پروژه درس داده کاوی ترم دوم سال تحصیلی ۱۴۰۰- ۱۳۹۹

سجاد ابراهیمی ۹۷۱۲۷۶۲۴۶۵ – محمد اسماعیلی ۹۶۱۲۷۶۲۱۵۰

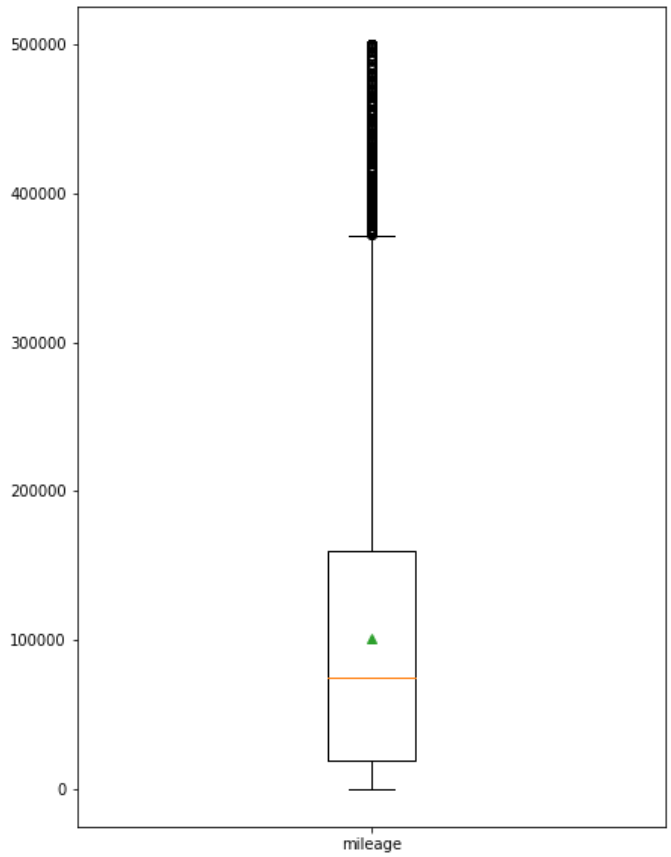
فاز اول پروژه: پیش پردازش

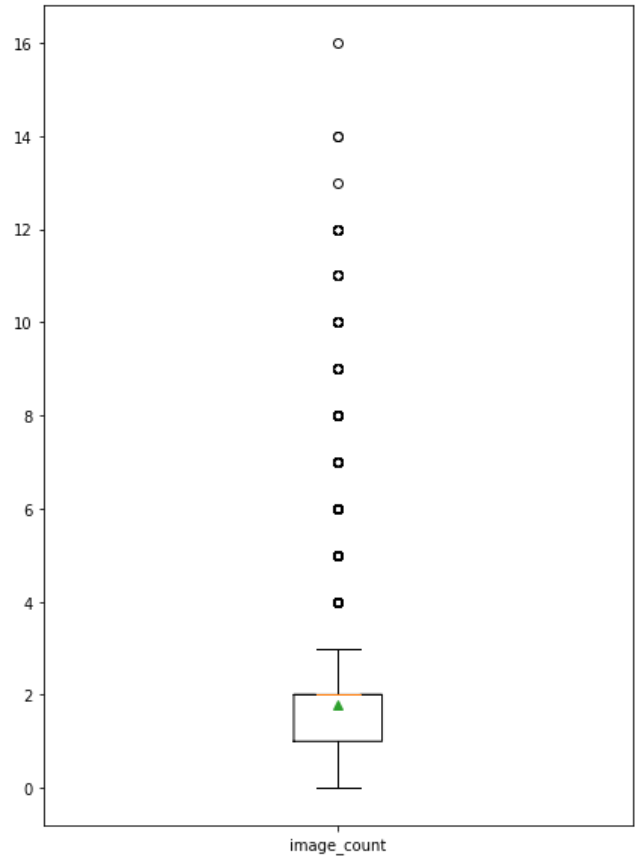
برای یک مجموعه داده که مربوط به دیتاست دیوار می باشد. موارد ذیل را برای این دیتاست ها انجام دهید.

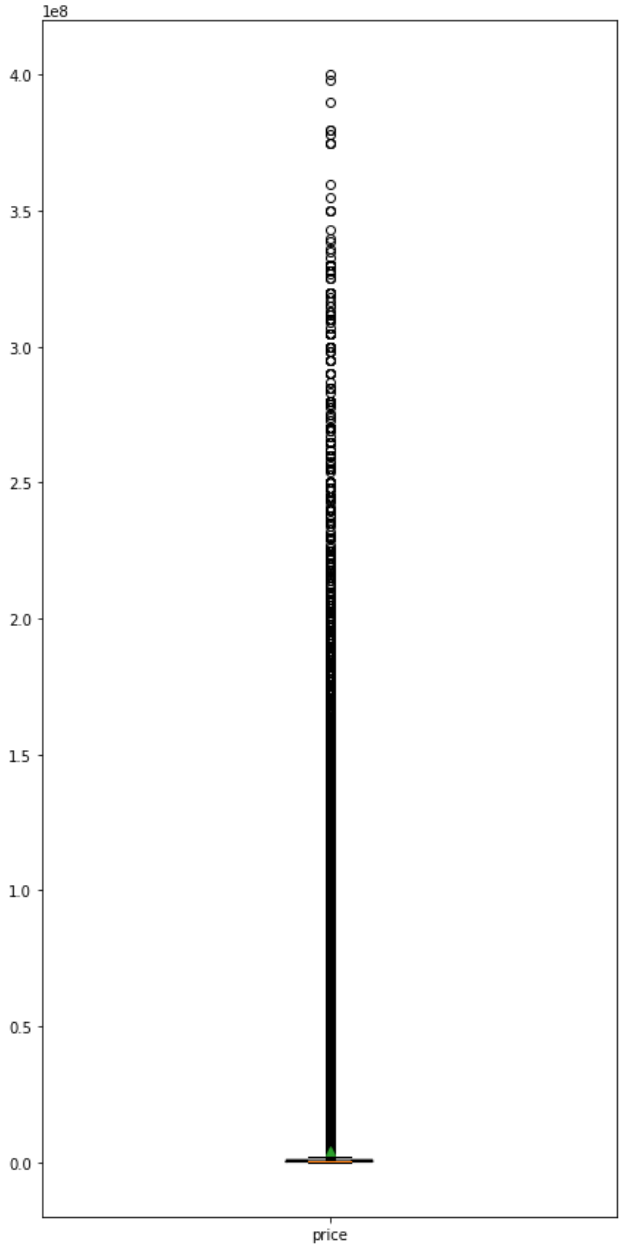
۱. ویژگی های مجموعه داده را طبق جدول زیر توصیف نمایید. سپس با رسم نمودار Box Plot مقادیر پرت هر ویژگی را شناسایی کنید.

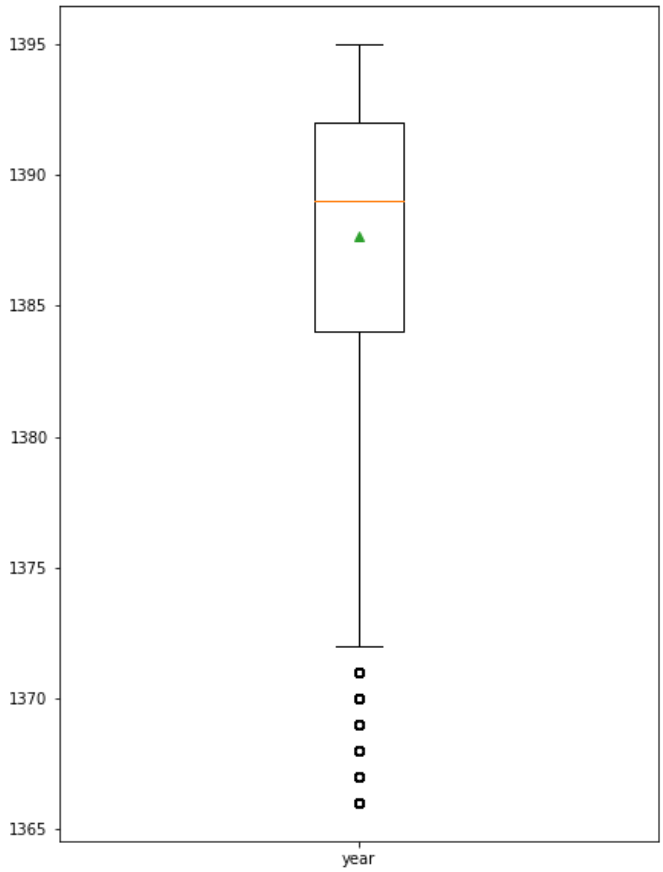
ردیف	نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean
۱	Archive_by_user	Binary	True - False	-	-	-
۲	Brand	Nominal	-	-	-	-
۳	Cat1	Nominal	-	-	-	-
۴	Cat2	Nominal	-	-	-	-
۵	Cat3	Nominal	-	-	-	-
۶	City	Nominal	-	-	-	-
۷	Created_at	Ordinal	-	-	-	-
۸	Desc	Nominal	-	-	-	-
۹	Id	Numeric	-	-	-	-
۱۰	Image_count	Numeric	0 – 16	0	16	1.7859
۱۱	Mileage	Numeric	0 - 500000	0	500000	100862.2919
۱۲	Platform	Nominal	web – mobile	-	-	-
۱۳	Price	Numeric	-1 - 400000000	1	400000000	4096271.8946
۱۴	Title	Nominal	-	-	-	-
۱۵	Type	Nominal	men – women boys - girls	-	-	-
۱۶	Year	Numeric	<1366 - 1395	1366	1395	1387.643

ردیف	نام ویژگی	Mode	Median	مقادیر پرت
۱	Archive_by_user	True	-	-
۲	Brand	پراید صندوق‌دار	-	-
۳	Cat1	for-the-home	-	-
۴	Cat2	furniture-and-home-decore	-	-
۵	Cat3	light	-	-
۶	City	Tehran	-	-
۷	Created_at	Saturday 12PM	-	-
۸	Desc	کاملا سالم	-	-
۹	Id	-	-	-
۱۰	Image_count	1	2	مقادیر بزرگتر از 3.5
۱۱	Mileage	200000	75000	مقادیر بزرگتر از 371500
۱۲	Platform	mobile	-	-
۱۳	Price	-1 (توافقی)	150000	مقادیر بزرگتر از 1455000
۱۴	Title	بوفه	-	-
۱۵	Type	women	-	-
۱۶	Year	1393	1389	مقادیر کوچکتر از 1372









۲. برای ویژگی ها (در صورت امکان)، قوانین **معتبر بودن** را تعریف نموده و میزان معتبر بودن رکوردها را براساس قوانین تعربف شده برای ویژگی ها ارزیابی کنید. چنانچه راهکاری برای برخورد با داده‌های غیر معتبر دارید، توضیح دهید.

۳. برای هر ویژگی مشخص نمایید با چه روشی می توان **صحت داده ها** را بصورت خودکار ارزیابی کرد. (به عنوان مثال در بخش نام استان، مقدار مشهد غیر صحیح است) راه حل خود را برای خودکاری سازی تشخیص داده‌های غیر صحیح برای هر ویژگی دارید بیان کنید.

ردیف	نام ویژگی	معتبر بودن داده	صحت داده
۱	Archive_by_user	استفاده کردن کاربر از مقادیر True/False	
۲	Brand	استفاده از کاراکترهای مجاز در نام‌گذاری اسم برند	
۳	Cat1	استفاده از کاراکترهای مجاز در نام‌گذاری نام مجموعه	موجود بودن در دیتابیس نام مجموعه‌ها
٤	Cat2	استفاده از کاراکترهای مجاز در نام‌گذاری نام مجموعه	موجود بودن در دیتابیس نام مجموعه‌ها
٥	Cat3	استفاده از کاراکترهای مجاز در نام‌گذاری نام مجموعه	موجود بودن در دیتابیس نام مجموعه‌ها
٦	City	فقط استفاده از حروف الفبا	موجود بودن نام در دیتابیس شهرها
۷	Created_at	سازگاری با عبارت منظم تاریخ‌های موجود در دیتاست	استفاده از مقادیر مجاز در هر قسمت از عبارت منظم برای قسمت روزهای هفته استفاده از هفت روز ممکن برای قسمت ساعت استفاده از اعداد صحیح بیشتر از صفر کمتر از ۱۲ برای قسمت وضعیت استفاده ازam/pm (البته در صورت صحیح نوشتن عبارت منظم ۱ و ۳ را در بخش معتبر بودن داده نیز میتوان چک کرد).
۸	Desc	استفاده از کاراکترهای مجاز تعریف شده توسط سیستم	
۹	Id	تنها استفاده از اعداد صحیح مثبت	-
۱۰	Image_count	استفاده از اعداد صحیح مثبت	-
۱۱	Mileage	استفاده از اعداد صحیح مثبت	مشخص کردن اینکه کدام مجموعه ها دارای این ویژگی میباشند و جلوگیری از ورود این مقدار برای اشیا مربوط به سایر مجموعه ها

۱۲	Platform	استفاده از حروف الفبا در نام پلتفرم	استفاده از نام پلتفرم‌های پشتیبانی شده توسط سیستم
۱۳	Price	استفاده از اعداد صحیح مثبت	
۱٤	Title	استفاده از حروف الفبا در موضوع	
۱۵	Type	استفاده از حروف الفبا	استفاده از موضوعات موجود در سامانه
۱٦	Year	استفاده از اعداد صحیح مثبت	تعیین حد پایین و در نظر گرفتن سال فعلی به عنوان حد بالا

۴. با استفاده از سه روش ارائه شده در کلاس، **کامل بودن** داده ها را ارزیابی کنید. بنظر شما کدام روش ارزیابی برای این مجموعه داده مناسب تر است؟ چرا؟

method	Value
1	0.0
2	0.8181
3	0.7583

با توجه به نتایج به دست آمده در هر یک از روش های مختلف و با توجه به نحوه عملکرد هر یک از روش ها روش سوم به دلیل بررسی تمام فیلد های موجود مطلوب ترین و منطقی ترین نتیجه را به ثبت می رساند.

در روش دوم نتیجه بستگی به ستون های انتخابی دارد که در بعضی حالات این مقدار تا عدد ۱/۰ نیز افزایش خواهد یافت. مقدار ثبت شده در جدول حاصل انتخاب ستون های تعداد عکس، توضیحات، و دسته بندی های اول تا سوم است.

۵. براساس دانشی که نسبت به دیتاست به دست آورده اید، ۵ موضوع چالشی که می‌توان بر روی این دادگان بررسی نمود را بیان نمایید.

i. یافتن شهرهایی که آگهی بیشتری از آنها ثبت می شود و تهیه سرور از دیتاسنتر های موجود در آن نقاط برای دسترسی سریعتر
ii. یافتن ساعات پیک استفاده از نرم افزار و در نظر گرفتن میزان مطلوب منابع برای آن ساعات
iii. بررسی قیمت گذاری های انجام شده بر روی محصولات و یافتن قیمت های خارج از عرف(به کمک محدوده داده های پرت) و اطلاع رسانی به مشتریان به جهت خرید با قیمت غیر واقعی

iv. یافتن پلتفرم های پراستفاده تر و تخصیص نیروی بیشتر برای بهبود آن پلتفرم
v. پیدا کردن دسته بندی های پر استفاده تر و تقسیم آنها به زیر دسته بندی های دقیق تر جهت بهبود تجربه کاربری