covariance_correlation

April 1, 2022

Student Name : Sartaj Ahmed Salman

Email: s2140019@edu.cc.uec.ac.jp

Phd Student At UEC Tokyo, Japan

Address: From Skardu, Pakistan

1 Correlation

What is Correlation?

Variables within a dataset can be related for lots of reasons.

Types: - Pearson's - Spearman's rho - kendall's tau

For example: - One variable could cause or depend on the values of another variable. - One variable could be lightly associated with another variable. - Two variables could depend on a third unknown variable.

Positive Correlation: both variables change in the same direction.

Neutral Correlation: No relationship in the change of the variables.

Negative Correlation: variables change in opposite directions.

2 Covariance

- Variables can be related by a linear relationship. This is a relationship that is consistently additive across the two data samples.
- This relationship can be summarized between two variables, called the covariance.
- The sign of the covariance can be interpreted as whether the two variables change in the same direction (positive) or change in different directions (negative).
- The magnitude of the covariance is not easily interpreted. A covariance value of zero indicates that both variables are completely independent.

[ ]:

```
# Import Libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

[ ]:

```
# Python code to demonstrate the
# use of numpy.cov
import numpy as np

x = [1.23, 2.12, 3.34, 4.5]

y = [2.56, 2.89, 3.76, 3.95]

# find out covariance with respect to columns
cov_mat = np.stack((x, y), axis = 0)
print("stacked values : " ,cov_mat)
print("Cov of the given values : ", np.cov(cov_mat))
```

```
stacked values :  [[1.23 2.12 3.34 4.5 ]
 [2.56 2.89 3.76 3.95]]
Cov of the given values :  [[2.03629167 0.9313    ]
 [0.9313     0.4498    ]]
```

3 Correlation instead of Covariance

[ ]:

```
# Loading data sets
kashti = sns.load_dataset("titanic")
phool = sns.load_dataset("iris")
ticket = pd.read_csv("Sample.csv")
```

[ ]:

```
kashti.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 16 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   891 non-null    int64
 1   survived     891 non-null    int64
 2   pclass       891 non-null    int64
 3   sex          891 non-null    object
 4   age          714 non-null    float64
 5   sibsp        891 non-null    int64
 6   parch        891 non-null    int64
 7   fare         891 non-null    float64
 8   embarked     889 non-null    object
 9   class        891 non-null    category
 10  who          891 non-null    object
 11  adult_male   891 non-null    bool
 12  deck         203 non-null    category
 13  embark_town  889 non-null    object
 14  alive        891 non-null    object
 15  alone        891 non-null    bool
```

```
dtypes: bool(2), category(2), float64(2), int64(5), object(5)
memory usage: 87.6+ KB
```

[ ]:

```python
ticket.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 5 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   purchase_days_before_daprture  30000 non-null  int64  
 1   airline                        30000 non-null  object 
 2   baggage_weight                 30000 non-null  float64
 3   baggage_pieces                 30000 non-null  int64  
 4   price                          30000 non-null  float64
dtypes: float64(2), int64(2), object(1)
memory usage: 1.1+ MB
```

[ ]:

```python
# Correlation
ticket.corr()
```

[ ]:

```
                               purchase_days_before_daprture  baggage_weight  \
purchase_days_before_daprture                       1.000000       -0.018565   
baggage_weight                                     -0.018565        1.000000   
baggage_pieces                                      0.002591       -0.164157   
price                                              -0.168927        0.167041   

                               baggage_pieces     price  
purchase_days_before_daprture        0.002591 -0.168927  
baggage_weight                      -0.164157  0.167041  
baggage_pieces                       1.000000  0.133475  
price                                0.133475  1.000000  
```

[ ]:

```python
# Pearson correlation
cor = ticket.corr(method='pearson') # When data is guassian
```

[ ]:

```python
# Spearman correlation
corr=ticket.corr(method='spearman') # When data is non-guassian
```

[ ]:

