تسک‌های پروژه:

۱) در گام اول، با پاسخ‌گویی به این سؤالات سعی کنید درک مطلوبی از دیتاست بدست آورید.

الف) چند متغیر از نوع اعداد صحیح و چند متغیر از نوع مقادیر صفر و یک وجود دارد؟

در جدول دیتاست ارائه شده، پنج ستون وجود دارد.

- دو ستون (userid و sum_gamerounds) دارای مقادیر عددی هستند. ستون userid دارای مقادیر عددی است اما مقادیرش در تحلیل و محاسبات پروژه لحاظ نمی‌شود و فقط از نظر داشتن یا نداشتن داپلیکیت و همچنین تعداد مقادیر بررسی می‌شود.
- دو ستون (retention_1 و retention_7) شامل مقادیر صفر و یکی هستند.
- ستون (version) دارای مقادیر string است اما فقط شامل دو دسته‌ی gate_30 و gate_40 است.

ب) آیا دیتایی از دست رفته است؟ (Missing Values) در صورت پاسخ مثبت، آن‌ها را با توجه به روش‌های جایگزینی، با مقداری مناسب پر کنید.

برای تمیز کردن دیتاست طبق مراحل زیر پیش می‌رویم:

- جای هیچ‌کدام از دیتاهای دیتاست خالی نیست. بنابراین نیازی به جایگزینی وجود ندارد.
- ستون‌های retention_1 و retention_7 بررسی شدند و مشخص شد فقط دارای مقادیر True و False هستند.
- ستون version هم بررسی شد و مشخص شد فقط دارای مقادیر gate_30 و gate_40 است.
- طبق منطقی که برای دیتاست توضیح داده شده است، وجود اعداد کوچک‌تر از صفر در ستون sum_gamerounds معتبر نیست. این مورد هم بررسی شد و مشخص شد تمام اعداد مثبت هستند.

- کاربرهایی که یکی موارد retention_1 یا retention_7 آن‌ها True باشد، مجموع راندهای بازی‌شان باید بیشتر از صفر باشد چون حداقل یک راند بازی کرده‌اند. بنابراین دیتای کاربرانی که این شرایط را دارند و sum_gamerounds آن‌ها صفر است Invalid است.

- همین‌طور کاربرهایی که retention_1 و retention_7 آن‌ها True است، مجموع راندهای بازی‌شان باید بیشتر از ۱ باشد. بنابراین دیتای کاربرانی که این شرایط را دارند و sum_gamerounds آن‌ها صفر و یک است Invalid هستند.

با اضافه کردن این دو شرط، این کاربران را از دیتاست حذف می‌کنیم.

اطلاعات مورد نیاز این تسک در جدول زیر آورده شده است:

| | Number of Values | Number of null values | Type of data |
|---|---|---|---|
| userid | 90,070 | 0 | int64 |
| version | 90,070 | 0 | object |

| | | | |
|---|---|---|---|
| sum_gamerounds | 90,070 | 0 | int64 |
| retention_1 | 90,070 | 0 | bool |
| retention_7 | 90,070 | 0 | bool |

ج) تعداد کاربران را پیدا کنید. و آماره‌های میانگین، واریانس، کمترین مقدار، چهارک اول و سوم را برای تعداد پارامتر sum_gamerounds پیدا کنید.

**تعداد idهای یونیک ۹۰۰۷۰ است و با توجه به تعداد سطرهای دیتاست متوجه می‌شیم تمامی useridها یونیک هستند.**

**آماره‌های مورد نیاز این تسک در جدول زیر آورده شده است:**

| | df |
|---|---|
| Count | 90,070 |
| Mean | 51.940 |
| Variance | 38,091.551 |
| Min | 0 |
| 25% | 5 |
| 75% | 51 |

**۲)** در این گام میخواهیم دیتاست مورد نیاز برای A/B Test را بدست بیاوریم. برای این کار به سوالات زیر پاسخ دهید:

الف) با توجه به مقدار version سطرهای مربوط به کاربران را جدا کنید و برای هر دو دسته، سوال ج از گام اول پروژه را مجددا پاسخ دهید.

**تعداد کاربران برای gate_30 عدد ۴۴۶۴۰ و برای gate_40 عدد ۴۵۴۳۰ است.**

**بقیه‌ی آماره‌های این تسک در جدول زیر آمده است:**

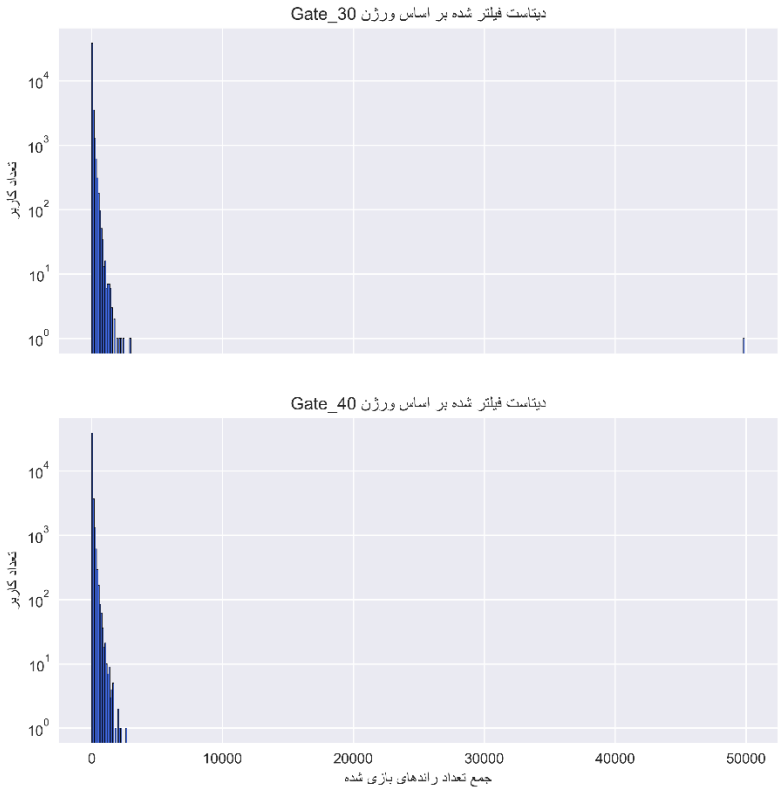
| | Gate_30 | Gate_40 |
|---|---|---|
| Count | 44,640 | 45,430 |
| Mean | 52.526 | 51.365 |
| Variance | 65,988.209 | 10,680.180 |
| Min | 0 | 0 |
| 25% | 5 | 5 |
| 75% | 50 | 52 |

ب) در این مرحله می خواهیم مقادیر نویز (Outliers)را از دیتاست حذف کنیم. دقت کنید همواره در هر دیتاستی ممکن است این مقادیر وجود داشته باشد، پس همواره مهم است که این کار را در شروع کار با هر دیتاستی انجام دهید. نکته مهم دیگر این است که این مقادیر معمولا یک درصد از کل دیتاست خواهد بود.

در ابتدا، توزیع ویژگی sum_gamerounds را برای هر دو دسته‌ی مرحله قبل انجام دهید. احتمالا در هر شکل، مقادیر خیلی زیادی در بازه‌ی اول قرار خواهد گرفت. ( این ویژگی دنیای طبیعی است )

ابتدا پلات‌های مربوط به این تسک را رسم می‌کنیم. با توجه به اختلاف زیاد تعداد راندهای بازی در میله‌های ابتدایی و انتهایی، برای آنکه در پلات تمامی میله‌ها مشخص باشند، از نمایش لگاریتمی استفاده شده است.

