پروژه درس مبانی داده کاوی

فاز اول

علی جعفری (9622762431)

سوال اول:

با توجه به کد و مشاهده دیتاست دیوار، جدول خواسته شده به صورت زیر پر شده است. بازه مقادیر، به علت طولانی بودن، در پایین جدول و تنها برای بعضی از ویژگی ها آورده شده است. زیرا بعضی از ویژگی ها (مانند id)، خاصیت یکتا بودن داشته و تعداد مقادیر مختلف آنها بسیار زیاد میشود. مقادیر پرت نیز تنها برای ویژگی های از نوع Numeric محاسبه شده است، زیرا همانطور که مشخص است، مقادیری مانند مینیمم، ماکسیمم و Median که برای رسم Box Plot به آنها نیاز است، در این دیتاست تنها برای مقادیر عددی تعریف شده و با معنی هستند. مقادیری هم که در جدول با خط تیره پر شده اند، به این نشانه هستند که آن مشخصه خاص، برای آن ویژگی خاص، بی معنی است. همچنین، از آوردن بازه مقادیر برای ویژگی های عددی پرهیز شده است. زیرا همان طور که مشخص است، مقادیر مینیمم و ماکسیمم این ویژگی ها اندازه گیری شده و بازه مقادیر ویژگی های عددی را میتوان با استفاده از مینیمم و ماکسیمم این ویژگی ها، بازنمایی کرد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ردیف | نام ویژگی | نوع | Min | Max | Mean | Mode | Median |
| 1 | archive\_by\_user | Symmetric Binary | - | - | - | True | - |
| 2 | brand | Nominal | - | - | - | پراید صندوق دار | - |
| 3 | cat1 | Nominal | - | - | - | for-the-home | - |
| 4 | cat2 | Nominal | - | - | - | furniture-and-home-decore | - |
| 5 | cat3 | Nominal | - | - | - | light | - |
| 6 | city | Nominal | - | - | - | Tehran | - |
| 7 | created\_at | Nominal | - | - | - | Saturday 12PM | - |
| 8 | desc | Nominal | - | - | - | کاملا سالم | - |
| 9 | id | Nominal | - | - | - | - | - |
| 10 | image\_count | Numeric | 0 | 16 | 1.78 | 0, 1 | 2 |
| 11 | mileage | Numeric | 0 | 500,000 | 100862.29 | 200000 | 75,000 |
| 12 | platform | Symmetric Binary | - | - | - | mobile | - |
| 13 | price | Numeric | 1 | 400,000,000 | 4,096,271.89 | 100,000 | 220,000 |
| 14 | title | Nominal | - | - | - | بوفه | - |
| 15 | type | Nominal | - | - | - | women | - |
| 16 | year | Numeric | 1366 | 1395 | 1387.64 | 1393 | 1389 |

**بازه مقادیر:**

1- archive\_by\_user: False, True

2- brand: سایر، پژو 405، نوکیا، اپل، سمند، سامسونگ، نیسان، پراید صندوق دار، ام وی ام، پراید هاچ بک، پژو پارس، پژو 206، پیکان، وانت، هوآوی، پژو روآ / آر دی، پژو 206 صندوق دار، ال جی، تیبا، زد تی ای، زانتیا، اچ پی، غیره، کیا، رنو، رانا، دوو، سونی، سونی اریکسون، هیوندای (غیره)، بلکبری، اچ تی سی، لنوو، تندر 90، تویوتا، دل، ایسوس، ایسر، لیفان، فراسو، آمازون، بنز، موتورولا، هیوندای سوناتا، بی ام و، ام اس آی، ویوسونیک، توشیبا، کامپک، فوجیتسو، سوزوکی، گیگابایت، ایلین ویر

3- cat1: for-the-home، vehicles، personal، electronic-devices، businesses، leisure-hobbies

4- cat2: furniture-and-home-decore، cars، baby-and-toys، parts-accessories، utensils-and-appliances، clothing-and-shoes، mobile-tablet، childrens-clothing-and-shoe، game-consoles-and-video-games، audio-video، building-and-garden، jewelry-and-watches، equipments-and-machinery، bicycle، animals، batch، musical-instruments، health-beauty، motorcycles، computers، sport-leisure، book-student-literature، utility، travel-packages، leisure-hobbies-toys، phone

5- cat3: sofa-armchair، antiques-and-art، heavy، personal-toys، cookware-tableware، light، clothing، mobile-phones، tv-projector، garden-and-patio، watches، offices، kitchen، farm-animals، cat، video-dvdplayer، shoes-belt-bag، industrial، tv-and-stereo-furniture، birds، guitar-bass-amplifier، beds-bedroom، carpets، mobile-tablet-accessories، fridge-and-freezer، lighting، tables-and-chairs، strollers-and-accessories، modem-and-network-equipment، jewelry، stereo-surround، camera-camcoders، training، storage، stove-and-heating، barbershop-and-beautysalon، dishwasher، parts-and-accessories، cafe-and-restaurant، microwave-stove، washer-dryer، educational، childrens-furniture، piano-keyboard، desktops، shop-and-cash، laptops، rhinestones، bathroom-wc-sauna، mp3-player، textile-ornaments، tablet، historical-objects، instrument-cleaning-tailoring، fish، accessories، camping-outdoor، traditional، child-car-seat، printer-scaner-copier، rodents-rabbits، coin-stamp، ball-sports، repair-tool، winter-sports، drums-percussion

6- city: Tehran، Mashhad، Karaj، Qom، Isfahan، Shiraz، Tabriz، Ahvaz، Kermanshah

7- created\_at: بازه مقادیر این ویژگی به علت زیاد بودن تعداد مقادیر این ویژگی و عددی نبودن این ویژگی، آورده نشده است.

8- desc: بازه مقادیر این ویژگی به علت زیاد بودن تعداد مقادیر این ویژگی و عددی نبودن این ویژگی، آورده نشده است.

9- id: بازه مقادیر این ویژگی به علت زیاد بودن تعداد مقادیر این ویژگی و عددی نبودن این ویژگی، آورده نشده است (البته جا دارد که در اینجا به این نکته اشاره کنم که با اینکه در ظاهر، ویژگی id، عددی به نظر میرسد، اما چون با این ویژگی بیشتر مانند یک اسم برخورد میشود تا یک عدد، و برای مثال id میتوانست یک رشته الفبایی یکتا نیز باشد و تغییری در نحوه برخورد با آن پیش نمی آمد، این ویژگی به صورت Nominal در نظر گرفته شده است).

10- image\_count: چون این ویژگی از نوع عددی است، بازه مقادیر این ویژگی با استفاده از مقادیر مینیمم و ماکسیمم این ویژگی، طبق جدول مشخص شده است و برابر با [0, 16] میباشد.

11- mileage: چون این ویژگی از نوع عددی است، بازه مقادیر این ویژگی با استفاده از مقادیر مینیمم و ماکسیمم این ویژگی، طبق جدول مشخص شده است و برابر با [0, 500000] میباشد.

12- platform: mobile, web

13- price: چون این ویژگی از نوع عددی است، بازه مقادیر این ویژگی با استفاده از مقادیر مینیمم و ماکسیمم این ویژگی، طبق جدول مشخص شده است و برابر با [1, 400000000] میباشد.

14- title: بازه مقادیر این ویژگی به علت زیاد بودن تعداد مقادیر این ویژگی و عددی نبودن این ویژگی، آورده نشده است.

15- type: women، girls، men، boys

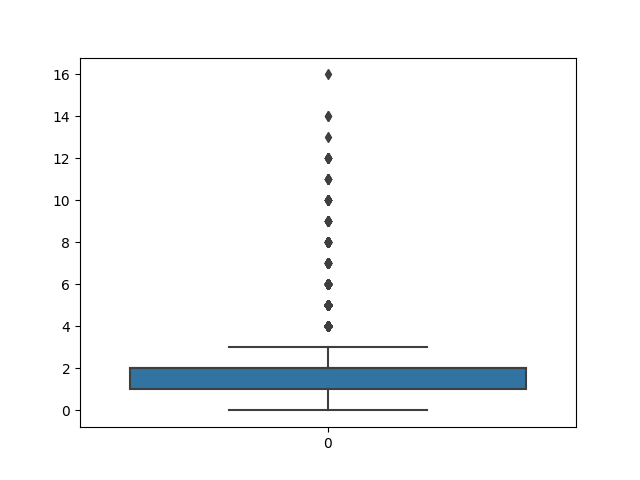
16- year: چون این ویژگی از نوع عددی است، بازه مقادیر این ویژگی با استفاده از مقادیر مینیمم و ماکسیمم این ویژگی، طبق جدول مشخص شده است و برابر با [1366, 1395] میباشد.

**مقادیر پرت:**

همانطور که در ابتدای گزارش گفته شد، و بنا بر دلایل گفته شده در ابتدای گزارش، مقادیر پرت تنها برای ویژگی های از نوع عددی محاسبه شده اند. طبق جدول، ویژگی های عددی دیتاست داده شده، عبارتند از image\_count، mileage، price، و year. در ادامه، Box Plot هر یک از این چهار ویژگی به ترتیب نمایش داده شده و مقادیر پرت هر یک از این چهار ویژگی، طبق Box Plot متناظر آورده شده است.

1- image\_count:

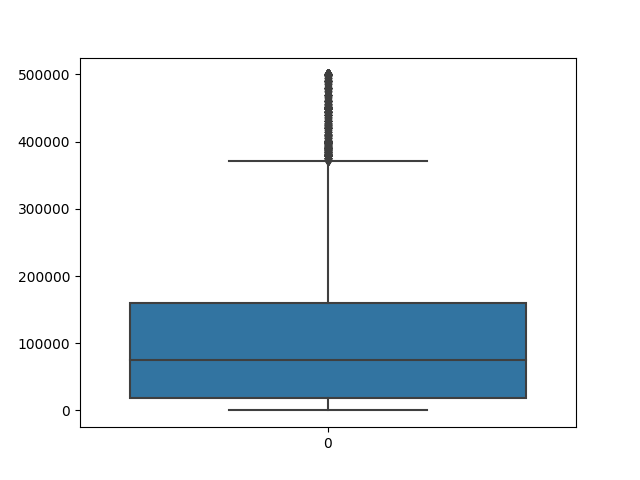
با توجه به تصویر 1، مقادیر image\_count با مقدار 4 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 1**

2- mileage:

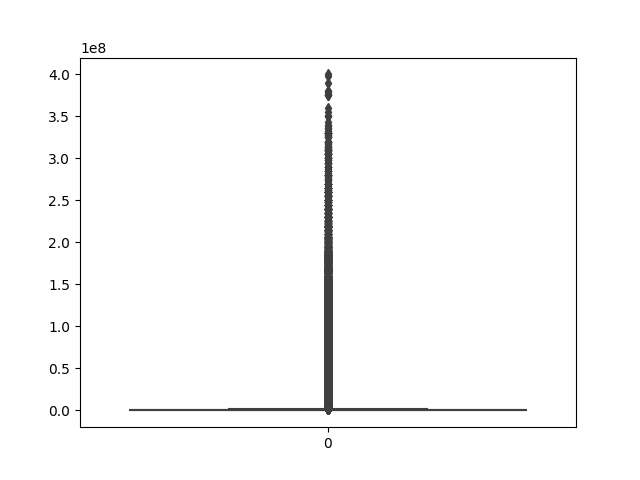
با توجه به تصویر 2، مقادیر mileage با مقدار 372000 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 2**

3- price:

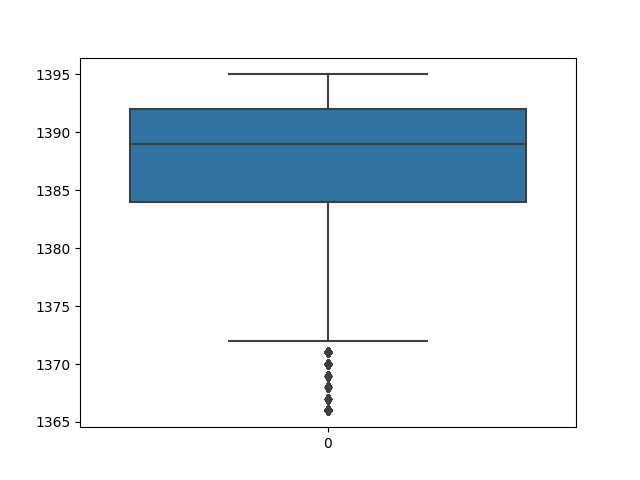
با توجه به تصویر 3، مقادیر price با مقدار 2070000 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 3**

4- year:

با توجه به تصویر 4، مقادیر year با مقدار 1371 به پایین، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 4**

سوال دوم:

قوانین معتبر بودن هر یک از ویژگی ها و به طور کلی، راهکار برخورد با مقادیر نامعتبر هر از ویژگی ها در ادامه آورده شده اند:

1- archive\_by\_user: اگر مقدار این ویژگی، برابر با یکی از حالات True یا False باشد، میگوییم که مقدار این ویژگی معتبر است. در غیر این صورت، مقدار این ویژگی نامعتبر خواهد بود.

2- brand: برای تشخیص معتبر بودن این ویژگی، میتوان یک لیست از اسم برند های مختلف تشکیل داد و نام برند های مختلف را در این لیست ذخیره کرد. در صورتی که brand وارد شده برای یک رکورد، در این لیست وجود داشت، میگوییم که برند معتبر است و در غیر این صورت، میگوییم که برند وارد شده نامعتبر است. میتوان حتی یک دیکشنری، علاوه بر لیست فوق، برای غلط های املایی رایج نام برند های مختلف تشکیل داد که در صورتی که مقدار یک برند با غلط املایی وارد شده بود، صورت اصلاح شده آن را ذخیره کنیم. یک راه دیگر نیز میتواند این باشد که مقدار وارد شده برای brand را به صورت اتوماتیک در اینترنت جستجو کنیم و با تفسیر اتوماتیک داده های دریافت شده از اینترنت، تعیین کنیم که آیا مقدار وارد شده برای brand واقعا متعلق به یک برند است یا خیر.

3- cat1: با فرض اینکه مقادیر ممکن برای این ویژگی، تنها شامل مقادیر تعریف شده هستند، که در بخش بازه مقادیر به آنها پرداخته شد، برای بررسی معتبر بودن مقدار این ویژگی برای یک رکورد، تنها لازم است بررسی کنیم که آیا با یکی از مقادیر تعریف شده برای این ویژگی مطابقت دارد یا خیر، و در تنها در صورت تطابق، مقدار ویژگی را معتبر در نظر بگیریم.

4- cat2: مانند مورد سوم.

5- cat3: مانند مورد سوم.

6- city: مانند مورد سوم.

7- created\_at: برای بررسی معتبر بودن مقدار این ویژگی، تنها لازم است بررسی کنیم که مقدار آن برابر با یکی از روز های هفته باشد و بعد از آن یک ساعت معتبر آمده باشد (یعنی یک ساعت در بازه [1, 12] و مقدار AM یا PM بعد از ساعت).

8- desc: برای بررسی معتبر بودن مقدار این ویژگی میتوانیم کلمات و جمله بندی را بررسی کنیم که آیا معتبر هستند یا خیر. برای بررسی معتبر بودن کلمات به کار رفته و جمله بندی میتوانیم یک classifier را برای این کار، برروی یک دیتاست بزرگ آموزش دهیم و یا از روش های دیگر مبتنی بر پردازش زبان طبیعی استفاده کنیم.

9- id: با بررسی دیتاست، مشخص شد که اکثر مقادیر ویژگی id، رشته های عددی با طول 14 هستند (و برخی از رشته های عددی id، طول کمتر دارند). بنابراین شاید بتوان گفت که برای بررسی معتبر بودن id، باید بررسی کنیم که آیا مقدار id، یک رشته عددی با طول 14 است یا خیر. هرچند که اطلاعات بیشتری درباره دیتاست در اختیار نداریم و با قطعیت نمیتوان گفت که رشته های عددی id با طول کمتر از 14، غیر معتبر هستند.

10- image\_count: برای بررسی معتبر بودن مقدار این ویژگی، تنها لازم است بررسی کنیم که مقدار این ویژگی، یک عدد بوده و عددی غیر منفی باشد.

11- mileage: مانند مورد دهم.

12- platform: با فرض اینکه مقادیر ممکن برای این ویژگی، تنها شامل مقادیر تعریف شده هستند، که در بخش بازه مقادیر به آنها پرداخته شد، برای بررسی معتبر بودن مقدار این ویژگی برای یک رکورد، تنها لازم است بررسی کنیم که آیا با یکی از مقادیر تعریف شده برای این ویژگی مطابقت دارد یا خیر، و در تنها در صورت تطابق، مقدار ویژگی را معتبر در نظر بگیریم.

13- price: برای بررسی معتبر بودن مقدار این ویژگی، تنها لازم است بررسی کنیم که مقدار این ویژگی، یک عدد بوده و عددی غیر منفی باشد.

14- title: برای بررسی معتبر بودن مقدار این ویژگی، میتوان یک classifier را برای این کار، برروی یک دیتاست بزرگ آموزش داد و یا از روش های دیگر مبتنی بر پردازش زبان طبیعی استفاده کرد.

15- type: با فرض اینکه مقادیر ممکن برای این ویژگی، تنها شامل مقادیر تعریف شده هستند، که در بخش بازه مقادیر به آنها پرداخته شد، برای بررسی معتبر بودن مقدار این ویژگی برای یک رکورد، تنها لازم است بررسی کنیم که آیا با یکی از مقادیر تعریف شده برای این ویژگی مطابقت دارد یا خیر، و در تنها در صورت تطابق، مقدار ویژگی را معتبر در نظر بگیریم.

16- year: برای بررسی معتبر بودن مقدار این ویژگی، لازم است بررسی کنیم که مقدار آن از سالی که اکنون در آن هستیم بیشتر نباشد، و همچنین میتوان یک بازه را تعریف نمود که مقدار این ویژگی برای یک رکورد، حداکثر تا مقدار آن بازه، از سالی که اکنون در آن هستیم کمتر باشد تا مقدار مربوط را معتبر در نظر بگیریم (برای مثال، مقدار این ویژگی حداکثر بتواند سی سال از سالی که اکنون در آن هستیم کمتر باشد).

سوال سوم:

برای تشخیص Syntactic Accuracy ویژگی های این دیتاست، باید همانند تشخیص Validity یا معتبر بودن هر یک از ویژگی ها، که در سوال دوم بررسی شد، عمل کنیم. در ادامه، به روش های بررسی Semantic Accuracy هر یک از ویژگی ها میپردازیم:

1- archive\_by\_user: برای این ویژگی، Semantic Accuracy به این معناست که اگر مقدار این ویژگی برابر با True بود، پست مورد نظر واقعا در آرشیو وجود داشته باشد و بالعکس. برای تعیین Semantic Accuracy این ویژگی و اینکه آیا پست واقعا آرشیو شده است یا خیر، میتوان متناظر با مقدار ویژگی، بررسی کرد که پست در آرشیو وجود دارد یا خیر.

2- brand: برای این ویژگی، Semantic Accuracy به این معناست که نام برند ثبت شده، واقعا تولید کننده کالای ثبت شده باشد. برای مثال، یک میز نمیتواند توسط شرکت نوکیا تولید شده باشد. برای بررسی Semantic Accuracy این ویژگی، میتوان یک دیکشنری از نام برند های مختلف تشکیل داد و نام هر برند را به تعدادی دسته بندی مپ کرد. سپس، دسته بندی کالای ثبت شده را، پس از بررسی و حصول اطمینان از صحت دسته بندی، با دسته بندی های مربوط به brand در دیکشنری مذکور، بررسی نمود و دید که آیا مطابقتی وجود دارد یا خیر.

3- cat1: برای این ویژگی، Semantic Accuracy به این معناست که کالای ثبت شده، واقعا به دسته بندی ثبت شده تعلق داشته باشد. این بررسی را میتوان با استفاده از آموزش دادن یک classifier برروی دیگر ویژگی های دیتاست، جهت پیش بینی دسته بندی یک کالای ثبت شده و اطمینان از صحت دسته بندی ثبت شده، انجام داد.

4- cat2: مشابه مورد سوم.

5- cat3: مشابه مورد سوم.

6- city: برای این ویژگی، Semantic Accuracy به این معناست که اگر برای مثال، برای یک رکورد، مقدار شهر برابر با تهران است، این رکورد واقعا از تهران درج شده باشد. شاید راه کاملا دقیقی برای تشخیص Semantic Accuracy برای این ویژگی وجود نداشته باشد. البته یک راه که به ذهنم رسید و شاید تا حدی بتوان با آن Semantic Accuracy این ویژگی را سنجید، این است که هنگام درج یک رکورد، اطلاعات لوکیشن را از کاربر به طور خودکار گرفته و رکورد را در دو دیتاست متفاوت ذخیره کنیم، و بعدا برای تشخیص Semantic Accuracy این ویژگی، دو رکورد ذخیره شده را از نظر این ویژگی با یکدیگر مقایسه کنیم. البته این روش مشکل افزونگی داده را دارد و باز هم احتمال خطا در آن تا حدی وجود دارد.

7- created\_at: مشابه مورد ششم، در اینجا به جای لوکیشن باید خود سرور، تاریخ و ساعت فعلی را در رکورد درج کرده و رکورد را در دو دیتاست متفاوت ذخیره کند.

سوال چهارم: در ادامه، معیار کامل بودن دیتاست با استفاده از سه روش گفته شده، سنجیده شده است.

**روش اول:** روش اول، بدین صورت بود که باید تنها تعداد ردیف هایی از دیتاست که دارای هیچ مقدار Null نبودند را شمرده و بر تعداد کل ردیف های دیتاست تقسیم کنیم. با استفاده از این روش، طبق کد پایتون، معیار کامل بودن دیتاست برابر با صفر بدست آمد. بنابراین میتوان نتیجه گرفت که ظاهرا هیچ ردیفی در این دیتاست نیست که دارای مقدار Null نباشد.

**روش دوم:** روش دوم، بدین صورت بود که ابتدا باید تنها ستون هایی خاص از دیتاست که بدرد کار ما میخورند را درنظر گرفته و سپس روشی مشابه روش اول را تکرار کنیم. برای محاسبه معیار کامل بودن طبق این روش، در اینجا تنها ستون های brand، cat1، cat2، cat3، city، created\_at، desc، platform، price، و title درنظر گرفته شده است. زیرا نظر من این بود که مهم ترین ستون های دیتاست، این ستون ها هستند و میتوان بسیاری از تصمیم گیری ها برروی دیتاست را، مانند پیش بینی قیمت یک محصول و یا خوشه بندی شهر ها بر اساس محصولاتی که در آنها به فروش میرسد، با استفاده از این ویژگی ها انجام داد و نیازی به ویژگی های دیگر نیست. با استفاده از این روش، طبق کد پایتون، معیار کامل بودن دیتاست برابر با 20.14 درصد بدست آمد.

**روش سوم:** روش سوم، بدین صورت بود که باید تعداد مجموع مقادیر ذخیره شده در دیتاست که دارای مقدار هستند را شمرد و بر تعداد مجموع گنجایش دیتاست (یعنی تعداد ردیف ها ضرب در تعداد ستون ها) تقسیم کرد. با استفاده از این روش، طبق کد پایتون، معیار کامل بودن دیتاست برابر با 78.68 درصد بدست آمد.

همچنین، به نظر من، روش دوم برای ارزیابی کامل بودن این دیتاست مناسب تر از سایر روش ها است. زیرا در اینجا روش اول، اطلاعات آنچنانی به ما نمیدهد و تنها متوجه میشویم که همه ردیف های دیتاست دارای مقدار Null هستند، روش سوم نیز از این نظر مناسب نیست که تعدادی از ویژگی های دیتاست، مقدار Null ندارند یا مقادیر Null آنها زیاد نیست، اما کاربرد خاصی در تصمیم گیری های مختلف نیز ندارند و در این روش صرفا شمرده میشوند و کامل بودن را افزایش میدهند. در حالی که اگر بخواهیم از ویژگی هایی که واقعا مورد نیاز ما هستند استفاده کنیم، نسبت مقادیر Null به کل مقادیر، بیشتر خواهد شد و درک واقع گرایانه تری نسبت به کامل بودن دیتاست بدست خواهیم آورد. بنابراین، روش دوم برای ارزیابی کامل بودن این دیتاست مناسب تر از سایر روش ها است.

سوال پنجم: میتوان پنج موضوع چالشی زیر را برروی این دیتاست بررسی نمود:

1- تعیین قیمت برای یک کالا

2- استخراج الگو های قرار دادن پست فروش کالا در شهر های مختلف

3- تعیین مسافت پیموده شده برای یک ماشین از روی قیمت آن

4- خوشه بندی کالا ها بر اساس قیمت آنها

5- تعیین قیمت برای یک کالا، تنها بر اساس توضیحات درج شده برای آن کالا