مراحل انجام داده کاوی

1. بارگیری کتابخانه ها و خواندن داده ها

1.1 بارگیری کتابخانه ها

1.2 خواندن داده ها

1.3 مقادیر از دست رفته

1.4 ویژگی های بی فایده

2. تجزیه و تحلیل اکتشافی داده ها (EDA)

2.1 توصیف دیتاست

2.2. توزیع هدف (tag)

2.3 توزیع ویژگی ها (click)

2.4 توزیع زمان

3. تجزیه و تحلیل با استفاده از t-test

1- بارگیری کتابخانه ها و خواندن داده ها

1- کتابخانه های مورد نیاز:

Pandas – Numpy –Matplotlib –Seaborn – Scipy

2- خواندن داده ها(نمایش 5 سطر اول)

مجموعه داده Click posts

	index	post_id	click_time	click_post_count
0	0	gXvC9ucx	5/7/2020 0:00	31
1	1	gXluVIHC	5/7/2020 0:00	19
2	2	gXxugqam	5/7/2020 0:00	13
3	3	gXy6Jn2Y	5/7/2020 0:00	14
4	4	gXo2A5D9	5/7/2020 0:00	38

مجموعه داده Reorders

	index	post_id	tag	reorder_time	post_publish_time	post_retire_time
0	0	gX1-5MmB	reorder	5/9/2020 10:00	5/8/2020 22:44	5/11/2020 14:12
1	1	gX1-WPeh	urgent-reorder	5/9/2020 20:39	5/8/2020 19:14	5/13/2020 19:07
2	2	gX1eSoEn	reorder	5/9/2020 12:48	5/8/2020 14:14	5/9/2020 12:52
3	3	gXkaGKA_	reorder	5/3/2020 13:24	4/25/2020 21:07	5/10/2020 12:35
4	4	gXkaGKA_	reorder	5/2/2020 11:08	4/25/2020 21:07	5/10/2020 12:35

3- ترکیب دو مجموعه داده بر اساس پست های خاص

	index_x	post_id	click_time	click_post_count	index_y	tag	reorder_time	post_publish_time	post_retire_time
0	0	gXvC9ucx	5/7/2020 0:00	31	3101	urgent-reorder	5/5/2020 12:30	5/4/2020 12:53	5/7/2020 0:56
64	1	gXluVlHC	5/7/2020 0:00	19	7293	urgent-reorder	5/5/2020 2:37	4/26/2020 14:35	5/13/2020 0:10
294	2	gXxugqam	5/7/2020 0:00	13	5399	reorder	5/8/2020 12:41	5/5/2020 11:18	5/11/2020 12:52
411	3	gXy6Jn2Y	5/7/2020 0:00	14	1133	reorder	5/7/2020 13:43	5/6/2020 16:06	5/8/2020 23:17
468	4	gXo2A5D9	5/7/2020 0:00	38	4150	reorder	5/1/2020 5:09	4/28/2020 15:37	5/9/2020 17:23

4- حذف ویژگی های (ستون ها) بی فایده

```
final_df = final_df.drop(columns=['index_x', 'index_y'])
```

```
[ ] final_df = final_df.reset_index()
```

```
[ ] final_df.tail(5)
```

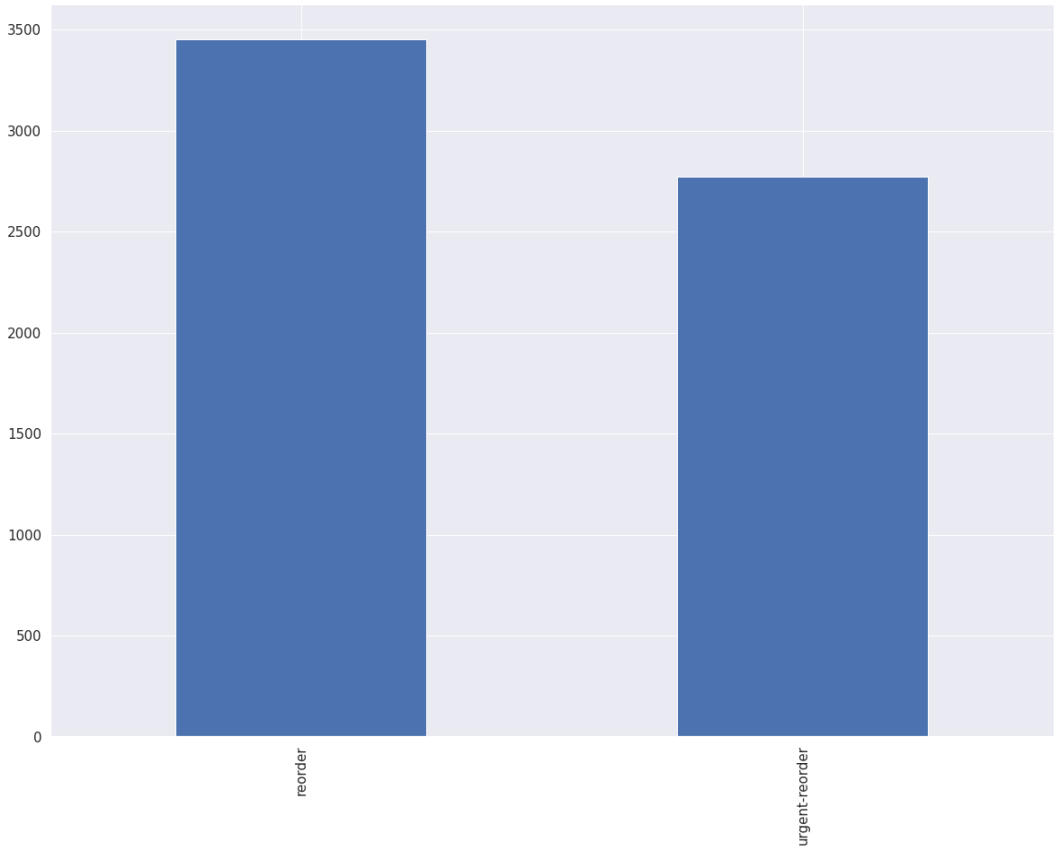
	post_id	click_time	click_post_count	tag	reorder_time	post_publish_time	post_retire_time
7320	gXzC9prE	5/7/2020 18:00	838	reorder	5/7/2020 16:05	5/7/2020 13:48	5/7/2020 16:28
7321	gXlypm15	5/3/2020 13:00	14	reorder	5/7/2020 10:20	4/26/2020 19:10	NaN
7322	gXkWywKl	5/1/2020 16:00	985	reorder	5/1/2020 8:24	4/25/2020 16:38	5/1/2020 11:32
7323	gXsmfEBU	5/2/2020 13:00	325	reorder	5/2/2020 11:53	5/2/2020 9:16	5/2/2020 14:28
7324	gXlWeLPP	5/3/2020 17:00	232	urgent-reorder	5/3/2020 14:30	4/27/2020 0:26	5/3/2020 22:25

2- تجزیه و تحلیل اکتشافی داده ها (EDA)

1- توزیع هدف (tags)

نمودار میله‌ای یا نواری (Bar Plot) نموداری است که داده‌های طبقه بندی شده را با میله‌های مستطیل شکل با ارتفاع یا طول متناسب با مقادیر ارائه شده نشان می‌دهد.

