



دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده برق و کامپیوتر



درس

## تحلیل داده و مصورسازی

دکتر محمدمبین صادقی – دکتر محمدرضا ابوالقاسمی

طراح تمرین: علی ایزدی

زمان بارگزاری تمرین: شنبه ۱۹ آذر

تمرین شماره ۵

موضوع تمرین: تحلیل داده‌های تجاری

نیمسال اول سال تحصیلی ۱۴۰۱ - ۱۴۰۲

## بخش اول) process mining

هدف از این بخش بررسی فرآیندهای مربوط به درمان بیماران و تحلیل آن ها است.

دادگان مورد استفاده در فایل PatientTreatment.csv قرار داده شده شرح ستون های آن به صورت زیر است:

- patient: شماره بیمار
- action: فرآیندهای مختلف در درمان بیمار که از first consult شروع شده و تا last consult ادامه پیدا میکند.
- resource: فرآیند مربوطه توسط کدام دکتر، تیم و یا بخش انجام شده است.
- DateTime: زمان مربوط به انجام هر فرآیند.

به هر کدام از سوالات زیر پاسخ داده و تحلیل خود را به صورت کامل برای هر سوال بیان کنید.

### سوالات

۱- بررسی داده:

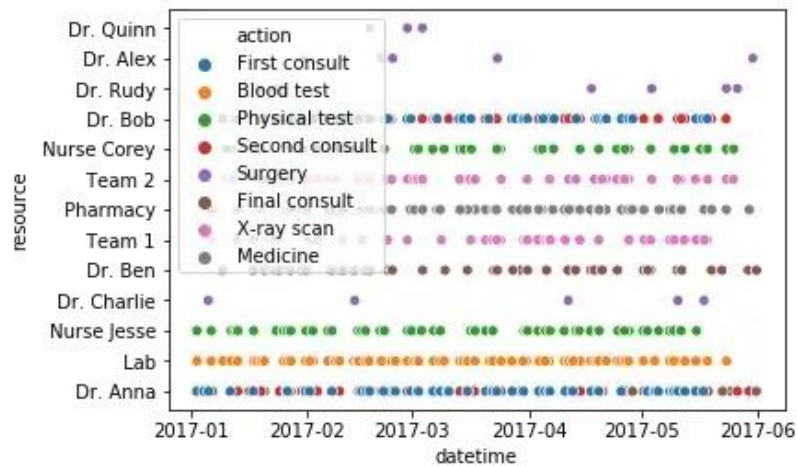
الف) تعداد بیماران یکتا را به دست آورید.

ب) تعداد هر یک از action های مختلف را به صورت صعودی در یک bar plot رسم کنید. کدام action ها بیشترین تعداد و کدام کمترین تعداد رخداد را دارند.

ج) تعداد هر یک از resource های مختلف را به صورت صعودی در یک bar plot رسم کنید. کدام resource بیشترین تعداد و کدام کمترین تعداد رخداد را دارند.

۲- در این سوال میخواهیم در یک نمودار بررسی کنیم که هر رخداد توسط چه resource ای و با چه action ای و در چه زمانی انجام شده است. پس از رسم نمودار آن را به دقت بررسی کنید و تحلیل های خود را از نتیجه توضیح دهید.

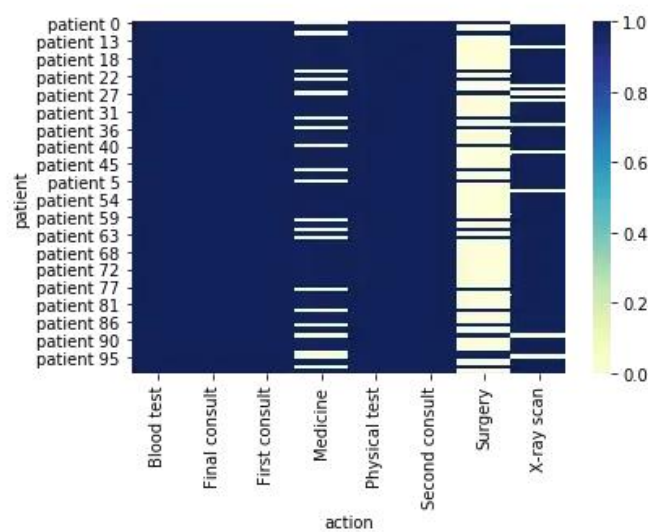
در واقع هدف رسیدن به نموداری مشابه نمودار شکل ۱ است. در محور افقی زمان، محور عمودی resource و هر action نیز با رنگ متفاوت برای هر رخداد مشخص شده است.



شکل ۱

۳- بررسی کنید کدام actionها برای همه بیماران انجام شده است و کدام یک برای تعدادی از بیماران انجام نشده است. نتیجه را در یک نمودار رسم کنید و تحلیل خود را برای actionهای مختلف ارائه دهید.

برای این که بتوان این اطلاعات را در یک نمودار رسم کرد میتوان از heatmap استفاده کرد. نتیجه شما باید نموداری مشابه نمودار شکل ۲ باشد که در آن محور افقی actionها و محور عمودی شماره بیماران است و اگر action انجام شده باشد با رنگ تیره و در غیر این صورت با رنگ روشن میشود.

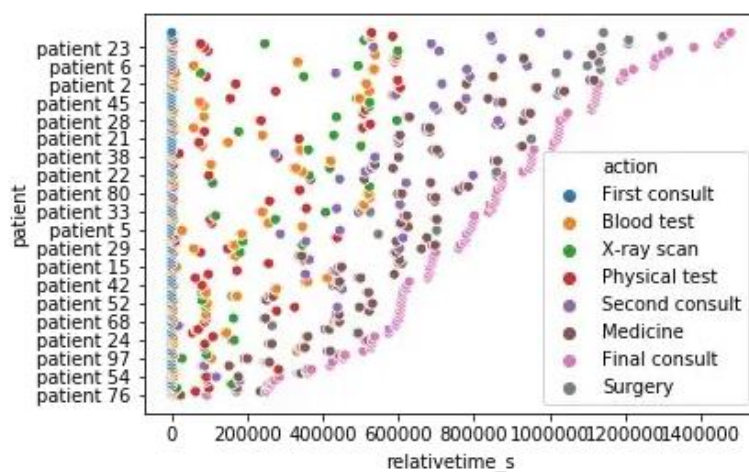


شکل ۲

۴- در این سوال میخواهیم در یک نمودار بررسی کنیم که هر **action** برای هر بیمار چند ثانیه طول کشیده است. پس از رسم نمودار تحلیل خود را از آن ارائه دهید.

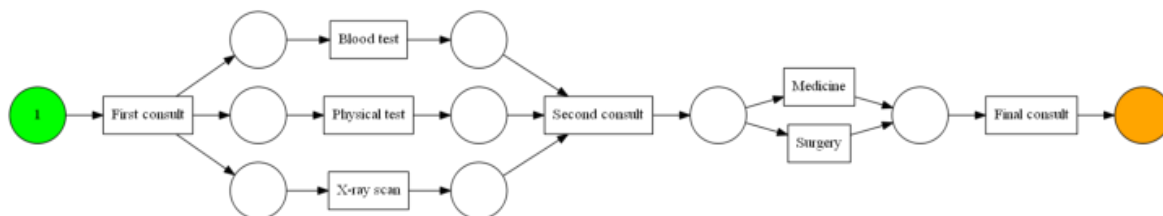
در واقع هدف رسیدن به نموداری مشابه نمودار شکل ۲ است. در محور افقی ثانیه‌های سپری شده، محور عمودی بیمار و هر **action** نیز با رنگ متفاوت برای هر رخداد مشخص شده است. هم چنین بیماران بر اساس کل زمان سپری شده در نشان مرتب شده اند.

راهنمایی: ابتدا باید برای هر بیمار اختلاف زمان رخداد هر **action** نسبت به اولین **action** را به دست آورید و آن را به ثانیه تبدیل کنید. سپس با استفاده از این ستون جدید نمودار شکل ۳ را رسم کنید.



شکل ۳

۵- هدف از این سوال بررسی ترتیب انجام **action** و تعداد بیماران برای هر کدام است. همان طور که انتظار می‌رود اولین و آخرین **action** باید **First consult** و **Final consult** باشند که برای هر ۱۰۰ بیمار انجام میشوند. اما **action** های دیگر ممکن است برای همه بیماران انجام نشوند. در واقع ترتیب انجام **action** ها مطابق با شکل زیر است:



گزارش کنید که در هر مرحله چه تعداد از ۱۰۰ بیمار **action** مربوطه را دریافت کرده اند. تحلیل خود را ارائه دهید. (تنها اعداد به دست آمده با استفاده از پیاده سازی خود را گزارش کنید و نیاز به رسم نمودار نیست.)



۶- برای بیمارانی که عمل جراحی انجام داده اند توزیع فاصله زمانی (برحسب دقیقه) بین مشورت دوم و انجام عمل جراحی را در یک نمودار رسم کنید و آن را تحلیل کنید.

## بخش دوم) business analytic

هدف از این بخش تحلیل کوهورت و نرخ بازگشت مشتریان برای فروش محصولات یک فروشگاه است.

دادگان مورد استفاده در فایل transaction.csv قرار داده شده شرح ستون های آن به صورت زیر است:

- transaction\_id
- product\_id
- customer\_id
- transaction\_date
- online\_order: آیا خرید به صورت آنلاین انجام شده است یا خیر؟
- product\_size: اندازه محصول در سه حالت small, medium, large

به هر کدام از سوالات زیر پاسخ داده و تحلیل خود را به صورت کامل برای هر سوال بیان کنید.

## سوالات

۱- هدف از این سوال به دست آوردن نرخ بازگشت است.

الف) نرخ بازگشت (retention rate) ماهیانه مشتری ها را به دست آورید و در یک نمودار رسم کنید.  
(در واقع باید محاسبه کنید چند درصد مشتری ها پس از گذشت یک ماه از خرید اولشان دوباره خرید انجام داده اند و این کار را برای دو، سه و ... ماه نیز انجام دهید و اعداد به دست آمده را در یک نمودار رسم کنید.)

ب) نرخ بازگشت ماهیانه مشتری ها را به صورت جداگانه برای خریدهای آنلاین و غیرآنلاین به دست آورید و در یک نمودار رسم کنید و آن ها را با هم مقایسه کنید.  
سطرهایی که مقادیر online\_order آن ها null است را از داده ها حذف کنید.



ج) نرخ بازگشت ماهیانه مشتری‌ها را به صورت جداگانه برای سائزهای مختلف محصولات به دست آورید و در یک نمودار رسم کنید و آن‌ها را با هم مقایسه کنید.  
سطرهایی که مقادیر `product_size` آن‌ها `null` است را از داده‌ها حذف کنید.

۲- جدول `cohort` ماهانه را به دست آورید و در یک نمودار رسم کنید. نمودار به دست آمده را تحلیل کنید.

## نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز شنبه ۱۹ آذر ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- خروجی مورد انتظار تمرین و هندزآن، در صورت انجام بخش امتیازی، ۱ فایل اجرا شده کتابچه جویپتر با فرمت `Notebook(.ipynb)` می باشد.
- هرگونه توضیحات و گزارش نویسی را به صورت `Markdown` داخل کتابچه جویپتر انجام دهید.
- 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:  
[aliizadi2030@gmail.com](mailto:aliizadi2030@gmail.com)