در این پروژه قصد داریم تا با هوش تجاری[1] آشنا شویم. در شرکت‌های تجاری وظیفه تحلیلگر داده‌های تجاری[2] پردازش لاگ‌ها، بدست آوردن متریک‌های مناسب، تحلیل متریک‌ها و دید[3] دادن به مدیران محصول[4] راجع به محصول است.

بخش اول

در این بخش شما باید از اکشن لاگ داده شده متریک‌های گفته شده را محاسبه کرده و تحلیل خود را بیان کنید.

اکشن لاگ داده شده مربوط به تراکنش‌های کاربران یک برنامه بانکی است که پس از ناشناس‌سازی[5] در اختیار شما قرار گرفته است.

در فایل Transactions.csv هر خط نشاندهنده یک تراکنش است و شرح ستون‌های آن به صورت زیر است:

- UserID آیدی کاربری که تراکنش را انجام داده است که برای هر کاربر یکتاست.
- ChannelID آیدی کانالی که تراکنش در آن رخ داده است.
- Date تاریخ تراکنش (شمسی) به صورت به هم چسبیده آورده شده است. یعنی 1398/03/24 به صورت 13980324 آورده شده است.
- Time زمان انجام تراکنش به صورت به هم چسبیده آورده شده است. یعنی 21:03:41 به صورت 210341 و همچنین 301 به معنای 00:03:01 است.
- Paid Amount ارزش تراکنش به ریال است.

در فایل Channels.csv اسم هر ChannelID آورده شده است.

تمیزکردن داده

تمیزسازی‌هایی که داده نیاز دارد را انجام دهید و هر کدام را شرح و علت آن را بیان کنید.

محاسبه متریک

متریک‌های گفته شده را بدست آورده و نمودارهای مربوطه را رسم و در گزارش بیاورید.

1. تعداد و ارزش تراکنش‌های روزانه برای ۳ ماه آخر
   (بازه ۳ ماهه نباید بصورت دستی مشخص شود و باید توسط کد پیدا شود)
2. تعداد و ارزش تراکنش‌های ساعتی برای ۴۸ ساعت آخر

[1] Business Intelligence
[2] Business Data Analyst
[3] Insight
[4] Product Managers
[5] Anonymized

3. تعداد مشتریان ماهیانه
4. درآمد هفتگی (فرض کنید ۱۰ درصد هر تراکنش را به عنوان کارمزد دریافت کنیم)
5. محاسبه نرخ ماندگاری[6] ماهیانه

این معیار مشخص میکند چند درصد کاربران فعال (دارای حداقل یک تراکنش) ماه گذشته این ماه برگشته‌اند (حداقل یک تراکنش انجام میدهند).

6. جدول کوهورت[7] ماهانه را برای کاربران بدست آورید (فرض کنید اولین تراکنش کاربر زمان نصب اپلیکیشن است)
7. تراکنش‌های غیر اول (تراکنش‌هایی که اولین تراکنش کاربر نیستند) بیشتر در چه کانال‌هایی رخ داده‌است، توزیع ۱۰ تای اول را بدست بیاورید.

تحلیل نتایج

تحلیل و دید[8] خود را به عنوان یک تحلیلگر داده در مورد هر یک از نمودارهای بالا بیان کنید. (روند[9]، تغییرات فصلی[10] و ...)

برای سوال ۷ محاسبه را برای تراکنش‌های اول انجام دهید و توزیع بدست آمده را مقایسه و تحلیل کنید. ماتریس جابه‌جایی بین Channel های تراکنش‌های اول و دوم را بدست آورده و به صورت هیت مپ[11] رسم کنید. هر المان این ماتریس (a_{ij} سطر i و ستون j) نشان‌دهنده این است که در چند درصد مواقع از Channel i که اولین پرداخت را داشته کاربر برای دومین پرداخت به Channel j رفته است.

بخش دوم

در این بخش شما باید لاگ سیستمی داده شده را پردازش کنید.

در فایل SSH.zip یک فایل SSH.log قرار گرفته است که شامل لاگ سیستمی مورد نظر است.

قسمتی از این لاگ را در شکل زیر مشاهده میکنید:

```
Dec 10 07:08:28 LabSZ sshd[24208]: reverse mapping checking getaddrinfo for ns.marryaldkfaczcz.com [173.234.31.186] failed - POSSIBLE BREAK-IN ATTEMPT!
Dec 10 07:08:28 LabSZ sshd[24208]: Invalid user webmaster from 173.234.31.186
Dec 10 07:08:28 LabSZ sshd[24208]: input_userauth_request: invalid user webmaster [preauth]
Dec 10 07:08:28 LabSZ sshd[24208]: pam_unix(sshd:auth): check pass; user unknown
Dec 10 07:08:28 LabSZ sshd[24208]: pam_unix(sshd:auth): authentication failure; logname= uid=0 euid=0 tty=ssh ruser= rhost=173.234.31.186
Dec 10 07:08:30 LabSZ sshd[24208]: Failed password for invalid user webmaster from 173.234.31.186 port 39257 ssh2
Dec 10 07:08:30 LabSZ sshd[24208]: Connection closed by 173.234.31.186 [preauth]
```

این لاگ به فرمت زیر است:

Timestamp LabSZ sshd[SessionID]: Event

[6] Retention Rate
[7] Cohort
[8] Insight
[9] Trend
[10] Seasonality
[11] Heatmap

1. شما باید با استفاده از RegEx[12] متغیرهای مشخص شده رنگ آبی در فرمت بالا را استخراج کرده و از این متغییرها یک دیتا فریم بسازید.
2. ابتدا IP ها و عددهای موجود در Event را با کاراکترهای ثابت (کاراکتر عدد و IP متفاوت باشد) جایگزین کنید سپس
   a. مشخص کنید چند نوع Event مختلف وجود دارد و توزیع آنها به چه صورت است.
   b. ماتریس جابه‌جایی بین Event ها را بدست آورده و به صورت هیت مپ[13] رسم کنید. هر المان این ماتریس (a_{ij} سطر i و ستون j) نشان‌دهنده این است که در چند درصد مواقع از Event i به Event j رفتیم.

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز جمعه 19 آذر ماه خواهد بود.
- انجام این تمرین به صورت یک نفره میباشد.
- لطفا هرگونه فرض در حل سوالات را در گزارش خود ذکر کنید.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

HW5_[Lastname]_[StudentNumber].zip

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

[12] Regular Expression
[13] Heatmap