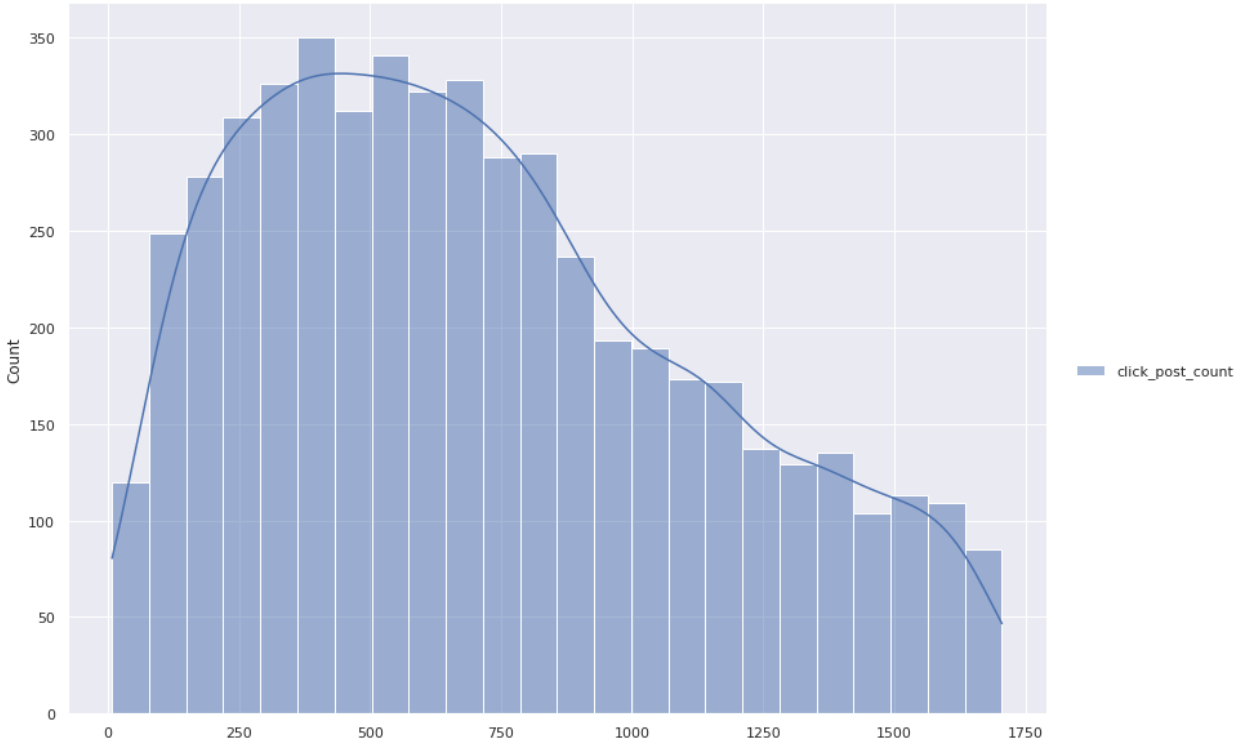
با توجه به نمودار پست های reorder با تعداد نزدیک 3500 مقادیر بیشتری از پست ها را نسبت به urgent-reorder به خود اختصاص داده اند.



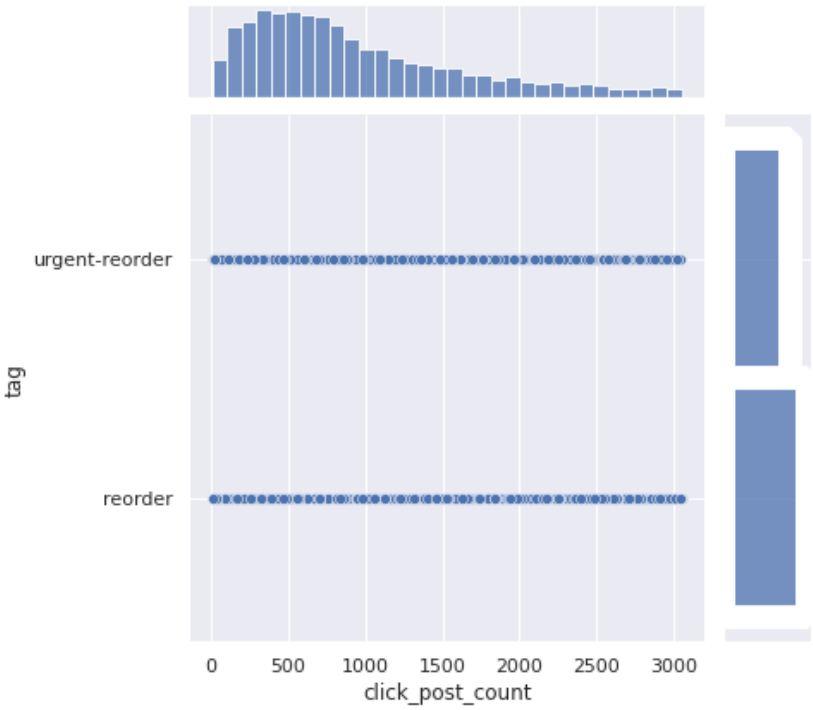
2- توزیع کلیک ها

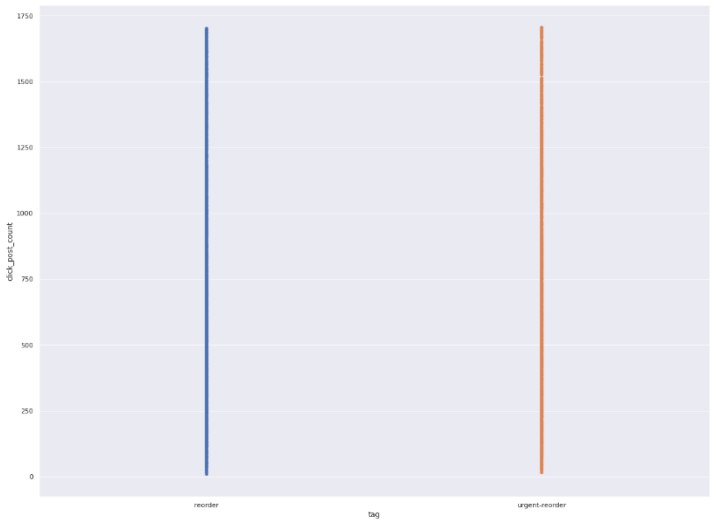
نمودار هیستوگرام یک نمایش دقیق از توزیع داده‌های عددی است. این نمودار تخمینی از توزیع احتمال متغیر پیوسته است .تفاوت نمودار هیستوگرام با نمودار میله‌ای در این است که یک نمودار میله‌ای رابطه‌ی دو متغیر را با هم نشان می‌دهد. اما هیستوگرام تنها به یک متغیر مربوط می‌شودخطی که مشاهده می کنید برآورد چگالی کرنل را نشان می دهد. این خط را با ارسال `False`به عنوان پارامتر`kde` ، می توان حذف کرد:

نمودار زیر محدوده توزیع کلیک ها برای هر پست خاص نمایش می دهد
Outlier ها برای زسم این توزیع حذف شده اند



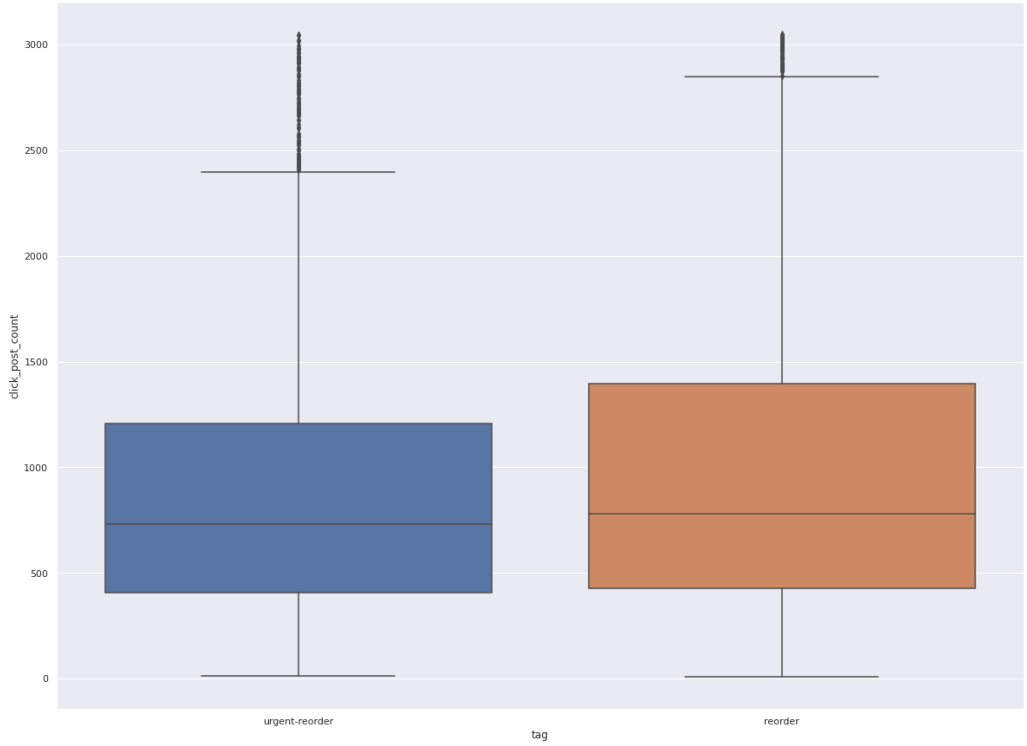
3- روابط بین تگ و تعداد کلیک در قالب نمودار های مختلف





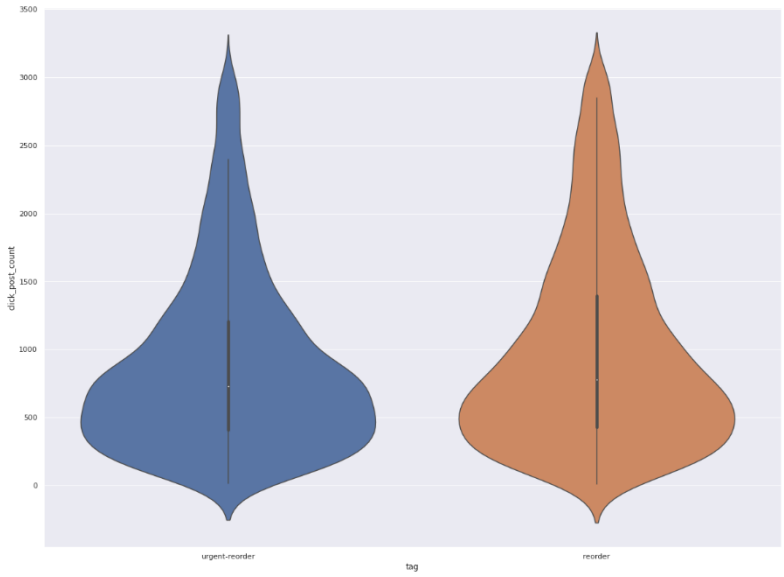
نمودار جعبه‌ای یک روش استاندارد برای نمایش توزیع داده‌ها است که براساس شاخص‌های آماری «کوچکترین مقدار(Minimum) «، »چارک اول(First Quartile -Q1) «، »میانه(Median) «، »چارک سوم (Third Quartile- Q3) «و «بزرگترین مقدار « (Maximum)ساخته شده است. همچنین این نمودار می‌تواند در مورد وجود داده‌های دورافتاده (Outlier) یا پرت، اطلاعاتی به شما بدهد و مقدار آن‌ها را تعیین کند. همچنین نشان دادن تقارن در داده‌ها از کارهایی این نمودار است.

همانطور که دیده می‌شود بین تعداد کلیک ها با تگ های reorder و urgantreorder ، تفاوت وجود دارد.



نمودار ویولنی مشابه نمودار جعبه‌ای، خلاصه‌ای از آمار داده شامل موارد زیر را نشان می‌دهد :

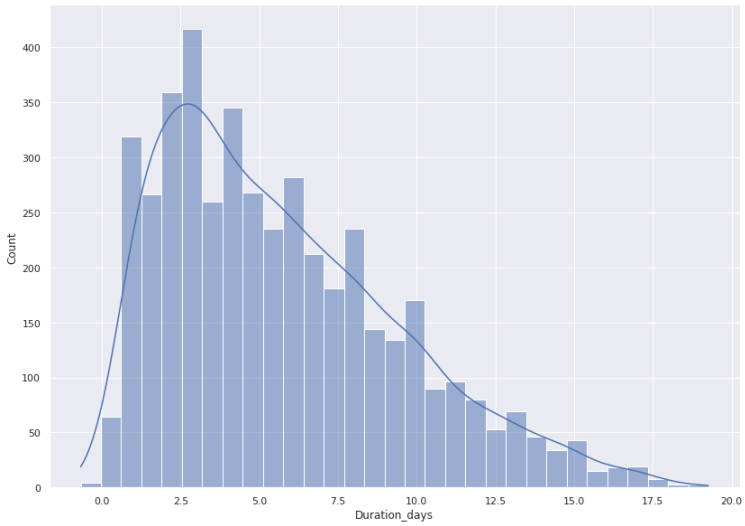
- خط عمودی کوچک وسط جعبه رسم شده داخل نمودار، نشان‌دهنده میانه است.
- خط ضخیم‌تر عمودی در مرکز شکل، نشان‌دهنده بازه میان چارکی است.
- خط باریک‌تر عمودی در مرکز، نشان دهنده ۹۵٪ فاصله اطمینان است.



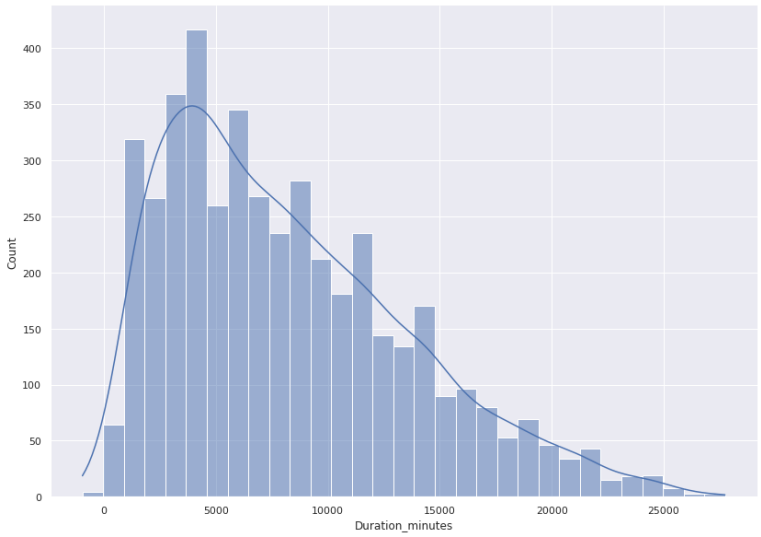
4- توزیع زمان

بررسی کل مدت زمان هر پست در مقایسه با تعداد پست ها

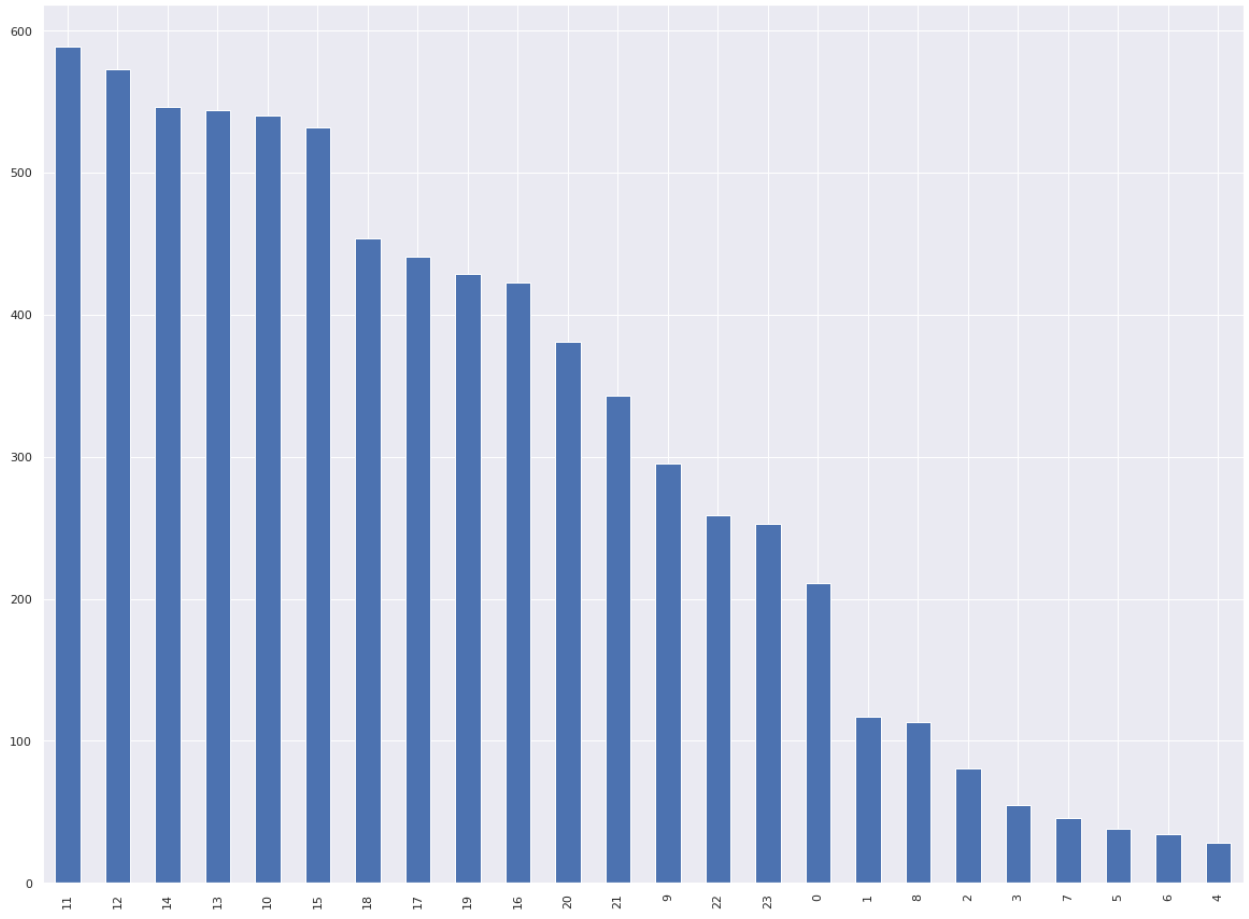
مدت زمان (روز)



مدت زمان (دقیقه)

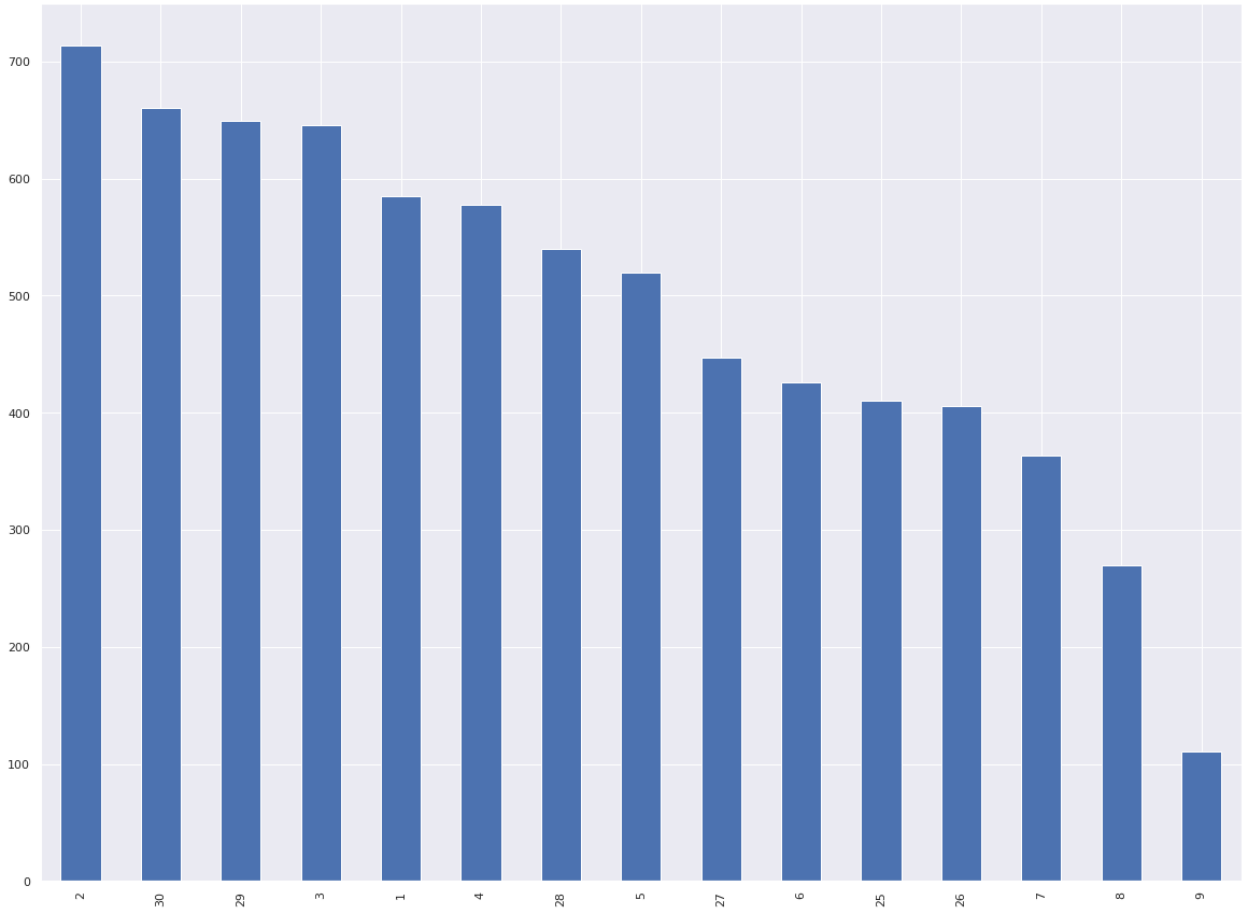


بررسی ساعات مختلف شبانه روز با بیش ترین تعداد پست



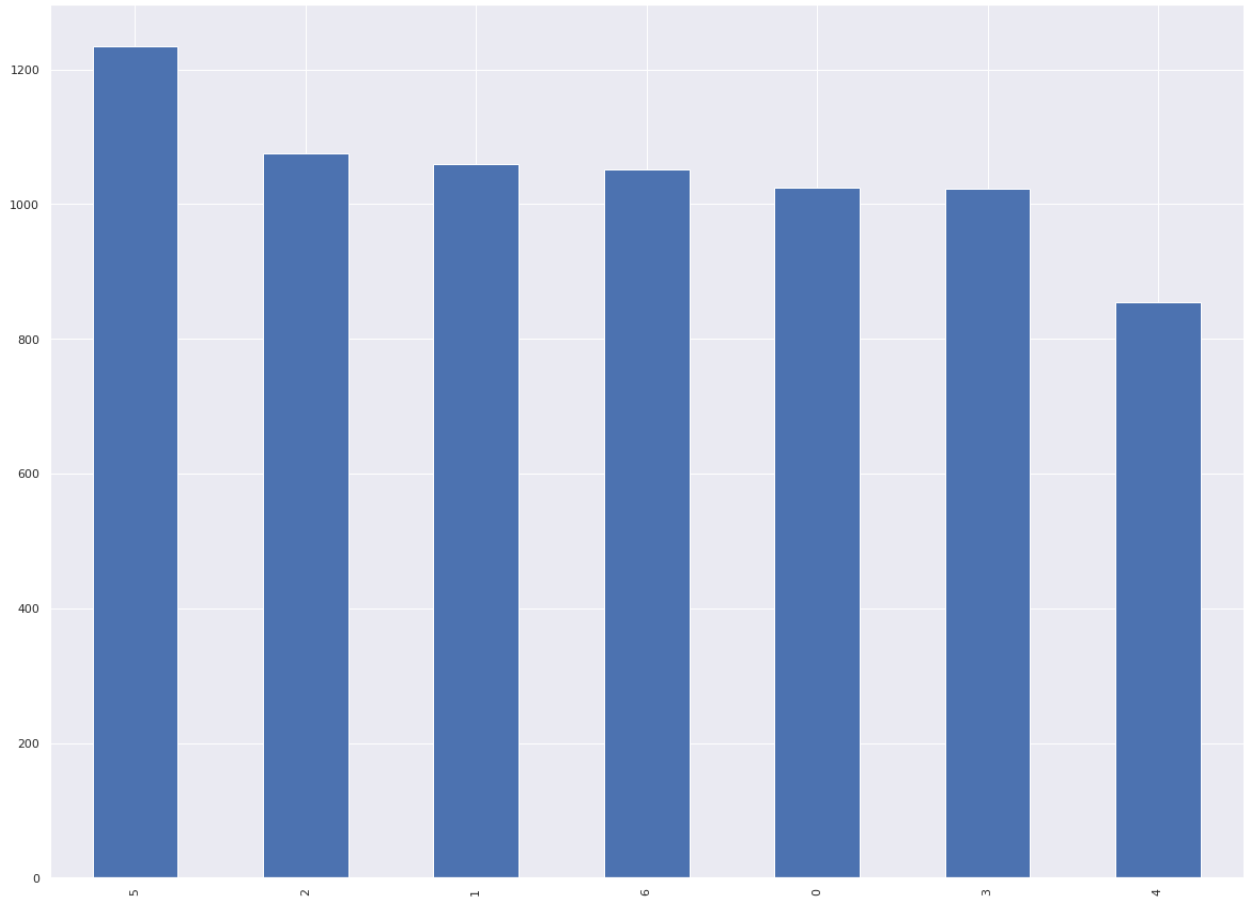
بررسی روز های ماه با بیش ترین تعداد پست

محور x شماره روز و محور y تعداد پست ها



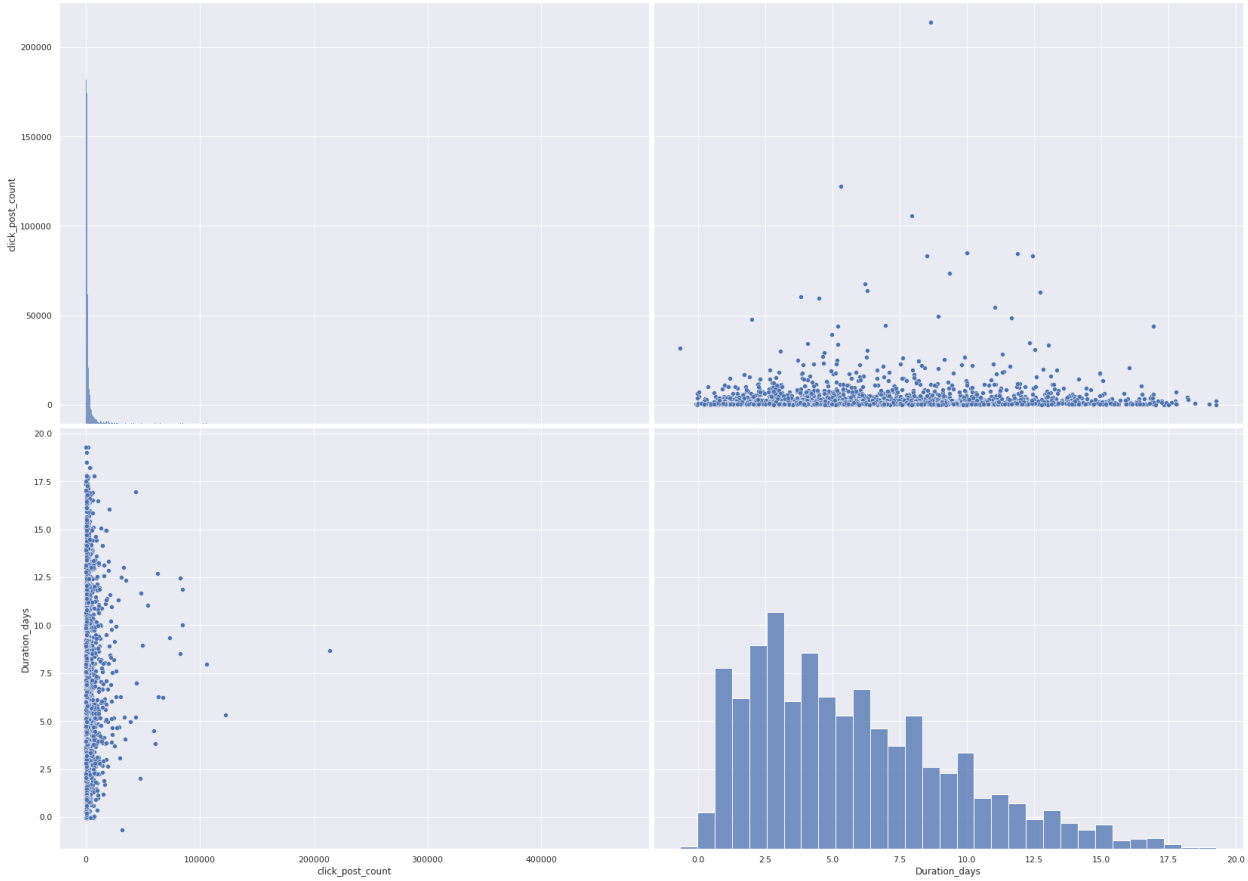
بررسی روز های هفته با بیش ترین تعداد پست

محور x شماره روز و محور y تعداد پست ها



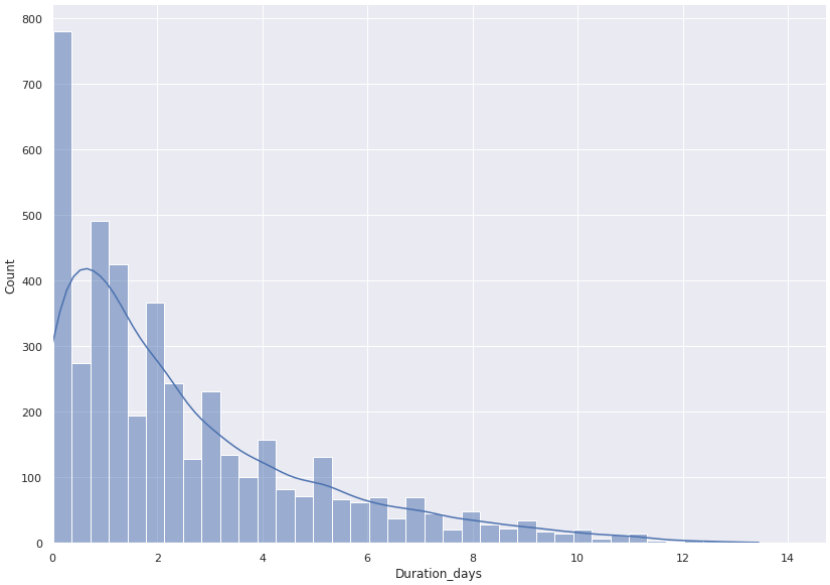
بررسی مدت زمان پست و کلیک های انجام شده

Pairplot، نوعی نمودار توزیعی است که اساساً به رسم یک نمودار مشترک برای کلیه ی ترکیبات ممکن ستون های عددی و بولی (Boolean)در دیتاست می پردازد.

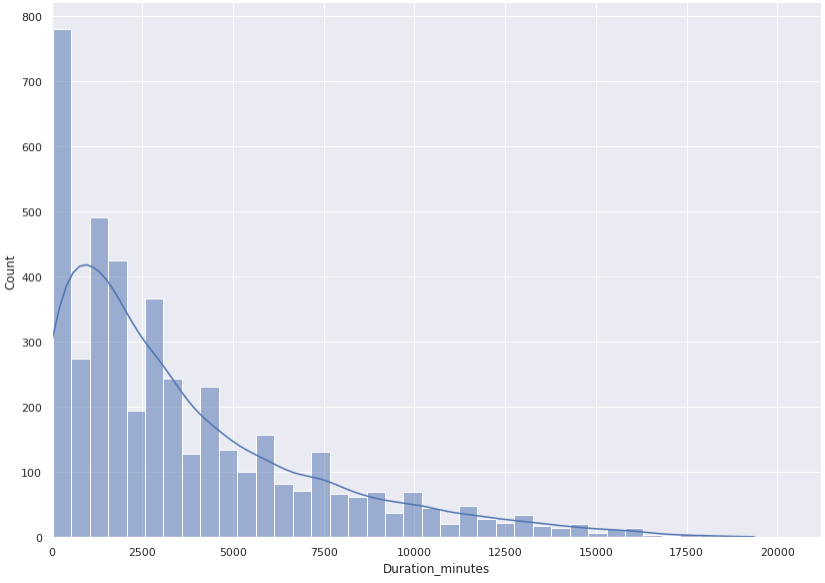


بررسی مدت زمان هر پست از زمان reorder یا urgent reorder در مقایسه با تعداد پست ها

مدت زمان (روز)

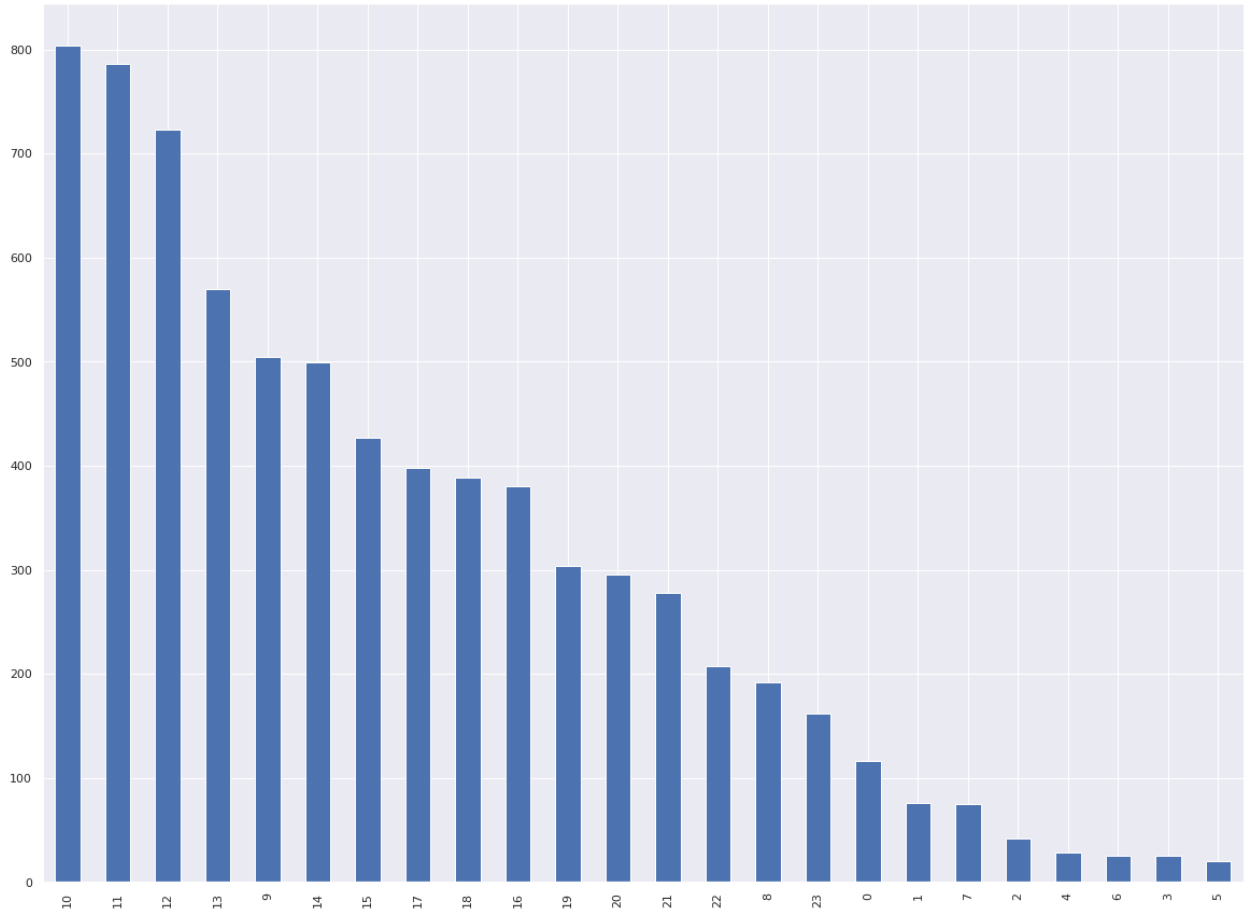


مدت زمان (دقیقه)



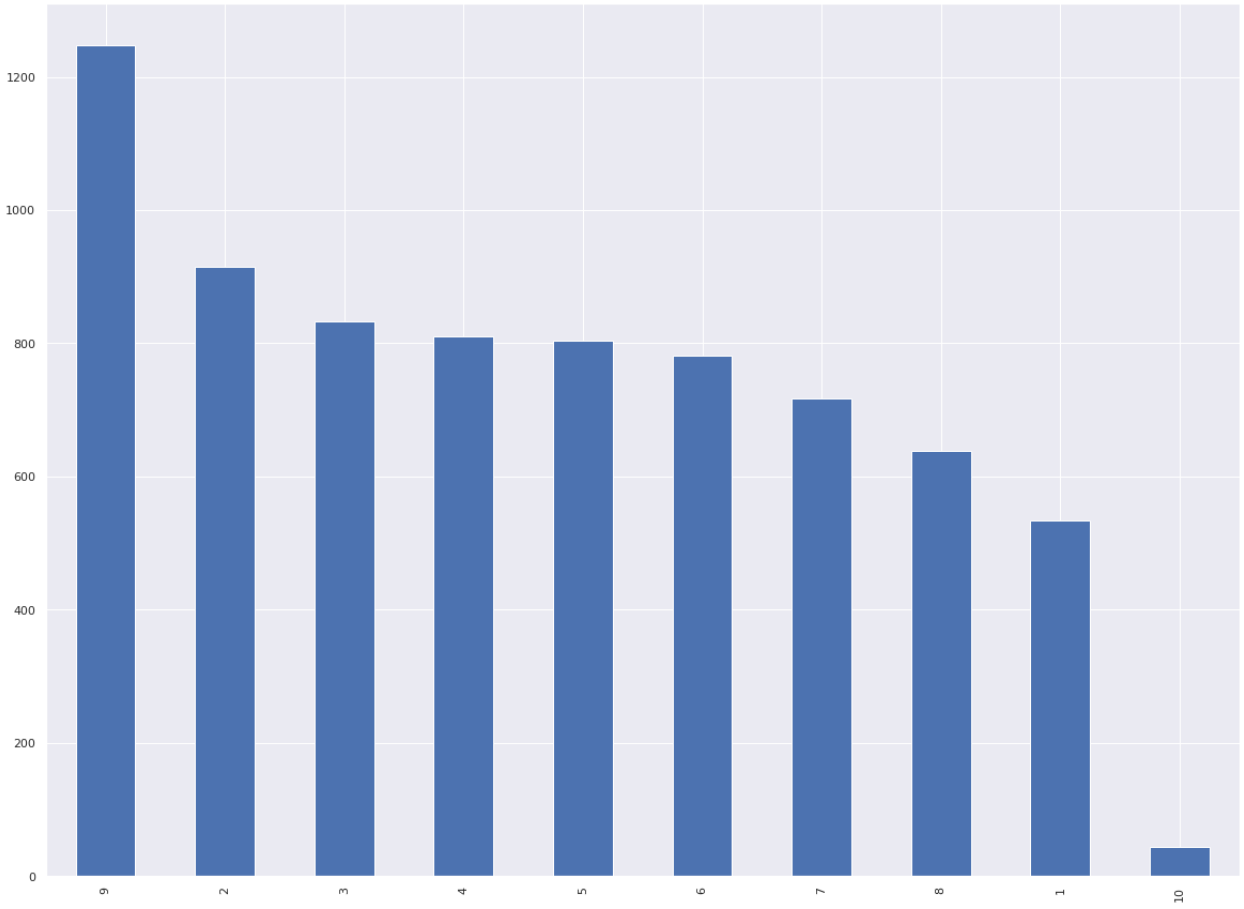
بررسی ساعات مختلف شبانه روز با بیش ترین تعداد پست

محور x شماره روز و محور y تعداد پست ها



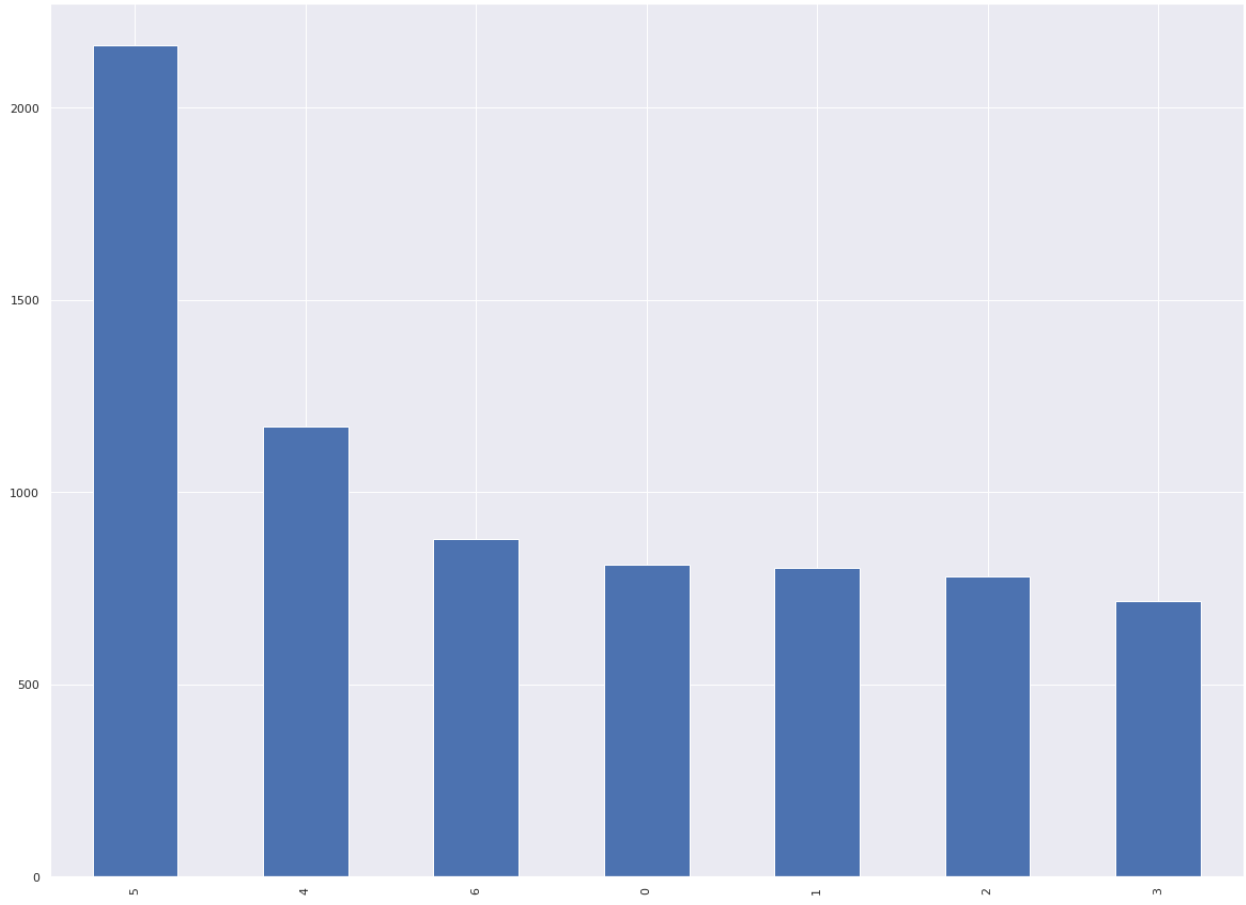
بررسی روز های ماه با بیش ترین تعداد پست

محور x شماره روز و محور y تعداد پست ها

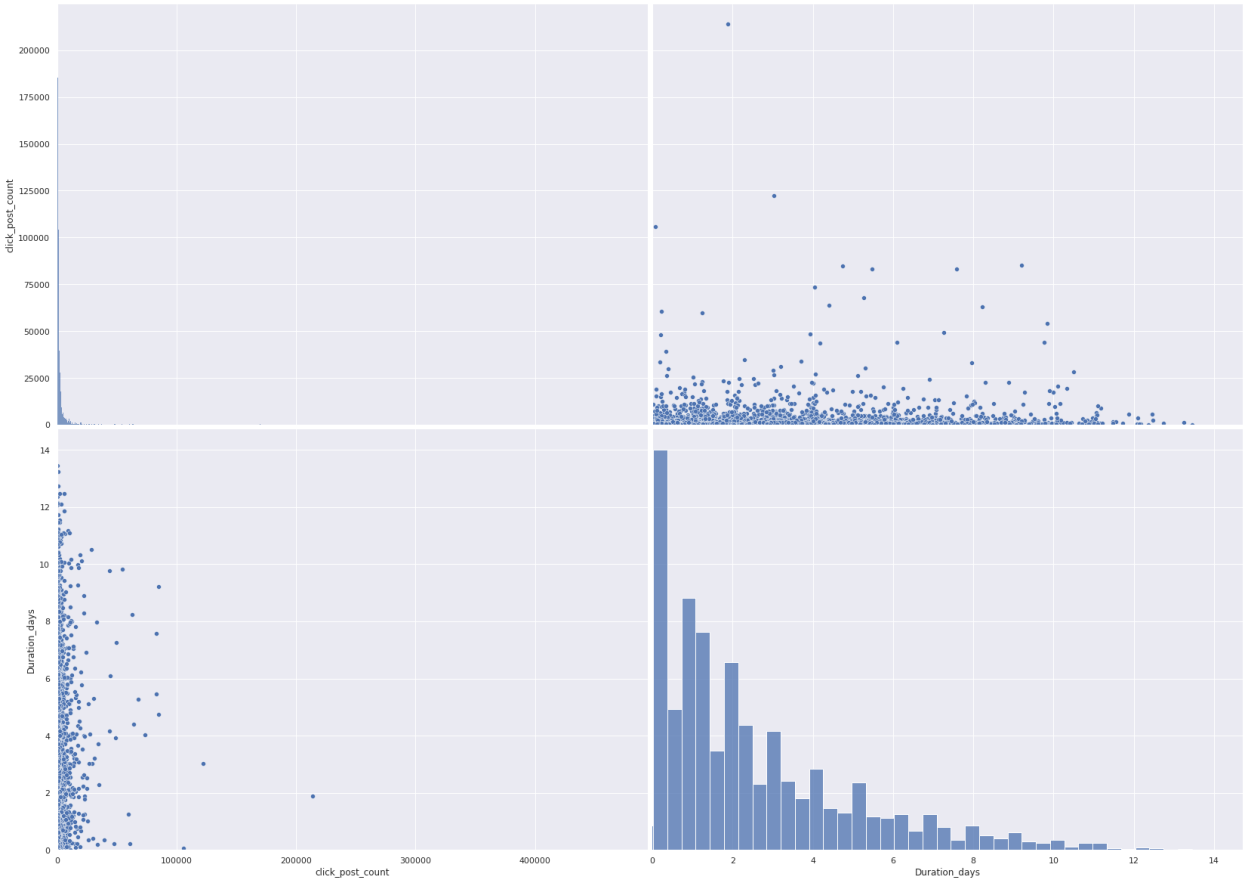


بررسی روز های هفته با بیش ترین تعداد پست

محور x شماره روز و محور y تعداد پست ها



بررسی مدت زمان پست (از زمان reorder)و کلیک های انجام شده

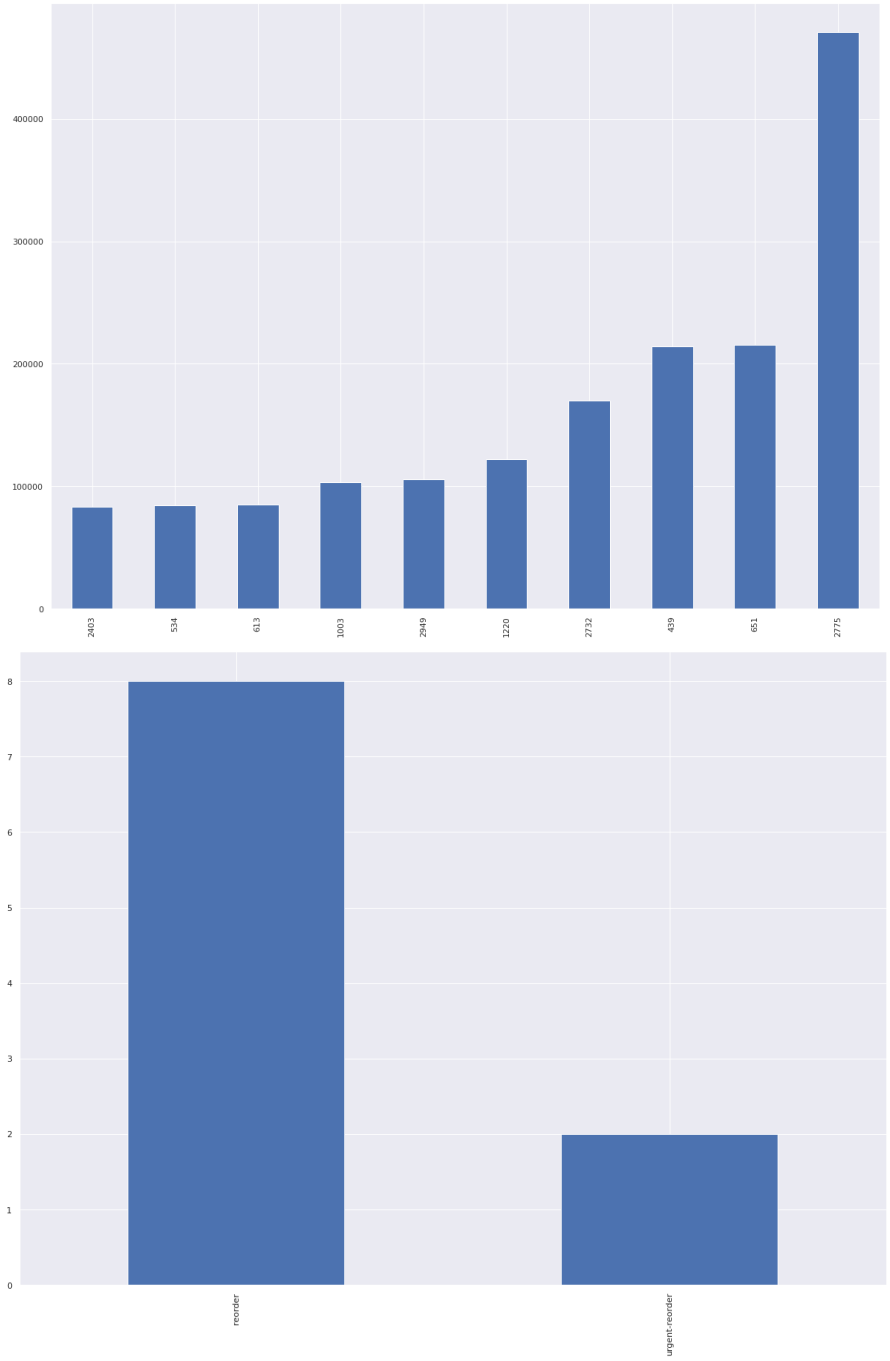


بررسی 10 پست برتر

در نمودار اول تعداد کلیک های 10 پست برتر را مشاهده می کنیم

در مودار دوم تگ های 10 پست برتر مایش داده شده که 8 پست دارای تگ reoderو 2 پست داری تگ urgantreoder است

محور x شماره پست را نشا می دهد و محور y تعداد کلیک



نتایج ttest

آزمون tدانشجویی به منظور اثبات مفید بودن ویژگی ها انجام شده است. آزمون tیکی از رایج ترین روش های آماری است که از واریانس ها برای بررسی احتمال تفاوت معنی داری در میانگین دو تگ مختلف استفاده می کند. آلفای 0.05به عنوان حد اعتبار در تحلیل استفاده می شود-.مقدار کمتر از 0.05فرضیه تهی را رد می کند و استنباط می کند که یک تفاوت قابل توجه وجود دارد

Ttest_indResult(statistic=-2.152244456214511, pvalue=0.03142323444791511)

با توجه به نمودار و نتایج در می یابیم که تگ reorder میانگی کلیک بیش تری نسبت یه تگ urgantreorder دارد

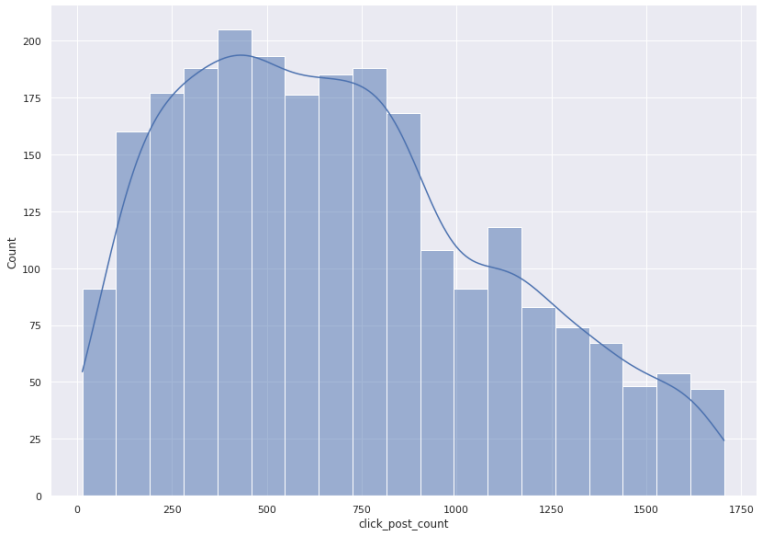


image 1 Reorder

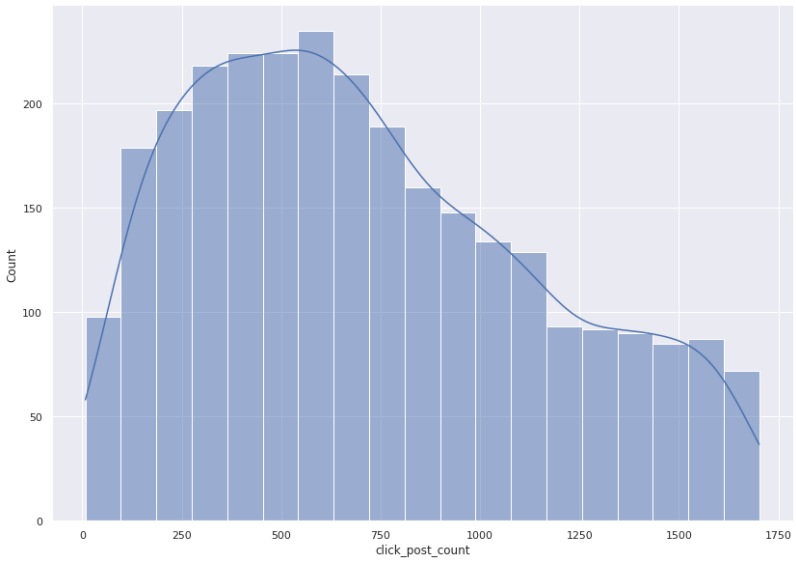


image 2 Urgant-Reorder