دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس

تحلیل داده

دکتر محمدامین صادقی-دکتر محمدرضا ابوالقاسمی

طراح تمرین: فرزانه حاتمی نژاد

زمان بارگذاری تمرین: چهارشنبه ۱۱ آبان

تمرین شماره ۳

موضوع تمرین: تحلیل داده و مصورسازی

نیمسال اول سال تحصیلی ۱۴۰۱ - ۱۴۰۲

بخش اول EDA and Visualization –

در این پروژه قصد داریم داده های دیتاست taxis را تحلیل و مصورسازی نماییم. این شکل از تحلیل را EDA یا Exploratory Data Analysis می نامند و همچنین به کمک مصور سازی نیز می توان اطلاعات مفیدی از دیتاست استخراج نمود.

با توجه به این امر، در این بخش به تحلیل و بررسی دیتاست مورد نظر می پردازیم. به سوالات زیر در نوت بوک پاسخ دهید. لازم به ذکر می باشد برای دریافت نمره ی کامل در این بخش،می بایست جواب ها و نمودار های خود را تحلیل نمایید.

سوالات

۱- ابتدا داده های لود شده را بررسی نمایید و چند سطر اول آن را نمایش دهید، مطمئن شوید لود دیتا به درستی صورت گرفته و نوع هر ویژگی را گزارش کنید.

۲- تعداد کل سفر های انجام شده را به دست آورید.

۳- تعداد 'pickup-zone' های یکتا را به دست آورید. (به کلمه ی یکتا دقت نمایید.)

۴- پرتکرار ترین 'dropoff-zone' به دست آورده و همراه با تعداد آن ها گزارش دهید. هدف به دست آوردن dropoff-zone می باشد که مقصد بیشتر سفر ها آن جاست.

a. پنج 'dropoff-zone' که بیشترین تکرار و پنج 'dropoff-zone' که کمترین تعداد را داشته اند، در دو نمودار مجزا به صورت sort شده رسم کنید.

b. دو نمودار را با یکدیگر مقایسه کرده و تحلیل کنید.

۵- درصد 'dropoff' و 'pickup' را در هر روز هفته (شنبه، یکشنبه و...) محاسبه کرده و در یک نمودار نمایش دهید. نمودار خود را تحلیل نمایید.

۶- تعداد 'pickup' و 'dropoff' بر اساس ساعت روز را محاسبه نموده و نمودار مناسب را رسم کرده و تحلیل نمایید.

۷- پر تکرار ترین روش پرداخت کرایه بر حسب روز های هفته را گزارش دهید. می خواهیم بدانیم در هر روز هفته کدام روش بیشتر استفاده شده است. و با نمودار مناسب نمایش داده و تحلیل کنید.

۸- تعداد سفر های مسافران به صورت گروه بندی شده به صورت زیر گزارش دهید:

a. هدف پیدا کردن تعداد مسافرانی است که تعداد سفر آن ها مقدار مشخصی دارد.(حداکثر ۲۰۰ سفر- بین ۲۰۰ تا ۵۰۰ سفر - حداقل ۵۰۰ سفر.)

b. پس از به دست آوردن دسته بندی آن را با نمودار مناسب نمایش داده و تحلیل نمایید.

۹- تعداد سفر هایی که 'dropoff-zone' مشخصی دارند در مقایسه با تعداد سفرهایی که 'dropoff-zone' مشخصی ندارند را گزارش دهید.

۱۰- پرتکرارترین رنگ تاکسی در بین سفرهایی با 'dropoff-zone' مشخص چه رنگی می باشد؟ نمودار مناسب رسم کنید.

بخش دوم -کاهش بعد

فرض نمایید می خواهیم در یک شرکت تاکسی آنلاین، مقدار کرایه هر سفر را بر اساس ویژگی هایی که داریم تخمین بزنیم. هدف آن است که ویژگی های موجود را کاهش دهیم. در ادامه به بررسی دو روش می پردازیم:

۱- می خواهیم چند ویژگی را انتخاب نماییم و بقیه را حذف کنیم.نمودار مناسب را رسم نموده و بگویید کدام ویژگی هارا انتخاب می کنید. نمودار خود را تحلیل کنید.

۲- راه حل دیگر برای کم کردن بعد های مورد نظر استفاده از روش هایی مانند PCA می باشد.

a. دیتا را به فضای دو بعدی برده و آن را رسم نمایید.

b. بهترین تعداد بعد را برای اعمال کردن روش PCA بر اساس واریانس تجمعی را می خواهیم به دست آوریم. نمودار آن را رسم نمایید.

c. پس از به دست آوردن تعداد بعد مناسب آن را بر روی دیتاست اعمال نمایید.

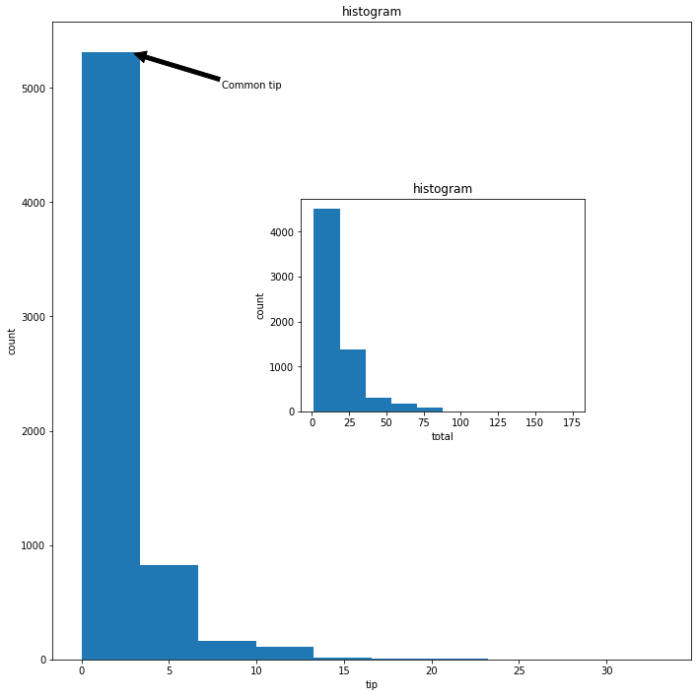
۳- هر کدام از دو روش بالا را به صورت مختصر توضیح داده، نتایج آن هارا مقایسه کنید. ومزایا و معیایب هر کدام را نام ببرید.

بخش سوم – استفاده از Figure

در این بخش می خواهیم با کاربرد Figure ونحوه ی استفاده از آن بیشتر آشنا شویم.Figure یک object است که تمامی المان های مربوط به plot را داراست. می توانیم از طریق آن محور های مختلفی تعریف نماییم. نمودار نهایی شما می بایست شبیه نمودار نشان داده شده در شکل ۱ باشد.

برای رسم نمودار مورد نظر مراحل زیر را انجام دهید:

- ابتدا یک figure با سایز (۱۰, ۱۰) تعریف نمایید.
- هیستوگرام tips را رسم نمایید.
- یک axes جدید درون نمودار قبل ایجاد نمایید.
- هیستوگرام total را رسم کنید.
- متداول ترین میزان tip را ب یک فلش نمایش دهید.



شکل ۱: استفاده از figure برای رسم نمودار

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز یکشنبه ۲۲ آبان ماه خواهد بود.
- در این تمرین تحلیل نمودار ها و پاسخ های به دست آمده از اهمیت زیادی برخوردار می باشد.
- انجام این تمرین به صورت یک نفره می باشد.
- خروجي مورد انتظار تمرین فایل jupiter می باشد.
- هر گونه توضیحات و گزارشنویسي را به صورت Markdown داخل کتابچه jupiter انجام دهید.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW3_[Lastname]_[StudentNumber].zip

برای مثال: HW3_Hatami_123456.zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:
farzaneh.hatami@ut.ac.ir