نام ونام خانوادگی : سارا رضایی
شماره دانشجویی: ۹۸۲۲۲۰۴۳
عنوان : گزارش و تحلیل دیتاست اول تمرین ۱ درس مبانی یادگیری ماشین

مقدمه

داده های مورد بررسی در این تمرین داده های فروش تلفن های همراه شرکت های مختلف است. در این داده ها رابطه بین ویژگی های مختلف تلفن همراه و کلاس قیمت تلفن بررسی میشود.

کلاینت قصد دارد مقدار تاثیر هر ویژگی را بررسی کند و موثرترین فیچر ها یافت شود. همچنین با استفاده از فیچر های موجود (مانند مقدار رم ، ظرفیت باطری ، حافظه و ....) مدل سازی برای پیش بینی کلاس قیمت انجام خواهد شد.

متد ها و مدل ها

از بلاک اول تا قبل از هدینگ تسک ۱ لود و بارگذاری داده ها از ککل انجام شده و shape و head دیتافریم برای مشاهده ی کلی داده ها چاپ شد. همچنین در بلاک آخر دیتافریم در یک دیتافریم دیگر کپی شده تا دیتافریم بدون تغییری برای اعمال پاک سازی با daskموجود باشد.

تسک اول preprocessing داده هاست که از **بلاک ۱۴** انجام شده است. در **بلاک ۱۴** مجموع داده های null در دیتاست محاسبه شد که تمامی ستون ها فاقد داده ی null بودند. همچنین زمان اجرای این سلول نیز محاسبه شده تا بعدا با کارایی در حالت استفاده از dask مقایسه شود.

در **بلاک۱۵و۱۶** با تکنیک IQR داده های پرت شناسایی شدند که با توجه به مقدار کم آن ها از حذف آنها چشم پوشی شد.

تسک دوم Exploratory Data Analysis که **از بلاک ۱۶** انجام شده است.در این قسمت اطلاعات آماری فیچر ها با استفاده از دستور describe نمایش داده شده و تایپ داده های هر ستون نیز بررسی شده.
همچنین ارتباط فیچر ها در نمودار های مختلف بررسی شده است که در بخش نتایج بیان میشود. در این بخش ارتباط کلاس قیمت با ویژگی های مختلف تلفن همراه از جمله توان باطری ، رم ، وزن گوشی و حافظه گوشی بررسی شده. همچنین داده های categorical درصد هر مقدار آنها نمایش داده شده است و در نهایت نمودار correlation فیچر ها رسم شده است.

تسک سوم آزمون های فرض که از **بلاک ۳۴** انجام شده است.

آزمون اول تاثیر گذاری داشتن تاچ اسکرین روی توان باطری بود که با استفاده از T-test 2samples بررسی و در نهایت پذیرفته شد.

آزمون دوم تاثیر نداشتن تعداد هسته های پردازنده ی تلفن همراه بر روی توان باطری بود که با ANOVA تست شد و در نهایت پذیرفته شد.

آزمون سوم تاثیرگذاری داشتن وایفای بر روی قیمت تلفن همراه بود که با Chi square تست شد و در نهایت پذیرفته شد.

آزمون چهارم عبارت است از اینکه میانگین رم تلفن ها برابر ۲۰۰۰ است که با T-test 1sample بررسی شد و در نهایت رد شد.

آزمون پنجم تاثیر نداشتن دو سیمکارته بودن بر روی مدت زمان مکالمه بود که با T-test 2samples بررسی شد و در نهایت پذیرفته شد.

تسک چهارم مدل سازی برای پیش بینی کلاس قیمت تلفن های همراه بر اساس فیچر های موجود میباشد که از **بلاک ۴۶** آماده سازی داده ها برای دادن به مدل ها انجام شده است.

در **بلاک ۴۶** فیچر target که price range است را از باقی فیچر ها جدا میکنیم.

و در **بلاک ۴۷** داده های test , train را با نسبت ۸۰ ۲۰ جدا شده است.

در **بلاک ۴۸** رابطه ی هر فیچر با قیمت با استفاده از تابع mutual info classif اندازه گیری شده است و در **بلاک ۴۹** این مقادیر برای بررسی بهتر سورت شده اند.

