به نام خدا

بوت کمپ تحلیل داده

گزارش پروژه اول

اعضای گروه

مهسا کارگر

امین شهنانی

جواد جهانگیری

علی بیات

زهرا الوندی

سرگروه

مهدی بانی

۱۹ بهمن ۱۴۰۱

فهرست مطالب

صورت پروژه

هدف این پروژه استخراج داده‌هایی از کافی شاپ های سراسر کشور و تحلیل آن هاست. برای انجام این پروژه اطلاعات رستوران های کشور را به دست آورده و پس ذخیره سازی آنها در پایگاه داده، اقدام به واکشی اطلاعات و تحلیل آنها می‌کنیم.

وبسایت **دلینو** به عنوان منبع اصلی این پروژه انتخاب شده است؛ بنابراین با بررسی و استخراج اطلاعات از همین سایت نیازمندیها و خواسته های پروژه را برآورده می‌کنیم. این پروژه در گام‌های زیر تعریف شده است:

- استخراج اطلاعات از سایت دلینو با web scraping
- طراحی شمای کلی جداول و ارتباط آنها یکی دیگر ذخیره در پایگاه داده Data Base
- تحلیل دیتاها و مصور سازی آن‌ها در Power BI

استخراج داده‌ها

با استفاده از فایل web_scrappinga.py عملیات web scraping را روی سایت دلینو انجام می‌دهیم و لینک هرکدام از رستوران‌ها از سایت دلینو بدست آمد. در فایل restaurant_links.txt لیست لینک‌های همه رستورانها برای انجام وب اسکرپینگ که توسط فایل web_scrapping.py بدست آمده است، قرار دارد.

با استفاده از لینکهای بدست آمده از فایل web_scrapping.py دیتاهای مربوط به هر کدام از رستورانهای بدست آمد و در فایل delino_api_data_csv.csv ذخیره شد. سپس لیست منوها را با کد موجود در فایل data_clean.ipynb از فایل delino_api_data_csv.csv و از ستون menu استخراج شد و در فایل menu.csv ذخیره گردید.

پیاده سازی دیتابیس

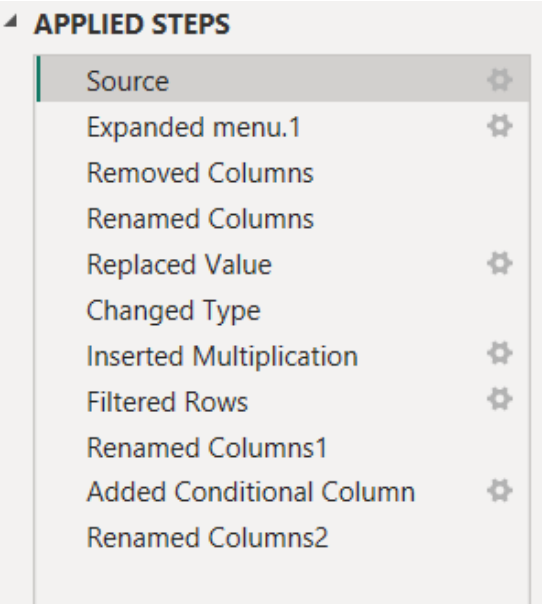
با استفاده از فایل csv_to_mysql.ipynb و با کتابخانه mysql.connector دیتاهای استخراج شده از سایت دلینو در دیتابیس لوکال قرار می‌گیرند. سپس با استفاده از فایل new_csv_to_mysql.ipynb تلاش شد تا با استفاده از کتابخانه SQLAlchemy دیتاهای موجود در فایل csv مربوط به اطلاعات رستوران‌ها و منوهای آن در دستابیس معرفی شده توسط کوئرا قرار بگیرد. در اتصال به دیتابس و ایجاد جداول موفق بودیم اما در درج اطلاعات ناموفق بودیم. فایل group_11_bk.nbakmysql نسخه بک آپ از دیتابیس لوکال می باشد.

