کمپانی‌های مختلف کمابیش ثابت می‌ماند و از الگویی مشابه الگوی دو نمودار قبلی پیروی می‌کند

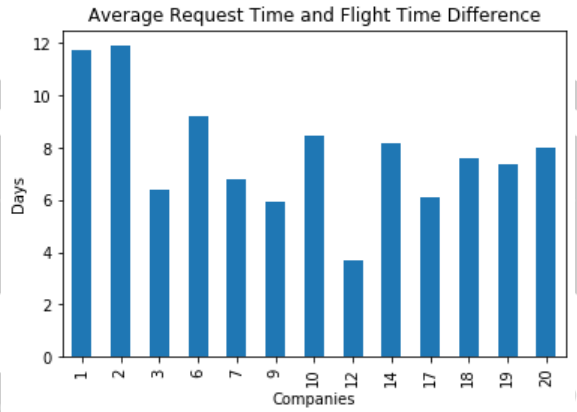




دو نمودار آخر تعداد پروازهای هر کمپانی از هر مبدا به هر مقصد را به ترتیب نشان می‌دهند. می‌توان به وضوح دید که این دو نمودار شباهت بسیاری به یکدیگر دارند؛ اما الزاما در هر فرودگاه؛ پرطرفدارترین شرکت از الگوی نمودارهای قبلی پیروی نمی‌کند و به نظر می‌آید که ارتباط معناداری وجود ندارد. با این وجود، از این دو نمودار دریافت می‌شود که بیش‌ترین تعداد پروازهای پرطرفدارترین شرکت، از محبوب‌ترین فرودگاه انجام می‌شود. با این حال نمی‌توان درباره‌ی وابستگی و ارتباط معنای‌دار این دو موضوع با قطعیت سخن گفت و باید داده‌های بیش‌تری بررسی شوند تا بتوان با اطمینان سخن گفت.

طرح پرسش

- **پرسش ۱:** فاصله‌ی میان تاریخ خرید بلیط و پرواز برای کمپانی‌های مختلف به چه صورت است؟ آیا ارتباطی بین محبوبیت ایرلاین و این اختلاف زمانی وجود دارد؟

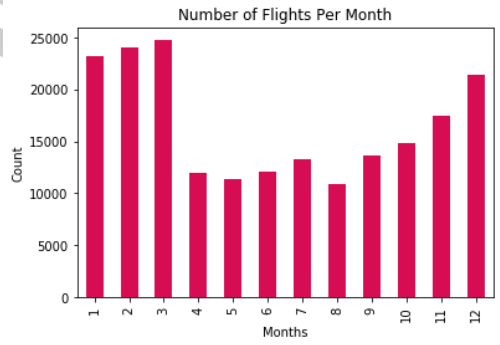


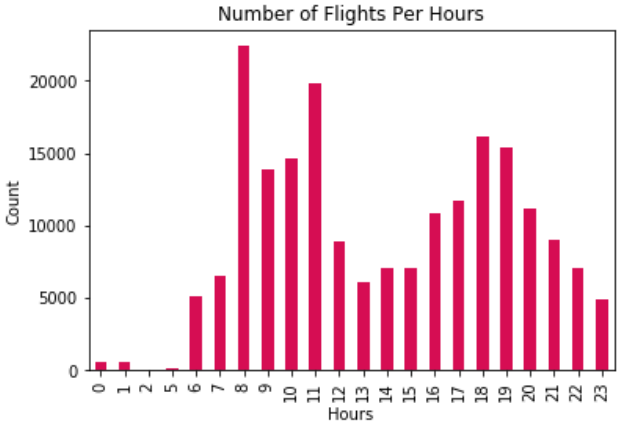
در نمودار بالا اختلاف زمانی میان تاریخ فروش بلیط و پرواز بر حسب شرکت را نشان می‌دهد. اگر چه به طور کلی این نمودار با محبوبیت ایرلاین ارتباط تنگاتنگی ندارد؛ اما می‌توان با در نظر گرفتن میانگین اختلاف زمانی سه ایرلاین محبوب(۳ و ۹ و ۱۹) و سه ایرلاین کم‌طرفدار(۱ و ۲و ۶)، می‌توان گفت هر چه ایرلاین پرطرفدارتر باشد، اختلاف زمانی دو تاریخ مذکور به طور میانگین کمتر است اما به هر حال برای اظهار نظر قطعی به داده‌های بیش‌تر و بررسی‌های پیشرفته‌تری نیاز هست.

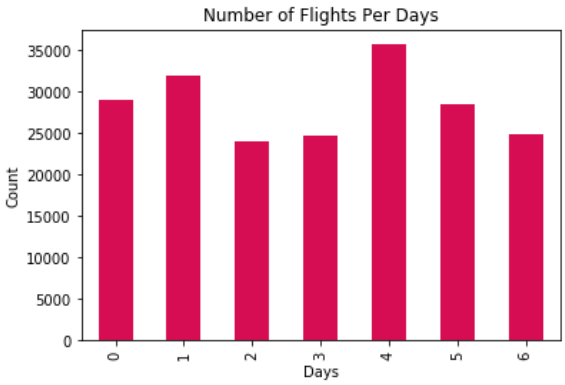
میانگین اختلاف زمانی ۳ ایرلاین پرطرفدار: ۶/۶۷

میانگین اختلاف زمانی ۳ ایرلاین کم‌طرفدار: ۱۰/۶۳

- **پرسش ۲:** بیش‌ترین پروازها در چه ماهی، چه روزی و چه ساعاتی از شبانه‌روز انجام می‌شود؟ آیا می‌توان حدس زد این داده‌ها متعلق به چه جایی هستند؟







با تحلیل این نمودارها می‌توان تا حدودی به رفتار اعضای جامعه پی برد؛ برای مثال پرطرفدارترین ماه سال(ماه سوم، مارس) می‌تواند نشانگر مناسبت و تعطیلات مهمی باشد که موجب می‌شود خریداران با فراغ بال به مسافرت بروند. از طرفی؛ ماه هشتم(سپتامبر) که کم‌ترین تعداد پرواز را دارد می‌تواند ماهی بسیار پرمشغله و پرکار باشد که مجالی برای مسافرت باقی نمی‌گذارد.

با توجه به ساعات پرطرفدار، می‌توان برداشت کرد که مشتریان ترجیح می‌دهند یا اول صبح پرواز کنند تا همان صبح یا نهایتا آخر روز به مقصدشان برسند یا این که پروازشان عصر باشد تا اواخر روز یا صبح روز بعد در مقصدشان باشند.

در مورد روزهای هفته می‌توان حدس زد که روز جمعه(که محبوب‌ترین روز است) روزی است که بسیاری از سفرهای کاری یا مسافرت‌های خود برمی‌گردند یا برای انجام کار یا شغلی به جایی دیگر می‌روند.

با توجه به الگوهای ذکر شده، می‌توان حدس زد که داده‌ها مربوط به پروازهای داخلی خود ایران باشند!