در **بلاک ۵۰** از این مقادیر برای انتخاب مجموعه پارامتر های مختلف استفاده شده است.

مجموعه اول شامل فیچر هایی ست که این عدد برایشان مقداری بیشتر از صفر است که شامل ۱۱ فیچر است.

مجموعه دوم شامل فیچر هایی ست که این عدد برایشان مقداری بیشتر از ۰,۰۰۵ است که شامل ۹ فیچر است.

مجموعه سوم شامل فیچر هایی ست که این عدد برایشان مقداری بیشتر از ۰,۰۲ است که شامل ۴ فیچر است.

اولین مدلی که آموزش داده شده است logistic regression است. از **بلاک ۵۴ تا ۵۹** این مدل با پارامتر های مختلف و مجموعه فیچر های مختلف آموزش داده شده است که بیشترین دقت این مدل۹۲,۲۵ درصد است که مربوط به مدل سوم با `max_iter=2000 , random_state= 50` و مجموعه فیچر سوم است. این مدل یک مدل با استراتژی OVA است.

دومین مدل که آموزش داده شده است naive bayes است. از **بلاک ۶۰ تا ۶۲** این مدل با پارامتر های مختلف و مجموعه فیچر های مختلف آموزش داده شده است که بیشترین دقت این مدل ۸۲,۲۵ درصد است که مربوط به مدل با پارامتر های دیفالت و مجموعه داده های اصلی است. این مدل یک مدل با استراتژی OVA است.

سومین مدل که آموزش داده شده است random forest است. از **بلاک ۶۳ تا ۶۷** این مدل با پارامتر های مختلف و مجموعه فیچر های مختلف آموزش داده شده است که بیشترین دقت این مدل ۹۲,۷۵درصد است که مربوط به مدل با `n_estimators = 100, random_state = 10` و مجموعه فیچر های دوم است. این مدل یک مدل با استراتژی OVA است.

تسک پنجم رسمconfusion matrix برای تمامی مدل ها بود که در زیر تمامی مدل ها رسم شد و مشاهده شد که مدل ها برای تمامی کلاس ها با یک دقت عمل نمیکنند و مثلا در این دیتاست معمولا کلاس ۳ کمترین دقت را داشت که احتمالا به دلیل نوع پراکندگی فیچر ها و مقدار فیچر های کلاس ۳ در داده های train بوده است.

تسک ششم بررسی توازن داده ها و ارایه راه حل هایی برای مشکل حل نامتوازن بودن داده ها بود که در فایل اصلی بررسی شد که داده ها متوازن هستند و ۳ راه حل Random Oversampling و Random Undersampling و SMOTE هم برای حل در صورت نامتوازن بودن ارایه شد.

تسک هفتم scale کردن داده ها بود که با دو متد `MinMaxScaler , StandardScaler` این کار صورت گرفت . برای این کار داده ها را روی داده های train را fit_transform میکنیم و داده هایtest را روی فیت کردن تاثیر نمیدهم و فقط با معیار قبلی transformمیکنیم و داده های scale شده به مدل های قبلی داده شد تا تفاوت در نتایج بررسی شود و مشاهده شد که دقت در تمامی مدل ها بالاتر رفت. از **بلاک ۶۸ تا ۷۵** مربوط به این قسمت میباشد.

تسک هشتم جداسازی داده های تست به نسبت ۲۰ ۸۰ بود که در تسک ۴ قبل از آموزش مدل ها انجام شد.

