PYTHON PROGRAMLAMAYA GİRİŞ

Hafta 10

PYTHON PROGRAMLAMA

- pandas, Python programlama dilinde kullanılan açık kaynak kodlu bir veri analizi kütüphanesidir.

PYTHON PROGRAMLAMA

- pandas temel veri yapıları:
- Series
- DataFrame
- Panel

PYTHON PROGRAMLAMA

- Üç veri yapısından bir tanesi Series'dir
- Series, aynı veri türünü saklayabilen, tek boyutlu bir vektör veya bir dizidir.

PYTHON PROGRAMLAMA

- pandas'daki ikinci veri yapısı olan dataframe satır, sütun ve indeks'ten meydana gelir.
- Dataframe'in her bir sütunu sadece bir tek veri türünü saklamaktadır.

PYTHON PROGRAMLAMA

- Dataset için link:

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

PYTHON PROGRAMLAMA

- name,school_rating,size,reduced_lunch,state_percentile_16,state_percentile_15,stu_teach_ratio,school_type,avg_score_15,avg_score_16,full_time_teachers,percent_black,percent_white,percent_asian,percent_hispanic
- Allendale Elementary School,5.0,851.0,10.0,90.2,95.8,15.7,Public,89.4,85.2,54.0,2.9,85.5,1.6,5.6
- Anderson Elementary,2.0,412.0,71.0,32.8,37.3,12.8,Public,43.0,38.3,32.0,3.9,86.7,1.0,4.9
- Avoca Elementary,4.0,482.0,43.0,78.4,83.6,16.6,Public,75.7,73.0,29.0,1.0,91.5,1.2,4.4

PYTHON PROGRAMLAMA

- # kütüphanelerin deklarasyonu
- import pandas as pd
- %matplotlib inline
- import matplotlib.pyplot as plt
- import seaborn as sns
- sns.set_style('darkgrid')
- sns.set(font_scale=1.5)

PYTHON PROGRAMLAMA

- # Datasetin GitHub'tan alınması
- df = pd.read_csv('https://raw.githubusercontent.com/LearnDataSci/article-resources/master/Essential%20Statistics/middle_tn_schools.csv')

PYTHON PROGRAMLAMA

- # dataframe'in ekranda gösterilmesi

- print(df)

PYTHON PROGRAMLAMA

- name school_rating size reduced_lunch \
- 0 Allendale Elementary School 5.0 851.0 10.0
- 1 Anderson Elementary 2.0 412.0 71.0
- 2 Avoca Elementary 4.0 482.0 43.0
- 3 Bailey Middle 0.0 394.0 91.0
- 4 Barfield Elementary 4.0 948.0 26.0
- .. ..
- 342 Winfree Bryant Middle School 3.0 611.0 57.0
- 343 Winstead Elementary School 5.0 515.0 8.0
- 344 Woodland Elementary 4.0 424.0 55.0
- 345 Woodland Middle School 5.0 866.0 2.0
- 346 Wright Middle 0.0 829.0 89.0

PYTHON PROGRAMLAMA

- # dataframe'in ekranda gösterilmesi

- print(df.tail(-1))

PYTHON PROGRAMLAMA

```
   name  school_rating   size  reduced_lunch  \
1              Anderson Elementary          2.0  412.0          71.0
2                Avoca Elementary          4.0  482.0          43.0
3                 Bailey Middle            0.0  394.0          91.0
4              Barfield Elementary          4.0  948.0          26.0
5    Barkers Mill Elementary School         4.0  893.0          48.0
..                    ...          ...    ...          ...
342  Winfree Bryant Middle School           3.0  611.0          57.0
343   Winstead Elementary School            5.0  515.0           8.0
344            Woodland Elementary          4.0  424.0          55.0
345     Woodland Middle School              5.0  866.0           2.0
346               Wright Middle            0.0  829.0          89.0
```

PYTHON PROGRAMLAMA

- # dataframe'in ekranda gösterilmesi

- print(df.head(-1))

