Ahmedabad University

Subject:- ML

Weekly Report Submission

GROUP: **DATA DYNAMOS**

Submitted to faculty: Prof. Mehul Raval

Date of Submission: 11/02/2023

Student Details

Roll No.	Name of the Student	Name of the Program
AU2040122	Aditi Vasa	Btech CSE
AU2040002	Shrey Somani	Btech CSE
AU2040196	Vandan Shah	Btech CSE
AU2040048	Ronit Shah	Btech CSE

2020-2021 (Monsoon Semester)

Report

Reference Paper:

https://cseweb.ucsd.edu/~jmcauley/pdfs/recsys18e.pdf

Need for studying the factors affecting size of cloths:

Online purchases of textiles, shoes, and clothes have increased by 23% just in the past year, accounting for more than half of all retail fashion purchases. The volume of returns, on the other hand, is a mixed bag: on average, 20% of purchases are returned, with that number increasing to 50% for pricey items. Retailers are out millions of dollars in missing sales, shipping, and processing expenses. It also has an impact on the environment.

The dynamics have changed in the past decade. From doing hours of trials in the changing rooms of malls to now spending hours on apps selecting return or exchange options.

One of the main causes of returns is the fits of the products. The answer lies in size and fit technologies.

This dataset studies various factors that are crucial in deciding whether the size of product is small, large or fit.

Printing data in JSON Format

In [3]:
```
import json


def load_data(fp):
    with open(fp) as fid:
        series = (pd.Series(json.loads(s)) for s in fid)
        return pd.concat(series, axis=1).T

renththerunway_fp = './renttherunway_final_data.json'
df = load_data(renththerunway_fp)
df.head()
```

Out[3]:

	fit	user_id	bust size	item_id	weight	rating	rented for	review_text	body type	review_summary	category	height	size	age	review_date
0	fit	420272	34d	2260466	137lbs	10	vacation	An adorable romper! Belt and zipper were a lit...	hourglass	So many compliments!	romper	5' 8"	14	28	April 20, 2016
1	fit	273551	34b	153475	132lbs	10	other	I rented this dress for a photo shoot. The the...	straight & narrow	I felt so glamourous!!!	gown	5' 6"	12	36	June 18, 2013
2	fit	360448	NaN	1063761	NaN	10	party	This hugged in all the right places! It was a ...	NaN	It was a great time to celebrate the (almost) ...	sheath	5' 4"	4	116	December 14, 2015
3	fit	909926	34c	126335	135lbs	8	formal affair	I rented this for my company's black tie award...	pear	Dress arrived on time and in perfect condition.	dress	5' 5"	8	34	February 12, 2014
4	fit	151944	34b	616682	145lbs	10	wedding	I have always been petite in my upper body and...	athletic	Was in love with this dress !!!	gown	5' 9"	12	27	September 26, 2016

Understanding Data

In [14]:
```
df.shape
```

Out[14]:
```
(192544, 15)
```

In [19]:
```
df.info()
#data types of columns
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192544 entries, 0 to 192543
Data columns (total 15 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   fit             192544 non-null  object
 1   user_id         192544 non-null  object
 2   bust size       174133 non-null  object
 3   item_id         192544 non-null  object
 4   weight          162562 non-null  object
 5   rating          192462 non-null  object
 6   rented for      192534 non-null  object
 7   review_text     192544 non-null  object
 8   body type       177907 non-null  object
 9   review_summary  192544 non-null  object
 10  category        192544 non-null  object
 11  height          191867 non-null  object
 12  size            192544 non-null  object
 13  age             191584 non-null  object
 14  review_date     192544 non-null  object
dtypes: object(15)
memory usage: 22.0+ MB
```

In [20]:
```
df.isnull().sum()
#checking for null values
```

Out[20]:
```
fit                 0
user_id             0
bust size       18411
item_id             0
weight          29982
rating             82
rented for         10
review_text         0
body type       14637
review_summary      0
category            0
height            677
size                0
age               960
review_date         0
dtype: int64
```

In [17]:
```
df.describe()
#mathematical information about data
```

Out[17]:

	fit	user_id	bust size	item_id	weight	rating	rented for	review_text	body type	review_summary	category	height	size	age	review_date
count	192544	192544	174133	192544	162562	192462	192534	192544	177907	192544	192544	191867	192544	191584	192544
unique	3	105571	106	5850	190	5	9	191031	7	154740	68	24	56	89	2274
top	fit	691468	34b	126335	130lbs	10	wedding	.	hourglass	Stylist Review	dress	5' 4"	8	31	June 15, 2016
freq	142058	436	27285	2241	14370	124537	57784	63	55349	977	92884	28012	40804	14522	844

In [18]:
```
df.duplicated().sum()
#checking the duplicate values
```

Out[18]:
```
189
```

CLEANING DATA

In [5]:
```
df=df.drop(['review_text','review_summary','review_date'],axis=1)
df
```

Out[5]:

	fit	user_id	bust size	item_id	weight	rating	rented for	body type	category	height	size	age
0	fit	420272	34d	2260466	137lbs	10	vacation	hourglass	romper	5' 8"	14	28
1	fit	273551	34b	153475	132lbs	10	other	straight & narrow	gown	5' 6"	12	36
2	fit	360448	NaN	1063761	NaN	10	party	NaN	sheath	5' 4"	4	116
3	fit	909926	34c	126335	135lbs	8	formal affair	pear	dress	5' 5"	8	34
4	fit	151944	34b	616682	145lbs	10	wedding	athletic	gown	5' 9"	12	27
...	...	...	...	...	...	...	...	...	...	...	...	...
192539	fit	66386	34dd	2252812	140lbs	10	work	hourglass	jumpsuit	5' 9"	8	42
192540	fit	118398	32c	682043	100lbs	10	work	petite	dress	5' 1"	4	29
192541	fit	47002	36a	683251	135lbs	6	everyday	straight & narrow	dress	5' 8"	8	31
192542	fit	961120	36c	126335	165lbs	10	wedding	pear	dress	5' 6"	16	31
192543	fit	123612	36b	127865	155lbs	10	wedding	athletic	gown	5' 6"	16	30

192544 rows × 12 columns

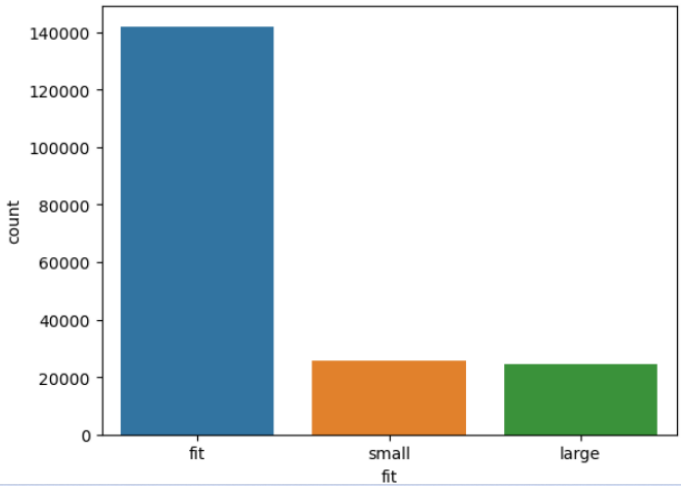
Exploratory Data Analysis to understand main features of data

In [4]:
```
import seaborn as sns
```

In [5]:
```
sns.countplot(df['fit'])
```

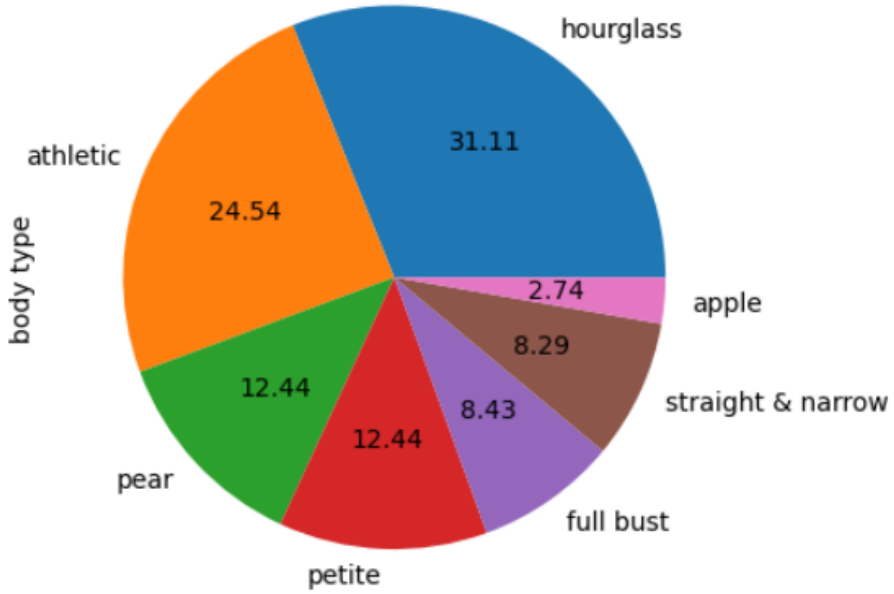
```
C:\Users\nirav\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword a
rg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit ke
yword will result in an error or misinterpretation.
  warnings.warn(
```

Out[5]: <AxesSubplot:xlabel='fit', ylabel='count'>



In [6]:
```
df['body type'].value_counts().plot(kind='pie',autopct='%.2f')
```

Out[6]: <AxesSubplot:ylabel='body type'>



TASK IN UPCOMING WEEK

- Decision on which data set to use among the selected two data sets.
- Clean and pre-process data and split the data into training and testing models.
- Apply two to three prediction algorithms for predicting cloth size and try to know which is better.