NPTEL – PYTHON FOR DATA SCIENCE

ASSIGNMENT 3 – SOLUTION

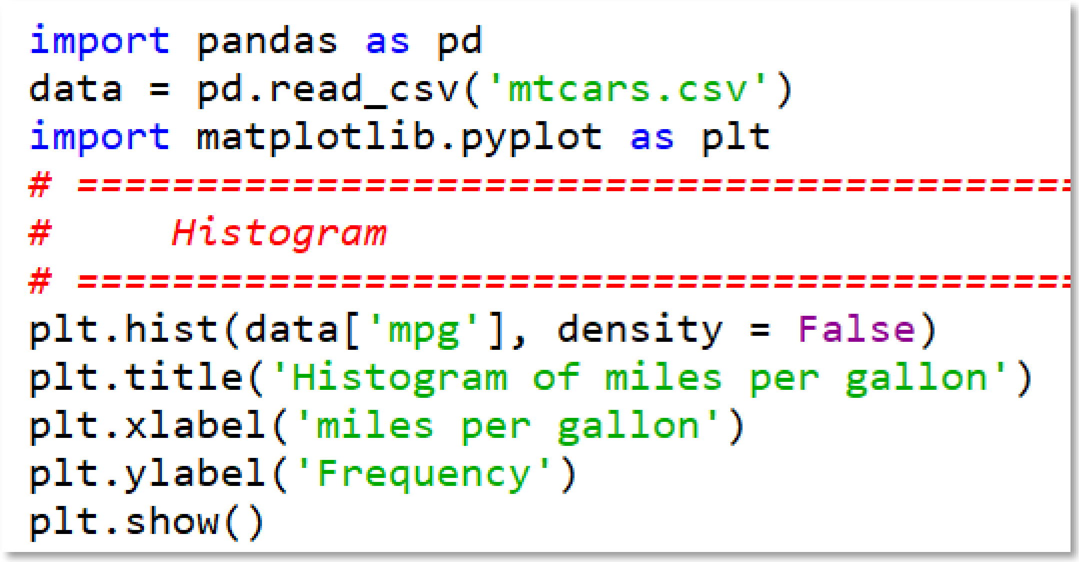
1. Both read_csv & read_table are used for reading a text file in python.

   Ans: d

2. Perhaps the simplest of all plots in the visualization is line plot. The command for line plot is plot ( ). By default, plot ( ) command from the matplotlib library gives a line plot.

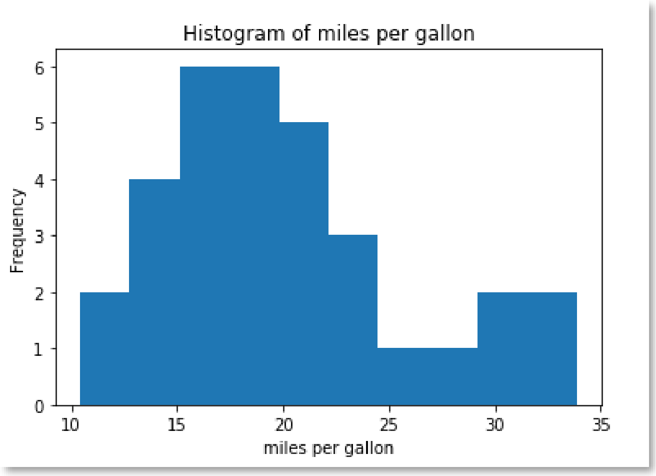
   Ans: a

3. **INPUT:**



```
import pandas as pd
data = pd.read_csv('mtcars.csv')
import matplotlib.pyplot as plt
# =========================================
#      Histogram
# =========================================
plt.hist(data['mpg'], density = False)
plt.title('Histogram of miles per gallon')
plt.xlabel('miles per gallon')
plt.ylabel('Frequency')
plt.show()
```

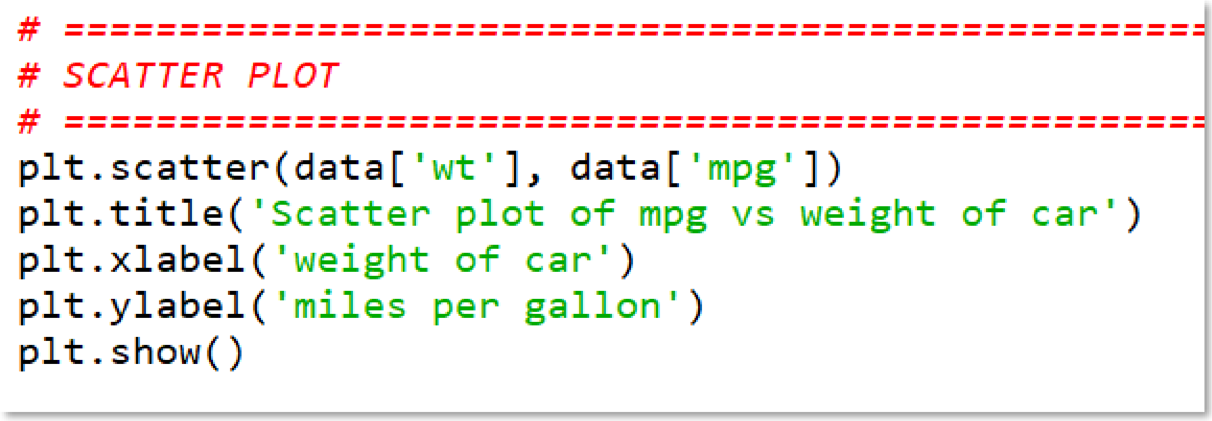
OUTPUT:



The interval 15 – 20 has the highest frequency

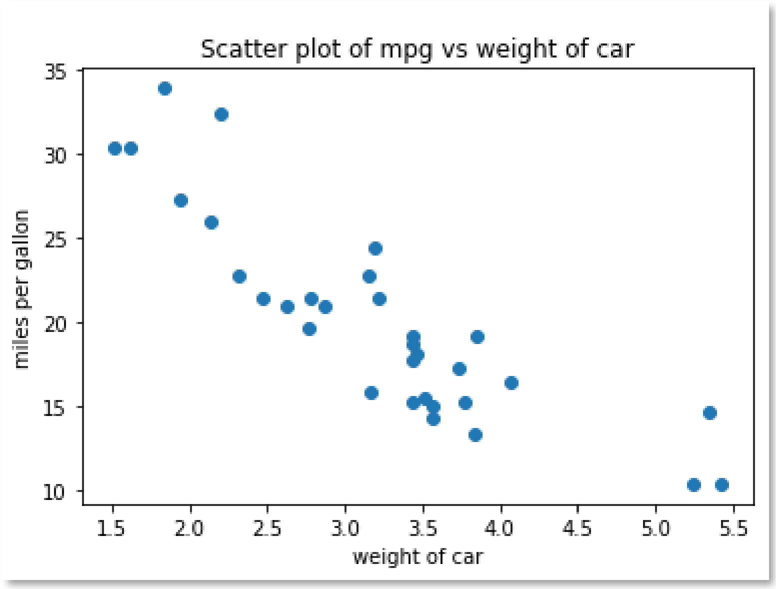
Ans: b

4. **INPUT:**



```
# ==================================================
# SCATTER PLOT
# ==================================================
plt.scatter(data['wt'], data['mpg'])
plt.title('Scatter plot of mpg vs weight of car')
plt.xlabel('weight of car')
plt.ylabel('miles per gallon')
plt.show()
```

OUTPUT:



Inference:

As weight of the car increases, the mpg decreases

Ans: a

5. The plot to show the relationship between two numerical variables is scatter plot. From seaborn library, regplot( ) is used to plot scatter plot.

Ans: d

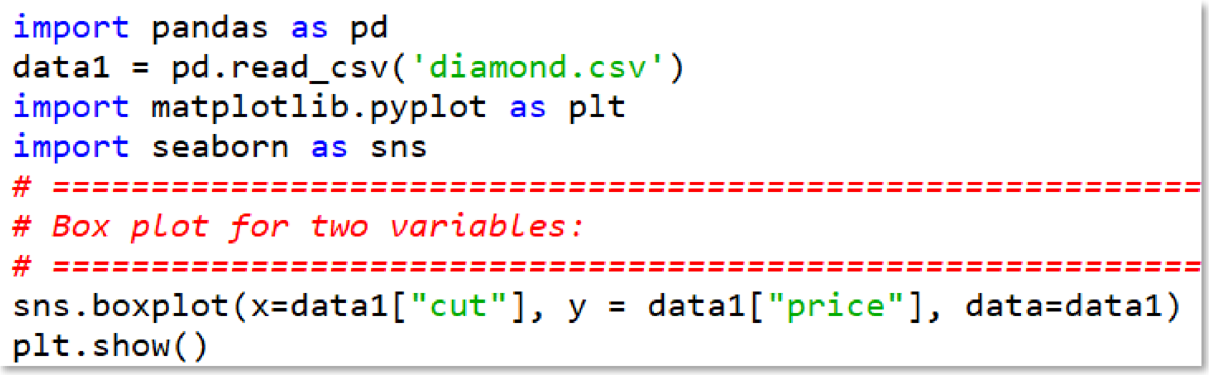
6. The lmplot( ) function combines regplot() and FacetGrid. It is intended as a convenient interface to plot scatter plots across conditional subsets of a dataset.