PYTHON PROGRAMLAMA

- name school_rating size reduced_lunch \
- 0 Allendale Elementary School 5.0 851.0 10.0
- 1 Anderson Elementary 2.0 412.0 71.0
- 2 Avoca Elementary 4.0 482.0 43.0
- 3 Bailey Middle 0.0 394.0 91.0
- 4 Barfield Elementary 4.0 948.0 26.0
- .. ..
- 341 Wilson Elementary School 4.0 800.0 25.0
- 342 Winfree Bryant Middle School 3.0 611.0 57.0
- 343 Winstead Elementary School 5.0 515.0 8.0
- 344 Woodland Elementary 4.0 424.0 55.0
- 345 Woodland Middle School 5.0 866.0 2.0

PYTHON PROGRAMLAMA

- # Sütun isimlerinin ekranda gösterilmesi

- print(df.head(o))

PYTHON PROGRAMLAMA

- Empty DataFrame
- Columns: [name, school_rating, size, reduced_lunch, state_percentile_16, state_percentile_15, stu_teach_ratio, school_type, avg_score_15, avg_score_16, full_time_teachers, percent_black, percent_white, percent_asian, percent_hispanic]
- Index: []

PYTHON PROGRAMLAMA

- # # ilk verinin ekranda gösterilmesi

- print(df.head(1))

PYTHON PROGRAMLAMA

- name school_rating size reduced_lunch \
- 0 Allendale Elementary School 5.0 851.0 10.0

- state_percentile_16 state_percentile_15 stu_teach_ratio school_type \
- 0 90.2 95.8 15.7 Public

- avg_score_15 avg_score_16 full_time_teachers percent_black \
- 0 89.4 85.2 54.0 2.9

- percent_white percent_asian percent_hispanic
- 0 85.5 1.6 5.6

PYTHON PROGRAMLAMA

- # sütun indeksleri

- print (df.columns)

PYTHON PROGRAMLAMA

- Index(['name', 'school_rating', 'size', 'reduced_lunch', 'state_percentile_16',
- 'state_percentile_15', 'stu_teach_ratio', 'school_type', 'avg_score_15',
- 'avg_score_16', 'full_time_teachers', 'percent_black', 'percent_white',
- 'percent_asian', 'percent_hispanic'],
- dtype='object')

PYTHON PROGRAMLAMA

- # Satır Range İndeks

- print (df.index)

PYTHON PROGRAMLAMA

- Ekran Çıktısı:

- RangeIndex(start=0, stop=347, step=1)

PYTHON PROGRAMLAMA

- # toplam veri sayısı

- print ("Toplam Veri Sayısı: ", df.size)

PYTHON PROGRAMLAMA

- Toplam Veri Sayısı: 5205

PYTHON PROGRAMLAMA

- # Satır ve sütun sayısı

- print ("Satır ve sütun sayısı: ", df.shape)

PYTHON PROGRAMLAMA

- Satır ve sütun sayısı: (347, 15)

PYTHON PROGRAMLAMA

- # school_rating sütunu incelenmesi

- print()
- print (df['school_rating'])

PYTHON PROGRAMLAMA

- 0 5.0
- 1 2.0
- 2 4.0
- 3 0.0
- 4 4.0
- ...
- 342 3.0
- 343 5.0
- 344 4.0
- 345 5.0
- 346 0.0
- Name: school_rating, Length: 347, dtype: float64

PYTHON PROGRAMLAMA

- # indeks numarası ile veri listesi

- print (df.iloc[342]) #indeks numarası

PYTHON PROGRAMLAMA

- name Winfree Bryant Middle School
- school_rating 3
- size 611
- reduced_lunch 57
- state_percentile_16 59.1
- state_percentile_15 65.2
- stu_teach_ratio 16.9
- school_type Public
- avg_score_15 61.4
- avg_score_16 57.7
- full_time_teachers 36
- percent_black 15.2
- percent_white 66.3
- percent_asian 1.5
- percent_hispanic 15.7
- Name: 342, dtype: object

