ass2-data-wrangling-ii

May 8, 2024

1 DATA SCIENCE AND BIG DATA ANALYTICS LABORATORY Data Wrangling II

Name-: More Pankaj Sampat

Branch-: TE Comp

Batch-: T2

Roll No-: 054

Date of Completion-: 1/3/2023

Problem Statement-: **C**reate an "Academic performance" dataset of students and perform the following operations using Python. 1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them. 2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them. 3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

[41]:
```
import pandas as pd
import numpy  as np
```

[42]:
```
dataframe=pd.read_csv('academic.csv',na_values='?')
dataframe
```

[42]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
0	M	KW	KuwaIT	lowerlevel	G-04	A
1	M	KW	KuwaIT	lowerlevel	G-04	A
2	M	KW	KuwaIT	lowerlevel	G-04	A
3	M	KW	KuwaIT	lowerlevel	G-04	A
4	M	KW	KuwaIT	lowerlevel	G-04	A
..	…	…	…	…	…	…
475	F	Jordan	Jordan	MiddleSchool	G-08	A
476	F	Jordan	Jordan	MiddleSchool	G-08	A
477	F	Jordan	Jordan	MiddleSchool	G-08	A
478	F	Jordan	Jordan	MiddleSchool	G-08	A
479	F	Jordan	Jordan	MiddleSchool	G-08	A

```
         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father         15.0                16
1           IT        F   Father         20.0                20
2           IT        F   Father         10.0                 7
3           IT        F   Father         30.0                25
4           IT        F   Father         40.0                50
..         …       …      …           …                 …
475  Chemistry        T   Father          5.0                 4
476    Geology        F   Father         50.0                77
477    Geology        S   Father         55.0                74
478    History        F   Father         30.0                17
479    History        S   Father         35.0                14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 …         …                   …
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       …              …   …
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]
```

[43]:
```
dataframe.isnull()
```

[43]:
```
     gender  NationalITy  PlaceofBirth  StageID  GradeID  SectionID  Topic  \
0     False        False         False    False    False      False  False
1     False        False         False    False    False      False  False
```

```
2     False        False         False    False    False      False  False
3     False        False         False    False    False      False  False
4     False        False         False    False    False      False  False
..      …            …             …        …        …          …    …
475   False        False         False    False    False      False  False
476   False        False         False    False    False      False  False
477   False        False         False    False    False      False  False
478   False        False         False    False    False      False  False
479   False        False         False    False    False      False  False

     Semester  Relation  raisedhands  VisITedResources  AnnouncementsView  \
0       False     False        False             False              False
1       False     False        False             False              False
2       False     False        False             False              False
3       False     False        False             False              False
4       False     False        False             False              False
..        …         …            …                 …                  …
475     False     False        False             False              False
476     False     False        False             False              False
477     False     False        False             False              False
478     False     False        False             False              False
479     False     False        False             False              False

     Discussion  ParentAnsweringSurvey  ParentschoolSatisfaction  \
0         False                  False                     False
1         False                  False                     False
2         False                  False                     False
3         False                  False                     False
4         False                  False                     False
..          …                      …                         …
475       False                  False                     False
476       False                  False                     False
477       False                  False                     False
478       False                  False                     False
479       False                  False                     False

     StudentAbsenceDays  Class
0                 False  False
1                 False  False
2                 False  False
3                 False  False
4                 False  False
..                  …    …
475               False  False
476               False  False
477               False  False
478               False  False
```

```
479               False  False

[480 rows x 17 columns]
```

[44]:
```
dataframe['Topic'] =  dataframe['Topic'].fillna('0')
print(dataframe)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04         A
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        M          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     …          …           …           …     …       …
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father         15.0                16
1           IT        F   Father         20.0                20
2           IT        F   Father         10.0                 7
3           IT        F   Father         30.0                25
4           IT        F   Father         40.0                50
..         …      …      …           …                 …
475  Chemistry        T   Father          5.0                 4
476    Geology        F   Father         50.0                77
477    Geology        S   Father         55.0                74
478    History        F   Father         30.0                17
479    History        S   Father         35.0                14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 …         …                   …
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
```

```
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       …                 …    …
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]
```

[45]:
```
print(dataframe.isnull().sum())
```

```
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     2
SectionID                   0
Topic                       0
Semester                    0
Relation                    8
raisedhands                 2
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

[46]:
```
dataframe['Relation'].dropna(axis=0, inplace=False)
```

[46]:
```
0      Father
1      Father
2      Father
3      Father
4      Father
        …
475    Father
476    Father
477    Father
478    Father
```

```
479    Father
Name: Relation, Length: 472, dtype: object
```

[47]:
```
print(dataframe.isnull().sum())
```

```
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     2
SectionID                   0
Topic                       0
Semester                    0
Relation                    8
raisedhands                 2
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

[48]:
```
dataframe = dataframe.dropna(subset=['Relation', 'GradeID'])
print(dataframe)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04         A
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        M          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father         15.0                16
1           IT        F   Father         20.0                20
2           IT        F   Father         10.0                 7
3           IT        F   Father         30.0                25
4           IT        F   Father         40.0                50
..         ...      ...      ...          ...               ...
475  Chemistry        T   Father          5.0                 4
```

```
476    Geology        F   Father         50.0               77
477    Geology        S   Father         55.0               74
478    History        F   Father         30.0               17
479    History        S   Father         35.0               14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 …          …                    …
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       …                 …   …
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[470 rows x 17 columns]
```

[49]:
```
print(dataframe.isnull().sum())
```

```
gender                    0
NationalITy               0
PlaceofBirth              0
StageID                   0
GradeID                   0
SectionID                 0
Topic                     0
Semester                  0
Relation                  0
raisedhands               2
VisITedResources          0
AnnouncementsView         0
Discussion                0
```

```
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

[50]:
```
dataframe['raisedhands'].replace(np.NaN,dataframe['raisedhands'].
 ↪mean(),inplace=True)

dataframe
```

```
/tmp/ipykernel_8758/3644195618.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe['raisedhands'].replace(np.NaN,dataframe['raisedhands'].mean(),inplac
e=True)
```

[50]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	15.0	16	2	20	Yes
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	20.0	20			
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	10.0	7			
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	30.0	25			
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	40.0	50			
..	…	…	…	…	…	…	…	…	…	…	…			
475	F	Jordan	Jordan	MiddleSchool	G-08	A	Chemistry	T	Father	5.0	4			
476	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	F	Father	50.0	77			
477	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	S	Father	55.0	74			
478	F	Jordan	Jordan	MiddleSchool	G-08	A	History	F	Father	30.0	17			
479	F	Jordan	Jordan	MiddleSchool	G-08	A	History	S	Father	35.0	14			

```
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 …           …                   …
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       …                 …   …
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[470 rows x 17 columns]
```

[51]:
```
print(dataframe.isnull().sum())
```

```
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 0
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

[52]:
```
dataframe.describe()
```

[52]:

	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	470.000000	470.000000	470.000000	470.000000
mean	46.831197	54.506383	37.659574	43.351064
std	30.833471	33.053793	26.536756	27.720855
min	0.000000	0.000000	0.000000	1.000000
25%	16.250000	20.000000	13.250000	20.000000
50%	50.000000	64.500000	33.000000	39.500000
75%	75.000000	84.000000	58.000000	70.000000
max	100.000000	99.000000	98.000000	99.000000

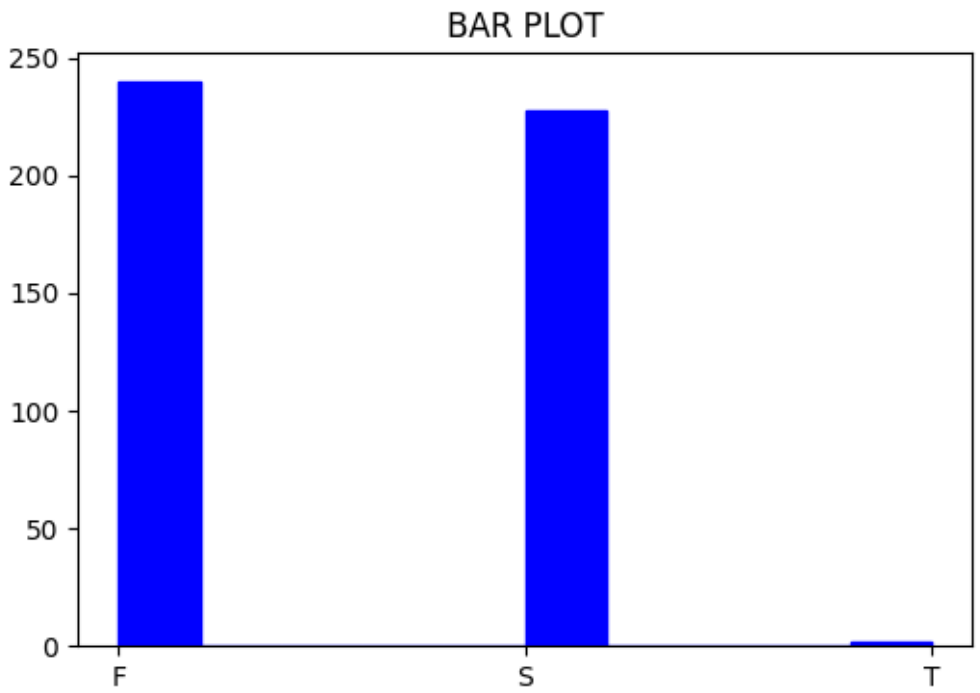
2 Outliers

[53]:
```
import matplotlib.pyplot as plt
```

[54]:
```
dataframe['Semester'].value_counts()
```

[54]:
```
F    240
S    228
T      2
Name: Semester, dtype: int64
```

[55]:
```
fig,x=plt.subplots(figsize=(6,4))
ax=plt.hist(dataframe['Semester'],color='b',edgecolor='b')
plt.title('BAR PLOT')
plt.show()
```



[56]:
```
dataframe[dataframe['Semester'].str.contains('T')==False]
dataframe
```

[56]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	\
0	M	KW	KuwaIT	lowerlevel	G-04	A	
1	M	KW	KuwaIT	lowerlevel	G-04	A	
2	M	KW	KuwaIT	lowerlevel	G-04	A	
3	M	KW	KuwaIT	lowerlevel	G-04	A	
4	M	KW	KuwaIT	lowerlevel	G-04	A	
..	…	…	…	…	…	…	
475	F	Jordan	Jordan	MiddleSchool	G-08	A	
476	F	Jordan	Jordan	MiddleSchool	G-08	A	
477	F	Jordan	Jordan	MiddleSchool	G-08	A	
478	F	Jordan	Jordan	MiddleSchool	G-08	A	
479	F	Jordan	Jordan	MiddleSchool	G-08	A	

	Topic	Semester	Relation	raisedhands	VisITedResources	\
0	IT	F	Father	15.0	16	
1	IT	F	Father	20.0	20	
2	IT	F	Father	10.0	7	
3	IT	F	Father	30.0	25	
4	IT	F	Father	40.0	50	
..	…	…	…	…	…	
475	Chemistry	T	Father	5.0	4	

```
476    Geology        F   Father         50.0               77
477    Geology        S   Father         55.0               74
478    History        F   Father         30.0               17
479    History        S   Father         35.0               14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 …         …                     …
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       …                 …   …
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[470 rows x 17 columns]
```

[57]:
```
dataframe['ParentAnsweringSurvey'].value_counts()
```

[57]:
```
Yes    262
No     208
Name: ParentAnsweringSurvey, dtype: int64
```

[58]:
```
dataframe.drop(['ParentAnsweringSurvey'], axis=1,inplace=True)
dataframe
```

```
/tmp/ipykernel_8758/22155142.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
dataframe.drop(['ParentAnsweringSurvey'], axis=1,inplace=True)
```

[58]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
0	M	KW	KuwaIT	lowerlevel	G-04	A
1	M	KW	KuwaIT	lowerlevel	G-04	A
2	M	KW	KuwaIT	lowerlevel	G-04	A
3	M	KW	KuwaIT	lowerlevel	G-04	A
4	M	KW	KuwaIT	lowerlevel	G-04	A
..	…	…	…	…	…	…
475	F	Jordan	Jordan	MiddleSchool	G-08	A
476	F	Jordan	Jordan	MiddleSchool	G-08	A
477	F	Jordan	Jordan	MiddleSchool	G-08	A
478	F	Jordan	Jordan	MiddleSchool	G-08	A
479	F	Jordan	Jordan	MiddleSchool	G-08	A

	Topic	Semester	Relation	raisedhands	VisITedResources
0	IT	F	Father	15.0	16
1	IT	F	Father	20.0	20
2	IT	F	Father	10.0	7
3	IT	F	Father	30.0	25
4	IT	F	Father	40.0	50
..	…	…	…	…	…
475	Chemistry	T	Father	5.0	4
476	Geology	F	Father	50.0	77
477	Geology	S	Father	55.0	74
478	History	F	Father	30.0	17
479	History	S	Father	35.0	14

