

گزارش پروژه تاکسی اینترنتی در نیویورک

گروه نِمُو: امید اسدی نیا، مریم خزائی، محمدرضا مشعوفی، حسام الدین میرجلیلی، و وحید نصیری منش به

تاریخ ۱۳ تیر ۱۴۰۳

خواسته یک: چکیده آماری

تمیزکاری داده

تجمیع داده‌ها

در تجمیع داده‌های تاکسی‌های زرد و سبز مسئله مهم تفاوت نام‌گذاری متغیرهای زمان‌های سفر این تاکسی‌ها، به ترتیب `lpep_..._datetime` و `tpep_..._datetime`، بود که نخست نام‌گذاری آن‌ها یکسان شد. پس از یکسان‌سازی متغیرهای زمان سفر هر دو گونه از تاکسی‌ها، با توجه به تعریف مسئله، سفرها در بازه زمانی آغاز سال ۲۰۲۰ تا پایان ماه مارس سال ۲۰۲۴ نگه داشته شدند.

پاکسازی مشاهدات

مهم‌ترین قواعدی که برای نگه‌داری مشاهدات سفرها اعمال شده‌اند، به قرار زیر است:

- مشاهداتی که زمان آغاز سفر پس از زمان پایان آن بوده‌اند، حذف شده‌اند
- با توجه به توزیع مقادیر متغیرهای شمار مسافر، مسافت پیموده‌شده، مدت زمان سفر، کرایه، هزینه‌های اضافه، انعام، عوارض، مالیات و هزینه کل سفر حداکثر ۰/۱٪ از مشاهدات آن‌ها پس از بررسی مقادیرشان حذف شدند که جزئیات نگه‌داشت مقادیر آن‌ها در ادامه آمده‌اند:

- متغیر `fare_amount` بین ۰ تا ۹۹ و مقادیر `NULL` نگه داشته شدند
- متغیر `extra` بین ۰ تا ۱۰ و مقادیر `NULL` نگه داشته شدند
- متغیر `mta_tax` بین ۰ تا ۳ و مقادیر `NULL` نگه داشته شدند

- متغیر tip_amount بین ۰ تا ۲۰ و مقادیر NULL نگه داشته شدند
 - متغیر tolls_amount بین ۰ تا ۲۰ و مقادیر NULL نگه داشته شدند
 - متغیر passenger_count بین ۱ تا ۶ و مقادیر NULL نگه داشته شدند؛ با توجه به اینکه در شهر نیویورک حداکثر تعداد مسافر مجاز ۵ نفر و در صورتی که یک مسافر کودک زیر ۷ سال باشد و در صندلی عقب بر روی پای فردی بنشیند، ۶ نفر است
 - متغیر trip_distance بزرگتر از ۰ و کوچکتر مساوی ۵۰۰ نگه داشته شدند
 - متغیر سرعت تاکسی از تقسیم مسافت پیموده شده بر زمان سفر ساخته شد. سپس سفرها با سرعت کوچکتر مساوی ۹۰ نگه داشته شدند
 - متغیرهای total_amount، improvement_surcharge، congestion_surcharge و airport_fee بزرگتر مساوی ۰ یا مقادیر NULL باشد.
 - سفرها با مسافت بیش از ۰/۳۱ مایل نگه داشته شدند
 - مدت زمانها بیش از یک دقیقه نگه داشته شدند
 - قواعد اعمال شده برای نگه داشتن مقادیر متغیرهای رسته‌ای^۱ عبارتند از:
 - مشاهدات با مقادیر ۰ و ۶ متغیر payment_type حذف شدند
 - مشاهدات با مقادیر ۵ و ۶ متغیر vendorID حذف شدند
 - مشاهدات با مقادیر ۹۹ متغیر ratecodeID حذف شدند
- در پایان مشاهدات تکراری بر پایه شش متغیر vendorID، unified_pickup_datetime، unified_dropoff_datetime، PULocationID و DOLocationID حذف شدند.
- با اعمال همه قواعد یادشده از ۱۴۷،۶۷۲،۴۸۰ مشاهده ۱۴۳،۵۸۴،۱۳۲ سفر نگه داشته شدند.

چکیده آماری و نمایی از داده‌ها

در جدول ۱ چکیده آماری متغیرهای شمار مسافر، مسافت، سرعت تاکسی، کرایه و هزینه‌های سفر ارائه شده‌اند.

جدول ۱: چکیده آماری مهم‌ترین متغیرهای داده تمیزکاری شده

Columns	min	max	mean	std	q25	median	q75	count	Null (%)
---------	-----	-----	------	-----	-----	--------	-----	-------	----------

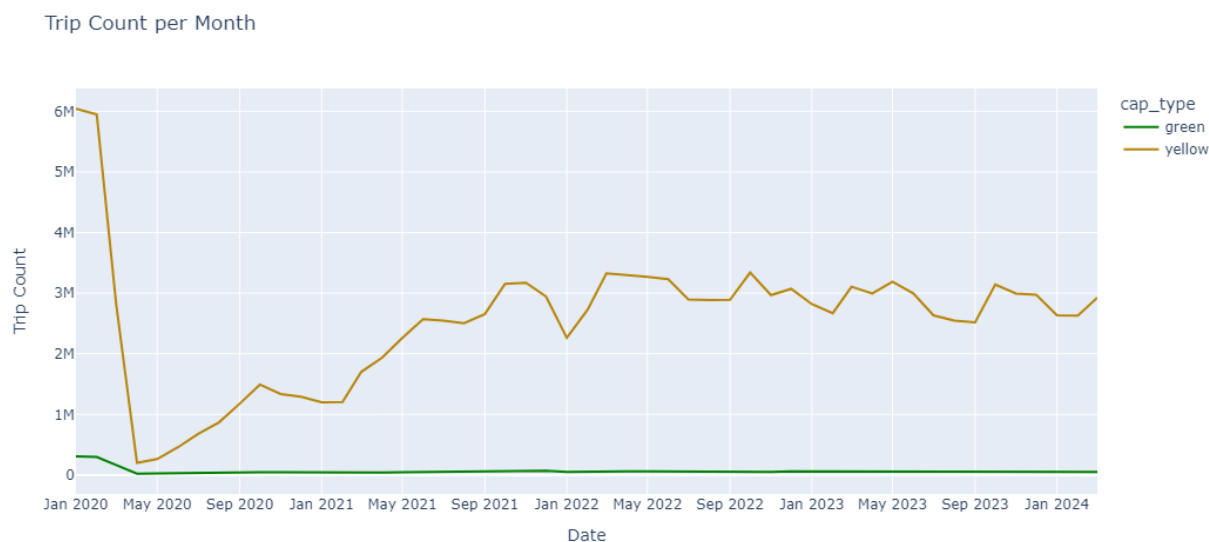
¹ Categorical

passenger_count	1	6	1.43	0.98	1	1	1	132,584,143	0
trip_distance	0.32	374.1	3.22	3.99	1.1	1.8	3.27	132,584,143	0
speed	0.004	90.00	11.84	6.33	7.81	10.26	13.89	132,584,143	0
fare_amount	-	99	15.05	12.95	7.47	10.68	16.83	132,584,143	0
total_amount	-	142.42	22.2	16.53	12.74	16.81	24.1	132,584,143	0

در نمودار ۱ شمار سفرهای ماهانه تاکسی‌های زرد و سبز در گذر زمان نشان داده شده است. چنان که نمایان است تعداد سفرهای

تاکسی‌های سبز در بازه چهار سال سال تغییرات خاصی نداشته اما تاکسی‌های زرد بعد از یک افت شدید دوباره رشد کرده و از

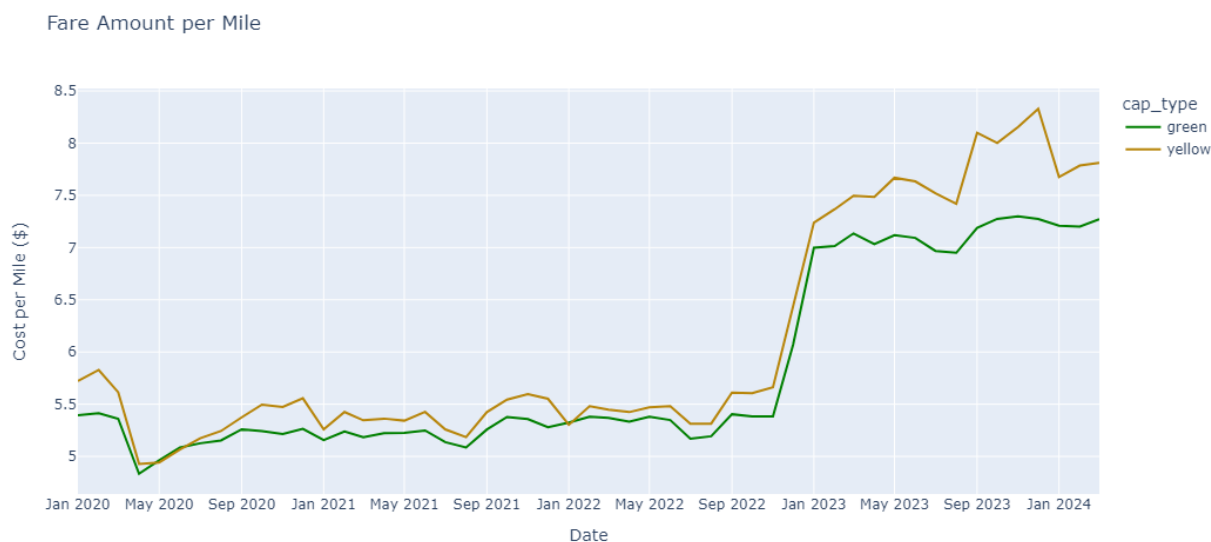
ماه سپتامبر ۲۰۲۱ تقریباً در یک بازه‌ی مشخص نوسان دارد.



نمودار ۱: شمار سفرهای ماهانه به تفکیک تاکسی‌های زرد و سبز

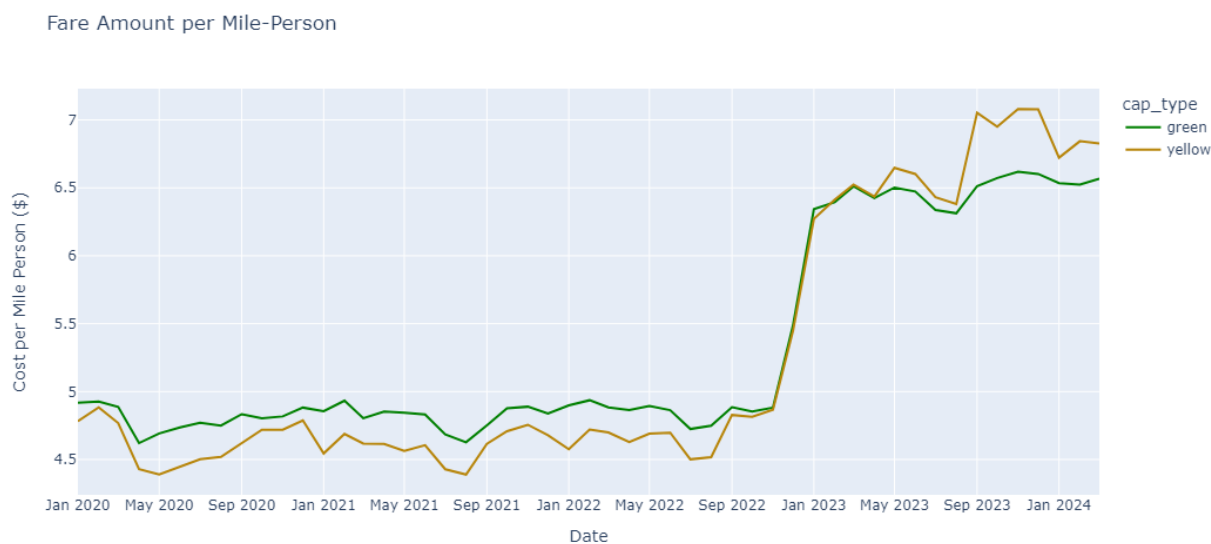
نمودار ۲ کرایه دریافت شده در هر ماه به ازای هر مایل پیمایش به تفکیک تاکسی‌های زرد و سبز را نشان داده که در اواخر سال

۲۰۲۲ کرایه برای هر دو نوع تاکسی‌ها افزایش شدیدی را تجربه کرده است.



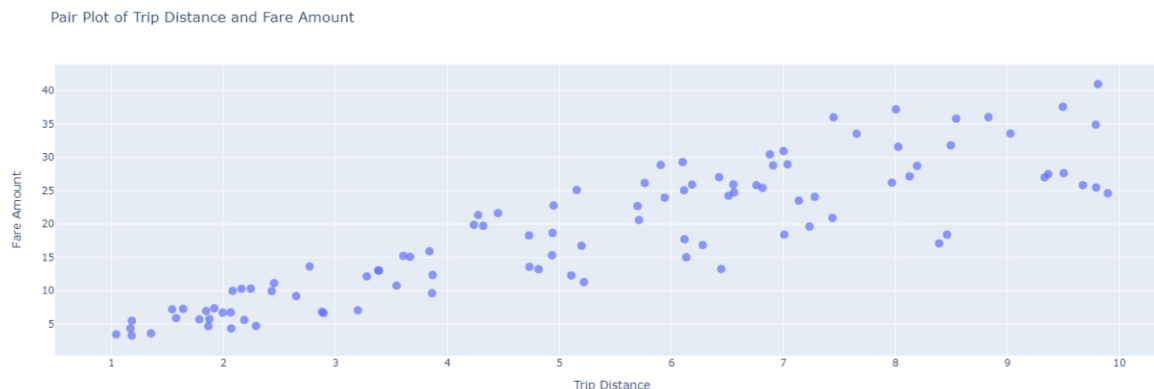
نمودار ۲: کرایه به ازای هر مایل به تفکیک تاکسی‌های زرد و سبز

نمودار ۳ میزان کرایه دریافتی از هر فرد به ازای هر مایل پیمایش را به تفکیک تاکسی‌های زرد و سبز نشان می‌دهد. همانند نمودار ۲ کرایه دریافتی در اواخر سال ۲۰۲۲ رشد شایان توجهی را تجربه می‌کند و پیش و پس از این دوره در بازه محدودی نوسان می‌کند.



نمودار ۳: کرایه به ازای هر مایل پیمایش و هر نفر مسافر به تفکیک تاکسی‌های زرد و سبز

با توجه به نمودار ۴ یک همبستگی مثبت میان مسافت سفرها و کرایه دریافتی قابل مشاهده است.



نمودار ۴: نمودار پراکندگی میان مسافت سفرها (مایل) و کرایه آن‌ها (دلار)

خواسته دو: راه‌اندازی محصول پرداخت اینترنتی

به منظور یافتن مسیرهای مناسب برای محصول پرداخت اینترنتی داده‌های یک سال اخیر، بازه زمانی ۰۱-۰۴-۲۰۲۳ تا ۳۱-۰۳-۲۰۲۴، نگه داشته شدند تا به‌روزترین الگوهای تردد تاکسی‌ها مورد استفاده قرار بگیرند. از سوی دیگر با توجه به ماهیت محصول تنها سفرهایی که پرداخت آن‌ها در این بازه به صورت آنلاین بود، نگه داشته شدند.

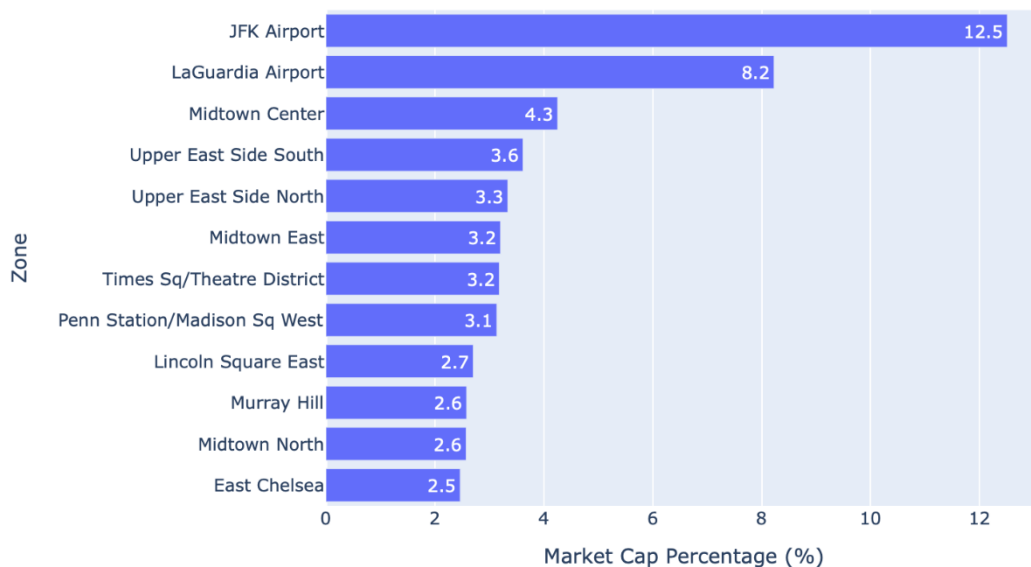
برای راه‌اندازی محصول یادشده نیاز است مبداهایی انتخاب شوند که تعداد تراکنش‌ها و حجم مبلغ تراکنش‌ها در آن‌ها بالا باشد تا بدین ترتیب با کسب سهم هر چه بیشتر از سفرهای آن‌ها درآمد ناشی از محصول افزایش یابد. به همین علت برای یافتن مبدأ مناسب یک متریک ترکیبی از شمار سفرها در هر مبدأ و میانه قیمت سفرها در آن مبدأ در نظر گرفته شد. هرچه شمار سفرها بیشتر باشد، تعداد تراکنش بالاتر می‌رود و هر چه میانه قیمت سفر در یک مبدأ بالاتر باشد، احتمال رخداد تراکنش‌ها با مبلغ بالاتر در آن مسیر بالاتر خواهد بود. بدین ترتیب متریک برای انتخاب مبدأ مناسب برای هر مبدأ به قرار زیر محاسبه شد:

شمار سفرها \times میانه مبلغ سفرها

لازم به ذکر است که هر چه این مقدار برای هر مبدأ بالاتر باشد، آن مبدأ برای راه‌اندازی محصول پرداخت اینترنتی مناسب‌تر خواهد بود.

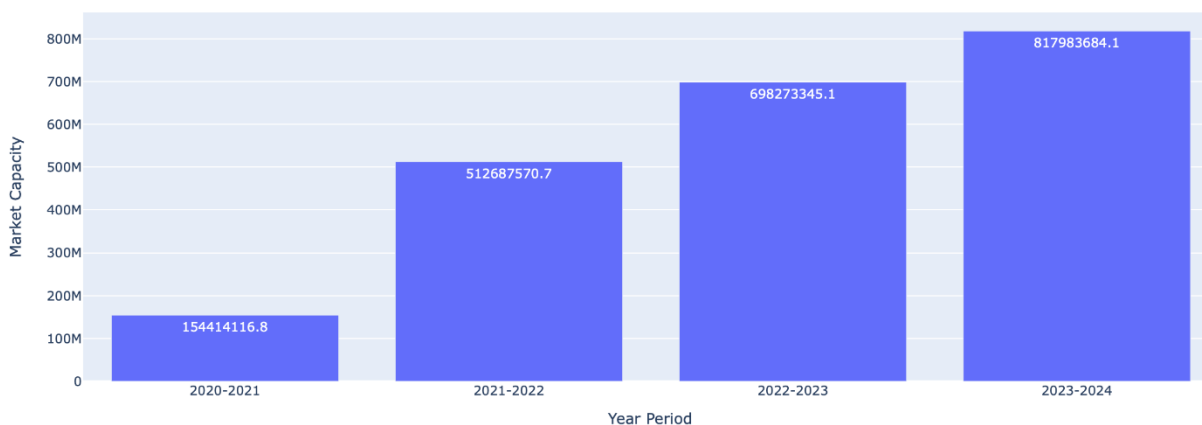
در ادامه با محاسبه متریک یادشده برای هر مبدأ ۱۲ مبدأ که بیشترین امتیاز را کسب کرده بودند، به عنوان محل‌های پیشنهادی انتخاب شدند. نمودار ۵ سهم بازار هر یک از مبدأهای پیشنهادی را نشان می‌دهد که مجموع سهم آن‌ها به بیش از ۵۰٪ می‌رسد.

Market Capacity Percentage by Zone In a Recent Year



نمودار ۵: سهم مبداهای پیشنهادی از ارزش بازار سفرهای پرداختی با کارت اعتباری در بازه آوریل ۲۰۲۳ تا پایان مارس ۲۰۲۴ در رابطه با میزان سود محصول BNPL ابتدا روند ارزش بازار در بازه‌های زمانی یک‌ساله مورد بررسی قرار گرفت. نمودار ۶ سهم ارزش سالانه بازار سفرهای پرداختی با کارت اعتباری را نمایش می‌دهد. با توجه به رشد مرکب سالانه^۲ ۷۴٪ ارزش این بازار می‌توان انتظار داشت که ارزش آن برای بازه یک‌ساله پس از مارس ۲۰۲۴ به رقم یک میلیارد و ۴۲۳ میلیون دلار برسد.

Market Capacity in Recent Years



نمودار ۶: ارزش بازار سفرهای پرداختی با کارت اعتباری (دلار)

² Compound Annual Growth Rate (CAGR)

با توجه به این که ۱۲ مبدأ انتخابی در مجموع ۵۲ درصد از سهم بازار را دربرگرفته‌اند و با فرض این که مجموع سهم آن‌ها از بازار در یک سال پیش رو دستخوش تغییر چندانی نشود، می‌توان انتظار داشت که با تصاحب ۱۰ درصد از تراکنش‌های این مبدأها حجم تراکنش‌های محصول پرداخت اینترنتی به بیش از ۷۴ میلیون دلار برسد. با لحاظ کارمزد سرویس برابر ۱/۵٪ درآمد محصول یادشده برای یک سال پس از مارس ۲۰۲۴ برابر یک میلیون و ۱۱۰ هزار دلار برآورد می‌شود.

خواسته سه: جایگاه تبلیغاتی

برای بررسی این خواسته از داده سفرهای یک سال اخیر یاری گرفته شده تا واپسین الگوهای تردد تاکسی‌ها در برآوردهای گزارش لحاظ شود.

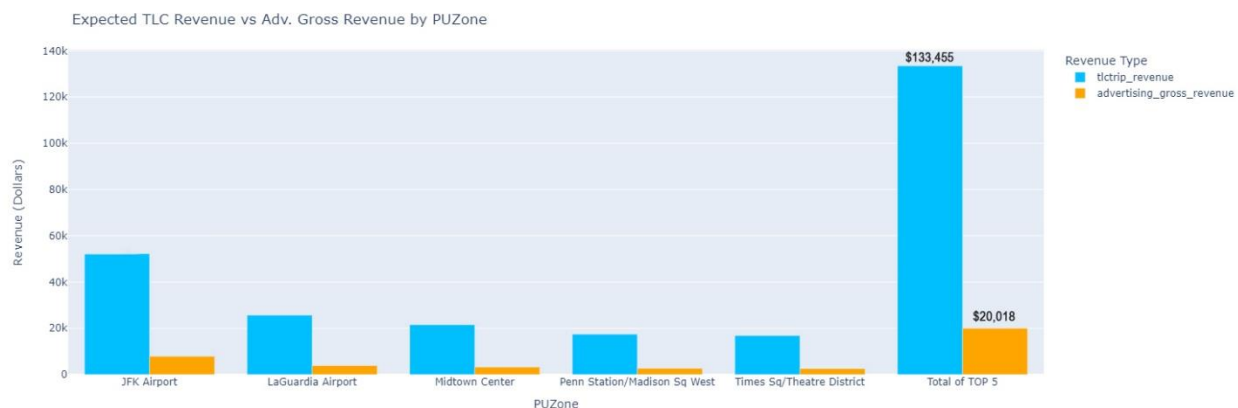
متریک تعریف شده برای برآورد پتانسیل تبلیغاتی هر سفر متریک مدت زمان سفر به کار گرفته می‌شود. با توجه به اینکه در موضوع تبلیغات دفعات نمایش آن مهم است، پس باید زمان هر سفر محاسبه شود تا بتوان شمار دفعات پخش آگهی‌ها را در هر سفر محاسبه نمود.

برای برآورد درآمدزایی این روش از تبلیغات فرض‌های زیر در نظر گرفته شده‌اند:

- میانگین هزینه تمام شده در ازای هر هزار نمایش تبلیغ (CPM) معادل ۳/۵۳ دلار است اما از آنجایی که توان شرکت در تعیین جامعه هدف (Targeting) کمتر است، نصف این مقدار یعنی ۱/۷۶ دلار در نظر گرفته شد
- میانگین مدت زمان ویدیوهای تبلیغاتی از ۱۰ تا ۲۰ ثانیه است. برای برآورد این میزان ۳۰ ثانیه لحاظ شده است
- آژانس‌های تبلیغ‌کننده معمولاً بین ۱۰ تا ۲۰ درصد بودجه تبلیغات را به عنوان سهم درآمدی خود در نظر می‌گیرند. برای برآورد سهم ۱۵٪ انتخاب شده است

با محاسبه مجموع مدت زمان سفرها برای تمامی محله‌ها و تقسیم آن بر ۳۰ ثانیه شمار دفعات نمایش ویدئوهای تبلیغاتی به دست می‌آیند. با تقسیم این تعداد بر هزار و ضرب آن در رقم ۱/۷۶ دلار درآمد شرکت تاکسی‌رانی به دست می‌آیند. درآمد آژانس تبلیغاتی ۱۵٪ درآمد شرکت تاکسی‌رانی لحاظ می‌شود. میزان درآمد شرکت تاکسی‌رانی و آژانس تبلیغاتی با توجه به داده‌های سفر به ترتیب ۲،۱۲۱،۸۳۸ و ۳۱۸،۲۷۵ دلار به دست می‌آیند.

افزون بر فرض‌های قبلی در نظر گرفته می‌شود که در ابتدا یک فاز آزمایشی به مدت ۳ ماه انجام می‌شود. این بار همان محاسبات بالا اما برای مدت ۳ ماه اخیر و به تفکیک محله‌ها انجام می‌شود. نمودار ۷ درآمد شرکت تاکسی‌رانی و آژانس تبلیغاتی را از تبلیغات به تفکیک هر محله نشان می‌دهد.



نمودار ۷: درآمد شرکت تاکسی‌رانی و آژانس تبلیغاتی از تبلیغات به تفکیک هر محله

خواسته چهار: قیمت‌گذاری سفر

برای ارزیابی عوامل مؤثر بر بهای سفر نمونه تصادفی ده درصد از کل داده‌ها گرفته شد. از آن‌جاکه به علت تورم کرایه‌های سفر در گذر زمان تغییر می‌کنند، در آغاز متغیر کرایه حقیقی از متغیر کل هزینه هر سفر، `total_amount`، ساخته شد؛ بدین ترتیب که کرایه‌های همه سال‌ها با توجه به تورم آمریکا به ارزش سال ۲۰۲۴ آورده شدند. جدول ۲ چکیده‌ای از مهم‌ترین متغیرهای نمونه تصادفی را ارائه می‌دهد:

جدول ۲: چکیده آماری از نمونه تصادفی ۱۰٪

Columns	count	mean	std	min	q25	median	q75	max
rlfare	13,366,471	24.11	17.45	0.01	14.22	18.46	26	161.18
distance	13,366,471	3.21	3.98	0.32	1.1	1.8	3.26	368.4
passenger_count	13,366,471	1.43	0.97	1.00	1.00	1.00	1.00	6.00

در راستای اندازه‌گیری اثرات عوامل گوناگون بر متغیر یادشده مدل خطی زیر طراحی شد:

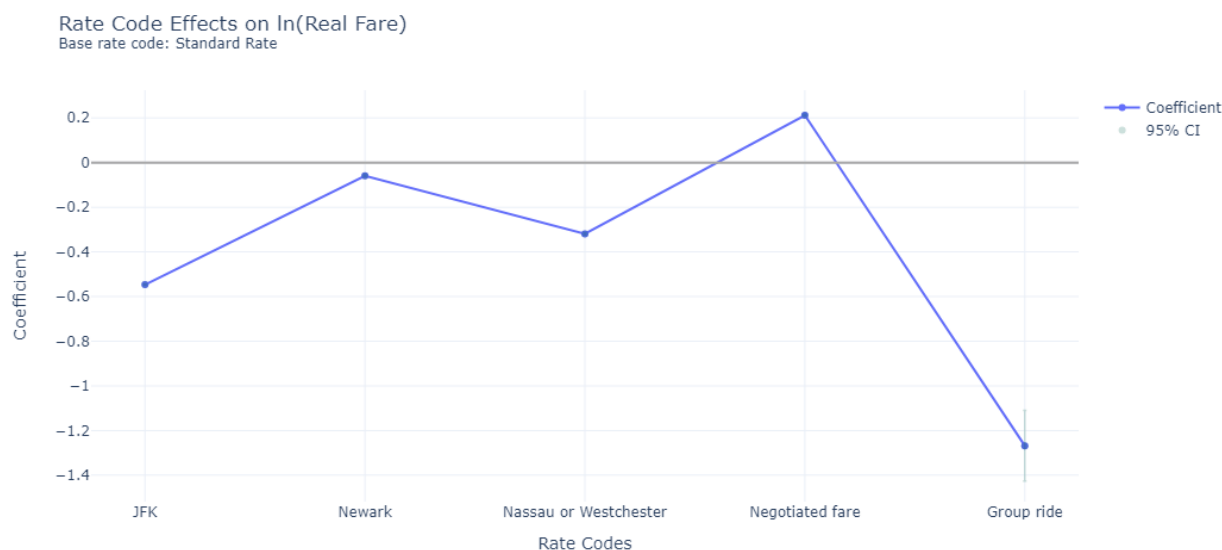
$$\ln(\text{real fare})_i = \text{distance}_i + \text{passenger count}_i + 1[\text{cab type}]_i + 1[\text{Rate Code}] + 1[\text{Time variables: day hour, weekday, month, year}]_i + \varepsilon_i$$

متغیر وابسته مدل لگاریتم طبیعی کل هزینه حقیقی سفر i است. متغیرهای توضیحی به ترتیب عبارتند از: مسافت، شمار مسافران، نوع تاکسی (= ۱ اگر تاکسی زرد باشد)، شیوه تعیین نرخ سفر، و متغیرهای رسته‌ای زمان که به ترتیب ساعات شبانه‌روز، روز هفته، ماه و سال هستند.

یافته‌ها

عوامل مؤثر بر بهای سفر

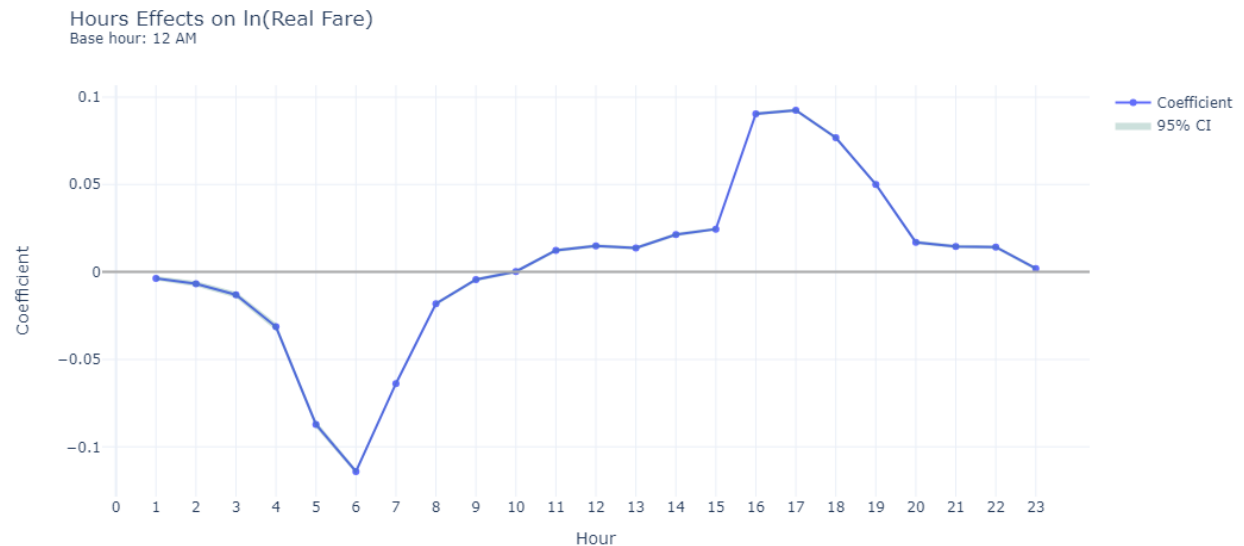
برآورد ضرایب مسافت، شمار مسافر و نوع تاکسی در رگرسیون خطی به ترتیب برابر $0/130$ ، $0/002$ و $0/176$ شد که هر سه در سطح 99% معنادار بوده‌اند. مقدار برآوردی ضریب مسافت نشان می‌دهد که با ثبات سایر شرایط در رگرسیون هر مایل پیمایش بیشتر با افزایش 13% کرایه حقیقی تاکسی همراه است. از سوی دیگر با توجه به مقدار برآوردی ضریب شمار مسافر می‌توان گفت افزایش مسافران تاکسی‌ها اثر شایان توجهی بر افزایش کرایه حقیقی ندارد که با شیوه درست گرفتن تاکسی‌ها همخوان است. در پایان کرایه تاکسی‌های زرد در مقایسه با تاکسی‌های سبز با ثبات سایر شرایط به طور میانگین $17/6\%$ بالاتر است.



نمودار ۸: اثرات روش‌های نرخ‌گذاری بر کرایه حقیقی سفر (روش پایه: روش استاندارد)

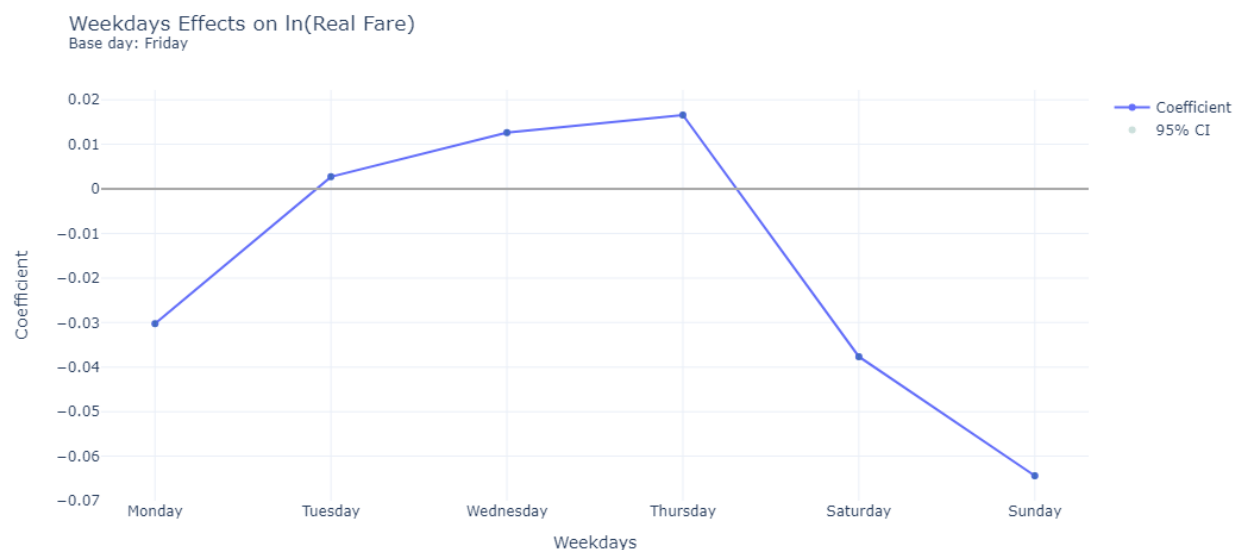
نمودار ۸ پیامد روش‌های نرخ‌گذاری بر کرایه حقیقی سفرها را نسبت به روش استاندارد را می‌سنجد. در حالی که همه روش‌های نرخ‌گذاری نسبت به روش استاندارد با ثبات دیگر شرایط به طور میانگین کرایه‌های حقیقی پایین‌تری دارند، کرایه‌های سفرهایی که با روش مذاکره‌ای تعیین شده‌اند، گران‌تر به شمار می‌آیند، موضوعی که می‌تواند نشانگر قدرت پایین‌تر چانه‌زنی سمت تقاضا در تعیین کرایه‌ها باشد.

الگوی تغییرات قیمت در گذر زمان



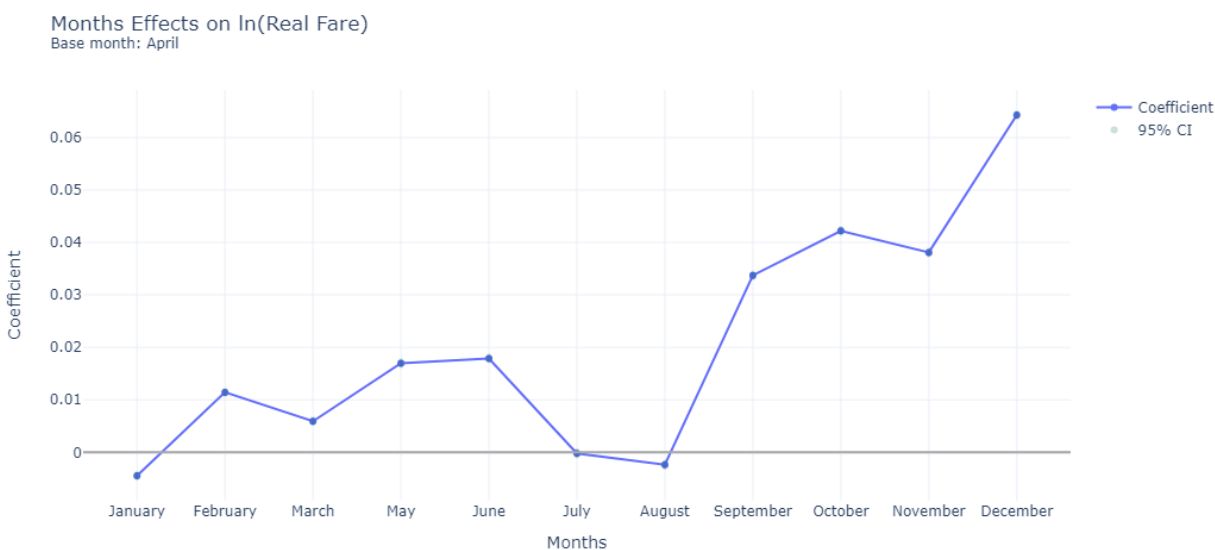
نمودار ۹: اثرات ساعات شبانه‌روز بر کرایه حقیقی سفر (ساعت پایه ۱۲ بامداد)

نمودار ۹ ضرایب برآوردی رگرسیون برای ساعات مختلف شبانه‌روز را نمایش می‌دهد. چنان‌که آشکار است با رفتن از ساعت ۱۲ بامداد به سوی ۶ صبح کرایه سفرها به طور میانگین با افت‌های فزاینده نسبت به ساعت پایه همراه است. اما با گذر از ساعت ۶ صبح تا ساعت ۱۰ صبح این افت کم‌رنگ‌تر می‌شود؛ به گونه‌ای که از ساعت ۱۰ صبح تا ساعت ۲۲ کرایه‌های حقیقی سفر به طور میانگین نسبت به ساعت ۱۲ بامداد با ثبات سایر شرایط به طور معناداری بالاترند. این افزایش در ساعت ۱۷ به اوج خود می‌رسد؛ به گونه‌ای که کرایه‌های حقیقی سفر در این ساعت با ثبات سایر شرایط نسبت به سفرهای ۱۲ بامداد به طور میانگین ۹/۲٪ گران‌ترند که این موضوع می‌تواند به علت افزایش تقاضای سفر افراد برای بازگشت به خانه در این ساعت باشد.



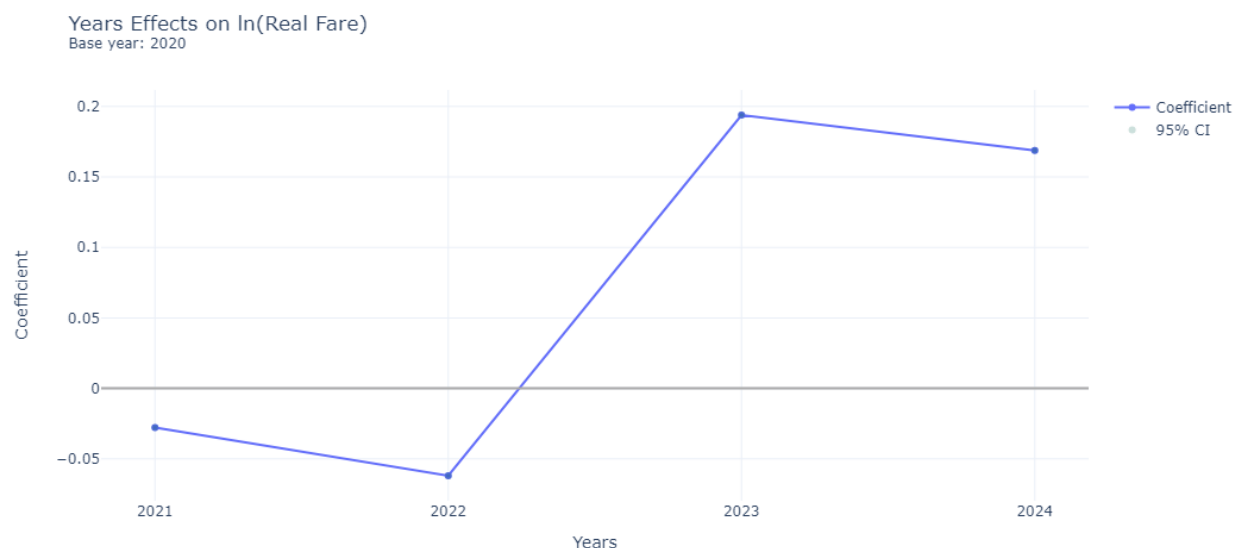
نمودار ۱۰: اثرات روزهای هفته بر کرایه حقیقی سفر (روز پایه: جمعه)

نمودار ۱۰ ضرایب برآوردی رگرسیون برای روزهای مختلف هفته را به تصویر می‌کشد. از روز دوشنبه تا پنج‌شنبه بهای حقیقی کرایه‌ها روندی فزاینده را تجربه می‌کنند؛ در حالی که کرایه‌های حقیقی سفر با ثبات سایر شرایط به طور میانگین از سفرهای روز جمعه ۳٪ ارزان‌ترند، آن‌ها با افزایش تدریجی در روز پنج‌شنبه به طور میانگین ۱/۷٪ از سفرهای روز جمعه گران‌تر به شمار می‌آیند. اما با رسیدن به روزهای شنبه و یکشنبه، به عنوان روزهای تعطیل هفته، بهای حقیقی سفرها افت شدیدی را تجربه می‌کند؛ به گونه‌ای که سفرهای روزهای شنبه و یکشنبه با ثبات دیگر شرایط به طور میانگین به ترتیب ۳/۸٪ و ۶/۴٪ از سفرهای روز جمعه ارزان‌ترند که می‌توان علت آن را در افت تقاضای سفر در این روزها جست.



نمودار ۱۱: اثرات ماه‌ها بر کرایه حقیقی سفر (ماه پایه: آوریل)

نمودار ۱۱ نوسانات کرایه حقیقی را در ماه‌های مختلف نسبت به ماه آوریل نشان می‌دهد. به طور کلی روندی که قابل مشاهده است، افزایش نرخ حقیقی سفرها با رفتن به ماه‌های پایانی سال است. با این حال با عبور از ماه اوت به ماه سپتامبر کرایه‌های حقیقی رشد بالایی را تجربه می‌کنند؛ در حالی که کرایه‌های حقیقی سفر در ماه اوت با ثبات دیگر شرایط به طور میانگین ۰/۲٪ از سفرهای ماه آوریل ارزان‌ترند، این مقدار برای سفرهای ماه سپتامبر به افزایش ۳/۴٪ تغییر می‌کند. با افزایش کرایه‌های حقیقی در ماه‌های بعدی این مقدار در ماه دسامبر به اوج خود می‌رسد؛ کرایه‌های حقیقی در ماه دسامبر به طور میانگین ۶/۴٪ از کرایه‌های ماه آوریل گران‌ترند. می‌توان احتمال داد که با فرارسیدن ماه‌های سرد سال و آغاز مدارس تقاضا برای تاکسی‌ها با جهش همراه می‌شود و این موضوع می‌تواند نوسانات کرایه را در گذر ماه‌ها توضیح دهد.



نمودار ۱۲: اثرات سال‌ها بر کرایه حقیقی سفر (سال پایه: ۲۰۲۰)

نمودار ۱۲ همان اثرات پیشین را به تفکیک سال‌های متفاوت نشان می‌دهد. چنان‌که آشکار است که کرایه حقیقی سفرها به طور میانگین در سال‌های ۲۰۲۱ و ۲۰۲۲ نسبت به سال ۲۰۲۰ پایین‌تر بوده است که می‌توان آن را به افت تقاضا در دوران همه‌گیری نسبت داد. در حالی که همه‌گیری به پایان می‌رسد و تقاضا متعاقباً افزایش می‌یابد، کرایه‌ها در سال‌های ۲۰۲۳ و ۲۰۲۴ افزایش می‌یابند که می‌توان آن را به رشد شدید تقاضا نسبت داد.

خواسته پنج: بازاریابی تاکسی اینترنتی

به منظور یافتن مسیرهای مناسب برای محصول پرداخت اینترنتی داده‌های یک سال اخیر، بازه زمانی ۰۱-۰۴-۲۰۲۳ تا ۳۱-۰۳-۲۰۲۴، نگه داشته شدند تا به‌روزترین الگوهای تردد تاکسی‌ها مورد استفاده قرار بگیرند.

بخش الف: پیشنهاد محل‌های استقرار اولیه

برای راه‌اندازی محصول یادشده نیاز است مبداهایی انتخاب شوند که تعداد سفر و حجم کرایه‌ها در آن‌ها بالا باشد تا بدین ترتیب با کسب سهم هر چه بیشتر از سفرهای آن‌ها درآمد پلتفرم افزایش یابد. به همین علت برای یافتن مبدأ مناسب یک متریک ترکیبی از شمار سفرها در هر مبدأ و میانه بهای سفرها در آن مبدأ در نظر گرفته شد. هرچه شمار سفرها بیشتر باشد، تعداد مسافران برای تبلیغ بازاریاب‌ها بیشتر خواهد بود و هر چه میانه قیمت سفر در یک مبدأ بالاتر باشد، احتمال روبه‌رو شدن با مسافرهایی که بهای

بیشتری برای سفرها می‌پردازند، بیشتر می‌شود. بدین ترتیب متریک برای انتخاب مبدأ مناسب برای هر مبدأ به قرار زیر محاسبه شد:

شمار سفرها \times میانه مبلغ سفرها

لازم به ذکر است که هر چه این مقدار برای هر مبدأ بالاتر باشد، آن مبدأ برای راه‌اندازی تاکسی اینترنتی مناسب‌تر خواهد بود. با توجه به این متریک پنج مبدأ که بیشترین مقدار را داشتند، انتخاب شدند که نام‌های آن‌ها به ترتیب بیشترین مقدار متریک عبارتند از: Upper East Side North و Upper East Side South، Midtown Center، LaGuardia Airport، JFK Airport. این پنج ناحیه در مجموع بیش از ۳۰٪ ارزش بازار سفرها را در یک سال اخیر به خود اختصاص داده بودند.

بخش ب: مسیرهای پرتردد

در راستای تعیین مسیرهای پرتردد برای هر کدام از محل‌های استقرار پیشنهادی شمار سفرها در یک سال اخیر به تفکیک مقصدها شمرده شدند و مقصدها با بیشترین شمار سفر به عنوان پرترددترین مسیر برای هر یک از پنج محل پیشنهادی انتخاب گردیدند. جدول ۳ پرترددترین مسیرها برای هر مبدأ آورده شده‌اند.

جدول ۳: پرترددترین مسیرهای مبدأهای پیشنهادی در بازه یک سال اخیر

Suggested location	Most crowded destination
JFK Airport	Times Sq/Theatre District
LaGuardia Airport	Times Sq/Theatre District
Midtown Center	Upper East Side South
Upper East Side North	Upper East Side South
Upper East Side South	Upper East Side North

برای بررسی این که آیا این مسیرهای پرتردد در طول روز تغییر می‌کنند، بازه شبانه‌روز به چهار بازه شش ساعته ۱۲ بامداد تا ساعت شش، ساعت ۶ تا ۱۲، ساعت ۱۲ تا ۱۸ و ساعت ۱۸ تا ساعت ۱۲ بامداد تقسیم شدند. سپس برای هر کدام از مبدأهای پیشنهادی در این بازه‌های زمانی پرترددترین مقصد استخراج شدند که نتایج آن در جدول ۴ ارائه شده است.

جدول ۴: مسیرهای پرتردد محل‌های پیشنهادی در بازه‌های زمانی مختلف شبانه‌روز

Suggested location	Time period	Most crowded destination
JFK Airport	12 AM – 6 AM	Outside of NYC

LaGuardia Airport	6 AM – 12 PM	Times Sq/Theatre District
	12 PM – 6 PM	Times Sq/Theatre District
	6 PM – 12 AM	Times Sq/Theatre District
	12 AM – 6 AM	Times Sq/Theatre District
	6 AM – 12 PM	Times Sq/Theatre District
	12 PM – 6 PM	Times Sq/Theatre District
	6 PM – 12 AM	Times Sq/Theatre District
	12 AM – 6 AM	LaGuardia Airport
Midtown Center	6 AM – 12 PM	Upper East Side South
	12 PM – 6 PM	Upper East Side South
	6 PM – 12 AM	Upper East Side North
	12 AM – 6 AM	East Harlem South
Upper East Side North	6 AM – 12 PM	Upper East Side South
	12 PM – 6 PM	Upper East Side South
	6 PM – 12 AM	Upper East Side South
	12 AM – 6 AM	Upper East Side North
Upper East Side South	6 AM – 12 PM	Upper East Side North
	12 PM – 6 PM	Upper East Side North
	6 PM – 12 AM	Upper East Side North

نمودار ۱۳ نشان می‌دهد که مسیرها به تفکیک مبدأهای پیشنهادی در بازه‌های زمانی شبانه‌روز چند بار به عنوان مسیر پرتدد تکرار

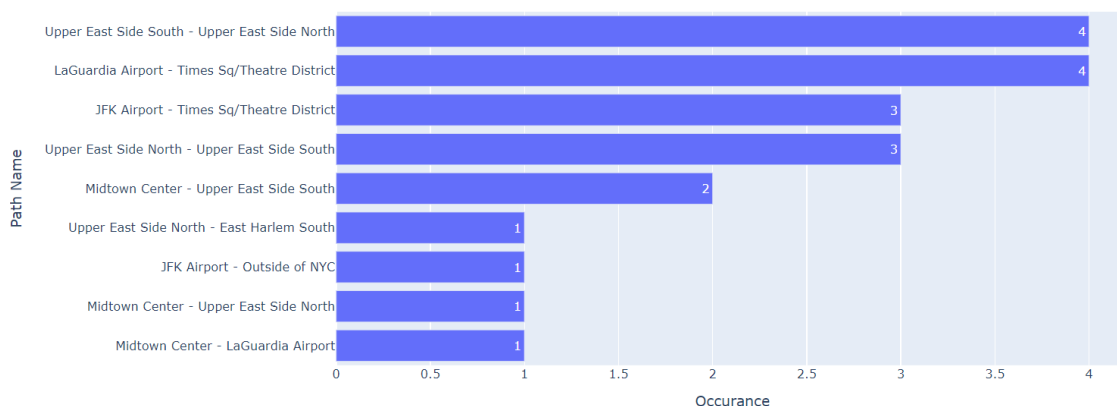
شدند؛ برای نمونه دو مسیر Upper East Side North – Upper East Side South و LaGuardia Airport – Times

sq/Theater District در هر چهار بازه زمانی شبانه‌روز به عنوان مسیر پرتدد به مبدأهای Upper East Side South و

LaGuardia Airport تعیین شده‌اند که نشان از این دارد که مسیرهای پرتدد از این دو مبدأ در طول روز تغییر چندانی به خود

نمی‌بینند.

Occurance of Path in different time periods



نمودار ۱۳: شمار تکرار مسیرها به عنوان مسیر پرتدد در بازه‌های زمانی شبانه‌روز

خواسته شش: قیمت‌گذاری غیرمنصفانه

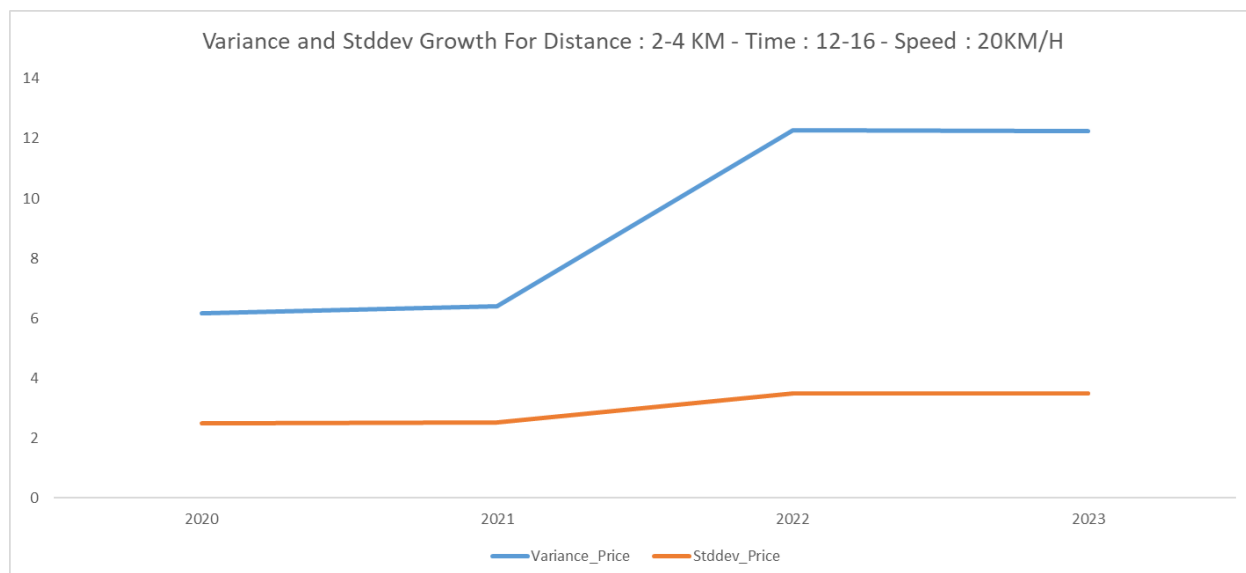
برای گروه‌بندی مسیرهای مشابه از دو معیار سرعت میانگین تاکسی و مسافت‌های مشابه یاری گرفته شده است. سرعت میانگین تاکسی که از تقسیم مسافت بر زمان سفر محاسبه شد، به بازه‌های ۱۰ کیلومتر بر ساعت تقسیم شدند. این متریک می‌تواند نمایانگر وضعیت ترافیک و سختی مسیر از جمله پیچ‌درپیچی راه باشد. مسافت‌های سفر نیز به بازه‌های دو کیلومتری تقسیم شدند. برای تعیین زمان‌های مشابه سفرها نیز شبانه‌روز به بازه‌های چهار ساعته تقسیم شدند؛ برای نمونه بازه ۱۲ تا ۴ صبح یک بازه زمانی در نظر گرفته شد.

بدین ترتیب یک مسیر مشابه در زمان مشابه سفرهایی را در بر می‌گیرد که در یک بازه سرعتی و مسافتی یکسان قرار دارند و از سوی دیگر در بازه یکسانی از شبانه‌روز رخ داده‌اند.

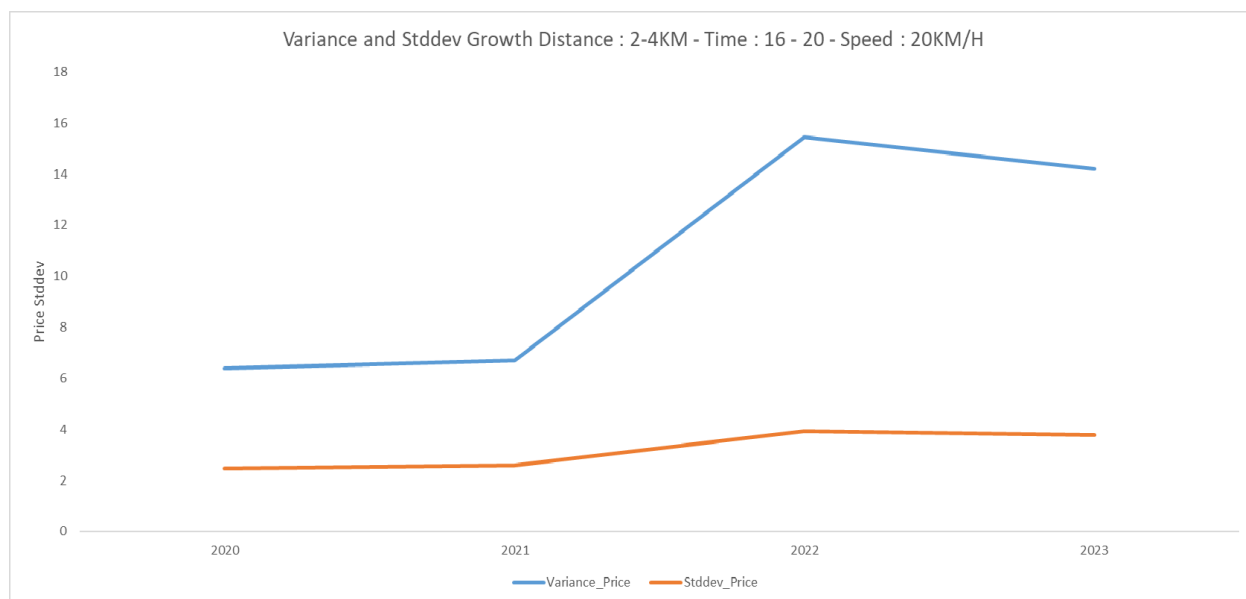
گزارش مبنای قیمت را هزینه کل سفر، `total_amount`، در نظر گرفته است.

برای ارزیابی بزرگی قیمت‌ها در مسیرهای مشابه از متریک‌های انحراف معیار و واریانس قیمت در مسیرهای مشابه در زمان‌های مشابه استفاده شد. این دو متریک هر چه برای یک مسیر و زمان مشابه بزرگتر باشند، نشان می‌دهند که تفاوت کرایه‌ها در یک مسیر و زمان مشابه بزرگتر است.

در راستای بررسی بزرگی تفاوت کرایه‌های مسیرهای مشابه در زمان‌های مشابه در گذر زمان می‌توان متریک تفاوت کرایه‌ها را برای هر زوج مسیر-زمان مشابه در هر سال محاسبه نمود.



نمودار ۱۴: روند بزرگی تفاوت کرایه در مسیرهای مشابه (مسافت: ۲-۴ کیلومتر، سرعت: ۲۰ کیلومتر بر ساعت) و زمان: ساعات ۱۲ تا ۱۶ برای نمونه در نمودار ۱۴ روند بزرگی تفاوت کرایه‌های برای مسیر مشابه، مسافت‌های دو تا چهار کیلومتر و سرعت ۲۰ کیلومتر در ساعت، در بازه زمانی ۱۲ تا ۱۶ نمایش داده شده است. چنان‌که آشکار است بزرگی تفاوت کرایه‌ها در گذر زمان در حال افزایش است، زیرا انحراف معیار کرایه‌ها در حال بزرگ‌تر شدن است. همین موضوع با توجه به نمودار ۱۵ برای همین مسیر مشابه اما در بازه ساعات ۱۶ تا ۲۰ نیز قابل مشاهده است. بخشی از این موضوع را می‌توان به تورم سالانه اقتصاد آمریکا نسبت داد که سبب می‌شود با بزرگ‌تر شدن اندازه کرایه‌ها انحراف معیار آن‌ها نیز افزایش یابد.



نمودار ۱۵: روند بزرگی تفاوت کرایه در مسیرهای مشابه (مسافت: ۲-۴ کیلومتر، سرعت: ۲۰ کیلومتر بر ساعت) و زمان: ساعات ۱۶ تا ۲۰

خواسته هفت: ناوگان ون

مبداهای پیشنهادی

برای شناسایی مبداهای مناسب برای تخصیص ون لازم است این موضوع را در نظر گرفت که مسافران ون لازم است با یکدیگر هم‌مسیر باشند. از این رو مسیرهایی که دارای بیشترین شمار سفر هستند، می‌توانند بهترین گزینه برای استقرار ون‌ها باشند. بنابراین متریک مدنظر برای شناسایی مبداهای شمار سفر در زوج مبدأ-مقصدهاست. هر چه مقدار این متریک بالاتر باشد، آن مسیر و مبدأ مربوط به آن گزینه بهتری برای تخصیص ون هستند.

برای استخراج این مبداهای نخست داده‌های سفر تاکسی‌ها به یک سال اخیر محدود شدند تا واپسین تحولات تردد مسافران در موقعیت‌های مکانی برای شناسایی مبداهای مناسب لحاظ شوند. با شمارش سفرها برای زوج مبدأ-مقصدها پنج محل به عنوان مبداهای مناسب شناسایی شدند که عبارتند از: Midtown Center، Upper East Side North، Upper East Side South، Lincoln Squar East و Upper West Side South.

برآورد تخفیف

از آن‌جا که ون‌ها شمار بیشتری از مسافران را جابه‌جا می‌کنند، از این رو می‌توان انتظار داشت که با سرشکن شدن هزینه‌های سفر میان شمار بیشتری از افراد کرایه کمتری سهم هر یک خواهد شد. یکی از بزرگترین هزینه‌های سفر به هزینه‌های مصرف سوخت بازمی‌گردد. فرض می‌شود که ون زمانی حرکت می‌کند که ظرفیت آن تکمیل شده است. از سوی دیگر بررسی داده‌ها نشان می‌دهد که بیش از ۷۵٪ سفرهای تاکسی‌ها با یک مسافر انجام شده است. ظرفیت هر ون ۱۰ نفر در نظر گرفته شده است. با توجه به این که ون‌ها به طور میانگین در هر مایل ۰/۰۵ گالن بو هر تاکسی در هر مایل ۰/۰۳ گالن بنزین مصرف می‌کند؛ بدین ترتیب با لحاظ بهای بنزین برابر با ۳/۱۵ دلار به ازای هر گالن می‌توان تفاوت سهم هر مسافر از هزینه‌های سوخت در تاکسی و ون را به قرار زیر برآورد نمود:

$$3/15 \times \left(0/03 - \frac{0/05}{10} \right) = 0/079 \frac{USD}{mile}$$

با توجه به محاسبات می‌توان گفت حداکثر تخفیفی که به مسافران ون می‌توان اعطا نمود، برابر با ۸ سنت در هر مایل است.

خواسته هشت: پیش بینی آینده

در آغاز برای هر تاریخ به تفکیک ساعات شبانه روز تعداد سفرهای انجام شده شمرده می شوند. شمار کل تاریخ-ساعاتها

۳۷۲۴۴ عدد است. با مرتب کردن داده بر پایه شمار سفرهای انجام شده یک درصد بالایی آن نگه داشته می شود.

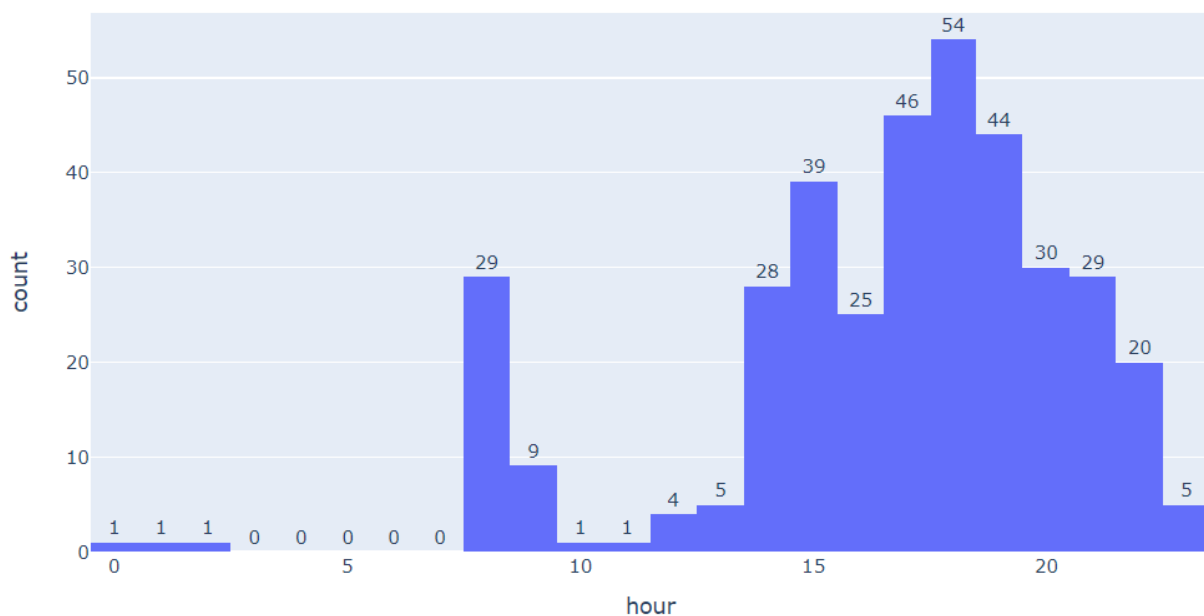
بیشترین تعداد سفر در تاریخ ۱۴-۰۲-۲۰۲۰ در ساعت ۱۸ اتفاق افتاده که شامل ۲۰۴۳۲ سفر بوده است.

برای ارزیابی عوامل مؤثر بر شمار بالای سفرها در یک درصد بالای داده ها در آغاز توزیع سفرها به تفکیک ساعات شبانه روز رسم

می شود. با توجه به نمودار ۱۶ بیشترین شمار روز-ساعاتها در ساعت اوج ترافیک نیویورک که طبق اطلاعات به دست آمده ۸ تا

۹ صبح و ۳ تا ۷ بعدازظهر است، رخ داده اند.

Distribution of Counts by Hour



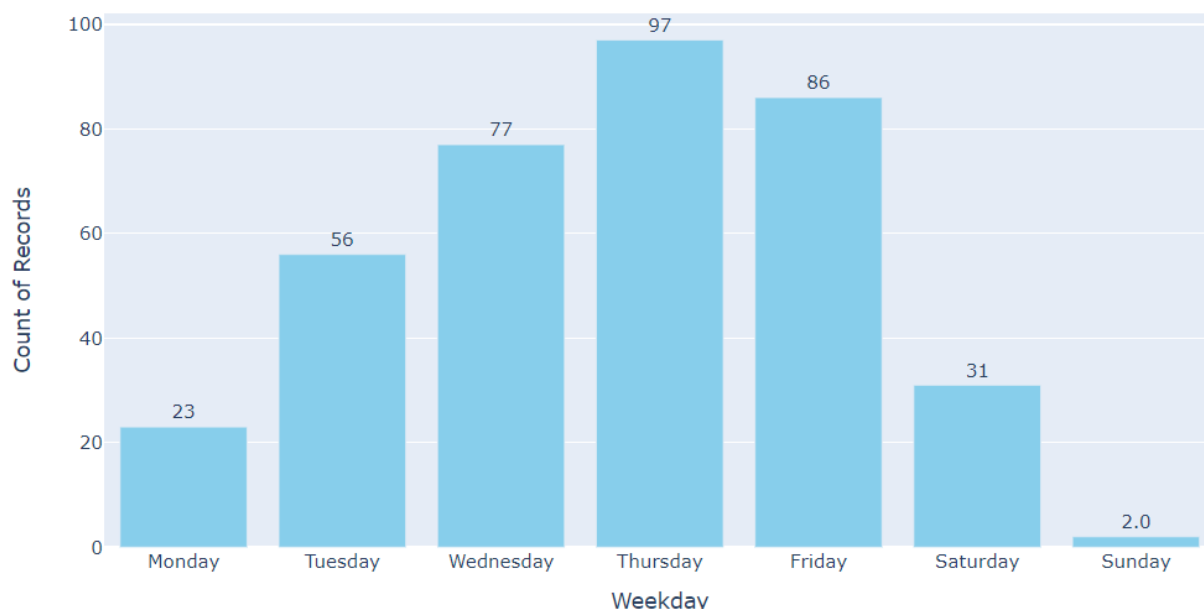
نمودار ۱۶: توزیع روز-ساعاتهای پرتدد به تفکیک ساعات شبانه روز

پس از آن توزیع شمار روز-ساعاتهای پرتدد به تفکیک روزهای هفته بررسی شدند. با توجه به نمودار ۱۷ بیشترین روز-

ساعاتهای پرتدد در روزهای پنجشنبه و جمعه و کمترین آن در روزهای شنبه، یکشنبه و دوشنبه است. به طور کلی ۸,۸۷ درصد

روز-ساعاتهای پرتدد مربوط به روزهای تعطیل و ۹۱/۱ درصد آن مربوط به روزهای کاری هفته است.

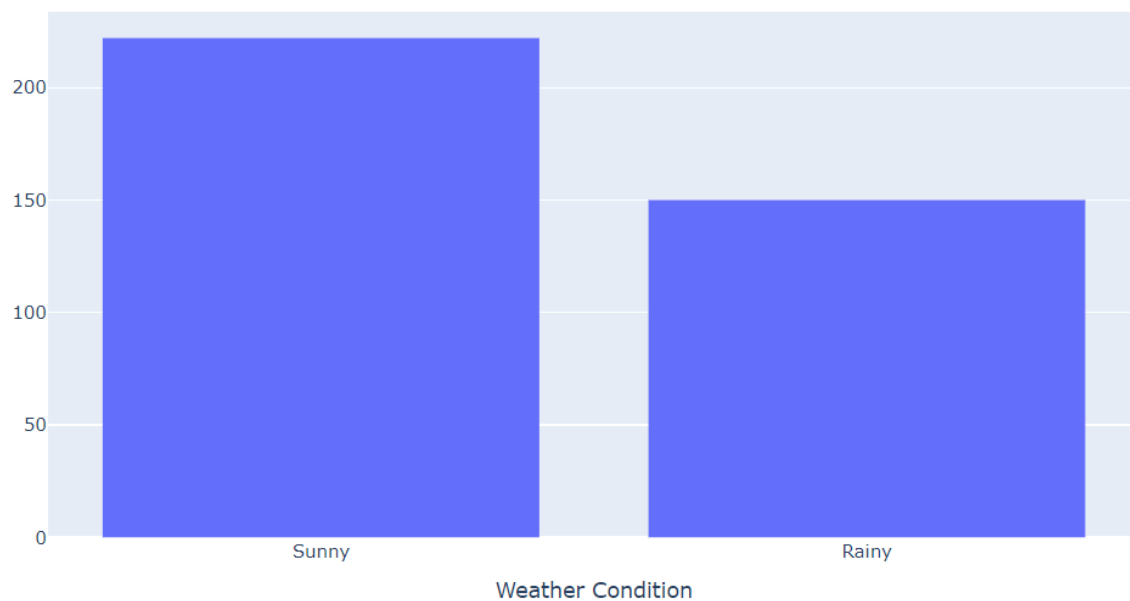
Distribution of Records by Weekday



نمودار ۱۷: توزیع روز-ساعت‌های پرتدد به تفکیک روزهای هفته

برای بررسی اثر وضعیت آب‌وهوا در تعیین روزهای پرتدد نخست اطلاعات بارش برای هر تاریخ به داده افزوده شدند. متغیر بارش به میزان بارش اشاره دارد که معمولاً در واحدهای میلی‌متر (mm) یا اینچ (in) اندازه‌گیری می‌شود. بارش شامل هر نوع رطوبتی است که از اتمسفر به سطح زمین می‌رسد، مانند باران، برف، تگرگ، و مه. اگر مقدار این سنجه بیشتر از یک باشد، نشان‌دهنده وجود بارش است. بر پایه این متغیر یک متغیر مجازی به نام وضعیت بارش ساخته شد که در صورت وجود بارش برابر یک باشد. نمودار ۱۸ شمار روز-ساعت‌های پرتدد را به تفکیک وضعیت بارش نشان می‌دهد. همان‌گونه که آشکار شمار قابل توجهی از روز-ساعت‌های پرتدد در شرایط بارانی رخ داده‌اند. ضریب همبستگی میان متغیر وضعیت بارش و شمار سفرها در این داده برابر ۰/۲۶ شد که نشان از همبستگی مثبت میان تردهای بیشتر در روز-ساعت‌های پرتدد در وضعیت‌های بارشی است.

Distribution of Records by Weather Condition



نمودار ۱۸: توزیع روز-ساعت‌های پرتردد به تفکیک وضعیت آب‌وهوا

در مجموع می‌توان گفت که در ساعات ترافیکی شبانه‌روز، روزهای کاری و وضعیت بارشی نسبت به روز و ساعات دیگر تقاضا برای سفر بیشتر است. بنابراین عرضه بیشتر در این ساعات می‌تواند موجب درآمدزایی بیشتر شود. لذا می‌توان برای تشویق راننده‌ها برای کار در این ساعات در پلتفرم مشوق‌هایی در نظر گرفت.