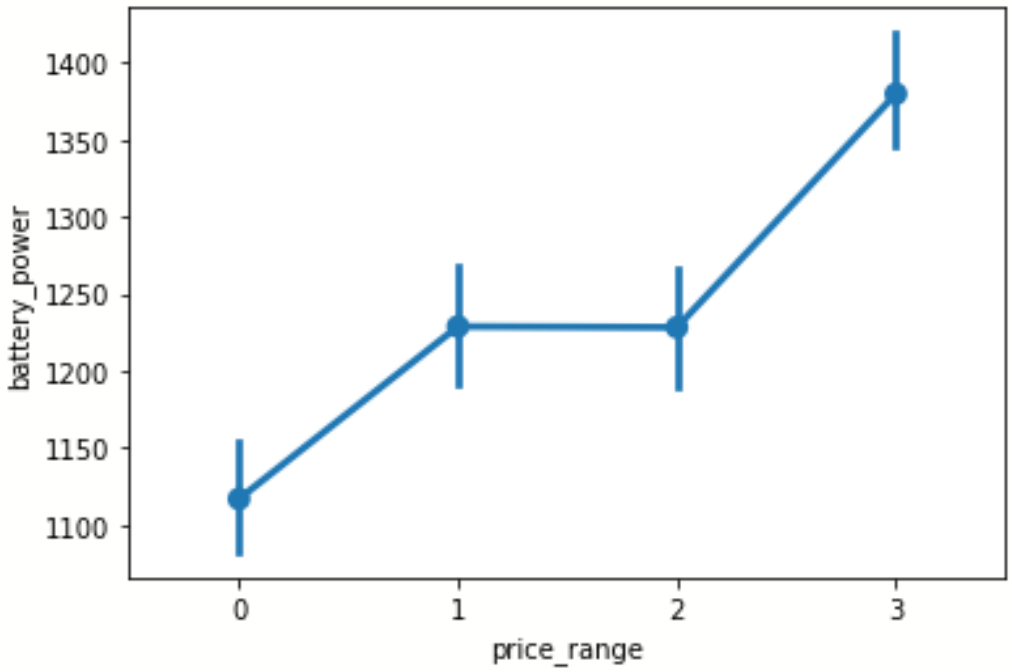
تسک نهم کاهش ابعاد فیچرها با متد PCA است که با ۵ pov مختلف این کاهش ابعاد انجام شد و پس از این کاهش ابعاد داده ها به مدل ها داده شدند و مشاهده شد که دقت مدل ها کاهش یافته است یا تغییری نکرده است. **بلاک ۷۶ تا ۸۵** مربوط به این قسمت است.

و تسک دهم ترکیب کلاس ها برای نامتوازن کردن داده ها است که این ترکیب در بلاک ۸۶ صورت گرفت. پس از آن از راه حل های random Undersampling و SMOTE به صورت ترکیبی برای متوازن سازی استفاده شد.

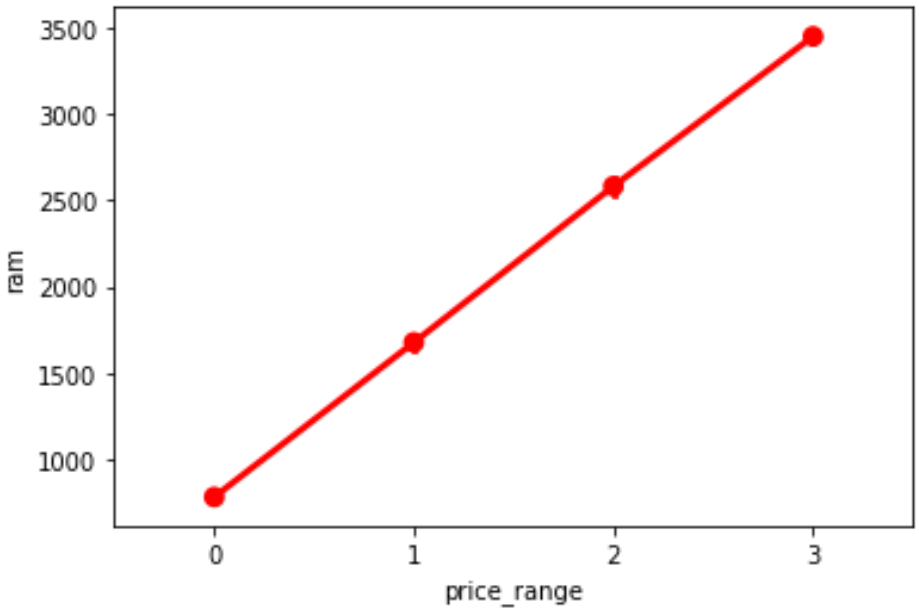
در قسمت **فعالیت های امتیازی** دو قسمت شمارش مقادیر null و چک کردن داده های پرت با استفاده از dask انجام شد و زمان اجرای آنها نیز بررسی شد که این زمان نسبت به حالت پردازش معمولی افزایش یافته بود که احتمالا به خاطر کوچک بودن دیتاست است.

نتایج نمودار ها

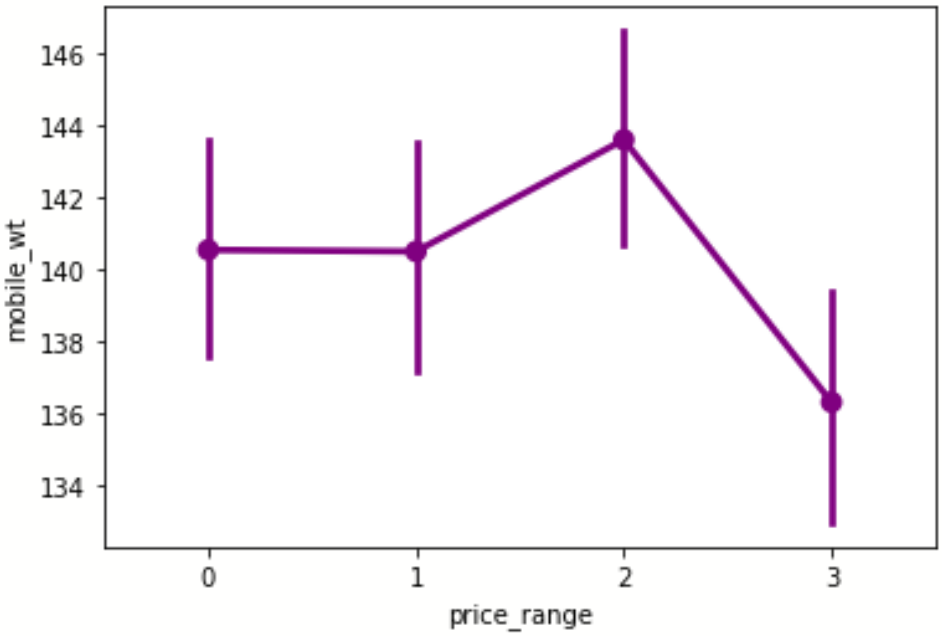
نمودار زیر رابطه بین قدرت باطری و کلاس قیمت را نشان میدهد که مشاهده میشود رابطه مستقیم وجود دارد و با افزایش قدرت باطری کلاس قیمت نیز افزایش میابد.



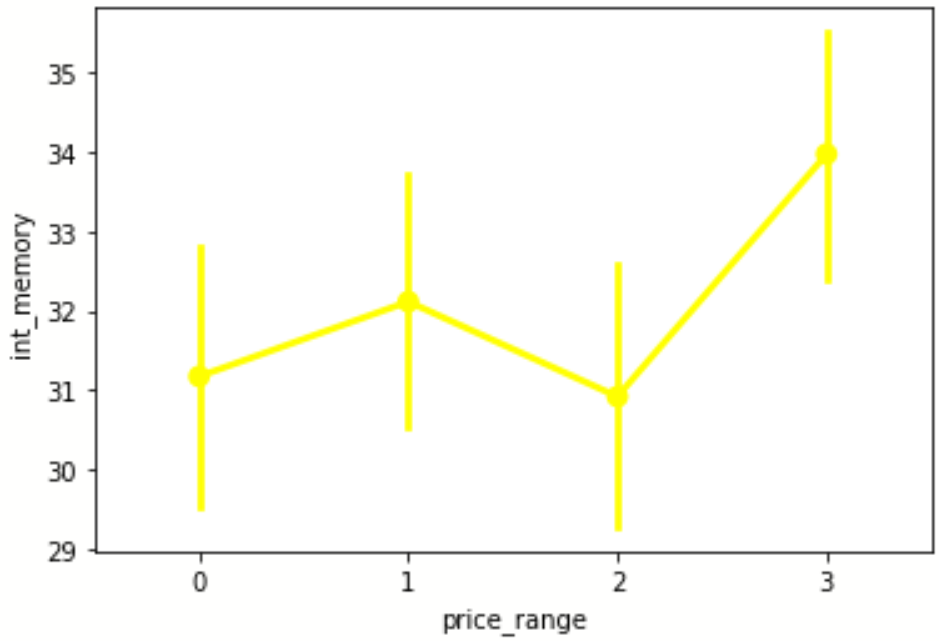
این نمودار هم نشان میدهد که افزایش رم در تلفن همراه رابطه مستقیم با افزایش کلاس قیمت تلفن همراه دارد.



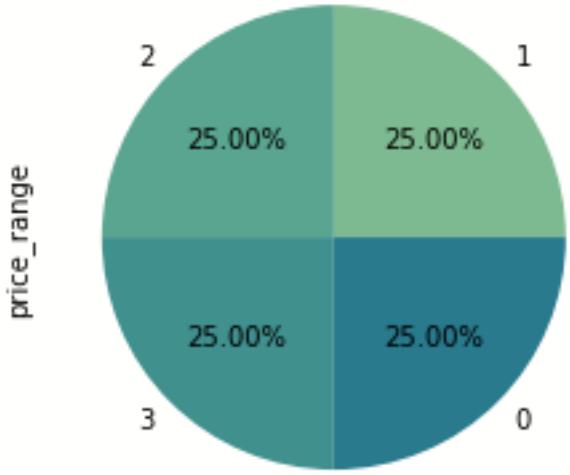
نمودار زیر رابطه ی وزن تلفن همراه و کلاس قیمت را نشان میدهد که با توجه به آن مشاهده میکنیم که از یک جایی به بعد افزایش وزن تلفن متوقف شده و تلفن های گران ترین کلاس مربوط به سبک ترین تلفن های همراه است.



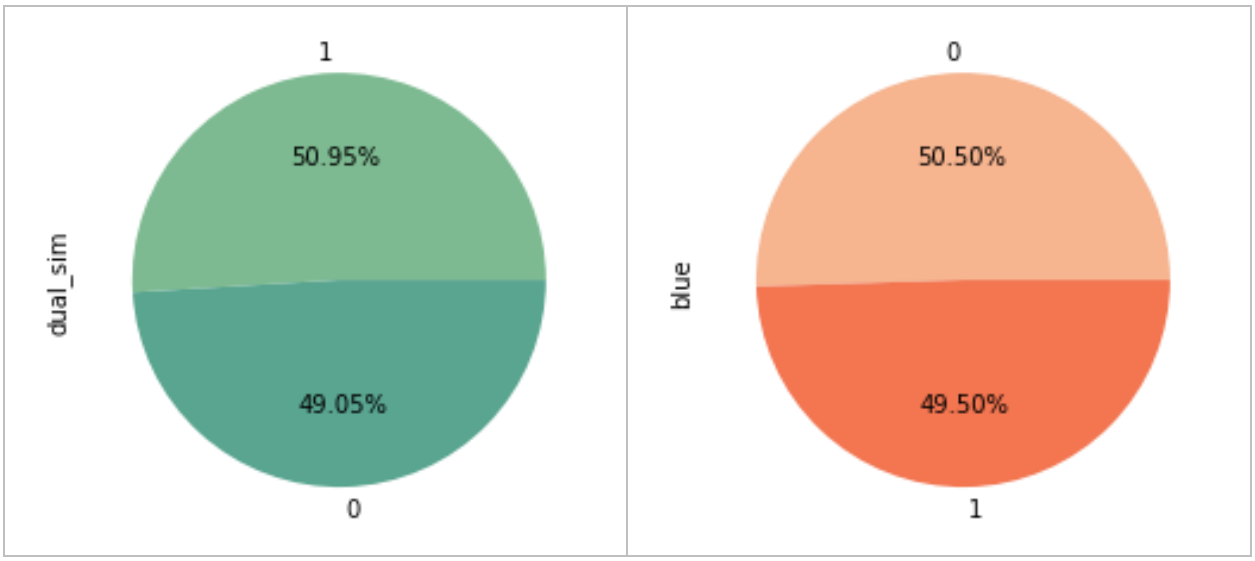
نمودار زیر رابطه ی حافظه ی تلفن همراه و کلاس قیمت را نشان میدهد که با توجه به پراکندگی تغییرات میتوان نتیجه گرفت که حافظه تاثیر زیاد و مستقیمی روی کلاس قیمت ندارد.

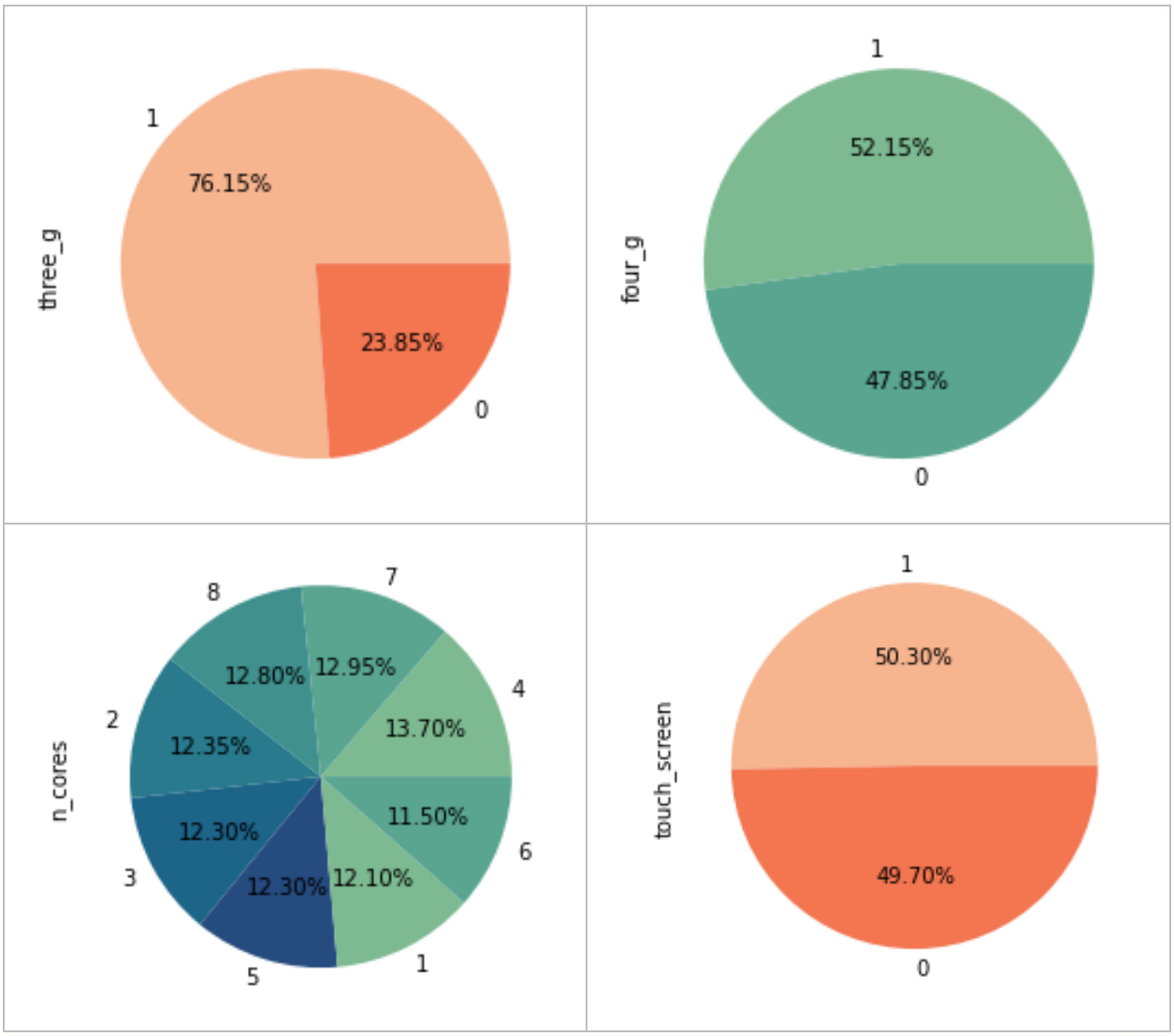


نمودار زیر **متوازن بودن داده ها** را نشان میدهد که مشاهده میکنیم از هر کلاس به میزان مساوری داده موجود است.

