```
from google.colab import drive
drive.mount ('/content/drive')
```

```
Mounted at /content/drive
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/xAPI-Edu-Data.csv')
df.head()
```

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64 
 10  VisITedResources          480 non-null    int64 
 11  AnnouncementsView         480 non-null    int64 
 12  Discussion                480 non-null    int64 
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

```
df.columns
```

```
Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID',
       'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands',
       'VisITedResources', 'AnnouncementsView', 'Discussion',
       'ParentAnsweringSurvey', 'ParentschoolSatisfaction',
       'StudentAbsenceDays', 'Class'],
      dtype='object')
```

```
missing_data_count = df.isnull().sum()
missing_data_count
```

```
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
```

```
df.fillna("no data found")
```

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

480 rows × 17 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64 
 10  VisITedResources          480 non-null    int64 
 11  AnnouncementsView         480 non-null    int64 
 12  Discussion                480 non-null    int64 
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

```
df.duplicated().sum()
```

```
2
```

```
df.describe()
```

	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	480.000000	480.000000	480.000000	480.000000
mean	46.775000	54.797917	37.918750	43.283333
std	30.779223	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.750000	20.000000	14.000000	20.000000
50%	50.000000	65.000000	33.000000	39.000000
75%	75.000000	84.000000	58.000000	70.000000
max	100.000000	99.000000	98.000000	99.000000

```
data=df.select_dtypes(include='int64')
```

```
data.head()
```

	raisedhands	VisITedResources	AnnouncementsView	Discussion
0	15	16	2	20
1	20	20	3	25
2	10	7	0	30
3	30	25	5	35
4	40	50	12	50

```
columns=data.columns


plt.figure(figsize=(18,12))
sns.heatmap(df.corr(),cbar=True,annot=True)
plt.show()
```

```
from sklearn.preprocessing import StandardScaler
Scaler = StandardScaler()
```

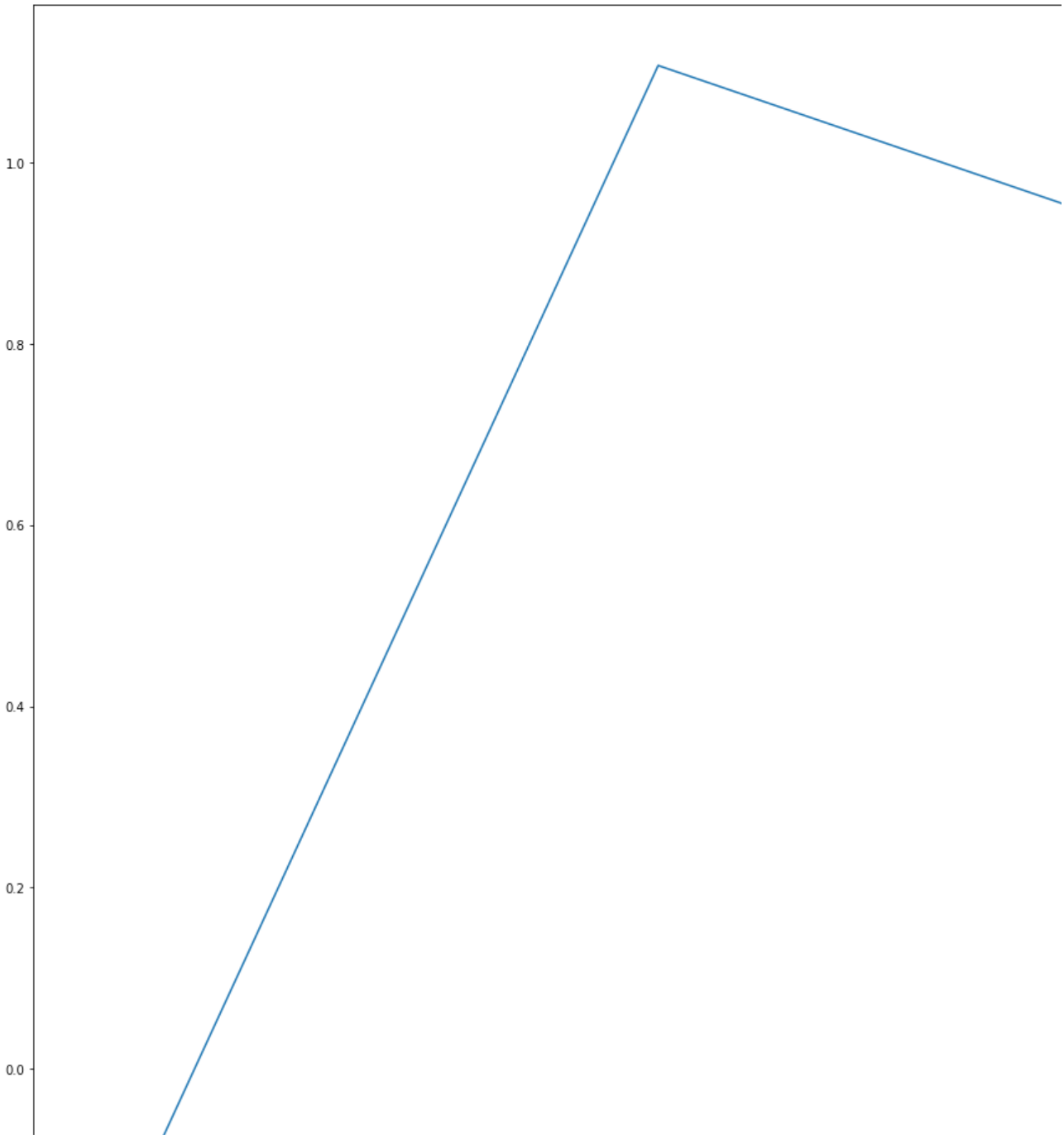

```
import pandas as pd
df = pd.DataFrame({
    'Company': ['A', 'A', 'A', 'B', 'B', 'B', 'B'],
    'Model': ['A1', 'A2', 'A3', 'B1', 'B2', 'B3', 'B4'],
    'Year': [2019, 2020, 2021, 2018, 2019, 2020, 2021],
    'Transmission': ['Manual', 'Automatic', 'Automatic', 'Manual', 'Automatic', 'Automatic
    'EngineSize': [1.4, 2.0, 1.4, 1.5, 2.0, 1.5, 1.5],
    'MPG': [55.4, 67.3, 58.9, 52.3, 64.2, 68.9, 83.1]
})
```

```
df.describe()
```

	Year	EngineSize	MPG
count	7.000000	7.000000	7.000000
mean	2019.714286	1.614286	64.300000
std	1.112697	0.267261	10.295468
min	2018.000000	1.400000	52.300000
25%	2019.000000	1.450000	57.150000
50%	2020.000000	1.500000	64.200000
75%	2020.500000	1.750000	68.100000
max	2021.000000	2.000000	83.100000

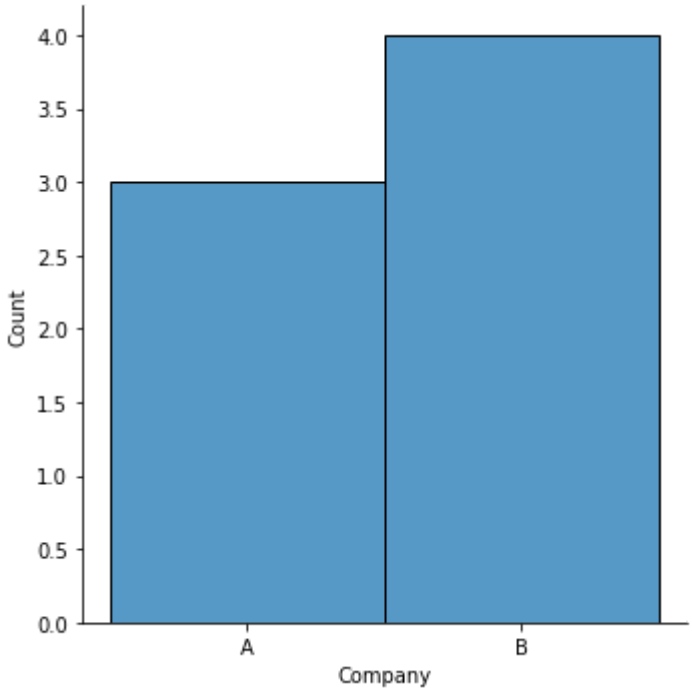
```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(18,20))
plt.plot(df.skew())
plt.show()
```

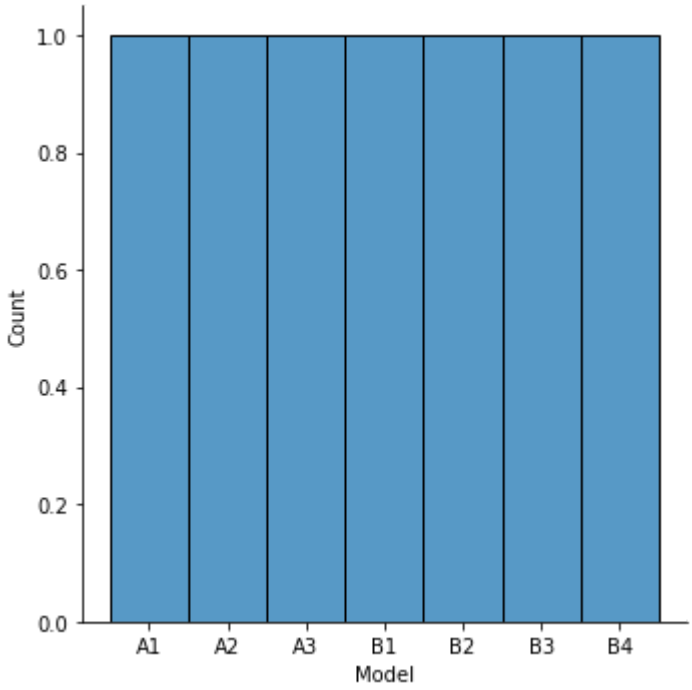
```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: FutureWarning: Droppi
  after removing the cwd from sys.path.
```

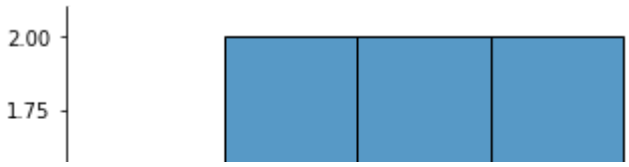


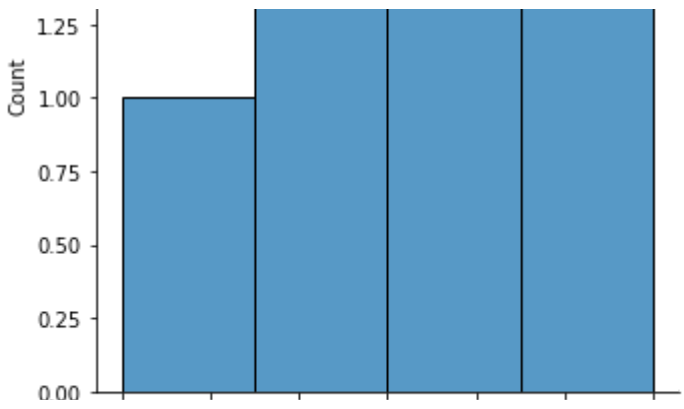

```
features_=df.columns.values[:]
fig=plt.figure(figsize=(20,10))
for columns, feature in enumerate(features_):
    sns.displot(df[feature])
plt.show()
```

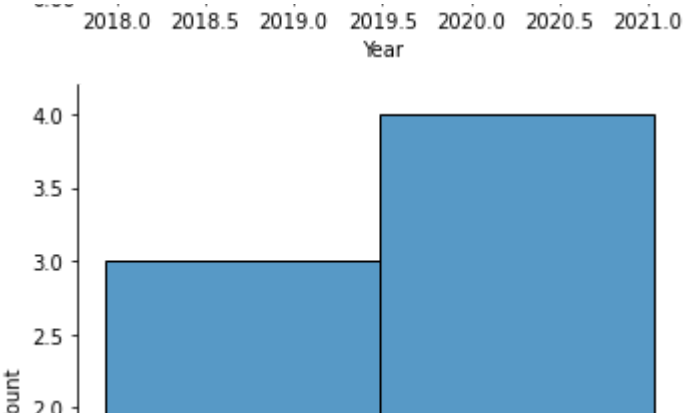
```
<Figure size 1440x720 with 0 Axes>
```










2018.0
2018.5
2019.0
2019.5
2020.0
2020.5
2021.0
Year
4.0
3.5
3.0
2.5
ount