پاسخ سوال اول بخش پیاده‌سازی تمرین اول مبانی داده‌کاوی

گزارش تحلیل و پیش‌پردازش داده‌های مربوط به دو فروشگاه لباس

الف) خواندن داده‌ها به کمک Pandas

ابتدا باید با استفاده از تابع read_json() داده‌ها را از فایل json خواند. در نتیجه این کار فایل به فرمت DataFrame درخواهد آمد. با استفاده از تابع head() می‌توان چند سطر اول این مجموعه داده را مشاهده کرد.

	item_id	category	size	quality	user_name	length	fit	cup size	user_id	waist	hips	bra size	bust	height	shoe size	shoe width	review_summary	review_text
0	123373	new	21	5.0	alexmeyer626	just right	fit	dd/e	875643	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	123373	new	7	5.0	Emily	just right	small	d	991571	29.0	38.0	34.0	36	5ft 6in	NaN	NaN	NaN	NaN
2	123373	new	13	3.0	sydneybraden2001	just right	small	b	587883	31.0	30.0	36.0	NaN	5ft 2in	NaN	NaN	NaN	NaN
3	123373	new	7	2.0	Ugggh	slightly long	small	b	395665	30.0	NaN	32.0	NaN	5ft 7in	9.0	NaN	NaN	NaN
4	123373	new	18	5.0	dberrones1	slightly long	small	b	944840	NaN	NaN	36.0	NaN	5ft 2in	NaN	NaN	NaN	NaN

ب) شرح اطلاعات آماری مجموعه داده

برای بدست آوردن اطلاعات آماری این مجموعه داده، می‌توان از تابع describe() استفاده کرد تا توصیفی آماری از متغیرهای عددی را مشاهده کنیم.

	item_id	size	quality	user_id	waist	hips	bra size	shoe size
count	82790.000000	82790.000000	82722.000000	82790.000000	2882.000000	56064.000000	76772.000000	27915.000000
mean	469325.229170	12.661602	3.949058	498849.564718	31.319223	40.358501	35.972125	8.145818
std	213999.803314	8.271952	0.992783	286356.969459	5.302849	5.827166	3.224907	1.336109
min	123373.000000	0.000000	1.000000	6.000000	20.000000	30.000000	28.000000	5.000000
25%	314980.000000	8.000000	3.000000	252897.750000	28.000000	36.000000	34.000000	7.000000
50%	454030.000000	12.000000	4.000000	497913.500000	30.000000	39.000000	36.000000	8.000000
75%	658440.000000	15.000000	5.000000	744745.250000	34.000000	43.000000	38.000000	9.000000
max	807722.000000	38.000000	5.000000	999972.000000	50.000000	60.000000	48.000000	38.000000

ج) Missing values

به سادگی می‌توان با استفاده از متد isnull() داده‌های دست رفته را شناسایی کرد و به کمک sum() تعداد کل missing valueها را برای هر ستون محاسبه کرد. طبق خروجی بدست آمده در مجموع ۱۲ ستون دارای مقادیر از دست رفته هستند که در شکل زیر می‌توان دید که این مقادیر مربوط به کدام ستون‌ها می‌باشند.

	missing value count
item_id	0
category	0
size	0
quality	68
user_name	0
length	35
fit	0
cup size	6255
user_id	0
waist	79908
hips	26726
bra size	6018
bust	70936
height	1107
shoe size	54875
shoe width	64183
review_summary	6725
review_text	6725

رویکرد در نظر گرفته شده برای برخورد با هر یک از ویژگی‌هایی که دارای missing value هستند در ادامه شرح داده می‌شود:

- Length: تعداد کل داده‌های از دست رفته برای این ویژگی ۳۵ عدد است؛ پس به سادگی می‌توان از آن‌ها صرف نظر کرد بدون آن که بر روی نتایج اثر منفی بگذارد.
- Quality: مشابه ویژگی length می‌توان از تاپل‌هایی که مقادیر ویژگی quality آن‌ها از دست رفته است چشم پوشی کرد؛ چرا که تنها ۶۸ سطر را شامل می شود. همچنین بازخوردی از مشتری برای این سطرها در نظر گرفته نشده است؛ پس نمی‌توان آن‌ها به درستی پر کرد.
- Review_summary: مقادیر از دست رفته این ستون را با مقدار Unknown می‌توان پر کرد.
- Review_text: مقادیر از دست رفته این ستون را با مقدار Unknown می‌توان پر کرد.
- Waist: ۹۶ درصد از داده‌های این ستون در تاپل‌های موجود از دست رفته‌اند؛ پس باید این ستون را به کلی حذف کرد.
- Hips: از آن جایی که مقادیر از دست رفته این ستون از ویژگی‌ها زیاد بوده و غیرقابل چشم‌پوشی است، برای این ستون می‌توان از دسته‌بندی بر اساس چندک و اختصاص صفت کیفی ترتیبی به آن‌ها استفاده کرد؛ سپس داده‌های از دست رفته را با مقدار Unknown پر کرد.
- Bust: این ویژگی نیز دارای نرخ missing value بالاییست (۶۸%) که مشابه ویژگی waist بهتر است حذف شود.
- Cup size: با بررسی برخی از سطرهایی که دارای missing value در ستون cup size هستند، به نظر می‌رسد تبدیل این ویژگی، به یک ویژگی categorical و پر کردن مقادیر از دست رفته با Unknown بهترین گزینه است.
- Height: با بررسی چندین سطر از داده‌ها که ستون height آن‌ها دارای missing value است، نمی‌توان به نتیجه‌ای مبنی بر روش درست پر کردن این فیلدها رسید و با توجه به درصد پایین وقوع missing value در این ستون به نظر می‌رسد می‌توان از این تاپل‌ها چشم‌پوشی کرد.
- Shoe size و Shoe width: درصد داده‌های از دست رفته هر دو ویژگی بسیار بالاست، اما مشابه waist یا bust امکان حذف این ستون‌ها وجود ندارد؛ چون اثرات منفی بر روی نتایج خواهد گذاشت. پس باید این ویژگی‌ها را به ویژگی‌های categorical تبدیل و از مقدار Unknown برای مقادیر از دست رفته استفاده کنیم.

- Bra size: با توجه به آن که اکثر سایزبندی‌ها در محدوده ۳۴ تا ۳۸ قرار دارند می‌توان این دسته از missingها را با استفاده از سنجه‌های شاخص مرکزی و به طور خاص با مقدار میانگین پر کرد (چون داده‌ها نرمال‌اند).

در این مرحله بهتر است دیگر روش‌های پالایش داده‌ها را نیز بکار برد که منجر به بهبود نتیجه تحلیل‌های ثانویه و آسان‌تر شدن کار با داده‌ها می‌شود:

- تبدیل ویژگی category به یک ویژگی دسته‌بندی شده
- با توجه به اطلاعات آماری بدست آمده و با استفاده از متد unique() واضح است که داده ۳۸ یک داده پرت برای ویژگی shoe size محسوب می‌شود:

```
In [46]: shoe_sizes = data['shoe size'].unique()
         shoe_sizes
```

```
Out[46]: array([ nan,  9. ,  8.5, 11. ,  7. ,  6. ,  8. ,  6.5, 10. ,  7.5,  5.5,
                 9.5, 10.5,  5. , 11.5, 38. ])
```

اگر به اطلاعات این تاپل نگاه کنیم، به نظر می‌رسد که رکوردی قابل قبول است و تنها سایز کفش غیرعادی محسوب می‌شود که ممکن است به اشتباه وارد شده باشد؛ پس این مقدار را null می‌کنیم.

	item_id	category	size	quality	user_name	length	fit	cup size	user_id	hips	bra size	height	shoe size	shoe width	review_summary	review_text
37313	416942	new	12	5.0	Catslittle	just right	fit	d	237498	XL	36.0	5ft 5in	38.0	average	Cardigans are best item Mod Cloth has	I love these cardigans, my favorite, good fabr...

- تبدیل ویژگی fit به یک ویژگی دسته‌بندی شده
- با مشاهده ستون height در این مجموعه داده به نظر می‌رسد که تبدیل مقادیر آن به داده‌های عددی و همچنین تبدیل آن‌ها به واحد سانتی‌متر ضروری است. پس برای اینکار از تابع parse_height() استفاده می‌کنیم و سپس به کمک متد apply و پاس دادن این تابع به عنوان ورودی مقادیر را به فرم مورد نظر در می‌آوریم.

```
def parse_height(height):
    feet_inches = height.replace('ft','').replace('in','').split(" ")
    if len(feet_inches) == 2:
        return inches_to_centimeters(12 * int(feet_inches[0]) + float(feet_inches[1]))
    return inches_to_centimeters(12 * int(feet_inches[0]))

def inches_to_centimeters(inches):
    return 2.54 * inches
```

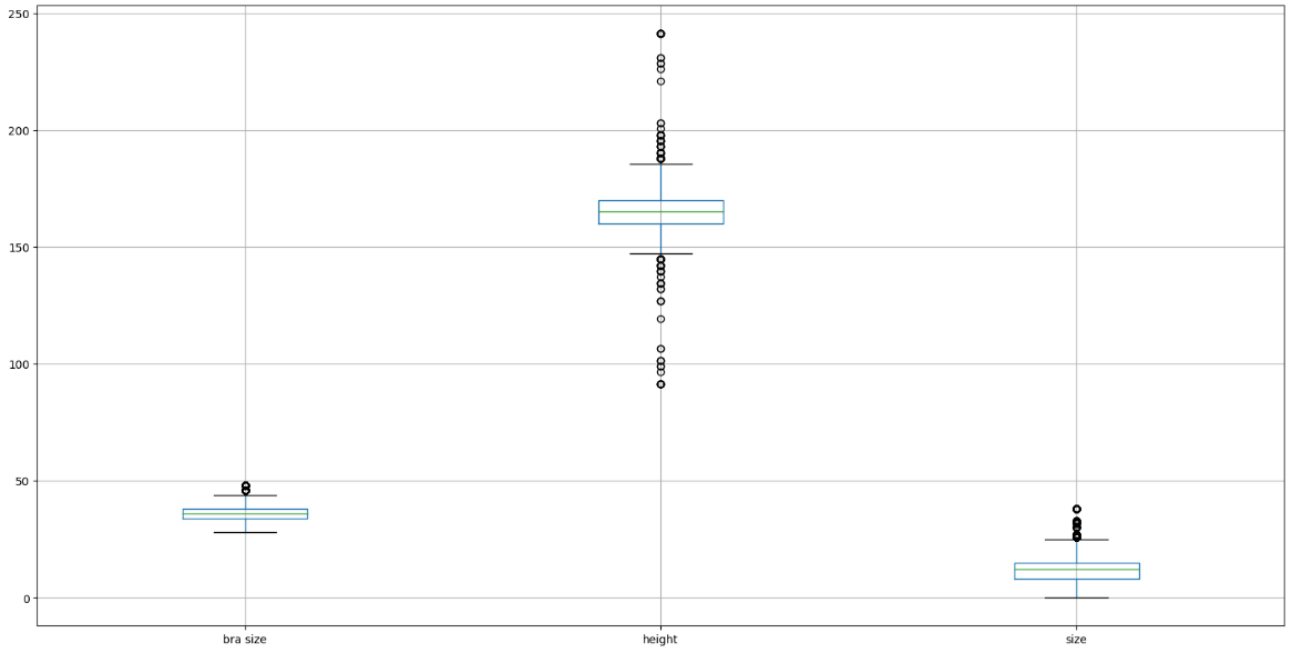
- تبدیل ویژگی quality به یک ویژگی دسته‌بندی شده
- در نهایت برای سادگی کار با این مجموعه داده می‌توان ویژگی user_name را حذف کرد؛ چرا که user_id به تنهایی اطلاعات مورد نظر ما برای شناسایی یک مشتری خاص را فراهم می‌کند.

در نتیجه عملیات پیش پردازش داده‌ها می‌بینیم که دیگر missing valueای وجود ندارد و می‌توان به سراغ رسم نمودارها و data visualization رفت.

```
Int64Index: 81594 entries, 1 to 82789
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   item_id         81594 non-null  int64
 1   category        81594 non-null  category
 2   size            81594 non-null  int64
 3   quality         81594 non-null  category
 4   length          81594 non-null  object
 5   fit             81594 non-null  category
 6   cup size        81594 non-null  category
 7   user_id         81594 non-null  int64
 8   hips            81594 non-null  category
 9   bra size        81594 non-null  float64
 10  height          81594 non-null  float64
 11  shoe size       81593 non-null  category
 12  shoe width      81594 non-null  category
 13  review_summary  81594 non-null  object
 14  review_text     81594 non-null  object
dtypes: category(7), float64(2), int64(3), object(3)
memory usage: 8.2+ MB
```

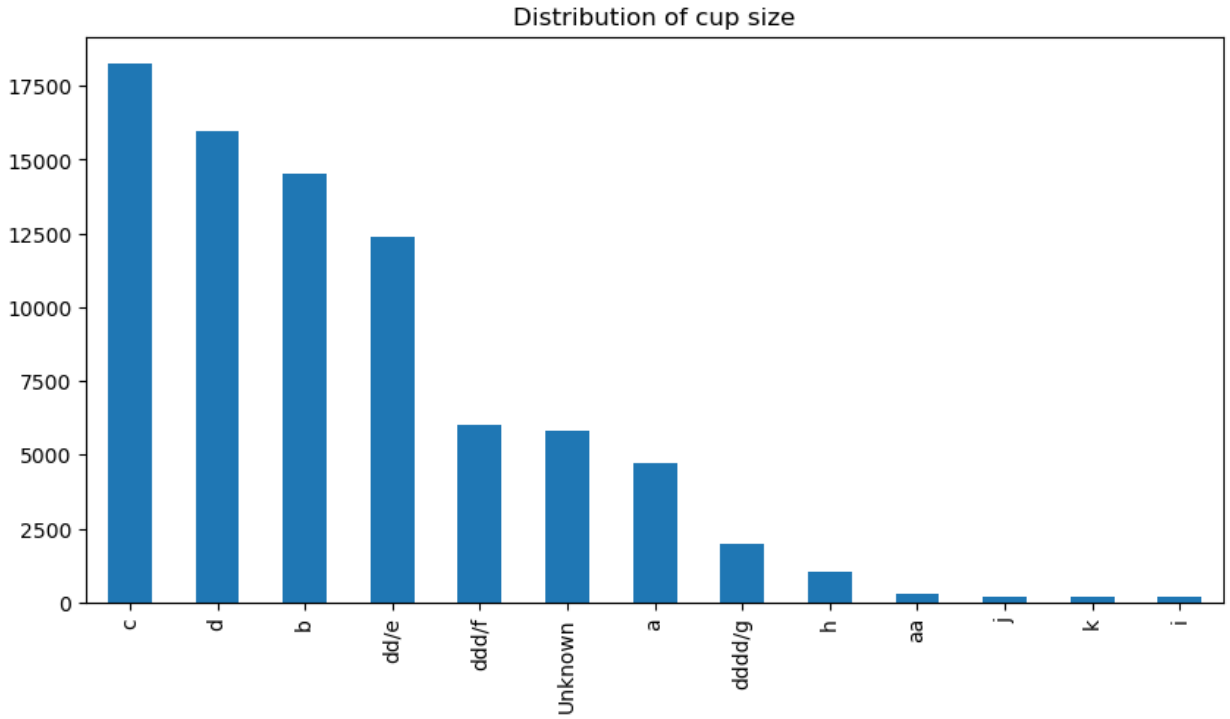
د) نمودار Boxplot برای ویژگی‌های عددی

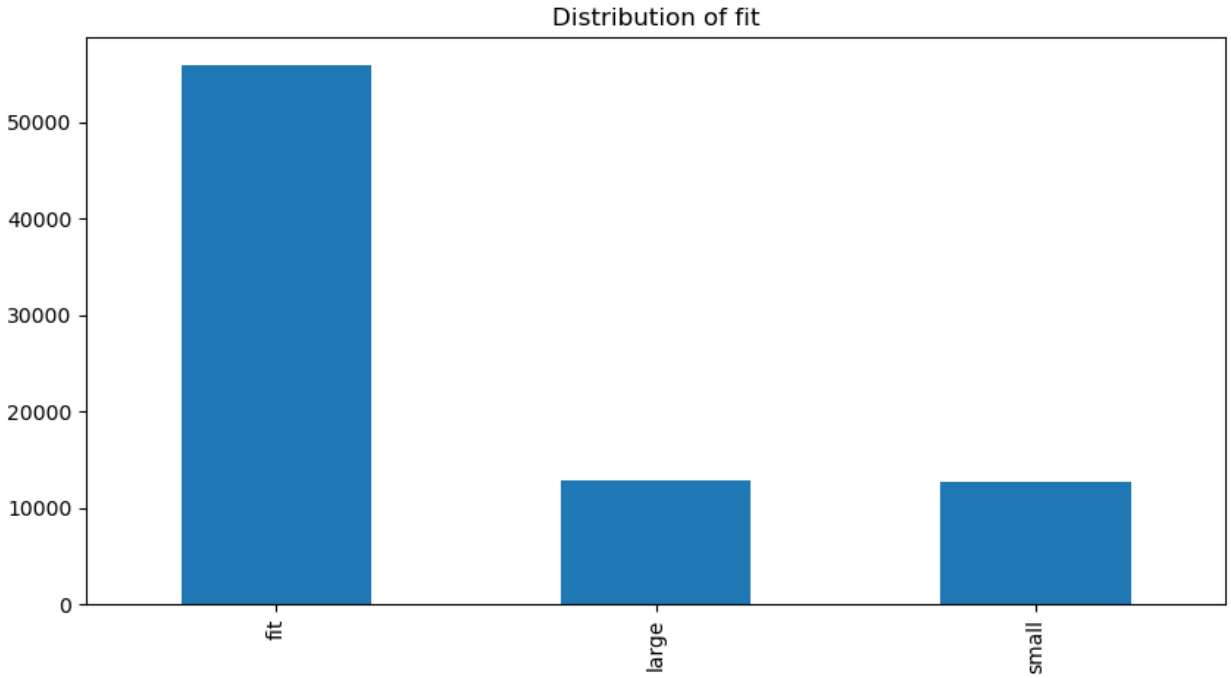
در این مرحله و پس از انجام عملیات پیش پردازش بر روی داده‌ها، bra size، size و height جز داده‌های عددی این مجموعه داده محسوب می‌شوند که نمودار boxplot برای آن‌ها در شکل زیر ترسیم شده است:

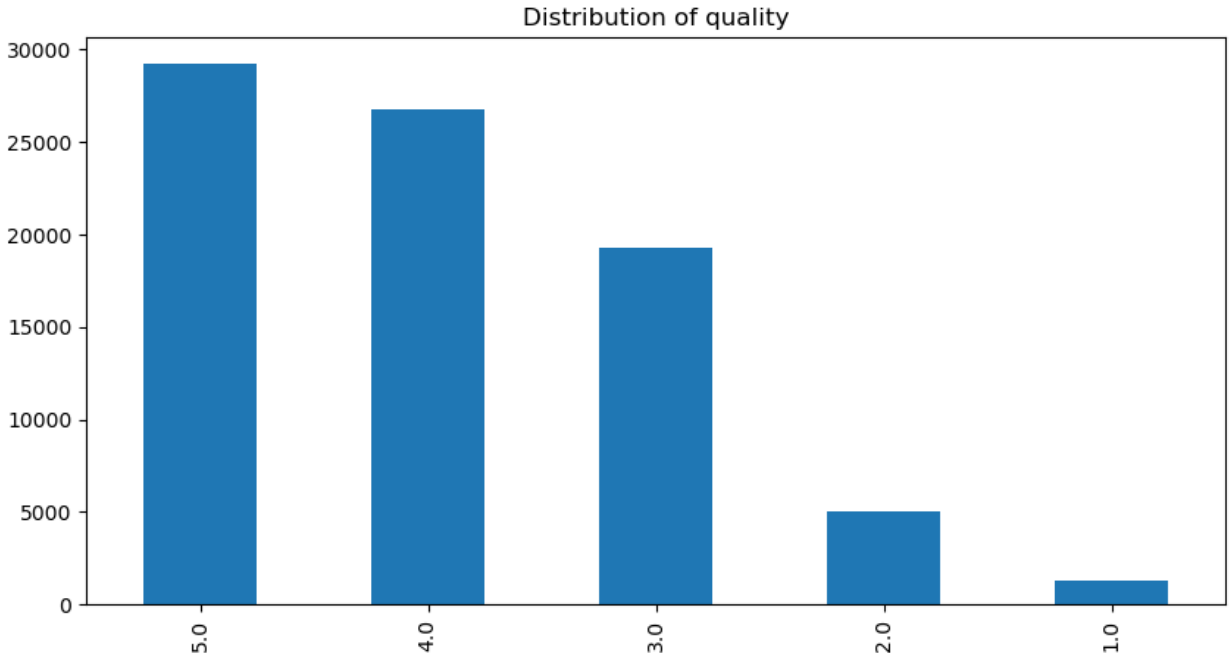


ه) نمودار توزیع

مطابق با خواسته سوال نمودار توزیع به ترتیب برای ۳ ویژگی cup size، fit و quality رسم شده است که نتایج آن در ادامه قابل مشاهده است.







و) نمودار category بر اساس length-feedback