تحلیل دیتاها در Power BI

تمیزکردن دیتاها

دیتا های دو فایل delino_api_data_csv و menu را با یکدیگر بر اساس دستور Merge و با ستون آیدی رستوران ها یکی می‌کنیم و جدول نهایی را Main_Table می‌نامیم. برای مرتب سازی داده‌ها، کارهای زیر را انجام می‌دهیم:

- حذف اسامی رستوران هایی که مشکل دار
- تغییر فرمت ستون امتیازها به نوع عددی
- ایجاد ستون Popularity (از حاصل ضرب امتیاز در تعداد نظرات بدست می‌آید.)
- ایجاد ستون Categorized Menu
- تغییر نام ستون های فایل خام اولیه
- حذف ستون های اضافه برای پردازش سریعتر داده ها در صفحه اصلی



این مراحل در بخش سمت راست Power Query قابل مشاهده است.

پس از انجام این کارها، گزینه Apply and Close را می‌زنیم.

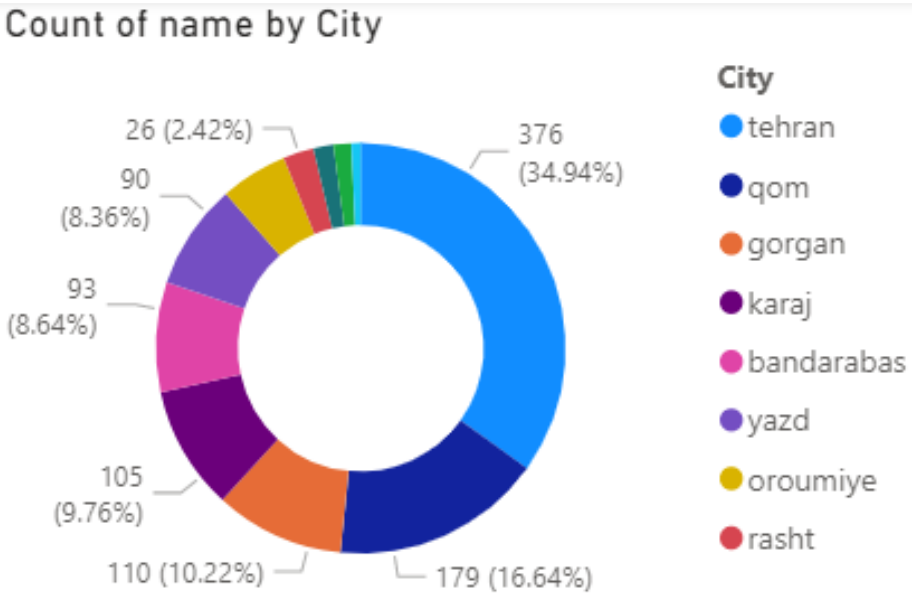
تحلیل و مصورسازی داده ها

- Count of city by name

این نمودار تعداد رستوران های هر شهر را به تفکیک روی نقشه کشوری نمایش می‌دهد. همانطور که مشخص است، رستوران های ثبت شده تهران بیشتر از سایر مناطق است.

City	Count of name
tehran	376
qom	179
gorgan	110
karaj	105
bandarabas	93
yazd	90
oroumiye	57
rasht	26
arak	17
gonbad kavoos	15
hamedan	8
Total	**1054**

نمودار دونات نیز با همین عنوان ترسیم شده است :

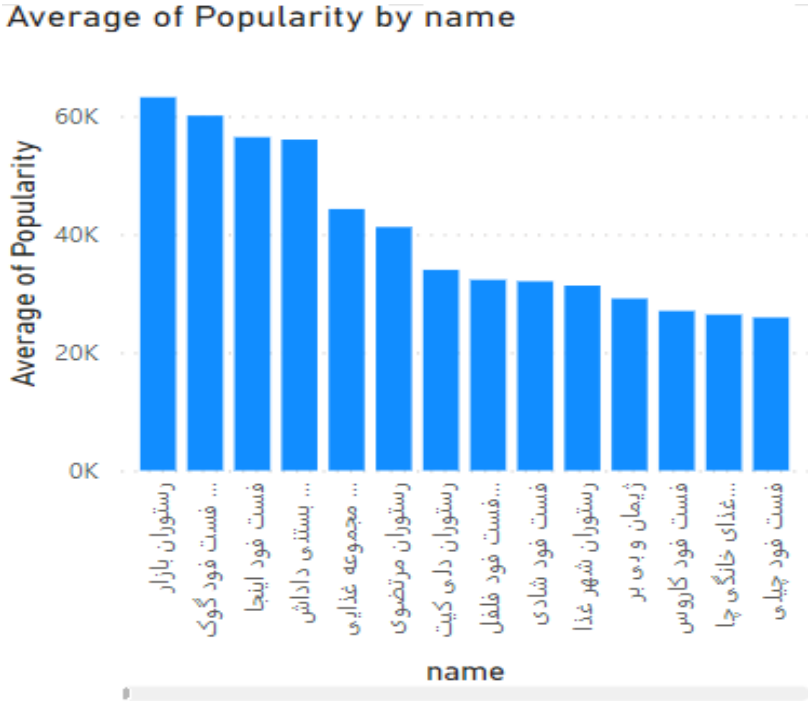


- Average of Rate by city

این نمودار (Funnel) ترتیب شهرهایی که رستوران هایی با بیشترین امتیاز دارند را از بالا به پایین نمایش می‌دهد.

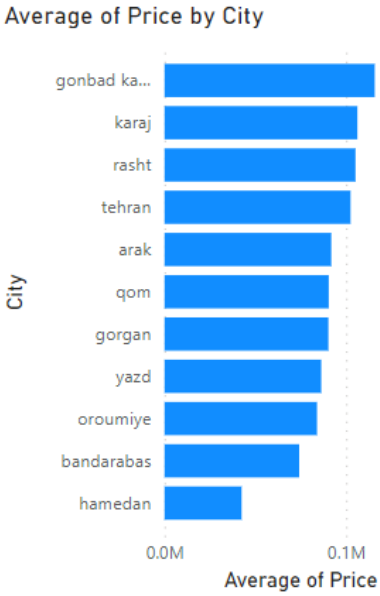
- Average of Popularity by name

این نمودار لیست رستوران هایی با بیشترین محبوبیت را نمایش می‌دهد.



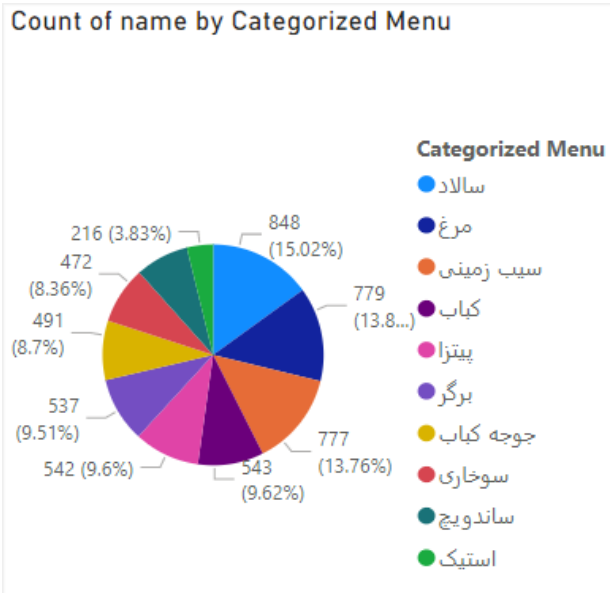
- Average of Price by city

این نمودار لیست شهرهایی با رستوران های گران قیمت را از بالا به پایین نشان می‌دهد.



- Count of name by Categorized Menu

این نمودار لیست غذاهایی با بیشترین سرو را در رستوران های مختلف به صورت یک نمودار دایره ای نشان می‌دهد.



این فایل با نام Visualisation.pbx در زیپ پروژه قرار داده شده است.

موفق و پیروز باشید