دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی صنایع و سیستم‌های مدیریت

عنوان:

شناسایی رویدادهای مهم کووید-۱۹ از طریق توییتر

نگارندگان:

پدرام پیرو اصفیا

مهدیار صادقی

سپیده کریمی

مهدی محمدی

استاد: دکتر مریم اشرفی

آذر ۱۴۰۰

فهرست مطالب

فهرست جداول

فهرست نمودارها

چکیده:

در طول پاندمی کرونا مردم نشان دادند تمایل دارند زمان بیشتری را در رسانه‌های اجتماعی نسبت به حالت عادی صرف کنند. این میل، پلتفرم‌های رسانه‌های اجتماعی مانند فیس بوک و توییتر را به منبع فعال اطلاعات تبدیل می‌کند که می‌توان با برنامه‌ریزی و اجرای مناسب، اطلاعات ارزشمندی را از این داده ها استخراج کرد. در این پروژه قصد داریم با استفاده از نمودارهای کنترل شوارتز، روشی ابتکاری برای شناسایی زمان رویدادهای جدید در دوران پاندمی کرونا ارائه دهیم.

فصل اول: مقدمه

با توسعه چشمگیر پلتفرم‌های رسانه‌های اجتماعی، افراد بیشتری متون آنلاین را در پلتفرم‌های مختلف ارسال می‌کنند تا نظرات خود را در مورد مسائل اجتماعی بیان کنند (Zhang, Xu, & Wan, 2012).در طول یک بحران، چه طبیعی و چه ساخته دست انسان، مردم تمایل دارند زمان نسبتاً بیشتری را در رسانه‌های اجتماعی نسبت به حالت عادی صرف کنند. با آشکار شدن بحران، پلتفرم‌های رسانه‌های اجتماعی مانند فیس بوک و توییتر به منبع فعال اطلاعات تبدیل می‌شوند(Imran, Castillo, Diaz, & Vieweg, 2015) زیرا این پلتفرم‌ها سریع‌تر از کانال‌های خبری رسمی اخبار را منتشر می‌کنند(Imran, Ofli, Caragea, & Torralba, 2020) ، کرونا هم از این دسته مستثنی نبود. بیماری کروناویروس (COVID-19) یک بیماری عفونی است که توسط یک کروناویروس تازه کشف‌شده ایجاد می‌شود. این بیماری از زمان کشف اولیه در ووهان چین به کشورهای متعددی در سراسر قاره‌ها گسترش‌یافته است و در ۱۱ مارس ۲۰۲۰ توسط سازمان بهداشت جهانی (WHO[1]) به‌عنوان یک بیماری همه‌گیر اعلام شد(WHO Director, 2020) .

در طول چنین رویدادهایی، مردم معمولاً وضعیت سلامتی و ایمنی خود را به اشتراک می‌گذارند، درباره وضعیت ایمنی عزیزان خود جست‌وجو می‌کنند و در ارتباط با رویدادهای جدید گفتگوهای زیادی انجام می‌دهند که این مکالمات در چنین پلتفرم‌های عمومی منجر به جمع‌آوری حجم زیادی از داده‌های اجتماعی می‌شود (Big Crisis Data: Social Media in Disasters and Time-Critical Situations, 2016).

(Andersen, Medaglia, & Henriksen, 2012) نشان دادند که توییتر، یک پلتفرم میکروبلاگینگ[2]، به دلیل طیف گسترده‌ای از برنامه‌های کاربردی، محبوب‌ترین پلتفرم رسانه‌های اجتماعی در میان سایر پلتفرم‌ها است. بررسی دقیق‌تر در مورد تسلط داده‌های توییتر دلایل خاصی را نشان داد، برای مثال، داده‌های توییتر را می‌توان به‌راحتی بر اساس جستجوی کلمه کلیدی یا هشتگ[3] استخراج کرد و همچنین می‌تواند بر اساس جستجوی جدول زمانی خاص پروفایل استخراج شود. این سهولت جمع‌آوری داده‌ها با استفاده از API[4] ها، تجزیه‌وتحلیل عمیق‌تر داده‌های توییتر را تسهیل می‌کند، که اغلب در پلتفرم‌های دیگر وجود ندارد.

به‌صورت عادی در هر هفته تعدادی توییت مرتبط با کرونا زده می‌شود؛ اما هنگامی‌که یک رویداد جدید رخ می‌دهد (کشف واریانت جدید، کشف واکسن جدید، افزایش آمار مبتلایان یا کشته‌شدگان و ...)، تعداد توییت های مرتبط

[1] World Health Organization
[2] Microblogging
[3] Hashtag
[4] Application Programming Interface

با کرونا در هر هفته افزایش چشمگیری می‌یابد. ما می‌توانیم با استفاده از تعداد توییت های مرتبط با کرونا این رویدادهای جدید و میزان واکنش مردم جهان به آن‌ها را شناسایی کنیم.

فصل دوم: شرح مسئله

همان‌طور که گفته شد توییتر یک پلتفرم میکروبلاگینگ است که به‌وسیله آن، کاربران محتوای اصلی را از طریق توییت کردن افکار، نظرات و اخبار به اشتراک می‌گذارند و با باز توییت کردن، علاقه‌مندی و پاسخ دادن به توییت های دیگران، با محتوای موجود درگیر می‌شوند. توییتر همچنین پلتفرمی برای انتقال دانش، باورها و نگرانی‌ها در مورد مسائل روز دنیا مانند همه‌گیری کووید -۱۹ است. ازاین‌رو، منبع خوبی برای بررسی افکار عمومی و میزان واکنش و بازخورد مردم به وقایع می‌باشد(Kwona & Grady, 2020) . اما تعداد توییت‌های منتشرشده در طول گسترش کووید -۱۹ در طول زمان ثابت نبوده است. عواملی همچون افزایش مبتلایان و فوتی‌ها، کاهش یا افزایش نگرانی‌ها، خبر کشف واکسن، خبر ممنوعیت یا ورود واکسن، عادی انگاری کرونا در طول زمان، فوت چهره‌های شناخته‌شده در اثر کرونا، شناسایی سویه‌های جدید و... می‌توانند بر تعداد توییت‌ها با موضوع کرونا تأثیر بگذارند.

برای شناسایی رفتار کاربران توییتر در قبال ویروس کرونا و مسائل مربوط به آن، پژوهشگران بسیاری روی این مسئله کارکرده‌اند. به‌عنوان‌مثال (Chen Lyu, Han, & Luli, 2021)در مقاله خود با استفاده از تحلیل داده‌های توییتری در مقیاس بزرگ، بیان می‌کند بحث عمومی مرتبط با واکسن کووید-۱۹ در توییتر عمدتاً ناشی از رویدادهای مهم در مورد واکسن‌های کووید-۱۹ است و موضوعات خبری فعال در رسانه‌های جریان اصلی را منعکس می‌کند؛ به‌عنوان‌مثال در حوالی ۱۱ آگوست ۲۰۲۰، زمانی که روسیه اولین واکسن کووید-۱۹ را تأیید کرد، تبدیل به موضوع موردبحث شد. با پیشرفت تجویز واکسن، موضوع آموزش در مورد دریافت واکسن به‌تدریج برجسته‌تر شد و پس از هفته اول ژانویه ۲۰۲۱ به بیشترین بحث تبدیل شد. همچنین (Wan, Li, Hutch, Naidech, & Luo, 2021) در مقاله‌ای، از توییت‌ها برای بررسی باورهای بهداشتی مرتبط با کرونا و بررسی عوامل تأثیرگذار مرتبط با نوسانات در باورهای سلامت در رسانه‌های اجتماعی استفاده کرده‌اند.

نگاه کلی به تعداد توییت‌ها در بازه زمانی ۸۱ هفته‌ای نشان داد تعداد توییت‌ها در برخی هفته‌ها رشد بسیار زیادی داشته و در بعضی موارد نیز با کاهش همراه بوده است. بنابراین در این پروژه با استفاده از نمودارهای کنترلی، اقدام به بررسی روند تعداد توییت‌های منتشرشده در بازه زمانی ۸۱ هفته‌ای می‌کنیم. با این کار متوجه خواهیم شد در کدام هفته‌ها تعداد توییت‌ها افزایشی یا کاهشی بوده و هرگاه تعداد توییت‌ها کمتر یا بیشتر از حدود کنترلی ما شد، به سراغ دیتاست روزانه آن هفته رفته و بررسی می‌کنیم که چه رویدادی در چه تاریخی باعث این تغییر غیر تصادفی شده است. این تحلیل، به ما در شناخت مسائل و اخباری از شیوع کرونا که بازخورد مردم جهان را به دنبال داشته‌اند و همچنین شناخت روندی که موجب کاهش توجه مردم به مسائل و اخبار مربوط به کووید-۱۹ بوده است، کمک می‌کند.

فصل سوم: روش حل مسئله

جمع‌آوری داده

توییتر دو نوع API را ارائه می‌دهد: جستجو API[5] و استریم API[6]. API استریم برای دسترسی به توییت‌ها از فید توییتر بلادرنگ[7] استفاده می‌کند. برای این مطالعه، API استریم از ۲۰ مارس ۲۰۲۰ استفاده می‌شود . (Lamsal, Design and analysis of a large-scale COVID-19 tweets dataset, 2021)

کلیدواژه‌ها

تعداد کلمات کلیدی از زمان شروع این مطالعه به‌طور مداوم در حال تغییر بوده است. جدول ۱ یک نمای کلی از کلیدواژه‌هایی که در حال حاضر استفاده می‌شوند را نشان می‌دهد. با رشد همه‌گیر، بسیاری از کلمات کلیدی جدید ظاهر شدند. در ۱۳ می ۲۰۲۰، توییتر همچنین فهرستی از ۵۶۴ کلمه کلیدی فیلتر چندزبانه مورداستفاده در نقطه پایانی استریم COVID-19 خود را منتشر کرد. API استریم به توسعه‌دهندگان اجازه می‌دهد تا از ۴۰۰ کلمه کلیدی برای فیلتر کردن جریان توییتر استفاده کنند.

در این پروژه از COVID19Tweets Dataset که از طریق API استریم و جدول کلیدواژه آمده استفاده شده است(Lamsal, Design and analysis of a large-scale COVID-19 tweets dataset, 2021) .

جدول ۱: بررسی اجمالی کلیدواژه‌ها تا ۱۷ جولای ۲۰۲۰

کلیدواژه‌ها	در حال استفاده از
corona, #corona, coronavirus, #coronavirus	March 20, 2020
covid, #covid, covid19, #covid19, covid-19, #covid-19, sarscov2, #sarscov2, sars cov2, sars cov 2, covid 19, #covid 19, #ncov, ncov, #ncov2019, ncov2019, 2019-ncov, #2019-ncov, #2019ncov, 2019ncov	April 18, 2020
pandemic, #pandemic, quarantine, #quarantine, flatten the curve, flattening the curve, #flatteningthecurve, #flattenthecurve, hand sanitizer, #handsanitizer, #lockdown, lockdown, social distancing, #socialdistancing, work from home, #workfromhome, working from home, #workingfromhome, ppe, n95, #ppe, #n95	May 16, 2020

[5] Search API
[6] Stream API
[7] Real-time twitter feed

به دلیل اینکه صرفاً تعداد این توییت‌ها (و نه محتوای آن‌ها) برای ما اهمیت داشت و تعداد توییت‌های روزانه از تاریخ ۲۰ مارس ۲۰۲۰ در سایت مربوط به دیتاست اعلام شده بود، ما اقدام به تهیه یک دیتاست از این اطلاعات کردیم(Lamsal, Coronavirus (COVID-19) Tweets Dataset, 2020) .

باید توجه داشت به دلیل تغییرات کلیدواژه‌ها تعداد داده‌ها تا تاریخ ۱۶ می در دو مرحله جهش داشت که باعث بی‌فایده بودن تحلیل تا قبل از آن تاریخ می‌شود؛ درنتیجه داده‌های قبل از آن تاریخ حذف شد و داده‌ها از تاریخ ۱۶ می ۲۰۲۰ تا ۳ دسامبر ۲۰۲۱ بررسی شدند.

پیش‌پردازش و گروه‌بندی داده

دیتاست موجود، تعداد توییت‌ها را به‌صورت روزانه در اختیار قرار داده بود. اما بنا به دو دلیل تعداد توییت های روزانه را به هفتگی تبدیل کردیم:

۱- افراد به دلیل مشغله زیاد و یا اهمیت برخی رویدادها ممکن است تا چند روز پس از رویداد در ارتباط با آن رویداد توییت بزنند؛ بنابراین ممکن است واکنش ها پس از یک رویداد مهم، تا چندین روز ادامه داشته باشد.

۲- بررسی هفتگی هوشمندانه‌تر از بررسی روزانه به نظر می‌رسد. همان‌گونه که پیش‌تر بیان شد، هنگامی‌که تعداد توییت های هفتگی از حدود کنترلی تجاوز کند، به جای همه روزها، فقط روزهای آن هفته را از دیتاست روزانه بررسی می‌کنیم و رویداد مربوطه را شناسایی می‌کنیم.

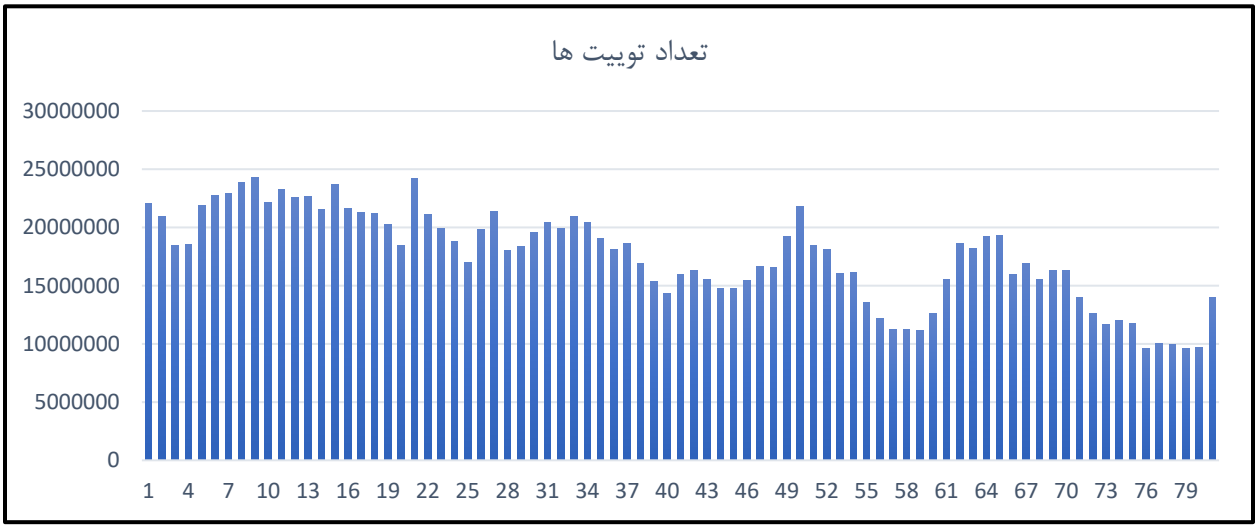
برای این کار از کتابخانه قدرتمند pandas زبان Python استفاده شد و تعداد توییت‌ها در هر هفته (که هفته‌ها از شنبه شروع می‌شود) به دست آمد (تاریخ بررسی‌شده شامل ۸۱ هفته کامل می‌باشد).

انتخاب نمودار کنترلی

با توجه به اینکه مشاهدات ما انفرادی است استفاده از نمودار $\bar{X}/R$ مجاز نیست (زیرا باید بین ۳ تا ۵ مشاهده داشته باشیم) بنابراین از نمودار I/MR استفاده شود بدین‌صورت که هر هفته یک عضو در نظر گرفته می‌شود و برای محاسبه انحراف معیار از تفاوت تعداد توییت‌های هر هفته استفاده می‌کنیم.

تقسیم دیتاست به دو بخش

دلایلی مانند کاهش قرنطینه، عادت مردم به زندگی در شرایط کرونایی، ساخت واکسن و کاهش نگرانی‌ها، موجب کاهش توییت‌ها با موضوع کرونا شده است؛ به‌طوری‌که از جایی به بعد، میانگین تعداد توییت‌ها به وضوح دچار تغییر می‌شود. با توجه به اهمیت این موضوع، ماکسیمم اختلاف میانگین‌ها را محاسبه کرده و هفته ۳۷ ام به‌عنوان محل شروع تغییر میانگین شناسایی شد. به‌این‌ترتیب دیتاست به ۲ بخش تقسیم شده و هر بخش جداگانه مورد بررسی قرار می‌گیرد.

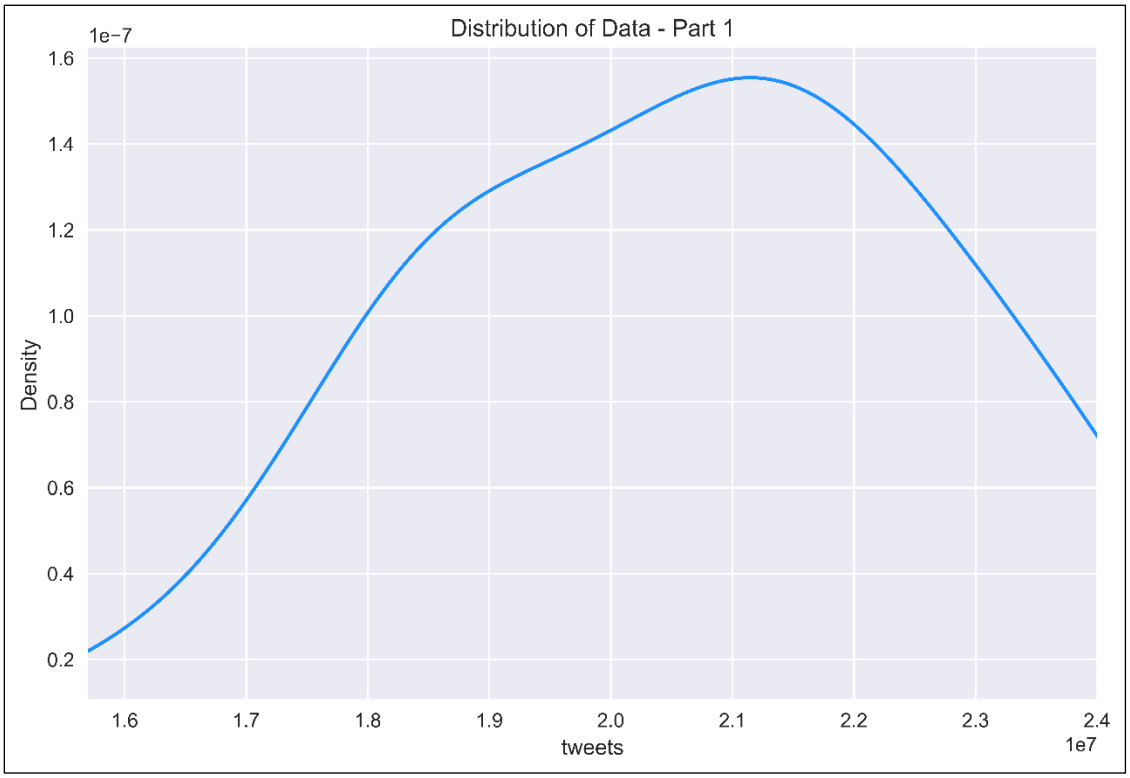


نمودار ۱: نمودار تعداد توییت های هر هفته

فصل چهارم: تحلیل قسمت اول دیتاست

بررسی نرمال بودن داده‌ها

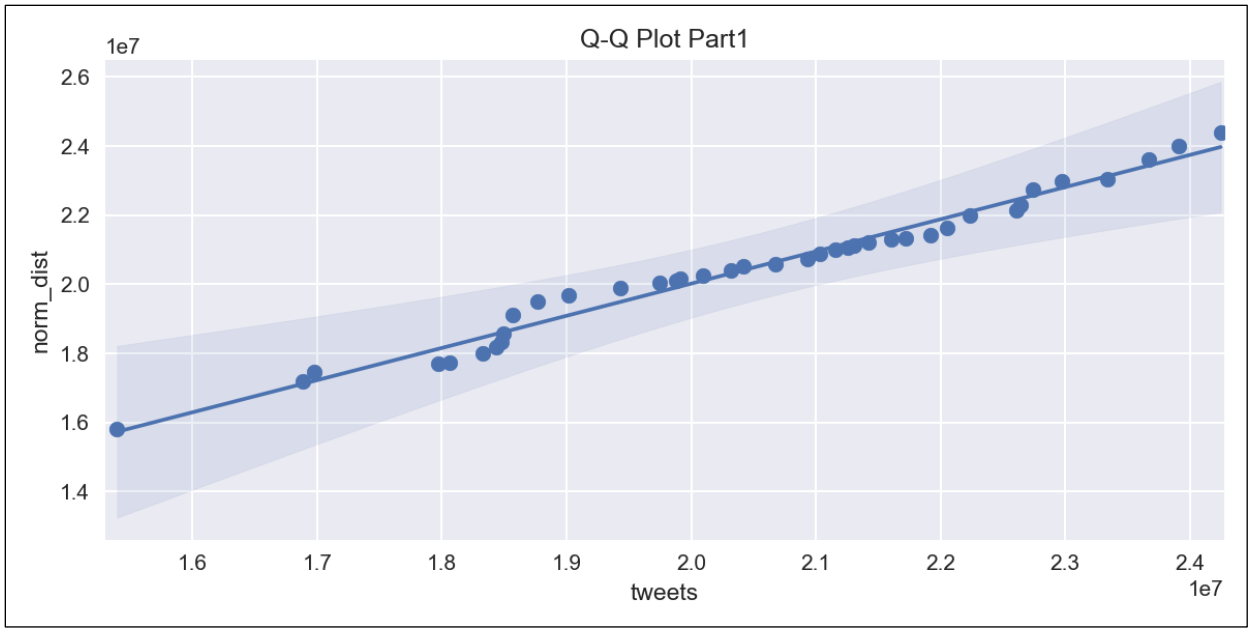
در ابتدای فرایند تحلیل به سراغ بررسی نرمال بودن یا نبودن توزیع بخش اول دیتاست خود می‌رویم. برای این کار می‌توان یا نمودار هیستوگرام یا kde[8] را رسم کرد. با توجه به اینکه نمودار kde انعطاف‌پذیری بیشتری را ارائه می‌دهد و به واقعیت نزدیک‌تر است (Wergieluk, 2020)، از کتابخانه‌های Seaborn و matplotlib در Python استفاده کرده و نمودار kde را جهت برآورد تابع چگالی احتمال دیتا رسم می‌کنیم:



نمودار ۲: نمودار kde تعداد توییت‌ها – بخش اول دیتاست

همان‌طور که مشاهده می‌شود نمودار رسم شده، شباهت زیادی به توزیع نرمال نداشته و دارای چولگی به سمت چپ است. به جهت بررسی بیشتر و دقیق‌تر با استفاده از کتابخانه‌ی matplotlib در Python به سراغ رسم نمودار Q-Q plot بافاصله اطمینان ۹۵٪ می‌رویم:

[8] Kernel Density Function



نمودار ۳: Q-Q plot برای توزیع تعداد توییت کاربران – بخش اول دیتاست

با توجه به اینکه نقاط در نمودار Q-Q plot با تقریب خوبی در حوالی نیمساز ربع اول و سوم قرارگرفته‌اند و از طرفی تمامی نقاط در بازه‌ی اطمینان ۹۵٪ قرار دارند، می‌توان نتیجه گرفت که مجموعه دیتای بخش اول از توزیع نرمال پیروی می‌کنند. بااین‌حال جهت اطمینان بیشتر نسبت به نرمال بودن توزیع داده‌ها با استفاده از زبان برنامه‌نویسی Python و کتابخانه‌ی Scipy و انجام دادن آزمون، فرض نرمال بودن داده‌ها را موردبررسی قراردادیم.

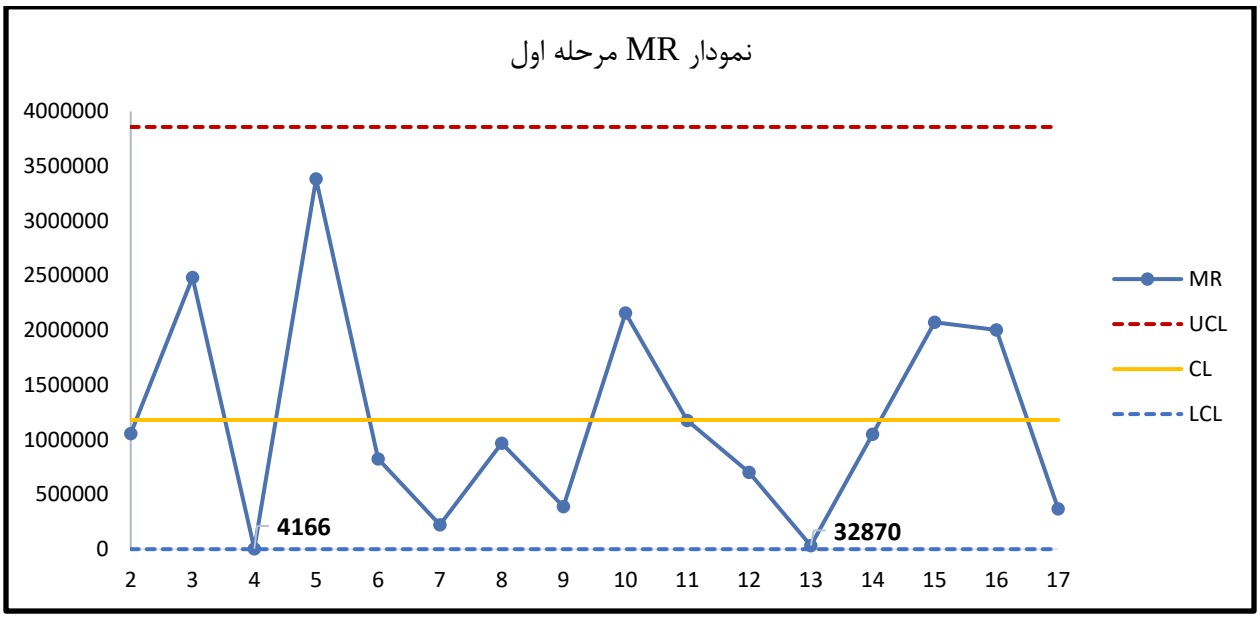
$$\begin{cases} H_0: \text{نمونه ها از توزیع نرمال پیروی می‌کنند} \\ H_1: \text{نمونه ها از توزیع نرمال پیروی نمی‌کنند} \end{cases}$$

که پس از انجام آزمون با سطح اطمینان ۵درصد، مقدار p-value برابر ۰.۶۷ به دست آمد و درنتیجه امکان رد فرض صفر وجود ندارد و بنابراین دیتای مجموعه اول از توزیع نرمال پیروی می‌کند. پس از اطمینان از نرمال بودن توزیع داده‌ها به تشکیل فاز اول می‌پردازیم.

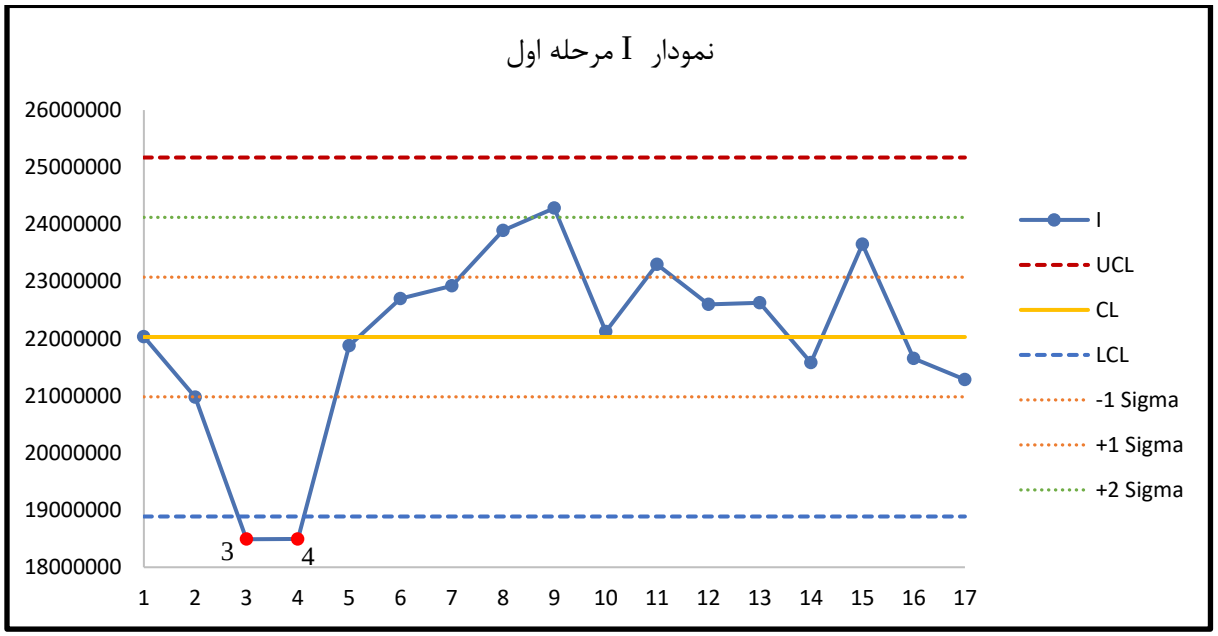
فاز اول نمودارها کنترلی MR و I

برای انجام این فاز، احتیاج داریم از دیتاهایی استفاده کنیم که بیشتر نسبت به تحت کنترل بودن و همچنین صحیح بودن آن‌ها مطمئن هستیم. بنا بر آگاهی‌ای که نسبت به دیتاست و اتفاقات رخداده در بازه‌های مختلف زمانی داشتیم، قسمت‌های ابتدایی یا به عبارتی ۱۷ هفته اول را به‌عنوان فاز اول در نظر گرفتیم و از این قسمت

برای تعیین حدود و انجام فاز اول پروژه استفاده می‌کنیم. طبق محاسبات موردنیاز برای محاسبه‌ی حدود کنترلی در MR و I ،که تمامی آن‌ها در فایل اکسل زمینه موجود می‌باشد، نمودارهای MR و I به شکل زیر است:

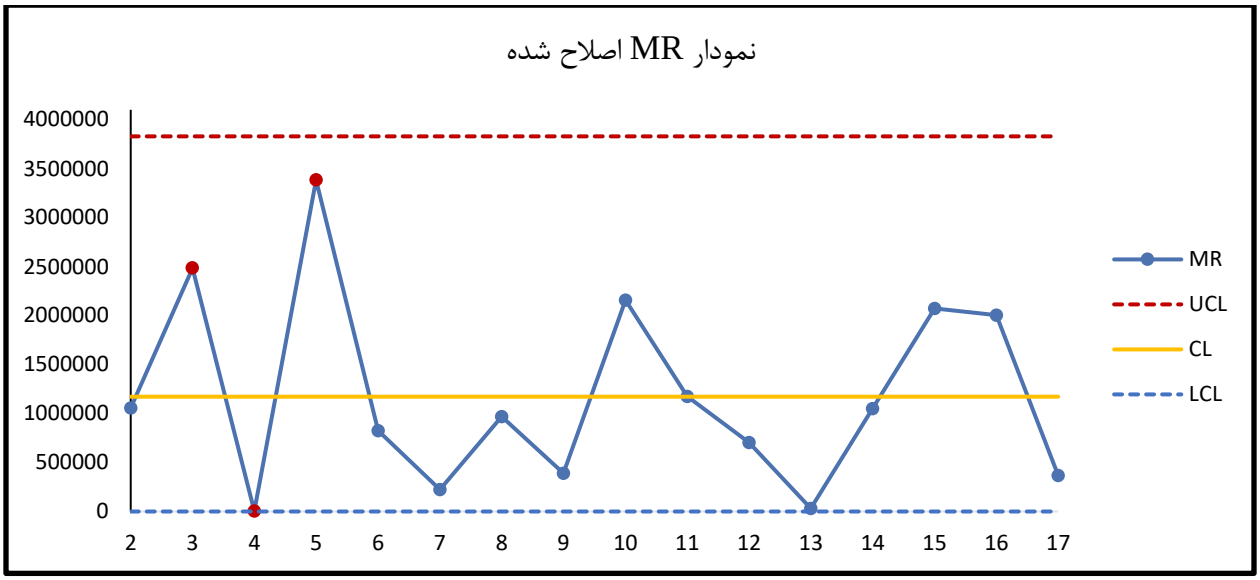


نمودار ۴: نمودار کنترلی MR برای فاز اول- بخش اول دیتاست

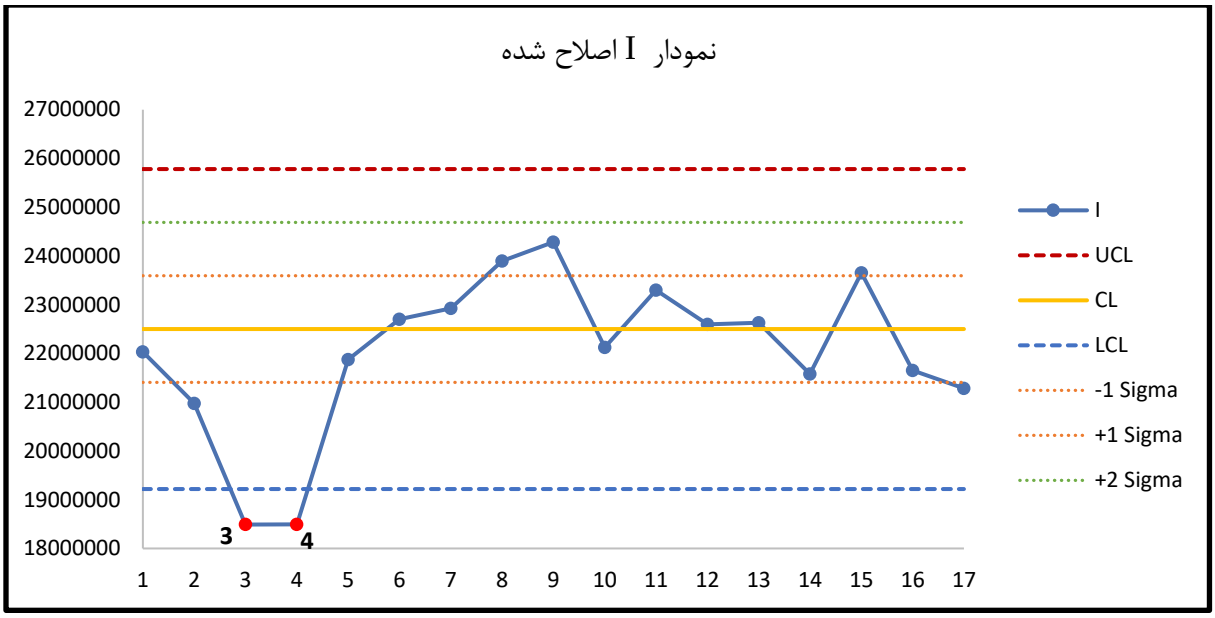


نمودار ۵: نمودار کنترلی I برای فاز اول - بخش اول دیتاست

همان‌طور که مشخص است، نمودار MR هشدار خارج از کنترل بودن دیتایی را اعلام نمی‌کند. ولی در نمودار I دونقطه خارج کنترل و پایین‌تر از LCL قرارگرفته‌اند. پس از بررسی‌های انجام‌شده مشخص شد در هفته‌ی سوم و چهارم تقریباً خبر یا حاشیه‌ای داغ درباره ویروس کرونا در خبرگزاری‌ها وجود ندارد و حتی اخباری مبنی بر وجود احتمال بازگشایی کالج‌ها یافت می‌شود(Thys, 2020) به همین دلیل شاهد کاهش تعداد توییت‌ها هستیم. به دلیل فاصله‌ی زیاد نقاط ۳و ۴ از حد پایین نمودار،آن‌ها را حذف و به بررسی و محاسبه مجدد پرداختیم. همچنین همان‌طور که در نمودار I مشخص است از هفته‌ی ۵ تا هفته‌ی ۹ شاهد یک‌روند افزایشی هستیم که پس از بررسی‌های صورت گرفته، مشخص شد در بازه‌ی فوق بسیاری از مردم آمریکا در اعتراض به کشته شدن جورج فلوید توسط افسر پلیس شروع به تظاهرات گسترده کردند و موجب افزایش نگرانی‌ها درباره احتمال شیوع دوباره کرونا به دلیل عدم رعایت شیوه‌نامه‌های بهداشتی شده‌اند؛ که در واکنش به این اتفاق تعداد توییت‌ها سیر صعودی به خود گرفته است(N. Bleich, 2020) . در ادامه نمودارهای MR و I را پس از حذف دیتاهای شماره ۳ و ۴ ملاحظه می‌کنید:



نمودار ۶: نمودار MR اصلاح‌شده فاز اول



نمودار ۷: نمودار I اصلاح‌شده فاز اول

همان‌طور که ملاحظه می‌کنید پس از حذف دیتاهای ۳ و۴ از محاسبات سایر نقاط در میان حدود قرارگرفته است و این به معنای پایان فاز اول محاسبات و آغاز فاز دوم می‌باشد.

حدود کنترلی به‌دست‌آمده برای فاز دوم به شرح زیر است:

جدول ۲: حدود کنترل نمودار MR فاز اول- بخش اول

حدود کنترل چارت MR		
UCL	CL	LCL
۳۸۲۸۹۳۹	۱۱۷۲۰۰۵	۰

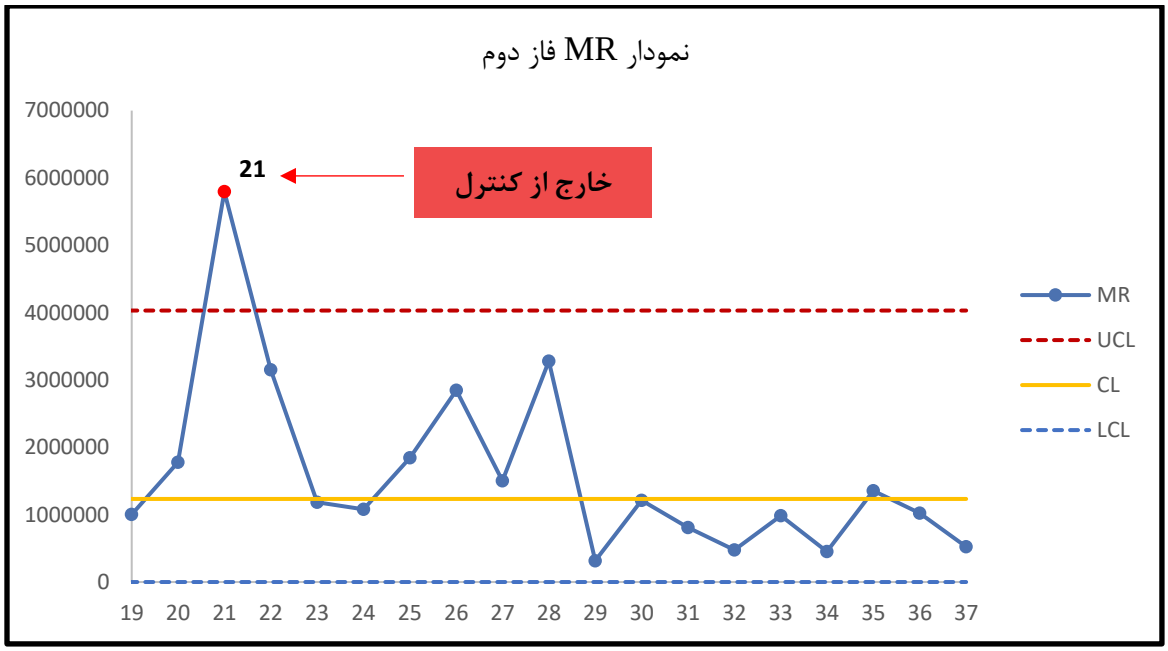
جدول ۳: حدود کنترل نمودار I فاز اول- بخش اول

حدود کنترل چارت I		
UCL	CL	LCL
۲۵۶۱۷۵۲۲	۲۲۵۰۰۴۸۸	۱۹۳۸۳۴۵۴

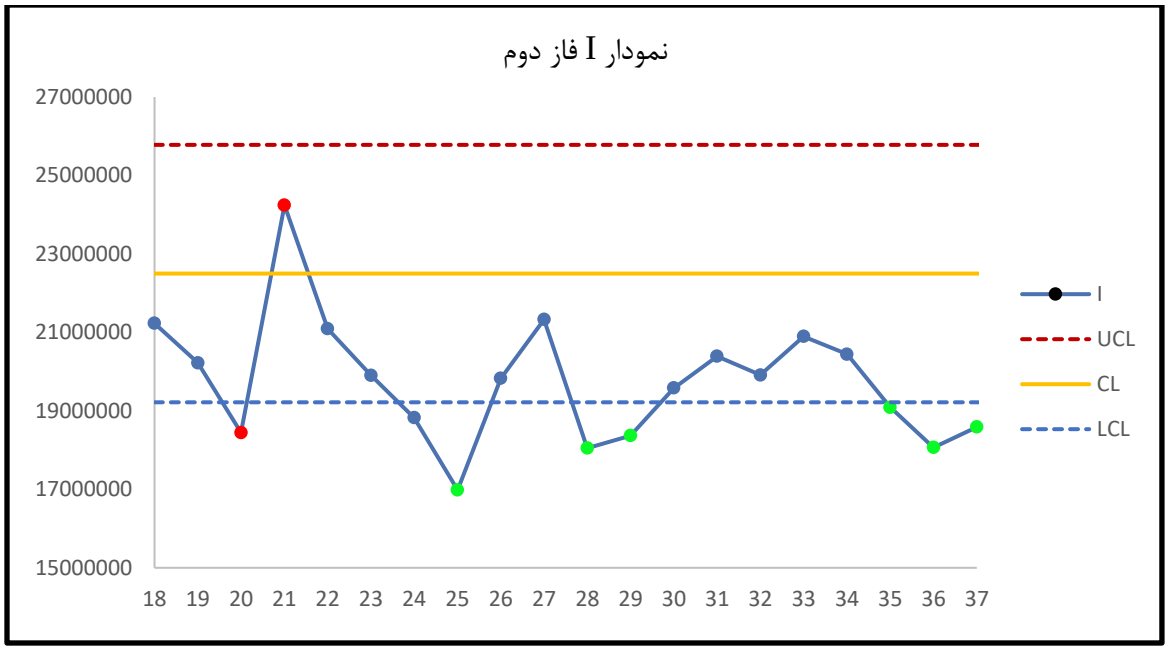
فاز دوم نمودارهای کنترلی MR و I

حال با توجه به حدود به‌دست‌آمده از فاز اول، می‌توانیم وارد فاز دوم شده و مسئله خود را ازنظر تحت کنترل بودن یا نبودن بررسی کنیم. با توجه به ماهیت موضوع در حال بررسی انتظار می‌رود در صورت منتشر شدن خبر یا

حاشیه‌ی تازه، شاهد اوج گرفتن تعداد توییت‌ها باشیم؛ البته همان‌طور که مشخص است این اوج گرفتن مربوط به یک بازه‌ی کوتاه بوده و پس از بازه‌ی فوق شاهد سیر نزولی تعداد توییت‌ها خواهیم بود. بنابراین ما به کمک حدود کنترلی به‌دست‌آمده در فاز اول، به سراغ رسم و تحلیل کنترل چارت‌های MR و I می‌رویم. همانند فاز اول ابتدا شروع به بررسی هفتگی دیتاها کرده و سپس در صورت بروز شرایط خارج از کنترل به بررسی روزانه هفته مورد نظر می‌پردازیم. در ادامه نمودارهای این بخش را مشاهده می‌کنید:



نمودار ۸: نمودار کنترلی MR برای فاز دوم



نمودار ۹: نمودار کنترلی I برای فاز دوم

پس از بررسی‌های صورت گرفته در دیتاست روزانه مشخص شد هشدار دریافتی، مربوط به اوایل اکتبر ۲۰۲۰ است. پس از جست‌وجو در اخبار مشخص شد، در بازه‌ی فوق تست کرونای دونالد ترامپ رئیس‌جمهور وقت آمریکا در بازه‌ی حدوداً یک ماه مانده به انتخابات ۲۰۲۰ آمریکا، مثبت شده و باعث ایجاد موجی از توییت‌ها در رابطه با این مضمون شده است(Lazzar, 2020) . توییت‌هایی با عناوینی مانند "چه کسی ترامپ را آلوده کرد؟"، "احتمال آلوده شدن بایدن به دلیل شرکت در مناظره با ترامپ"، "انتقادات به کاخ سفید به دلیل بسنده کردن به تست کرونا و عدم رعایت سایر پروتکل‌های بهداشتی"(Weintraub, 2020) .

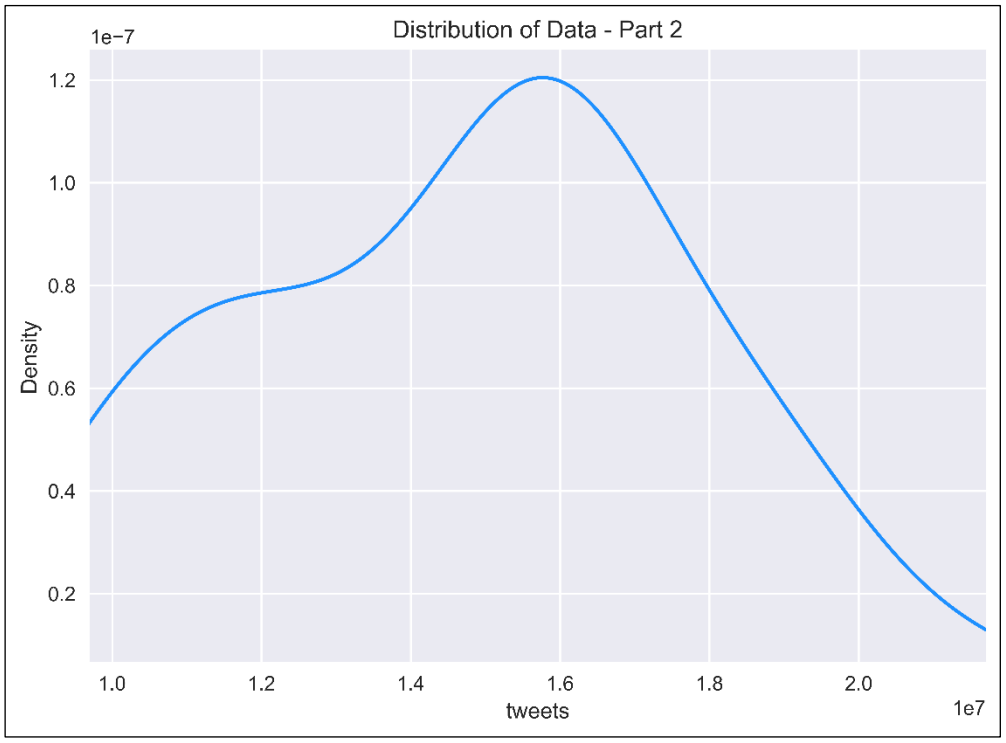
پس از ریشه‌یابی علت هشدار خارج از کنترل در هفته‌ی ۲۱ ، به بررسی سایر نقاط خارج از کنترل در نمودار I پرداختیم. همان‌طور که از اخبار مربوط به بازه‌های زمانی فوق برداشت می‌شود، اهمیت مردم به مقوله‌ی کرونا به‌شدت کاهش‌یافته، به‌طوری حتی درصد رعایت پروتکل‌های بهداشتی توسط مردم کاهش چشم‌گیری داشته است(Allen & Lipsitch, 2020) . همین مسئله نشان‌دهنده‌ی چرایی پایین‌تر از حد کنترل قرار گرفتن تعداد توییت‌ها در هفته‌های نشان داده شده در نمودار I است. کرونا در حال عادی شدن برای مردم است و این موضوع به این معناست که تعداد توییت‌ها درباره‌ی این موضوع سیر نزولی به خود گرفته و فقط در اثر اتفاقات خاص (مانند کرونا گرفتن ترامپ) یا ایجاد حواشی‌ای مانند گمانه‌زنی‌ها درباره پیک‌های بعدی (Weintraub, 2020) ، مردم به اظهارنظر در توییتر خواهند پرداخت. به‌طور مثال در تاریخ ۱۰ نوامبر(loffe, 2020)، واشنگتن‌پست خبری مبنی نزدیک بودن زمان فرارسیدن موج جدید را منتشر می‌کند یا در همان حوالی گمانه‌زنی‌هایی مربوط به کشف واکسن شکل می‌گیرد و ما پس از چند هفته نزول، شاهد ایجاد روند صعودی در هفته‌های ۲۶و ۲۷ هستیم

(Weintraub, 2020). و یا از هفته‌ی ۳۰ تا ۳۳یک‌روند افزایشی ملاحظه می‌کنیم که پس از بررسی اخبار مشخص شد، در این بازه (یعنی اوایل دسامبر ۲۰۲۰) اخبار مربوط به کشف واکسن‌های فایزر و مدرنا منتشرشده و درنتیجه تعداد توییت‌ها را بالابرده است (Branswell, 2020) ؛ و در هفته‌های بعد همان‌طور که مشاهده می‌کنید پس از چند هفته صعود دوباره تعداد توییت‌ها سیر نزولی به خود گرفته است.

فصل پنجم: تحلیل قسمت دوم دیتاست

بررسی نرمال بودن داده‌ها

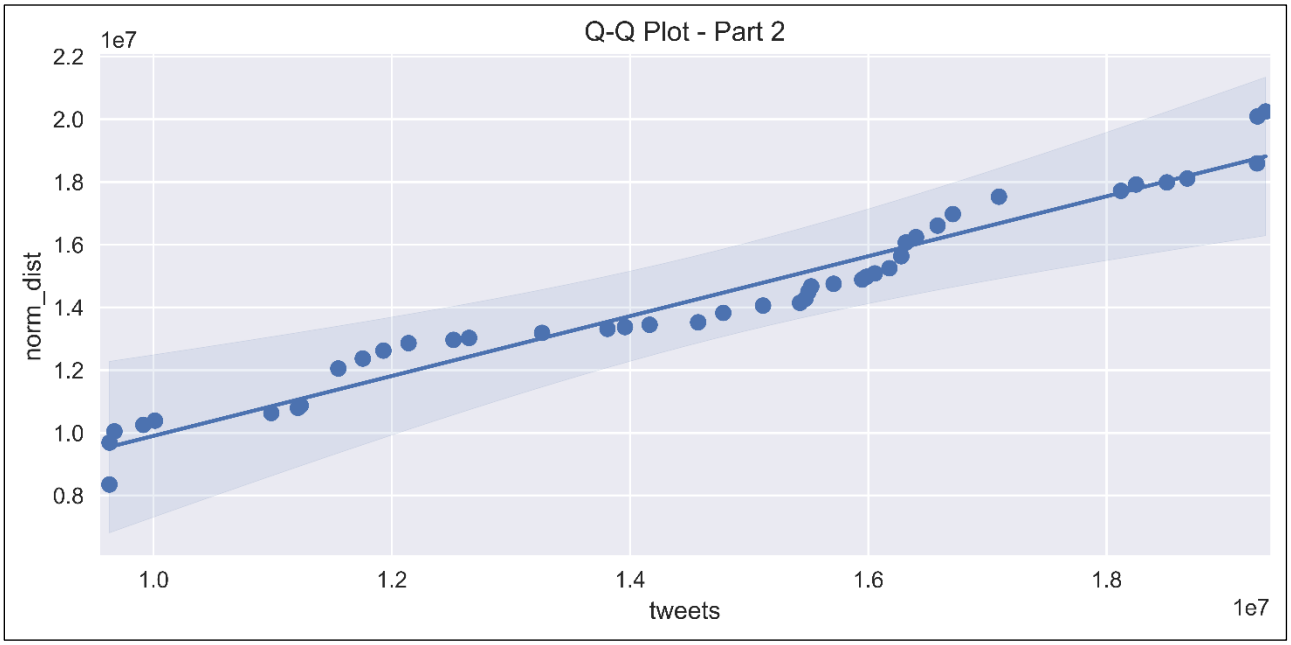
در این بخش به بررسی داده‌های بخش دوم دیتاست(۴۴هفته پایانی) می‌پردازیم. به‌طور مشابه با بخش اول، ابتدا به بررسی نرمال بودن توزیع دیتاست با استفاده از نمودار kde کتابخانه‌های seaborn و matplotlib میپردازیم:



نمودار ۹: نمودار kde تعداد توئیت‌ها – بخش دوم دیتاست

با توجه به نمودار رسم شده، نمودار چولگی به سمت راست دارد؛ همچنین قله‌ی کوچکی در سمت چپ مشاهده می‌شود. به همین علت فرض نرمال بودن را باید دقیق‌تر بررسی نمود.

به این منظور، نمودار Q-Q plot بافاصله اطمینان ۹۵٪ با استفاده از کتابخانه‌ی matplotlib در Python را رسم می‌کنیم:

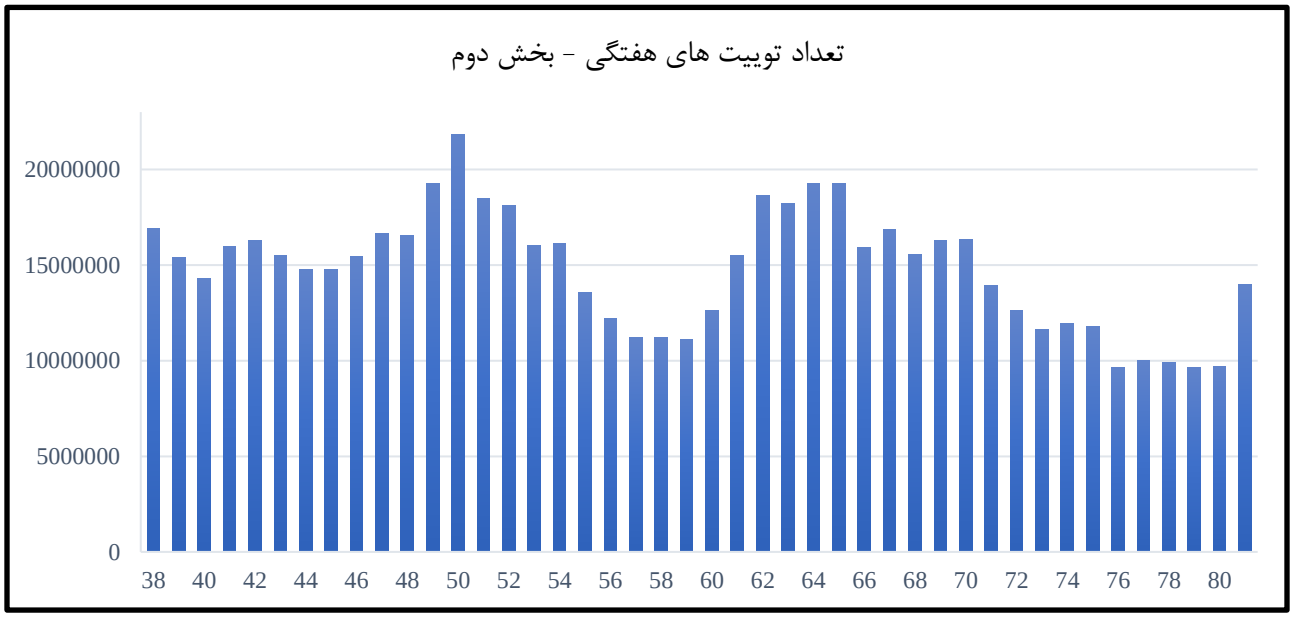


نمودار ۱۰: Q-Q plot برای توزیع تعداد توئیت‌ها- بخش دوم دیتاست

نمودار Q-Q plot نشان می‌دهد نقاط تقریب نسبتاً خوبی در حوالی نیمساز ربع اول و سوم قرارگرفته‌اند و تقریب توزیع نرمال برای داده‌ها، تقریب بدی نیست. بااین‌حال جهت اطمینان بیشتر مجدداً با استفاده از زبان برنامه‌نویسی Python و کتابخانه‌ی Scipy و انجام دادن آزمون، فرض نرمال بودن داده‌ها را موردبررسی قراردادیم. آزمون فرض زیر را انجام می‌دهیم.

$$\begin{cases} H_0: \text{نمونه ها از توزیع نرمال پیروی می‌کنند} \\ H_1: \text{نمونه ها از توزیع نرمال پیروی نمی‌کنند} \end{cases}$$

مقدار p-value محاسبه‌شده ۰.۵۶ است که نشان از عدم رد فرض صفر و نرمال بودن داده‌ها دارد.

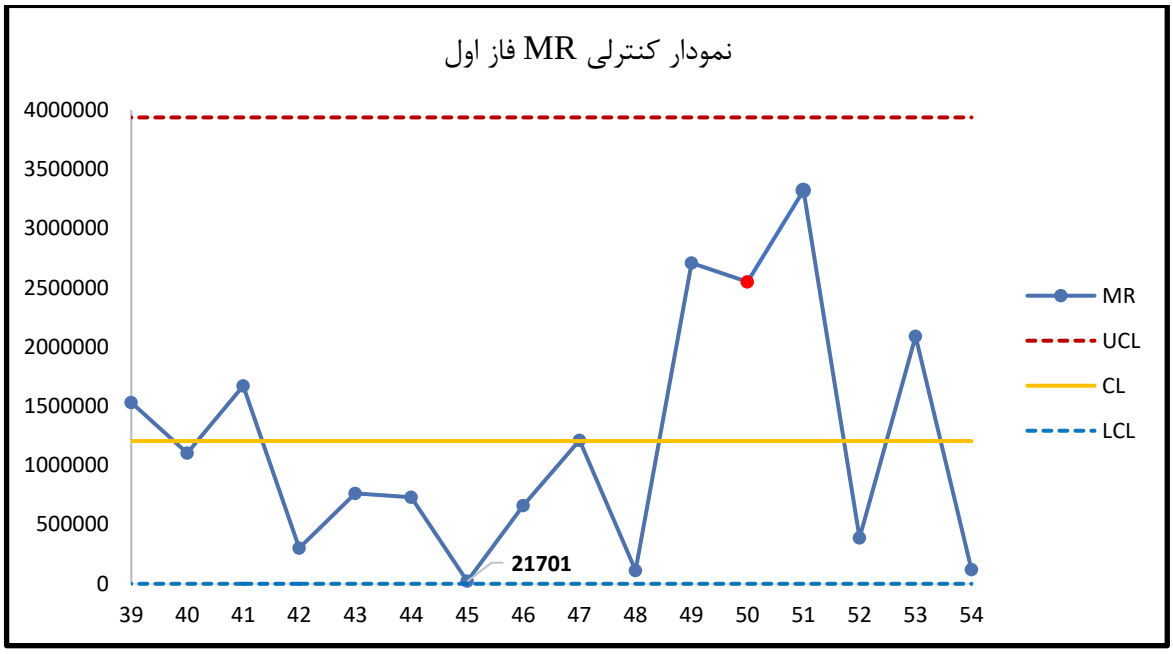


نمودار ۳: نمودار میله‌ای تعداد توئیت های هفتگی – بخش دوم دیتاست

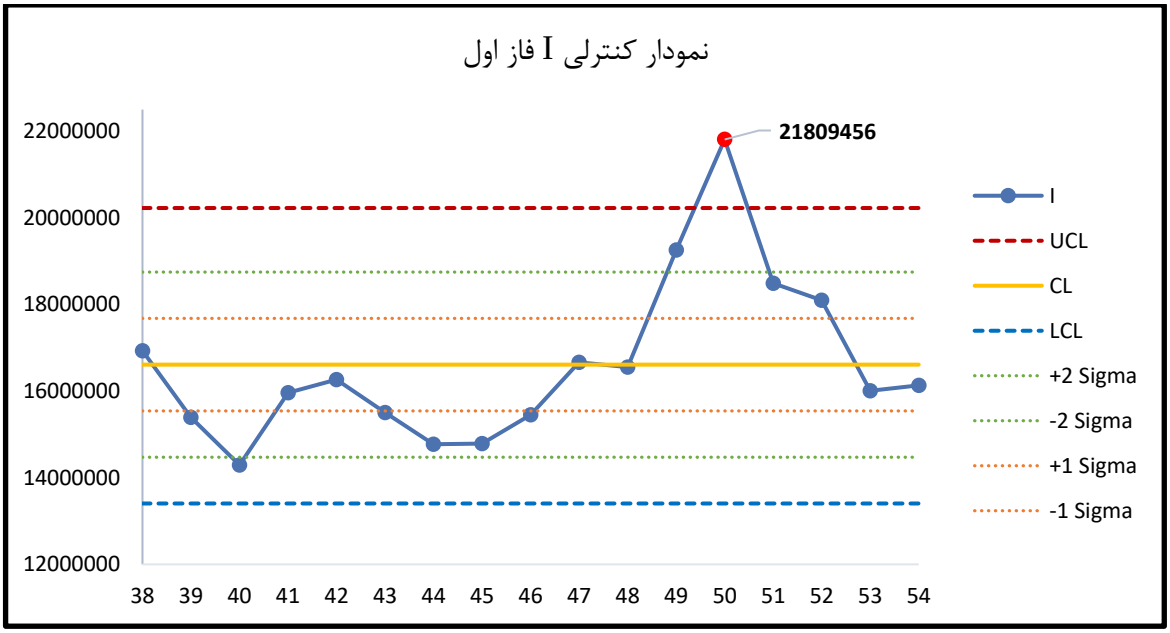
بررسی تعداد توییت در این ۴۴ هفته، نشان از روندی سینوسی دارد که به‌تنهایی، حاکی از عدم تحت کنترل بودن فرایند است؛ چراکه یکی از دلایل اصلی رد فرض تحت کنترل بودن یک فرآیند، مشاهده روند در نمودار است و باید دلیل این صعود و نزول در تعداد توییت‌ها را بررسی کنیم. اما جهت اطمینان بیشتر و اینکه صرفاً طبق مشاهدات نتیجه‌گیری نکرده باشیم، نمودار کنترلی نیز رسم می‌کنیم.

فاز اول نمودارها کنترلی MR و I

برای تشکیل فاز اول از داده‌های ۵۴ هفته اول که با احتمال بیشتری تحت کنترل هستند، استفاده می‌کنیم.
پس از محاسبه حدود کنترل، نمودارهای MR و I به شرح زیر هستند:



نمودار ۱۲: نمودار کنترلی MR برای فاز اول- بخش دوم دیتاست



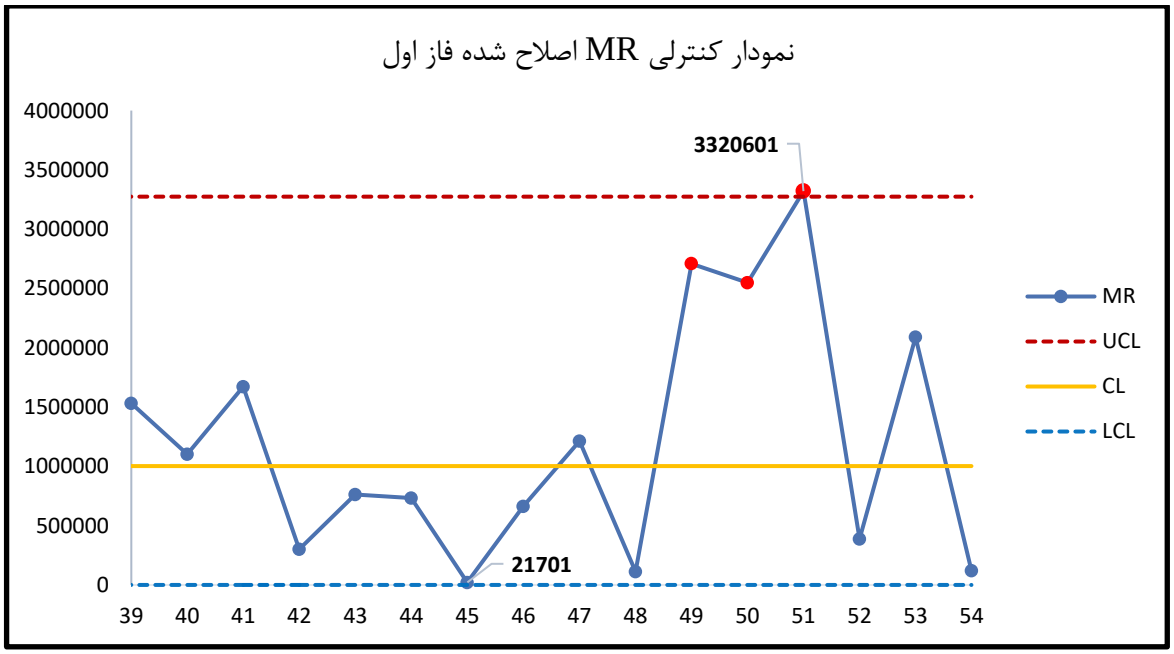
نمودار ۱۳: نمودار کنترلی I برای فاز اول- بخش دوم دیتاست

مشاهده می‌شود که در هفته ۵۰ام، نمونه‌ای خارج از حد UCL در نمودار I افتاده است. همچنین از هفته ۳۹ تا ۴۶، با توجه به قوانین حساس سازی وسترن الکتریک، ۸ نقطه متوالی در پایین CL هستند که یک

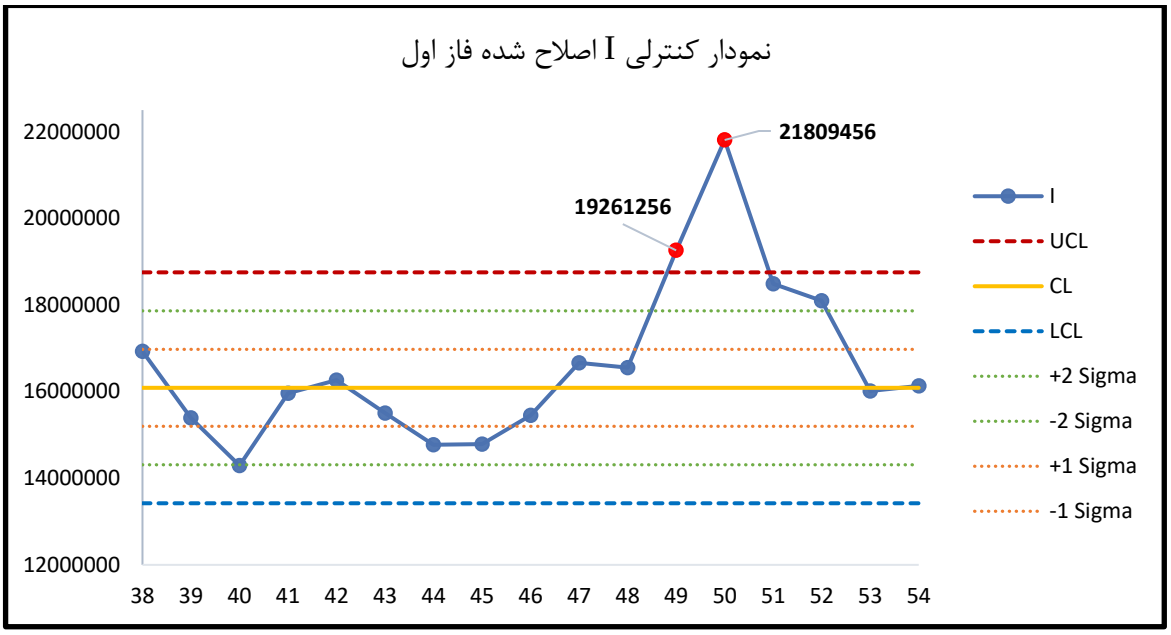
هشدار برای تحت کنترل نبودن فرآیند محسوب می‌شود. بنابراین باید حدودی که به دست آورده‌ایم را تصحیح کرده و صرفاً نمونه‌های خارج از حدود را روی نمودار نگهداریم.

بررسی جزئی دیتاست روزانه در هفته ۵۰ام، تعداد توییت‌های منتشرشده در ۲۹ آپریل ۲۰۲۱ را نزدیک به ۴،۸۰۰،۰۰ مورد بیان می‌کند. به همین علت، روزهای نزدیک به این روز را موردبررسی قرار داده تا علت هشدار را بیابیم.

با استناد به آمار منتشرشده از سازمان بهداشت جهانی (WHO) تعداد مبتلایان و جان‌باختگان به ویروس کرونا در این روز بیشترین مقدار خود را تابه‌حال داشته است. می‌توانید این داشبورد را ببینید (WHO Coronavirus (COVID-19) Dashboard, 2020). با توجه بیشتر به نمودار، متوجه می‌شویم که نقطه پیک این نمودار، هفته ۴۹ام دیتاست را نیز شامل می‌شود؛ به‌عبارت‌دیگر افزایش تعداد توییت در هفته ۴۹ام، به علت افزایش شمار جانباختگان و مبتلایان است. حال با شناسایی دلیل تحت کنترل نبودن نمونه ۴۹ و ۵۰ام، این دو نمونه را از محاسبات حدود کنترل خارج کرده و به محاسبه مجدد حدود کنترل می‌پردازیم. در ادامه نمودارهای MR و I را پس از حذف نمونه‌های ۴۹ و ۵۰ ملاحظه می‌کنید:



نمودار ۱۴: نمودار کنترلی MR اصلاح شده فاز اول



نمودار ۱۵: نمودار کنترلی I اصلاح شده فاز اول

با رسم حدود تصحیح‌شده می‌بینیم که نمونه ۵۱ام از حد UCL نمودار MR رد شده، اما با بررسی بیشتر دلیلی برای این اتفاق پیدا نمی‌کنیم، از طرف دیگر به دلیل اینکه فاصله از نمونه تا حد UCL زیاد نیست، از این تفاوت چشم‌پوشی کرده و حدود را اصلاح نمی‌کنیم و به فاز دوم می‌رویم.

حدود کنترلی فاز اول برای بخش دوم از قرار زیر است:

جدول ۴: حدود کنترل نمودار MR فاز اول- بخش دوم

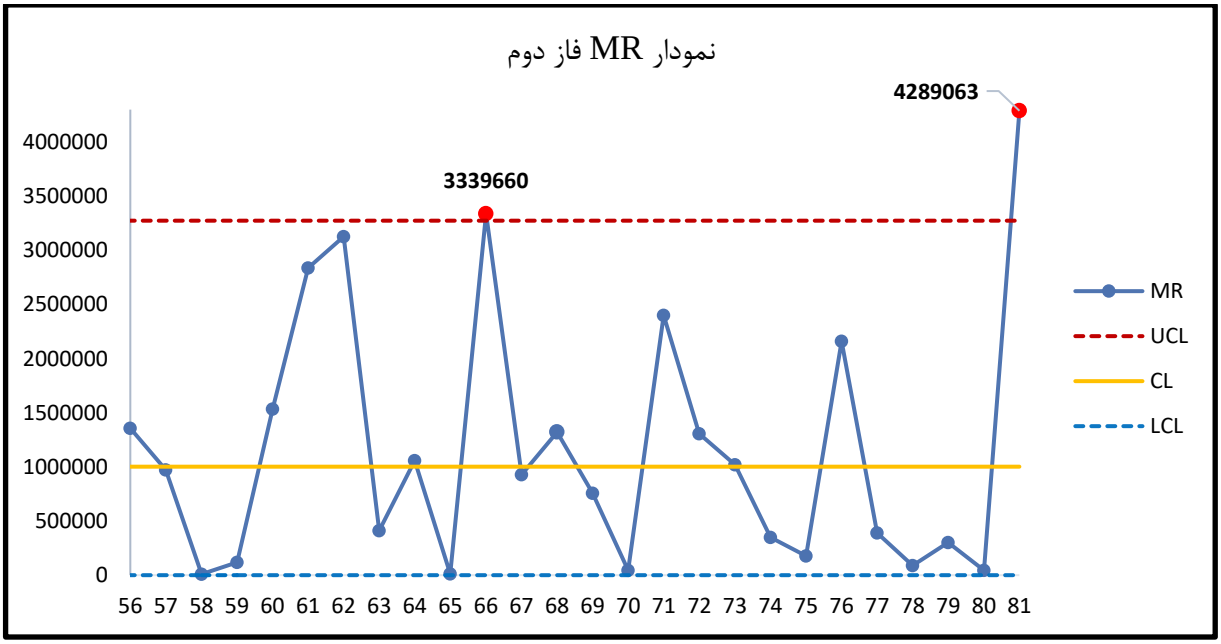
حدود کنترل چارت MR		
LCL	CL	UCL
۰	۱۰۰۲۲۲۳	۳۲۷۴۲۶۴

جدول ۵: حدود کنترل I فاز اول- بخش دوم

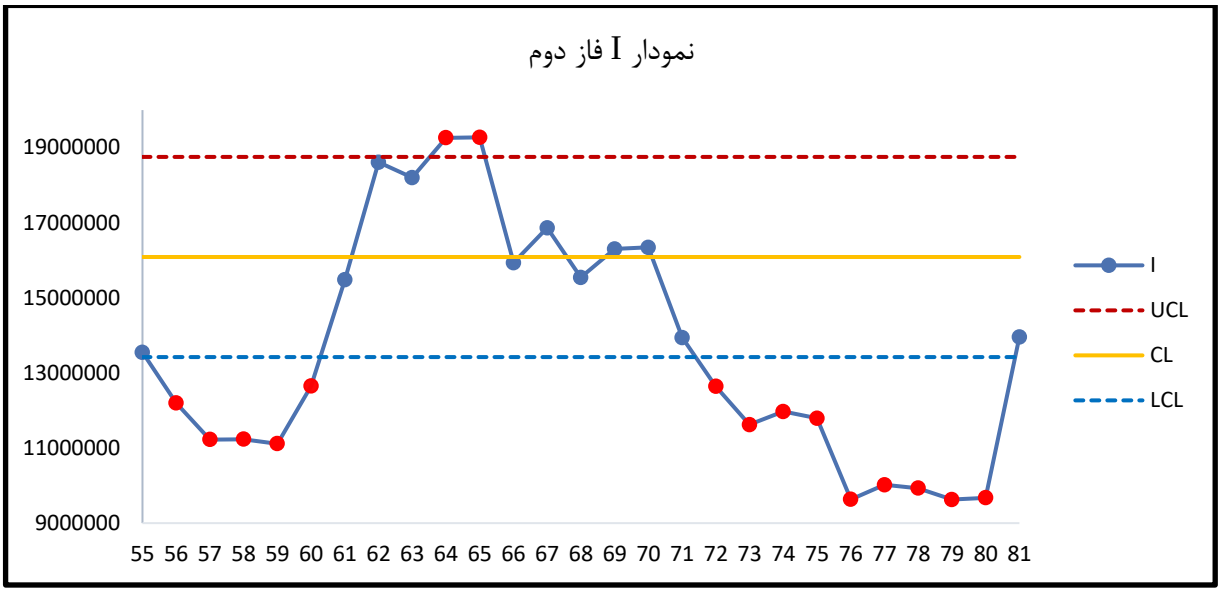
حدود کنترل چارت I		
LCL	CL	UCL
۱۳۴۲۲۴۹۱	۱۶۰۸۷۹۷۹	۱۸۷۵۳۴۶۷

فاز دوم نمودارهای کنترلی MR و I

طبق تحلیلی که در شروع بخش دوم داشتیم، انتظار داریم تعداد توییت‌های هفتگی برای این بخش تحت کنترل نباشد، حتی برای فاز اول نیز به سختی توانستیم حدود کنترلی را بپذیریم. با در نظر گرفتن حدود فاز اول برای فاز دوم که شامل هفته‌های ۵۵ تا ۸۱ می‌شود، داریم:



نمودار ۱۶: نمودار کنترلی MR برای فاز دو

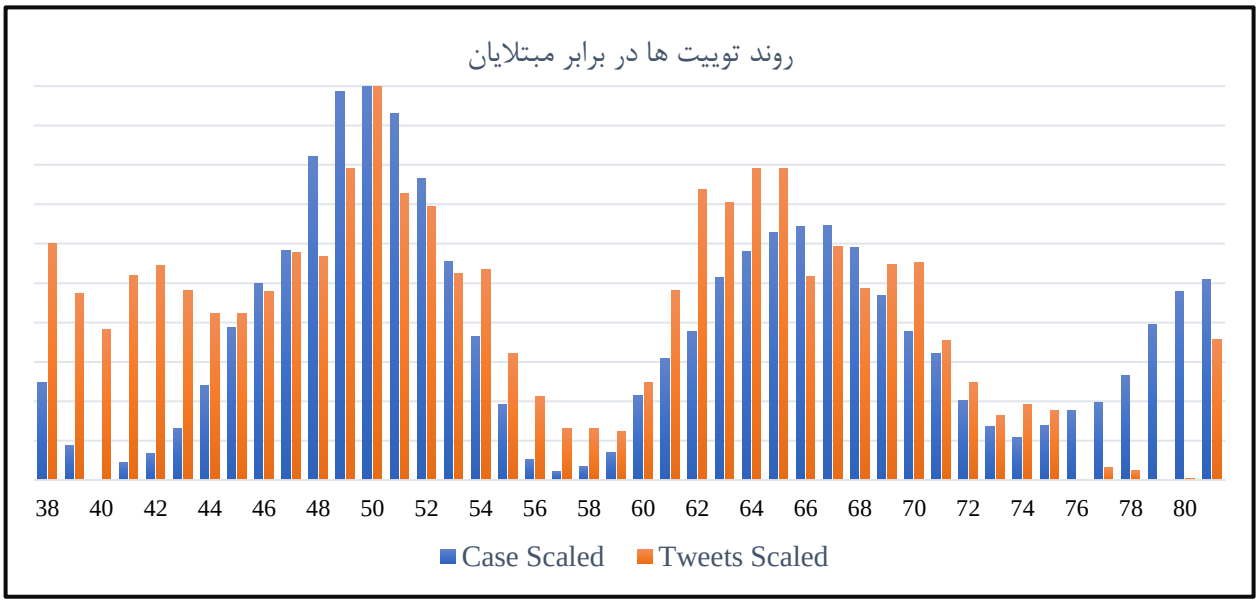


نمودار ۱۷: نمودار کنترلی I برای فاز دو

همان‌گونه که انتظار داشتیم، به علت روند سینوسی که در بخش دوم وجود دارد، نمونه‌های زیادی خارج از حدود افتاده‌اند که تعداد این نمونه‌ها، ۱۷ تا است و صرفاً ۱۰ نقطه در داخل حدود افتاده‌اند که همان ۱۰ نقطه نیز زیرمجموعه روند سینوسی هستند که پیش‌تر صحبت شد.

در گام بعد به پیدا کردن علت سینوسی بودن تعداد توییت‌ها از هفته ۳۸ به بعد می‌پردازیم. به این منظور، روند کلی تعداد توییت‌ها را با ابتدا با شمار مبتلایان و سپس با شمار فوتی‌ها که در داشبورد سازمان بهداشت جهانی قرارگرفته است، مقایسه می‌کنیم. دیتای مربوط به مجموع تعداد مبتلایان و جان‌باختگان به تفکیک روز در وب‌سایت این سازمان قرارگرفته است. با استفاده از این دیتا و همچنین به دست آوردن مجموع تعداد مبتلایان و جان‌باختگان کل کشورها، و پس‌ازآن به دست آوردن این مقادیر برای هر هفته پس از ۱۶ می ۲۰۲۰ با استفاده از نرم‌افزار پایتون، ترند تعداد جان‌باختگان و مبتلایان بعد از هفته ۳۸ام را با تعداد توییت‌های هفتگی هفته‌ی ۳۸ام به بعد مقایسه می‌کنیم.

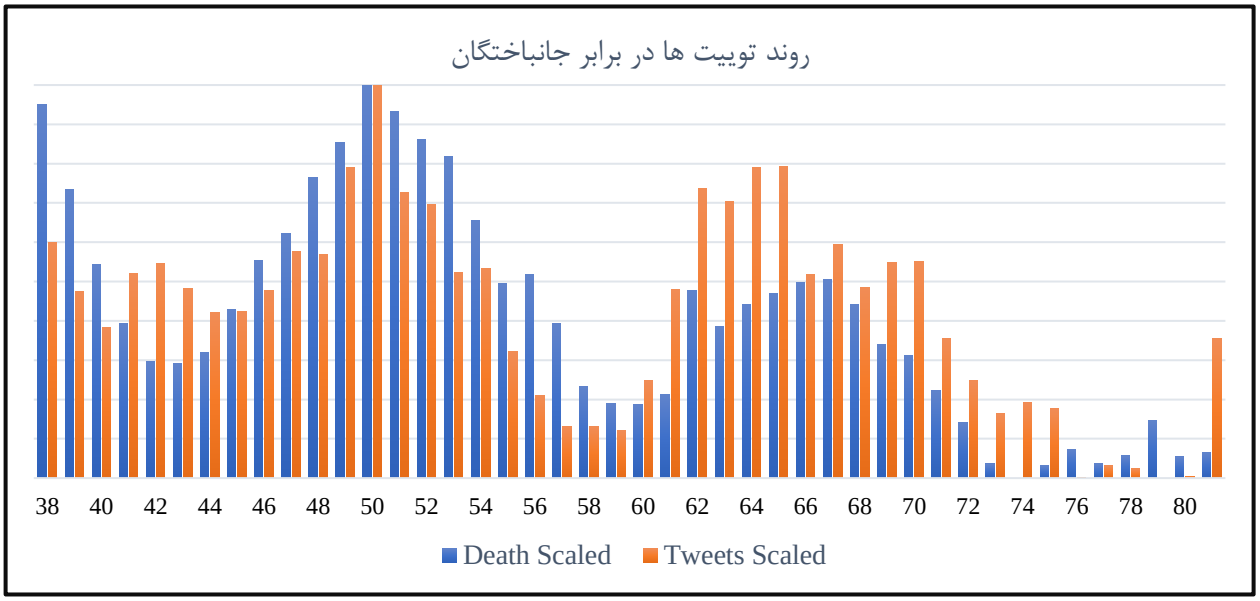
ابتدا با تعداد مبتلایان شروع می‌کنیم. برای مقایسه بهتر تعداد مبتلایان و تعداد توییت‌ها در نمودار میله‌ای، با استفاده از روش Min-Max Normalization، داده‌ها را اسکیل کرده و سپس نمودار میله‌ای را رسم می‌کنیم:



نمودار ۱۸: نمودار میله ای مقایسه روند مبتلایان با تعداد توئیت ها

همان‌طور که از نمودار مشهود است، در بسیاری موارد با افزایش تعداد مبتلایان کرونا، شمار توییت‌ها نیز افزایش میابد اما هفته‌هایی نیز وجود دارد که این موضوع را رد می‌کند. همچنین ضریب همبستگی بین این دو برابر ۶۵.۹۱٪ است که همبستگی خوبی را نشان می‌دهد.

اما با مشاهده رابطه‌ی بین تعداد جان‌باختگان اسکیل شده و توییت‌ها داریم:



نمودار ۱۹: نمودار میله‌ای مقایسه روند جانباختگان با تعداد توئیت‌ها

از این نمودار نتیجه می‌گیریم که روندهای سینوسی که در هر یک از نمودارها وجود دارد، کاملاً بر همدیگر منطبق هستند و با کاهش تعداد توییت در هفته، تعداد جان‌باختگان کاهش و با افزایش تعداد توییت، تعداد جان‌باختگان افزایش می‌یابد. با محاسبه همبستگی به عدد ۷۴.۷۹٪ می‌رسیم که به نسبت عدد بالایی می‌باشد و می‌توان تعداد توییت و جان‌باختگان را به‌شدت همبسته(Pearson Correlation, n.d.) در نظر گرفت.

موضوع بعدی که وجود دارد، این است که به‌صورت کلی، تعداد توییت‌های هفتگی بعد از هفته ۳۸ام، از ۳۷ هفته اول خیلی کمتر است به‌قدری که مجبور شدیم برای تحلیل و بررسی تعداد توییت‌ها، جایی که اختلاف میانگین ماکس می‌شود (که همان هفته ۳۷ام باشد) به‌عنوان جایی که میانگین عوض‌شده است در نظر بگیریم.

علتی که برای این موضوع یافتیم این بود که به‌صورت کلی و بدون در نظر گرفتن هفته‌هایی که مرگ‌ومیر به پیک خود رسیده بود، به علت تمام شدن قرنطینه و عادی شدن کرونا، مردم مدت‌زمان کمتری را در توییتر می‌گذرانند و نگرانی‌های سابق را در رابطه با کرونا ندارند.

هدف از مطالعه فوق بررسی ارتباط بین تعداد توییت های مرتبط با کرونا و وقایع اتفاق افتاده در دوران همه‌گیری کرونا بود.از بخش های قبل می توان نتیجه گرفت، تعداد توییت ها و اتفاقات، ارتباط نزدیکی بایکدیگر داشته و می توان از روی تغییرات یکی از آنها، تغییرات دیگری را تا حدودی پیش‌بینی کرد.

منابع

(n.d.). Retrieved from https://covid19.who.int/

Allen, J., & Lipsitch, M. (2020, October 8). *Americans, we can fight COVID-19 and save lives now. Wear a mask!* Retrieved from usatoday: https://www.usatoday.com/story/opinion/2020/10/08/wear-mask-fight-covid-19-and-save-lives-now-medical-experts-column/5907452002/

Andersen, K., Medaglia, R., & Henriksen, H. (2012). Social Media in Public Health Care: Impact Domain Propositions. *Government Information Quarterly*, 462-469.

Big Crisis Data: Social Media in Disasters and Time-Critical Situations. (2016). Cambridge University Press.

Branswell, H. (2020, December 2). *The Covid-19 vaccines are a marvel of science. Here's how we can make the best use of them*. Retrieved from statnews: https://www.statnews.com/2020/12/02/how-society-can-make-the-most-of-covid-19-vaccines/

Chen Lyu, J., Han, E. L., & Luli, G. K. (2021). COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis . *J Med Internet Res*, 26.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 1-38.

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Information Processing & Management*, 102-261.

Kwona, J., & Grady, C. (2020). Defining facets of social distancing during the COVID-19 pandemic: Twitter analysis. *Journal of Biomedical Informatics*.

Lamsal, R. (2020, 10 12). *Coronavirus (COVID-19) Tweets Dataset*. Retrieved from IEEEDataport: https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 2790–2804.

Lazzar, K. (2020, October 2). *Trump's diagnosis is an indictment of his COVID-19 response*. Retrieved from bostonglobe: https://www.bostonglobe.com/2020/10/02/nation/heres-what-epidemiologists-are-saying-about-president-trump-getting-coronavirus/

loffe, J. (2020, November 10). *The virus doesn't care who won the presidential election*. Retrieved from washingtonpost: https://www.washingtonpost.com/outlook/2020/11/10/biden-politicized-virus-trump/?arc404=true

منابع

N. Bleich, S. (2020, June 4). *Protest, demand change — but, please, put on your mask*. Retrieved from TheHill: https://thehill.com/opinion/healthcare/501222-protest-demand-change-but-please-put-on-your-mask

Pearson Correlation. (n.d.). Retrieved from Sciencedirect: https://www.sciencedirect.com/topics/computer-science/pearson-correlation/

Thys, F. (2020, June 5). *To Reopen Campus, Colleges Prepare To Take On Contagious Students*. Retrieved from wbur: https://www.wbur.org/news/2020/06/05/college-plan-contagious-students

Wan, H., Li, Y., Hutch, M., Naidech, A., & Luo, Y. (2021). Using Tweets to Understand How COVID-19–Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study . *J Med Internet Res*.

Weintraub, K. (2020, October 2). *'Not completely out of the woods': Biden's negative test doesn't mean he can't be positive in coming days*. Retrieved from usatoday: https://www.usatoday.com/story/news/2020/10/02/bidens-covid-19-test-friday-doesnt-mean-clear-coronavirus/3593167001/

Weintraub, K. (2020, November 10). *There may be a COVID-19 vaccine by the end of the year, but 'normality' may not come until end of 2021*. Retrieved from usatoday: https://www.usatoday.com/story/news/health/2020/11/10/covid-19-vaccine-willingness-needed-to-end-pandemic/3516649001/

Weintraub, K. (2020, October 7). *What can we expect from a winter COVID-19 second wave? No one knows for sure, but there is reason for hope and concern*. Retrieved from usatoday: https://www.usatoday.com/story/news/health/2020/10/07/covid-winter-second-wave-may-bad-experts-warn-continued-vigilance/5873746002/

Wergieluk, J. (2020, April 30). *Histograms vs. KDEs Explained*. Retrieved from Towardsdatascience: https://towardsdatascience.com/histograms-vs-kdes-explained-ed62e7753f12

WHO Coronavirus (COVID-19) Dashboard. (2020). Retrieved from WHO: https://covid19.who.int/

WHO Director. (2020, March 11). General’s opening remarks at the media briefing on COVID19.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, 10283-10291.