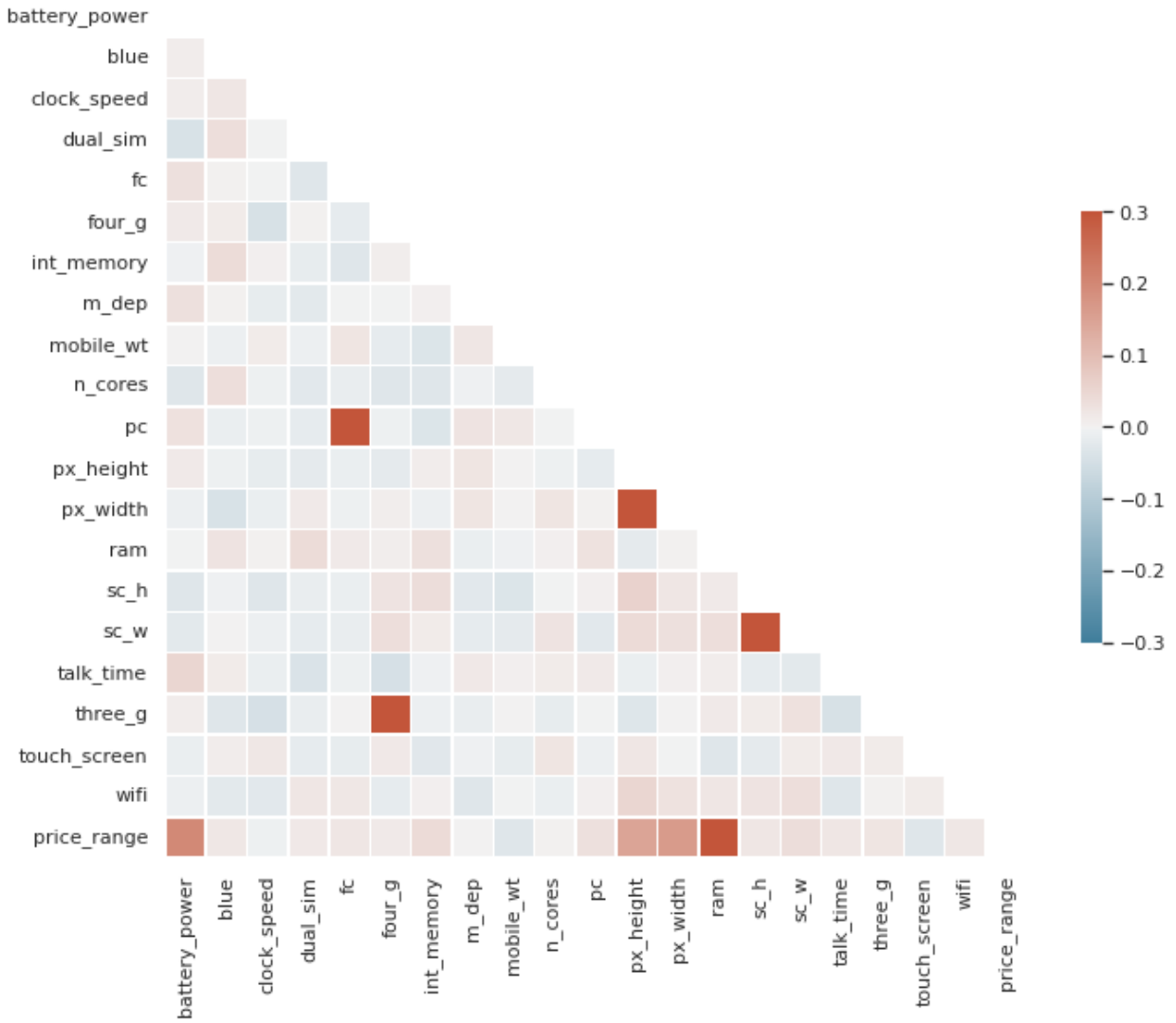


نمودار هایی که در ادامه نمایش داده شده توزیع مقادیر مختلف داده ها ی categorical را نمایش میدهد




three_g
1
76.15%
23.85%
0
four_g
1
52.15%
47.85%
0
n_cores
8
7
12.80%
12.95%
4
2
13.70%
12.35%
11.50%
6
12.30%
3
12.30%
12.10%
5
1
touch_screen
1
50.30%
49.70%
0

نمودار آخر نمودار correlation است که ارتباط فیچر ها با یکدیگر و با price range را نشان میدهد.



طبق این نمودار چهار فیچر ram, px_width, px_height, battery power بیشترین ارتباط و تاثیر را بر قیمت دارند.