	AnnouncementsView	Discussion	ParentschoolSatisfaction
0	2	20	Good
1	3	25	Good
2	0	30	Bad
3	5	35	Bad
4	12	50	Bad
..	…	…	…
475	5	8	Bad
476	14	28	Bad
477	25	29	Bad
478	14	57	Bad
479	23	62	Bad

	StudentAbsenceDays	Class
0	Under-7	M
1	Under-7	M
2	Above-7	L
3	Above-7	L
4	Above-7	M

```
..                 …   …
475            Above-7     L
476            Under-7     M
477            Under-7     M
478            Above-7     L
479            Above-7     L

[470 rows x 16 columns]
```

3 Normalization

[59]:
```
dataframe['raisedhands']=dataframe['raisedhands']/ dataframe['raisedhands'].
 ↪max()
dataframe['VisITedResources']=dataframe['VisITedResources']/␣
 ↪dataframe['VisITedResources'].max()
```

```
/tmp/ipykernel_8758/586398258.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe['raisedhands']=dataframe['raisedhands']/
dataframe['raisedhands'].max()
/tmp/ipykernel_8758/586398258.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe['VisITedResources']=dataframe['VisITedResources']/
dataframe['VisITedResources'].max()
```

[60]:
```
dataframe[['raisedhands','VisITedResources']].head(10)
```

[60]:

	raisedhands	VisITedResources
0	0.15	0.161616
1	0.20	0.202020
2	0.10	0.070707
3	0.30	0.252525
4	0.40	0.505051
5	0.42	0.303030
6	0.35	0.121212
7	0.50	0.101010
8	0.12	0.212121
9	0.70	0.808081

4 Data Transformation

[61]:
```
dataframe['GradeID'].value_counts()
```

[61]:
```
G-02    142
G-08    116
G-07     98
G-04     48
G-06     32
G-11     13
G-12      9
G-09      5
G-10      4
G-05      3
Name: GradeID, dtype: int64
```

[62]:
```
dataframe.GradeID.replace({"G-02":1, "G-04":2, "G-05":3,"G-06":4,"G-07":
 ↪5,"G-08":6,"G-09":7,"G-10":8,"G-11":9,"G-12":10},inplace=True)
dataframe['GradeID'].value_counts()
```

```
/tmp/ipykernel_8758/2636104831.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe.GradeID.replace({"G-02":1, "G-04":2, "G-05":3,"G-06":4,"G-07":5,"G-0
8":6,"G-09":7,"G-10":8,"G-11":9,"G-12":10},inplace=True)
```

[62]:
```
1     142
6     116
5      98
2      48
4      32
9      13
10      9
7       5
8       4
3       3
Name: GradeID, dtype: int64
```

[63]:
```
dataframe['GradeID']
```

[63]:
```
0      2
1      2
2      2
3      2
4      2
```

```
      ..
475    6
476    6
477    6
478    6
479    6
Name: GradeID, Length: 470, dtype: int64
```

[64]:
```
dataframe['StudentAbsenceDays'].value_counts()
```

[64]:
```
Under-7    283
Above-7    187
Name: StudentAbsenceDays, dtype: int64
```

[65]:
```
dataframe.StudentAbsenceDays.replace({"Under-7":0,"Above-7":1},inplace=True)
dataframe['StudentAbsenceDays'].value_counts()
```

```
/tmp/ipykernel_8758/1125095110.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe.StudentAbsenceDays.replace({"Under-7":0,"Above-7":1},inplace=True)
```

[65]:
```
0    283
1    187
Name: StudentAbsenceDays, dtype: int64
```

[66]:
```
dataframe['StudentAbsenceDays']
```

[66]:
```
0      0
1      0
2      1
3      1
4      1
      ..
475    1
476    0
477    0
478    1
479    1
Name: StudentAbsenceDays, Length: 470, dtype: int64
```

5 Binning

[67]:
```
dataframe
```

[67]:
```
     gender NationalITy PlaceofBirth       StageID  GradeID SectionID  \
0         M          KW       KuwaIT    lowerlevel        2         A
1         M          KW       KuwaIT    lowerlevel        2         A
2         M          KW       KuwaIT    lowerlevel        2         A
3         M          KW       KuwaIT    lowerlevel        2         A
4         M          KW       KuwaIT    lowerlevel        2         A
..      ...         ...          ...           ...      ...       ...
475       F      Jordan       Jordan  MiddleSchool        6         A
476       F      Jordan       Jordan  MiddleSchool        6         A
477       F      Jordan       Jordan  MiddleSchool        6         A
478       F      Jordan       Jordan  MiddleSchool        6         A
479       F      Jordan       Jordan  MiddleSchool        6         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         0.15          0.161616
1            IT        F   Father         0.20          0.202020
2            IT        F   Father         0.10          0.070707
3            IT        F   Father         0.30          0.252525
4            IT        F   Father         0.40          0.505051
..          ...      ...      ...          ...               ...
475  Chemistry        T   Father         0.05          0.040404
476    Geology        F   Father         0.50          0.777778
477    Geology        S   Father         0.55          0.747475
478    History        F   Father         0.30          0.171717
479    History        S   Father         0.35          0.141414

     AnnouncementsView  Discussion ParentschoolSatisfaction  \
0                    2          20                     Good
1                    3          25                     Good
2                    0          30                      Bad
3                    5          35                      Bad
4                   12          50                      Bad
..                 ...         ...                      ...
475                  5           8                      Bad
476                 14          28                      Bad
477                 25          29                      Bad
478                 14          57                      Bad
479                 23          62                      Bad

     StudentAbsenceDays Class
0                     0     M
1                     0     M
2                     1     L
```

```
3                     1     L
4                     1     M
..                  …   …
475                   1     L
476                   0     M
477                   0     M
478                   1     L
479                   1     L

[470 rows x 16 columns]
```

[68]:
```
dataframe['Discussion'].value_counts()
```

[68]:
```
70    23
40    22
33    20
50    18
30    16
      ..
65     1
76     1
55     1
73     1
62     1
Name: Discussion, Length: 90, dtype: int64
```

[69]:
```
min_val=dataframe['Discussion'].min()
max_val=dataframe['Discussion'].max()
print(min_val)
print(max_val)
```

```
1
99
```

[70]:
```
bins=np.linspace(min_val,max_val,4)
bins
```

[70]:
```
array([ 1.        , 33.66666667, 66.33333333, 99.        ])
```

[71]:
```
labels=['small','medium','big']
```

[72]:
```
dataframe['bins']=pd.
  ↪cut(dataframe['Discussion'],bins=bins,labels=labels,include_lowest=True)
```

```
/tmp/ipykernel_8758/3369377855.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataframe['bins']=pd.cut(dataframe['Discussion'],bins=bins,labels=labels,inclu
de_lowest=True)
```

[73]:
```
dataframe['bins'].value_counts()
```

[73]:
```
small     216
medium    127
big       127
Name: bins, dtype: int64
```

[74]:
```
dataframe.head()
```

[74]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic
0	M	KW	KuwaIT	lowerlevel	2	A	IT
1	M	KW	KuwaIT	lowerlevel	2	A	IT
2	M	KW	KuwaIT	lowerlevel	2	A	IT
3	M	KW	KuwaIT	lowerlevel	2	A	IT
4	M	KW	KuwaIT	lowerlevel	2	A	IT

	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView
0	F	Father	0.15	0.161616	2
1	F	Father	0.20	0.202020	3
2	F	Father	0.10	0.070707	0
3	F	Father	0.30	0.252525	5
4	F	Father	0.40	0.505051	12

	Discussion	ParentschoolSatisfaction	StudentAbsenceDays	Class	bins
0	20	Good	0	M	small
1	25	Good	0	M	small
2	30	Bad	1	L	small
3	35	Bad	1	L	medium
4	50	Bad	1	M	medium

[75]:
```
dataframe.describe()
```

[75]:

	GradeID	raisedhands	VisITedResources	AnnouncementsView
count	470.000000	470.000000	470.000000	470.000000
mean	3.904255	0.468312	0.550570	37.659574
std	2.438450	0.308335	0.333877	26.536756
min	1.000000	0.000000	0.000000	0.000000
25%	1.000000	0.162500	0.202020	13.250000
50%	5.000000	0.500000	0.651515	33.000000
75%	6.000000	0.750000	0.848485	58.000000
max	10.000000	1.000000	1.000000	98.000000

```
       Discussion  StudentAbsenceDays
count  470.000000          470.000000
mean    43.351064            0.397872
std     27.720855            0.489980
min      1.000000            0.000000
25%     20.000000            0.000000
50%     39.500000            0.000000
75%     70.000000            1.000000
max     99.000000            1.000000
```

[76]:
```
dataframe
```

[76]:
```
    gender NationalITy PlaceofBirth       StageID  GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel        2         A
1        M          KW       KuwaIT    lowerlevel        2         A
2        M          KW       KuwaIT    lowerlevel        2         A
3        M          KW       KuwaIT    lowerlevel        2         A
4        M          KW       KuwaIT    lowerlevel        2         A
..     …          …           …            …      …       …
475      F      Jordan       Jordan  MiddleSchool        6         A
476      F      Jordan       Jordan  MiddleSchool        6         A
477      F      Jordan       Jordan  MiddleSchool        6         A
478      F      Jordan       Jordan  MiddleSchool        6         A
479      F      Jordan       Jordan  MiddleSchool        6         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father         0.15          0.161616
1           IT        F   Father         0.20          0.202020
2           IT        F   Father         0.10          0.070707
3           IT        F   Father         0.30          0.252525
4           IT        F   Father         0.40          0.505051
..         …      …       …           …                …
475  Chemistry        T   Father         0.05          0.040404
476    Geology        F   Father         0.50          0.777778
477    Geology        S   Father         0.55          0.747475
478    History        F   Father         0.30          0.171717
479    History        S   Father         0.35          0.141414

     AnnouncementsView  Discussion ParentschoolSatisfaction  \
0                    2          20                     Good
1                    3          25                     Good
2                    0          30                      Bad
3                    5          35                      Bad
4                   12          50                      Bad
..                 …         …                      …
475                  5           8                      Bad
476                 14          28                      Bad
```

```
477                 25          29                      Bad
478                 14          57                      Bad
479                 23          62                      Bad

     StudentAbsenceDays Class    bins
0                     0     M   small
1                     0     M   small
2                     1     L   small
3                     1     L  medium
4                     1     M  medium
..                  …   …      …
475                   1     L   small
476                   0     M   small
477                   0     M   small
478                   1     L  medium
479                   1     L  medium

[470 rows x 17 columns]
```

[77]:
```
dataframe.to_csv('output_II.csv')
```

[78]:
```
dataframe=pd.read_csv("output_II.csv")
dataframe
```

[78]:
```
     Unnamed: 0 gender NationalITy PlaceofBirth       StageID  GradeID  \
0             0      M          KW       KuwaIT    lowerlevel        2
1             1      M          KW       KuwaIT    lowerlevel        2
2             2      M          KW       KuwaIT    lowerlevel        2
3             3      M          KW       KuwaIT    lowerlevel        2
4             4      M          KW       KuwaIT    lowerlevel        2
..          …    …          …           …            …      …
465         475      F      Jordan       Jordan  MiddleSchool        6
466         476      F      Jordan       Jordan  MiddleSchool        6
467         477      F      Jordan       Jordan  MiddleSchool        6
468         478      F      Jordan       Jordan  MiddleSchool        6
469         479      F      Jordan       Jordan  MiddleSchool        6

    SectionID      Topic Semester Relation  raisedhands  VisITedResources  \
0           A         IT        F   Father         0.15          0.161616
1           A         IT        F   Father         0.20          0.202020
2           A         IT        F   Father         0.10          0.070707
3           A         IT        F   Father         0.30          0.252525
4           A         IT        F   Father         0.40          0.505051
..        …       …       …      …          …                …
465         A  Chemistry        T   Father         0.05          0.040404
466         A    Geology        F   Father         0.50          0.777778
467         A    Geology        S   Father         0.55          0.747475
```

```
468         A    History        F   Father         0.30          0.171717
469         A    History        S   Father         0.35          0.141414

     AnnouncementsView  Discussion ParentschoolSatisfaction  \
0                    2          20                      Good
1                    3          25                      Good
2                    0          30                       Bad
3                    5          35                       Bad
4                   12          50                       Bad
..                 …          …                        …
465                  5           8                       Bad
466                 14          28                       Bad
467                 25          29                       Bad
468                 14          57                       Bad
469                 23          62                       Bad

     StudentAbsenceDays Class    bins
0                     0     M   small
1                     0     M   small
2                     1     L   small
3                     1     L  medium
4                     1     M  medium
..                  …   …      …
465                   1     L   small
466                   0     M   small
467                   0     M   small
468                   1     L  medium
469                   1     L  medium

[470 rows x 18 columns]
```

[ ]:

[ ]:

[ ]: