4_Life_Expectancy_Exploratory_Data_Analysis_PART2

January 29, 2023

Table of Contents

1 Life_Expectancy_WHO_UN_Analysis_Modeling

1.1 Exploratory_Data_Analysis_PART_2

To: Magnimind

From: Matt Curcio, matt.curcio.ri@gmail.com

Date: 2023-01-29

Re: NOTEBOOK #4

---

[1]:
```
# Common Python Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

import warnings
warnings.simplefilter(action='ignore', category=UserWarning)

# Statistics related library
from scipy import stats
```

[2]:
```
!ls *.csv
```

```
Clean_LE_Data_FEng_4.csv      Life_Expectancy_Data.csv  y_test.csv
Clean_LE_Data_Post_EDA_3.csv  x_test.csv                y_train.csv
Clean_LE_Data_w_Means_2.csv   x_train.csv
```

[3]:
```
filename = 'Clean_LE_Data_FEng_4.csv'

df = pd.read_csv(filename, header=0)

df.head(3)
```

[3]:

	Country	Year	Status	LifeExpectancy	AdultMort	EtOH	PercExpen	\
0	Afghanistan	2015	0	65.0	263.0	0.01	71.279624	
1	Afghanistan	2014	0	59.9	271.0	0.01	73.523582	
2	Afghanistan	2013	0	59.9	268.0	0.01	73.219243	

	Measles	BMI	lt5yD	Polio	TotalExpen	DTP	HIV	Thin1_19y	Income	\
0	1154	19.1	83	6.0	8.16	65.0	0.1	17.2	0.479	
1	492	18.6	86	58.0	8.18	62.0	0.1	17.5	0.476	
2	430	18.1	89	62.0	8.13	64.0	0.1	17.7	0.470	

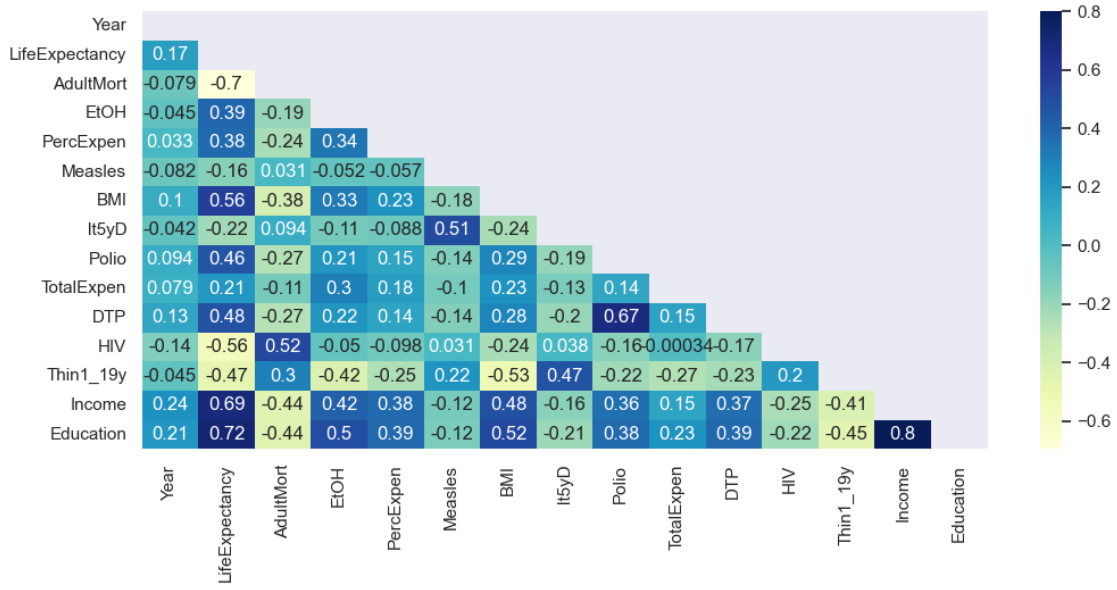
	Education	Region
0	10.1	2
1	10.0	2
2	9.9	2

[4]:
```
# Convert to categorical
df['Status'] = pd.Categorical(df['Status'])
df['Region'] = pd.Categorical(df['Region'])

print(f'\nFile "{filename}" has ', df.shape[0], 'observations &', df.shape[1],␣
 ↪'features.\n')
```

```
File "Clean_LE_Data_FEng_4.csv" has  2928 observations & 18 features.
```

1.1.1 Re-check correlations of numeric vars

[5]:
```
# creating triangular mask
mask = np.triu(np.ones_like(df.corr()))

# plotting a triangle correlation heatmap
sns.set(rc={"figure.figsize":(12, 5)}) #width=3, #height=4
dataplot = sns.heatmap(df.corr(), cmap="YlGnBu", annot=True, mask=mask)
plt.show()
```



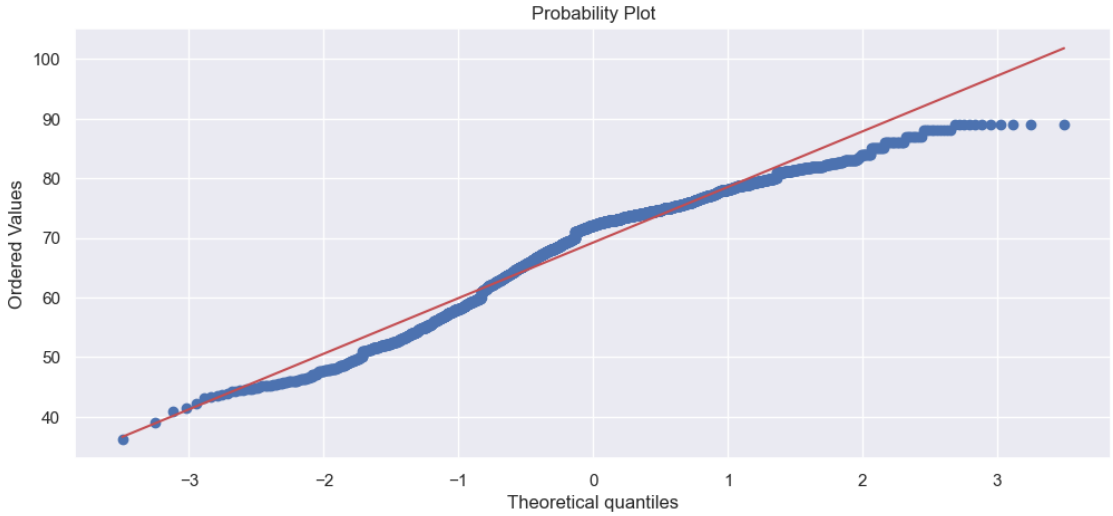
1.1.2 NOTE #3:

- The absolute values of correlations between Life Expectancy and {Income:0.69} and {Education:0.72} are greatest.

Correlation with respect to Life Expectancy Listed By Absolute Magnitude

Feature	Value
'Education'	0.72
'Income'	0.69
'AdultMort'	0.7
'HIV'	0.56
'BMI'	0.56
'DTP'	0.48
'Thin1_19y'	0.47
'Polio'	0.46
'EtOH'	0.39
'PercExpen	0.38
'lt5yD'	0.22
'TotalExpen'	0.21
'Measles'	0.16

[6]:
```
#Get QQ-plot of LifeExpectancy
fig = plt.figure()
res = stats.probplot(df['LifeExpectancy'], plot=plt)
plt.show()
```


Probability Plot
Ordered Values
Theoretical quantiles
100
90
80
70
60
50
40
−3
−2
−1
0
1
2
3

[ ]: