Семинарска работа по „Вовед во Наука на Податоците“

Тема: Пронаоѓање разни типови корелации меѓу податоци
Код: `github.com/DimeJovanovski/vnp-seminarska`
Видео одбрана: `youtu.be/xo5_lHQYrPA`

Автор: Димитар Јовановски, 203099

01.09.2023

Вовед

Во оваа семинарска работа ќе се врши анализа врз пет различни податочни множества(временски серии) извлечени од Yahoo Finance кои во себе имаат податоци за изминатите 4 години.

- Bitcoin (BTC) `finance.yahoo.com/quote/BTC-USD/`
- Grayscale Bitcoin Trust (GBTC) `finance.yahoo.com/quote/GBTC/`
- Riot Blockchain Inc. (RIOT) `finance.yahoo.com/quote/RIOT/`
- Square Inc. (SQ) `finance.yahoo.com/quote/SQ/`
- Tesla Inc. (TSLA) `finance.yahoo.com/quote/TSLA/`

Крајната цел е да се види меѓу кои од овие податоци и колку има корелација. Ќе се испитува Pearson, Spearman корелации и корелација со зададен window.

Пристап на работа

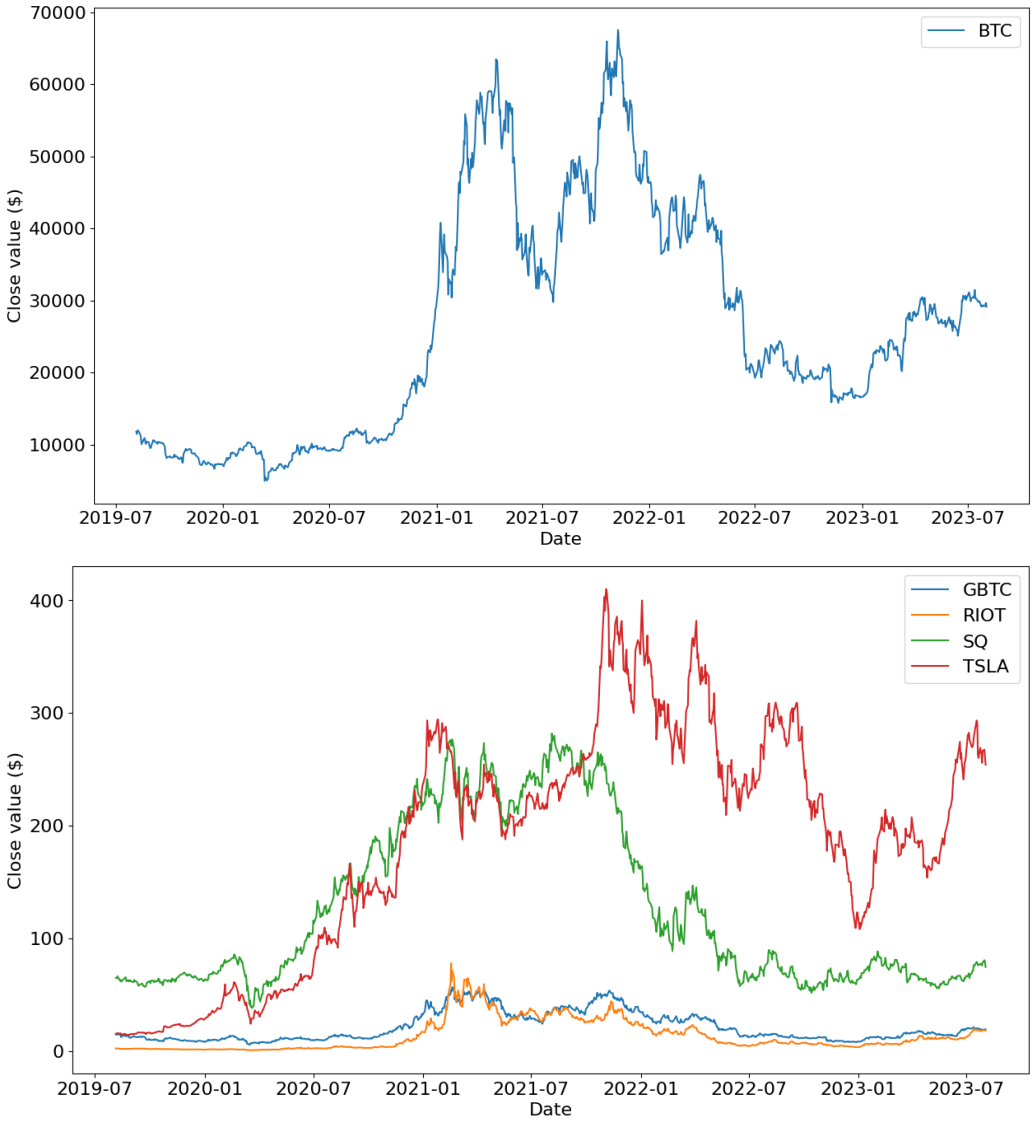
Податоците се во .csv формат и за работа со нив ќе се користи Pandas библиотеката. Сите податочни множества се состојат од вкупно 7 колони, а тие се: Date, Open, High, Low, Close, Adj Close и Volume. За нашите анализи од сите овие потребни ни се само Date и Close. За полесна понатамошна работа, ќе ги споиме податоците на сите овие податочни множества во едно т.ш. ќе се реименуваат Close колоните на секое податочно множество посебно во поопсни имиња(за да се избегне преклопување на имиња) и ќе се спојат сите според датумите во Date колоната.

