Statistics for Data Science.

- URL: https://www.learndatasci.com/tutorials/data-science-statistics-using-python/ (https://www.learndatasci.com/tutorials/data-science-statistics-using-python/)

By: Sagun Shakya

- GITAM Institute of Science

Description of the problem:

- A public school administrator makes statistical research regarding the poor performance of the students in the school.

Data Set:

- https://raw.githubusercontent.com/LearnDataSci/article-resources/master/Essential%20Statistics/middle_tn_schools.csv (https://raw.githubusercontent.com/LearnDataSci/article-resources/master/Essential%20Statistics/middle_tn_schools.csv)

Parameters used with their description:

- **Variable: Definition**
- name : The name of the school
- school_rating : The school's rating
- size : The school's student count
- reduced_lunch : The percentage of students that got enrolled in reduced lunch
- state_percentile_16: The school's percentile in 2016
- state_percentile_15: The school's percentile in 2015
- stu_teach_ratio : The school's student to teacher ratio
- school_type : The type of school (public, private, magnet, alternative, etc)
- avg_score_15 : The school's average test score for 2015
- avg_score_16 : The school's average test score for 2016
- full_time_teachers : The school's total full time teachers
- percent_black : Percentage of black students at the school
- percent_white : Percentage of white students at the school
- percent_asian : Percentage of asian students at the school
- percent_hispanic : Percentage of hispanic students at the school

- NOTE:
  - **reduced_lunch** is a variable measuring the average percentage of students per school enrolled in a federal program that provides lunches for students from lower-income households.

Types of statistics used:

- 1. **Descriptive statistics**:
  - identify patterns in the data.
  - two measures used to describe the data: central tendency and deviation.
- 2. **Inferential Statistics**:
  - allow us to make hypotheses (or inferences) about a sample that can be applied to the population.

In [1]:
```
#imports for the problem.

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import copy
import os
```

In [2]:
```
#Changing the working directory to the folder containing the .csv file.

os.chdir('D:\Sagun Shakya\Python\Basic-Statistics-and-Visualization-master')
```

In [3]:
```
df = pd.read_csv('middle_tn_schools.csv', index_col = 'name')
school = copy.deepcopy(df)                #creating a virtual copy of the original data
```

Checking for null values.

In [4]:
```
school.isnull().sum()
```

Out[4]:
```
school_rating          0
size                   0
reduced_lunch          0
state_percentile_16    0
state_percentile_15    6
stu_teach_ratio        0
school_type            0
avg_score_15           6
avg_score_16           0
full_time_teachers     0
percent_black          0
percent_white          0
percent_asian          0
percent_hispanic       0
dtype: int64
```

In [5]:
```
#dropping null values.
school.dropna(inplace = True)
school.isnull().sum()
```

Out[5]:
```
school_rating          0
size                   0
reduced_lunch          0
state_percentile_16    0
state_percentile_15    0
stu_teach_ratio        0
school_type            0
avg_score_15           0
avg_score_16           0
full_time_teachers     0
percent_black          0
percent_white          0
percent_asian          0
percent_hispanic       0
dtype: int64
```

Dimension of the dataframe.

In [6]:
```
school.shape
```

Out[6]:
```
(341, 14)
```

In [7]:
```
sub = school[['reduced_lunch', 'school_rating']]
```

Statistical Summary for the numerical data.

In [8]:
```
school.describe()
```

Out[8]:

	school_rating	size	reduced_lunch	state_percentile_16	state_percentile_15	stu_teach_ratio
count	341.000000	341.000000	341.000000	341.000000	341.000000	341.000000
mean	2.982405	704.923754	49.920821	59.063930	58.249267	15.459824
std	1.685487	400.970851	25.379175	32.445386	32.702630	5.728366
min	0.000000	53.000000	2.000000	0.500000	0.600000	7.300000
25%	2.000000	424.000000	30.000000	31.700000	27.100000	13.700000
50%	3.000000	606.000000	50.000000	67.600000	65.800000	15.000000
75%	4.000000	851.000000	71.000000	88.200000	88.600000	16.600000
max	5.000000	2314.000000	98.000000	99.800000	99.800000	111.000000

- **reduced_lunch** is a good proxy for household income which, in turn, must be correlated with the schools' performance.

- Using **groupby** and **describe** method.

In [12]:
```
sub = school[['reduced_lunch', 'school_rating']]
sub.groupby(['school_rating']).describe()
```

Out[12]:

	reduced_lunch							
	count	mean	std	min	25%	50%	75%	max
school_rating								
0.0	42.0	83.500000	8.903959	53.0	79.25	86.0	90.00	98.0
1.0	38.0	74.894737	11.943085	53.0	65.00	74.5	84.75	98.0
2.0	43.0	63.976744	11.933323	37.0	54.50	62.0	73.50	88.0
3.0	56.0	50.285714	13.550866	24.0	41.00	48.5	63.00	78.0
4.0	85.0	40.458824	16.002643	4.0	30.00	41.0	50.00	86.0
5.0	77.0	21.610390	17.766879	2.0	8.00	19.0	30.00	87.0

In [13]:
```
sub.corr()
```

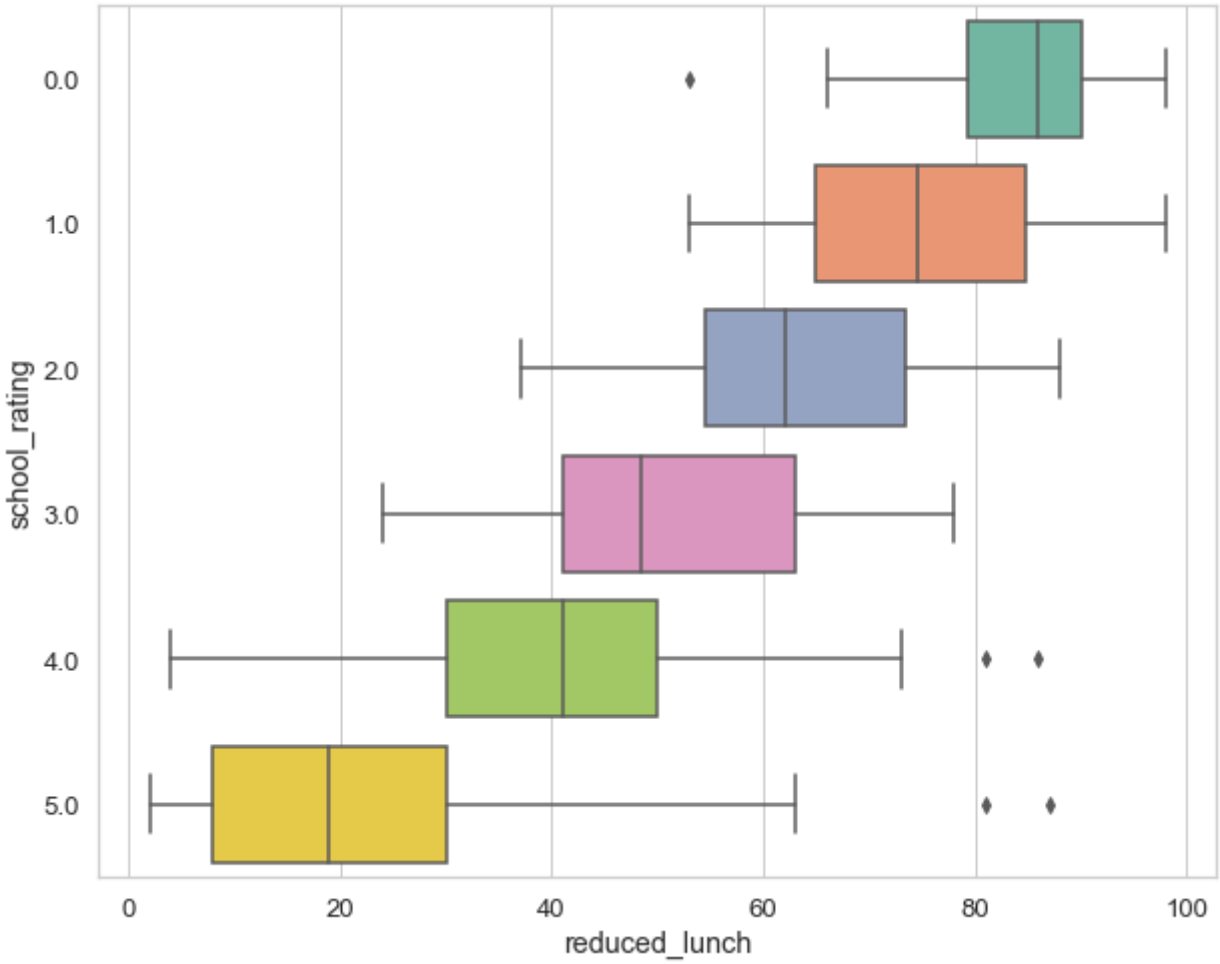
Out[13]:

	reduced_lunch	school_rating
reduced_lunch	1.000000	-0.818037
school_rating	-0.818037	1.000000

Creating a Box - and - Whisker Plot.

In [14]:
```
plt.figure(figsize = (10,8))
sns.set(style = 'whitegrid', font_scale = 1.2)

sns.boxplot(x = 'reduced_lunch', y = 'school_rating', data = school, palette = 'Set2',
plt.show()
```



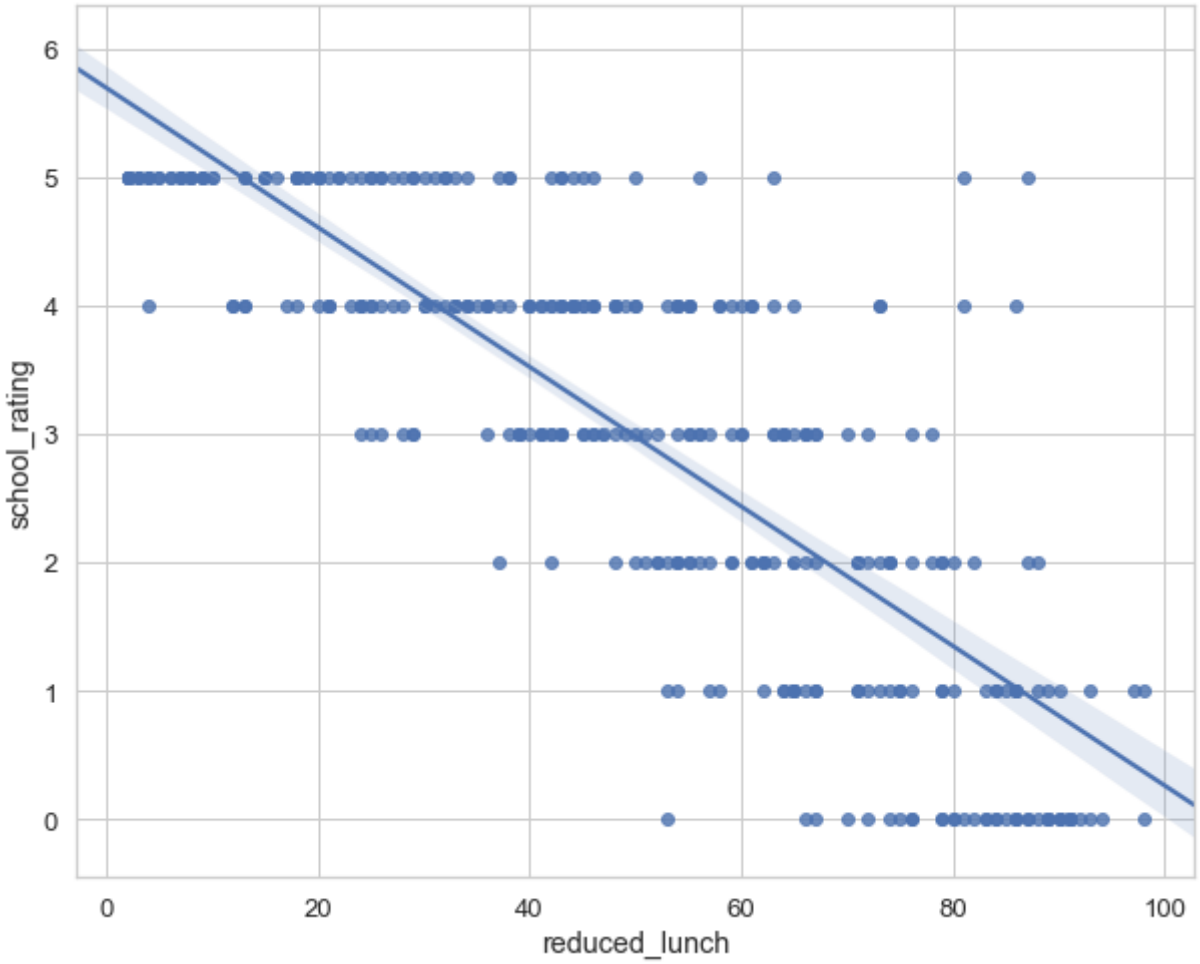
Conclusion:

- The higher rated schools have lower proportions of students having the 'reduced_lunch' scheme.

Linear Correlation between 'reduced_lunch' and 'school_rating'.

In [16]:
```
plt.figure(figsize = (10,8))
sns.set(style = 'whitegrid', font_scale = 1.2)

sns.regplot(y = school['school_rating'], x= school['reduced_lunch'], scatter= True, mar
plt.show()
```



Conclusion:

- There exists a negative correlation between these two variables meaning that as the number of students having the 'reduced_lunch' scheme increases in a school, the rating can decrease (can be set up for hypothesis testing in the later stages of the same research).

Creating a heatmap (correlation matrix) for checking which variables further affect the school_rating.

In [17]:
```
school.corr()
```

Out[17]:

	school_rating	size	reduced_lunch	state_percentile_16	state_percentile_15	stu_tea
school_rating	1.000000	0.174402	-0.818037	0.985476	0.937817	
size	0.174402	1.000000	-0.268493	0.165095	0.162887	
reduced_lunch	-0.818037	-0.268493	1.000000	-0.819148	-0.825085	-
state_percentile_16	0.985476	0.165095	-0.819148	1.000000	0.949694	
state_percentile_15	0.937817	0.162887	-0.825085	0.949694	1.000000	
stu_teach_ratio	0.199151	0.140678	-0.201650	0.181639	0.141066	
avg_score_15	0.941336	0.161788	-0.839536	0.949197	0.991847	
avg_score_16	0.982491	0.136475	-0.820004	0.994116	0.946101	
full_time_teachers	0.114626	0.966971	-0.201282	0.110210	0.109569	
percent_black	-0.606631	-0.136791	0.561948	-0.587216	-0.564929	-
percent_white	0.656134	0.092990	-0.674078	0.643632	0.612183	
percent_asian	0.156934	0.189146	-0.220604	0.145926	0.181822	
percent_hispanic	-0.387348	-0.017239	0.499779	-0.395358	-0.371708	-

In [18]:
```
cor = school.corr().columns.values
print(cor)
print(type(cor))
```

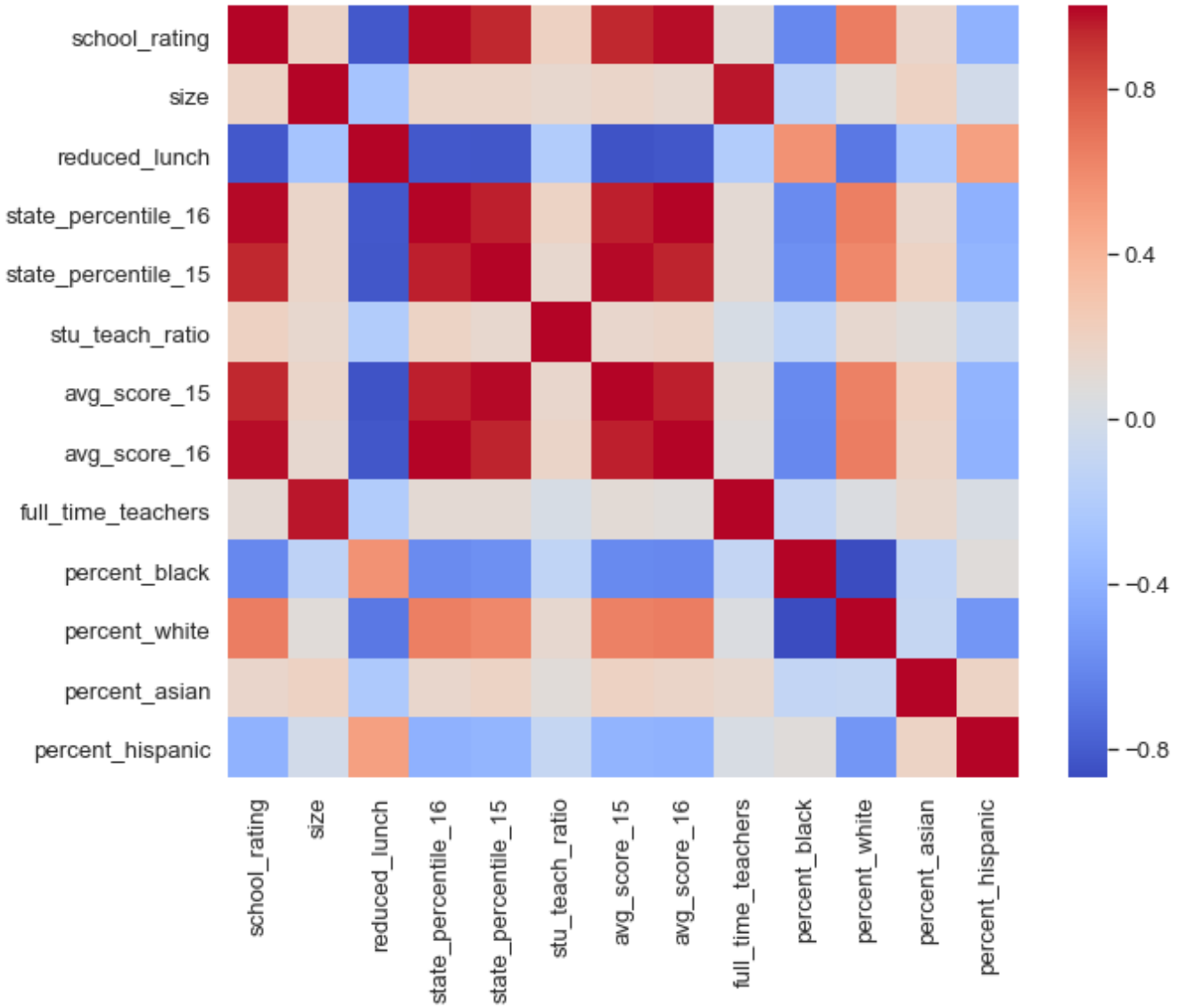
```
['school_rating' 'size' 'reduced_lunch' 'state_percentile_16'
 'state_percentile_15' 'stu_teach_ratio' 'avg_score_15' 'avg_score_16'
 'full_time_teachers' 'percent_black' 'percent_white' 'percent_asian'
 'percent_hispanic']
<class 'numpy.ndarray'>
```

In [19]:

```
plt.figure(figsize = (10,8))
sns.set(style = 'whitegrid', font_scale = 1.2)

sns.heatmap(data=school.corr(), xticklabels=cor, yticklabels=cor, cmap = 'coolwarm')
plt.show()

'''
Description of the figure:
- Red cells indicate positive correlation.
- Blue cells indicate negative correlation.
- White cells indicate no correlation.
- The darker the colors, the stronger the correlation (positive or negative) between th
  cmap = 'coolwarm'.
'''
```



Out[19]: "\nDescription of the figure:\n- Red cells indicate positive correlation.\n- Blue cell
s indicate negative correlation.\n- White cells indicate no correlation. \n- The darke
r the colors, the stronger the correlation (positive or negative) between those two va
riables as indicated by\n cmap = 'coolwarm'.\n"

The End.

Prepared by: Sagun Shakya

- GITAM Institute of Science.