# 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-:** Create 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	Α	
	1	M	KW	KuwaIT	lowerlevel	G-04	Α	
	2	M	KW	KuwaIT	lowerlevel	G-04	Α	
	3	M	KW	KuwaIT	lowerlevel	G-04	Α	
	4	M	KW	KuwaIT	lowerlevel	G-04	Α	
		•••	•••	•••		•••		
	475	F	Jordan	Jordan	MiddleSchool	G-08	Α	
	476	F	Jordan	Jordan	MiddleSchool	G-08	Α	
	477	F	Jordan	Jordan	MiddleSchool	G-08	Α	
	478	F	Jordan	Jordan	MiddleSchool	G-08	Α	
	479	F	Jordan	Jordan	MiddleSchool	G-08	Α	

```
Topic Semester Relation raisedhands
                                                         VisITedResources
                              F
      0
                   IT
                                  Father
                                                   15.0
                                                                         16
                   ΙT
                              F
                                                   20.0
      1
                                  Father
                                                                         20
      2
                   ΙT
                                  Father
                                                   10.0
                                                                          7
      3
                   ΙT
                              F
                                  Father
                                                   30.0
                                                                         25
      4
                   IT
                              F
                                  Father
                                                   40.0
                                                                         50
                              Т
           Chemistry
                                                    5.0
                                                                          4
      475
                                  Father
      476
              Geology
                              F
                                  Father
                                                   50.0
                                                                        77
      477
                              S
              Geology
                                  Father
                                                   55.0
                                                                         74
                                                   30.0
      478
              History
                                  Father
                                                                         17
      479
              History
                                  Father
                                                   35.0
                                                                         14
            AnnouncementsView Discussion ParentAnsweringSurvey
      0
                                                                Yes
                                         20
                             3
                                         25
                                                                Yes
      1
      2
                             0
                                         30
                                                                 No
      3
                             5
                                         35
                                                                 No
                                         50
      4
                            12
                                                                 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
      1
                                Good
                                                  Under-7
                                                               М
      2
                                 Bad
                                                  Above-7
                                                               L
      3
                                 Bad
                                                  Above-7
                                                               L
      4
                                 Bad
                                                  Above-7
      475
                                                  Above-7
                                 Bad
                                                               L
      476
                                 Bad
                                                  Under-7
                                                               М
      477
                                 Bad
                                                  Under-7
                                                               М
      478
                                 Bad
                                                  Above-7
                                                               L
      479
                                                  Above-7
                                 Bad
                                                               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	e False	Fals	e False	False	
3	False	False	False	e False	Fals	e False	False	
4	False	False	False	e False	Fals	e False	False	
	•••	•••	•••		•••			
475	False	False	False	e False	Fals	e False	False	
476	False	False	False	e False	Fals	e False	False	
477	False	False	False	e False	Fals	e False	False	
478	False	False		e False		e False	False	
479	False	False	False	e False	Fals	e False	False	
	_							,
0			raisedhands	VisiTedRes		Announcemen		\
0	False	False	False		False		False	
1	False	False	False		False		False	
2 3	False False	False	False		False		False	
3 4		False False	False		False		False False	
			False		False		raise	
 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	
113	raise	raise	raise		Tarse		Tarse	
	Discussion	n ParentA	nsweringSurvey	y Parentso	hoolSat	isfaction \		
0	Fals		False			False		
1	Fals	е	False	Э		False		
2	Fals	е	False	Э		False		
3	Fals	е	False	е		False		
4	Fals	е	False	е		False		
	•••		•••			•••		
475	Fals	е	False	е		False		
476	Fals	е	False	е		False		
477	Fals	е	False	Э		False		
478	Fals	е	False	Э		False		
479	Fals	е	False	е		False		
•	StudentAb	•	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 Nat	tionalITy F	PlaceofBirth	${\tt StageID}$	${\tt GradeID}$	SectionID	\
О	M	KW	KuwaIT	lowerlevel	G-04	Α	
1	M	KW	KuwaIT	lowerlevel	G-04	Α	
2	M	KW	KuwaIT	lowerlevel	G-04	Α	
3	M	KW	KuwaIT	lowerlevel	G-04	Α	
4	M	KW	KuwaIT	lowerlevel	G-04	Α	
	•••	•••	•••		•••		
475	F	Jordan	Jordan	MiddleSchool	G-08	Α	
476	F	Jordan	Jordan	MiddleSchool	G-08	Α	
477	F	Jordan	Jordan	MiddleSchool	G-08	Α	
478	F	Jordan	Jordan	MiddleSchool	G-08	Α	
479	F	Jordan	Jordan	MiddleSchool	G-08	A	
	Tonio	. Semester	Relation ra	aisedhands Vi	sITedRes	ources \	
0	I.		Father	15.0		16	
1	I.		Father	20.0		20	
2	I.		Father	10.0		7	
3	I.		Father	30.0		25	
4	I.		Father	40.0		50	
		•••	•••	•••	•••		
475	Chemistry	у Т	Father	5.0		4	
476	Geology	y F	Father	50.0		77	
477	Geology	y S	Father	55.0		74	
478	History	y F	Father	30.0		17	
479	History	y S	Father	35.0		14	
	Announcer	nentsView	Discussion H	ParentAnswering	gSurvey	\	
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		

 ${\tt ParentschoolSatisfaction~StudentAbsenceDays~Class}$ 

```
0
                                           Under-7
                          Good
                                                        М
1
                          Good
                                           Under-7
                                                        М
2
                           Bad
                                           Above-7
                                                        L
3
                           Bad
                                           Above-7
                                                        L
4
                                                        М
                           Bad
                                           Above-7
. .
475
                           Bad
                                           Above-7
                                                        L
                                           Under-7
476
                           Bad
                                                        Μ
477
                           Bad
                                           Under-7
                                                        Μ
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
                              0
StageID
GradeID
                              2
SectionID
                              0
Topic
                              0
Semester
                              0
Relation
                              8
                              2
raisedhands
VisITedResources
                              0
AnnouncementsView
                              0
Discussion
                              0
ParentAnsweringSurvey
                              0
ParentschoolSatisfaction
                              0
StudentAbsenceDays
                              0
{\tt Class}
                              0
dtype: int64
```

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

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

478

Father

479 Father

Name: Relation, Length: 472, dtype: object

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

gender 0 0 NationalITy ${\tt PlaceofBirth}$ 0  ${\tt StageID}$ 0 GradeID 2  ${\tt SectionID}$ 0 0 Topic Semester 0 Relation 8 raisedhands 2 VisITedResources 0 AnnouncementsView 0 Discussion 0 ParentAnsweringSurvey 0 ParentschoolSatisfaction 0 StudentAbsenceDays 0  ${\tt Class}$ 0 dtype: int64

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

	gender	Nati	lonalITy	PlaceofB	irth	S	tageID	${\tt GradeID}$	Sectio	nID	\
0	M		KW	Ku	waIT	lowe	rlevel	G-04		Α	
1	M		KW	Ku	waIT	lowerlevel		G-04		Α	
2	M		KW	Ku	waIT	lowe	rlevel	G-04		Α	
3	M		KW	Ku	waIT	lowe	rlevel	G-04		Α	
4	M		KW	Ku	waIT	lowe	rlevel	G-04		Α	
			•••			•••	•••				
475	F		Jordan	Jo:	rdan	Middle	School	G-08		Α	
476	F		Jordan	Jo:	rdan	Middle	School	G-08		Α	
477	F		Jordan	Jo:	Jordan		School	G-08		Α	
478	F		Jordan	Jo:	Jordan		School	G-08		Α	
479	F		Jordan	Jo:	rdan	MiddleSchool		G-08		Α	
	To	pic	Semester	Relation	n ra	isedhan	ds Vi	sITedRes	ources	\	
0		IT	F	Fathe:	r	15	.0		16		
1		IT	F	Fathe:	r	20	.0		20		
2		IT	F	Fathe:	r	10	.0		7		
3		IT	F	Fathe:	r	30	.0		25		
4		IT	F	Fathe:	r	40	.0		50		
			•••	•••		•••		•••			
475	Chemis	stry	7	Fathe:	r	5	.0		4		

476	Geology	F	Father	50.0		7	77	
477	Geology	S	Father	55.0		7	74	
478	History	F	Father	30.0		1	۱7	
479	History	S	Father	35.0			L <b>4</b>	
	<b>3</b>							
	Announcements	/iew ]	Discussion	ParentAnswer	ingSurvev	\		
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			
	ParentschoolSat	isfac	tion Studer	ntAbsenceDays	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	o rows x 17 colu	ımns]						
: pri	.nt(dataframe.is	nu11()	) gim())					
, pr	.iic (davarrame, rs	IIuII ( )	7. Sulii ( ) )					
gen	der		0					
_	ionalITy		0					
	ceofBirth		0					
	geID		0					
	deID		0					
	tionID		0					
Top			0					
_	ester		0					
	ation		0					
	sedhands		2					
	ITedResources		0					
	ouncementsView		0					
	cussion		0					

[49]

```
ParentAnsweringSurvey 0
ParentschoolSatisfaction 0
StudentAbsenceDays 0
Class 0
```

dtype: int64

0

```
[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(),inplace=True)

[50]:		gender N	${f NationalITy}$	Placeofl	Birth	Sta	geID	${\tt GradeID}$	Sectio	nID	\
	0	M	KW	Κı	uwaIT	lowerl	evel	G-04		Α	
	1	M	KW	Κı	uwaIT	lowerl	evel	G-04		Α	
	2	M	KW	Κı	uwaIT	lowerl	evel	G-04		Α	
	3	M	KW	Κı	uwaIT	lowerl	evel	G-04		Α	
	4	M	KW	Kι	uwaIT	lowerl	evel	G-04		Α	
		•••	•••			•••	•••	•••			
	475	F	Jordan	Jo	ordan	MiddleSc	hool	G-08		Α	
	476	F	Jordan	Jo	ordan	MiddleSc	hool	G-08		Α	
	477	F	Jordan	Jo	ordan	MiddleSc	hool	G-08		Α	
	478	F	Jordan	Jo	ordan	MiddleSc	hool	G-08		Α	
	479	F	Jordan	Jo	ordan	MiddleSc	hool	G-08		Α	
		Top	oic Semeste	r Relatio	on ra	isedhands	Vis	sITedReso	ources	\	
	0		IT	F Fath	er	15.0			16		
	1		IT	F Fath	er	20.0			20		
	2		IT	F Fath	er	10.0			7		
	3		IT	F Fath	er	30.0			25		
	4		IT	F Fath	er	40.0			50		
				•••		•••		•••			
	475	Chemist	cry	T Fathe	er	5.0			4		
	476	Geolo	ogy	F Fath	er	50.0			77		
	477	Geolo		S Fathe	er	55.0			74		
	478	Histo		F Fathe	er	30.0			17		
	479	Histo	•	S Fathe	er	35.0			14		
			Ť								
	AnnouncementsView			Discus	sion P	arentAnsw	ering	Survey	\		
		mmouncomonob v 10w					,	, ,			

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

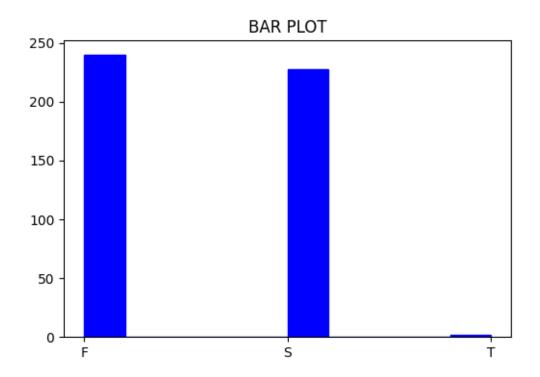
[470 rows x 17 columns]

## [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
${\tt ParentAnsweringSurvey}$	0
${\tt ParentschoolSatisfaction}$	0
${\tt StudentAbsenceDays}$	0
Class	0
dtype: int64	

```
[52]: dataframe.describe()
             raisedhands
[52]:
                          VisITedResources
                                             AnnouncementsView Discussion
              470.000000
                                 470.000000
                                                     470.000000
                                                                 470.000000
      count
      mean
               46.831197
                                  54.506383
                                                      37.659574
                                                                  43.351064
               30.833471
                                                      26.536756
                                                                  27.720855
      std
                                  33.053793
      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
              100.000000
                                  99.000000
                                                      98.000000
                                                                  99.000000
      max
```

#### 2 Outliers



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

[56]:		gender Na	ationalITy	PlaceofBir	th Stag	eID GradeID	SectionID	\
	0	M	KW	Kuwa	IT lowerle	evel G-04	Α	
	1	M	KW	Kuwa	IT lowerle	vel G-04	A	
	2	M	KW	Kuwa	IT lowerle	evel G-04	A	
	3	M	KW	Kuwa	IT lowerle	evel G-04	A	
	4	M	KW	Kuwa	IT lowerle	evel G-04	A	
		•••	•••	•••	•••			
	475	F	Jordan	Jord	an MiddleSch	ool G-08	A	
	476	F	Jordan	Jord	an MiddleSch	ool G-08	Α	
	477	F	Jordan	Jord	an MiddleSch	ool G-08	A	
	478	F	Jordan	Jord	an MiddleSch	ool G-08	A	
	479	F	Jordan	Jord	an MiddleSch	lool G-08	А	
		Topi	ic Semester	Relation	raisedhands	VisITedReso	ources \	
	0	_	IT F		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	Chemisti	rv T	Father	5.0		4	

```
50.0
      476
             Geology
                              F
                                  Father
                                                                        77
      477
                              S
                                  Father
                                                   55.0
                                                                        74
             Geology
                              F
      478
             History
                                  Father
                                                   30.0
                                                                        17
      479
                                  Father
                                                   35.0
             History
                                                                        14
           AnnouncementsView Discussion ParentAnsweringSurvey
      0
                                         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
                                                 Under-7
      0
                                Good
                                Good
                                                 Under-7
                                                              М
      1
      2
                                 Bad
                                                 Above-7
                                                              Τ.
      3
                                 Bad
                                                 Above-7
                                                              L
      4
                                                 Above-7
                                                              М
                                 Bad
      475
                                 Bad
                                                 Above-7
                                                              L
      476
                                                 Under-7
                                 Bad
      477
                                 Bad
                                                 Under-7
                                                              М
      478
                                 Bad
                                                 Above-7
                                                              L
      479
                                                 Above-7
                                                              L
                                 Bad
      [470 rows x 17 columns]
[57]: dataframe['ParentAnsweringSurvey'].value_counts()
[57]: Yes
              262
      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	NationalIT	y I	PlaceofBirth	Stage	eID	GradeID	SectionID	\
	0	M	K/	W	KuwaIT	lowerle	vel	G-04	A	
	1	M	K/	W	KuwaIT	lowerle	vel	G-04	A	
	2	М	K	W	KuwaIT	lowerle	vel	G-04	A	
	3	М	K	W	KuwaIT	lowerle	vel	G-04	A	
	4	M	K	W	KuwaIT	lowerle	vel	G-04	A	
		•••	***		•••		••	•••		
	475	F	Jordan		Jordan			G-08	A	
	476	F	Jordan		Jordan			G-08	A	
	477	F	Jordan		Jordan			G-08	A	
	478	F	Jordan		Jordan			G-08	A	
	479	F	Jorda	n	Jordan	MiddleSch	ool	G-08	A	
		То	pic Semest	er	Relation r	aisedhands	Vis	sITedResc	ources \	
	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	
	• •	<b></b> .		_				•••	_	
	475	Chemis	•	T	Father	5.0			4	
	476	Geo]		F	Father	50.0			77	
	477	Geo]		S		55.0			74	
	478	Hist	•	F S		30.0			17 14	
	479	Hist	Lory	D	Father	35.0			14	
		Annour	ncementsVie	W	Discussion	Parentschoo!	lSat	cisfactio	on \	
	0			2	20			Goo	od	
	1		;	3	25			Goo	od	
	2			0	30			Ва		
	3			5	35			Ва		
	4		1:	2	50			Ва	ıd	
	 475		<b></b> !	5	<b></b> 8			<b></b> Ва	nd	
	476		14		28			Ba		
	477		2!		29			Ba		
	478		14		57			Ba		
	479		2:		62			Ва		
		O+ 1 :	- Al D	_ ,	<b>01</b>					
	0	student	AbsenceDay							
	0		Under-		M M					
	1		Under-		M T					
	2 3		Above-		L					
			Above-' Above-'		L M					
	4		ADOVe-	ı	M					

```
475
                     Above-7
                                 L
      476
                     Under-7
                                 М
      477
                     Under-7
      478
                     Above-7
                     Above-7
      479
                                 Τ.
      [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']/

```
[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

dataframe['VisITedResources'].max()

#### 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
      Name: GradeID, dtype: int64
[63]: dataframe['GradeID']
[63]: 0
             2
      1
             2
      2
             2
      3
             2
             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
           187
      Name: StudentAbsenceDays, dtype: int64
[66]: dataframe['StudentAbsenceDays']
[66]: 0
             0
      1
             0
      2
             1
      3
             1
      4
             1
      475
             1
      476
      477
             0
      478
             1
      479
      Name: StudentAbsenceDays, Length: 470, dtype: int64
```

# 5 Binning

[67]:	data	aframe								
[67]:		gender	NationalI7	у	PlaceofBirth	Stage	eID	GradeID Secti	onID	\
	0	М	ŀ	W	KuwaIT	lowerle	vel	2	A	
	1	М	ŀ	W	KuwaIT	lowerle	vel	2	A	
	2	М	ŀ	W	KuwaIT	lowerle	vel	2	Α	
	3	М	ŀ	W	KuwaIT	lowerle	vel	2	A	
	4	М	ŀ	W	KuwaIT	lowerle	vel	2	Α	
		•••	•••		•••	•••		•••		
	475	F	Jorda	n	Jordan	MiddleScho	ool	6	Α	
	476	F	Jorda	ın	Jordan	MiddleScho	ool	6	Α	
	477	F	Jorda	n	Jordan	MiddleScho	ool	6	A	
	478	F	Jorda	n	Jordan	MiddleScho	ool	6	A	
	479	F	Jorda	n	Jordan	MiddleScho	ool	6	A	
		То	nic Semest	or	Relation ra	isedhands	Viq	ITedResources	\	
	0	10	IT	F		0.15	VIS	0.161616	`	
	1		IT	F		0.20		0.202020		
	2		IT	F		0.10		0.070707		
	3		IT	F		0.30		0.252525		
	4		IT	F		0.40		0.505051		
			11	1	rather	0.40		0.00001		
	 475	Chemis	 trv	Т	 Father	0.05		0.040404		
	476	Geol	•	F		0.50		0.777778		
	477	Geol		S		0.55		0.747475		
	478	Hist		F		0.30		0.171717		
	479	Hist	•	S		0.35		0.141414		
			-							
		Announ	cementsVie	W	Discussion P	arentschool	lSat			
	0			2	20			Good		
	1			3	25			Good		
	2			0	30			Bad		
	3			5	35			Bad		
	4			.2	50			Bad		
	 475		•••	5	<b></b> 8			 Bad		
	476		1	4	28			Bad		
	477			25	29			Bad		
	478			4	57			Bad		
	479			23	62			Bad		
		Q4 1	+ A b = P		(1)					
	0	Studen	tAbsenceDa	•						
	0			0	M					
	1			0						
	2			1	L					

```
4
                             1
                                   Μ
      475
                                   L
                             1
      476
                             0
                                   М
      477
                             0
                                   М
      478
                                   L
                             1
      479
                                   L
                             1
      [470 rows x 16 columns]
[68]: dataframe['Discussion'].value_counts()
[68]: 70
            23
      40
            22
      33
            20
      50
            18
      30
            16
      65
      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.
       Gout(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
```

3

1

L

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
                 127
      big
      Name: bins, dtype: int64
[74]:
     dataframe.head()
                                                         GradeID SectionID Topic
[74]:
        gender NationalITy PlaceofBirth
                                               StageID
      0
                                            lowerlevel
                                                                2
             Μ
                         ΚW
                                   KuwaIT
                                                                           Α
                                                                                IT
                                                                2
      1
             М
                         KW
                                   KuwaIT
                                            lowerlevel
                                                                           Α
                                                                                IT
      2
                                                                2
                         KW
                                                                                IT
             M
                                   KuwaIT
                                            lowerlevel
                                                                           Α
      3
                         KW
                                            lowerlevel
                                                                2
                                                                           Α
                                                                                IT
             Μ
                                   KuwaIT
      4
             M
                         KW
                                   KuwaIT
                                            lowerlevel
                                                                                IT
        Semester Relation
                             raisedhands
                                          VisITedResources
                                                               AnnouncementsView
      0
                F
                    Father
                                    0.15
                                                    0.161616
                                                                                2
                F
                                    0.20
                                                                                3
      1
                    Father
                                                    0.202020
      2
                F
                    Father
                                    0.10
                                                    0.070707
                                                                                0
      3
                F
                                                                                5
                    Father
                                    0.30
                                                    0.252525
      4
                F
                    Father
                                     0.40
                                                    0.505051
                                                                               12
         Discussion ParentschoolSatisfaction StudentAbsenceDays Class
                                                                                bins
      0
                  20
                                           Good
                                                                    0
                                                                          Μ
                                                                               small
                  25
                                           Good
                                                                    0
      1
                                                                           М
                                                                               small
      2
                  30
                                                                    1
                                                                           L
                                                                               small
                                            Bad
      3
                  35
                                            Bad
                                                                    1
                                                                           L
                                                                              medium
      4
                  50
                                                                    1
                                                                              medium
                                            Bad
[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
               10.000000
                                                                     98.000000
      max
                              1.000000
                                                  1.000000
```

	Coun	0 110.000		170.00000	•			
	mean	43.3510	064	0.39787	2			
	std	27.7208		0.48998				
	min	1.0000		0.00000				
	25%	20.0000		0.00000				
	50%	39.5000		0.00000				
	75%	70.000		1.00000				
	max	99.000	000	1.00000	00			
[76]:	data	frame						
[76]:		gender Nat:	ionalITy P	laceofBirth	StageID	GradeID Section	onID	\
	0	M	KW	KuwaIT	lowerlevel	2	Α	
	1	М	KW	KuwaIT	lowerlevel	2	Α	
	2	М	KW	KuwaIT	lowerlevel	2	A	
	3	M	KW	KuwaIT	lowerlevel	2	A	
	4	M	KW	KuwaIT	lowerlevel	2 	Α	
	475	F	Jordan	Jordan	MiddleSchool	6	Α	
	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	Α	
		Topic	Semester	Relation ra	isedhands Vi	sITedResources	\	
	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		Father	0.35	0.141414		
		Announceme	entsView	Discussion P	arentschoolSa	tisfaction \		
	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		

Discussion StudentAbsenceDays

count 470.000000

470.000000

```
478
                            14
                                         57
                                                                    Bad
      479
                            23
                                         62
                                                                    Bad
            StudentAbsenceDays Class
                                          bins
      0
                              0
                                     М
                                         small
                              0
      1
                                     М
                                         small
      2
                              1
                                     L
                                         small
      3
                              1
                                     L
                                        medium
      4
                              1
                                     М
                                        medium
      475
                                     L
                                         small
                              1
      476
                              0
                                     М
                                         small
      477
                              0
                                     М
                                         small
      478
                              1
                                     L
                                        medium
      479
                                        medium
                              1
      [470 rows x 17 columns]
[77]: dataframe.to_csv('output_II.csv')
[78]: dataframe=pd.read_csv("output_II.csv")
      dataframe
[78]:
            Unnamed: O gender NationalITy PlaceofBirth
                                                                  StageID
                                                                           GradeID
      0
                      0
                             М
                                         KW
                                                   KuwaIT
                                                              lowerlevel
                                                                                  2
                                                                                  2
      1
                      1
                                         KW
                                                   KuwaIT
                                                              lowerlevel
      2
                      2
                             М
                                         KW
                                                   KuwaIT
                                                              lowerlevel
                                                                                  2
      3
                      3
                                         KW
                                                   KuwaIT
                                                              lowerlevel
                                                                                  2
                             М
      4
                      4
                                                                                  2
                             М
                                         KW
                                                   KuwaIT
                                                              lowerlevel
      465
                             F
                                                            MiddleSchool
                                                                                  6
                   475
                                     Jordan
                                                   Jordan
      466
                             F
                                     Jordan
                                                   Jordan
                                                            MiddleSchool
                                                                                  6
                   476
      467
                   477
                             F
                                     Jordan
                                                   Jordan
                                                            MiddleSchool
                                                                                  6
      468
                             F
                   478
                                     Jordan
                                                   Jordan
                                                            MiddleSchool
                                                                                  6
      469
                   479
                             F
                                     Jordan
                                                   Jordan
                                                            MiddleSchool
                                                                                  6
          SectionID
                           Topic Semester Relation
                                                      raisedhands
                                                                    VisITedResources
      0
                              ΙT
                                         F
                                              Father
                                                              0.15
                                                                              0.161616
      1
                   Α
                              IT
                                         F
                                              Father
                                                              0.20
                                                                              0.202020
      2
                   Α
                              IT
                                         F
                                              Father
                                                              0.10
                                                                              0.070707
                                              Father
      3
                   Α
                              IT
                                                              0.30
                                                                              0.252525
      4
                   Α
                              IT
                                         F
                                              Father
                                                              0.40
                                                                              0.505051
                      Chemistry
                                              Father
                                                              0.05
                                                                              0.040404
      465
                   Α
                                         Τ
      466
                   Α
                         Geology
                                         F
                                              Father
                                                              0.50
                                                                              0.777778
      467
                   Α
                         Geology
                                              Father
                                                              0.55
                                                                              0.747475
```

477

25

29

Bad

```
468
                        History
                                             Father
                                                             0.30
                                                                             0.171717
                  Α
                                        F
     469
                  Α
                        History
                                        S
                                             Father
                                                             0.35
                                                                             0.141414
           AnnouncementsView Discussion ParentschoolSatisfaction \
     0
                                        20
                                                                  Good
                            3
     1
                                        25
                                                                  {\tt Good}
     2
                            0
                                        30
                                                                   Bad
     3
                            5
                                         35
                                                                   Bad
     4
                           12
                                        50
                                                                   Bad
     . .
                            5
     465
                                         8
                                                                   Bad
     466
                           14
                                        28
                                                                   Bad
     467
                           25
                                        29
                                                                   Bad
     468
                           14
                                        57
                                                                   Bad
     469
                           23
                                        62
                                                                   Bad
          StudentAbsenceDays Class
                                         bins
     0
                             0
                                        small
     1
                             0
                                    М
                                        small
     2
                             1
                                    L
                                         small
     3
                             1
                                    L
                                       medium
     4
                             1
                                       medium
     465
                                        small
                                    L
                             1
     466
                                        small
                             0
                                    Μ
     467
                                        small
                             0
     468
                                       medium
                             1
                                    L
     469
                             1
                                    L
                                       medium
     [470 rows x 18 columns]
[]:
[]:
```

[]: