پروژه پایانی درس مبانی یادگیری آماری

استاد درس: دکتر نیک‌آبادی

دانشجو: الهه محمدی

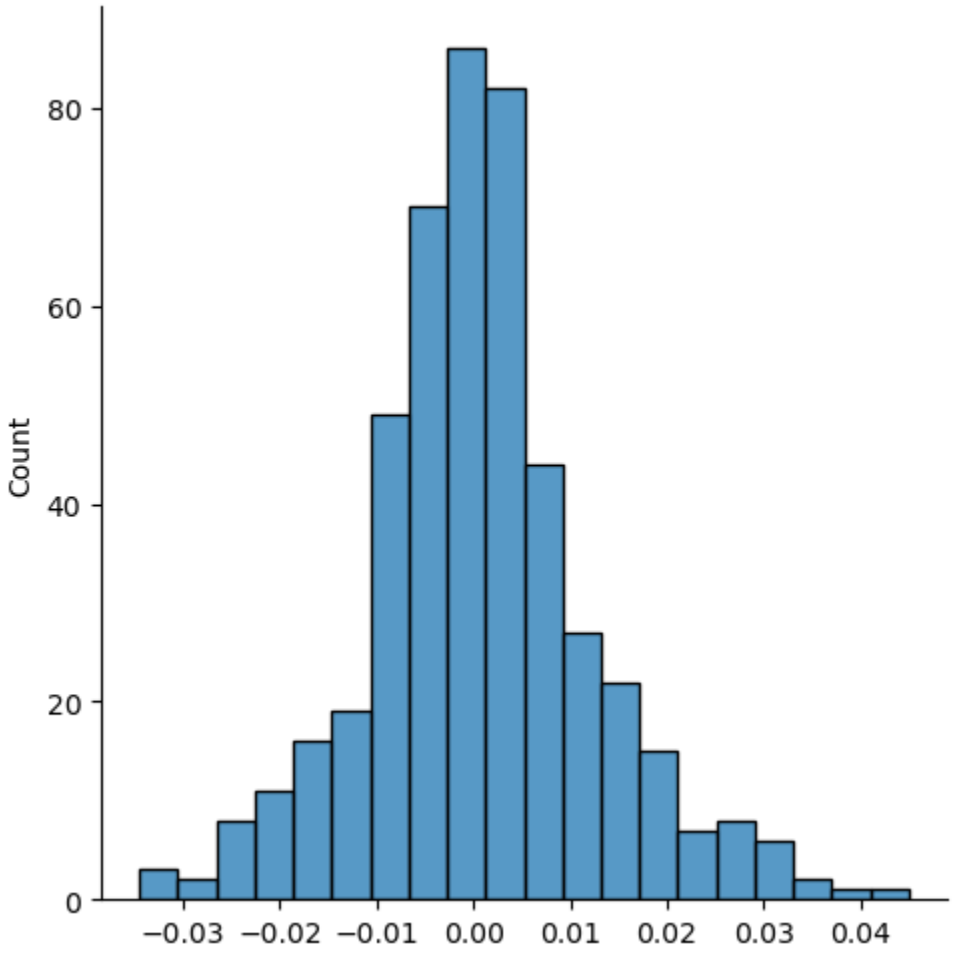
شماره دانشجویی: ۴۰۰۱۳۱۹۱۹

بهمن ۱۴۰۱

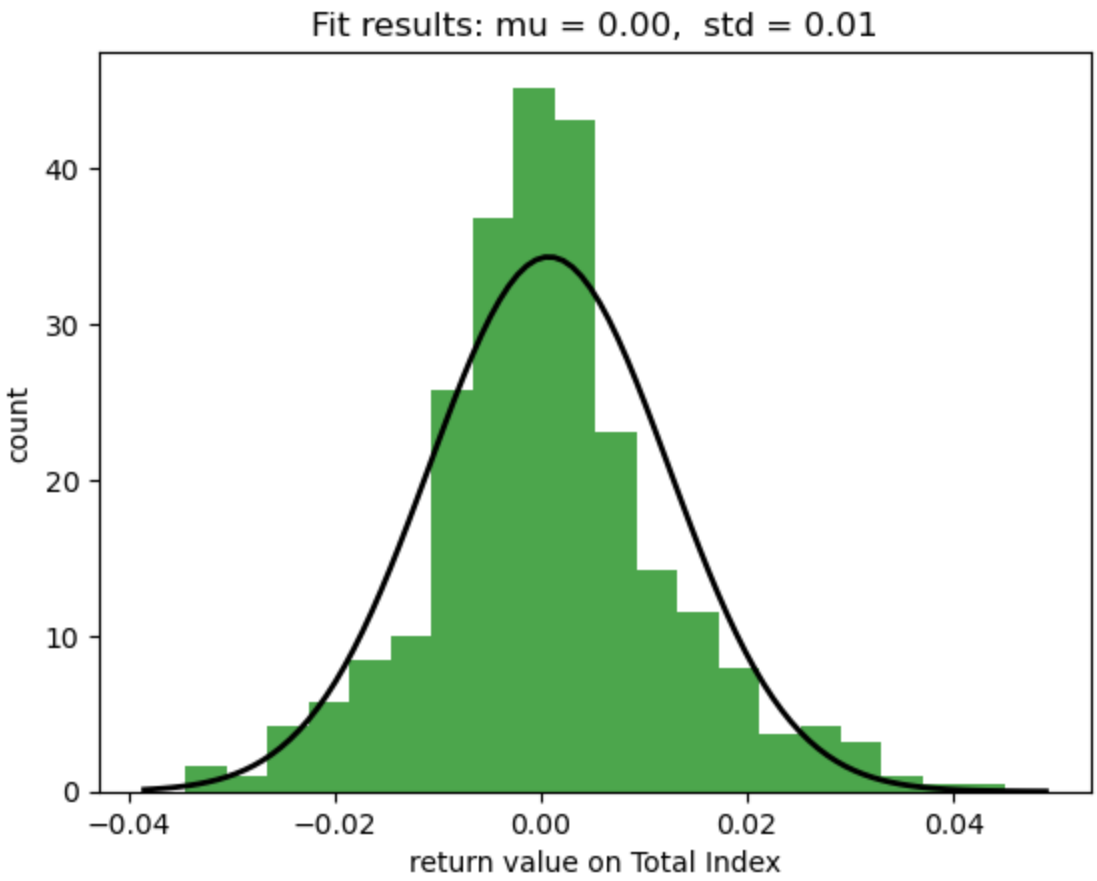
سوال ۱:

الف) توزیع احتمالاتی مقادیر بازده شاخص کل را به دو روش پارامتری (با استفاده از یک توزیع نرمال) و غیرپارامتری (هیستوگرام گسسته) تخمین زده و نتایج را تحلیل کنید.

فایل TotalIndex.csv در پوشه Data حاوی داده دو ساله شاخص کل در بازه یک بهمن ۱۳۹۹ تا یک بهمن ۱۴۰۱ است. این فایل را خوانده و با محاسبه مقدار بازده برای هر دو مقدار متوالی از این فایل، در روش غیر پارامتری، هیستوگرام گسسته آن را رسم می‌کنیم که به صورت زیر خواهد بود.



همچنین برای ارائه یک تخمین پارامتریک از جنس توزیع نرمال بر روی مقادیر بازده محاسبه شده، از کتابخانه scipy.stats.norm استفاده می‌کنیم. نتیجه کار به صورت زیر خواهد بود.



همچنان که از تصویر فوق مشخص است، مقدار بازده شاخص کل از یک توزیع نرمال دقیق تبعیت نمی‌کند و اساسا شکل پراکندگی داده‌ها در سمت چپ مقدار ۰ با پراکندگی سمت راست آن متفاوت است. همچنین قله توزیع نرمال تخمین زده شده با قله نمودار هیستوگرام به درستی تطبیق ندارند. همچنین یک تست فرضیه با کتابخانه scipy.stats.normaltest اجرا شد که نتیجه آن موید همین مسئله است که مقدار بازده شاخص کل از یک توزیع نرمال تبعیت نمی‌کند.

ب) متقارن یا نامتقارن بودن توزیع را با استفاده از آزمونهای مربوطه بررسی کرده و آن را تحلیل کنید.

برای این‌که متقارن بودن این توزیع را بررسی کنیم، از فرض متقارن بودن داده‌ها در توزیع نرمال استفاده می‌کنیم. با توجه به بخش قبلی سوال، این داده از یک توزیع نرمال تبعیت نمی‌کند و بنابراین در تست‌های سنجش نرمال بودن توزیع، فرض صفر که فرض نرمال بودن توزیع داده است رد می‌شود. همچنین با اجرای تست‌هایی که فاکتور skewness را مستقیما اندازه گیری می‌کند نتیجه باز هم نشان‌دهنده عدم تقارن توزیع داده است. در این راستا تست کولموگروف برای بررسی تقارن توزیع داده با فرض یک توزیع نرمال اجرا می‌شود که با کوچکتر به دست

آمدن p-value از ۰.۰۵، ما تصمیم می‌گیریم این توزیع از توزیع نرمال تبعیت نمی‌کند و در نتیجه متقارن نیست. (تست scipy.stats.skewtest و scipy.stats.normaltest هم مورد ارزیابی قرار گرفت و نتیجه همان بود)

ج) میانگین و واریانس بازده هر یک از سهم‌ها در کل بازه مشخص شده را محاسبه کرده و آنها را با یکدیگر مقایسه کنید. نتایج را بر اساس وضعیت تغییرات قیمت سهم‌های مورد نظر در بازه مذکور تفسیر کنید.

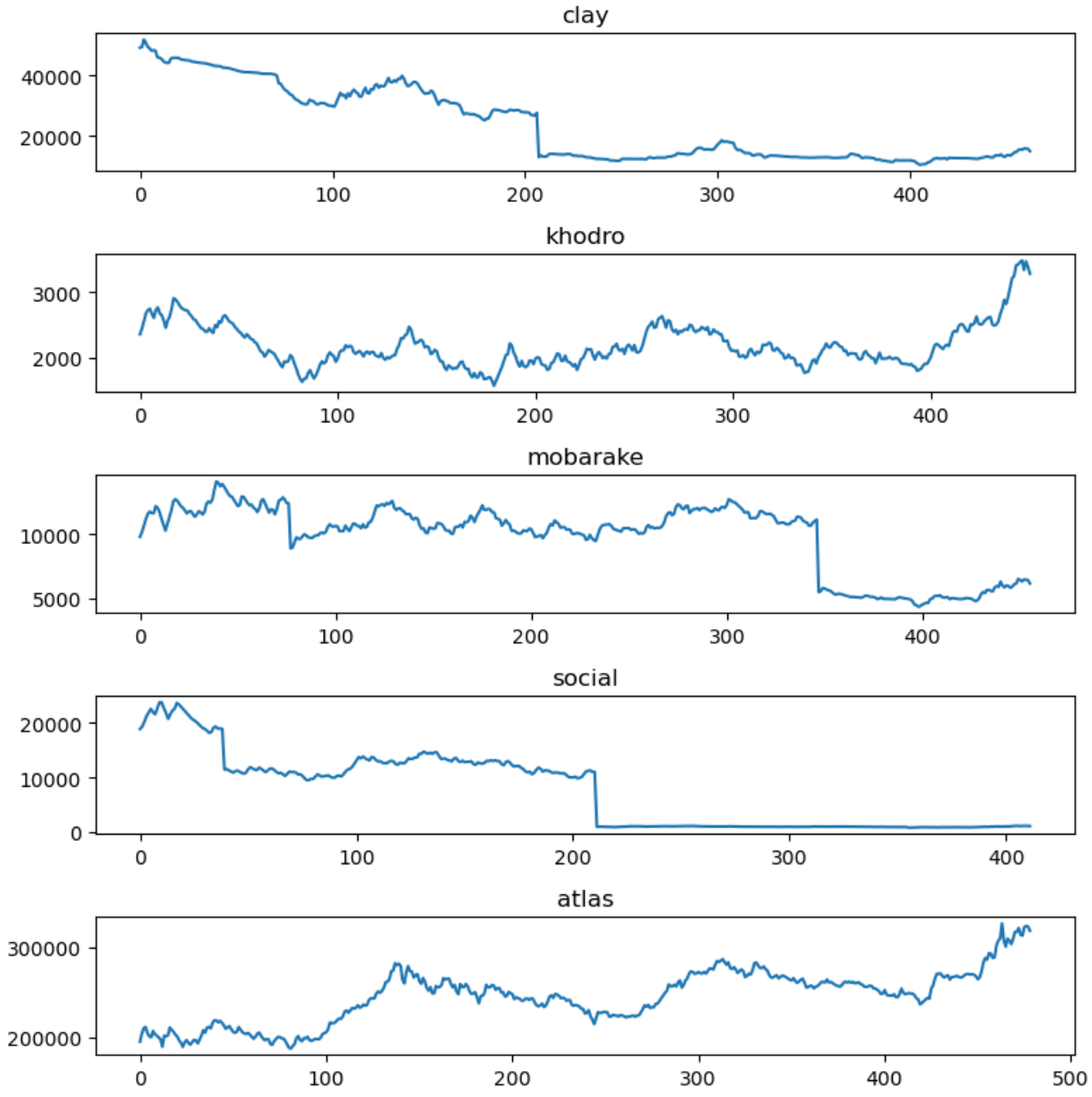
میانگین و واریانس محاسبه شده بر روی بازدهی سهم‌ها در کل بازه دوساله به صورت زیر بدست آمد.

```
clay_mean= -0.001894 clay_var= 0.000982
khodro_mean= 0.001092 khodro_var= 0.000718
mobarake_mean= -0.000241 mobarake_var= 0.001215
social_mean= -0.002678 social_var= 0.002941
atlas_mean= 0.001138 atlas_var= 0.000226
```

```
clay_mean= -0.001909 clay_var= 0.000984
khodro_mean= 0.000912 khodro_var= 0.000714
mobarake_mean= -0.000434 mobarake_var= 0.001212
social_mean= -0.002861 social_var= 0.002948
atlas_mean= 0.001043 atlas_var= 0.000222
```

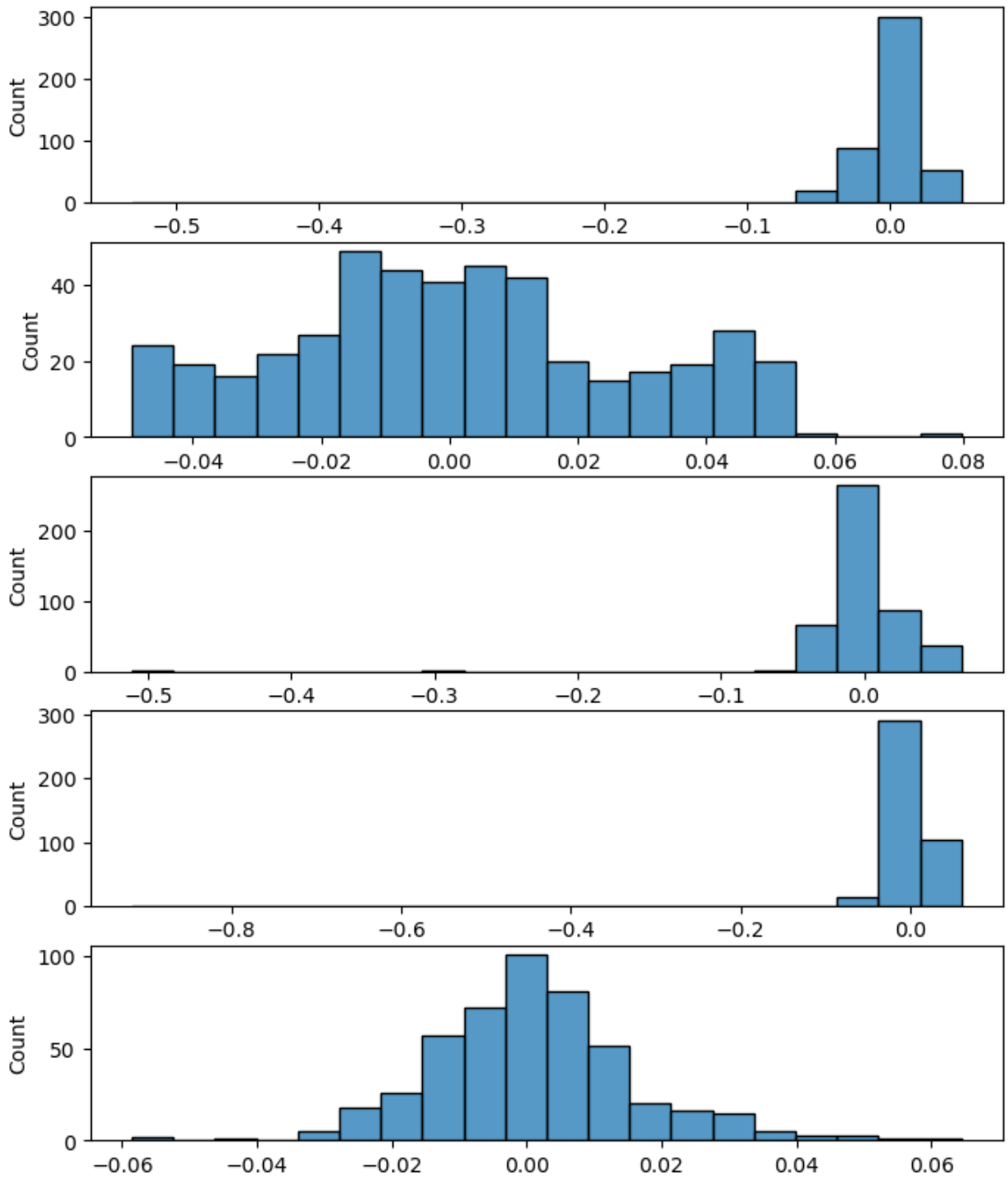
با توجه به اینکه قیمت پایانی سهم در روز بعد به اندازه منفی ۵ درصد تا مثبت ۵ درصد قیمت پایانی روز قبل تغییر می‌کند مقادیر به دست آمده در بالا را آنالیز می‌کنیم. در مقادیر میانگین، سهم‌های ایران خودرو و صندوق اطلس مقادیر مشابهی دارند و سهم شستا میانگین بازدهی منفی در حدود -0.28% دارد. به این ترتیب متوسط تغییرات قیمت روی سهم شستا بیشتر از بقیه بوده و بعد از آن سهم کخاک میانگین تغییرات قیمت بیشتری را تجربه کرده است. اگر تغییرات قیمت سهم‌های مختلف را مانند شکل زیر رسم کنیم، این موضوع که تغییرات قیمتی دو سهم شستا و کخاک از سایرین بیشتر است، به وضوح دیده می‌شود. برای مثال در سهم صنایع خاک چینی ایران (کخاک)، در بازه یک هفته‌ای ۷ دسامبر ۲۰۲۱ تا ۱۵ دسامبر ۲۰۲۱ قیمت پایانی سهم از حدود ۲۶۰۰۰ به حدود ۱۳۰۰۰ نزول میکند:) این مورد در مورد سهم شستا نزول شدیدتری را نشان میدهد. چون در

شروع این بازه دو ساله قیمت پایانی در حدود ۲۰۰۰۰ بوده و بعد از گذشت حدود یکسال به نزدیک ۱۰۰۰ نزول می‌کند و یکسال باقیمانده را در حوالی همین عدد سیر می‌کند.



بیشترین واریانس تغییرات بازدهی روی سهام شستا بوده (در حدود 0.3%) و بعد از آن بیشترین واریانس روی تغییرات بازدهی سهام فولاد مبارکه و کخاک است (در حدود 0.1%).

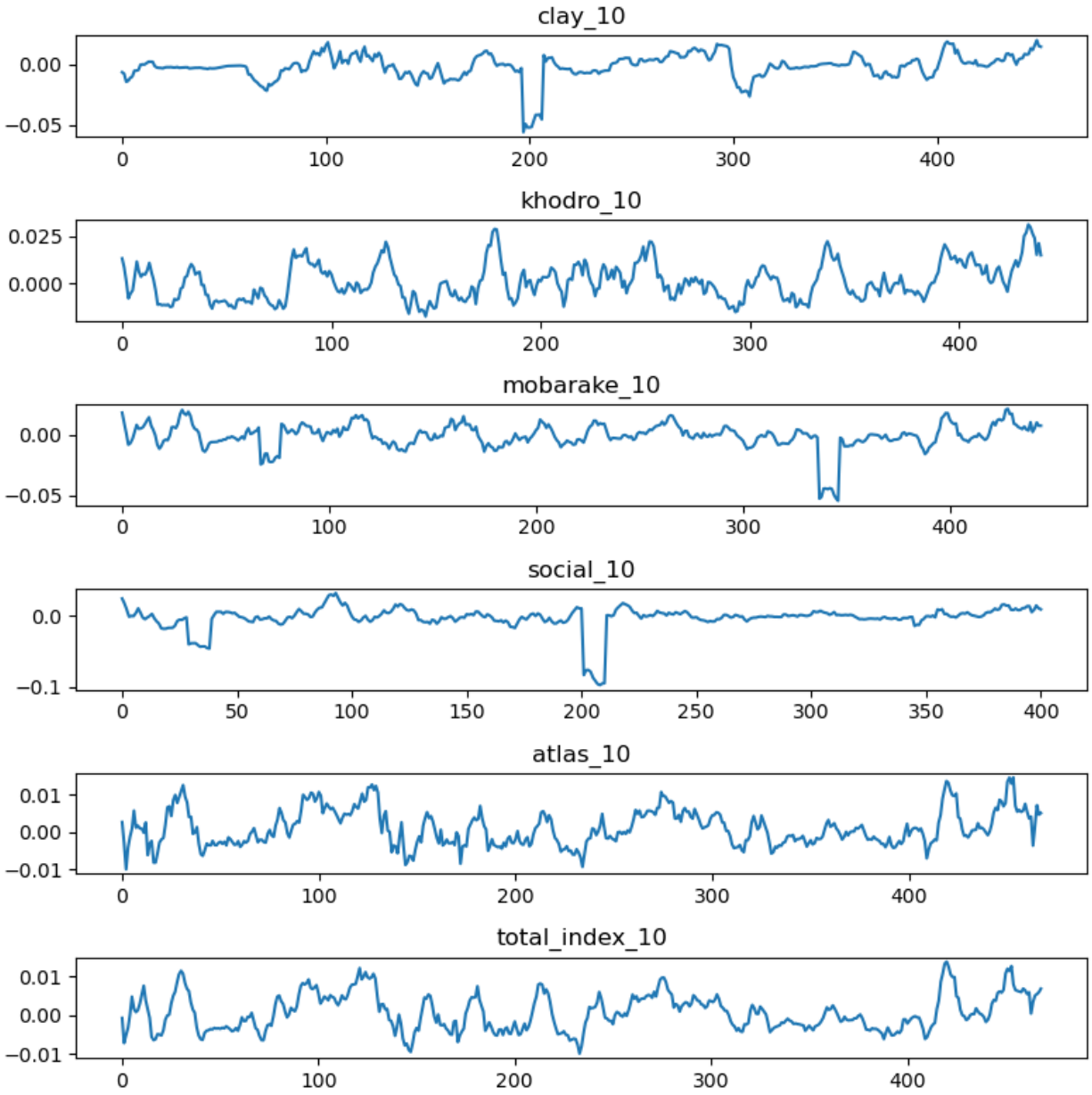
هیستوگرام تغییرات روی میزان بازدهی این سهم‌ها نیز در شکل زیر نشان داده شده است.


Count
300
200
100
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Count
40
20
0
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Count
200
100
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Count
300
200
100
0
−0.8
−0.6
−0.4
−0.2
0.0
Count
100
50
0
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06

با مقایسه نتایج به دست آمده از میانگین و واریانس بازده هر یک از سهم‌ها در بازه دو ساله و مقایسه آن با نمودار اول در بالا یعنی نمودار تغییرات قیمت سهم‌ها، به این نتیجه می‌رسیم که بازه دو ساله به خوبی تغییرات قیمت را نشان نداده است. چون با وجود افت های شدید، مثلا میانگین بازدهی‌ها در سهام شستا در حدود -0.26% شده که به نظرم خیلی گویای مسئله کاهش شدید ارزش سهام شستا نیست.

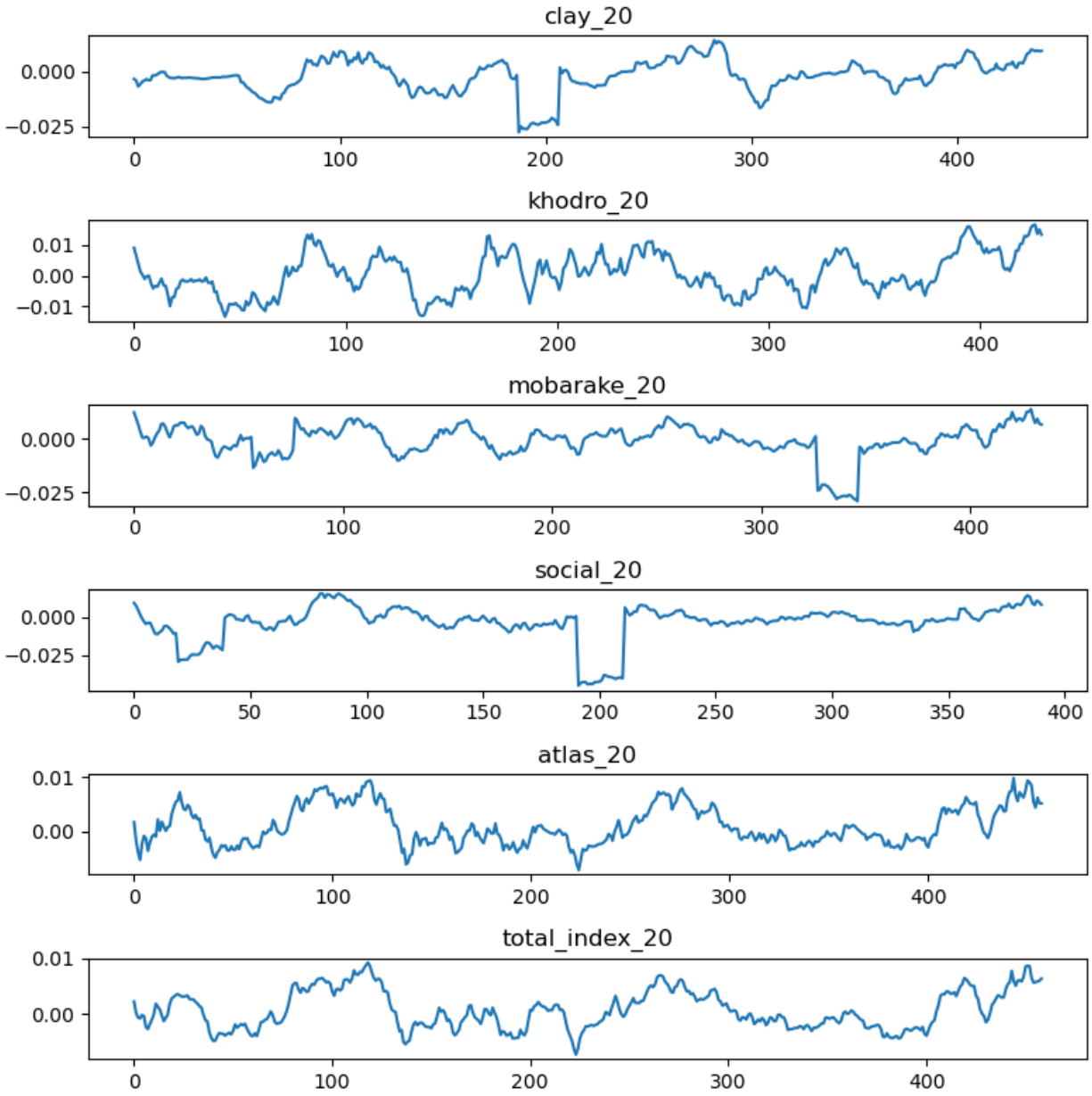
د) تغییرات میانگین و واریانس بازده شاخص کل و هر یک از سهم‌ها در پنجره‌های زمانی مشخص(۱۰، ۲۰ و ۵۰ روزه) در طول زمان را بررسی کنید.

برای بازه ۱۰ روزه:



سهم‌های کخاک، مبارکه و شستا در بازه‌هایی دچار افزایش سرمایه شده‌اند و در نتیجه قیمت پایانی افت زیادی داشته و سپس روی قیمت جدیدشان ادامه داده‌اند. در نتیجه در بازه‌هایی که تغییرات قیمت به تازگی اتفاق افتاده، کاهش زیاد میانگین بازدهی‌ها را می‌بینیم. در مورد خودرو میانگین بازدهی‌ها بیشتر اوقات مثبت بوده است. صندوق اطلس روی میانگین بازدهی‌ها نوسان دارد. اما بازه این نوسان‌ها کوچک بوده و به نظر میرسد نمودار سهم صندوق اطلس به نموداری که در اینجا برای شاخص کل داریم نزدیکتر باشد، خصوصا اینکه بازه نوسان این دو نمودار یکسان به نظر می آید. سهام خودرو نیز به نظر میرسد با نوسان ۰.۰۲ درصد، روند نسبتا مثبتی داشته باشد.

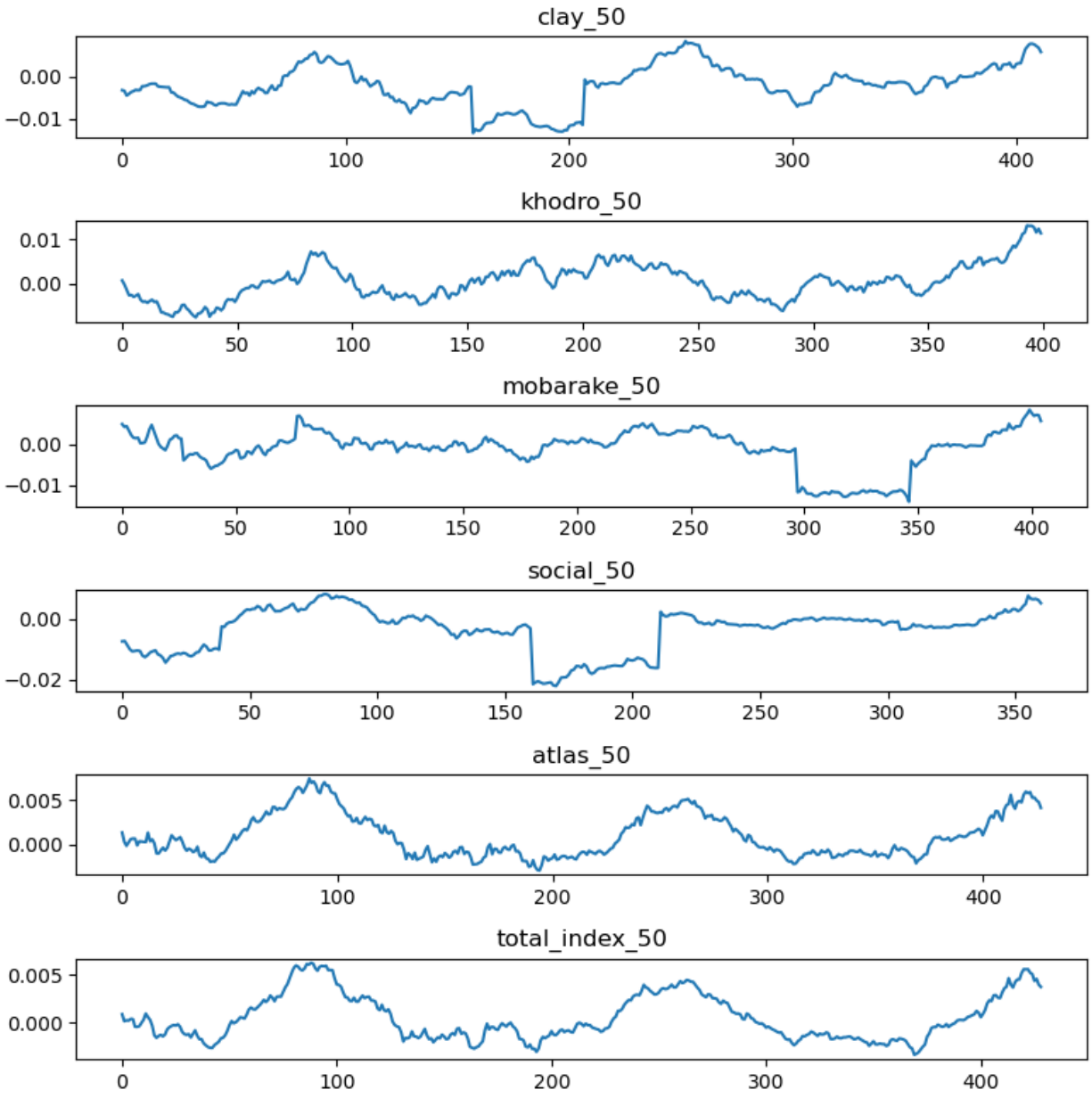
برای بازه ۲۰ روزه:



مشابه تفسیر بالا، در سهم‌های مبارکه، شستا و کخاک، افت زیاد میانگین تغییرات قیمت پایانی در بازه‌ای از زمان را می‌بینیم. با توجه به بزرگتر شدن بازه میانگین، بازه زمانی این رویداد، بزرگتر از حالت قبلی به چشم می‌آید. اما با توجه به بزرگتر شدن پنجره زمانی، از نظر مقدار عددی، کوچکتر از حالت قبلی شده است. برای مثال در مورد سهم کخاک، مقدار میانگین در بازه شروع افت قیمت پایانی، برای پنجره زمانی ۱۰ روزه به عدد منفی ۰.۰۵ درصد رسیده و برای پنجره زمانی ۲۰ روزه به عدد منفی ۰.۰۲ درصد رسیده است. در مورد سهم خودرو به نظر میرسد در یکسال اخیر روبه رشد بوده (به بازه پس از روز ۲۰۰ ام دقت کنید) و بازه تغییرات میانگین تغییرات قیمت،

کوچکتر از قبلی به دست می‌آید. در مورد سهم مبارکه به نظر میرسد در حدو ۲ ماه اخیر روبه رشد بوده(در حدود ۰.۰۱ درصد). اما در طی ۲ سال قبل تا حدود ۲ ماه قبل، میانگین تغییرات قیمت پایانی، بیشتر اطراف صفر نوسان داشته است. همچنین باز هم می‌بینیم روند تغییرات میانگین بازدهی روی سهم اطلس شبیه به روند تغییرات میانگین بازده شاخص کل بوده است و حتی از نظر واریانس تغیرات میانگین هم مشابه هم بوده اند.

برای بازه ۵۰ روزه:

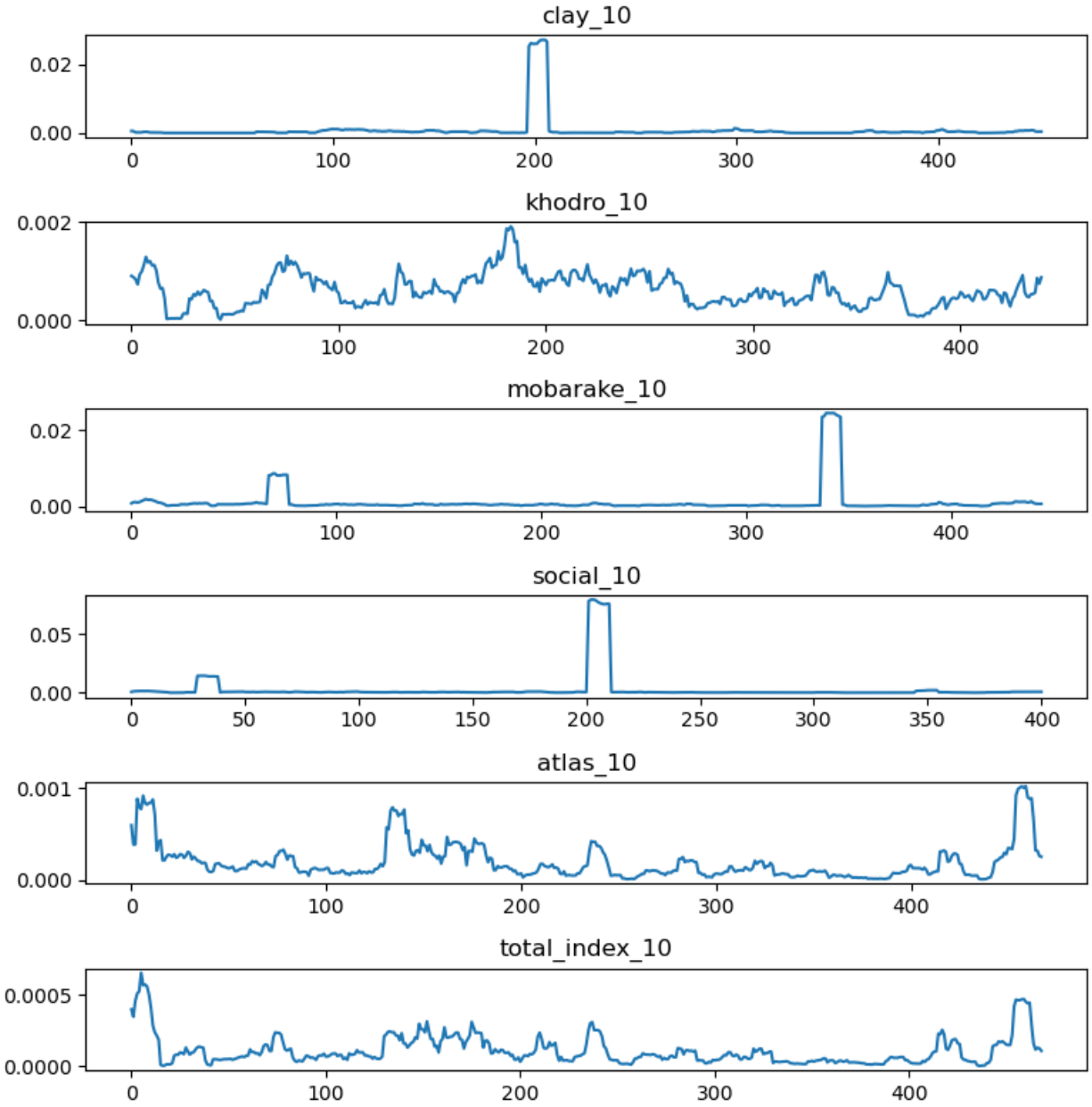


در پنجره زمانی ۵۰ روزه، نگاشت تغییرات سهم اطلس با تغییرات قیمت شاخص کل محسوس‌تر دیده می‌شود. بازه افت قیمت در سهم‌های کخاک، شستا و مبارکه طولانی‌تر شده، اما مقدارا به صفر نزدیکتر شده‌اند. همچنین باز هم صعودی شدن سهم خودرو در ۵۰ رکورد اخیر داده‌ها دیده می‌شود.

صعودی شدن روندها در طی ۲ ماه اخیر تقریبا در تمام سهم‌ها و نیز شاخص کل هم دیده می‌شود.

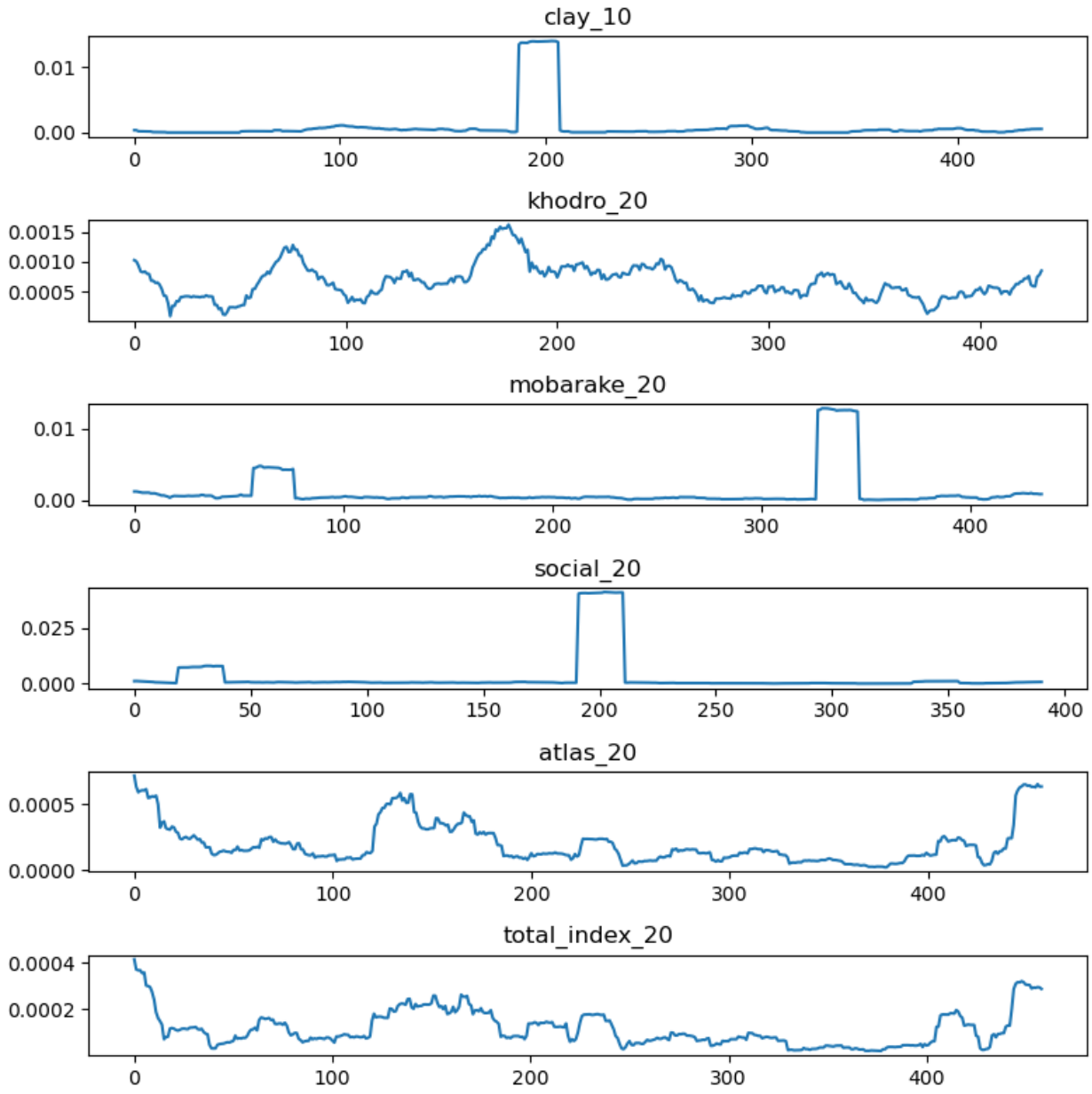
حالا واریانس تغییرات را بررسی می‌کنیم.

برای بازه ۱۰ روزه:

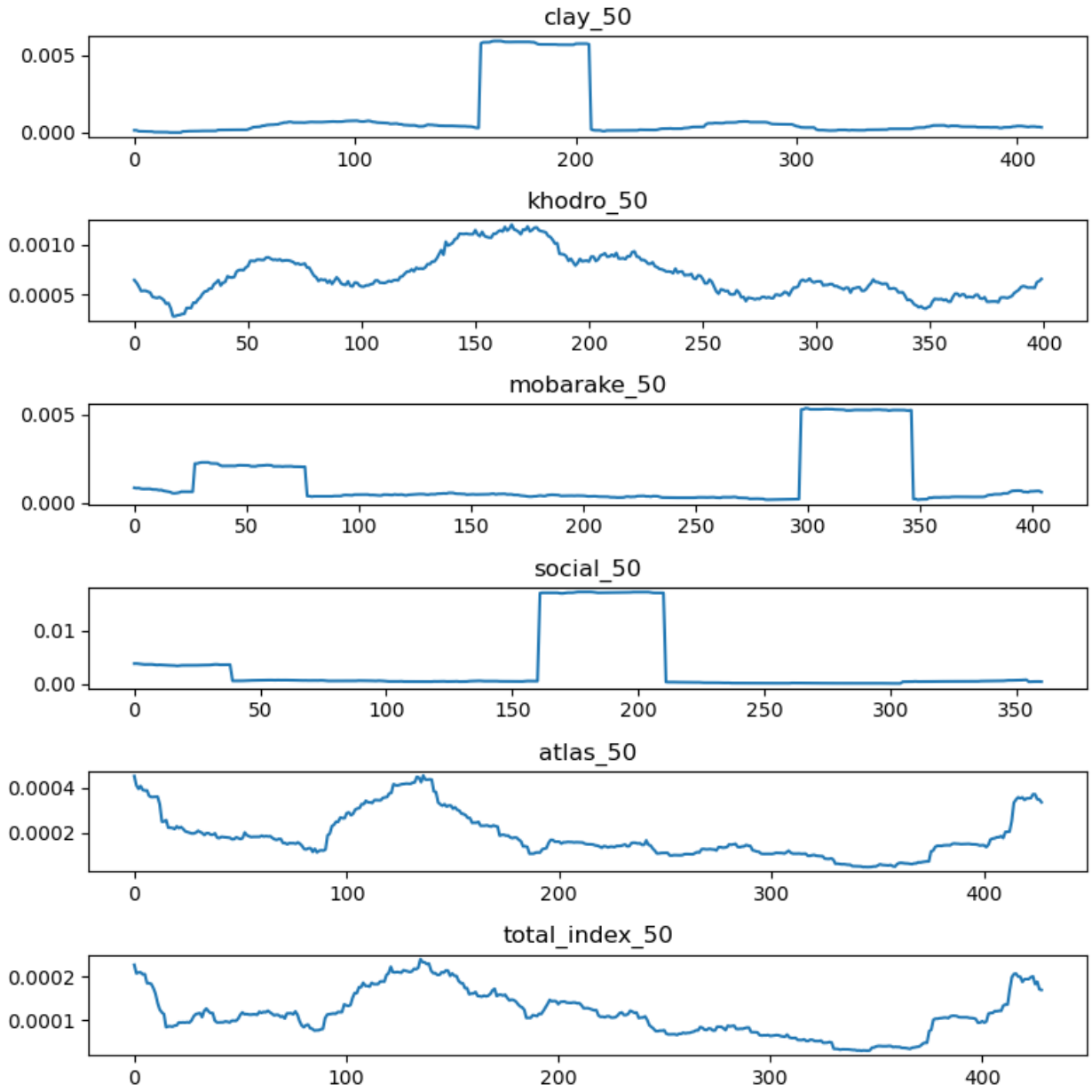


واریانس سهم‌های شستا و مبارکه و کخاک در زمان تغییر ناگهانی قیمت، نوسانی را نشان داده، البته چون بازه ۱۰ روزه است، تاثیر این رویداد در نمودار کوتاه بوده است. در سهم‌های مبارکه و شستا دو تغییر قیمت دیده می‌شود. یعنی در طی ابتدا تا انتهای ۱۴۰۰ دو مرتبه افزایش سهم انجام داده است. سهم اطلس نیز بیشتر مواقع مشابه با تغییرات شاخص کل رفتار کرده است.

برای بازه ۲۰ روزه:



برای بازه ۵۰ روزه:



واریانس تغییرات سهم اطلس و شاخص کل خیلی به هم شبیه هستند. همچنین واریانس سهم‌های شستا و مبارکه و کخاک در زمان تغییر ناگهانی قیمت، تغییر ناگهانی را نشان داده و با تثبیت شدن روی قیمت پایانی جدید، واریانس تغییرات بسیار کمتر می‌شود. به نظر میرسد سهم‌های خودرو و اطلس در طی دو ماه اخیر رو به مثبت جلو می‌روند.

سوال دوم: بررسی همبستگی‌ها

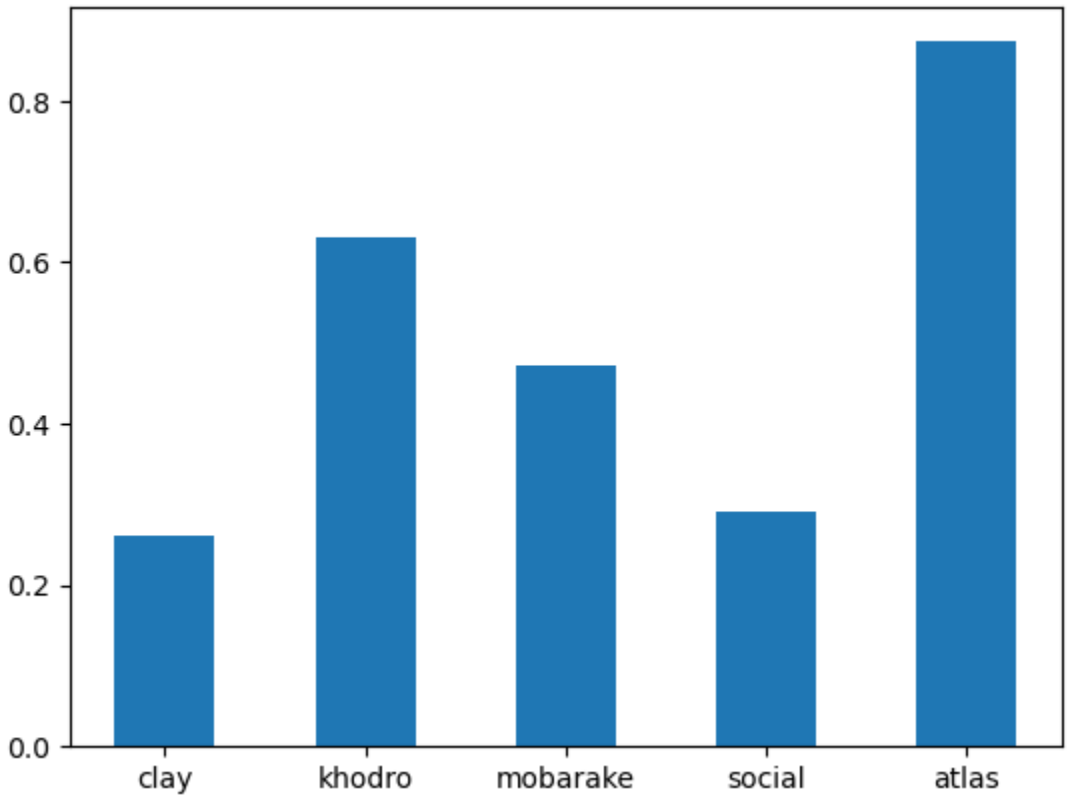
الف) میزان خودهمبستگی (autocorrelation) بازده شاخص کل لگ‌های زمانی مختلف (۱ تا ۵ روز) را محاسبه و تحلیل کنید.

میزان خودهمبستگی بازده شاخص کل در تاخیرهای زمانی ۱ تا ۵ روز به صورت زیر به دست می‌آید. بیشترین میزان خودهمبستگی در ۱ روز بعد و کمترین آن در ۵ روز بعد بوده است. نتایج نشان می‌دهد برای پیش‌بینی بازده شاخص کل، خیلی نمی‌توان از مقادیر این شاخص در چندین روز قبل استفاده کرد. میزان خودهمبستگی در تاخیرهای زمانی ۲ تا ۵ روز بعد به خوبی گویای این مسئله است.

```
array([ 0.30153472, -0.00441886,  0.09146403,  0.01421556, -0.03447565])
```

ب) میزان همبستگی(Correlation) بازده شاخص کل با بازده هر یک از سهم‌های مورد نظر را محاسبه کنید.

میزان همبستگی بازده شاخص کل با بازده سایر سهم‌ها در روند دوساله انتخاب شده به صورت زیر خواهد بود.

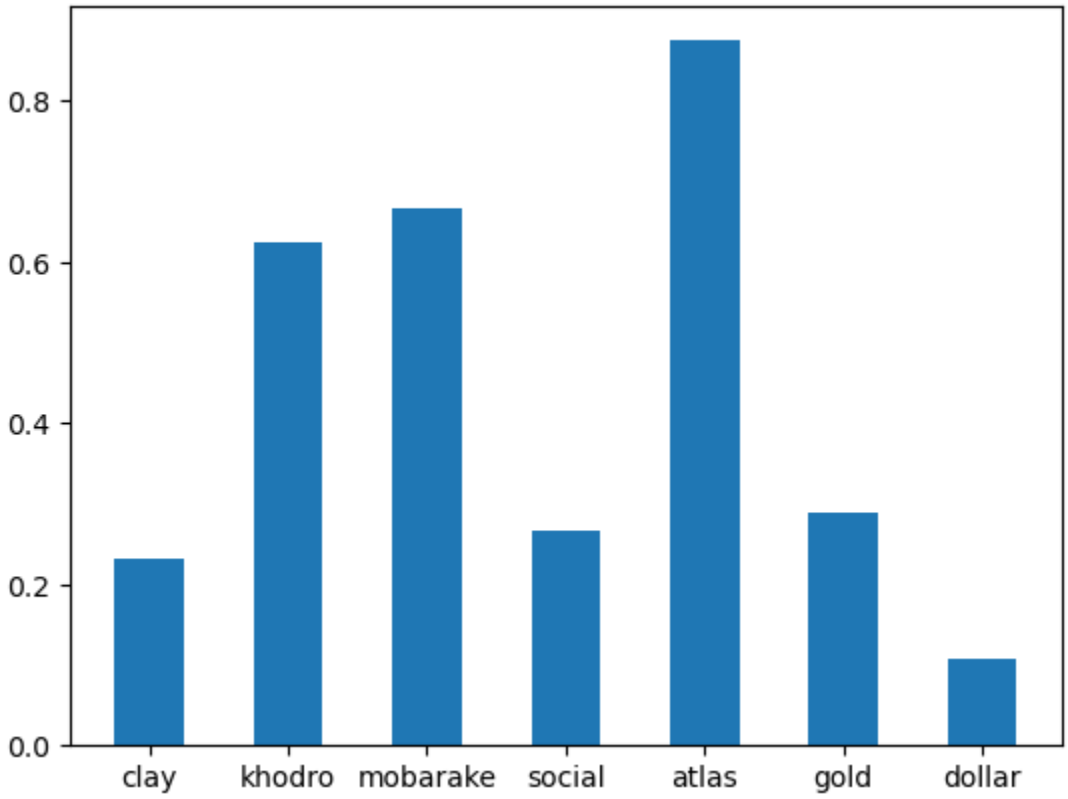


طبق نتایجی که در جدول فوق دیده می‌شود بیشترین همبستگی بازده شاخص کل با بازدهی سهم صندوق اطلس است و از آنجا که p-value نظیر با این تست کمتر از ۰.۰۵ است، بنابراین فرض صفر که به غیرهمبسته بودن خطی این دو سیگنال اشاره می‌کند رد می‌شود. با توجه به جدول، ترتیب همبستگی به ترتیب برای سهم اطلس، خودرو، مبارکه بوده و دو سهم شستا و کخاک در درجات بعدی همبستگی هستند.

برای محاسبه میزان همبستگی از تابع scipy.stats.pearsonr استفاده کردیم تا ضریب همبستگی پیرسون را پیدا کنیم. این ضریب ارتباط خطی بین دو متغیر را نشان می‌دهد و یک مقدار p-value نیز برای تست فرض صفر غیرهمبسته بودن برمی‌گرداند. مقدار این ضریب بین -۱ تا +۱ تغییر می‌کند و صفر بودن آن یعنی هیچ ارتباط خطی ندارند.

ج) همبستگی بازده شاخص کل با بازده طلا و دلار را مقایسه کنید(اختیاری)

پس از اشتراک گیری روی تاریخ‌ها، خروجی همبستگی ها با لحاظ کردن دادگان بازده طلا و بازده دلار به صورت زیر به دست می‌آید.



۳) پیشبینی مقدار بازده

در آزمایش‌های این بخش از داده‌های ۰ تا t برای آموزش و از داده‌های t تا T (اندیس آخرین روز) برای آزمون استفاده کنید.

الف) یک مدل رگرسیون خطی آموزش دهید که بر اساس اطلاعات چند روز بازده شاخص کل، مقدار بازده روز بعد را پیشبینی کند. در مورد نحوه انتخاب اطلاعات چند روز قبل از نتایج قسمت قبل کمک بگیرید.

با توجه به نتایج بخش خودهمبستگی در سوال ۲، من دو آزمایش یکبار با داده یک روز قبل و یکبار با دادگان ۳ روز قبل برای پیش بینی داده روز بعدی انجام دادم. نتیجه این آزمایشات نشان داد، داده یک روز قبل نتیجه بهتری در آماره MSE بدست می‌آورد.

```
Mean Squared Error:  0.00010733745204005 13
```

ب) مدل دسته‌بندی کننده آماری آموزش دهید که تنها روند تغییرات شاخص کل (یعنی مثبت یا منفی بودن بازده شاخص) را پیشبینی کند. دقت مدل را در پیشبینی یک، دو، سه و چهار روز بعد بررسی کنید.

برای این بخش از مدل logistic regression استفاده کردیم. چون با یک خروجی باینری سروکار داریم. به ازای هر یک از خروجی‌های پیش‌بینی نیز یک مدل مجزا آموزش داده شد. یعنی برای مثال برای پیش‌بینی یک روز بعد، از یک مدل و برای پیش‌بینی دو روز بعد از یک مدل دیگر استفاده شد.

دقت پیش‌بینی مدل:

```
result for 1 next steps:
Predicted   0   1
Actual
0          45  30
1          22  47
********************************************************************************
result for 2 next steps:
Predicted  0   1
Actual
0          3  72
1          5  63
********************************************************************************
result for 3 next steps:
Predicted   0   1
Actual
0          13  62
1          18  50
********************************************************************************
result for 4 next steps:
Predicted  0   1
Actual
0          4  71
1          7  61
********************************************************************************
```

دقت در یک روز بعد

$$Accuracy = (45+47)/(144) = 92/144=0.63$$

دقت پیش‌بینی مدل برای دو روز بعد:

$$Accuracy = 66/143=0.46$$

دقت پیش‌بینی مدل برای سه روز بعد:

$$Accuracy = 63/142 = 0.44$$

دقت پیش‌بینی مدل برای چهار روز بعد:

$$Accuracy = 65/141 = .46$$

ج) یک مدل ترکیبی آموزش دهید که با استفاده از مقادیر قبلی بازده شاخص کل و سهم‌های مد نظر، مقدار بازده شاخص کل و مثبت یا منفی بودن آن را پیش بینی کند.

با توجه به آزمایشاتی که در سوال ۲ انجام شد از تمام سهم‌ها به همراه خود شاخص کل با لگ زمانی ۱۰ استفاده می‌کنیم. نتیجه آزمایش برای معیار کمترین مربع خطا 0.0001435 بدست می‌آید. همچنین خروجی این مدل برای تعیین مثبت یا منفی بودن شاخص به صورت زیر خواهد بود:

```
[[26 24]
 [19 30]]
```

و به این ترتیب صحت در حدود ۵۶ درصد خواهد بود.

د) مدل رگرسیون Lasso را برای پیشبینی بازده شاخص کل با استفاده از اطلاعات ده روز قبل بازده شاخص کل به کار بگیرید. توانایی این مدل در انتخاب ویژگی‌های مطلوب را بررسی کنید.

رگرسیون lasso یک الگوریتم یادگیری ماشینی است که می‌تواند برای انجام رگرسیون خطی و در عین حال کاهش تعداد ویژگی‌های استفاده شده در مدل استفاده شود. این مدل برای جلوگیری از بیش برازش مدل استفاده می‌شود. همچنین برای انتخاب ویژگی با تنظیم برخی از ضرایب به عدد صفر استفاده می‌شود. رگرسیون lasso شکلی از رگرسیون خطی است، به این ترتیب که یک پارامتر منظم‌سازی در مجموع قدرمطلق وزن‌ها ضرب شده و به تابع لاس که least squares است اضافه می‌شود. رگرسیون lasso ، رگرسیون خطی منظم‌شده نیز نامیده می‌شود. ایده القای جریمه در برابر پیچیدگی، با افزودن عبارت منظم‌سازی به دست می‌آید به نحوی که با افزایش مقدار پارامتر منظم‌سازی، وزن‌ها کاهش می‌یابد (و در نتیجه جریمه ایجاد می‌شود) تا هدف کلی مجموع حداقل مربعات حفظ شود. برای تعیین مقدار منظم‌سازی از پارامتر آلفا در ورودی این تابع استفاده می‌شود. هر چه مقدار آلفا بیشتر باشد یعنی جریمه سنگین‌تری برای بیشتر شدن تعداد پارامترهای مدل داریم و در نتیجه تعداد ویژگی کمتری در مدل استفاده خواهد شد. رابطه رگرسیون lasso در زیر آورده شده است.

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

Parameters: $\theta = \{\theta_0, \theta_1, \theta_2, \ldots \theta_n\}$

Features: $x = \{x_0, x_1, x_2, \ldots x_n\}$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{m} \sum_{j=1}^{n} |\theta_j|$$

در ادامه مدل Lasso را اجرا گرفته و در انتها برای اندازه‌گیری میزان فیت شدن مدل روی داده‌ها از فراخوانی model.score(X_test, y_test) استفاده می‌کنیم. یک بودن این مقدار به معنای آن است که مدل به خوبی روی متغیر هدف پیش‌بینی می‌کند، اما صفر بودن آن به معنای نداشتن هرگونه اطلاعات مدل درباره متغیر هدف

است. این آماره میتواند مثبت یا منفی هم باشد که منفی بودن آن به معنای همبستگی منفی بین متغیر هدف و مدل است. هر چه این مقدار به صفر نزدیکتر باشد به معنای بدتر بودن پیش‌بینی مدل در توصیف متغیر هدف است. ما تابع score را هم برای دادگان آموزش و تست استفاده می‌کنیم و هرچه خروجی این امتیاز برای داده آموزش بیشتر از این امتیاز به ازای داده تست باشد، بیش برازش بیشتری در مدل اتفاق افتاده است. همچنین خروجی mean squared error را نیز بر روی داده تست و داده آموزش به دست آوردیم.

در مجموعه آزمایشاتی که برای تعیین آلفا انجام شد، مقدار ۰.۰۰۱ برای آلفا انتخاب شد. کوچکتر از آن هم می‌تواند استفاده شود اما در این صورت، دیگر هیچ ضریبی برای ویژگی‌ها صفر به دست نمی‌آید. ضرایب به دست آمده برای ویژگی‌ها به صورت زیر به دست آمد.

```
Test score: 0.013577399396172418
Train score: 0.12806630324382706
Mean Squared Error On Test :  0.00011240646475867021
Mean Squared Error On Train :  0.00011732165167691157

array([ 0.        ,  0.        , -0.        ,  0.        , -0.        ,
       -0.        , -0.        ,  0.00109083, -0.        ,  0.00282932])
```

طبق این نتایج لگ های اول و سوم در پیش‌بینی مدل موثرتر بوده اند و می‌توان از سایر لگ‌های زمانی صرف نظر کرد. این نتیجه موید همان نتیجه خودهمبستگی در سوال ۲ است.