به نام خدا

تمرین اول داده‌کاوی پیشرفته

مهدی فقهی

۴۰۱۷۲۲۱۳۶

Problem 1

Indicate with reasons if each of the following tasks is a Data Mining task or not

If they are ,classify what kind of a data mining task is

1. Dividing a company's customers by gender.

تسک داده کاوی نیست. این فقط شامل دستکاری و دسته بندی داده های اساسی بر اساس یک ویژگی است.

2. Dividing the customers of a company according to their profitability.

بله، این یک کار داده کاوی است. نوع وظیفه داده کاوی: طبقه بندی یا classification است.

3. Count how many emails are tagged with spam.

این یک کار داده کاوی نیست. این یک عملیات شمارش ساده است که شامل تجزیه و تحلیل الگوها یا روابط در داده ها نمی شود

4. Calculating the total sales of a company.

این یک کار داده کاوی نیست زیرا یک محاسبه ساده است که می تواند با استفاده از عملیات حسابی اساسی انجام شود. داده کاوی شامل کشف الگوها و روابط در داده ها است که ممکن است به راحتی با استفاده از روش های سنتی قابل مشاهده نباشد.

5. Predicting the future stock price of a company using past records.

این یک کار داده کاوی است و به تحت عنوان predictive modeling است . هدف پیش بینی رویدادهای آینده بر اساس داده های گذشته است. در این مورد، داده های گذشته، قیمت سهام گذشته شرکت است. این کار شامل تجزیه و تحلیل این داده ها و ایجاد یک مدل پیش بینی کننده است که می تواند قیمت سهام آینده را پیش بینی کند.

6. Find the answer to each question by analyzing the corresponding image.

وظیفه داده کاوی نیست. این وظیفه تجزیه و تحلیل تصویر برای استخراج اطلاعات از یک تصویر است. این به داده کاوی مربوط نمی شود زیرا هیچ مجموعه داده ساختاریافته یا بدون ساختاری در تجزیه و تحلیل وجود ندارد.

7. Monitoring the heart rate of a patient for abnormalities.

این یک کار داده کاوی نیست. این یک وظیفه نظارتی برای اهداف بهداشتی است که شامل جمع آوری داده های فیزیولوژیکی در طول زمان از یک بیمار با استفاده از یک دستگاه پزشکی است. هدف شناسایی هرگونه انحراف از محدوده طبیعی و انجام اقدامات مناسب، مانند هشدار دادن به ارائه دهنده مراقبت های بهداشتی یا شروع یک فوریت پزشکی است.

8. Monitoring seismic waves for earthquake activities.

این یک کار داده کاوی نیست. این وظیفه به جای تجزیه و تحلیل و استخراج الگوها یا بینش ها از داده ها، شامل نظارت بر پدیده های فیزیکی در جهان طبیعی است.

9. Determining the association rules of market transactions.

این جز وظایف داده کاوی است که به عنوان association rule mining شناخته می شود. این شامل کشف ارتباط و روابط بین موارد مختلف در یک مجموعه داده است، در این مورد، معاملات بازار. برای شناسایی الگوها و کشف بینش‌های پنهانی که می‌توان از آنها برای بهبود تصمیم گیری تجاری استفاده کرد، مفید است.

10. Summarization of arguments about a certain topic.

این جز وظایف داده کاوی نیست، زیرا شامل تجزیه و تحلیل یک مجموعه داده بزرگ برای کشف الگوها یا روابط نیست. بلکه شامل خلاصه کردن استدلال های موجود در مورد یک موضوع خاص است که جز زیرمجموعه‌های NLP است .

11. Sorting a student database based on student identification numbers.

داده کاوی نیست. این یک کار مرتب سازی ساده بر اساس یک ویژگی است و شامل کشف الگوها یا روابط در داده ها نمی شود.

12. Predicting the outcomes of tossing a (fair) pair of dice.

این یک کار داده کاوی نیست. این یک مسئله نظریه احتمال است.

13. Extracting the frequencies of a sound wave.

این کار داده کاوی نیست. استخراج فرکانس های یک موج صوتی در حوزه پردازش سیگنال قرار می گیرد. این شامل تجزیه و تحلیل و تبدیل سیگنال در حوزه زمان به حوزه فرکانس است.

14. Save the identification numbers in a database.

.این کار داده کاوی نیست بلکه یک فعالیت ذخیره سازی داده است

15. Look up the phone number in the phone directory.

این کار داده کاوی نیست. این یک کار جستجوی ساده است که در آن شماره تلفن در فهرست تلفن جستجو می شود.

16. Query a Web search engine for information about "Iran University of Science and Technology".

این وظیفه یک کار داده کاوی نیست. این یک کار ساده بازیابی اطلاعات است که می تواند با جستجو در موتور جستجو انجام شود.

17. Certain names are more prevalent in certain US locations.

به صورت مستقیم کار داده کاوی نیست، بلکه یک کار تجزیه و تحلیل آماری است. این شامل تجزیه و تحلیل داده های جمعیتی موجود برای شناسایی ارتباط بین نام های خاص و مکان های خاص ایالات متحده است. بنابراین، می توان آن را به عنوان یک کار Descriptive Data Mining task طبقه بندی کرد که شامل یافتن الگوها و روابط درون داده های موجود است.

18. Group together similar documents returned by the search engines according to their context..

این یک کار داده کاوی است و یک Text Mining task است . این کار شامل گروه بندی اسناد مشابه بر اساس زمینه آنها است که به تجزیه و تحلیل داده های متنی نیاز دارد. متن کاوی شکلی از داده کاوی است که شامل فرآیند تجزیه و تحلیل و استخراج اطلاعات و بینش ارزشمند از داده های متنی است. در این مورد، این کار مستلزم استفاده از تکنیک های پردازش زبان طبیعی (NLP) برای شناسایی شباهت ها در زمینه اسناد و گروه بندی آنها بر اساس آن است.

Problem 2

Classify the following attributes as binary, discrete, or continuous. Also, classify them with reasons as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio because it has all four properties.

a. Time in terms of AM or PM.

Binary and qualitative (nominal).

این ویژگی را می توان به عنوان باینری طبقه بندی کرد زیرا فقط دو مقدار ممکن دارد، یا AM یا PM. علاوه بر این، کیفی است زیرا مقادیر عددی نیستند، بلکه برچسب هایی هستند که برای نشان دادن دوره های زمانی استفاده می شوند.

b. Brightness as measured by a light meter.

Continuous, quantitative (ratio)

روشنایی که توسط نورسنج اندازه گیری می شود، یک مقدار عددی دقیق و قابل اندازه گیری را در مقیاس مداوم ارائه می دهد. یک ratio scale است زیرا دارای یک نقطه صفر معنی دار (تاریکی کامل) است.

c. Brightness as measured by people's judgments.

Attribute: Discrete

Type of attribute: Continuous

Qualitative/Quantitative: Qualitative (ordinal)

روشنایی یک ویژگی پیوسته است زیرا می تواند هر مقداری را در یک محدوده به خود بگیرد. با این حال، با قضاوت افراد سنجیده می شود، که ذهنی است و می تواند از فردی به فرد دیگر متفاوت باشد ولی این توضیحات معمولا با چند صفت خاص بیشتر قابلیت توضیح ندارند و نمی‌توانند آن پیوستگی که داده‌دارد را لحاظ کند لذا به صورت گسسته با چند صفت شناخته می‌شوند توسط انسان‌ها . بنابراین داده های کیفی است. قضاوت‌ها ترتیبی هستند، زیرا می‌توانند به ترتیب رتبه‌بندی شوند (به عنوان مثال روشن‌تر، روشن، نه خیلی روشن).

d. Satisfaction of customers. (Unsatisfied, Neutral, Satisfied)

Binary, qualitative (nominal)

رضایت مشتریان یک ویژگی باینری است زیرا تنها دو نتیجه ممکن دارد (ناراضی یا راضی). همچنین یک ویژگی qualitative است زیرا دسته ها هیچ ترتیب ذاتی یا ارزش عددی ندارند.

e. Bronze, Silver, and Gold medals as awarded at the Olympics.

Attribute: Discrete

Qualitative: Ordinal

چندتا اسم گسسته داریم که بین آن‌ها بر حسب نوع مدل یک نوع ترتیب و جایگاه وجود دارد .

f. Height of a person. (tall, short)

Attribute :Binary

Qualitative: Ordinal (tall, short)

اگر فقط دو تا اسم tall , short را داشته باشیم یک سری صفت کیفی هستند ، می‌توانیم به صورت Binary در نظرشان بگیریم چون یک نفر یا tall است یا short و از آن جهت که پشت این صفت‌ها یک ترتیب است می‌توانیم آن را به صورت ordinal در نظر بگیریم .

g. Height of a person as measured by centimeters.

Classification: Continuous

quantitative : ratio

ارتفاع یک ویژگی پیوسته است زیرا می تواند هر مقداری را در یک محدوده (یعنی بین 0 تا 300 سانتی متر) بگیرد. هیچ گسست طبیعی یا نقطه برشی در مقیاس اندازه گیری وجود ندارد.

همچنین می توان آن را به عنوان یک ویژگی کمی طبقه بندی کرد زیرا یک اندازه گیری عددی است.

همچنین :

it could be classified as a ratio attribute because there is an inherent zero point (i.e., no height) that indicates a lack of the attribute being measured

h. The temperature of a home. (cold, hot, warm)

Type: Discrete

Classification: Qualitative, Ordina

i. The temperature of a home as measured by a thermometer.j. Military rank.

Type: Discrete,

Classification :qualitative, ordinal

درجه نظامی یک متغیر طبقه‌بندی است که از تعداد محدودی دسته‌بندی مجزا تشکیل شده است مقوله های درجه نظامی مانند یک سلسله مراتب یا نظم است و بنابراین یک متغیر ترتیبی است.

k. University rank.

Type: Discrete,

Classification :qualitative, ordinal

رتبه دانشگاه فقط می تواند تعداد محدودی از مقادیر را بگیرد (گسسته)، و رتبه بندی را می توان مرتب کرد (ترتیبی).

l. Angles as measured in degrees between 0◦ and 360◦.

Type: Continuous

Qualitative/Quantitative: Quantitative (ratio)

زاویه ها می توانند هر مقداری بین 0 تا 360 درجه داشته باشند و با هر درجه دقت قابل اندازه گیری هستند. بنابراین یک متغیر پیوسته است. علاوه بر این، درجه ها دارای یک نقطه صفر ثابت هستند (زاویه ای به معنی 0 درجه است) و تفاوت بین مقادیر معنی دار است و آن را به یک متغیر نسبت تبدیل می کند.

m. Height above sea level.

Type of Attribute: Continuous.

Type of Measurement: Quantitative Ratio.

ارتفاع از سطح دریا یک ویژگی پیوسته است زیرا می تواند هر مقدار عددی را در یک محدوده داشته باشد
ویژگی ارتفاع از سطح دریا را می توان در مقیاس نسبت اندازه گیری کرد زیرا دارای یک نقطه صفر واقعی است (یعنی سطح دریا). نسبت ها را نیز می توان محاسبه کرد، به عنوان مثال، یک نقطه دو برابر ارتفاع نقطه دیگر است اگر مقدار آن دو برابر باشد.

n. Number of patients in a hospital.

Type of Attribute: Continuous.

Qualitative/Quantitative classification: Quantitative (Ratio)

تعداد بیماران در یک بیمارستان می تواند هر عدد صحیح مثبت باشد. علاوه بر این، ویژگی دارای یک نقطه صفر واقعی است، زیرا مقدار صفر نشان دهنده غیبت کامل بیماران است و آن را به یک متغیر نسبت کمی تبدیل می کند.

o. ISBN numbers for books. (Look up the format on the Web.)

- گسسته: هر شماره ISBN منحصر به فرد است و در مقیاس پیوسته قرار نمی گیرد. علاوه بر این، تعداد ارقام ثابت و قابل شمارش است.

qualitative (nominal)

- کیفی: اسمی: هر شماره ISBN یک شناسه منحصر به فرد است و دلالت بر ترتیب یا بزرگی ندارد.
- در این مورد تفسیر ترتیبی وجود ندارد زیرا اعداد شابک دارای رتبه و ترتیب ذاتی نیستند.
- در این مورد هیچ تفسیر فاصله ای وجود ندارد زیرا هیچ نقطه صفر یا واحد اندازه گیری معناداری برای اعداد شابک وجود ندارد.
- در این مورد هیچ تفسیر نسبتی وجود ندارد زیرا نقطه صفر یا واحد اندازه گیری معناداری برای اعداد شابک وجود ندارد.

p. Ability to pass light in terms of the following values: opaque, translucent, transparent.

Attribute: Ability to pass light

Classification: Categorical -

Type: Binary (with three categories: opaque, translucent, transparent)

Nature: Qualitative (ordinal)

ویژگی "توانایی عبور نور" یک متغیر طبقه بندی است، زیرا دسته بندی های مجزا از اشیاء را بر اساس سطح شفافیت آنها توصیف می کند. گسسته است زیرا دارای سه دسته انحصاری و جامع است. این متغیر کیفی است زیرا دسته ها عددی نیستند بین توانایی اجزا برای عبور دادن نور یک ترتیب وجود دارد بین این قدرت عبور .

q. Distance from the center of campus.

type: Continuous

Quantitative : numerical

فاصله می تواند هر مقداری را در مقیاس پیوسته بگیرد، می توان آن را به واحدهای بسیار کوچک مانند اینچ یا سانتی متر اندازه گرفت.

فاصله کمی است زیرا یک کمیت قابل اندازه گیری است و می تواند به عنوان یک مقدار عددی بیان شود. علاوه بر این، فاصله می تواند فاصله باشد زیرا اندازه گیری از یک نقطه دلخواه (مرکز دانشگاه) شروع می شود و هیچ نقطه صفر واقعی وجود ندارد.

r. Density of a substance in grams per cubic centimeter.

Type: Continuous

Quantitative: Ratio

چگالی می تواند هر مقداری را در محدوده مشخصی به خود بگیرد و با استفاده از ابزارهای علمی با دقت زیادی اندازه گیری شود. گرم و سانتی متر مکعب هر دو واحد اندازه گیری هستند که مقادیر دقیق قابل اندازه گیری را نشان می دهند. نسبت بین دو اندازه گیری در این واحدها معنادار است (یعنی وزن در واحد حجم یک مقدار دقیق است). به طور کلی، این ویژگی دارای تفسیر کمی، پیوسته و مقیاس نسبت است.

s. Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

Type: Discrete (با فرض اینکه هر عدد چک کت یک عدد صحیح باشد)

Qualitative vs Quantitative: Qualitative (Nominal)

شماره چک کت یک شناسه منحصربه‌فرد است که به هر کتی که تحویل داده می‌شود اختصاص داده می‌شود. گسسته است زیرا مجموعه محدودی از مقادیر عددی (اعداد کامل) را به خود می گیرد و کیفی است زیرا مبتنی بر اندازه گیری عددی نیست، بلکه یک مقدار مقوله ای است که به عنوان یک شناسه عمل می کند.

Problem 3

What is the difference between the definition of classification and clustering? Name four applications of classification and two applications of clustering with their goal and approach.

طبقه بندی فرآیند دسته بندی داده ها به گروه ها یا کلاس های از پیش تعریف شده بر اساس ویژگی ها یا شناسه های از پیش تعریف شده است. در مقابل، خوشه‌بندی فرآیند گروه‌بندی مجموعه‌های داده خام بر اساس شباهت‌ها یا عدم شباهت‌های آن‌ها بدون دانش قبلی از گروه‌ها یا کلاس‌ها است.

طبقه بندی یک تکنیک یادگیری نظارت شده است که در آن یک الگوریتم یاد می گیرد که داده ها را بر اساس مجموعه ای از نمونه های برچسب گذاری شده به دسته های از پیش تعریف شده طبقه بندی کند. هدف از طبقه بندی، گروه بندی داده ها به دسته ها یا کلاس های مجزا، با استفاده از مجموعه ای از قوانین از پیش تعیین شده است. رویکرد آن مبتنی بر یادگیری نظارت شده است، جایی که مدل بر روی داده‌های برچسب گذاری شده آموزش داده می‌شود و سپس روی داده‌های جدید و بدون برچسب اعمال می‌شود. خوشه بندی یک تکنیک یادگیری بدون نظارت است که در آن یک الگوریتم خوشه ها را در داده ها بر اساس شباهت های بین مشاهدات مختلف شناسایی می کند. هدف از خوشه‌بندی، شناسایی گروه‌های طبیعی درون داده‌ها، بدون کلاس‌ها یا برچسب‌های از پیش تعیین‌شده است. رویکرد آن مبتنی بر یادگیری بدون نظارت است، جایی که مدل بر روی خود داده ها آموزش داده می شود، بدون هیچ گونه دانش قبلی از دسته ها.

چهار کاربرد طبقه بندی عبارتند از:

1. فیلتر کردن ایمیل: هدف از فیلتر کردن ایمیل، طبقه بندی ایمیل ها به دسته های از پیش تعریف شده، مانند ایمیل های اسپم، تبلیغاتی یا شخصی است. این رویکرد می‌تواند از الگوریتم‌های مبتنی بر قانون تا الگوریتم‌های یادگیری ماشینی متفاوت باشد.

2. تشخیص پزشکی: هدف تشخیص پزشکی، طبقه بندی بیماران به بیماری ها یا شرایط خاص است. این رویکرد می تواند از تکیه بر علائم بالینی تا استفاده از تصویربرداری یا داده های ژنتیکی متفاوت باشد.

3. تشخیص تصویر: هدف از تشخیص تصویر، طبقه بندی تصاویر به دسته های از پیش تعریف شده مانند اشیا، صحنه ها یا چهره ها است. الگوریتم های یادگیری عمیق معمولا برای این برنامه استفاده می شود.

4. Fraud detection - معاملات را بر اساس مجموعه ای از قوانین از پیش تعریف شده به عنوان مشروع یا تقلبی دسته بندی می کند.

دو کاربرد خوشه بندی عبارتند از:

1. Market segmentation: هدف از تقسیم‌بندی بازار، گروه‌بندی مشتریان به بخش‌هایی بر اساس شباهت‌هایشان در رفتار، ترجیحات یا جمعیت‌شناسی است. این رویکرد می تواند از استفاده از الگوریتم های خوشه بندی تا تخصص دامنه متفاوت باشد.

2. Social network analysis: هدف از تجزیه و تحلیل شبکه های اجتماعی، گروه بندی افراد در جوامع بر اساس تعاملات اجتماعی، علایق یا نظرات آنهاست. این رویکرد می تواند از استفاده از الگوریتم های مبتنی بر شبکه تا پردازش زبان طبیعی متفاوت باشد.

Problem 4

Name all types of sampling and explain their advantages and disadvantages.

Is random sampling without replacement a good method? Explain your answer.

1.Simple Random Sampling

نمونه‌گیری تصادفی ساده، روشی آماری برای انتخاب نمونه از جامعه‌ای است که هر یک از اعضای آن شانس مساوی برای قرار گرفتن در نمونه را دارند. این روش به دلیل مزایا و معایب آن یک روش نمونه گیری پرکاربرد در تحقیقات است.

مزایای:

- **unbiased:**

هر یک از اعضای جامعه شانس مساوی برای انتخاب شدن دارند،
بنابراین نتیجه تحت تاثیر ترجیحات محقق قرار نمی گیرد.

- **representative:**

نمونه گیری تصادفی ساده تضمین می کند که نمونه به دست آمده بازتاب واقعی جامعه است
و تعمیم نتایج به جمعیت بزرگتر را آسان تر می کند.

- **easy to implement:**

این روش برای پیاده سازی ساده است و نیازی به دانش یا تخصص خاصی ندارد.

- **ensures accuracy:**

نمونه گیری تصادفی ساده احتمال خطا در فرایند انتخاب نمونه را کاهش می دهد
و در نتیجه از دقت بیشتر نتایج مطالعه اطمینان می دهد.

معایب:

- **can be costly:**

نمونه‌برداری تصادفی ساده مستلزم جامع بودن چارچوب نمونه‌گیری است، که می‌تواند پرهزینه و زمان‌بر باشد.

- **can be impractical:**

اگر حجم جامعه بزرگ باشد، انتخاب یک نمونه تصادفی ساده می تواند غیر عملی باشد
و به منابع قابل توجهی برای به دست آوردن حجم نمونه مورد نیاز نیاز دارد.

- **could lead to under-representation:**

نمونه‌گیری تصادفی ساده در صورتی که دسترسی به برخی از اعضای جامعه سخت باشد کارساز نخواهد بود،
که منجر به حضور کمتر برخی از گروه‌ها می‌شود.

- **may not capture the variability:**

نمونه‌گیری تصادفی ساده ممکن است تغییر پذیری(variability) در جامعه را نشان ندهد، زیرا نمونه‌های انتخاب شده ممکن است نماینده زیرگروه‌های مختلف جامعه نباشند.

2.Sampling without replacement

مزایای:

- نمونه نماینده جامعه است زیرا هر عضو شانس مساوی برای انتخاب شدن دارد.
- این اطمینان را ایجاد می کند که هر عنصر فقط یک بار انتخاب شده است، و بنابراین، از تکرار جلوگیری می کند.

معایب:

- نمونه برداری بدون جایگزینی می تواند پیچیده تر و زمان برتر از نمونه برداری با جایگزینی باشد.
- اگر جامعه همگن نباشد و نمونه برداری بدون جایگزینی انجام شود، ممکن است خطر سوگیری در نمونه وجود داشته باشد. در چنین مواردی، نمونه ممکن است به طور دقیق ویژگی های جامعه را منعکس نکند.

3.Sampling with replacement

این روش معمولاً در آمار و تحقیقات، به ویژه در شرایطی که نمونه کوچک است یا جمعیت زیاد است، استفاده می شود.

مزایای:

- روش نمونه گیری ساده و سرراست است که به تخصص فنی کمی نیاز دارد.
- می توان آن را برای نمونه هایی با هر اندازه ای اعمال کرد و آن را به یک تکنیک همه کاره تبدیل می کند که می تواند در مطالعات کوچک و بزرگ استفاده شود.
- امکان نمونه گیری مکرر را فراهم می کند، به این معنی که یک مورد می تواند چندین بار انتخاب شود. این می تواند در انواع خاصی از مطالعات که مشاهدات مکرر لازم است مفید باشد.

معایب:

- نمونه ممکن است به طور دقیق جامعه را نشان ندهد، زیرا برخی موارد ممکن است به دلیل فرآیند انتخاب تصادفی بیش از حد یا کمتر ارائه شوند.
- می تواند منجر به خطای نمونه برداری بیشتری شود، به این معنی که نمونه ممکن است از دقت کمتری نسبت به نمونه ای که از طریق روش های دیگر به دست آمده است، باشد.
- می تواند نتیجه گیری معنادار را دشوارتر کند، زیرا انتخاب مکرر موارد مشابه ممکن است نتایج را منحرف کند.
- ممکن است برای مطالعاتی که حجم نمونه کوچک است مناسب نباشد، زیرا انتخاب مکرر اقلام می تواند منجر به همپوشانی بیشتر بین نمونه ها شود.

4.Stratified sampling

زمانی مناسب است که می خواهید اطمینان حاصل کنید که ویژگی های خاص به طور متناسب در نمونه نشان داده شده است. شما جمعیت خود را به اقشار تقسیم می کنید (به عنوان مثال، تقسیم بر جنسیت یا نژاد)، و سپس به طور تصادفی از هر یک از این زیر گروه ها انتخاب می کنید.

مزایای:

- نمونه گیری طبقه ای تضمین می کند که نمونه نماینده جمعیت بزرگتر همگن است .
- این روش سوگیری نمونه گیری را کاهش می دهد که می تواند زمانی رخ دهد که گروه های خاصی در نمونه بیش از حد یا کمتر حضور داشته باشند.
- دقت تخمین ها را می توان افزایش داد اگر اقشار به خوبی تعریف شده باشند و اندازه نمونه از هر طبقه متناسب با اندازه لایه باشد.
- به استفاده بهینه از منابع کمک می کند زیرا حجم نمونه از هر زیرگروه متناسب با اندازه زیرگروه است.

معایب:

- نیاز به دانش قبلی از زیر گروه هایی دارد که جمعیت را تقسیم می کنند و اگر این زیر گروه ها به خوبی تعریف نشده باشند، نتایج ممکن است نادرست باشد.
- نمونه برداری طبقه ای می تواند زمان بر و پرهزینه باشد، زیرا مستلزم ایجاد زیرگروه ها و به دست آوردن نمونه از هر زیرگروه است.
- اطمینان از اینکه نمونه های به دست آمده از هر زیرگروه نماینده جامعه هستند می تواند دشوار باشد.

5.Systematic sampling :

نمونه گیری سیستماتیک یک تکنیک آماری است که در آن یک نقطه شروع تصادفی انتخاب می کنیم و سپس هر kth عنصر را از جامعه/نمونه به عنوان نمونه نماینده انتخاب می کنیم. به عنوان مثال، اگر نیاز به نظرسنجی از 100 نفر دارید و جمعیت آن 1000 نفر است، می توانید از نمونه گیری سیستماتیک با انتخاب یک عدد تصادفی بین 1 تا 10، فرض کنید 5، و سپس انتخاب هر 10 نفر استفاده کنید که نتیجه آن انتخاب افراد شماره 5، 15، 25، 35 و غیره

مزیت نمونه گیری سیستماتیک این است که نسبت به نمونه گیری تصادفی موثرتر و کارآمدتر است زیرا باعث کاهش سوگیری در فرآیند انتخاب و صرفه جویی در زمان و منابع می شود.

نقطه ضعف این است که ممکن است منجر به الگوها یا دوره‌هایی در نمونه شود که نماینده جامعه نیستند. بنابراین، انتخاب نقطه شروع و فاصله نمونه برداری مناسب برای جلوگیری از این سوگیری ضروری است.

6.Cluster sampling :

نمونه گیری خوشه ای یک روش نمونه گیری است که در آن جامعه به گروه های مجزای متعددی که به عنوان خوشه شناخته می شوند، تقسیم می شود. سپس این خوشه ها به صورت تصادفی انتخاب می شوند و همه افراد یا واحدهای درون آن خوشه های انتخاب شده در نمونه گنجانده می شوند. این روش اغلب زمانی مورد استفاده قرار می گیرد که بررسی هر فرد یا واحد در جمعیت امکان پذیر نباشد، یا زمانی که جمعیت به طور گسترده در یک منطقه جغرافیایی بزرگ پراکنده شده است. نمونه گیری خوشه ای اغلب در تحقیقات بازار، نظرسنجی افکار عمومی و مطالعات اجتماعی و علمی استفاده می شود. همچنین در مطالعات اپیدمیولوژی استفاده می‌شود، جایی که خوشه‌ها مناطق جغرافیایی با خطرات بهداشتی مشابه را نشان می‌دهند. نمونه‌گیری خوشه‌ای این **مزیت** را دارد که در مقایسه با سایر روش‌های نمونه‌گیری زمان‌بر و کم‌هزینه‌تر است، و زمانی که منابع محدود در دسترس باشد، آن را ایده‌آل می‌کند.

نقطه ضعف این است به اندازه روش‌های دیگر نماینده کل جامعه نباشد و اگر خوشه‌ها به خوبی انتخاب نشده باشند، خطر سوگیری وجود دارد.

قسمت دوم سوال آیا random sampling without replacement روشی مناسب است :

نمونه گیری تصادفی بدون جایگزینی بسته به زمینه و هدف نمونه گیری می تواند روش خوبی باشد. این روش تضمین می کند که هر آیتم در جامعه شانس برابری برای انتخاب شدن برای نمونه دارد و پس از انتخاب یک آیتم، از جامعه حذف می شود و نمی توان دوباره انتخاب کرد. این می تواند در شرایطی که جامعه محدود است و حجم نمونه در مقایسه با حجم جامعه نسبتاً کوچک است مفید باشد. همچنین می‌تواند با اطمینان از اینکه هر آیتم شانس برابری برای انتخاب دارد، به جلوگیری از سوگیری در فرآیند نمونه‌گیری کمک کند. با این حال، ممکن است در همه موارد مناسب نباشد، به عنوان مثال، زمانی که جمعیت بسیار زیاد است، و انتخاب هر مورد بدون جایگزین ممکن است غیر عملی باشد.

Problem 5

One of the most important steps in data mining that takes a lot of time from users is data preprocessing. In the exercise, we are going to pre-process and visualize the data. For this purpose, the Corona disease dataset "owid-covid-data.csv" has been considered for this .exercise, and you can download this dataset from the link below

Answer each of the following questions according to the Corona disease dataset:

1. One of the steps taken at the beginning of all data analysis projects is data quality assessment, during which a general view of the ratio of fields and their values is obtained and familiarity with the data takes place.
Calculate each of the following for 10 arbitrary fields from the Corona disease dataset.

- **Number of rows without value (Null)**
- **Maximum and minimum values of each column along with its country name**
- **Median and mean of each column**
- **Checking the invalid values of each column (negative data and out of range,**
- **etc.**

روش خواندن از فایل :

```
import pandas as pd

# read csv file
df = pd.read_csv('owid-covid-data.csv')
# print first 5 rows of data
print(df.head())
```

دیدن تعداد سطرهای که هیچ مقدار null ندارند :

```
num_rows_without_null = df.notnull().all(axis=1).sum()

print("Number of rows without null values:", num_rows_without_null)
```

```
Number of rows without null values: 0
```

ماکزیمم و مینیمم هر ستون متناسب با کشورش :

```
# group the dataframe by 'country'
grouped = df.groupby('iso_code')

# find the maximum and minimum values of each column for each country
result = grouped.agg(['max', 'min'])

# print the result
print(result)
```

پیدا کردن میانگین و میانه هر ستون :

```
# compute median and mean of each column
medians = df.median()
means = df.mean()
```

پیدا کردن ستون‌های که عددی هستند و مقادیر منفی دارند :

```
min_values = df.describe().loc['min']
invalid_cols = min_values[min_values < 0].index.tolist()
print(invalid_cols)
```

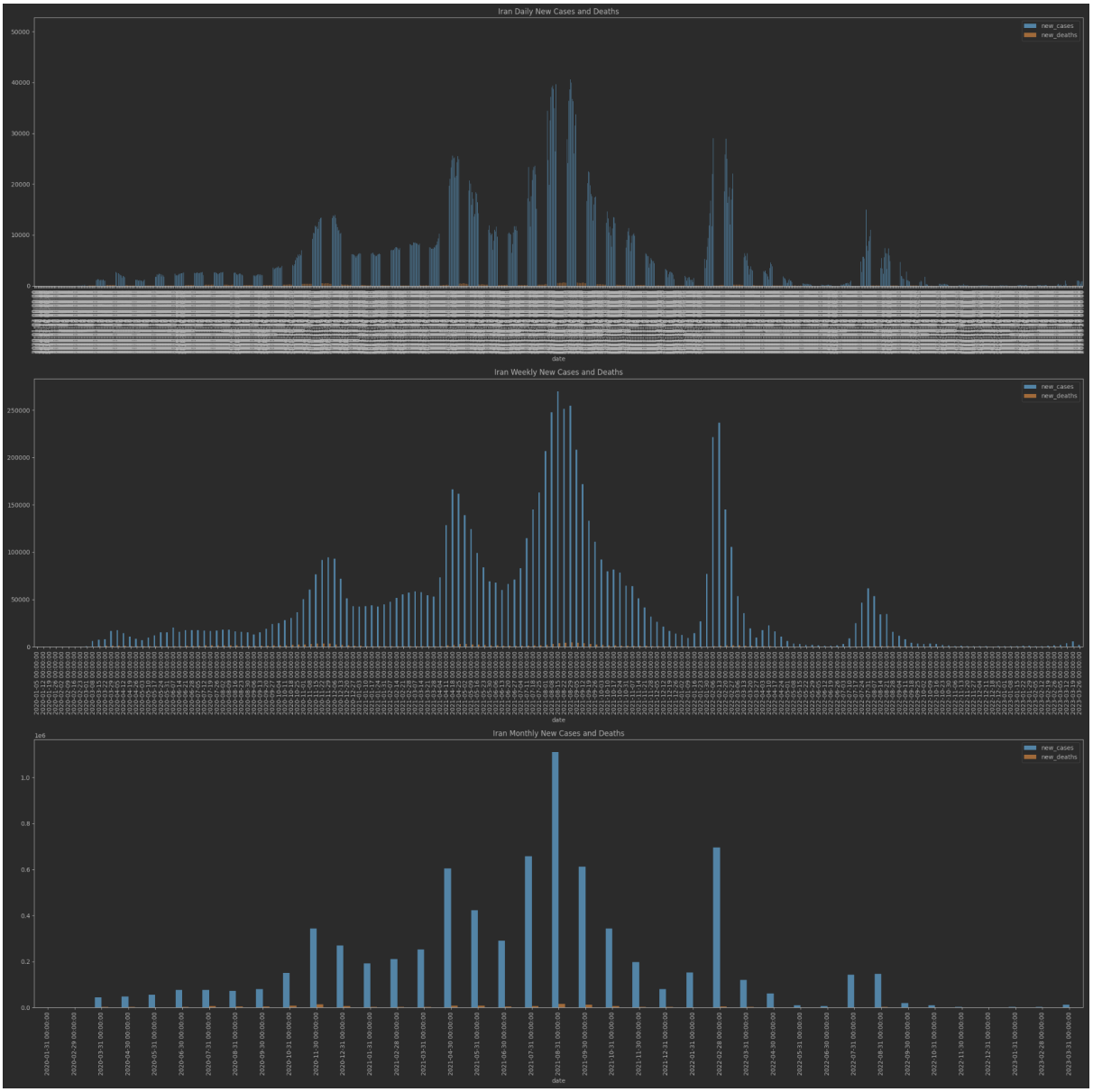
```
['reproduction_rate', 'excess_mortality_cumulative_absolute', 'excess_mortality_cumulative', 'excess_mortality',
'excess_mortality_cumulative_per_million']
```

Draw a bar chart for the values of new_cases and new_deaths columns of Iran in daily, weekly and.

.monthly intervals

```python
import matplotlib.pyplot as plt

# Filter the DataFrame to only include rows where the country is Iran
iran_data = df[df['iso_code']=='IRN']

# Convert date column to datetime format
iran_data['date'] = pd.to_datetime(iran_data['date'])

# Set date column as index
iran_data.set_index('date', inplace=True)

# Resample data to daily, weekly and monthly intervals
daily_df = iran_data.resample('D').sum()[['new_cases', 'new_deaths']]
weekly_df = iran_data.resample('W').sum()[['new_cases', 'new_deaths']]
monthly_df = iran_data.resample('M').sum()[['new_cases', 'new_deaths']]

# Plot the bar charts
fig, (ax1, ax2, ax3) = plt.subplots(3, 1, figsize=(25, 25))

daily_df.plot(kind='bar', ax=ax1)
ax1.set_title('Iran Daily New Cases and Deaths')

weekly_df.plot(kind='bar', ax=ax2)
ax2.set_title('Iran Weekly New Cases and Deaths')

monthly_df.plot(kind='bar', ax=ax3)
ax3.set_title('Iran Monthly New Cases and Deaths')

plt.tight_layout()
plt.show()
```

که حاصل برابر است با :


Iran Daily New Cases and Deaths
new_cases
new_deaths
50000
40000
30000
20000
10000
0
date
Iran Weekly New Cases and Deaths
new_cases
new_deaths
250000
200000
150000
100000
50000
0
date
Iran Monthly New Cases and Deaths
1e6
new_cases
new_deaths
1.0
0.8
0.6
0.4
0.2
0.0
date

In the new_cases column of France, negative values can be seen, assuming that these data are errors, what is your best suggestion to deal with these data? State two methods of finding missing data (missing values), implement one method. Create and report suggested values to replace these data

حذف داده‌های که می‌دانیم اشتباه هستند ساده‌ترین راه و گاها مطمئن‌ترین راه است .

اما یکی از راه‌های مقابله با مقادیر منفی در ستون new cases فرانسه، جایگزینی آنها با صفر (0) است، با این فرض که مقادیر منفی نشان می‌دهد که موارد گزارش‌شده برای تصحیح خطای قبلی به سمت پایین بازبینی شده‌اند. از طرف دیگر، اگر گمان می رود که مقادیر منفی ممکن است نشان دهنده تفاوت بین موارد تایید شده جدید و موارد بازیابی شده باشد، آن مقادیر منفی را می توان با افزودن قدر مطلق آن عدد منفی به ستون بازیابی شده تصحیح کرد. گزینه دیگر بررسی منبع داده ها و تماس با ارائه دهنده برای تایید صحت اعداد گزارش شده است. اگر الگوی ثابتی از مقادیر منفی ذکر شود، فرضیات یا روش‌شناسی خاصی که می‌تواند به مقادیر منفی منجر شود باید مورد بررسی و بازنگری قرار گیرد. صرف نظر از این، ضروری است که این خطاها را با احتیاط بررسی کنیم تا از تفسیر نادرست داده ها یا نتیجه گیری نادرست بر اساس اطلاعات نادرست یا ناقص جلوگیری شود.

1. Pairwise Deletion: در این روش، داده های از دست رفته در هر جا که ظاهر شوند به سادگی نادیده گرفته می شوند و تجزیه و تحلیل با استفاده از تمام داده های موجود انجام می شود.

2. Imputation: در این روش داده های گمشده با مقادیر تخمین زده شده بر اساس داده های موجود پر می شود.

این روش ها شامل استفاده از آزمون های آماری برای شناسایی مقادیر گمشده بر اساس توزیع آنها در مجموعه داده است. نمونه هایی از این آزمون ها عبارتند از: آزمون کولموگروف-اسمیرنوف، آزمون شاپیرو-ویلک و آزمون اندرسون-دارلینگ.

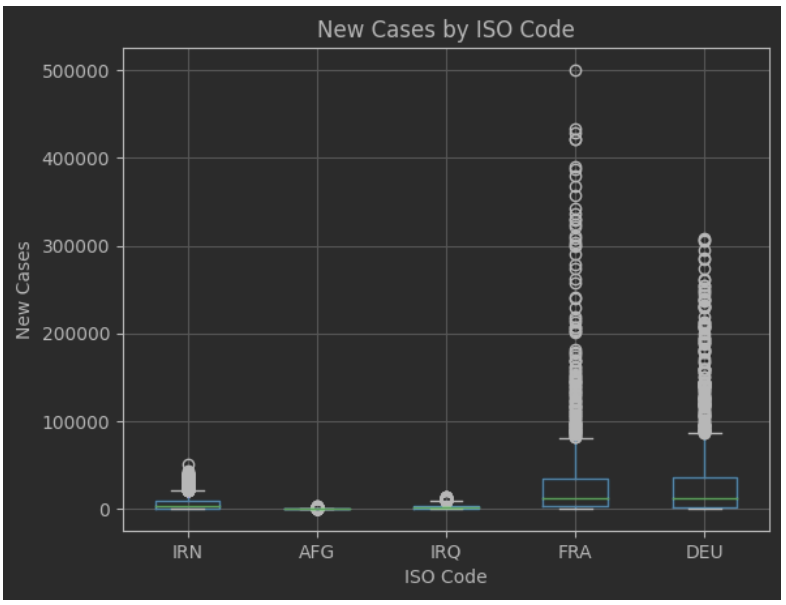
می‌توانید روش‌های آماری مانند میانگین، میانه و نسبت حالت را برای جایگزینی مقادیر گمشده با تخمینی بر اساس داده‌های موجود اعمال کنید. روش دیگر، می‌توانید از روش‌های یادگیری ماشین برای پیش‌بینی مقادیر گمشده بر اساس الگوهای موجود در داده‌های موجود استفاده کنید.

```python
# identifying columns with missing data
missing_columns = df.columns[df.isnull().any()].tolist()

# applying mean imputation to fill the missing values
for column in missing_columns:
    try:
        df[column] = df[column].astype(float)
        mean = df[column].mean()
        print(f'columns {column} : replacement {mean}')
        df[column].fillna(value=mean, inplace=True)
    except:
        df[column] = df[column].astype(str)
        max = df[column].max()
        print(f'columns {column} : replacement {max}')
        df[column].fillna(value=max, inplace=True)
```

```
columns total_deaths : replacement 77985.917032869
columns new_deaths : replacement 99.55503028665828
columns new_deaths_smoothed : replacement 99.94719390214632
columns total_cases_per_million : replacement 82039.19716272318
columns new_cases_per_million : replacement 168.4267084991773
columns new_cases_smoothed_per_million : replacement 169.11012451854666
columns total_deaths_per_million : replacement 783.1037924701335
columns new_deaths_per_million : replacement 1.0571037850530407
columns new_deaths_smoothed_per_million : replacement 1.0612106528098058
columns reproduction_rate : replacement 0.9114953710968147
columns icu_patients : replacement 750.0985205007536
columns icu_patients_per_million : replacement 17.738968502968227
columns hosp_patients : replacement 4222.2903974825085
columns hosp_patients_per_million : replacement 143.43759216033732
```

a. For the values of the new_cases column, draw a box whisker diagram of the data of Iran and two neighboring countries of Iran and two European countries in one diagram. Note that the middle is also visible. As in Figure 2:

```
# filter for iso_code IRN, AFG, IRQ, FRA and DEU

iso_codes = ["IRN", "AFG", "IRQ", "FRA", "DEU"]

df_filtered = df[df['iso_code'].isin(iso_codes)]

# create a list of new cases corresponding to each iso_code
new_cases_by_iso_code = [df_filtered[df_filtered['iso_code'] == iso]['new_cases'] for iso in iso_codes]

new_pandas = pd.DataFrame({'IRN':new_cases_by_iso_code[0],
                           'AFG':new_cases_by_iso_code[1],
                           'IRQ':new_cases_by_iso_code[2],
                           'FRA':new_cases_by_iso_code[3],
                           'DEU':new_cases_by_iso_code[4]})
# create a box whisker diagram using matplotlib


pand_plt = new_pandas.boxplot(column = iso_codes)

# add title and axis labels
plt.title("New Cases by ISO Code")
plt.xlabel("ISO Code")
plt.ylabel("New Cases")

# show the plot
plt.show()
```



b. Calculate and report the value of Q1, Q3, IQR, top whisker and bottom whisker for the new_cases column values of Iran.

```
# filter data for IRN
df_irn = df[df['iso_code']=='IRN']

# calculate Q1, Q3, and IQR
Q1 = df_irn['new_cases'].quantile(0.25)
Q3 = df_irn['new_cases'].quantile(0.75)
IQR = Q3 - Q1

# calculate top and bottom whiskers
top_whisker = Q3 + 1.5*IQR
bottom_whisker = Q1 - 1.5*IQR

# print the results
print("Q1: ", Q1)
print("Q3: ", Q3)
print("IQR: ", IQR)
print("Top whisker: ", top_whisker)
print("Bottom whisker: ", bottom_whisker)
```

```
Q1:  543.5
Q3:  8421.25
IQR:  7877.75
Top whisker:  20237.875
Bottom whisker:  -11273.125
```

c. Find the 10 most outlier data using box whisker plot. The most outlier data means those that have the greatest distance from the top whisker and bottom whisker.

```
"""
To find the 10 most outlier data using the box whisker plot, we can first calculate the interquartile range (IQR) as:

IQR = Q3 - Q1,

where Q1 is the 25th percentile or the first quartile and Q3 is the 75th percentile or the third quartile.

Then, we can define the outlier threshold as:

upper = Q3 + 1.5 * IQR

lower = Q1 - 1.5 * IQR

Any data point that falls outside these upper and lower thresholds would be considered as an outlier.

We can use NumPy to calculate the quartiles and IQR, and then use a boolean mask to filter the outlier data points. The 10 most extreme outliers can
 then be sorted based on their distances from the top and bottom whiskers and displayed.

"""
```

```
iso_codes = ["IRN", "AFG", "IRQ", "FRA", "DEU"]
# calculate quartiles and IQR for each iso_code
quartiles = []
for iso in iso_codes:
    iso_data = df_filtered[df_filtered['iso_code'] == iso]['new_cases']
    q1, q3 = np.percentile(iso_data, [25, 75])
    quartiles.append({'iso_code': iso, 'q1': q1, 'q3': q3, 'iqr': q3 - q1})

# define outlier threshold and filter outlier data points
outliers = []
for q in quartiles:
    upper = q['q3'] + 1.5 * q['iqr']
    lower = q['q1'] - 1.5 * q['iqr']
    iso_data = df_filtered[df_filtered['iso_code'] == q['iso_code']]['new_cases']
    iso_outliers = iso_data[(iso_data > upper) | (iso_data < lower)]
    outliers.extend(list(zip([q['iso_code']] * len(iso_outliers), iso_outliers)))

# sort outliers by distance from whiskers and display top 10
outliers_sorted = sorted(outliers, key=lambda x: abs(quartiles[iso_codes.index(x[0])]['q1'] - x[1]), reverse=True)
print("Top 10 most extreme outliers:")

IRN_number = 0
AFG_number = 0
IRQ_number = 0
FRA_number = 0
DEU_number = 0
```

```python
for i, o in enumerate(outliers_sorted):

    if o[0] == 'IRN' and IRN_number < 10:
        IRN_number += 1
        print(f"{i+1}. ISO code: {o[0]}, value: {o[1]}")

    elif o[0] == 'AFG' and AFG_number < 10:
        AFG_number += 1
        print(f"{i+1}. ISO code: {o[0]}, value: {o[1]}")

    elif o[0] == 'IRQ' and IRQ_number < 10:
        IRQ_number += 1
        print(f"{i+1}. ISO code: {o[0]}, value: {o[1]}")

    elif o[0] == 'FRA' and FRA_number < 10:
        FRA_number += 1
        print(f"{i+1}. ISO code: {o[0]}, value: {o[1]}")

    elif o[0] == 'DEU' and DEU_number < 10:
        DEU_number += 1
        print(f"{i+1}. ISO code: {o[0]}, value: {o[1]}")
```

```
Top 10 most extreme outliers:
1. ISO code: IRN, value: 50228.0
2. ISO code: IRN, value: 42541.0
3. ISO code: IRN, value: 41194.0
4. ISO code: IRN, value: 40808.0
5. ISO code: IRN, value: 40623.0
6. ISO code: IRN, value: 39983.0
7. ISO code: IRN, value: 39819.0
8. ISO code: IRN, value: 39619.0
9. ISO code: IRN, value: 39357.0
10. ISO code: IRN, value: 39174.0
```

Problem 6

After getting familiar with the ARMA model, in this section we are trying to implement ARIMA using python and statsmodels library and use it to predict the probable distance that a person is going to cover in the upcoming days. The steps to perform the task are explained below

1. Data Preprocessing

In this exercise, the dataset includes latitude and longitude, time, and an accumulated distance for each row. Specifically, we intend to use the available information to predict the distance traveled in future days.

1.1. Importing data

- Read the data from the file.
- Perform necessary preprocessing on the data.
- Since the distance in the dataset is stored cumulatively, it is necessary to calculate the distance for each day so that we can use the date and distance of that day to calculate the distance of future days using the ARIMA model.

```
import pandas as pd

# read csv file
df = pd.read_csv('ARIMA-dataset.csv')
# print first 5 rows of data
df.head()
```

5 rows × 4 columns

	time	lat	lon	dev_acc_d
0	2022-03-19 15:12:54	35.724380	51.386521	0.00000
1	2022-03-19 15:12:59	35.723492	51.385984	0.00000
2	2022-03-19 15:13:04	35.723234	51.386712	0.07172
3	2022-03-19 15:13:09	35.723138	51.387657	0.15773
4	2022-03-19 15:13:14	35.723082	51.388778	0.25916

```
# use isnull() and sum() to count the number of null values in each column
null_counts = df.isnull().sum()
print(null_counts)
```

```
time            0
lat             0
lon             0
dev_acc_d    1564
dtype: int64
```

```
df.shape
```

در قدم بعد مقادیر null حذف شده از ستون dev_acc_d را پر می‌کنم به این صورت که اگر بین آیتم ۵ تا ۱۰ خالی باشد سعی می‌کنم تابع چگالی را از آیتم 5 به علاوه یک مقدار ثابت برابر بکنم تا وقتی که به آیتم 10 میرسم حاصل آن برابر با جمع این مقدار ثابت و مقدار جدیدی باشد که برای آیتم 9 اختیار کرده‌ایم .

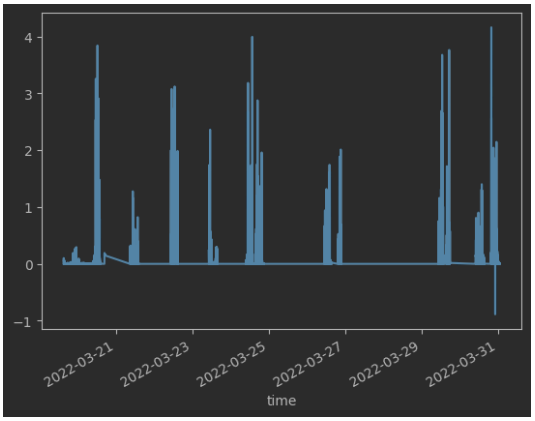
```
def make_dev_acc_d_full_null(data_frame):

    data_dev_acc_d = data_frame['dev_acc_d']
    new_list = []
    item_before_null = False
    number_null = 0
    see_last_item = None
    for item in data_dev_acc_d:

        if pd.isna(item):
            item_before_null = True
            number_null += 1

        elif not item_before_null:
            new_list.append(item)
            see_last_item = item

        else:
            item_before_null = False
            item_must_add = (item - see_last_item)/number_null
            for rank in range(number_null):
                new_list.append(item_must_add*(rank+1)+see_last_item)

            number_null = 0
            see_last_item = item
            new_list.append(item)

    data_frame['dev_acc_d'] = new_list
    return data_frame
```

سپس داده‌های که به صورت تجمیعی هست را به حالت غیر تجمیعی در می‌آورم و مسافت بر حسب هر روز را پیدا می‌کنم .

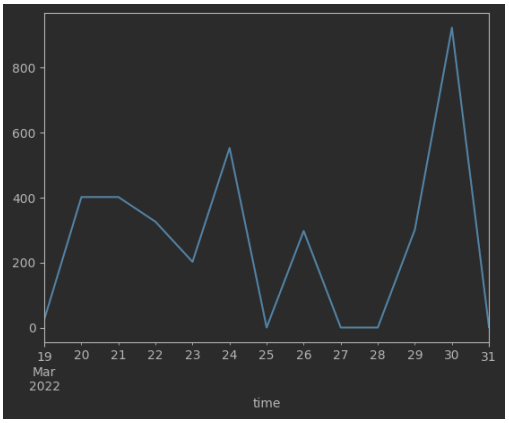
```
def distance_for_each_day(data_frame):

    data_dev_acc_d = data_frame['dev_acc_d']
    new_list = []
    before_item = data_dev_acc_d[0]

    for item in data_dev_acc_d:

        new_list.append(item-before_item)

        before_item = item

    data_frame['distance_for_each_day'] = new_list
    return data_frame
```

سپس نمودار مسافت برای هر روز را به صورت روزانه هفتگی و ماهانه می کشیم .

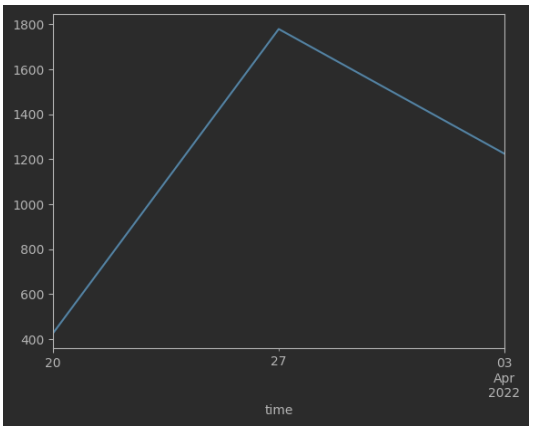
روزانه



هفتگی

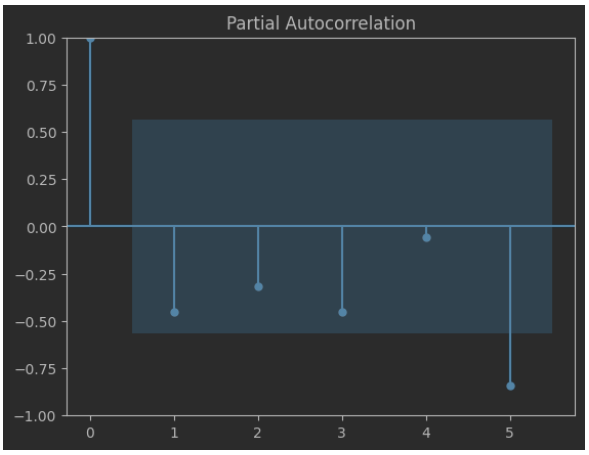


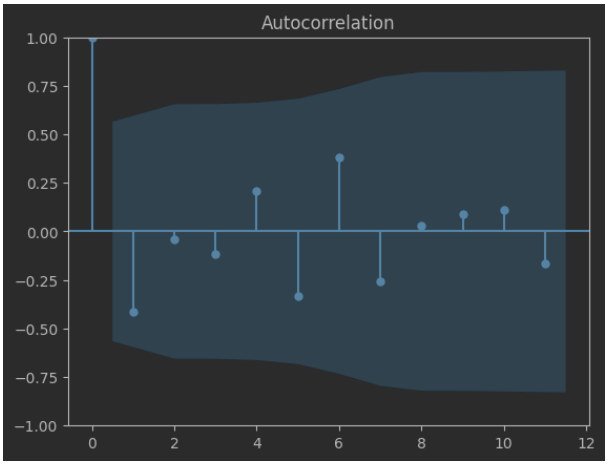
ماهانه



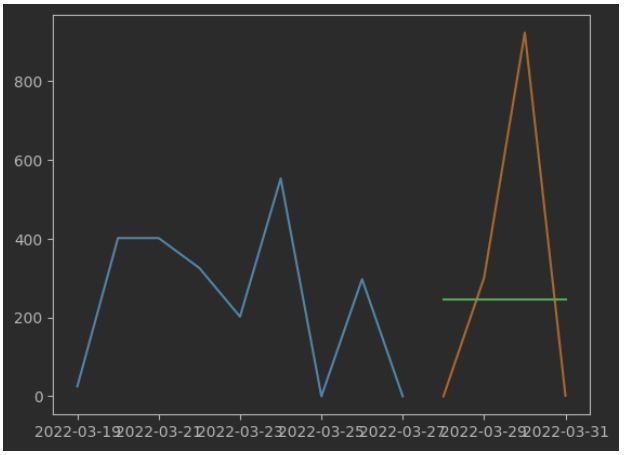
سپس از ما خواسته شده بود که پارامترهای (q,d,p) را تخمین بزنیم .

برای تخمین صحیح این مقادیر دو نمودار pacf , acf می‌تواند کمک کننده باشد .





هرچند که می‌شد `auto arima` استفاده کرد که خود بهترین مدل را پیدا کند .برای مثال قسمت نارنجی رنگ از نمودار هفتگی را به کمک آن به خوبی پیش بینی کردیم از شکل اصلی .(داده مربوط به train قسمت آبی رنگ است)

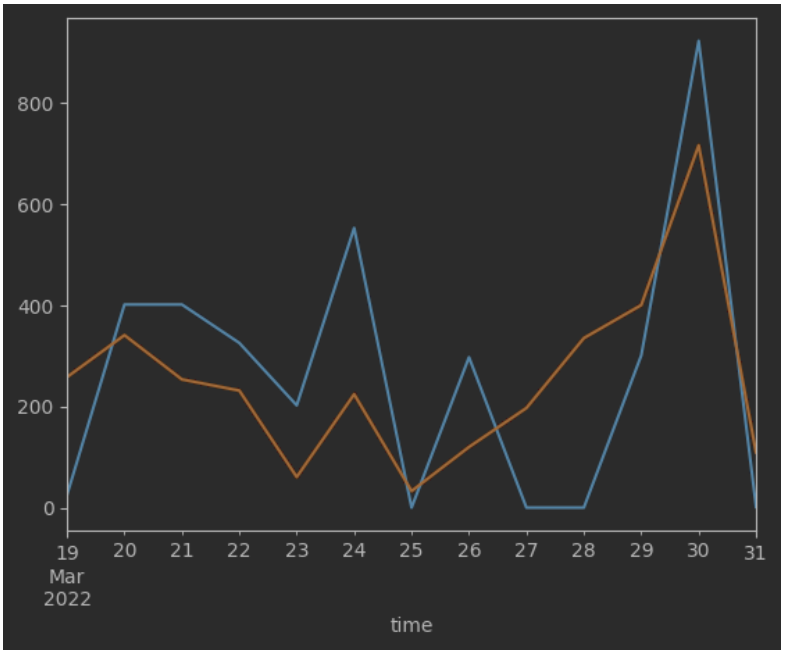


اما از آنجا که خواسته بود خودمان بهترین مدل را برحسب q , p , d تخمینی از سمت خودمان پیدا کنیم نیاز بود نمودارهای بالا کشیده شود .

همانطور که از نمودار هفتگی مشخص است داده‌های ما ترند ندارند و خود در حالت stationary هستند پس نیاز نیست که از این دادگان مشتق گرفته شود پس d را برابر با صفر در نظر می‌گیریم .

برای p نیز عدد ۳ را در نظر گرفتم یعنی به سه تا از دادگان پیشین برای داده فعلی نگاه کند .چون هر سه ماه روند در حال عوض شدن است و می‌توان روند را برحسب آن تخمین زد .

برای q نیز همین عدد سه را انتخاب کردم .



نتیجه حاصل نتیجه قابل قبولی مانند آنچه که به صورت اتوماتیک بدست آمد نیست ولی تا حدود خوبی رفتار مدل را شبیه سازی کرده است .

با افزایش q, p مدل نزدیکتر به به مدل اصلی می‌شد یعنی هرچقدر به صورت دستی و صرفا برای بازی کردن با مدل این متغیرها را بالاتر میبردم مدل در حالت overfit قرار می‌گرفت و با تغییر d و قرار دادن مقدار برای آن به غیر از صفر مدل به طور کل خراب می‌شد و توانایی خود را برای تطبیق با نمودار را به کلی از دست می‌داد .