پلات‌ها به صورت زیر است:




حالا، برای ویژگی sum_gamerounds در کل دیتاست مشخص کنید که ۹۹ درصد اعداد از چه عددی کمتر است؟

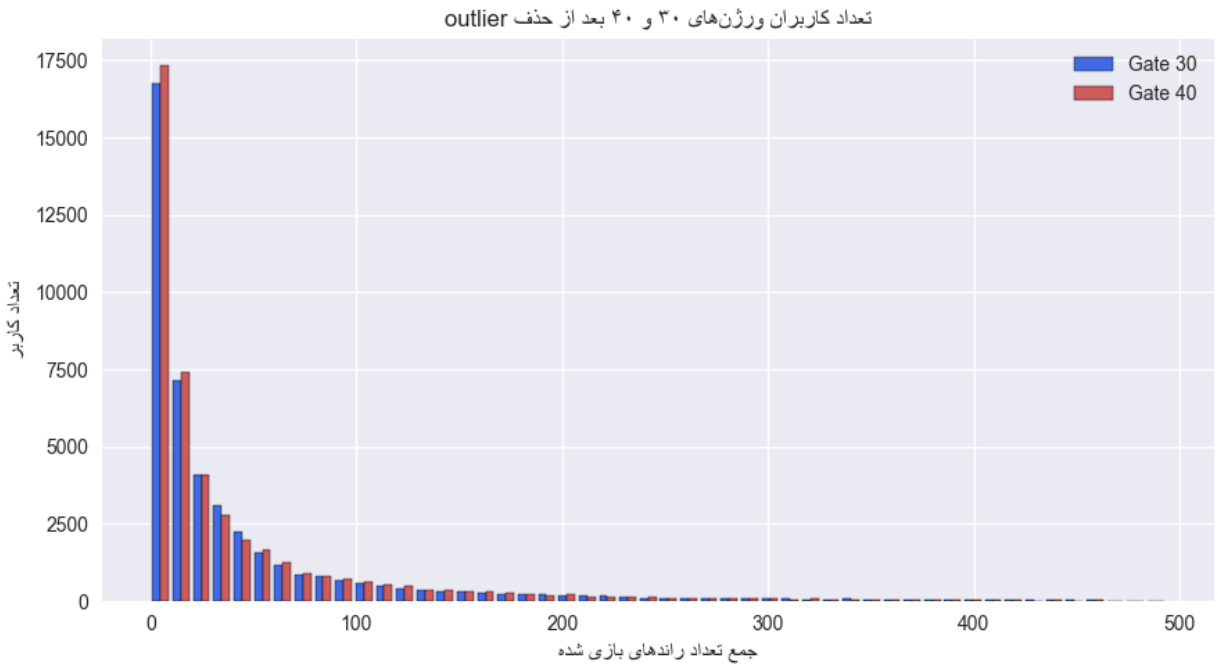
ابتدا مقدار outlier برای ستون sum_gamerounds را محاسبه می‌کنیم. عدد خروجی کد ۴۹۳ است.

مقدار بالاتر از این مقدار را که شامل ۱٪ از کل دیتا می‌شود از دیتاست حذف می‌کنیم.

احتمالا مقادیر بالای این عدد outlier خواهند بود. یک بار دیگر نمودارها را بکشید. ( دقت کنید که بازهم اکثریت مقادیر در میله اول خواهد بود )

دیتاستی را که مقادیر outlier از آن حذف شده‌اند مجددا بر اساس ورژن‌های gate_30 و gate_40 فیلتر کرده و پلات‌ها را دوباره رسم می‌کنیم.

پلات‌های ورژن‌های gate_30 و gate_40 بدون وجود outlier به صورت زیر است:

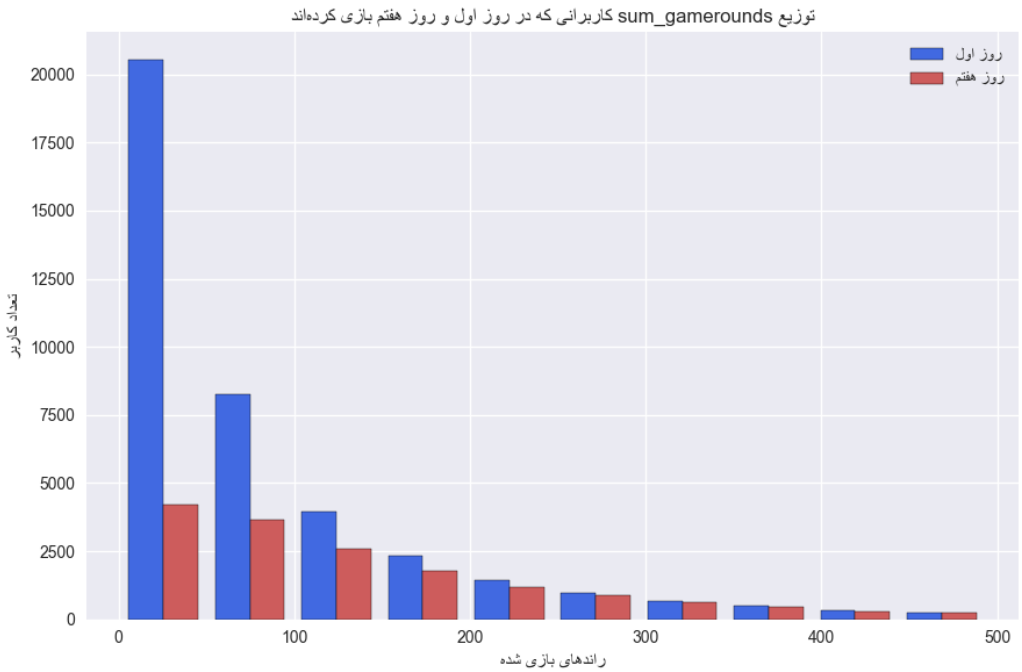



۳) در این گام میخواهیم بیشتر وارد جزییات شویم و به مساله retention بپردازیم:

الف) در ابتدا آیا توضیحی برای اینکه چرا اکثریت دیتاها در میله اول هستند دارید؟ برای پاسخ به این سوال این بار نموداری بکشید که محور x راندهای بازی و محور y تعداد کاربران در هر راند باشد.

رفتار طبیعی در انجام بازی‌ها و استفاده از اپلیکیشن‌ها همین‌طور است. تعداد زیادی از کاربران زمان کمی از اپلیکیشن استفاده می‌کنند یا تعداد دفعات بازی کردن آن‌ها کم است. کاربرانی که جذب بازی یا اپلیکیشن می‌شوند و به استفاده کردن از آن ادامه می‌دهند، دفعات بازی کردن یا زمان استفاده‌شان بیشتر می‌شود. بنابراین طبیعی است که نمودار ما در این دیتاست دارای میله‌های اول بلند باشد و به تدریج طول میله‌هایش کم‌تر شود.

برای مثال پلات‌های زیر بر اساس مجموع راندهای بازی کاربرانی که در روز اول  یا روز هفتم بازی کرده‌اند رسم شده است:




ب) حالا به مساله retention می‌پردازیم. دو ویژگی retention_1 و retention_7 را در نظر بگیرید. چه تعداد از کاربران یک روز بعد از نصب آمده و بازی کرده‌اند؟ چه تعداد از کاربران بعد از هفت روز از نصب، بازی را انجام داده‌اند؟ در هر دو سوال، پاسخ چه درصدی از کل کاربران بوده است؟ پاسخ این سوالات را برای تمامی دیتاست، و گروه‌های متفاوت کاربران بر اساس version بدست بیاورید.

پاسخ تمامی سوال‌های مطرح شده در جدول زیر آمده است:

| | Number of Columns | Number of True retention_1 | Percentage retention_1 | Number of True retention_7 | Percentage retention_7 |
|---|---|---|---|---|---|
| Version gate_30 | 44,194 | 19,556 | 44.250 | 8,058 | 18.233 |
| Version gate_40 | 44,978 | 19,627 | 43.636 | 7,828 | 17.404 |
| Total | 89,172 | 39,183 | 43.940 | 15,886 | 17.815 |

ج) حالا آماره‌های میانه، میانگین، واریانس و ماکزیمم را برای گروه‌های مختلف - بنا بر ویژگی‌های version و retention - بدست بیاورید. ( یکبار version با retention_1 و بار دیگر version با retention_7 )

آماره‌های مربوط به این تسک در جدول زیر آورده شده است:

| Column1 | Gate_30_ret_1 | Gate_40_ret_1 | Gate_30_ret_7 | Gate_40_ret_7 |
|---|---|---|---|---|
| Median | 46 | 48 | 98 | 104 |
| Mean | 80.908 | 81.139 | 131.168 | 134.184 |
| Variance | 8,102.107 | 7,996.587 | 12,092.789 | 11,982.194 |
| Max | 493 | 493 | 493 | 493 |

## ۴) حال به انجام A/B Test می‌پردازیم:

فرضیه‌ای که در این مساله می‌خواهیم به آن بپردازیم این است آیا انتقال اولین مانع از مرحله ۳۰ به مرحله ۴۰، تاثیری بر rentention کاربران داشته است یا خیر؟ با توجه به مطالبی که در طول درس خوانده‌اید، و شهودی که تا اینجای کار از مساله و دیتاست بدست آورده‌اید، به این سوال پاسخ دهید :)

برای انجام این تست، ابتدا فرضیه‌ی صفر خود را مشخص می‌کنیم.

$h_0$ = انتقال اولین مانع از مرحله‌ی ۳۰ به مرحله‌ی ۴۰ تاثیری بر Retention در روز اول نداشته است.

ابتدا p-value را معادل ۹۵٪ قرار می‌دهیم و با این فرض z critical را از روی جدول به دست می‌آوریم. عدد به دست آمده ۱.۹۶ است.

نوع توزیع retentionها را برنولی در نظر می‌گیریم.

سپس مقدار p را از فرمول زیر به دست می‌آوریم:

$$p = \frac{nt_1 + nt_2}{n_1 + n_2} = \frac{19556 + 19627}{44194 + 44978} = 0.4394$$

بعد این مقدار را در فرمول محاسبه z-score قرار می‌دهیم:

با توجه به فرض اولیه $\mu_1$ را برابر با $\mu_2$ در نظر می‌گیریم.

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (\mu_1 - \mu_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(44.250 - 43.636) - 0}{\sqrt{0.4394(1-0.4394)\left(\frac{1}{44194} + \frac{1}{44978}\right)}} = 184.705$$

$\widehat{p_1}$ = درصد کاربرهای gate_30 که ret_1 آن‌ها True است.

$\widehat{p_2}$ = درصد کاربرهای gate_40 که ret_1 آن‌ها True است.

$nt_1$ = تعداد کاربران gate_30 که ret_1 آن‌ها True است.

$nt_2$ = تعداد کاربران gate_40 که ret_1 آن‌ها True است.

$n_1$ = تعداد کل کاربران gate_30.

$n_2$ = تعداد کل کاربران gate_40.

با توجه به اختلاف z score به دست آمده و z critical، می‌توانیم نتیجه‌گیری کنیم که انتقال اولین مانع روی retention روز اول تاثیر داشته است.

از همین روش برای A/B Test روز هفتم استفاده می‌کنیم:

$h_0$ = انتقال اولین مانع از مرحله‌ی ۳۰ به مرحله‌ی ۴۰ تاثیری بر Retention در روز هفتم نداشته است.

$$p = \frac{nt_1 + nt_2}{n_1 + n_2} = \frac{8058 + 7828}{44194 + 44978} = 0.1781$$

با توجه به فرض اولیه $\mu_1$ را برابر با $\mu_2$ در نظر می‌گیریم.

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (\mu_1 - \mu_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(18.233 - 17.404) - 0}{\sqrt{0.1781(1-0.1781)\left(\frac{1}{44194} + \frac{1}{44978}\right)}} = 323.504$$

در روز هفتم هم به این نتیجه می‌رسیم که انتقال مانع از مرحله‌ی ۳۰ به ۴۰ روی retention تاثیر داشته است.