python4_data_analysis_2

February 4, 2024

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
file="https://bit.ly/BDA_Week_5_dataset"
```

[3]:
```
data=pd.read_csv(file, na_values=" ")
```

[4]:
```
data.head()
```

[4]:

	ID	age	gender	product	cost	selling_price	quantity
0	1	52	Male	E	425.0	485.0	5.0
1	2	32	Female	S	342.0	350.0	4.0
2	3	35	Female	S	222.0	233.0	2.0
3	4	50	Male	P	929.0	936.0	7.0
4	5	43	Female	E	343.0	359.0	NaN

[5]:
```
data.shape
```

[5]:
```
(15, 7)
```

[8]:
```
print(f" The sample size is {data.shape[0]}")
print(f" The number of variables is {data.shape[1]}")
```

```
 The sample size is 15
 The number of variables is 7
```

[9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15 entries, 0 to 14
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   ID             15 non-null     int64
 1   age            15 non-null     int64
 2   gender         14 non-null     object
 3   product        14 non-null     object
 4   cost           14 non-null     float64
```

```
 5   selling_price  14 non-null     float64
 6   quantity       13 non-null     float64
dtypes: float64(3), int64(2), object(2)
memory usage: 968.0+ bytes
```

[10]:
```
data.columns
```

[10]:
```
Index(['ID', 'age', 'gender', 'product', 'cost', 'selling_price', 'quantity'],
dtype='object')
```

[11]:
```
data.index
```

[11]:
```
RangeIndex(start=0, stop=15, step=1)
```

[12]:
```
data.describe()
```

[12]:

	ID	age	cost	selling_price	quantity
count	15.000000	15.000000	14.000000	14.000000	13.000000
mean	8.000000	41.333333	510.142857	537.785714	5.076923
std	4.472136	11.298968	249.715178	263.053425	1.934836
min	1.000000	22.000000	222.000000	233.000000	2.000000
25%	4.500000	34.500000	342.250000	352.250000	4.000000
50%	8.000000	42.000000	419.500000	458.000000	5.000000
75%	11.500000	47.500000	631.000000	665.750000	7.000000
max	15.000000	69.000000	1000.000000	1089.000000	8.000000

[13]:
```
data.head(3)
```

[13]:

	ID	age	gender	product	cost	selling_price	quantity
0	1	52	Male	E	425.0	485.0	5.0
1	2	32	Female	S	342.0	350.0	4.0
2	3	35	Female	S	222.0	233.0	2.0

[15]:
```
#absolute frequency

data['gender'].value_counts()
```

[15]:
```
Female    9
Male      5
Name: gender, dtype: int64
```

[16]:
```
data['product'].value_counts()
```

[16]:
```
E    5
P    5
S    4
Name: product, dtype: int64
```

[17]:
```
#relative frequency

data['product'].value_counts(normalize=True)
```

[17]:
```
E    0.357143
P    0.357143
S    0.285714
Name: product, dtype: float64
```

[19]:
```
select_columns=['gender', 'quantity']
data[select_columns]
```

[19]:

	gender	quantity
0	Male	5.0
1	Female	4.0
2	Female	2.0
3	Male	7.0
4	Female	NaN
5	Female	3.0
6	Male	7.0
7	Female	4.0
8	Male	7.0
9	Female	3.0
10	Female	4.0
11	Male	7.0
12	Female	5.0
13	Female	NaN
14	NaN	8.0

[20]:
```
data
```

[20]:

	ID	age	gender	product	cost	selling_price	quantity
0	1	52	Male	E	425.0	485.0	5.0
1	2	32	Female	S	342.0	350.0	4.0
2	3	35	Female	S	222.0	233.0	2.0
3	4	50	Male	P	929.0	936.0	7.0
4	5	43	Female	E	343.0	359.0	NaN
5	6	49	Female	P	240.0	249.0	3.0
6	7	38	Male	P	1000.0	1089.0	7.0
7	8	29	Female	E	405.0	423.0	4.0
8	9	22	Male	S	660.0	689.0	7.0
9	10	46	Female	P	307.0	325.0	3.0
10	11	43	Female	E	414.0	431.0	4.0
11	12	36	Male	E	544.0	596.0	7.0
12	13	42	Female	S	500.0	519.0	5.0
13	14	34	Female	NaN	NaN	NaN	NaN
14	15	69	NaN	P	811.0	845.0	8.0

[28]:
```
#select the first 3 rows and only first 6 variables

data.loc[:'2', :'selling_price']
```

[28]:

	ID	age	gender	product	cost	selling_price
0	1	52	Male	E	425.0	485.0
1	2	32	Female	S	342.0	350.0
2	3	35	Female	S	222.0	233.0

[25]:
```
data.iloc[:3,:6]
```

[25]:

	ID	age	gender	product	cost	selling_price
0	1	52	Male	E	425.0	485.0
1	2	32	Female	S	342.0	350.0
2	3	35	Female	S	222.0	233.0

[29]:
```
#select the first 5 rows

data.iloc[:5]
```

[29]:

	ID	age	gender	product	cost	selling_price	quantity
0	1	52	Male	E	425.0	485.0	5.0
1	2	32	Female	S	342.0	350.0	4.0
2	3	35	Female	S	222.0	233.0	2.0
3	4	50	Male	P	929.0	936.0	7.0
4	5	43	Female	E	343.0	359.0	NaN

[27]:
```
data.loc[:'4']
```

[27]:

	ID	age	gender	product	cost	selling_price	quantity
0	1	52	Male	E	425.0	485.0	5.0
1	2	32	Female	S	342.0	350.0	4.0
2	3	35	Female	S	222.0	233.0	2.0
3	4	50	Male	P	929.0	936.0	7.0
4	5	43	Female	E	343.0	359.0	NaN

[32]:
```
#select the last row

data.iloc[-1]
```

[32]:
```
ID                15
age               69
gender           NaN
product            P
cost           811.0
selling_price  845.0
quantity         8.0
```

```
Name: 14, dtype: object
```

[34]:
```
#select the second last row

data.iloc[-2]
```

[34]:
```
ID                  14
age                 34
gender          Female
product            NaN
cost               NaN
selling_price      NaN
quantity           NaN
Name: 13, dtype: object
```

[35]:
```
#filtering

#select cost row with <300

filter = data['cost']<300

data[filter]
```

[35]:

	ID	age	gender	product	cost	selling_price	quantity
2	3	35	Female	S	222.0	233.0	2.0
5	6	49	Female	P	240.0	249.0	3.0

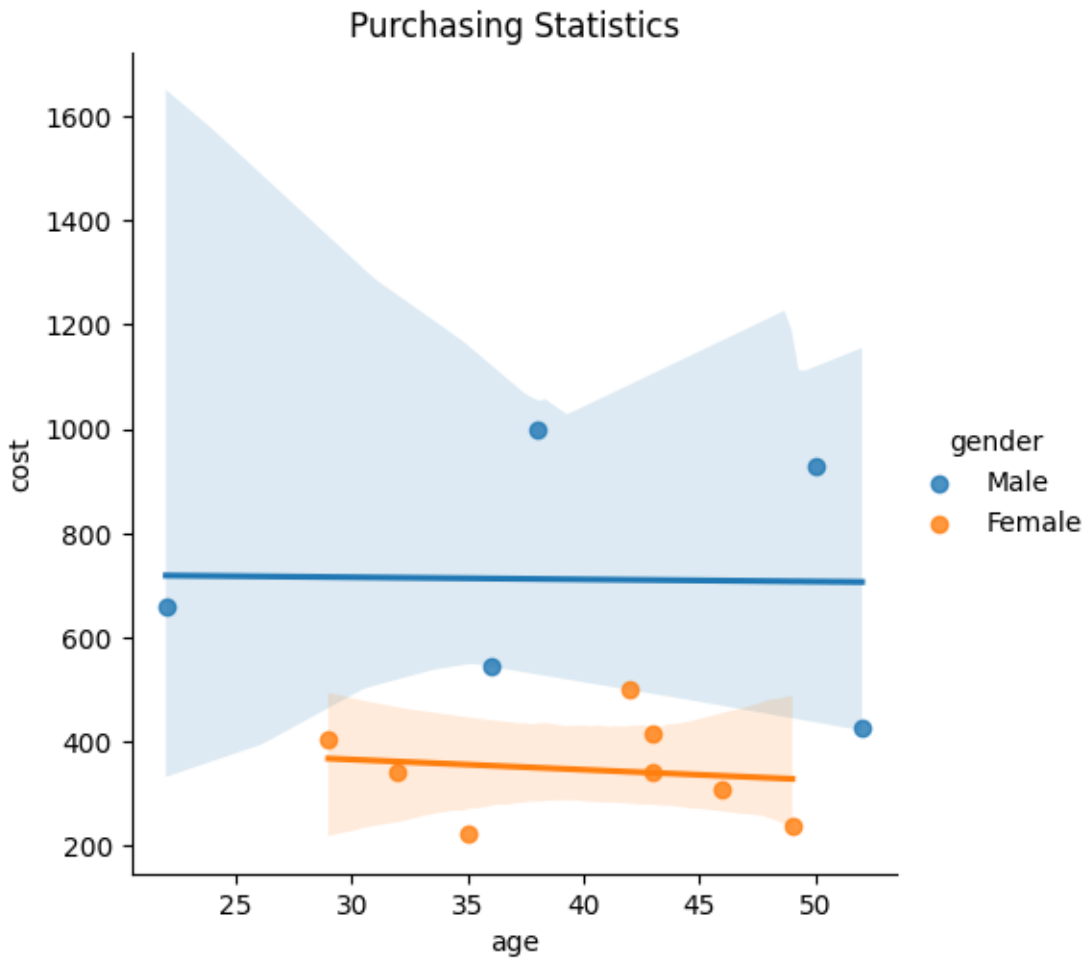
[38]:
```
#find cost between 300 and 400

filter_2=(data['cost']>300) & (data['cost']<400)

data[filter_2]
```

[38]:

	ID	age	gender	product	cost	selling_price	quantity
1	2	32	Female	S	342.0	350.0	4.0
4	5	43	Female	E	343.0	359.0	NaN
9	10	46	Female	P	307.0	325.0	3.0

[50]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

[54]:
```
sns.lmplot(data=data, x="age", y='cost', hue='gender')
plt.title("Purchasing Statistics")
plt.show()
```


Purchasing Statistics
cost
1600
1400
1200
1000
800
600
400
200
25
30
35
40
45
50
age
gender
Male
Female

[ ]: