dup

November 25, 2019

[4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import LabelEncoder
```

[5]:
```
df_census = pd.read_csv('CensusData.csv')
```

[27]:
```
bool_series = df_census.duplicated(subset=None, keep=False)
bool_series
```

[27]:
```
0        False
1        False
2        False
3        False
4        False
5        False
6        False
7        False
8        False
9        False
10       False
11       False
12       False
13       False
14       False
15       False
16       False
17       False
18       False
19       False
20       False
21       False
22       False
23       False
24       False
```

```
25       False
26       False
27       False
28       False
29       False
         ...
32531    False
32532    False
32533    False
32534    False
32535    False
32536    False
32537    False
32538    False
32539    False
32540    False
32541    False
32542    False
32543    False
32544    False
32545    False
32546    False
32547    False
32548    False
32549    False
32550    False
32551    False
32552    False
32553    False
32554    False
32555    False
32556    False
32557    False
32558    False
32559    False
32560    False
Length: 32561, dtype: bool
```

[28]:
```
df_census_unique = df_census[~bool_series]
```

[30]:
```
df_census_unique
```

[30]:
```
       Age          workclass  fnlwgt      education  education-num  \
0       39          State-gov   77516      Bachelors             13
1       50   Self-emp-not-inc   83311      Bachelors             13
2       38            Private  215646        HS-grad              9
3       53            Private  234721           11th              7
4       28            Private  338409      Bachelors             13
```

5	37	Private	284582	Masters	14
6	49	Private	160187	9th	5
7	52	Self-emp-not-inc	209642	HS-grad	9
8	31	Private	45781	Masters	14
9	42	Private	159449	Bachelors	13
10	37	Private	280464	Some-college	10
11	30	State-gov	141297	Bachelors	13
12	23	Private	122272	Bachelors	13
13	32	Private	205019	Assoc-acdm	12
14	40	Private	121772	Assoc-voc	11
15	34	Private	245487	7th-8th	4
16	25	Self-emp-not-inc	176756	HS-grad	9
17	32	Private	186824	HS-grad	9
18	38	Private	28887	11th	7
19	43	Self-emp-not-inc	292175	Masters	14
20	40	Private	193524	Doctorate	16
21	54	Private	302146	HS-grad	9
22	35	Federal-gov	76845	9th	5
23	43	Private	117037	11th	7
24	59	Private	109015	HS-grad	9
25	56	Local-gov	216851	Bachelors	13
26	19	Private	168294	HS-grad	9
27	54	?	180211	Some-college	10
28	39	Private	367260	HS-grad	9
29	49	Private	193366	HS-grad	9
...	...	...	...	...	...
32531	30	?	33811	Bachelors	13
32532	34	Private	204461	Doctorate	16
32533	54	Private	337992	Bachelors	13
32534	37	Private	179137	Some-college	10
32535	22	Private	325033	12th	8
32536	34	Private	160216	Bachelors	13
32537	30	Private	345898	HS-grad	9
32538	38	Private	139180	Bachelors	13
32539	71	?	287372	Doctorate	16
32540	45	State-gov	252208	HS-grad	9
32541	41	?	202822	HS-grad	9
32542	72	?	129912	HS-grad	9
32543	45	Local-gov	119199	Assoc-acdm	12
32544	31	Private	199655	Masters	14
32545	39	Local-gov	111499	Assoc-acdm	12
32546	37	Private	198216	Assoc-acdm	12
32547	43	Private	260761	HS-grad	9
32548	65	Self-emp-not-inc	99359	Prof-school	15
32549	43	State-gov	255835	Some-college	10
32550	43	Self-emp-not-inc	27242	Some-college	10
32551	32	Private	34066	10th	6

```
32552   43            Private   84661      Assoc-voc             11
32553   32            Private  116138        Masters             14
32554   53            Private  321865        Masters             14
32555   22            Private  310152   Some-college             10
32556   27            Private  257302     Assoc-acdm             12
32557   40            Private  154374        HS-grad              9
32558   58            Private  151910        HS-grad              9
32559   22            Private  201490        HS-grad              9
32560   52       Self-emp-inc  287927        HS-grad              9

                Marital-status          Occupation     Relationship  \
0                Never-married        Adm-clerical    Not-in-family
1           Married-civ-spouse     Exec-managerial          Husband
2                     Divorced   Handlers-cleaners    Not-in-family
3           Married-civ-spouse   Handlers-cleaners          Husband
4           Married-civ-spouse      Prof-specialty             Wife
5           Married-civ-spouse     Exec-managerial             Wife
6        Married-spouse-absent       Other-service    Not-in-family
7           Married-civ-spouse     Exec-managerial          Husband
8                Never-married      Prof-specialty    Not-in-family
9           Married-civ-spouse     Exec-managerial          Husband
10          Married-civ-spouse     Exec-managerial          Husband
11          Married-civ-spouse      Prof-specialty          Husband
12               Never-married        Adm-clerical        Own-child
13               Never-married               Sales    Not-in-family
14          Married-civ-spouse        Craft-repair          Husband
15          Married-civ-spouse    Transport-moving          Husband
16               Never-married     Farming-fishing        Own-child
17               Never-married   Machine-op-inspct        Unmarried
18          Married-civ-spouse               Sales          Husband
19                    Divorced     Exec-managerial        Unmarried
20          Married-civ-spouse      Prof-specialty          Husband
21                   Separated       Other-service        Unmarried
22          Married-civ-spouse     Farming-fishing          Husband
23          Married-civ-spouse    Transport-moving          Husband
24                    Divorced        Tech-support        Unmarried
25          Married-civ-spouse        Tech-support          Husband
26               Never-married        Craft-repair        Own-child
27          Married-civ-spouse                   ?          Husband
28                    Divorced     Exec-managerial    Not-in-family
29          Married-civ-spouse        Craft-repair          Husband
...                        ...                 ...             ...
32531            Never-married                   ?    Not-in-family
32532       Married-civ-spouse      Prof-specialty          Husband
32533       Married-civ-spouse     Exec-managerial          Husband
32534                 Divorced        Adm-clerical        Unmarried
32535            Never-married     Protective-serv        Own-child
```

32536	Never-married	Exec-managerial	Not-in-family	
32537	Never-married	Craft-repair	Not-in-family	
32538	Divorced	Prof-specialty	Unmarried	
32539	Married-civ-spouse	?	Husband	
32540	Separated	Adm-clerical	Own-child	
32541	Separated	?	Not-in-family	
32542	Married-civ-spouse	?	Husband	
32543	Divorced	Prof-specialty	Unmarried	
32544	Divorced	Other-service	Not-in-family	
32545	Married-civ-spouse	Adm-clerical	Wife	
32546	Divorced	Tech-support	Not-in-family	
32547	Married-civ-spouse	Machine-op-inspct	Husband	
32548	Never-married	Prof-specialty	Not-in-family	
32549	Divorced	Adm-clerical	Other-relative	
32550	Married-civ-spouse	Craft-repair	Husband	
32551	Married-civ-spouse	Handlers-cleaners	Husband	
32552	Married-civ-spouse	Sales	Husband	
32553	Never-married	Tech-support	Not-in-family	
32554	Married-civ-spouse	Exec-managerial	Husband	
32555	Never-married	Protective-serv	Not-in-family	
32556	Married-civ-spouse	Tech-support	Wife	
32557	Married-civ-spouse	Machine-op-inspct	Husband	
32558	Widowed	Adm-clerical	Unmarried	
32559	Never-married	Adm-clerical	Own-child	
32560	Married-civ-spouse	Exec-managerial	Wife	

	Race	Sex	capital-gain	capital-loss	\
0	White	Male	2174	0	
1	White	Male	0	0	
2	White	Male	0	0	
3	Black	Male	0	0	
4	Black	Female	0	0	
5	White	Female	0	0	
6	Black	Female	0	0	
7	White	Male	0	0	
8	White	Female	14084	0	
9	White	Male	5178	0	
10	Black	Male	0	0	
11	Asian-Pac-Islander	Male	0	0	
12	White	Female	0	0	
13	Black	Male	0	0	
14	Asian-Pac-Islander	Male	0	0	
15	Amer-Indian-Eskimo	Male	0	0	
16	White	Male	0	0	
17	White	Male	0	0	
18	White	Male	0	0	
19	White	Female	0	0	

20		White	Male	0	0
21		Black	Female	0	0
22		Black	Male	0	0
23		White	Male	0	2042
24		White	Female	0	0
25		White	Male	0	0
26		White	Male	0	0
27	Asian-Pac-Islander		Male	0	0
28		White	Male	0	0
29		White	Male	0	0
...		...	...	...	...
32531	Asian-Pac-Islander		Female	0	0
32532		White	Male	0	0
32533	Asian-Pac-Islander		Male	0	0
32534		White	Female	0	0
32535		Black	Male	0	0
32536		White	Female	0	0
32537		Black	Male	0	0
32538		Black	Female	15020	0
32539		White	Male	0	0
32540		White	Female	0	0
32541		Black	Female	0	0
32542		White	Male	0	0
32543		White	Female	0	0
32544		Other	Female	0	0
32545		White	Female	0	0
32546		White	Female	0	0
32547		White	Male	0	0
32548		White	Male	1086	0
32549		White	Female	0	0
32550		White	Male	0	0
32551	Amer-Indian-Eskimo		Male	0	0
32552		White	Male	0	0
32553	Asian-Pac-Islander		Male	0	0
32554		White	Male	0	0
32555		White	Male	0	0
32556		White	Female	0	0
32557		White	Male	0	0
32558		White	Female	0	0
32559		White	Male	0	0
32560		White	Female	15024	0

	hours-per-week	native-country	Income
0	40	United-States	<=50K
1	13	United-States	<=50K
2	40	United-States	<=50K
3	40	United-States	<=50K

4	40	Cuba	<=50K
5	40	United-States	<=50K
6	16	Jamaica	<=50K
7	45	United-States	>50K
8	50	United-States	>50K
9	40	United-States	>50K
10	80	United-States	>50K
11	40	India	>50K
12	30	United-States	<=50K
13	50	United-States	<=50K
14	40	?	>50K
15	45	Mexico	<=50K
16	35	United-States	<=50K
17	40	United-States	<=50K
18	50	United-States	<=50K
19	45	United-States	>50K
20	60	United-States	>50K
21	20	United-States	<=50K
22	40	United-States	<=50K
23	40	United-States	<=50K
24	40	United-States	<=50K
25	40	United-States	>50K
26	40	United-States	<=50K
27	60	South	>50K
28	80	United-States	<=50K
29	40	United-States	<=50K
...	...	...	...
32531	99	United-States	<=50K
32532	60	United-States	>50K
32533	50	Japan	>50K
32534	39	United-States	<=50K
32535	35	United-States	<=50K
32536	55	United-States	>50K
32537	46	United-States	<=50K
32538	45	United-States	>50K
32539	10	United-States	>50K
32540	40	United-States	<=50K
32541	32	United-States	<=50K
32542	25	United-States	<=50K
32543	48	United-States	<=50K
32544	30	United-States	<=50K
32545	20	United-States	>50K
32546	40	United-States	<=50K
32547	40	Mexico	<=50K
32548	60	United-States	<=50K
32549	40	United-States	<=50K
32550	50	United-States	<=50K

```
32551              40   United-States   <=50K
32552              45   United-States   <=50K
32553              11          Taiwan   <=50K
32554              40   United-States    >50K
32555              40   United-States   <=50K
32556              38   United-States   <=50K
32557              40   United-States    >50K
32558              40   United-States   <=50K
32559              20   United-States   <=50K
32560              40   United-States    >50K

[32514 rows x 15 columns]
```

[14]:
```
bool_series
```

[14]:
```
False    32514
True        47
dtype: int64
```

[43]:
```
#before removing duplication
print ("before removing duplication ")
df_census.shape
```

```
before removing duplication
```

[43]:
```
(32561, 15)
```

[45]:
```
#after removing duplication
print ("after removing duplication ")
df_census_unique.shape
```

```
after removing duplication
```

[45]:
```
(32514, 15)
```

[48]:
```
#data integration
dataset1 = pd.read_csv('CensusData.csv')
```

[49]:
```
dataset2 = pd.read_csv('CensusData.csv')
```

[54]:
```
combined_dataset = pd.concat([dataset1,dataset2])
```

[55]:
```
combined_dataset.shape
```

[55]:
```
(65122, 15)
```

[5]:
```
#histogram
data = pd.read_csv('diamonds.csv')
x = data.cut
```

[ ]:
```
plt.hist(x,orientation='vertical')
plt.title('cut histogram')
```

```
plt.xlabel('cut')
plt.ylabel('frequency')
cut = ('ideal','premium','good','very good','fair')
index = np.arange(len(cut))
plt.xticks(index,cut,rotation=90)
plt.show()
```

[ ]:
```
plt.hist(x,color='blue',orientation='vertical')
plt.title('cut histogram')
plt.xlabel('cut')
plt.ylabel('frequency')
cut = ('ideal','premium','good','very good','fair')
# index = np.arange(len(cut))
# plt.xticks(index,cut,rotation=90)
plt.show()
```

[ ]:
```
plt.hist(x,bins=[0,10,20,30,40,50,60,70,80,90,99])
```

[7]:
```
data
```

[7]:

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
5	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
10	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
11	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
12	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
13	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
14	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
15	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
16	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68
17	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29	2.70
18	0.30	Good	J	SI1	63.8	56.0	351	4.23	4.26	2.71
19	0.30	Very Good	J	SI1	62.7	59.0	351	4.21	4.27	2.66
20	0.30	Good	I	SI2	63.3	56.0	351	4.26	4.30	2.71
21	0.23	Very Good	E	VS2	63.8	55.0	352	3.85	3.92	2.48
22	0.23	Very Good	H	VS1	61.0	57.0	353	3.94	3.96	2.41
23	0.31	Very Good	J	SI1	59.4	62.0	353	4.39	4.43	2.62
24	0.31	Very Good	J	SI1	58.1	62.0	353	4.44	4.47	2.59
25	0.23	Very Good	G	VVS2	60.4	58.0	354	3.97	4.01	2.41
26	0.24	Premium	I	VS1	62.5	57.0	355	3.97	3.94	2.47
27	0.30	Very Good	J	VS2	62.2	57.0	357	4.28	4.30	2.67

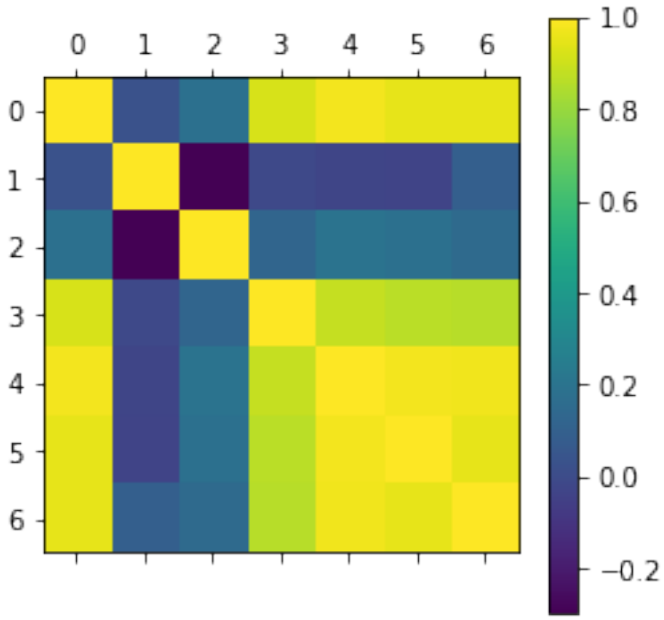
```
28      0.23  Very Good     D     VS2   60.5   61.0    357  3.96  3.97  2.40
29      0.23  Very Good     F     VS1   60.9   57.0    357  3.96  3.99  2.42
...      ...        ...   ...     ...    ...    ...    ...   ...   ...   ...
53910   0.70    Premium     E     SI1   60.5   58.0   2753  5.74  5.77  3.48
53911   0.57    Premium     E      IF   59.8   60.0   2753  5.43  5.38  3.23
53912   0.61    Premium     F    VVS1   61.8   59.0   2753  5.48  5.40  3.36
53913   0.80       Good     G     VS2   64.2   58.0   2753  5.84  5.81  3.74
53914   0.84       Good     I     VS1   63.7   59.0   2753  5.94  5.90  3.77
53915   0.77      Ideal     E     SI2   62.1   56.0   2753  5.84  5.86  3.63
53916   0.74       Good     D     SI1   63.1   59.0   2753  5.71  5.74  3.61
53917   0.90  Very Good     J     SI1   63.2   60.0   2753  6.12  6.09  3.86
53918   0.76    Premium     I     VS1   59.3   62.0   2753  5.93  5.85  3.49
53919   0.76      Ideal     I    VVS1   62.2   55.0   2753  5.89  5.87  3.66
53920   0.70  Very Good     E     VS2   62.4   60.0   2755  5.57  5.61  3.49
53921   0.70  Very Good     E     VS2   62.8   60.0   2755  5.59  5.65  3.53
53922   0.70  Very Good     D     VS1   63.1   59.0   2755  5.67  5.58  3.55
53923   0.73      Ideal     I     VS2   61.3   56.0   2756  5.80  5.84  3.57
53924   0.73      Ideal     I     VS2   61.6   55.0   2756  5.82  5.84  3.59
53925   0.79      Ideal     I     SI1   61.6   56.0   2756  5.95  5.97  3.67
53926   0.71      Ideal     E     SI1   61.9   56.0   2756  5.71  5.73  3.54
53927   0.79       Good     F     SI1   58.1   59.0   2756  6.06  6.13  3.54
53928   0.79    Premium     E     SI2   61.4   58.0   2756  6.03  5.96  3.68
53929   0.71      Ideal     G     VS1   61.4   56.0   2756  5.76  5.73  3.53
53930   0.71    Premium     E     SI1   60.5   55.0   2756  5.79  5.74  3.49
53931   0.71    Premium     F     SI1   59.8   62.0   2756  5.74  5.73  3.43
53932   0.70  Very Good     E     VS2   60.5   59.0   2757  5.71  5.76  3.47
53933   0.70  Very Good     E     VS2   61.2   59.0   2757  5.69  5.72  3.49
53934   0.72    Premium     D     SI1   62.7   59.0   2757  5.69  5.73  3.58
53935   0.72      Ideal     D     SI1   60.8   57.0   2757  5.75  5.76  3.50
53936   0.72       Good     D     SI1   63.1   55.0   2757  5.69  5.75  3.61
53937   0.70  Very Good     D     SI1   62.8   60.0   2757  5.66  5.68  3.56
53938   0.86    Premium     H     SI2   61.0   58.0   2757  6.15  6.12  3.74
53939   0.75      Ideal     D     SI2   62.2   55.0   2757  5.83  5.87  3.64

[53940 rows x 10 columns]
```

[9]:
```
data.corr()
```

[9]:
```
          carat     depth     table     price         x         y         z
carat  1.000000  0.028224  0.181618  0.921591  0.975094  0.951722  0.953387
depth  0.028224  1.000000 -0.295779 -0.010647 -0.025289 -0.029341  0.094924
table  0.181618 -0.295779  1.000000  0.127134  0.195344  0.183760  0.150929
price  0.921591 -0.010647  0.127134  1.000000  0.884435  0.865421  0.861249
x      0.975094 -0.025289  0.195344  0.884435  1.000000  0.974701  0.970772
y      0.951722 -0.029341  0.183760  0.865421  0.974701  1.000000  0.952006
z      0.953387  0.094924  0.150929  0.861249  0.970772  0.952006  1.000000
```

[14]:
```
plt.matshow(data.corr())
# plt.xticks(range(len(data.columns)),data.columns)
# plt.yticks(range(len(data.columns)),data.columns)
plt.colorbar()
```

[14]:
```
<matplotlib.colorbar.Colorbar at 0xdf1f8b4630>
```



[6]:
```
correlation_matrix = data.corr().round(2)
plt.show()
plt.savefig("heatmap.png")
plt.clf
plt.close()
```

[ ]:
```

```