Correlation Analysis

February 9, 2023

[11]:
```
#Readthedata
import pandas as pd
df=pd.read_csv(r"movies.csv")
df
```

[11]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0
…	…	…	…	…	…	…	…	…	…	…	…	…
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks				
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz				
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai				
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall				
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia				

```
…                      …               …              …          …
7663           Joseph Ebanks      Shannon Bond   United States     7000.0
7664             Lisa Huston  Michael Saquella   United States        NaN
7665            Lynno Lovert      Onyama Laura   United States    58750.0
7666           James Randall     Christina Roz   United States    15000.0
7667            Pereko Mosia  Siyabonga Mabaso    South Africa        NaN

           gross                    company  runtime
0      46998772.0              Warner Bros.    146.0
1      58853106.0          Columbia Pictures   104.0
2     538375067.0                  Lucasfilm   124.0
3      83453539.0         Paramount Pictures    88.0
4      39846344.0             Orion Pictures    98.0
…             …                          …      …
7663          NaN                        NaN    90.0
7664          NaN  Cactus Blue Entertainment    90.0
7665          NaN           Embi Productions     NaN
7666          NaN                        NaN   120.0
7667          NaN               PK 65 Films   102.0

[7668 rows x 15 columns]
```

[10]:
```
#Lets check missing values
import pandas as pd
df=pd.read_csv(r"movies.csv")
df.isna().sum()
```

[10]:
```
name           0
rating        77
genre          0
year           0
released       2
score          3
votes          3
director       0
writer         3
star           1
country        3
budget      2171
gross        189
company       17
runtime        4
dtype: int64
```

[9]:
```
#Lets remove missing values
import pandas as pd
df=pd.read_csv(r"movies.csv")
```

```
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df
```

[9]:

```
                                                 name     rating      genre  \
0                                     The Shining          R      Drama
1                                   The Blue Lagoon          R  Adventure
2     Star Wars: Episode V - The Empire Strikes Back         PG     Action
3                                       Airplane!          PG     Comedy
4                                     Caddyshack           R     Comedy
…                                                 …        …         …
7648                          Bad Boys for Life          R     Action
7649                        Sonic the Hedgehog         PG     Action
7650                                  Dolittle         PG  Adventure
7651                       The Call of the Wild        PG  Adventure
7652                          The Eight Hundred  Not Rated     Action

      year                         released  score      votes  \
0     1980      June 13, 1980 (United States)    8.4   927000.0
1     1980       July 2, 1980 (United States)    5.8    65000.0
2     1980      June 20, 1980 (United States)    8.7  1200000.0
3     1980       July 2, 1980 (United States)    7.7   221000.0
4     1980      July 25, 1980 (United States)    7.3   108000.0
…     …                                 …    …        …
7648  2020   January 17, 2020 (United States)    6.6   140000.0
7649  2020  February 14, 2020 (United States)    6.5   102000.0
7650  2020   January 17, 2020 (United States)    5.6    53000.0
7651  2020  February 21, 2020 (United States)    6.8    42000.0
7652  2020    August 28, 2020 (United States)    6.8     3700.0

             director                  writer               star  \
0     Stanley Kubrick             Stephen King     Jack Nicholson
1      Randal Kleiser  Henry De Vere Stacpoole     Brooke Shields
2      Irvin Kershner           Leigh Brackett        Mark Hamill
3        Jim Abrahams             Jim Abrahams        Robert Hays
4        Harold Ramis       Brian Doyle-Murray        Chevy Chase
…                 …                        …                  …
7648     Adil El Arbi             Peter Craig         Will Smith
7649      Jeff Fowler               Pat Casey       Ben Schwartz
7650   Stephen Gaghan           Stephen Gaghan  Robert Downey Jr.
7651    Chris Sanders            Michael Green      Harrison Ford
7652          Hu Guan                  Hu Guan    Zhi-zhong Huang

             country       budget        gross  \
0     United Kingdom   19000000.0   46998772.0
1      United States    4500000.0   58853106.0
```

```
2      United States   18000000.0  538375067.0
3      United States    3500000.0   83453539.0
4      United States    6000000.0   39846344.0
…                …          …           …
7648   United States   90000000.0  426505244.0
7649   United States   85000000.0  319715683.0
7650   United States  175000000.0  245487753.0
7651          Canada  135000000.0  111105497.0
7652           China   80000000.0  461421559.0

                                  company  runtime
0                             Warner Bros.    146.0
1                          Columbia Pictures    104.0
2                                  Lucasfilm    124.0
3                         Paramount Pictures     88.0
4                             Orion Pictures     98.0
…                                        …      …
7648                       Columbia Pictures    124.0
7649                      Paramount Pictures     99.0
7650                      Universal Pictures    101.0
7651                    20th Century Studios    100.0
7652  Beijing Diqi Yinxiang Entertainment    149.0

[5424 rows x 15 columns]
```

[ ]:
```
#Lets check for duplicates
import pandas as pd
df=pd.read_csv(r"movies.csv")
df1=df.duplicated(['name'])
df1
```

[8]:
```
#Lets change the datatype
import pandas as pd
df=pd.read_csv(r"movies.csv")
df1=df.convert_dtypes()
df1.dtypes
```

[8]:
```
name         string
rating       string
genre        string
year          Int64
released     string
score       Float64
votes         Int64
director     string
writer       string
star         string
```

```
country      string
budget        Int64
gross         Int64
company      string
runtime       Int64
dtype: object
```

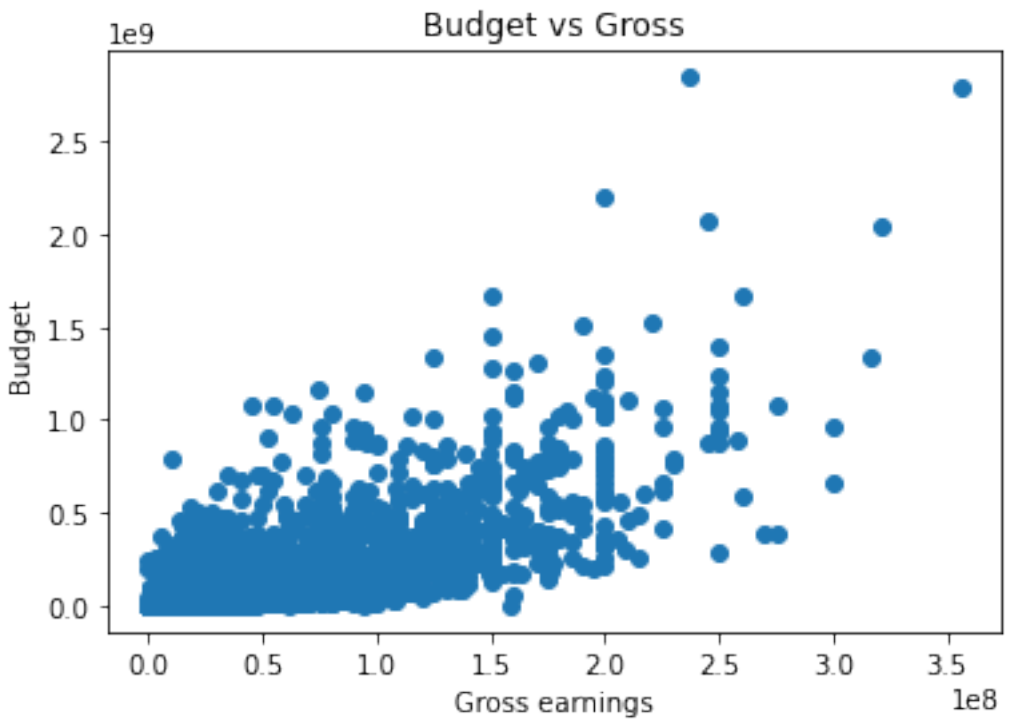
[7]:
```
#Lets sort the data
import pandas as pd
df=pd.read_csv(r"movies.csv")
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
df2
```

