Descriptive_Statistics_pandas_seaborn

May 6, 2020

1 Descriptive statistical analysis using Pandas and Seaborn

[4]:
```
# I will be using a data set consisting of countries and their 2016 median␣
 ↪income.

import pandas as pd

df = pd.read_csv("C:/Users/Pedro Santos/Documents/learn_code/Python for Data␣
 ↪Science/dataAnalyticsPortFolio/datasets/europe-datasets/median_income_2016.
 ↪csv")
df.head()
```

[4]:
```
     country  median_income
0    Belgium          21335
1   Bulgaria           6742
2    Czechia          12478
3    Denmark          21355
4    Germany          21152
```

[5]:
```
df.info()
'''
We can see that all objects exist, or rather, are not null.
'''
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 2 columns):
country          32 non-null object
median_income    32 non-null int64
dtypes: int64(1), object(1)
memory usage: 592.0+ bytes
```

[57]:
```
# describe() gives us a nice summary of the data which includes the mean.
df.describe()
```

[57]:
```
       median_income
count      32.000000
```

```
mean     15972.343750
std       6640.636617
min       4724.000000
25%      10190.500000
50%      16205.000000
75%      21161.250000
max      28663.000000
```

2 Measurements of center and data visualization

[10]:
```python
# Mean is not a very good indicator due to outliers
print(df['median_income'].mean())
```

```
15972.34375
```

[52]:
```python
# Checking for outliers
import seaborn as sb

sb.set(style="whitegrid")
bs = sb.boxplot(x=df['median_income'])

'''
The boxplot does not show any outliers and the data seems to be normally␣
 ↪distributed.
'''
```

[52]:
```
'\nThe boxplot does not show any outliers and the data seems to be normally
distributed.\n'
```

[56]:

```
'''
We can verify normal distribution with, for example, a q-q plot, which
compares our data with a Gaussian distribution (or normal distribution)
'''
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot

# q-q plot
qqplot(df['median_income'], line='s')
pyplot.show()

'''
The data can be considered normally distributed
'''
```

[45]:

```
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('seaborn-whitegrid')

plt.scatter(df['median_income'], df['country'], marker='o')

'''
The scatter plot shows us that the data points are dispersed with
some trend lines, for example around 10000 euros and 20000 euros.
With the lowest median income country being Romania and the highest Luxembourg.
'''
```

[45]:

```
'\nThe scatter plot shows us that the data points are dispersed with\nsome trend
lines, for example around 10000 euros and 20000 euros. \n'
```

[63]:

```
# One useful metric is the variation of the data - standard deviation
# (Spread of the data from their mean)
std_dev = df['median_income'].std()
std_dev
```

[63]:

```
6640.636617409193
```

[67]:

```
# Also the famous histogram...
sb.distplot(df['median_income'])
```

[67]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2817c7d2b00>
```

[ ]:

```
```