

## به نام خداوند بخشندۀ مهریان

تعیین موقعیت مکانی مناسب ایجاد کسب وکار جدید با استفاده از داده‌های مکانی شهر، توزیع جمعیت و کسب وکارهای فعلی

محمدمهدی فاریابی      میلاد آقاجوهری

۹۴۱۰۵۴۷۴      ۹۳۱۰۱۹۵۱

فاز پایانی پروژه: جمع‌آوری داده - تحلیل اکتشافی - بررسی آماری - مدلسازی - توسعه‌ی سیستم توصیه

## جمع آوری داده‌ها

برای جمع آوری داده‌های این تمرین در ابتدا سعی در کرال کردن داده‌های نقشه‌ی گوگل کردیم که با محدودیت‌های google maps API که به هر کاربر عادی اجازه‌ی ۲۰۰ radar scan را در طول روز می‌دهد روپروردیم و نتوانستیم این داده‌ها را crawl کنیم. بعد از مدتی جستجو با وبسایت <http://tehrantrafficmap.tehran.ir/> آشنا شدیم. داده‌های این وبسایت توسط شهرداری تهران به صورت سازمانی جمع‌آوری شده و به نظر داده‌ای کامل تر از داده‌ی گوگل است!

کرال کردن داده‌ی این سایت بسیار ساده بود اما داده‌ها رمزگذاری شده بودند که بعد از مشقت فراوان بالاخره موفق به یافتن روش رمزگذاری داده‌ها شدیم. داده‌ها توسط فرایندی تبدیل متنی می‌شدند و در نهایت هم توسط سیستم مختصاتی EPSG:3857 ذخیره شده بودند.

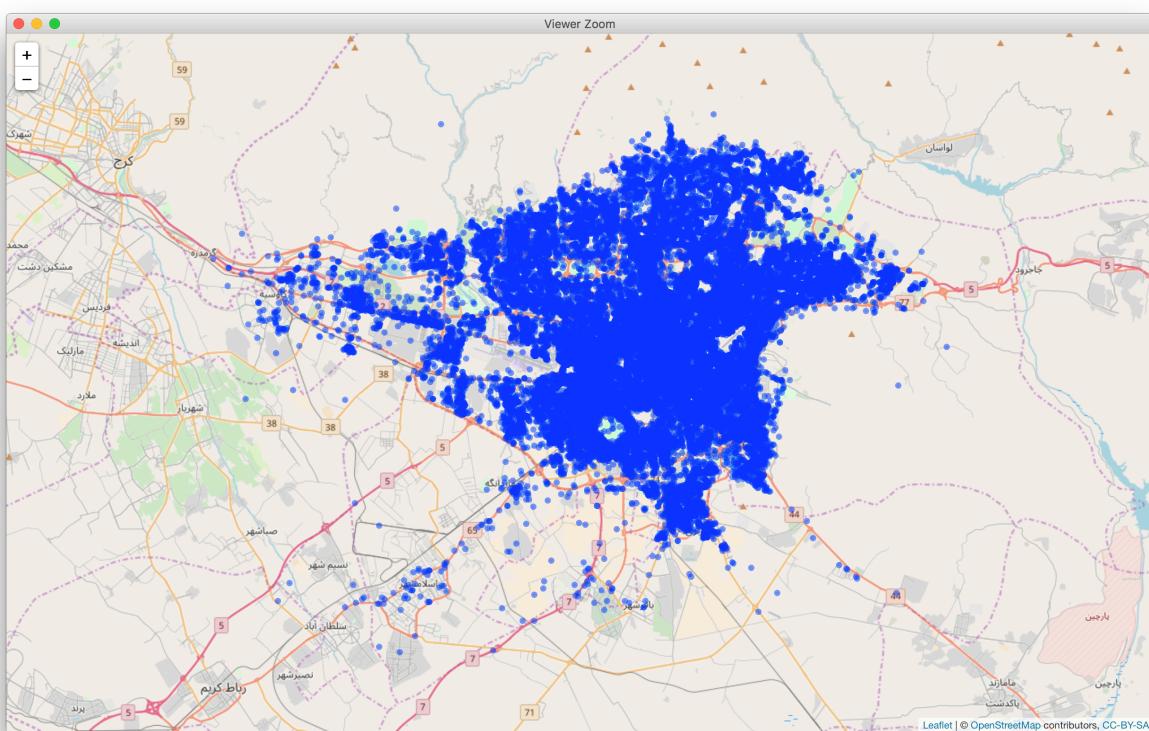
در نهایت پس از چند روزی تلاش موفق به نوشتمن یک کرالر کامل برای داده‌های این وبسایت شدیم.

```
(.venv) [mohammadmahdi:~/Desktop/tehran_data_parse]$ python3 crawl_data.py
~~~> Starting ...
~~~> crawling Restaurant ...
~~~> Restaurant Done in 48.84928011894226 seconds.
~~~> crawling FastFood ...
~~~> FastFood Done in 70.75677394866943 seconds.
~~~> crawling Kitchen ...
~~~> Kitchen Done in 13.713966131210327 seconds.
~~~> crawling Tabachi ...
~~~> Tabachi Done in 12.731689929962158 seconds.
~~~> crawling KababShop ...
~~~> KababShop Done in 31.190496683120728 seconds.
~~~> crawling CoffeeShop ...
```

از آنجایی که این داده متن باز نیست به دلیل رعایت حق مالک داده از ارائه‌ی کد کرالر در کنار گزارش معدویریم.

در نهایت کد کرالر داده‌ها را به صورت لیستی ای tuple های پایتونی در اختیار ما قرار می‌دهد که آن را با کد دیگری با استفاده از کتابخانه‌ی pandas به فرم DataFrame تبدیل می‌کنیم.

کد تبدیل به DataFrame به ضمیمه‌ی این گزارش کار ارسال می‌شود. داده‌ی جمع‌آوری شده نیز به همراه گزارش ارسال می‌شود.



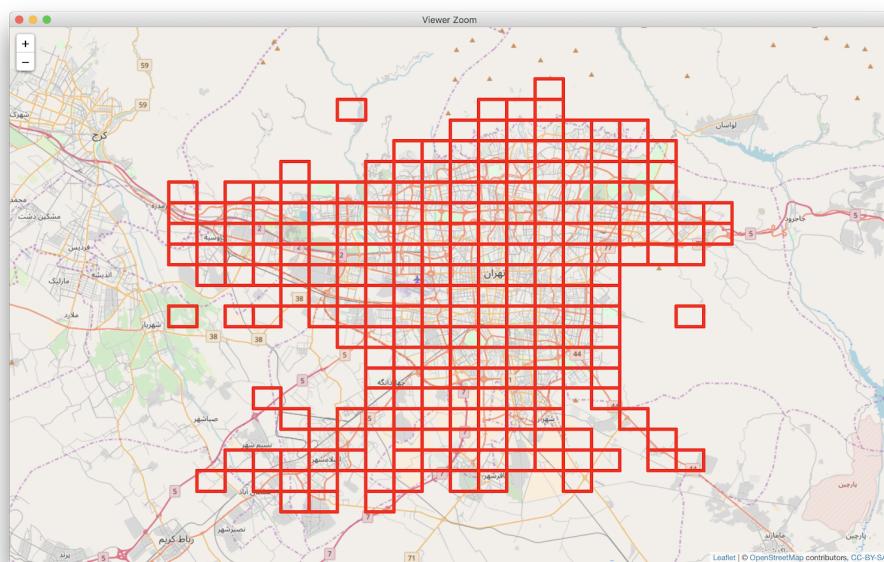
تمام نقاط این داده روی نقشه‌ی شهر تهران

## بصري سازي دادهها

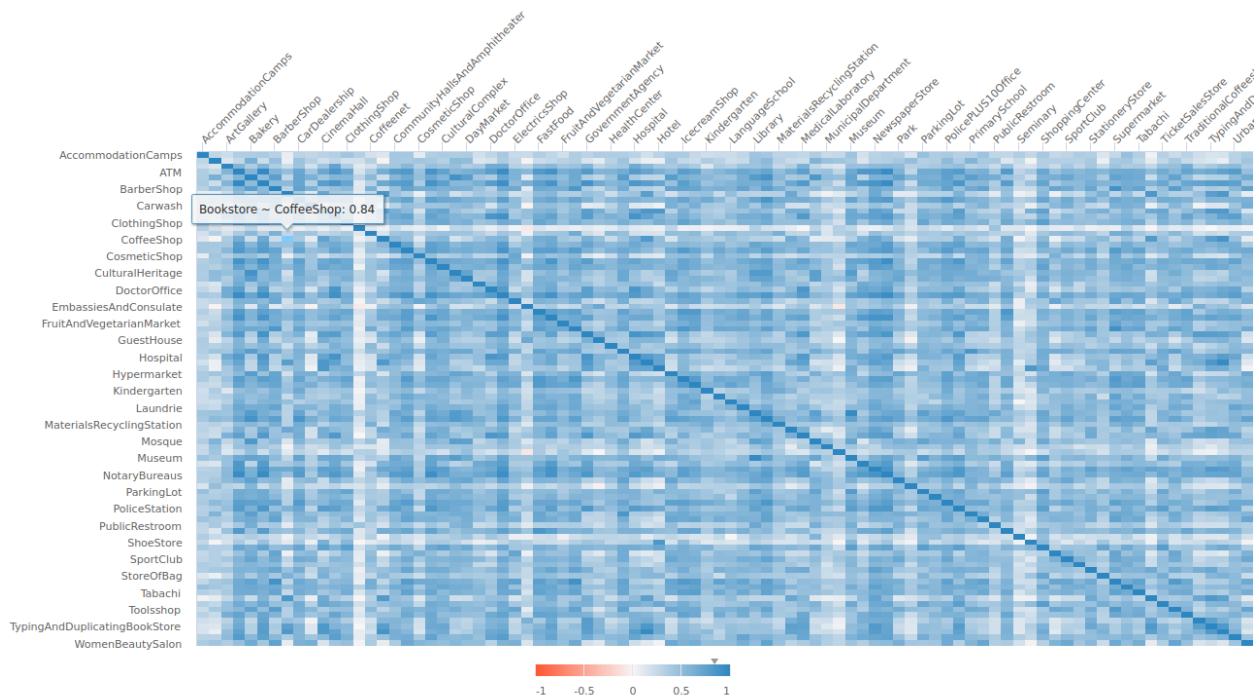
برای بصري سازی دادههای استخراج شده در این فاز از ابزارهای بصري سازی مانند leaflet و plotly، highchartr، ggplot2 استفاده کرده‌ایم.

### تحلیل اکتشافی دادهها: ضرایب همبستگی

ما ابتدا شهر تهران را به مناطقی مستطیلی شکل تقسیم میکنیم تصویرسازی این مستطیلی سازی در زیر قابل مشاهده است:



سپس در هر منطقه تعداد کسب و کارهای مختلف از قبیل رستوران، مدرسه، کافی شاپ، بیمارستان و ... را محاسبه می‌کنیم. حال می‌توان بین حضور مشاغل مختلف ماتریس همبستگی را حساب کنیم تا حسی از این بگیریم که چه مشاغلی در کنار هم دیگر زیاد تکرار شده‌اند. این گامی ابتدایی برای این خواهد بود که وضعیت قرارگیری مشاغل در کنار هم‌دیگر را بسنجدیم. تصویرسازی تعاملی ماتریس همبستگی در لینک زیر قابل دسترسی است. عکسی از آن در این تصویر گنجانده شده است:



همانطور که دیده می‌شود اولین چیزی که در این تصویر جلب توجه می‌کند عدم وجود همبستگی‌های منفی است که گویا بدین معنی است که مشاغل اینطور نیستند که حضور یکی مانع حضور دیگری باشد و برای مثال رستوران داشتن یک منطقه دلیل بر کم بودن مدرسه در آن منطقه شوند بلکه در واقع به این صورت هستند که در صورت وجود کتاب‌فروشی در یک منطقه به احتمال زیاد در آن منطقه کافی‌شایپ هم وجود دارد (همانطور که در تصویر هم دیده می‌شود ضریب همبستگی بین این دو 0.84 است).

### تحلیل اکتشافی داده‌ها: بررسی نام‌های هر کسب و کار

از آنجایی که پژوهشی تحلیل داده‌ی ما در واقع برای کمک به کسب و کارها شروع به کار کرده است، یکی از کمک‌های بزرگ می‌تواند کمک به انتخاب نام برای کسب و کارها باشد. در حال حاضر بسیاری از کسب و کارها از اسم‌های مشابه استفاده می‌کنند و این نوعی یکنواختی و همنزگی به آن‌ها داده که برای کسب و کاری که می‌خواهد تمایز دیده شود سمی مهلك است. برای مثال اگر شما می‌خواهید رستورانی متفاوت بزنید احتمالاً نام‌های نایب و دیزی و شاندیز برای شما تمایز ایجاد نمی‌کنند (مگر اینکه شعبه‌ای از آن‌ها باشید یا بخواهید از اعتبار آن‌ها استفاده کنید). برای اینکه حسی نسبت به این داشته باشیم که چه کلماتی در اسم رستوران‌ها یا کسب و کارهای مختلف بیشتر استفاده می‌شود تا بتوانیم اسم مناسب‌تری انتخاب کنیم نمودارهایی شبیه نمودار زیر مفید هستند:

## نامهای پرکاربرد در رستوران‌های تهران

ترک جو نزیباد بادک بلو ایاس تهیه  
کندو بلو طغ زدایی جوان  
جی پر دیس آنها  
کتاب آنا / پیالا پیان  
عربی نایب سنتره سنتور  
تالار دیزی با غ طلا نی کتابی  
مللی خان رفتاری  
تبیرین پا یاخت هانی تهران بلان  
ضیافت زنجیر بهار قرنج  
پارسیان

پر استفاده‌ترین نام‌ها برای مراکز درمانی-بهداشتی

۱۰



در بالا پر استفاده ترین نام‌ها در کافی‌شایپ‌های تهران و در پایین پر استفاده ترین نام‌ها در مراکز مدارس (مدارس موسیقی و غیره هم منظور هستند) تهران را می‌بینید:



## تحلیل اکتشافی داده‌ها: بررسی وضعیت شهر تهران از نظر تعداد اماکن مختلف

در ابتدا شهودی به دست می‌آوریم که از هر کدام از مکان‌ها و کسب و کارها چند عدد در شهر تهران موجود هستند.



همانطور که می‌بینید البته آنچه در اینجا آمده است الزاماً مکان نیست و کسب و کارهای مختلف هستند که البته مشکلی هم برای ما ایجاد نمی‌کنند.

## تحلیل اکتشافی داده‌ها: بررسی وضعیت نامهای مختلف در شهر تهران

این قسمت البته بیشتر برای این است که حسی از فضای شهر تهران بگیریم، نامهای پر استفاده در اماکن تهران بر روی برج میلاد:



تحلیل اکتشافی داده‌ها: بررسی مقایسه‌ای نام‌ها در پایین شهر و بالا شهر تهران

این‌ها نام‌های پراستفاده در بالا شهر تهران هستند:

ـ زیبایی پارسیان بربری خمینی اجتماعی  
ـ جعفر میوه روز سامان پروتئینی  
ـ فرهنگی پلیس + ۱ ابوالفضل زبان  
ـ شهدای آموزش مطبوعات فرهنگ محمد  
ـ پلیس حسن شماره مسکن علی دولت  
ـ ابن خدمات بهداشتی فروش آقتصاد  
ـ حلیم بلطف فروشی خودرو سپه  
ـ مهرآباد ملت تهران آمام  
ـ تجارت ملی حاشید مجاز  
ـ الله پیش استناد کلینیک دفتر  
ـ رضا تحریر شهر رسمی منطقه رفاه  
ـ برق مردانه صادرات مواد سایپا  
ـ حسین ازدواج ورزشی ایران پژوهشکی  
ـ انصار بزرگ بنزین فود بدکی جنوب  
ـ بانوان دکتر بستنی بازارگار طلاق  
ـ امور حضرت خوان تخصصی علمی  
ـ پرورش پوشانگ الزمان بهداشت  
ـ نوین اسلامی کشاورزی بزرگ کوروش

و این‌ها نام‌های پراستفاده در پایین شهر تهران هستند:

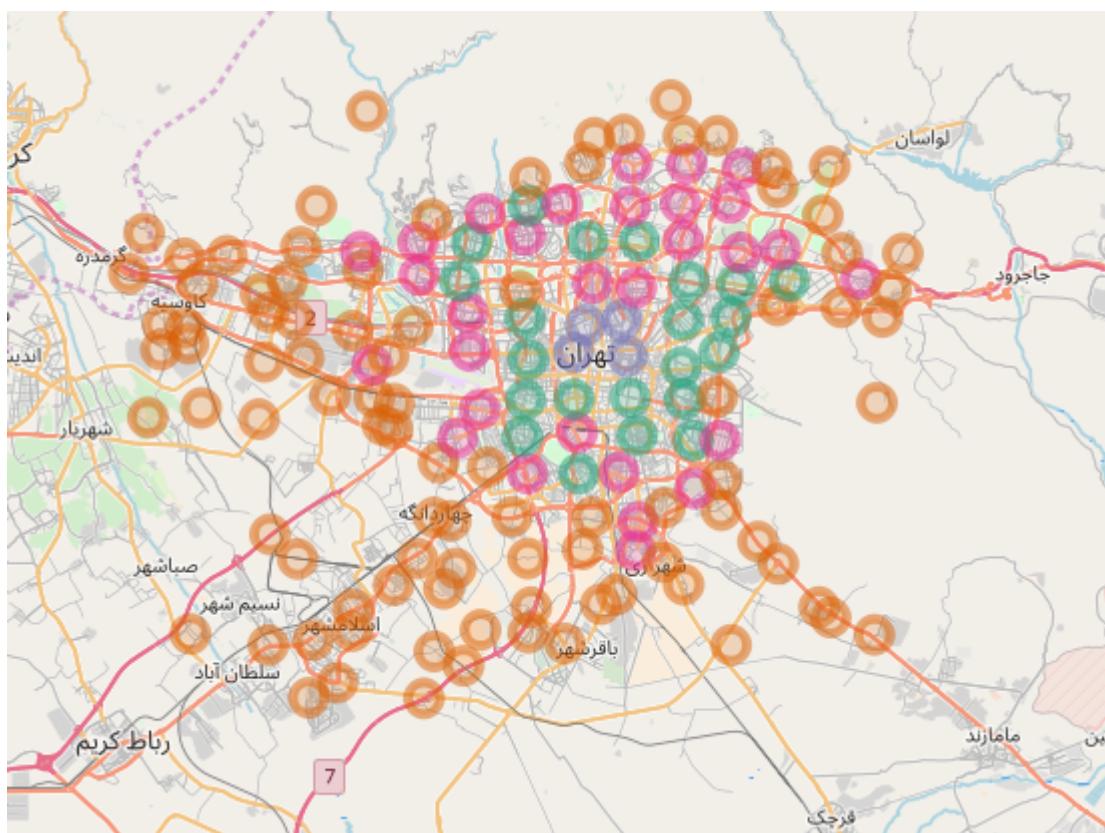
ـ درمانی پیش دولت پلاک انتظامی  
ـ پروتئینی پلیس + ۱ سرمایه شبانه فرهنگ  
ـ فانتزی پلیس + ۱ سرمایه شبانه فرهنگ  
ـ شهرک روزی بیمارستان انصار شویی  
ـ ترمه بار بستنی شماره رفاه امور خوان  
ـ واحد اسلامی پژوهشکی نوین شاب  
ـ بربری آینده صادرات روز نوین شاب  
ـ سنتی پارسیان بهداشتی خدمات سینما  
ـ بليط علوم ملت رسمی سبزی میوه نور  
ـ علمی مسکن فود بانک ملی دکتر کفش  
ـ عمومی تهران استناد شهر سپه  
ـ مواد پاسارگاد فروشی های  
ـ امام خودرو ایران ورزشی استان  
ـ مطبوعات شهید دانشگاه پوشان  
ـ حلیم فروش تخصصی محله بنزین  
ـ توسعه موسیقی کشاورزی تحریر مرکز  
ـ منطقه فرهنگی مردانه دندانپزشکی  
ـ پژوهش آزاد قوامیں مرکزی ایرانیان  
ـ گردشگری جمهوری ایرانیان

تفاوت‌ها بسیار جالب هستند اما از آن‌ها می‌گذریم.

# تحلیل آماری

دسته‌بندی داده‌ی اماکن نقاط مختلف شهر با استفاده از کاهش بعد و استفاده از الگوریتم‌های clustering

روش کار به این صورت است که همانطور که در بالا گفته شد که تهران را به بخش‌های مستطیلی تقسیم کرده‌ایم و برای هر بخش تعداد انواع مختلف اماکن را محاسبه کرده‌ایم. حال بر روی همین داده از الگوریتم PCA برای کاهش بعد استفاده کرده‌ایم و سپس روی این داده‌ی کاهش بعد یافته از الگوریتم kmeans برای تقسیم مناطق تهران به چندین بخش استفاده کرده‌ایم که نتیجه‌ی زیر به دست آمده است. هدف از اجرای این کار این بوده است که بتوانیم مناطق تهران را به چندین بخش متفاوت تقسیم کنیم. تفسیر ما از مدل به دست آمده این است که مناطق کم جمعیت و کم تراکم تهران که فقیر نشین بوده و بانک کم داشته‌اند جداسده، مناطقی همانند ولی عصر که پر از انواع اماکن هستند جدا شده، مناطقی که کسب و کارها کم هستند اما بانک‌ها خیلی زیاد هستند مثل بالای شهر جدا شده و مناطق عادی و معتدل هم جدا شده‌اند.



## تحلیل آماری

ساخت سیستم توصیه‌ی مکان ایجاد کسب و کار جدید با دریافت نوع کسب و کار با توجه به داده‌ی شهری تهران

هدف پژوهشی ما از ابتدا توسعه‌ی سیستمی برای پیشنهاد مکان و موقعیت جغرافیایی مناسب برای ایجاد یک کسب و کار جدید بوده است. برای این منظور ابتدا تهران را همانند شکل ابتدای گزارش کار به چندین بلوک تقسیم بندی می‌کنیم. فرض می‌کنیم هر بلوک نماینده‌ی یک محل باشد. حال بررسی می‌کنیم که در هر محل چه کسب و کارهایی وجود دارد و از هر کدام چند تا در آن محل مشغول به فعالیت هستند. داده را spread می‌کنیم و به ازای هر نوع کسب و کار یک ستون به آن اضافه می‌کنیم. هر سطر داده معادل یک محل شهر تهران خواهد بود. این داده را به عنوان یک سبد transaction استفاده خواهیم کرد که هر item آن یک کسب و کار و هر transaction آن یک محل شهر خواهد بود.

به ازای تمامی خانه‌های غیرصفر این جدول TRUE و به ازای سایر خانه‌ها FALSE می‌گذاریم.

الگوریتم Apriori را برای یافتن rule های association اجرا می‌کنیم. این الگوریتم به ما تعدادی rule خواهد داد. این rule ها را فیلتر می‌کنیم طوری که تنها rule هایی باقی بمانند که کسب و کار مورد نظرمان در سمت راست آنها قرار دارد. این rule ها را بر حسب lift مرتب می‌کنیم. و ۱۰ rule با lift بالا را انتخاب می‌کنیم.

هر کدام از این rule ها تعدادی کسب و کار در سمت چپ و کسب و کار هدف ما را در سمت راستشان دارند. در محل‌های شهر تهران دنبال محل‌هایی می‌گردیم که تمامی کسب و کارهای سمت چپ rule ها در آنها حضور داشته باشند. سپس این محل را به این صورت ارزش‌گذاری می‌کنیم که محل‌هایی که تعداد کسب و کار هدف ما در آنها کمتر است ارزش بالاتری داشته باشند.

در نهایت محل‌ها را بر اساس ارزش به ۳ دسته‌ی سبز و آبی و قرمز تقسیم بندی می‌کنیم که به ترتیب از پر ارزش ترین به کم ارزش ترین قرار دارند.