[7]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000	James Cameron	James Cameron	Sam Worthington	United States	237000000
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000	Anthony Russo	Christopher Markus	Robert Downey Jr.	United States	356000000
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000
6663	Star Wars: Episode VII - The Force Awakens	PG-13	Action	2015	December 18, 2015 (United States)	7.8	876000	J.J. Abrams	J.J. Abrams	Daisy Ridley	United States	245000000
7244	Avengers: Infinity War	PG-13	Action	2018	April 27, 2018 (United States)	8.4	897000	Anthony Russo	Christopher Markus	Robert Downey Jr.	United States	321000000
…	…	…	…	…	…	…	…	…	…	…	…	…
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500	Francesca Gregorini	Tatiana Von Furstenberg	Rooney Mara	United States	1000000
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900	Stephen Cornwell	Stephen Cornwell	Kevin Gage	United States	10000000
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000	John Fawcett	Karen Walton	Katharine Isabelle	Canada	2000000
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300	Charles Band	Dennis Paoli	Robin Wright	United States	1100000
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800	George Huang	George Huang	Peter Scolari	United States	3000000

```
6663          Lawrence Kasdan       Daisy Ridley  United States  245000000
7244       Christopher Markus  Robert Downey Jr.  United States  321000000
…                         …                  …              …          …
5640  Tatiana von Fürstenberg        Rooney Mara  United States    3000000
2434       Wallace C. Bennett       Brad Johnson  United States    5000000
3681             Karen Walton      Emily Perkins         Canada    5000000
272             Alan J. Adler    Robert Glaudini  United States     800000
3203                Andy Burg       Will Friedle  United States   15000000

          gross                    company  runtime
5445  2847246203     Twentieth Century Fox      162
7445  2797501328            Marvel Studios      181
3045  2201647264     Twentieth Century Fox      194
6663  2069521700                 Lucasfilm      138
7244  2048359754            Marvel Studios      149
…          …                        …      …
5640        5073          Two Prong Lesson       96
2434        2970          Trimark Pictures       97
3681        2554  Copperheart Entertainment      108
272         2270          Embassy Pictures       85
3203         309                  Daybreak       85

[5424 rows x 15 columns]
```

[6]:
```
#Lets build a scatter plot for finding correlation between 2 variables
#BudgetvsGross
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
df=pd.read_csv(r"movies.csv")
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
plt.scatter(x=df2['budget'],y=df2['gross'])
plt.title("Budget vs Gross")
plt.xlabel('Gross earnings')
plt.ylabel("Budget")
plt.show()
```



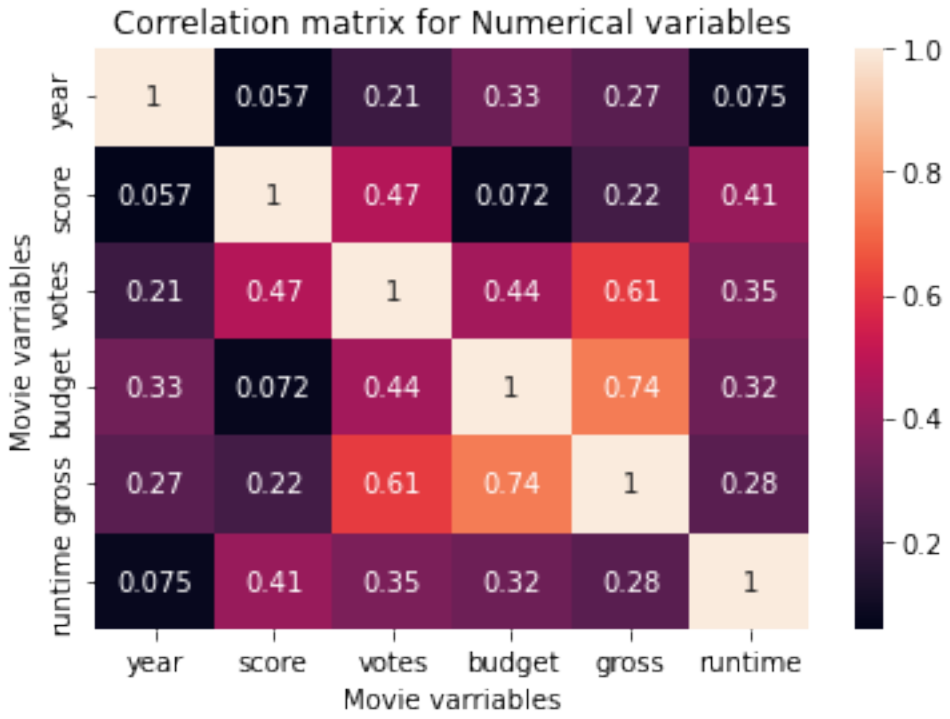
[5]:
```
#Lets see correlation
import pandas as pd
df=pd.read_csv(r"movies.csv")
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
df2.corr()
```

[5]:

	year	score	votes	budget	gross	runtime
year	1.000000	0.056506	0.206161	0.327961	0.274395	0.075173
score	0.056506	1.000000	0.474349	0.072155	0.222709	0.414145
votes	0.206161	0.474349	1.000000	0.439757	0.614808	0.352331
budget	0.327961	0.072155	0.439757	1.000000	0.740263	0.318718
gross	0.274395	0.222709	0.614808	0.740263	1.000000	0.275830
runtime	0.075173	0.414145	0.352331	0.318718	0.275830	1.000000

[3]:
```
#Heatmap using Seaborn

import seaborn as sns
import pandas as pd
import matplotlib
```

```
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv(r"movies.csv")
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
correlation_matrix=df2.corr()
sns.heatmap(correlation_matrix,annot=True)
plt.title("Correlation matrix for Numerical variables")
plt.xlabel('Movie varriables')
plt.ylabel('Movie varriables')
plt.show()
```



[2]:
```
#Lets give random nos for categorical variables
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv(r"movies.csv")
```

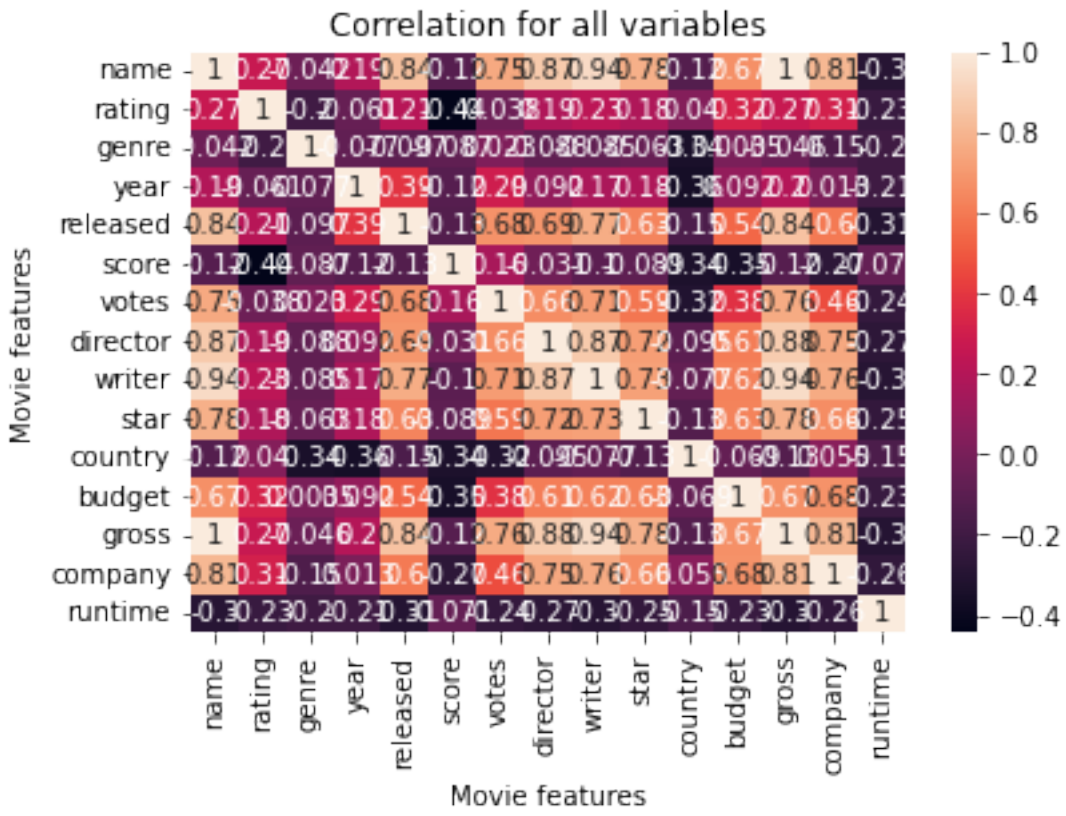
```python
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
df3=df2.apply(lambda x : x.factorize()[0]).corr()
df3.corr()
```

[2]:

```
              name    rating     genre      year  released     score  \
name      1.000000  0.266689 -0.041663  0.191235  0.838888 -0.118011
rating    0.266689  1.000000 -0.200016 -0.060886  0.206638 -0.443450
genre    -0.041663 -0.200016  1.000000 -0.077350 -0.096746 -0.087029
year      0.191235 -0.060886 -0.077350  1.000000  0.389136 -0.118942
released  0.838888  0.206638 -0.096746  0.389136  1.000000 -0.131005
score    -0.118011 -0.443450 -0.087029 -0.118942 -0.131005  1.000000
votes     0.753220 -0.038114  0.023205  0.290286  0.675363  0.159483
director  0.872280  0.186998 -0.087741  0.091649  0.688479 -0.031097
writer    0.936214  0.225412 -0.084628  0.166163  0.767280 -0.103326
star      0.776690  0.180688 -0.062643  0.178117  0.631075 -0.088823
country  -0.123338  0.040325 -0.342084 -0.359606 -0.154935 -0.335882
budget    0.669379  0.321675  0.003461  0.091511  0.538285 -0.350662
gross     0.999703  0.266789 -0.046036  0.198613  0.842300 -0.116594
company   0.805526  0.311023 -0.147318  0.012534  0.597389 -0.270704
runtime  -0.296209 -0.228822 -0.202450 -0.205859 -0.305633 -0.070643

             votes  director    writer      star   country    budget  \
name      0.753220  0.872280  0.936214  0.776690 -0.123338  0.669379
rating   -0.038114  0.186998  0.225412  0.180688  0.040325  0.321675
genre     0.023205 -0.087741 -0.084628 -0.062643 -0.342084  0.003461
year      0.290286  0.091649  0.166163  0.178117 -0.359606  0.091511
released  0.675363  0.688479  0.767280  0.631075 -0.154935  0.538285
score     0.159483 -0.031097 -0.103326 -0.088823 -0.335882 -0.350662
votes     1.000000  0.656699  0.710722  0.587888 -0.318496  0.379410
director  0.656699  1.000000  0.870429  0.716884 -0.094737  0.607285
writer    0.710722  0.870429  1.000000  0.727466 -0.076826  0.620617
star      0.587888  0.716884  0.727466  1.000000 -0.132049  0.625534
country  -0.318496 -0.094737 -0.076826 -0.132049  1.000000 -0.068780
budget    0.379410  0.607285  0.620617  0.625534 -0.068780  1.000000
gross     0.758017  0.876263  0.939156  0.781667 -0.127075  0.670479
company   0.461515  0.749002  0.759049  0.656142  0.055429  0.678584
runtime  -0.241187 -0.270285 -0.295992 -0.248435 -0.152094 -0.227779

             gross   company   runtime
name      0.999703  0.805526 -0.296209
rating    0.266789  0.311023 -0.228822
genre    -0.046036 -0.147318 -0.202450
year      0.198613  0.012534 -0.205859
```

```
released  0.842300  0.597389 -0.305633
score    -0.116594 -0.270704 -0.070643
votes     0.758017  0.461515 -0.241187
director  0.876263  0.749002 -0.270285
writer    0.939156  0.759049 -0.295992
star      0.781667  0.656142 -0.248435
country  -0.127075  0.055429 -0.152094
budget    0.670479  0.678584 -0.227779
gross     1.000000  0.805456 -0.299973
company   0.805456  1.000000 -0.257365
runtime  -0.299973 -0.257365  1.000000
```

[15]:
```
#Lets plot the correlation

import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv(r"movies.csv")
df.dropna(subset=["budget"],inplace=True)
df.dropna(subset=["gross"],inplace=True)
df.dropna(subset=["rating"],inplace=True)
df1=df.convert_dtypes()
df2=df1.sort_values(by=['gross'],ascending=False)
df3=df2.apply(lambda x : x.factorize()[0]).corr()
corr_matrix=df3.corr()
sns.heatmap(corr_matrix,annot=True)
plt.title('Correlation for all variables')
plt.xlabel("Movie features")
plt.ylabel('Movie features')
plt.figure(figsize=(1000,500))
plt.show()
```




```
<Figure size 72000x36000 with 0 Axes>
```

[ ]: