به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس مبانی یادگیری آماری

استاد نیک‌آبادی

پروژه پایانی

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

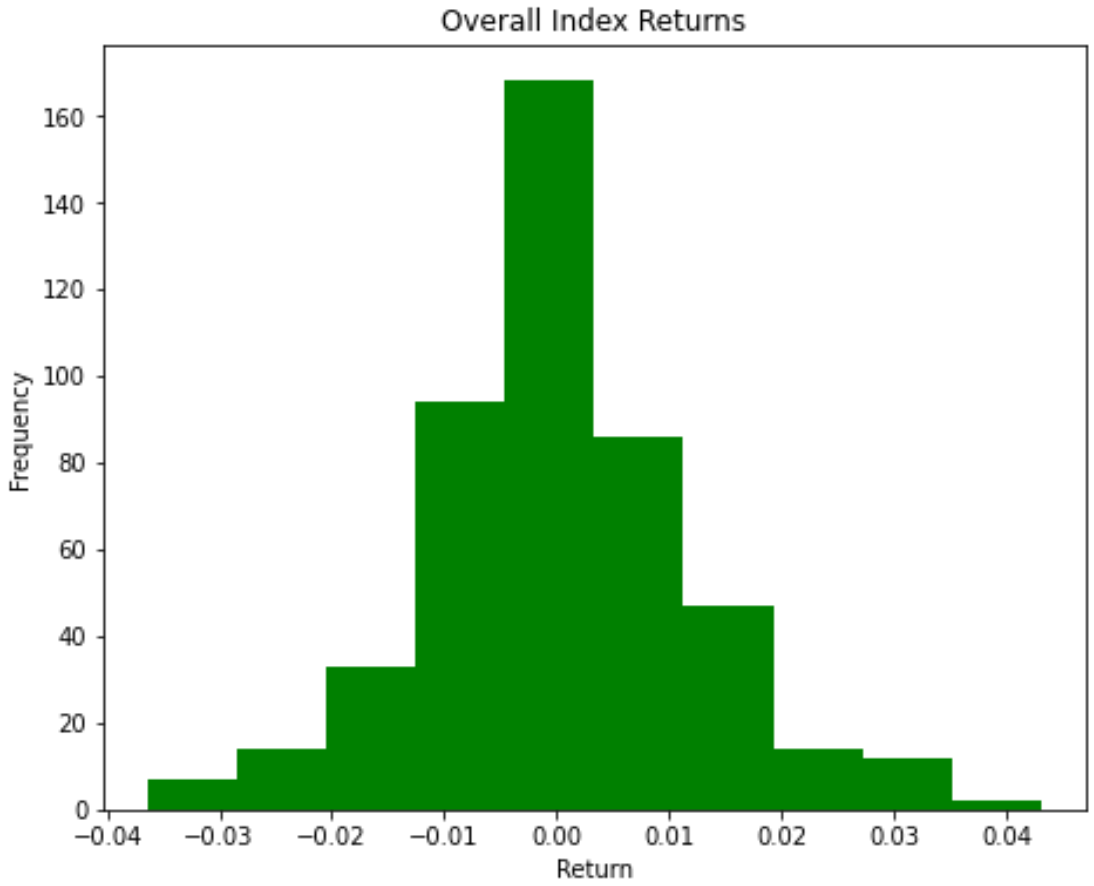
ملاحظات کلی

مطابق درخواست پروژه باید داده‌ها در یک بازه دوسالانه استفاده کرد. بازه مدنظر در این پروژه از ابتدای سال ۲۰۲۱ تا انتهای سال ۲۰۲۲ لحاظ شده است. اطلاعات مربوط به قیمت سهام‌ها به صورت مستقیم از خود سایت رسمی بورس و اطلاعات شاخص، دلار و طلا با استفاده از کتابخانه finpy-tse بدست آمده است.

سوال ۱: بررسی توزیع احتمالاتی مقادیر بازده

الف) برای بدست آوردن توزیع احتمالاتی بازده شاخص کل با روش پارامتری میانگین و واریانس بازده‌ها را محاسبه کردم. اگر بازده را به صورت درصد محاسبه کنیم به توزیع نرمال با میانگین 0.0265 و واریانس 1.439 می‌رسیم.

با روش غیرپارامتری هم به چیزی مشابه با نمودار زیر می‌توان رسید.



متناسب با چه توزیع پارامتری و چه غیر پارامتری می‌توان متوجه شد در بازه دو ساله در نظر گرفته شده میانگینا بورس متعادل بوده است. همچنین می‌توان دید که به ندرت شاخص بورس بازده ۵٪ یا -۵٪ داشته است که این نشان می‌دهد که در اکثر روزها تمام سهم‌ها همزمان در بیشترین حالت مثبت یا منفی قرار نداشتند و همواره برخی از سهام‌ها برخلاف روال بازار پیش رفته‌اند.

ب) ابتدا میانه نمونه‌ها را بررسی کردم که بسیار به صفر نزدیک بود ($۳ * ۱۰^{-۵}$). لذا اگر قرار باشد که توزیع متقارن باشد، این تقارن حول صفر رخ خواهد داد. برای بررسی متقارن بودن یا نبودن از آزمون Wilcoxon signed rank[1] استفاده می‌کنم. فرض null این هست که توزیع متقارن است و بررسی خواهیم کرد که آیا می‌توان این فرض را رد کرد یا خیر. چون تعداد داده‌ها ۴۷۷ تاست، فرض نرمال بودن برقرار است. مطابق محاسبات مقدار -۰.۲۶ برای z بدست آمده است. اگر مقدار ۵٪ را برای p-value در نظر بگیریم، این عدد با حد آستانه ۱.۶۵ فاصله زیادی دارد و لذا فرض پایه رد نمی‌شود. پس می‌توان گفت که توزیع تا حد قابل قبولی متقارن است.

چنین نتیجه‌ای همسو با نتایج قسمت الف نشان می‌دهد که دو سال درنظر گرفته شده بازار در شرایط تعادل قرار داشته است و بازدهی مثبت شدید یا منفی شدیدی را در مجموع بازه نداشته است.

ج) در جدول زیر میانگین و واریانس سهم‌های مدنظر آورده شده است.

	میانگین	واریانس
فولاد	-0.273	32.302
شستا	-2.925	3006.196
کخاک	-0.388	31.527
اطلس	0.078	2.273
خودرو	-0.052	7.212

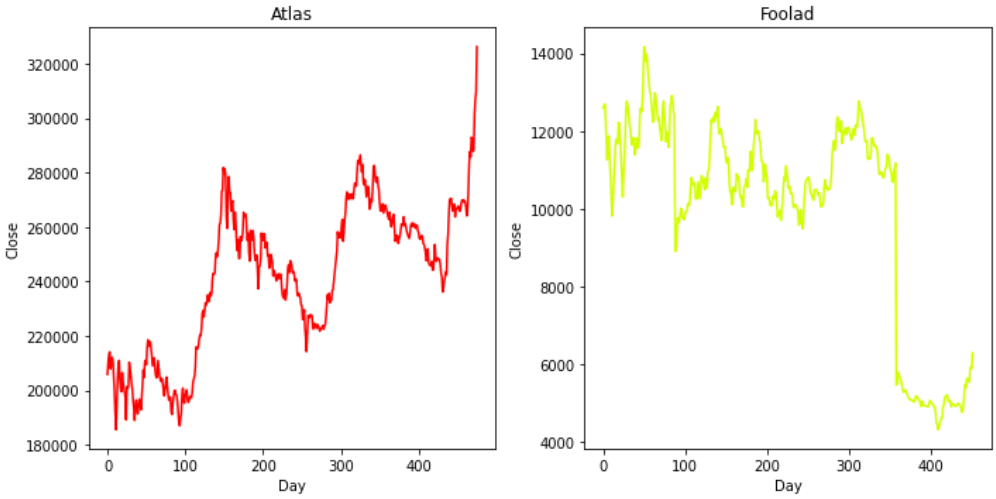
[1] https://mathcracker.com/wilcoxon-signed-ranks

همچنین تصمیم گرفتم که یک بار هم مقادیر بازدهی بیشتر از ۱۰٪ را حذف کنم و مجدد اعداد میانگین و واریانس را بدست آورم که جدول زیر بدست آمد:

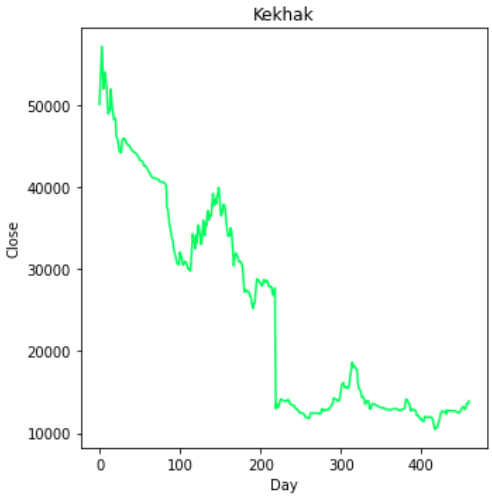
	میانگین	واریانس
فولاد	0.044	4.700
شستا	-0.018	4.942
کخاک	-0.143	3.865
اطلس	0.078	2.273
خودرو	-0.052	7.212

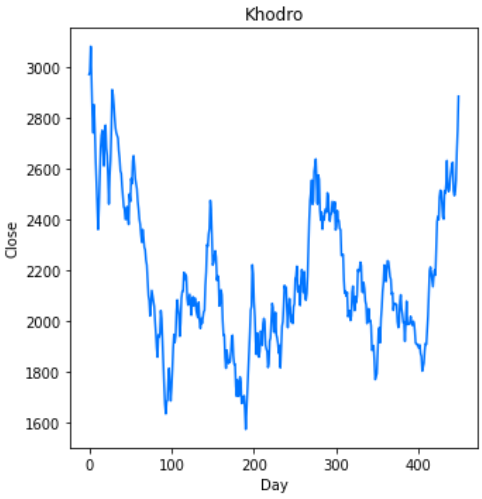
مقادیر بازدهی بیشتر از ۱۰٪ گاهی اوقات ناشی از افزایش سرمایه و مسائل نظیر آن است که اگر اعداد تعدیل‌شده می‌بود نباید دیده می‌شد ولی چون الان اعداد تعدیل‌شده نیستند ممکن است ما را گمراه کنند. در عین حال ممکن است برخی از سهم‌ها برخی روزها بدون دامنه نوسان باز شده باشند که در این حالت نباید آن مقدار بازدهی حتی اگر زیاد بود را حذف کرد. برای تحلیل در ادامه جدول دوم بیشتر ملاک خواهد بود چراکه اعداد منطقی‌تری دارد.

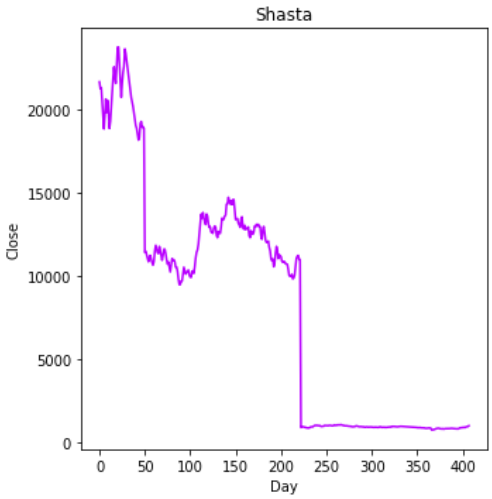
برای تفسیر اعداد دو جدول فوق نیاز است که نمودار تغییرات قیمت هر یک از سهام‌ها را داشته باشیم. همچنین برای تفسیر واریانس به نموداری نیاز داریم که تغییرات قیمت را بیشتر در خود داشته باشد. لذا نمودار هیستوگرام بعدی را ترسیم کردم که مقادیر بازدهی سهام‌های مختلف است. برای آنکه قابل مشاهده باشد تنها درصدهای بازدهی بین -۱۰٪ تا +۱۰٪ را نگه داشتم. طبیعتا هر چه مقادیر حول صفر بیشتر باشد (در شرایطی که میانگین نزدیک به صفر است)، واریانس کمتر خواهد بود.

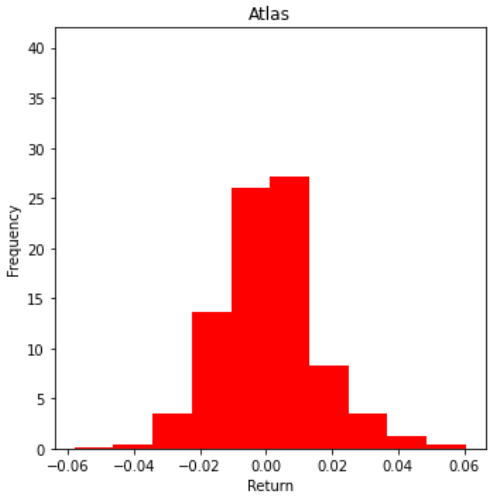

Atlas
Close
Day
320000
300000
280000
260000
240000
220000
200000
180000
0
100
200
300
400

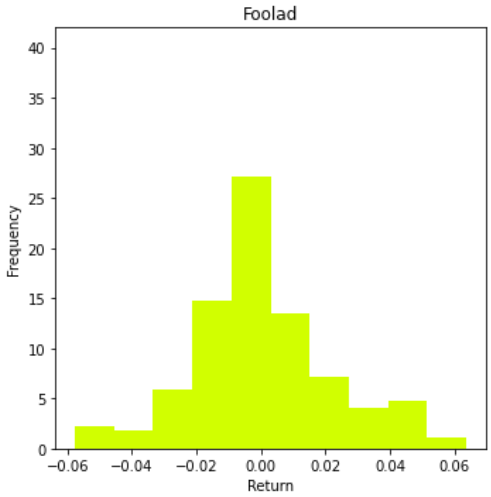
Foolad
Close
Day
14000
12000
10000
8000
6000
4000
0
100
200
300
400

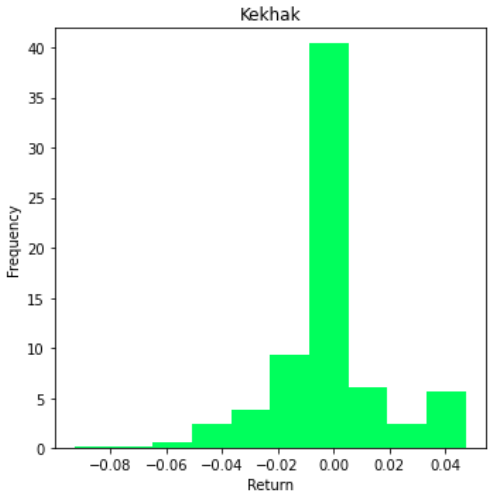

Kekhak
Close
Day
50000
40000
30000
20000
10000
0
100
200
300
400

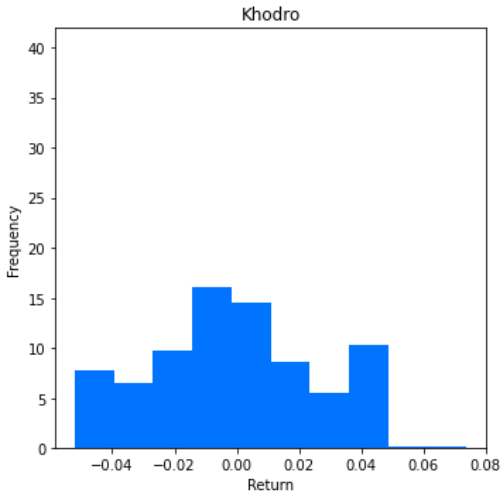

Khodro
Close
Day
3000
2800
2600
2400
2200
2000
1800
1600
0
100
200
300
400

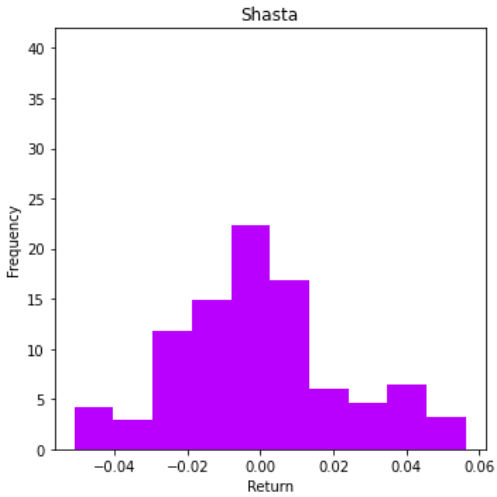

Shasta
Close
Day
20000
15000
10000
5000
0
0
50
100
150
200
250
300
350
400


Atlas
Frequency
Return


Foolad
Frequency
Return


Kekhak
Frequency
Return


Khodro
Frequency
Return


Shasta
Frequency
Return

همانطور که از جداول بر می‌آید صندوق اطلس با مدیریت مناسب سبد سهام توانسته است بیشترین میانگین بازدهی را داشته باشد. در حین حال و احتمالا با چینش سهام‌های متنوع واریانس تغییر قیمت خود را نسبت به تک سهم‌های بررسی شده کمینه کند که خود باعث کاهش ریسک و البته از دست دادن سودهای بسیار زیاد می‌شود.

سهم کخاک و شستا در این دو سال بازدهی مناسبی نداشته است که چنین چیزی هم از میانگین و هم از نمودار قابل ملاحظه است. در مورد سهم کخاک همچنین می‌بینیم که بازدهی صفر درصد را در روزهای زیادی تجربه کرده است ولی در مجموع روزهای منفی بر روزهای مثبت غلبه کرده است اما در مورد شستا چنین نیست و تغییرات بازده بیشتر بوده است.

سهم خودرو فراز و فرود قیمتی شدیدی داشته است. چنین امری از واریانس زیاد آن بر می‌آید. با دیدن نمودار تغییرات قیمت هم نوسان‌های زیاد قابل مشاهده است. به علاوه در نمدار هیستوگرام خودرو هم برخلاف سایر سهم‌ها حالت نرمال دیده نمی‌شود و بدین ترتیب بازدهی‌های با فاصله از صفر زیاد مشاهده شده است که همه این‌ها ریسک بالای این سهم را نشان می‌دهد.

در مورد سهم فولاد مورد خیلی خاصی را نمی‌توان گزارش کرد و تقریبا متعادل بوده است و نمودار هیستوگرام آن هم شبیه به شاخص کل است.

د) به ترتیب برای شاخص کل، اطلس، فولاد، کخاک، خودرو و شستا نموداری شامل تغییرات میانگین و واریانس آورده شده است.

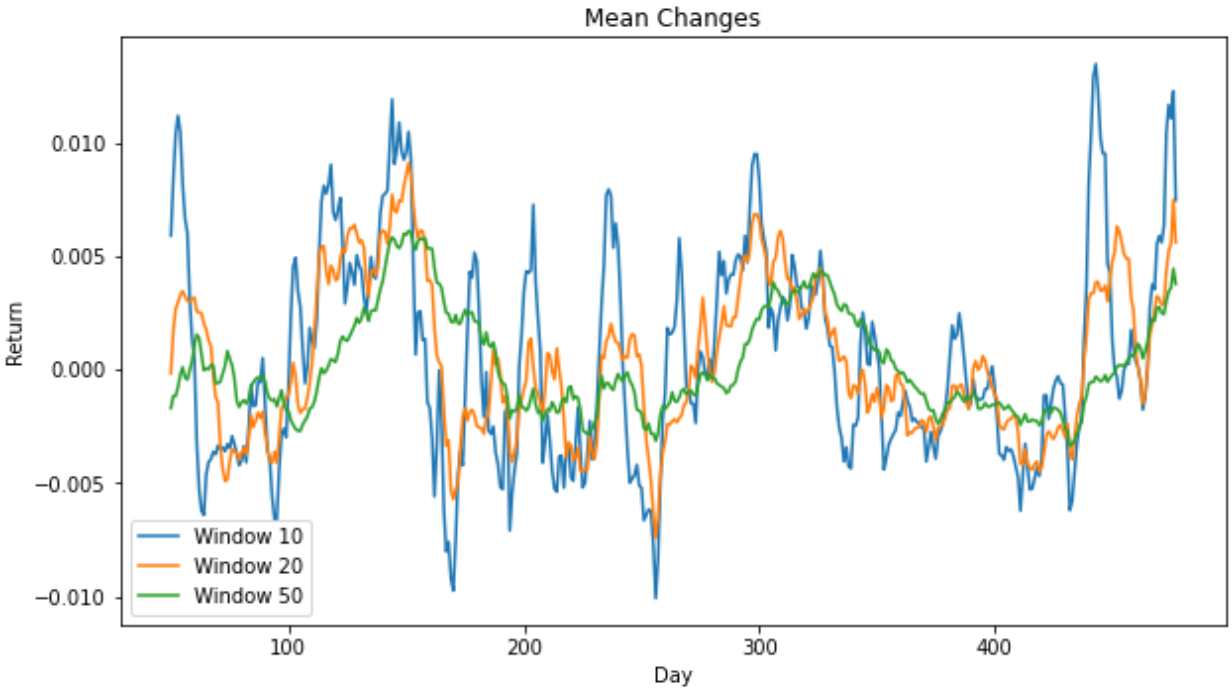
با بررسی نمودارها می‌توان موارد زیر را دریافت:

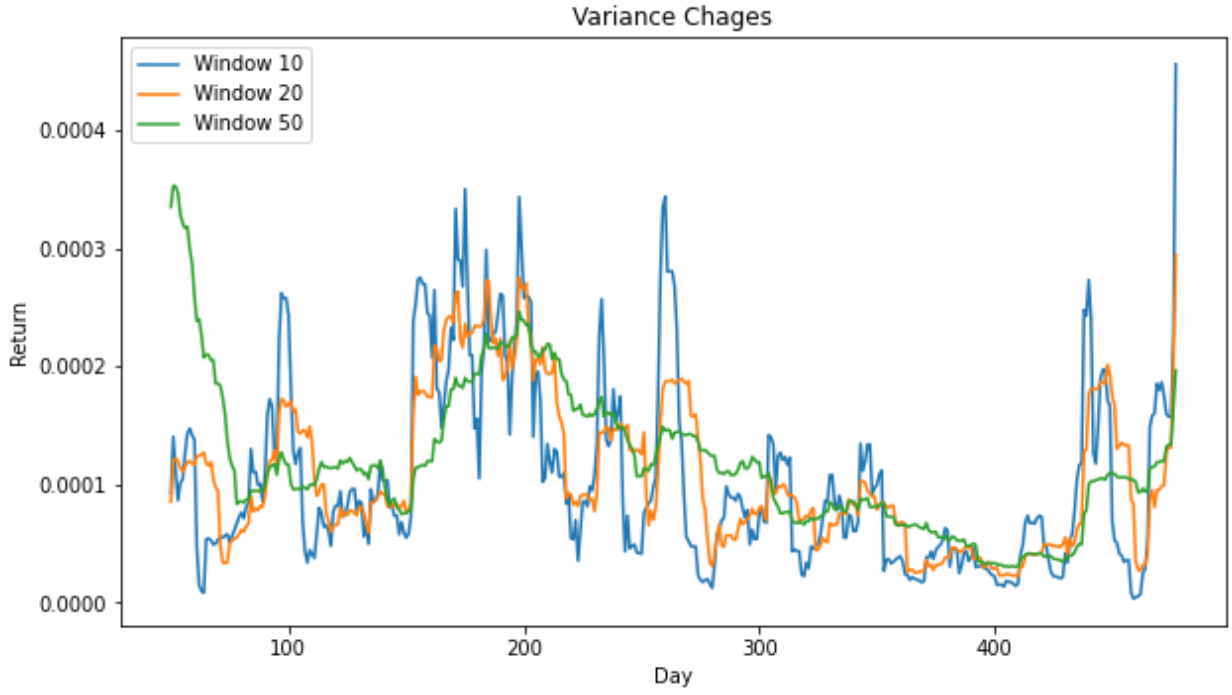
- مطابق انتظار تغییرات در پنجره ده روزه نسبت به پنجره بیست روزه و تغییرات در پنجره بیست روزه نسبت به پنجره پنجاه روزه شدیدتر است. همچنین تغییرات شدید اثرات طولانی مدت‌تر بر روی پنجره‌های بلند مدت‌تر دارند. چراکه هر چقدر که بازه طولانی‌تر باشد نوسانات جزئی قیمتی کم تاثیرتر می‌شود و صرفا روال کلی

حرکت قیمتی باقی می‌ماند و در عین حال تغییرات مهم در پنجره‌های بلند مدت‌تر ثبت می‌شود.

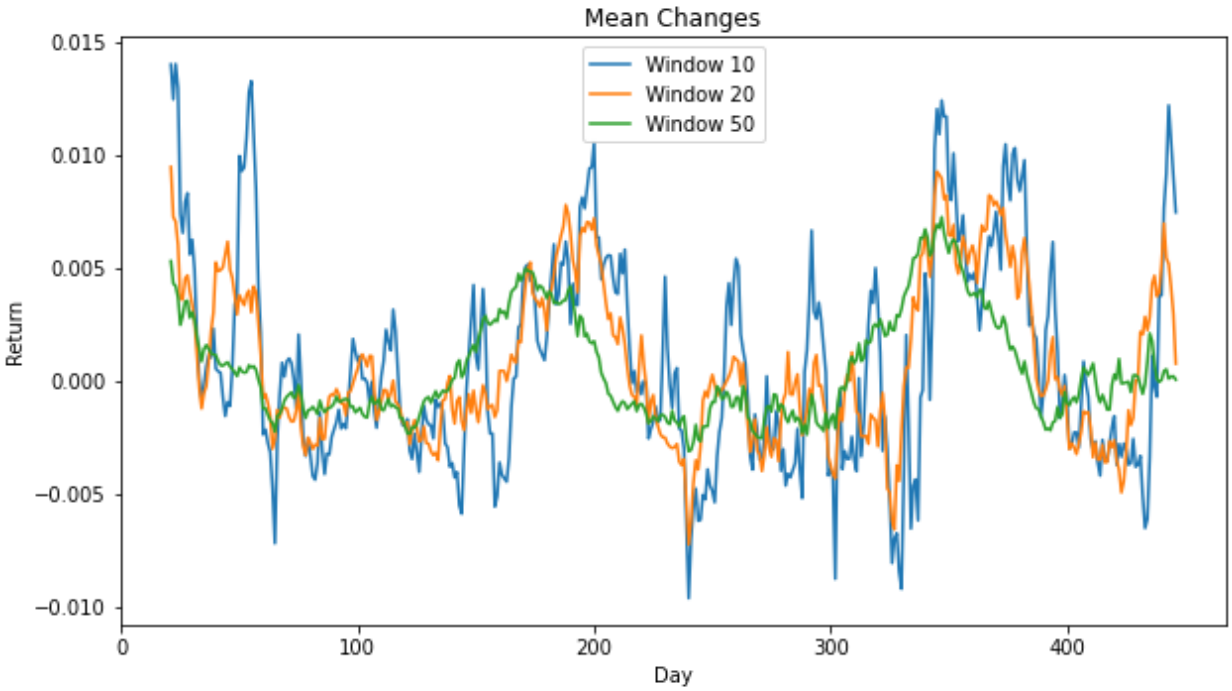
- در نمودار شاخص کل و برخی از سهم‌ها هر کجا واریانس در پنجره ده روزه به قله خود رسیده است، روند میانگین تغییر پیدا کرده است. یعنی یا روند میانگین از قله مثبت خود به سمت اعداد منفی حرکت کرده است یا بالعکس. چنین چیزی معقول است چراکه وقتی واریانس به قله برسد، یعنی هم اعداد بازدهی مثبت زیاد و هم بازدهی منفی زیاد در پنجره وجود داشته است که در نتیجه نشان‌گر تغییر روند است.
- به دلیل برخی از تغییرات شدید قیمتی که پیش‌تر در مورد آن صحبت کردیم، قسمت واریانس برخی از سهم‌ها دچار مشکل شده است. چنین چیزی در مورد میانگین به این شدت رخ نداده است چراکه واریانس نسبت به داده‌های نویز حساس‌تر است.

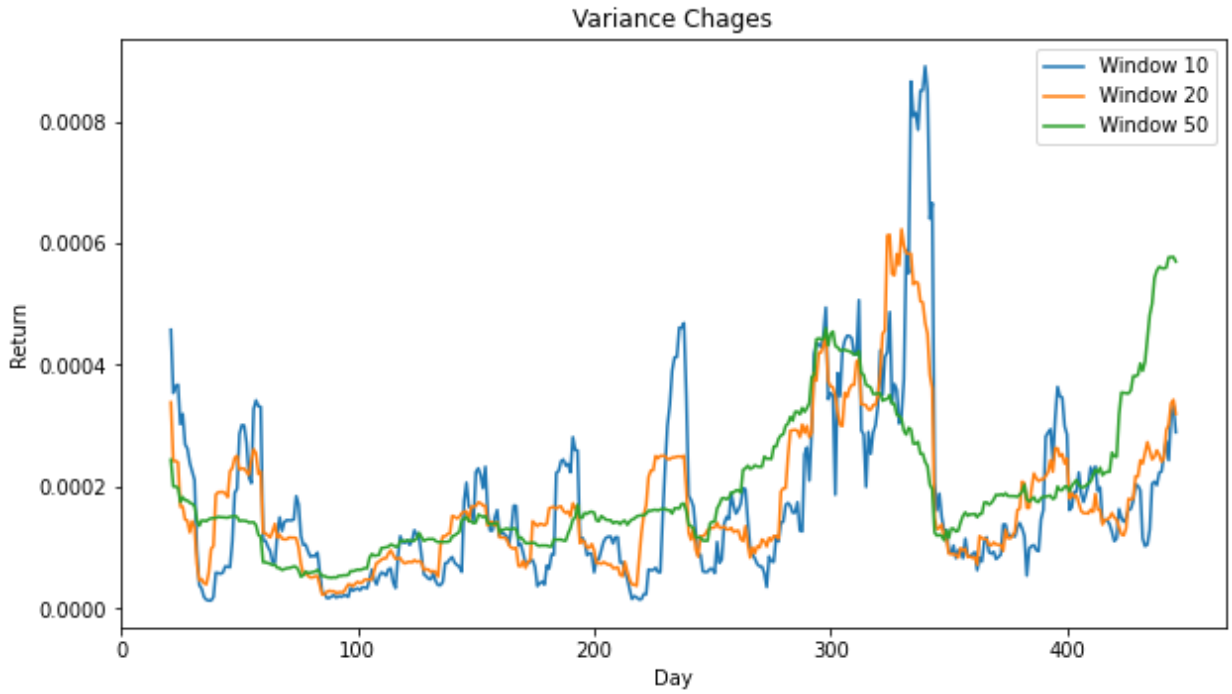
Rolling Window On Overall Index



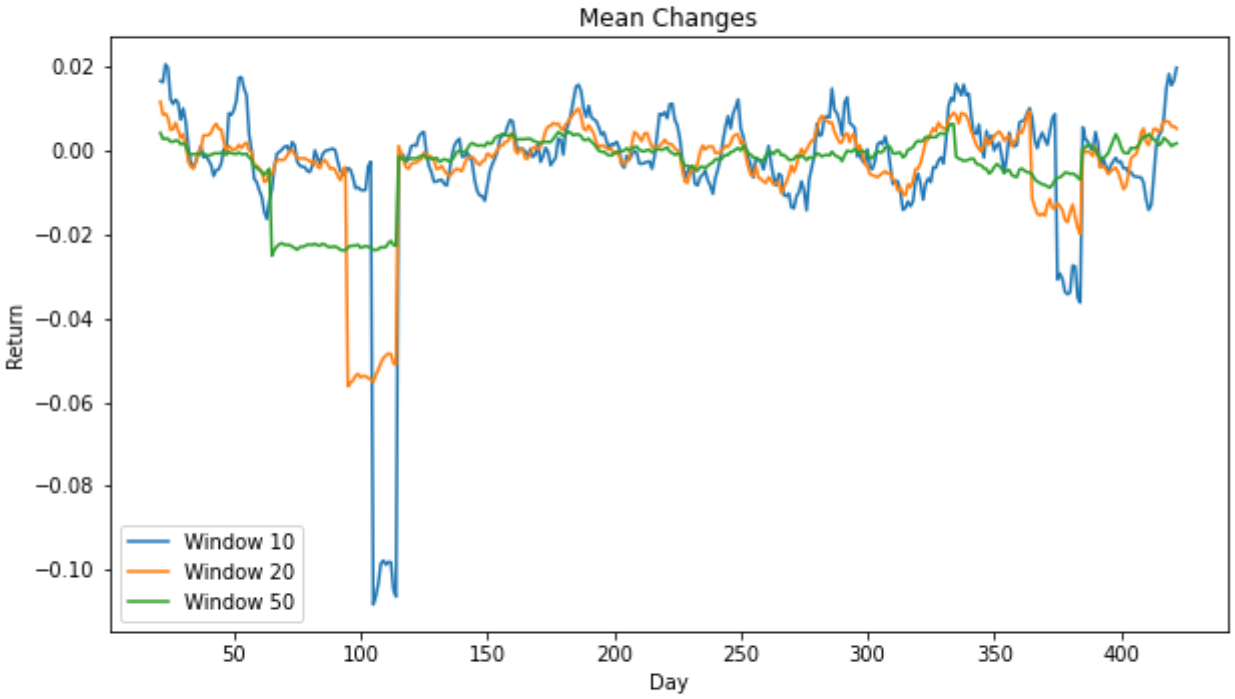


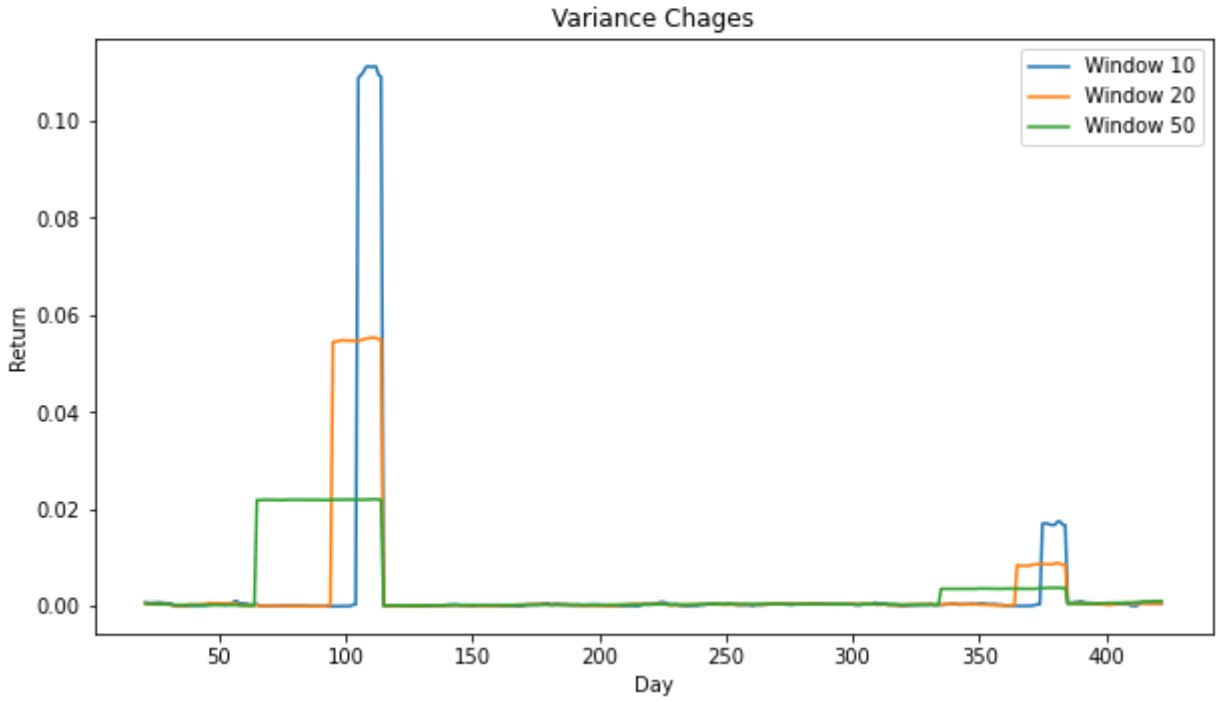
Rolling Window On Atlas



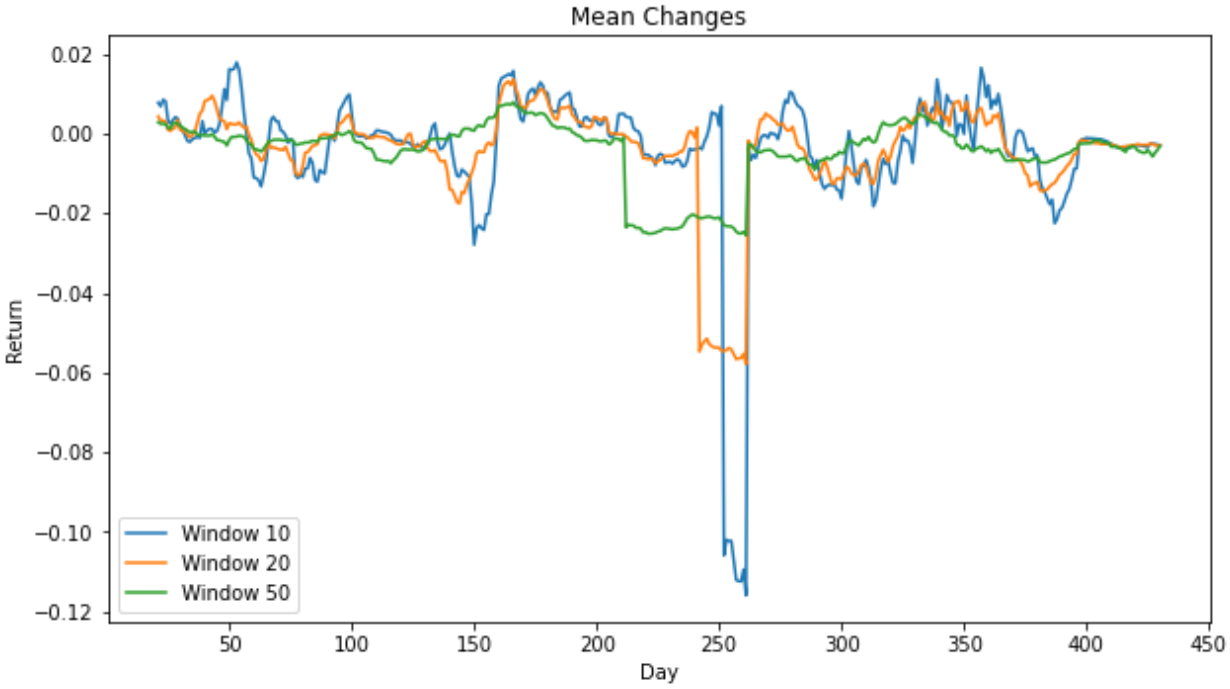


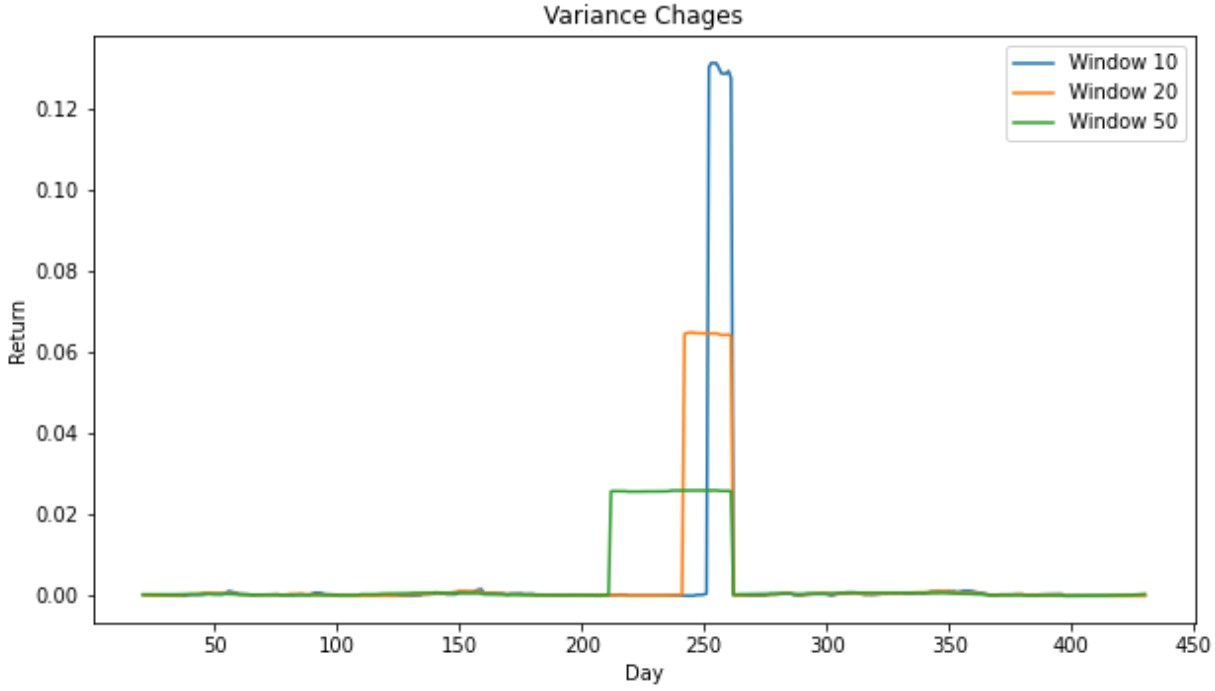
Rolling Window On Foolad



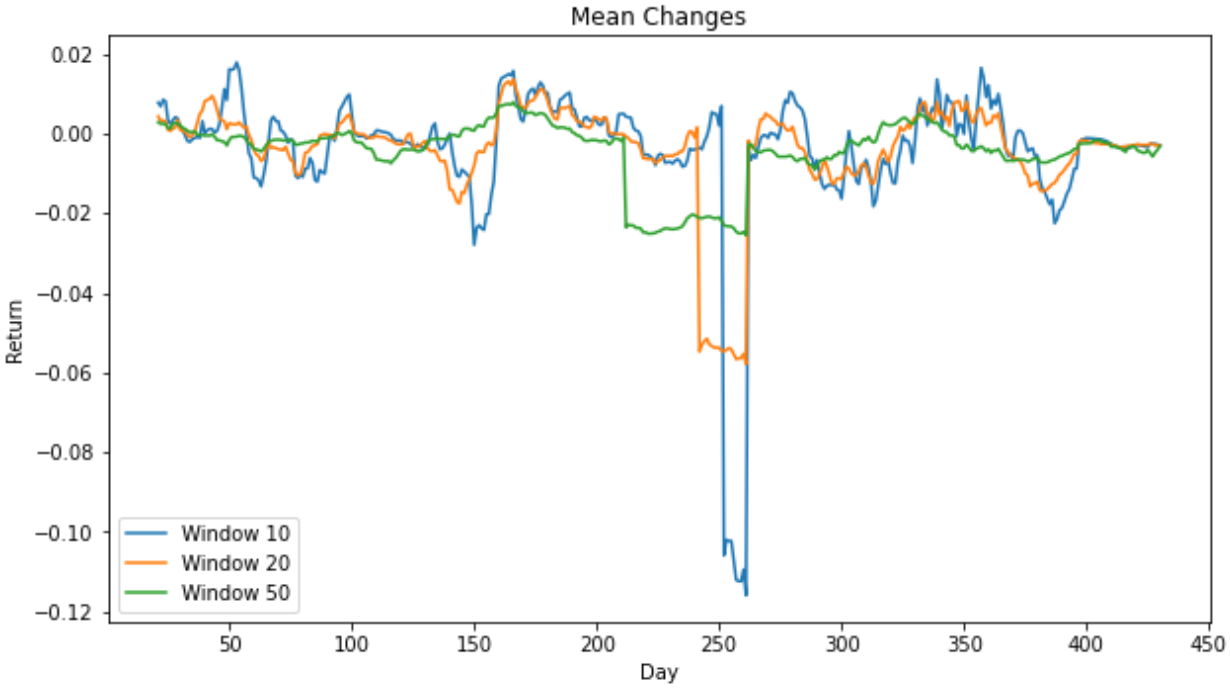


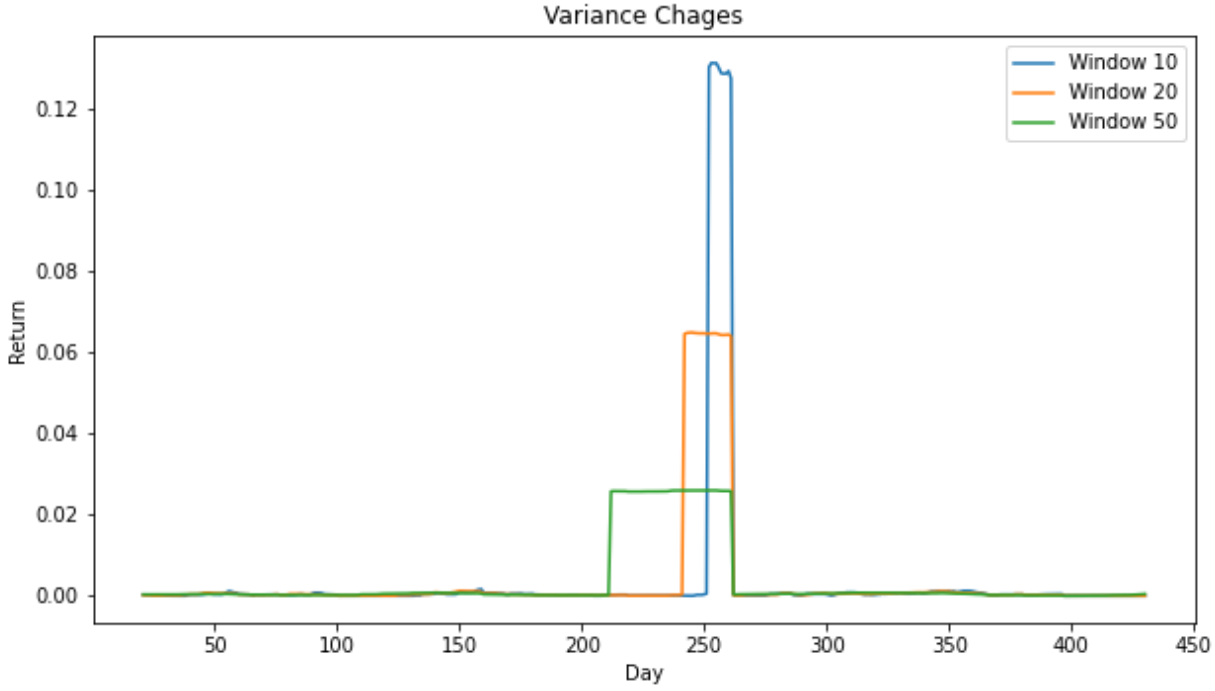
Rolling Window On Kekhak



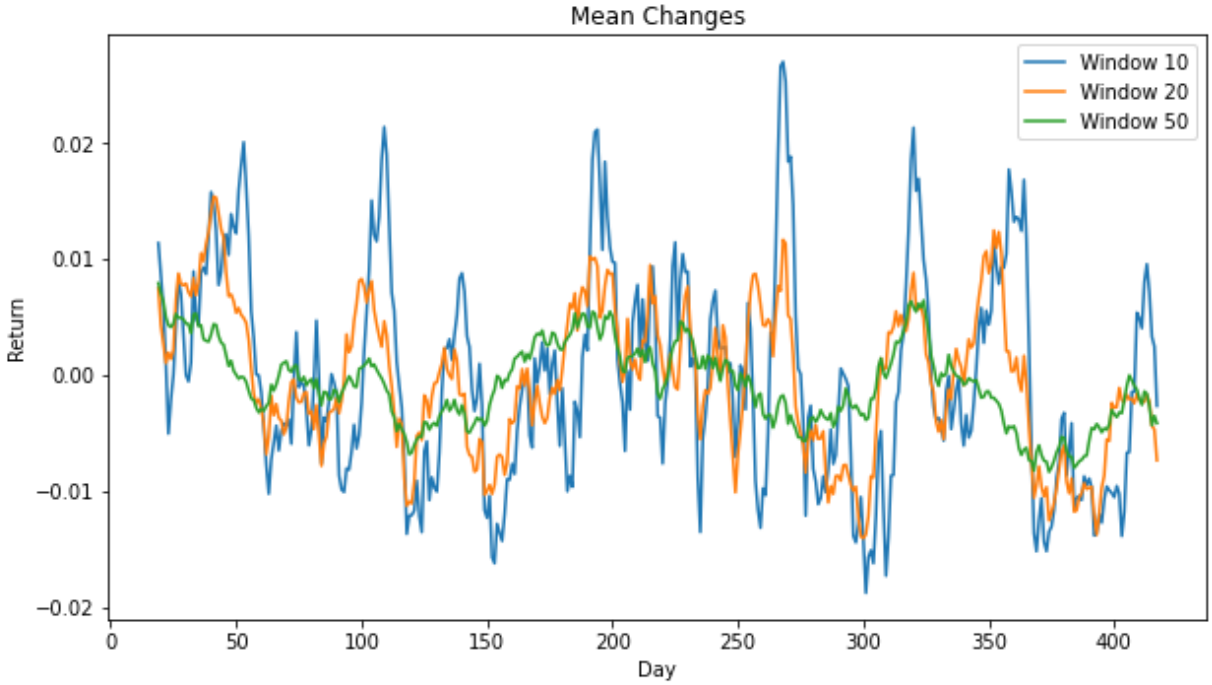


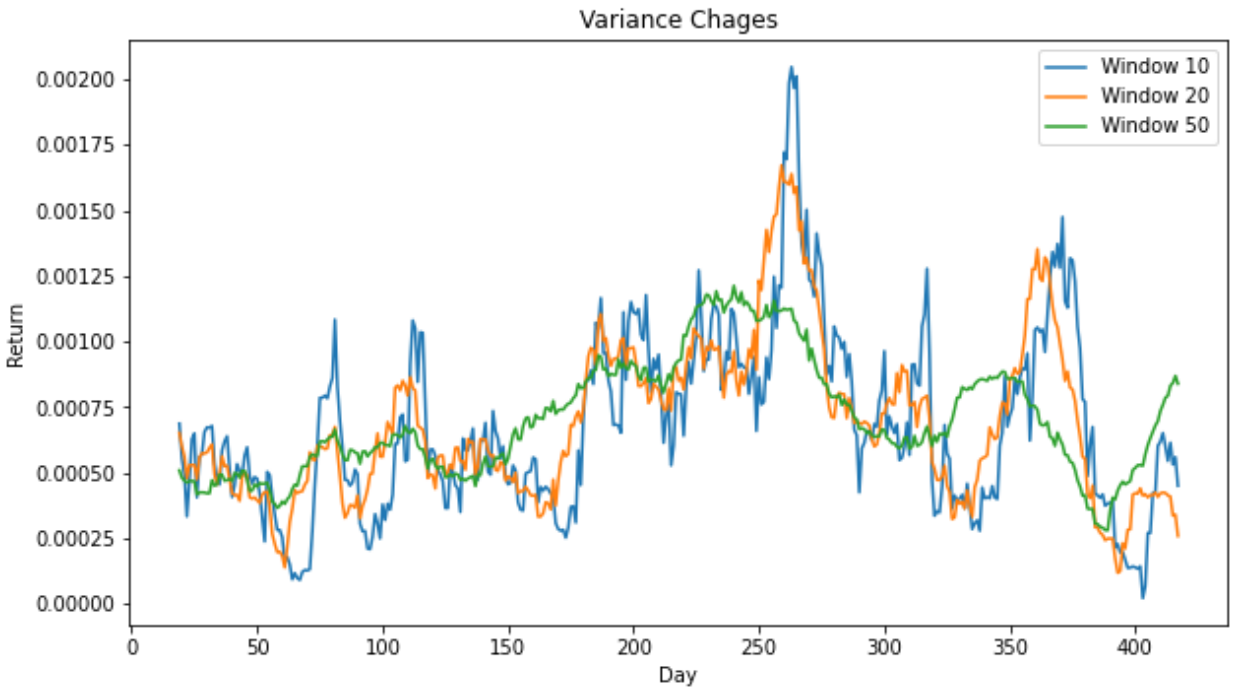
Rolling Window On Kekhak



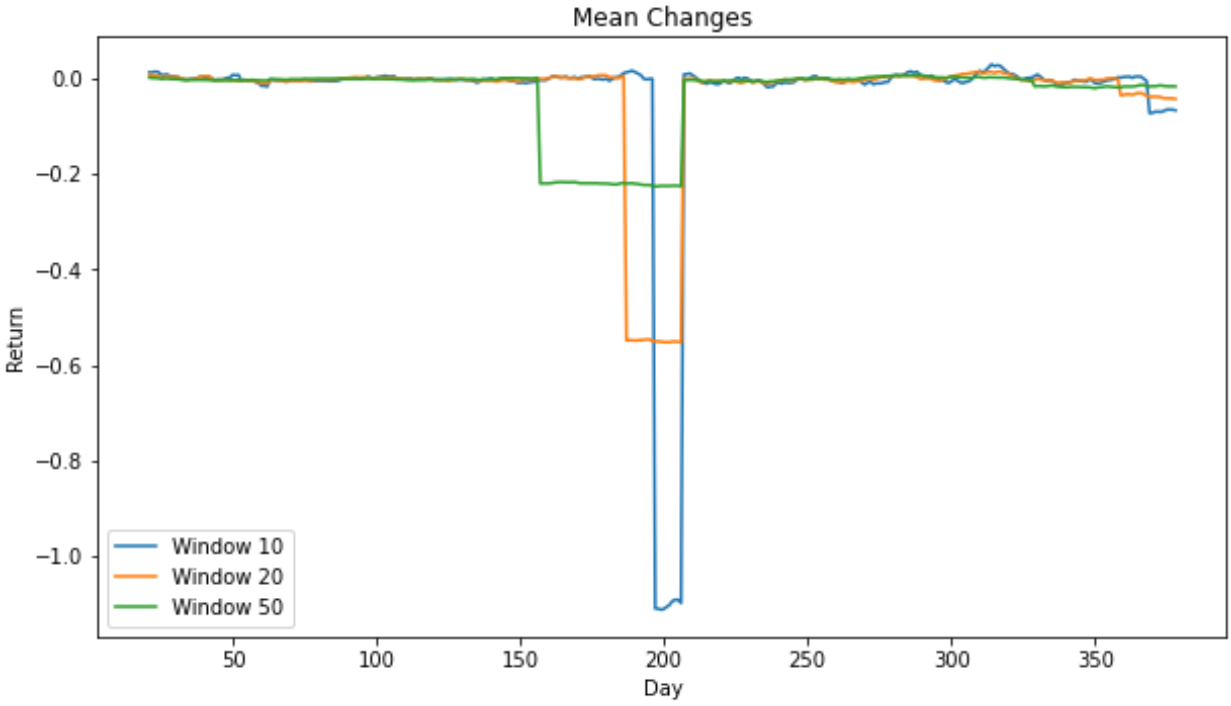


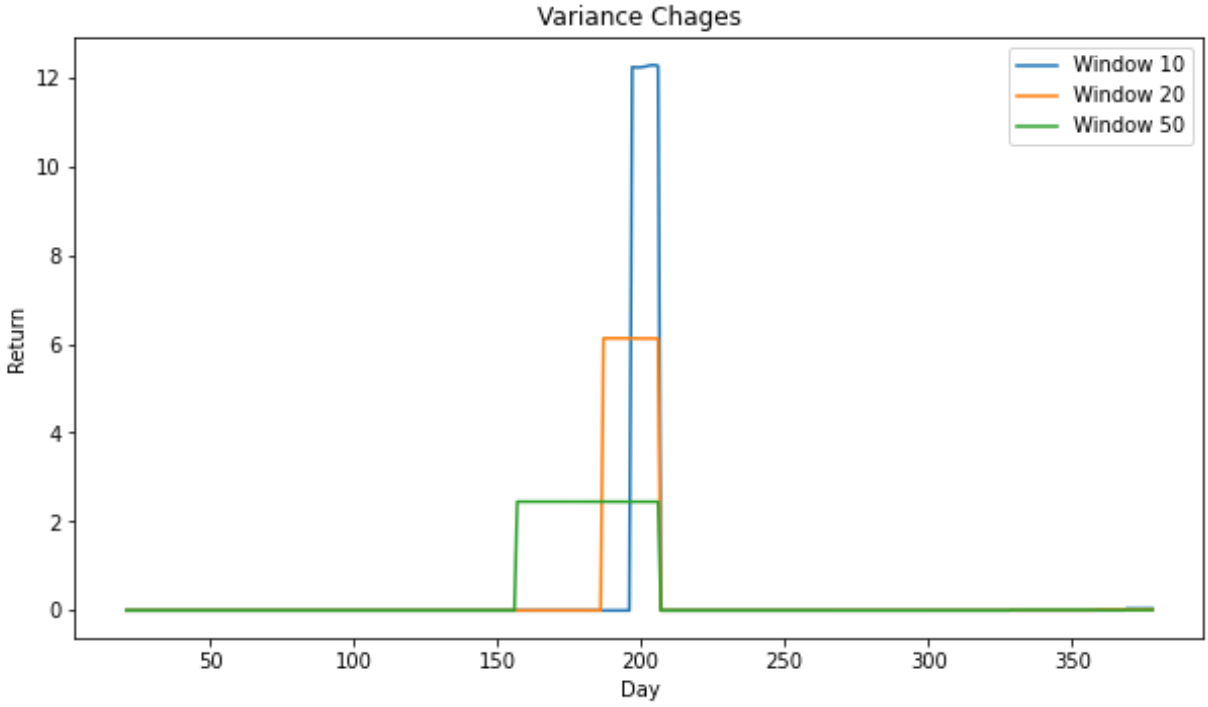
Rolling Window On Khodro





Rolling Window On Shasta





سوال ۲: بررسی همبستگی‌ها

الف) در جدول زیر مقادیر خودهمبستگی شاخص برای مقدار لگ ۱ تا ۵ روز آورده شده است:

لگ	خودهمبستگی
۱	۰.۳۴۲
۲	۰.۰۲۶
۳	۰.۰۶۸
۴	-۰.۰۰۸
۵	-۰.۰۲۹

همانطور که مشخص است خودهمبستگی شاخص با یک روز تاخیر مقدار قابل توجهی دارد و نشان می‌دهد مقدار بازدهی یک روز می‌تواند تاثیر مهمی روی روز بعد بگذارد ولی این تاثیر برای روزهای بعد بسیار کاهش پیدا می‌کند. نکته جالب‌تر این جا است که خودهمبستگی شاخص با لگ ۵ عددی منفی است؛ احتمالا نشان می‌دهد که مقدار بازدهی نوسان زیادی دارد و اگر یک روز بازدهی مثبتی وجود داشت، در چند روز آینده تقریبا جبران می‌شود!

ب) همبستگی شاخص کل با هر یک از سهام‌های مذکور در جدول زیر آورده شده است:

سهام	همبستگی
اطلس	۰.۸۵۹
فولاد	۰.۳۰۸
کخاک	۰.۱۴۴
خودرو	۰.۶۶۷
شستا	۰.۰۳۵

همانطور که مشخص است سهم اطلاس به عنوان یک صندوق و احتمالا با داشتن ترکیبی از سهم‌های عمدتا شاخص‌ساز توانسته است همبستگی بسیار بالایی با شاخص

کل داشته باشد. پس از آن سهم خودرو همبستگی بالایی داشته است. احتمالا به دلیل لیدر بودن این سهم و تاثیر روانی آن بر کلیات بورس است. از آن طرف سهام‌های شستا و کخاک تقریبا همبستگی خاصی با شاخص کل ندارند.

ج) قیمت دلار از طریق کتابخانه finpy-tse قابل دریافت بود ولی قیمت طلا خیر. در سایت رسمی شرکت بورس هم قیمت طلا به صورت چندساله موجود نبود. با توجه به همبستگی قیمت شدید صندوق‌های مبتنی بر طلای موجود در بورس و قیمت طلا تصمیم گرفتم قیمت یکی از صندوق‌های مبتنی بر طلا را به عنوان قیمت طلا استفاده کنم. به بیان دقیق‌تر قیمت صندوق «طلای لوتوس» را مدنظر گرفتم.

همبستگی شاخص کل با دلار و طلا در جدول زیر آورده شده است:

بازار	همبستگی
دلار	۰.۱۸۷
طلا	۰.۲۷۰

همانطور که مشخص است شاخص کل بورس هم دلار و هم با طلا همبستگی مثبت دارد. طلا نسبت به دلار همبستگی بیشتر با شاخص کل دارد. با وجود آنکه دلار و طلا دو بازار رقیب بورس به حساب می‌آیند اما به دلیل تورم بسیار بالا و کاهش ارزش ریال، تمام قیمت‌ها از جمله قیمت طلا، دلار و سهام به طور کلی افزایش پیدا می‌کند و بالعکس.

سوال ۳: پیش‌بینی مقدار بازده

مطابق با روال معمول ۸۰٪ داده‌ها برای آموزش و ۲۰٪ برای آزمون مورد استفاده است.

الف) با توجه به اطلاعات میزان خودهمبستگی به نظر می‌رسد که از روی بازدهی روز قبل بیشترین حدس را می‌توان در مورد بازدهی روز فعلی داشت ولی با این حال ممکن است داشتن اطلاعات سایر روزها هم کمک به بهبود جزئی خروجی مدل داشته باشد. ابتدا فقط یک مدل رگرسیون خطی ساده که تنها از بازده روز قبل استفاده می‌کرد را آموزش دادم. برای بدست آوردن مقدار بتا از معادله نرمال (Normal Equation) استفاده کردم.

فرمول مدل مطابق با زیر با خطای MSE برابر با $1.21 * 10^{-4}$ بدست آمد:

$$R = 0.364 * R_1 + 6.04 * 10^{-5}$$

در این رابطه R مقدار بازده امروز و R_1 بازده روز قبل است. همچنین اگر میانگین بازده‌های دیده شده در زمان آموزش را به عنوان پیش‌بینی ارائه می‌دادیم به خطای MSE برابر با $1.26 * 10^{-4}$ می‌رسیدیم که نشان می‌دهد مدل ما از مدل بدیهی تنها کمی بهتر است.

یک بار هم همین مدل را با استفاده از بازده پنج روز قبل آموزش دادم. در این حالت خطا به $1.28 * 10^{-4}$ افزایش یافت که نشان می‌دهد همان مدل قبلی بهتر است. نهایتا مدل جدید به این شکل قابل بیان است:

$$R = 0.465 * R_1 - 0.252 * R_2 + 0.185 * R_3 - 0.109 * R_4 + 0.002R_5 + 2.141 * 10^{-4}$$

ب) برای پیشبینی روند بازدهی با توجه به نتایج قسمت قبل تنها از مقدار بازدهی روز قبل استفاده می‌شود. می‌توان برای هر دو کلاس دو مدل نرمال در نظر گرفت. سپس تعلق هر داده را به هر یک از دو کلاس سنجید؛ نهایتا برای پیش‌بینی بهتر می‌توان فراوانی هر یک از دو کلاس را هم در نظر گرفت که بدین ترتیب عملا یک دسته‌بند بیز خواهیم داشت.

توزیع نرمالی که بر روی داده‌های مثبت آموزش پیدا کرده است دارای میانگین ۰.۰۰۳ و واریانس ۰.۰۰۰۱ و توزیعی نرمالی که بر روی داده‌های منفی آموزش پیدا کرده است دارای میانگین ۰.۰۰۳- و ورایانس ۰.۰۰۰۱ است. نهایتا توجه کنید که تعداد داده‌های مثبت در بازه آموزش برابر با ۱۹۳ و تعداد داده‌های منفی برابر با ۱۸۷ بوده است.

مقدار Accuracy پس از آموزش مدل برابر با ۵۸.۳۳٪ بدست آمد. این صحت چندان زیاد نیست ولی از حالت پایه بیشتر است؛ چنانچه می‌خواستیم بیشترین کلاس را به عنوان پیش‌بینی برای تمام داده‌ها ذکر کنیم به عدد ۵۳.۱۲٪ می‌رسیدیم.

حال قصد داریم صحت پیش‌بینی مدل با استفاده از بازدهی چند روز قبل را بررسی کنیم. در جدول زیر صحت مدل برای حالات مختلف خواسته شده آورده شده است.

چند روز بعد	صحت
۱	۵۸.۳۳٪
۲	۴۹.۴۷٪
۳	۶۳.۱۵٪
۴	۵۶.۸۶٪

به طور کلی می‌توان دید که پیش‌بینی برای دو یا چهار روز بعد چندان موفقیت آمیز نیست. تنها می‌توان دید که پیش‌بینی برای ۳ روز بعد حتی از ۱ روز بعد هم بهتر بوده است. چنین چیزی مطابق انتظار من نبوده و احتمالا به صورت تصادفی بدست آمده است.

ج) برای این قسمت از سه سهم با همبستگی بالا با شاخص یعنی اطلس، خودرو و فولاد و همچنین صندوق طلای لوتوس استفاده می‌کنم. توجه کنید که برای این قسمت مجموعه‌داده کمی متفاوت از قسمت قبل‌های قبل است چراکه باید تمام ویژگی‌ها موجود باشد و داده‌های آموزش و تست برای این هدف کمتر خواهند شد.

برای پیشبینی بازدهی با مدل رگرسیون و با استفاده از بازدهی شاخص روز قبل و چهار سهم و صندوق یادشده خطای MSE در این حالت به عدد $9.84 * 10^{-5}$ می‌رسد که به

اندازه قابل توجهی از مدل مبتنی بر تنها شاخص کل بهتر است. رابطه مدل برازش‌شده عبارت است از:

$$R = 2.78 * 10^{-1} * R_{overall} + 1.10 * 10^{-1} * R_{Gold} - 1.06 * 10^{-2} * R_{Khodro} + 3.45 * 10^{-3} * R_{Foolad} + 3.67 * 10^{-2} * R_{Atlas} + 1.06 * 10^{-4}$$

برای دسته‌بندی آماری روال کلی مانند قبل است. نکته حالب این جالست که اگر برای مجموعه‌داده فعلی تنها از بازدهی شاخص کل روز قبل استفاده کنیم میزان Accuracy به عدد ۶۹.۴۱٪ می‌رسد! احتمالا علت تفاوت ده درصدی بین این صحت با صحت قسمت ب تفاوت مجموعه‌داده باشد. همچنین اگر از اطلاعات سایر سهام‌ها استفاده کنیم میزان Accuracy به عدد ۴۹.۴۱٪ کاهش پیدا می‌کند. احتمالا این کاهش جدی به دلیل تفاوت رفتار جدی سهم‌های در نظر گرفته شده در بازه مدنظر باشد. مثلا مشاهده شد که با حذف بازدهی فولاد صحت به عدد ۶۲.۳۵٪ رسیده است. در نمودارهای سوال ۱ هم دیدیم که سهم فولاد در اواخر بازه دوساله روال صعودی‌ای و متفاوتی از سایر قسمت‌های بازه‌اش داشته است.

د) می‌دانیم که Lasso عملا از یک منظم‌ساز استفاده کرده است. هر چقدر ابرپارامتر متناسب با آن را بیشتر کنیم تاثیرش بیشتر می‌شود و بالعکس. مطابق با نتایج زمانی که مقدار ابرپارامتر برابر با صفر در نظر گرفته شده است (رگرسیون عادی) خطای MSE برابر است با $1.24 * 10^{-4}$ با افزایش مقدار ابرپارامتر مشاهده می‌شود که مقدار خطای MSE حتی بیشتر هم می‌شود و به عدد $1.27 * 10^{-4}$ می‌رسد. بنابراین مدل رگرسیون Lasso برای این مساله و این ویژگی‌ها کارآمد نیست.