```python
cor
```

[ ]:

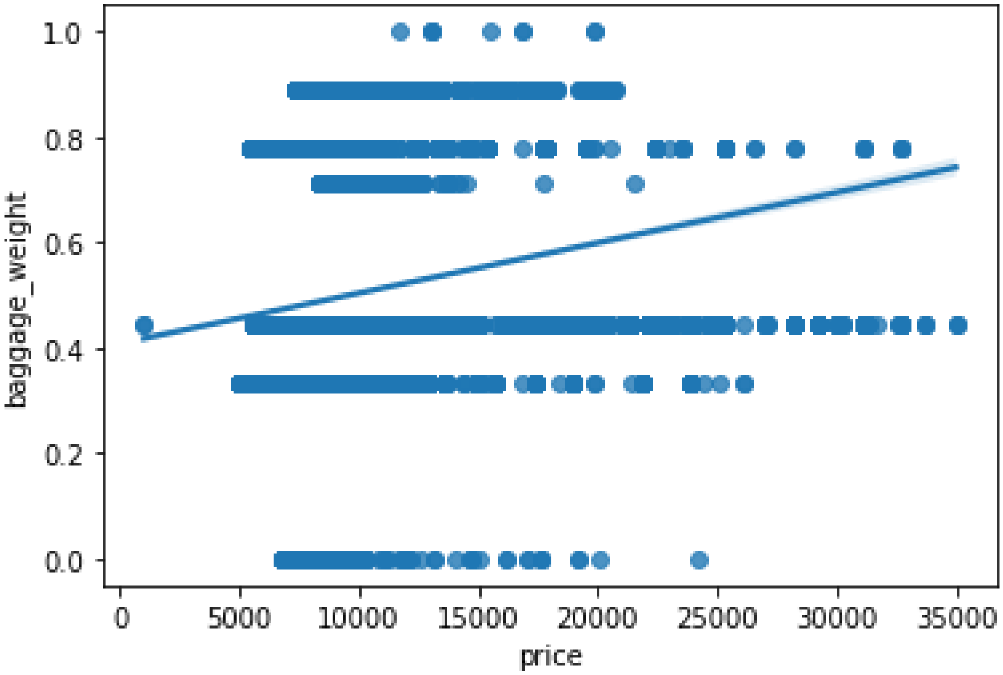
```
                               purchase_days_before_daprture  baggage_weight  \
purchase_days_before_daprture                       1.000000       -0.018565   
baggage_weight                                     -0.018565        1.000000   
baggage_pieces                                      0.002591       -0.164157   
price                                              -0.168927        0.167041   

                               baggage_pieces     price  
```

```
purchase_days_before_daprture        0.002591 -0.168927
baggage_weight                      -0.164157  0.167041
baggage_pieces                       1.000000  0.133475
price                                0.133475  1.000000
```

[ ]:
```python
# Positive correlation
sns.regplot(ticket['price'],ticket['baggage_weight'], data=ticket)
```

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

[ ]:
```
<AxesSubplot:xlabel='price', ylabel='baggage_weight'>
```

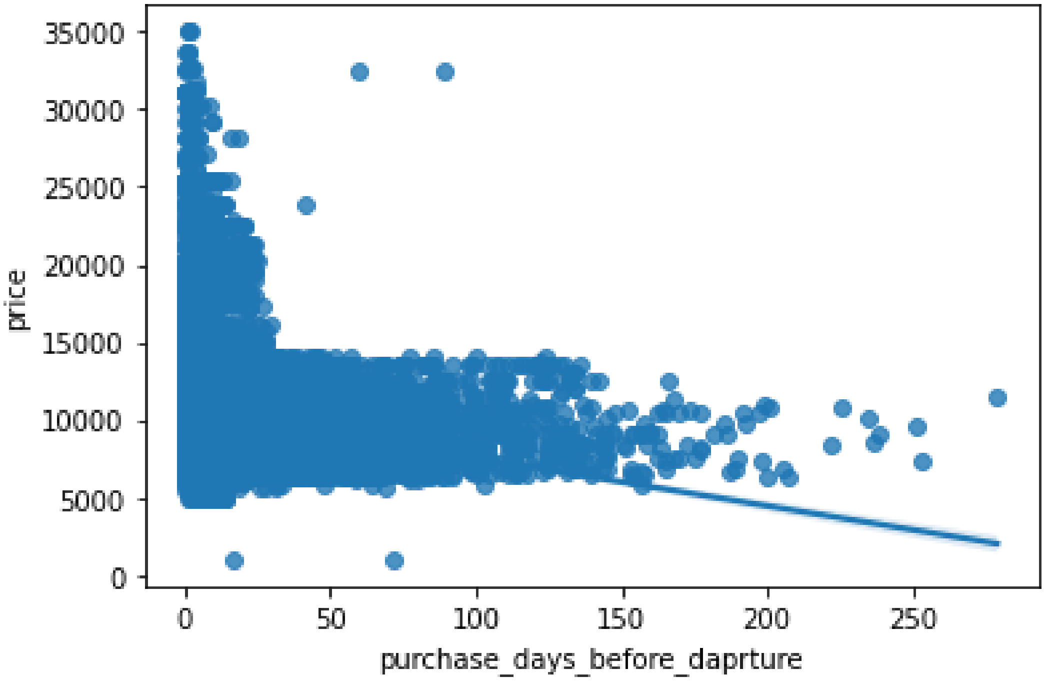


[ ]:
```python
# Nagative correlation
sns.regplot(ticket['purchase_days_before_daprture'],ticket['price'],␣
 ↪data=ticket)
```

```
C:\Users\Sartaj\anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
```

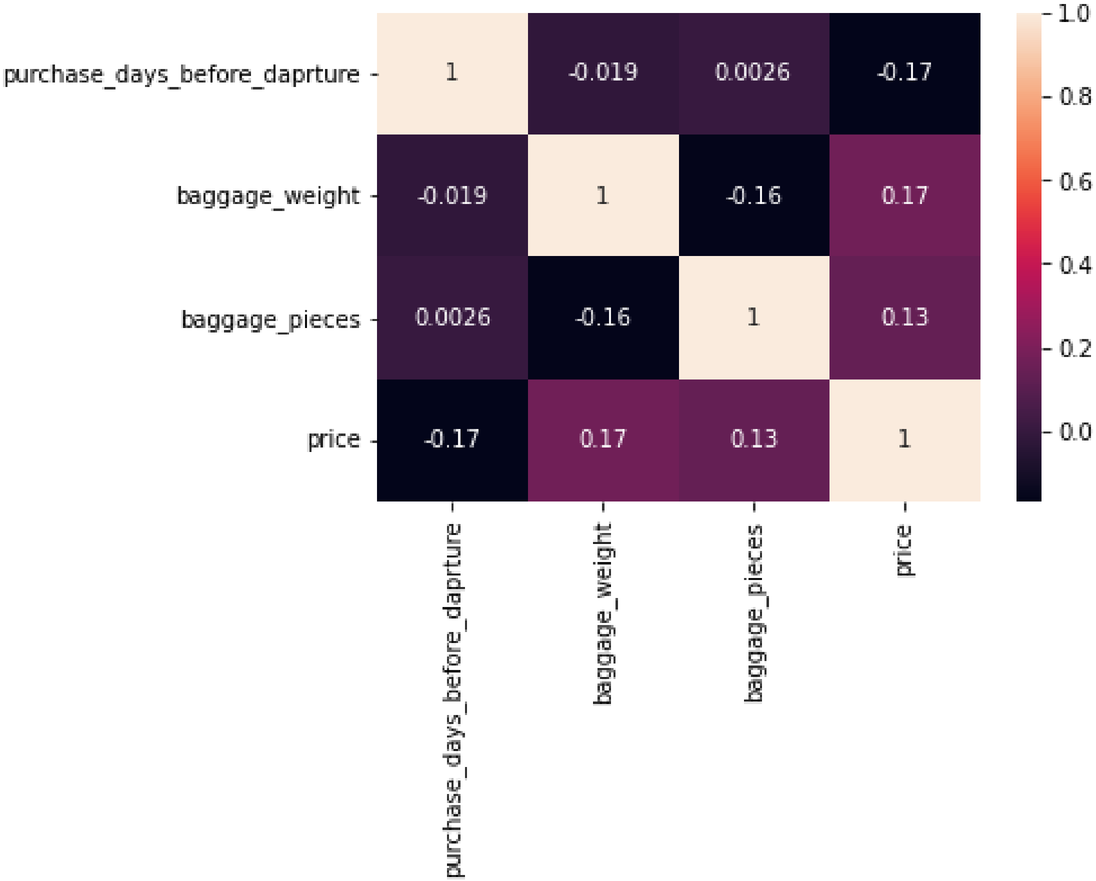
```
 warnings.warn(
```

[ ]:
```
<AxesSubplot:xlabel='purchase_days_before_daprture', ylabel='price'>
```



[ ]:
```
# Heatmap
sns.heatmap(cor, annot=True)
```

[ ]:
```
<AxesSubplot:>
```

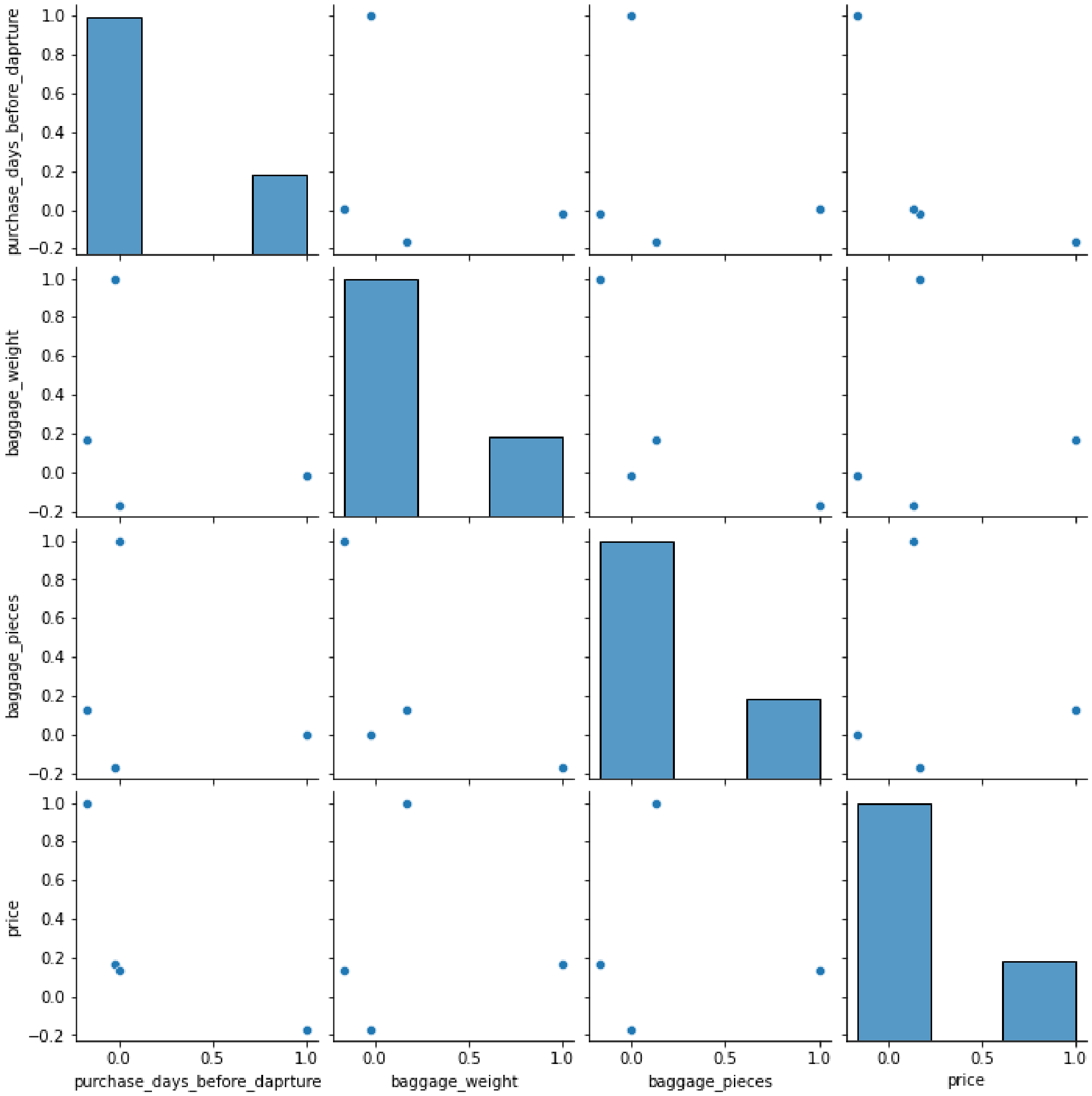


```
[ ]: cor.style.background_gradient('coolwarm')
```

```
[ ]: <pandas.io.formats.style.Styler at 0x1f40e1ee700>
```

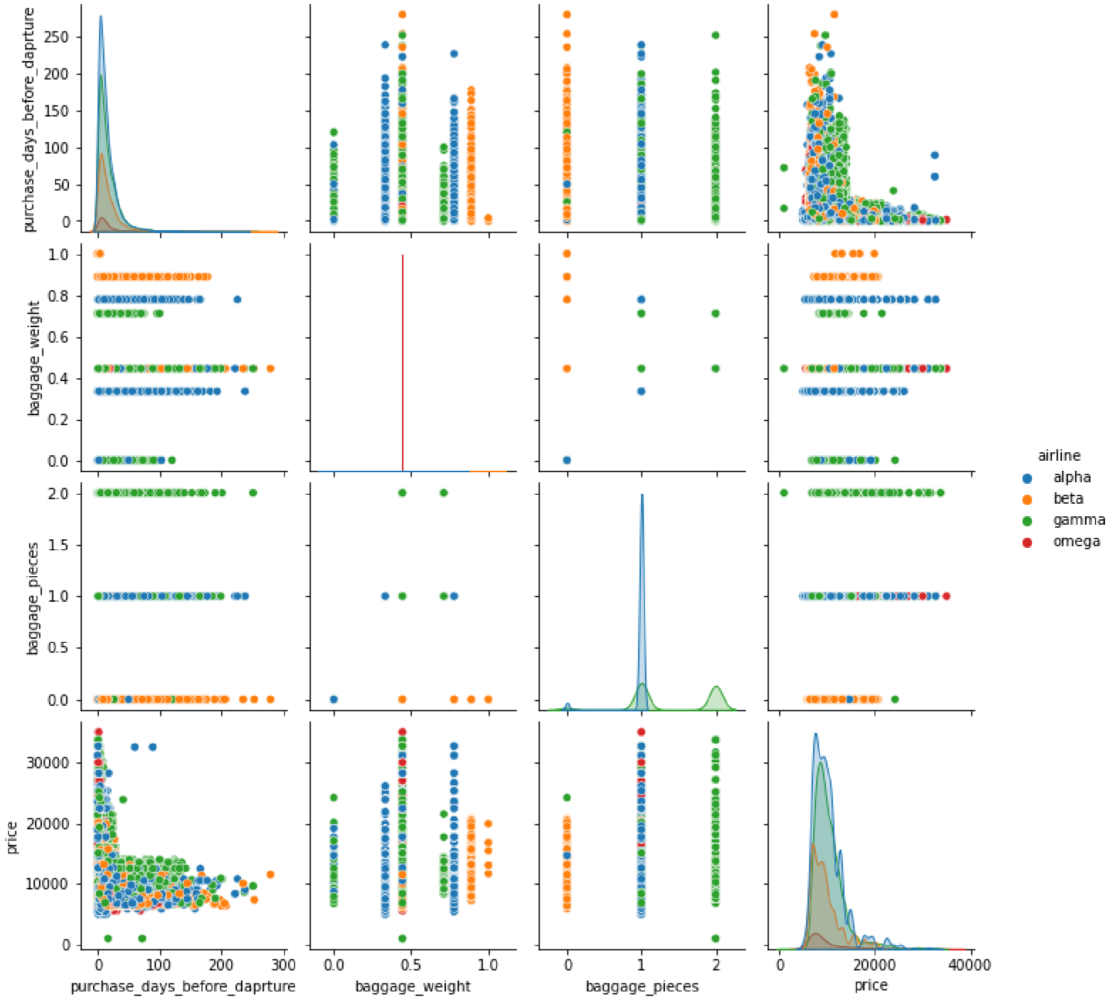
```
[ ]: # Pair plot
     sns.pairplot(cor)
```

```
[ ]: <seaborn.axisgrid.PairGrid at 0x1f40eb12850>
```



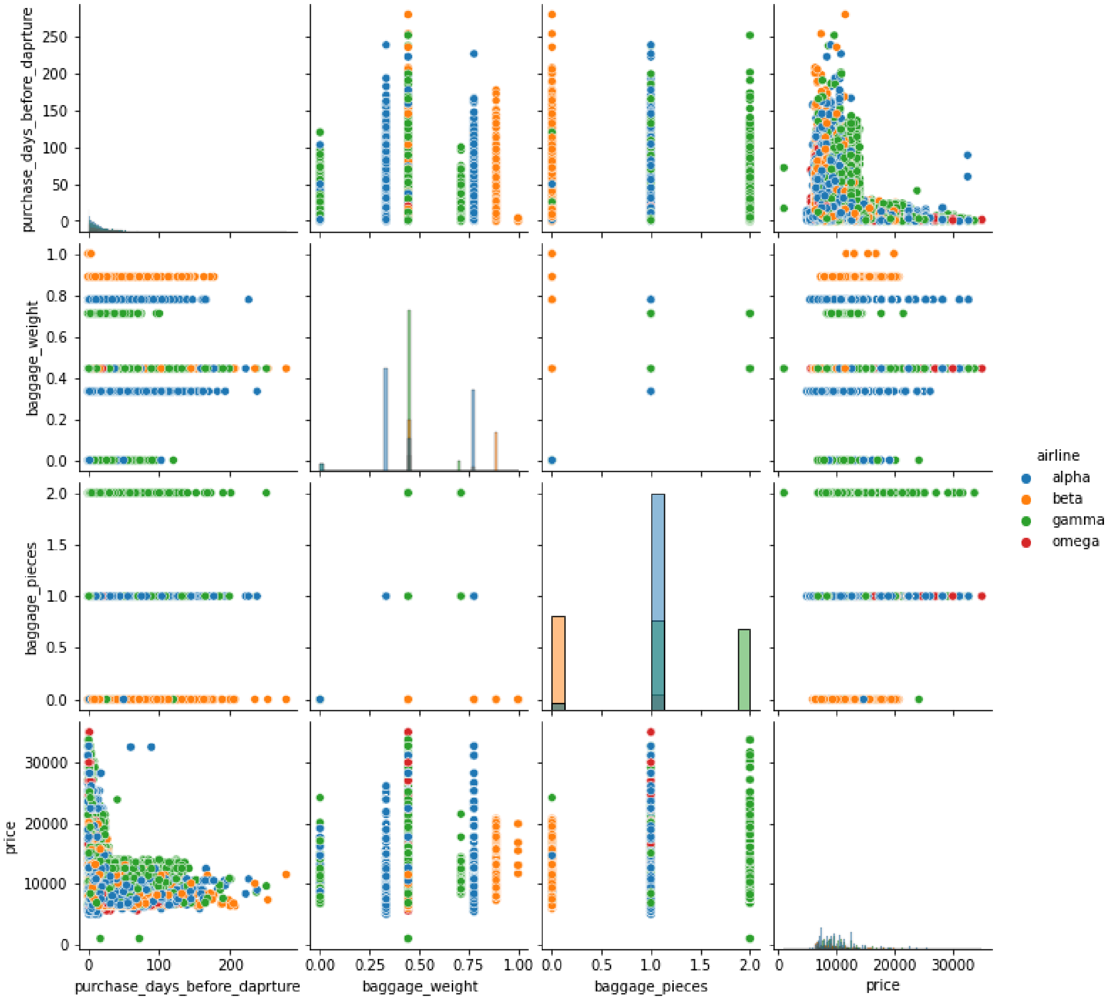
```
[ ]: # Pair plot
     sns.pairplot(ticket, hue='airline')
```

```
[ ]: <seaborn.axisgrid.PairGrid at 0x1f412def670>
```



```
[ ]: # Pair plot with histogram
     sns.pairplot(ticket, hue='airline', diag_kind='hist')
```

```
[ ]: <seaborn.axisgrid.PairGrid at 0x1f41340f160>
```

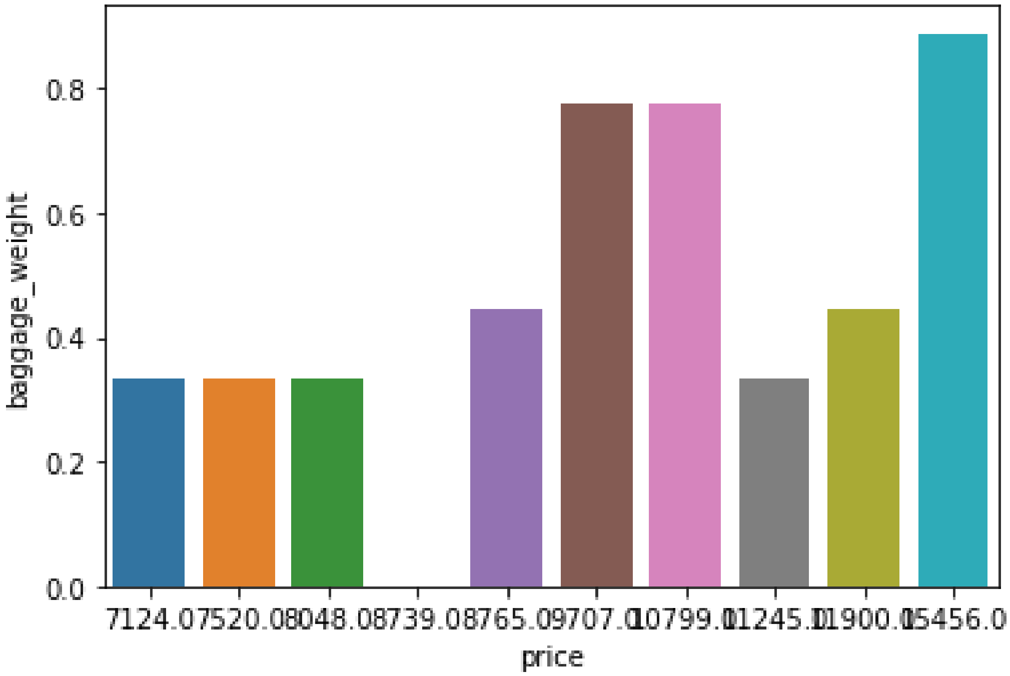


```
[ ]: samp = ticket.sample(10)
```

```
[ ]: sns.barplot(samp['price'],samp['baggage_weight'], data=samp)
```

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

```
[ ]: <AxesSubplot:xlabel='price', ylabel='baggage_weight'>
```

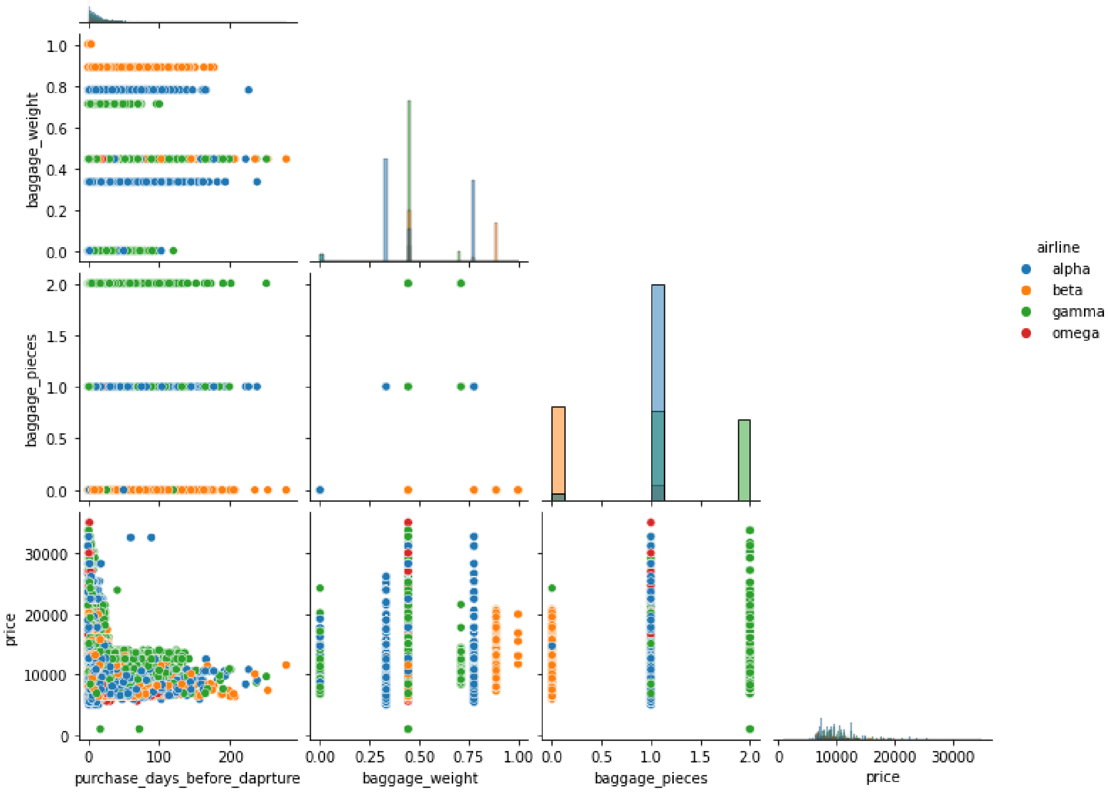


```
[ ]: # Using Scipy stats
     from scipy.stats import pearsonr
     corr, _ = pearsonr(ticket['price'], ticket['baggage_weight'])
     print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.167
```

```
[ ]: # Pair plot with histogram
     sns.pairplot(ticket, hue='airline', diag_kind='hist', corner=True)
```

```
[ ]: <seaborn.axisgrid.PairGrid at 0x1f4158affd0>
```



```
[ ]: phool.head()
```

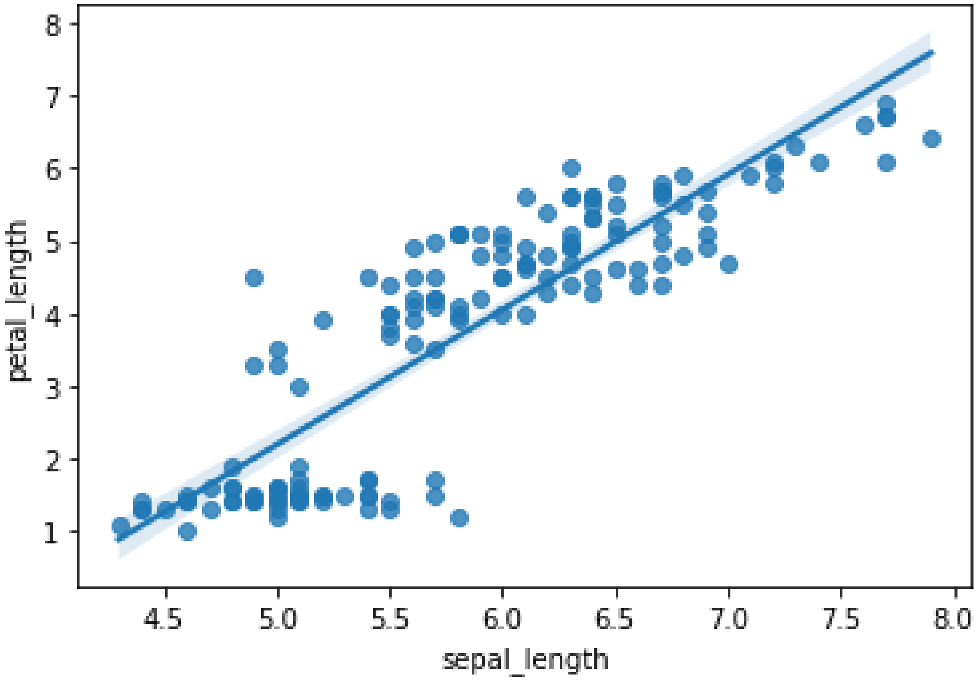
```
[ ]:    sepal_length  sepal_width  petal_length  petal_width species
     0           5.1          3.5           1.4          0.2  setosa
     1           4.9          3.0           1.4          0.2  setosa
     2           4.7          3.2           1.3          0.2  setosa
     3           4.6          3.1           1.5          0.2  setosa
     4           5.0          3.6           1.4          0.2  setosa
```

```
[ ]: # Positive correlation
     sns.regplot(phool['sepal_length'],phool['petal_length'], data=phool)
```

```
C:\Users\Sartaj\anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
```

```
      warnings.warn(
```

```
[ ]: <AxesSubplot:xlabel='sepal_length', ylabel='petal_length'>
```


petal_length
sepal_length