Date	Close_BTC	Close_GBTC	Close_RIOT	Close_SQ	Close_TSLA
2019-08-05	11805.65332	14.6	2.24	64.849998	15.221333
2019-08-06	11478.168945	14.77	2.08	64.599998	15.383333
2019-08-07	11941.96875	14.97	2.12	65.0	15.561333

Табела 1: Изглед на финално податочно множество при спојување на сите податоци

Пронаоѓање на Pearson корелации

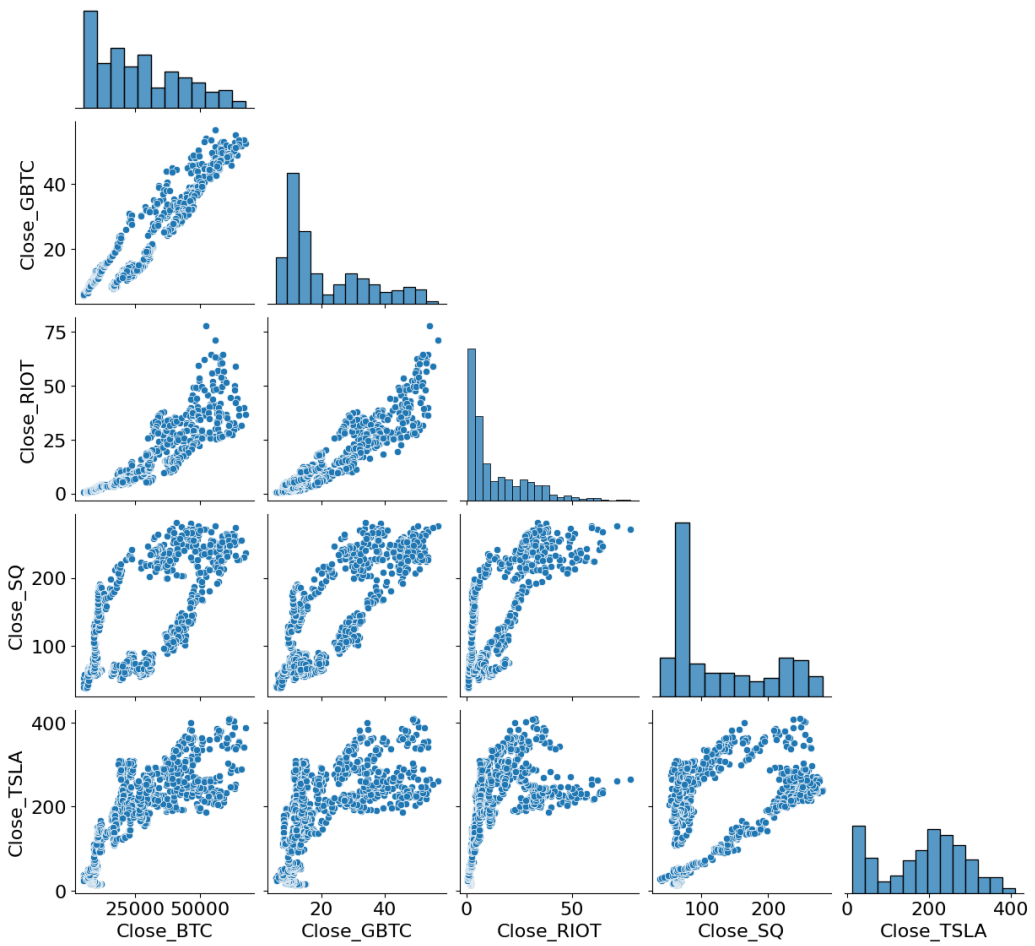
Пред да се пресметува корелацијата меѓу атрибутите на податочните множества, ќе се направи проста визуелизација на истите податоци за да се види нивна иницијална корелација и состојба.



Слика 1: Споредба на Closing вредностите меѓу Биткоин и останатите

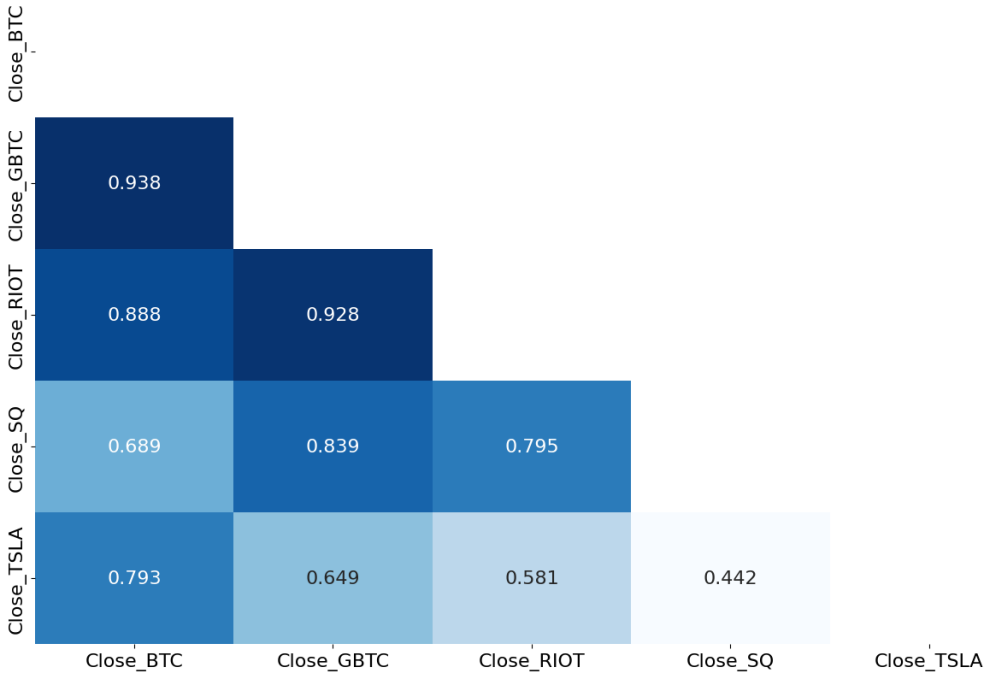
Уште од првичниот приказ со голо око може да се види корелација меѓу овие вредности иако сеуште не е јасно колку е таа силна. По прецизен начин да се претстави коорелацијата меѓу сите овие атрибути во една

визуелизација е со расејувачка матрица со помош на Seaborn python библиотеката со функцијата pairplot(). Сега појасно може да се види кооорелацијата. Grayscale Bitcoin Trust (GBTC) има особено силна кооорелација со Bitcoin (BTC) и Riot Blockchain Inc. (RIOT), што има смисла затоа што тие се здружение кои имаат големи инвестиции во Биткоин и се што е поврзано со криптовалутата.



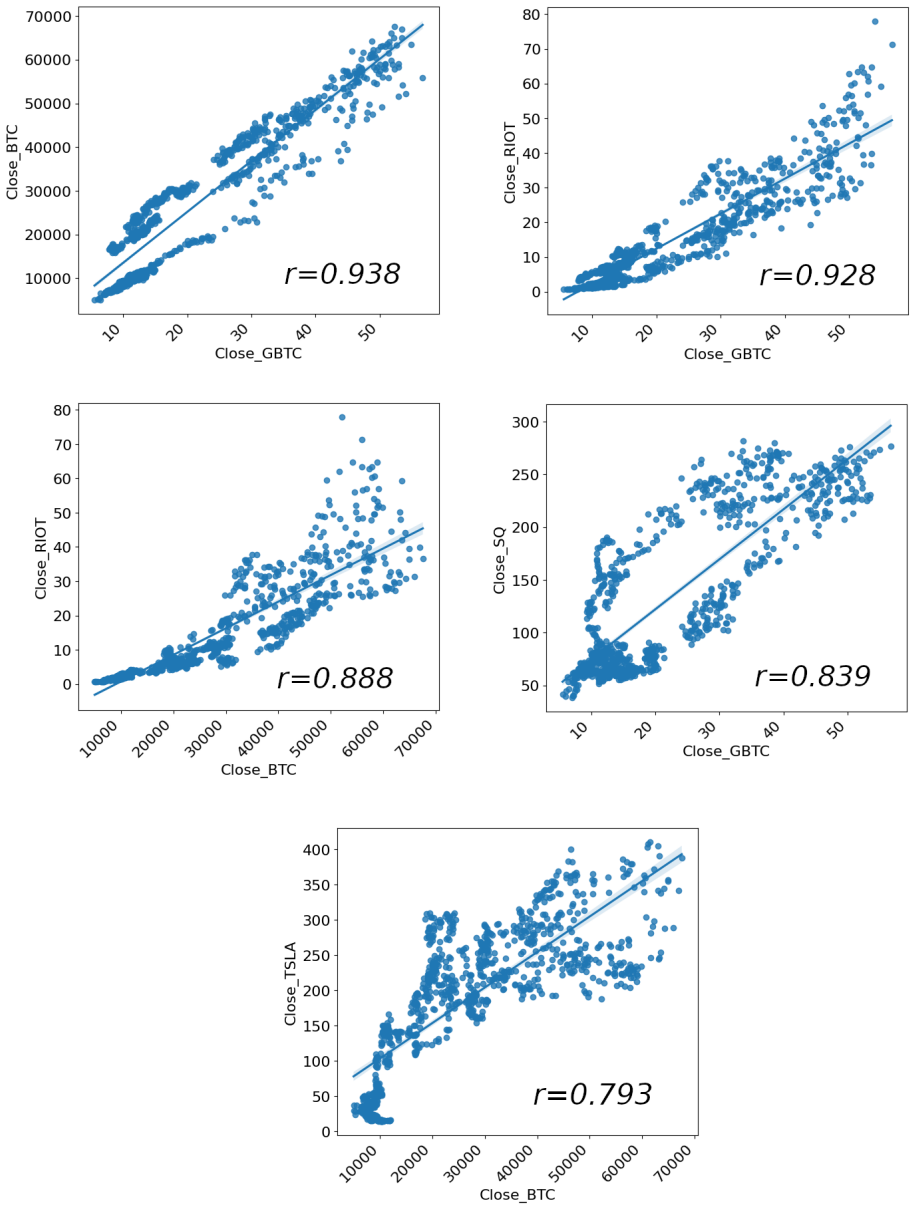
Слика 2: Матрица на расејување за Close вредностите на атрибутите

Но, за да ја видиме точната вредност на кооорелација меѓу атрибутите ќе ја искористиме варијација од претходната функција, наречена heatmap(). Со неа во точни вредности може да се види силата на Pearson корелација меѓу сите овие атрибути.



Слика 3: Матрица со Pearson корелации меѓу Closing атрибутите

Како што се претпостави претходно, највисоката Pearson коорелација има Grayscale Bitcoin Trust (GBTC) со Bitcoin (BTC), со коефициент од r=0.938 и со Riot Blockchain Inc. (RIOT) со коефициент од r=0.928.
Исто така, може да се забележи силна коорелација меѓу Riot Blockchain Inc. (RIOT) и Bitcoin (BTC) со коефициент од r=0.888. Ова не е случајно, затоа што оваа компанија се занимава со копање на крипто валути и нивна размена што значи дека нивниот приход директно влијае од вредноста на ваквите валути.
Доволно силна е и коорелацијата меѓу Square Inc. (SQ) и Grayscale Bitcoin Trust (GBTC) со коефициент од r=0.839. Ова е главно затоа што двете компании работат со истата, што значи дека се директно зависни од флуктуациите на нејзината вредност.
Според Pearson коорелацијата може да се заклучи дека вредноста на Биткоин-от има силна поврзаност со вредноста на берзата на сите овие организации/компании, но најмалку е со Tesla Inc. и Square Inc.(но сепак има влијание).



Слика 4: Дијаграми со линеарна регресија кои ги покажуваат атрибутите со најголема Pearson коефициент на коорелација каде се гледа дека Биткоин ин/директно има влијание врз вредноста на берзизте на Grayscale Bitcoin Trust, Riot Blockchain, Square и Tesla

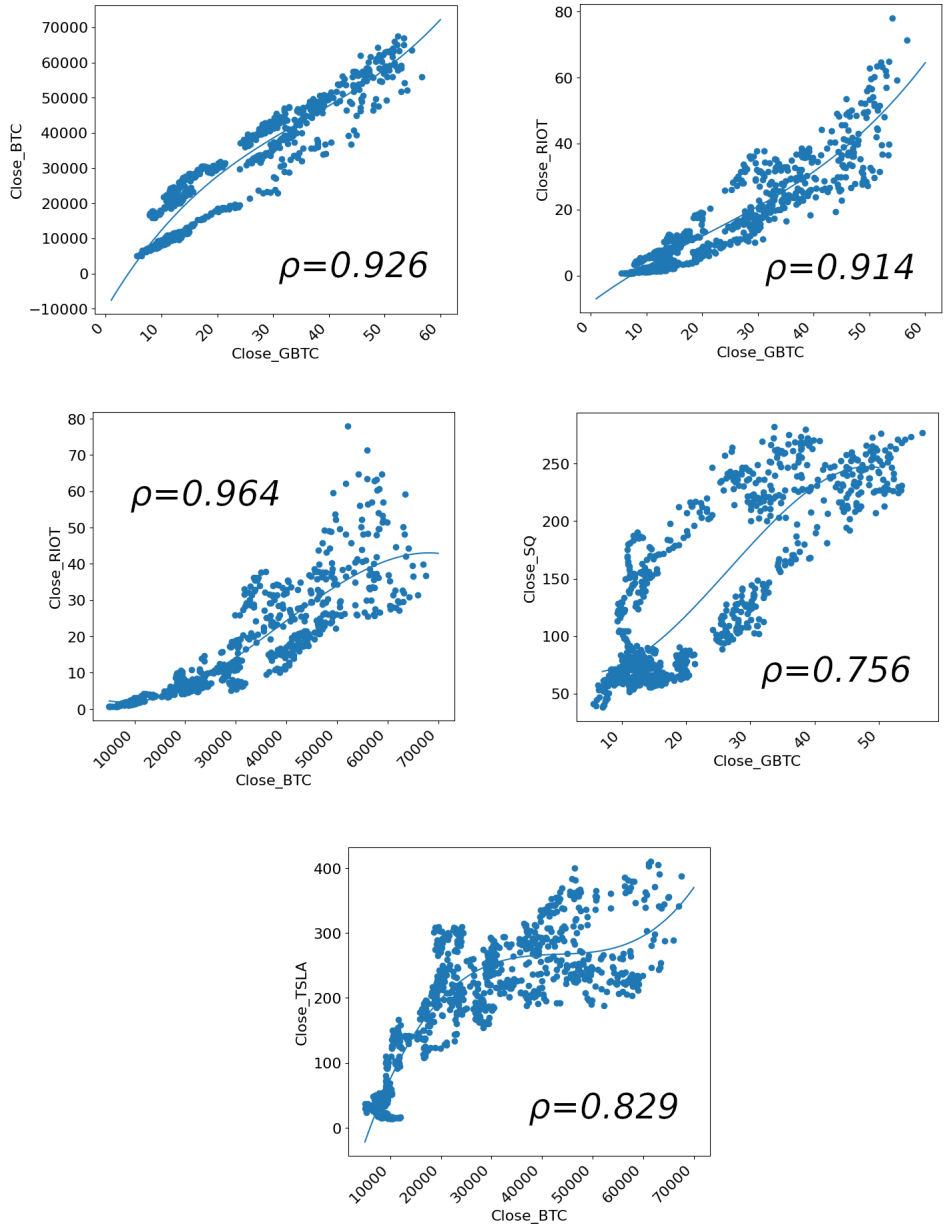
Пронаоѓање на Spearman корелации

Затоа што повеќето атрибути во случајов немаат нормална дистрибуција, користењето на Spearman алгоритмот за откривање на нивната севкупна корелација е посоодветен.
Врската меѓу атрибутите ја опишува преку монотонична функција т.ш. првин ги рангира податоците на двата атрибута и потоа со тие вредности се пресметува Pearson корелација.
Крајниот Spearman коефициент на корелација може да е вредност во интервалот -1 $\leq p \leq$ 1.

Value	Strength
0.00-0.19	very weak
0.20-0.39	weak
0.40-0.59	moderate
0.60-0.79	strong
0.80-1.0	very strong

Табела 2: Сила на врска според Spearman коефициент

Во Python, ова лесно може да се направи со stats.spearmanr() методот што припаѓа на библиотеката SciPy. Како аргументи ги зема само податоците на двата атрибути за кои сакаме да видиме поврзаност.



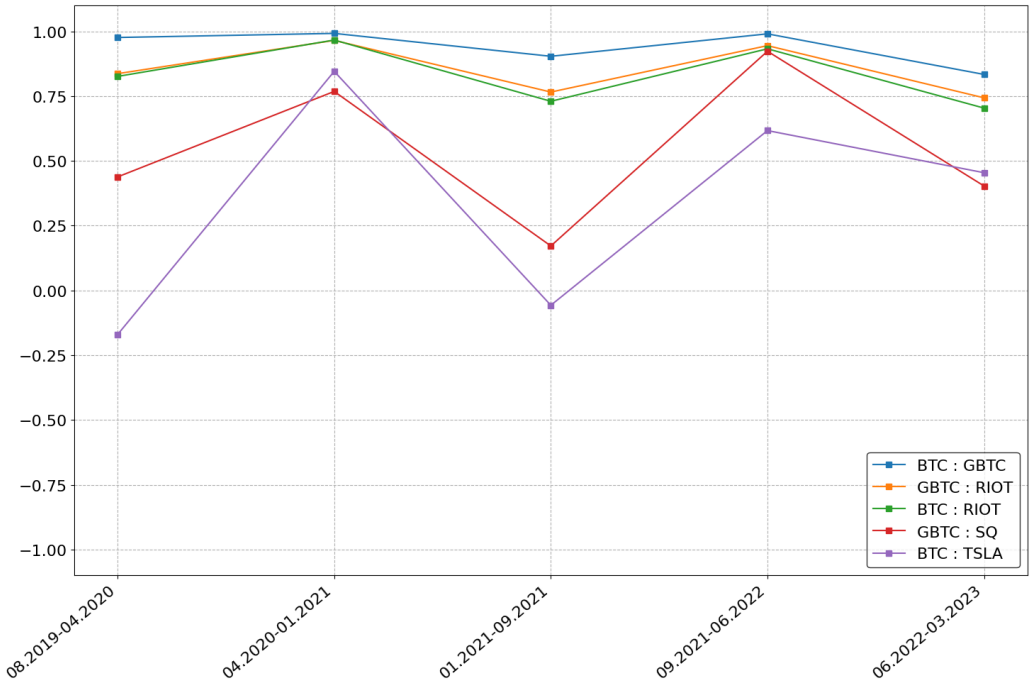
Слика 5: Дијаграми чии регресии носат сличен заклучок за висока корелација меѓу атрибутите, но Spearman корелациските коефициентите сепак ни кажуваат дека има отстапувања особено кај последните 2 графика

Корелации со rolling window

Како последен тип на пресметка на коорелација меѓу атрибутите, ќе се искористи Pearson методот со лизгачки прозорец/интервал. Вака може многу по-прецизно да се заклучи не само севкупното ниво на коорелација(како до сега), туку периодично да заклучиме кои периоди коорелацијата меѓу кои атрибути опаѓа, расте, стагнира сл.
Затоа што нема конкретно правило за одредување на големината на лизгачкиот прозорец, ќе земеме груби претпоставки/бројки и ќе се испробаат неколку големини, а целта е да се извлечат што можно повеќе заклучоци.Ова ќе го постигнеме со Pandas методата .rolling(n).corr() т.ш. лизгачкиот прозорец е со големина n.
Како водич за големина на прозорецот може да ја користиме Слика 1 т.ш. сакаме да избереме што можно поголем прозорец, а со што помалку шум т.е. не премал(затоа што губиме детали), но не и преголем(затоа што има премногу шум и се губи поентата на rolling window методот).
Битно е да се знае дека покрај тоа што прозорецот е со големина од X, тоа се X вредности од „Date“, а не X дена. Ова е затоа што во податоците често има временски дупки т.е. има доста денови во годината кои се испуштени и ова е нормално.

(1) **Ако прозорецот е со големина од 180** тоа значи дека ќе се пресметува коорелација за секој можен интервал од околу 180 дена.
Ако се погледнат резултатите од rolling window анализирата меѓу атрибутите „Close BTC“ и „Close GBTC“, може да се забележи дека коорелацијата е во главно многу силна, но доживува пад во периодот меѓу Јуни 2022 и Март 2023 година.
Ако се погледнат резултатите од rolling window анализирата меѓу атрибутите „Close GBTC“ и „Close RIOT“, може да се забележи исто така силна корелација во повеќето ви интервалите, но се приметуваат и поголеми падови на силата на корелација во периодите меѓу Јануари и Септември 2021 и Јуни 2022 и Март 2023 година.
Скоро идентична приказна е коорелацијата меѓу атрибутите „Close BTC“ и „Close RIOT“ каде наглите падови на коорелација е во истите временски интервали.
Заразлика од претходните примери, голема разлика може да се види кај поврзаноста меѓу атрибутите „Close GBTC“ и „Close SQ“. Овде може да се видат доста нагли падови во нивото на корелација меѓу атрибутите, особено во периодите меѓу Август 2019 и Април 2020 година, Јануари и Септември 2021 година и меѓу Јуни 2022 и Март 2023 година. Ова кажува многу, затоа што претходно на Слика 4, видовме дека коефици-

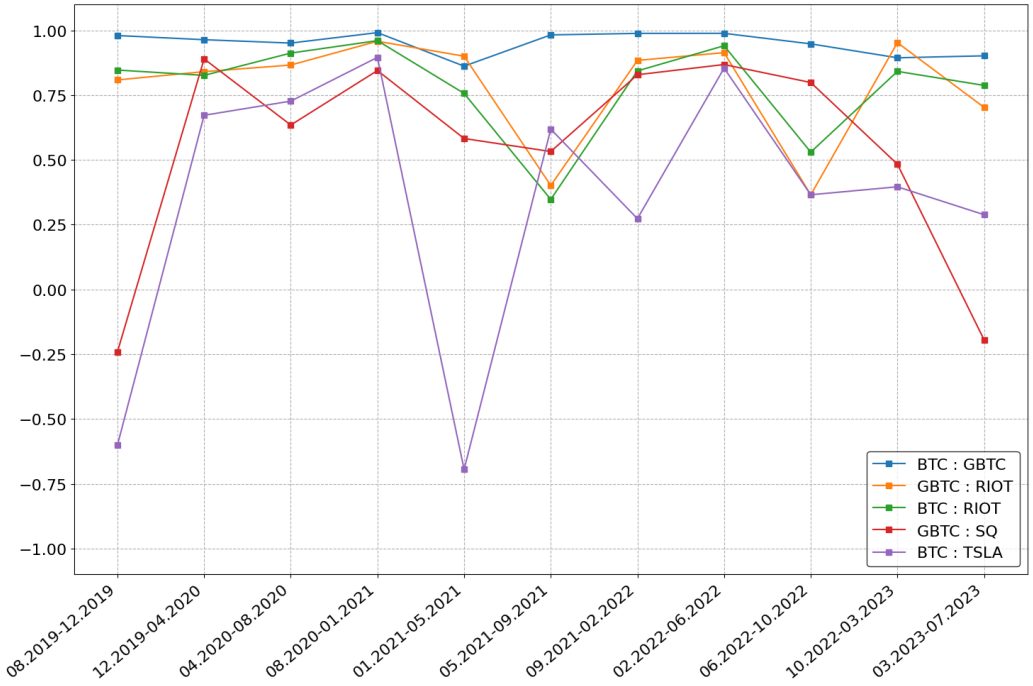
ентот на коорелација меѓу овие атрибути е 0.839(што е прилично висок), а анализата со rolling window покажува дека повеќето од времето корелацијата е ниска. Значи дека резултатот од првичната севкупна Pearson корелациона анализа е со доста голема наклонетост(bias).
Скоро истата приказна е за коорелацијата меѓу атрибутите „Close BTC“ и „Close TSLA“ каде може да се види дека најголема коорелација имало во периодот Април 2020 и Јануари 2021 година(периодот кога Елон Маск на Тесла решава да купи огромна количина на Биткоин, да дозволи купување на Тесла автомобили со валутата итн).



Слика 6: Анализа на нивото на корелација меѓу дадените атрибути со rolling window со вредност од 180

(2) **Ако прозорецот е со големина од 90** тоа значи дека ќе се пресметува коорелација за секој можен интервал од околу 90 дена.
Во овој случај на Слика 7 може многу по-прецизно да се види нивото на корелација меѓу атрибутите, но сепак заклучокот е сличен како со претходната анализа. Повеќе од јасна е високиот степен на корелација меѓу Биткоин и Grayscale Bitcoin Trust (GBTC). Ова има потполна смисла затоа што ова здружение работи со оваа криптовалута и е ранлива на нејзината состојба. **Корелацијата меѓу овие два атрибути е најсилна и најстабилна од сите други во овој пример.**

Заразлика од нејзе, компанијата Riot Blockchain Inc. (RIOT) не е толку зависна и корелирана со целите на оваа криптовалута, иако има моменти кога истата има висок коефициент на корелација(меѓу 2020-2021 и 2023 година). А затоа што берзата на Grayscale Bitcoin Trust (GBTC) е толку поврзана со таа на Биткоин, нормално е дека и Grayscale Bitcoin Trust (GBTC) ќе има скоро исти нивоа на корелација со со Riot Blockchain Inc. (RIOT) како со самиот Биткоин.
Од понудените, убедливо најниски корелациии со Биткоин имаат Square Inc. (SQ) и Tesla Inc. (TSLA), но тоа не значи дека тие немаат никаква корелација. Ова е затоа што заразлика од Riot Blockchain Inc. (RIOT) и Grayscale Bitcoin Trust (GBTC), овие две компании не работат само со криптовалути, туку нудат услуги поврзани со целосно други полиња(Тесла произведува автомобили, а Square е платежен систем кој работи и со стандардни валути пр. долар).



Слика 7: Анализа на нивото на корелација меѓу дадените атрибути со rolling window со вредност од 90