Ans: c

7. A box-and-whisker plot shows the visual representation of the statistical five number summary using a method that is a function of the inter-quartile range.

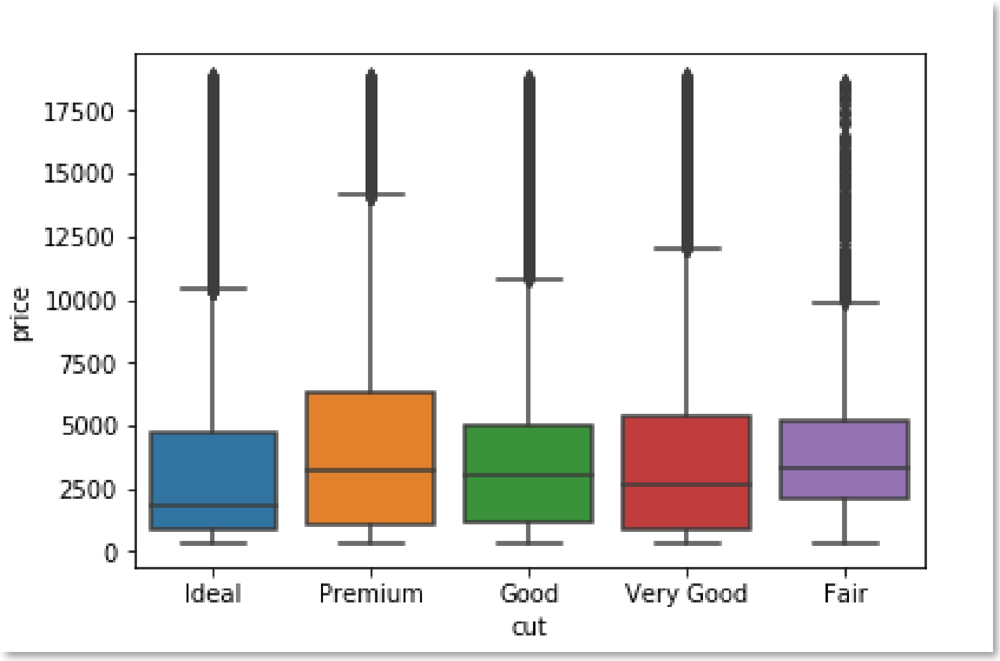
Ans: b

8. **INPUT:**

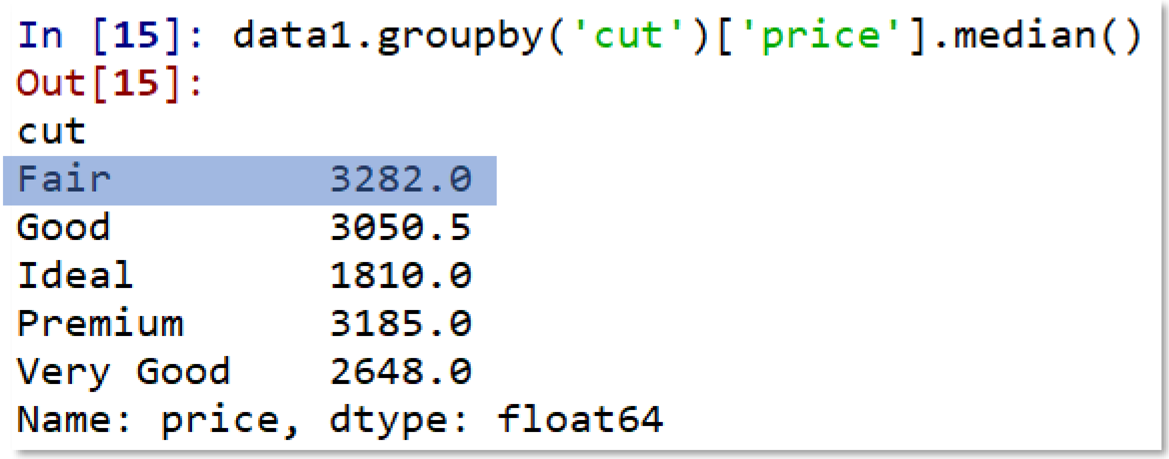


```python
import pandas as pd
data1 = pd.read_csv('diamond.csv')
import matplotlib.pyplot as plt
import seaborn as sns
# ==========================================================
# Box plot for two variables:
# ==========================================================
sns.boxplot(x=data1["cut"], y = data1["price"], data=data1)
plt.show()
```

OUTPUT:



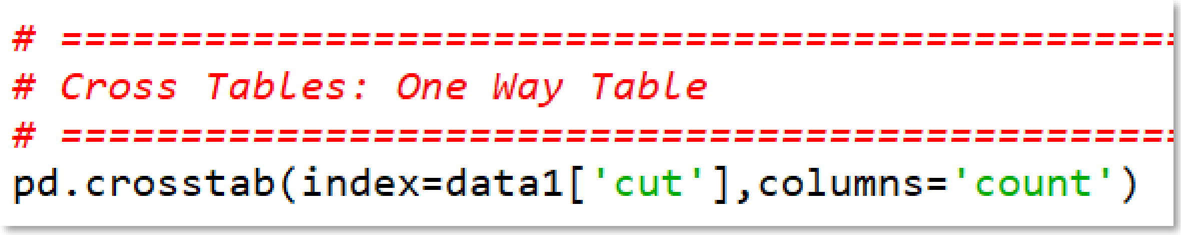
REFERENCE:



```
In [15]: data1.groupby('cut')['price'].median()
Out[15]:
cut
Fair          3282.0
Good          3050.5
Ideal         1810.0
Premium       3185.0
Very Good     2648.0
Name: price, dtype: float64
```

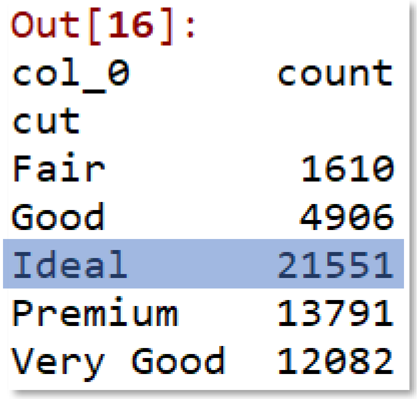
Ans: d

9. **INPUT:**



```
# ==============================================
# Cross Tables: One Way Table
# ==============================================
pd.crosstab(index=data1['cut'],columns='count')
```

OUTPUT:



```
Out[16]:
col_0       count
cut
Fair         1610
Good         4906
Ideal       21551
Premium     13791
Very Good   12082
```

Ans: d

10. The probability of two different events occurring at the same time is known as joint probability

Ans: c

11. The command to detect NaN (null) values in pandas dataframes are

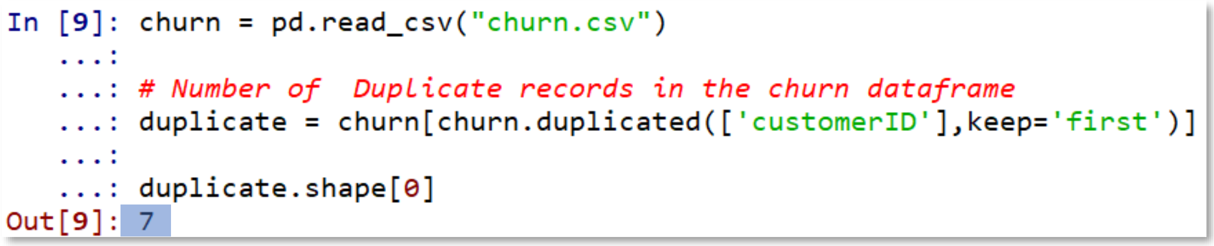
isna( ) or .isnull( )

Ans: d

12. DataFrame.column_name.dtypes, DataFrame.column_name.ftypes, and DataFrame.column.dtype are used to identify the data type of a column in a dataframe

    Ans: d

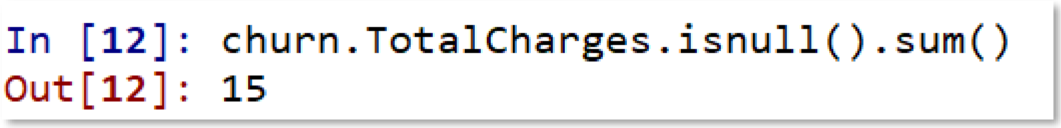
13. **CODE:**



```
In [9]: churn = pd.read_csv("churn.csv")
   ...:
   ...: # Number of  Duplicate records in the churn dataframe
   ...: duplicate = churn[churn.duplicated(['customerID'],keep='first')]
   ...:
   ...: duplicate.shape[0]
Out[9]: 7
```

Ans: a

14. **CODE:**



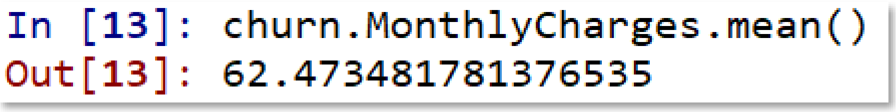
```
In [12]: churn.TotalCharges.isnull().sum()
Out[12]: 15
```

There are 15 records missing in the variable ***TotalCharges***

Ans: c

15. The average monthly charge paid by the customer for the services he/she has signed up for is $ 62.47

CODE:

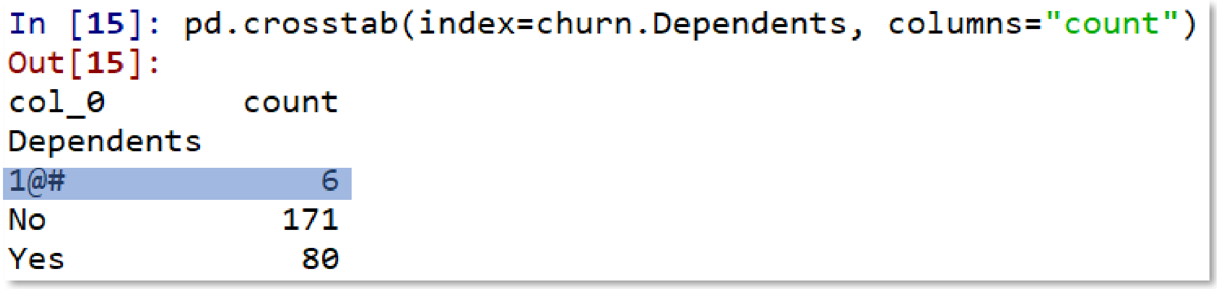


```
In [13]: churn.MonthlyCharges.mean()
Out[13]: 62.473481781376535
```

Ans: b

16. Under the variable ***Dependents*** of churn dataframe, there are 6 records that have ***“1@#”***

CODE:

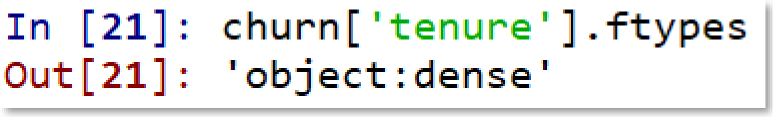


```
In [15]: pd.crosstab(index=churn.Dependents, columns="count")
Out[15]:
col_0         count
Dependents
1@#               6
No              171
Yes              80
```

Ans: b

17. The data type of the variable ***tenure*** from the churn dataframe is ‘Object’

CODE:

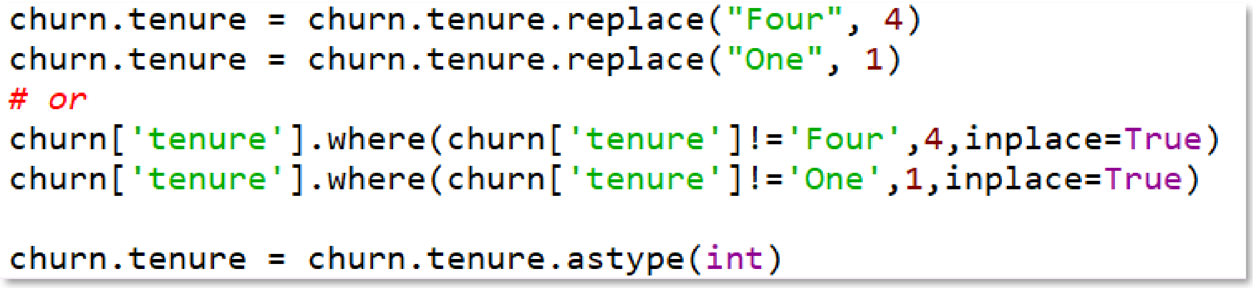


```
In [21]: churn['tenure'].ftypes
Out[21]: 'object:dense'
```

Ans: d

18. Pandas.Dataframe.where(), pandas.Dataframe.replace and numpy.where() can be used to replace ‘Four’ by 4 and ‘One’ by 1 under the variable ***“tenure”***

CODE:



```
churn.tenure = churn.tenure.replace("Four", 4)
churn.tenure = churn.tenure.replace("One", 1)
# or
churn['tenure'].where(churn['tenure']!='Four',4,inplace=True)
churn['tenure'].where(churn['tenure']!='One',1,inplace=True)

churn.tenure = churn.tenure.astype(int)
```

Ans: d

19. The Pearson correlation coefficient value ranges from -1 to 1

Ans: b

20. Indentation is used to mark the beginning of sequence of operations in control structures

Ans: c