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

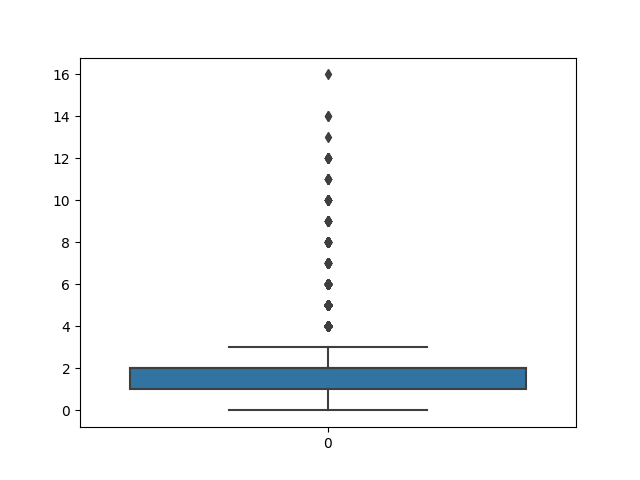
16- year: چون این ویژگی از نوع عددی است، بازه مقادیر این ویژگی با استفاده از مقادیر مینیمم و ماکسیمم این ویژگی، طبق جدول مشخص شده است و برابر با [1366, 1395] میباشد.

**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:

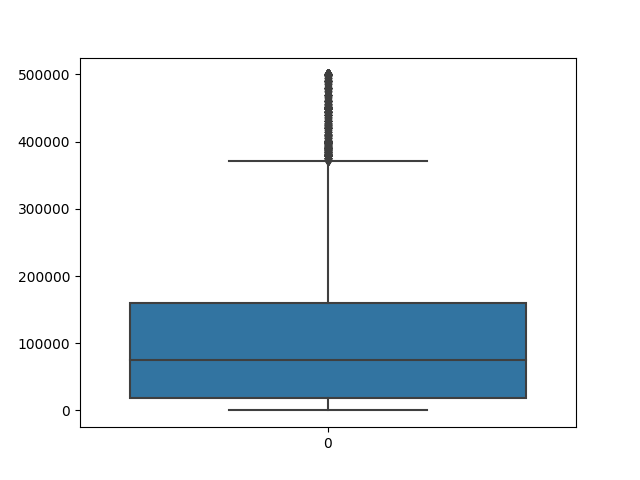
با توجه به تصویر 1، مقادیر image\_count با مقدار 4 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 1**

2- mileage:

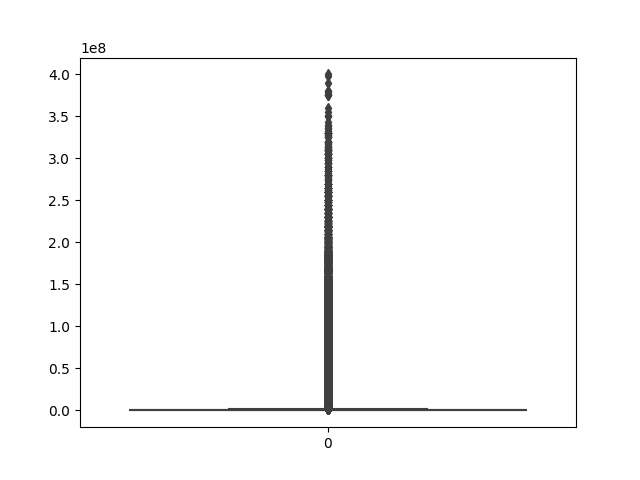
با توجه به تصویر 2، مقادیر mileage با مقدار 372000 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**تصویر 2**

3- price:

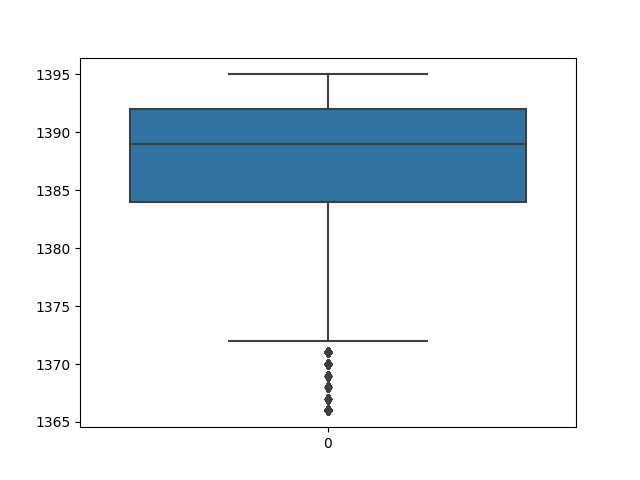
با توجه به تصویر 3، مقادیر price با مقدار 2070000 به بالا، به عنوان مقدار پرت در نظر گرفته شده اند:



**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.