PYTHON PROGRAMLAMA

- # Filtreleme işlemi(Filtering):

- print (df[df.school_rating > 3][['name', 'school_type']])

PYTHON PROGRAMLAMA

- name school_type
- 0 Allendale Elementary School Public
- 2 Avoca Elementary Public
- 4 Barfield Elementary Public
- 5 Barkers Mill Elementary School Public
- 6 Barksdale Elementary Public
- .. ..
- 336 White House Middle School Public
- 341 Wilson Elementary School Public
- 343 Winstead Elementary School Public
- 344 Woodland Elementary Public
- 345 Woodland Middle School Public

- [164 rows x 2 columns]

PYTHON PROGRAMLAMA

- # Filtreleme işlemi(Filtering):

- print (df[df.school_rating > 4] [['school_rating','name', 'school_type']])

PYTHON PROGRAMLAMA

	school_rating	name	school_type
0	5.0	Allendale Elementary School	Public
7	5.0	Beech Elementary	Public
13	5.0	Blackman Elementary School	Public
20	5.0	Brentwood High School	Public
21	5.0	Brentwood Middle School	Public
..	...	...	...
302	5.0	Trinity Elementary	Public
309	5.0	Union Elementary School	Public
314	5.0	Walnut Grove Elementary	Public
343	5.0	Winstead Elementary School	Public
345	5.0	Woodland Middle School	Public

[78 rows x 3 columns]

PYTHON PROGRAMLAMA

- # DataFrame'deki ilk 5 verinin listesi

- print (df.head())

PYTHON PROGRAMLAMA

- name school_rating size reduced_lunch \
- 0 Allendale Elementary School 5.0 851.0 10.0
- 1 Anderson Elementary 2.0 412.0 71.0
- 2 Avoca Elementary 4.0 482.0 43.0
- 3 Bailey Middle 0.0 394.0 91.0
- 4 Barfield Elementary 4.0 948.0 26.0

PYTHON PROGRAMLAMA

- # DataFrame'deki son 5 verinin listesi:

- print (df.tail())

PYTHON PROGRAMLAMA

- name school_rating size reduced_lunch \
- 342 Winfree Bryant Middle School 3.0 611.0 57.0
- 343 Winstead Elementary School 5.0 515.0 8.0
- 344 Woodland Elementary 4.0 424.0 55.0
- 345 Woodland Middle School 5.0 866.0 2.0
- 346 Wright Middle 0.0 829.0 89.0

PYTHON PROGRAMLAMA

- # Dataset hakkında genel bir bilgi

- print (df.info())

PYTHON PROGRAMLAMA

- <class 'pandas.core.frame.DataFrame'>
- RangeIndex: 347 entries, 0 to 346
- Data columns (total 15 columns):
- name 347 non-null object
- school_rating 347 non-null float64
- size 347 non-null float64
- reduced_lunch 347 non-null float64
- state_percentile_16 347 non-null float64
- state_percentile_15 341 non-null float64
- stu_teach_ratio 347 non-null float64
- school_type 347 non-null object
- avg_score_15 341 non-null float64
- avg_score_16 347 non-null float64
- full_time_teachers 347 non-null float64
- percent_black 347 non-null float64
- percent_white 347 non-null float64
- percent_asian 347 non-null float64
- percent_hispanic 347 non-null float64
- dtypes: float64(13), object(2)
- memory usage: 40.8+ KB
- None

PYTHON PROGRAMLAMA

- # Dataset hakkında genel bir bilgi

- print(df.info(verbose=False))

PYTHON PROGRAMLAMA

- <class 'pandas.core.frame.DataFrame'>
- RangeIndex: 347 entries, 0 to 346
- Columns: 15 entries, name to percent_hispanic
- dtypes: float64(13), object(2)
- memory usage: 40.8+ KB
- None

PYTHON PROGRAMLAMA

- info() Metodu
- info() metodu, dataset ile ilgili olarak satır ve sütun sayısını, null olmayan verilerin sayısını göstermektedir.
- Her sütundaki veri türü de bu metot gösterilmektedir.
- dataFrame'in kullandığı bellek miktarı da gösterilmektedir.

PYTHON PROGRAMLAMA

- # genel istatistiki bilgi

- df.describe()

PYTHON PROGRAMLAMA

	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
count	347.000000	347.000000	347.000000	347.000000	341.000000	347.000000	341.000000	347.000000	347.000000	347.000000	347.000000	347.000000	347.000000
mean	2.968300	699.472622	50.279539	58.801729	58.249267	15.461671	57.004692	57.049856	44.939481	21.197983	61.673487	2.642651	11.164553
std	1.690377	400.598636	25.480236	32.540747	32.702630	5.725170	26.696450	27.968974	22.053386	23.562538	27.274859	3.109629	12.030608
min	0.000000	53.000000	2.000000	0.200000	0.600000	4.700000	1.500000	0.100000	2.000000	0.000000	1.100000	0.000000	0.000000
25%	2.000000	420.500000	30.000000	30.950000	27.100000	13.700000	37.600000	37.000000	30.000000	3.600000	40.600000	0.750000	3.800000
50%	3.000000	595.000000	51.000000	66.400000	65.800000	15.000000	61.800000	60.700000	40.000000	13.500000	68.700000	1.600000	6.400000
75%	4.000000	851.000000	71.500000	88.000000	88.600000	16.700000	79.600000	80.250000	54.000000	28.350000	85.950000	3.100000	13.800000
max	5.000000	2314.000000		98.000000	99.800000	99.800000	111.000000	99.000000	98.900000	140.000000			

PYTHON PROGRAMLAMA

- # Sütun için genel istatistiki bilgi

- df["school_rating"].describe()

PYTHON PROGRAMLAMA

- count 347.000000
- mean 2.968300
- std 1.690377
- min 0.000000
- 25% 2.000000
- 50% 3.000000
- 75% 4.000000
- max 5.000000
- Name: school_rating, dtype: float64

PYTHON PROGRAMLAMA

- # Sütun için genel istatistiki bilgi

- df["school_rating"].describe(include="all")

PYTHON PROGRAMLAMA

- count 347.000000
- mean 2.968300
- std 1.690377
- min 0.000000
- 25% 2.000000
- 50% 3.000000
- 75% 4.000000
- max 5.000000
- Name: school_rating, dtype: float64

PYTHON PROGRAMLAMA

- # İki sütun için genel istatistiki bilgi

- df[["school_rating", "size"]].describe()

PYTHON PROGRAMLAMA

	school_rating	size
count	347.000000	347.000000
mean	2.968300	699.472622
std	1.690377	400.598636
min	0.000000	53.000000
25%	2.000000	420.500000
50%	3.000000	595.000000
75%	4.000000	851.000000
max	5.000000	2314.000000

PYTHON PROGRAMLAMA

- # Pandas groupby methodu

PYTHON PROGRAMLAMA

- df[['reduced_lunch', 'school_rating']].groupby(['school_rating']).describe()

PYTHON PROGRAMLAMA

reduced_lunch								
	count	mean	std	min	25%	50%	75%	max
school_rating								
0.0	43.0	83.581395	8.813498	53.0	79.50	86.0	90.00	98.0
1.0	40.0	74.950000	11.644191	53.0	65.00	74.5	84.25	98.0
2.0	44.0	64.272727	11.956051	37.0	54.75	62.5	74.00	88.0
3.0	56.0	50.285714	13.550866	24.0	41.00	48.5	63.00	78.0
4.0	86.0	41.000000	16.681092	4.0	30.00	41.5	50.00	87.0
5.0	78.0	21.602564	17.651268	2.0	8.00	19.0	29.75	87.0

PYTHON PROGRAMLAMA

- # İki sütun arasındaki korelasyonun bulunması

- df[['reduced_lunch', 'school_rating']].corr()

PYTHON PROGRAMLAMA

- | | reduced_lunch | school_rating |
- | reduced_lunch | 1.000000 | -0.815757 |
- | school_rating | -0.815757 | 1.000000 |

PYTHON PROGRAMLAMA

- Dataset grafikleri:

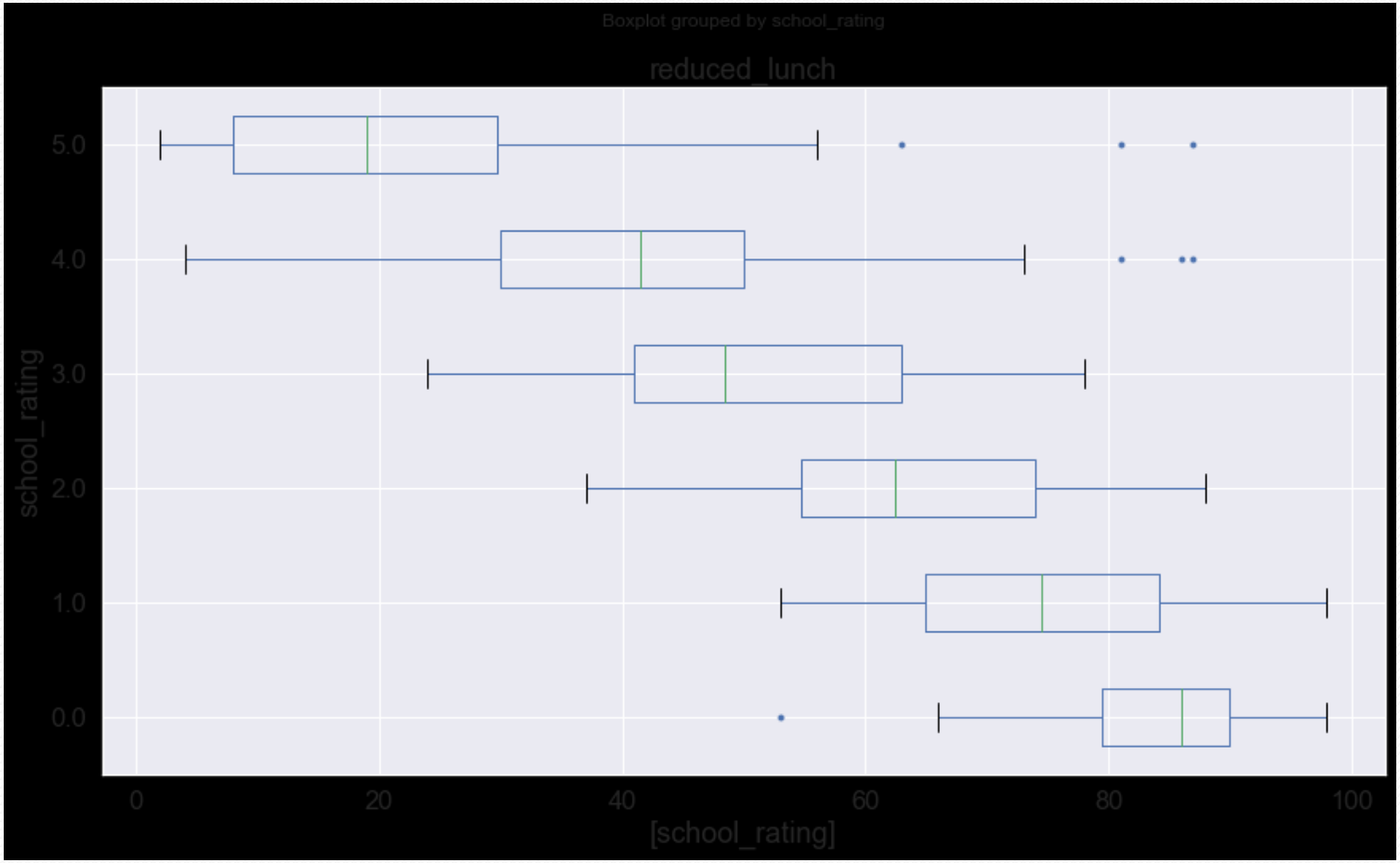
PYTHON PROGRAMLAMA

- Box-and-Whisker Plot ortalamadan uzaklığı göstermektedir.

PYTHON PROGRAMLAMA

- fig, ax = plt.subplots(figsize=(14,8))
- ax.set_ylabel('school_rating')
- # iki sütun ile grafik oluşturulması
- _ = df[['reduced_lunch', 'school_rating']].boxplot(by='school_rating', figsize=(13,8), vert=False, sym='b.', ax=ax)

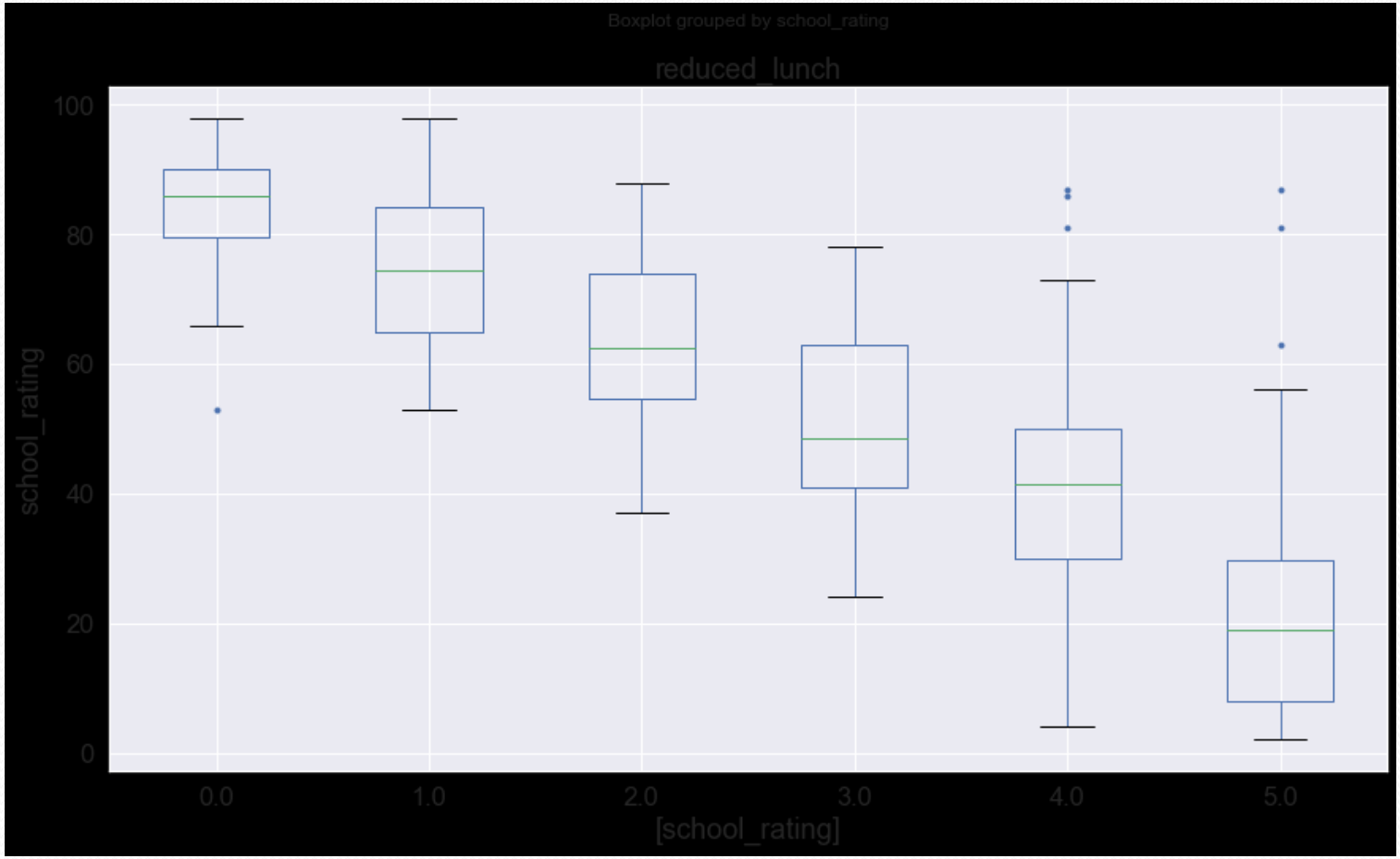
PYTHON PROGRAMLAMA



PYTHON PROGRAMLAMA

- fig, ax = plt.subplots(figsize=(14,8))
- ax.set_ylabel('school_rating')
- # iki sütun ile dikey grafik oluşturulması
- _ = df[['reduced_lunch', 'school_rating']].boxplot(by='school_rating', figsize=(13,8), vert=True, sym='b.', ax=ax)

PYTHON PROGRAMLAMA

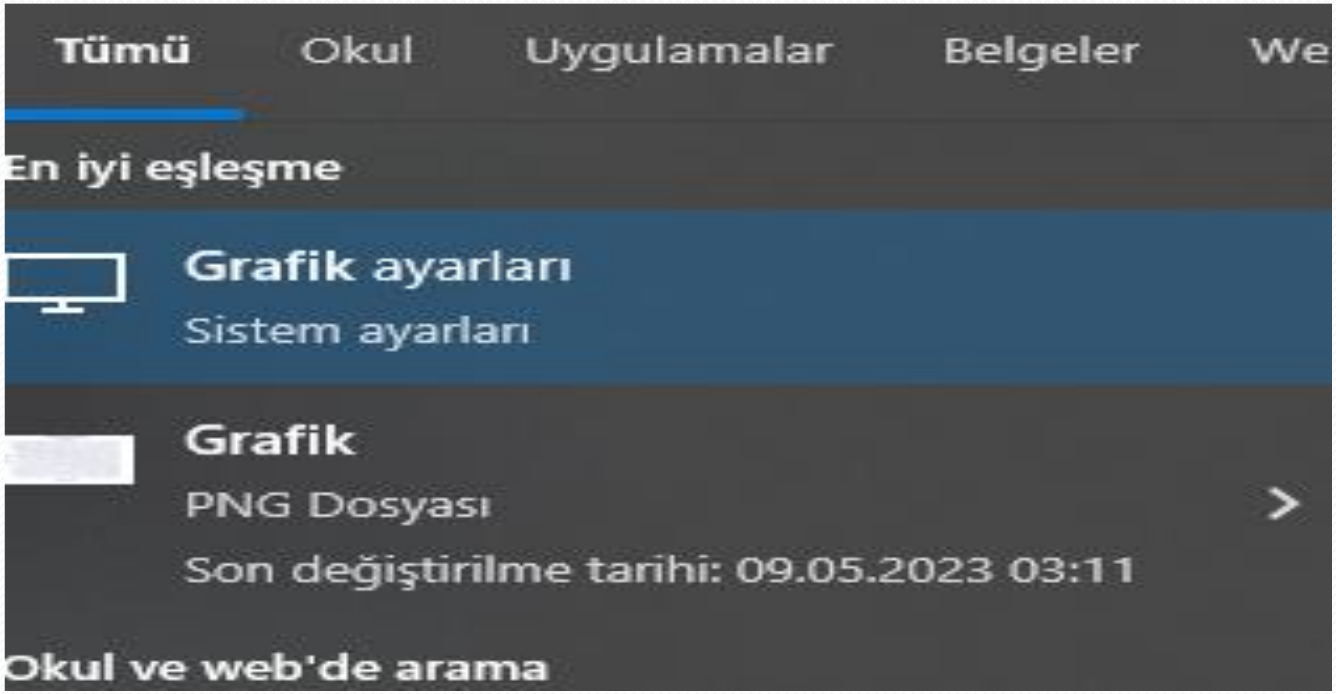


PYTHON PROGRAMLAMA

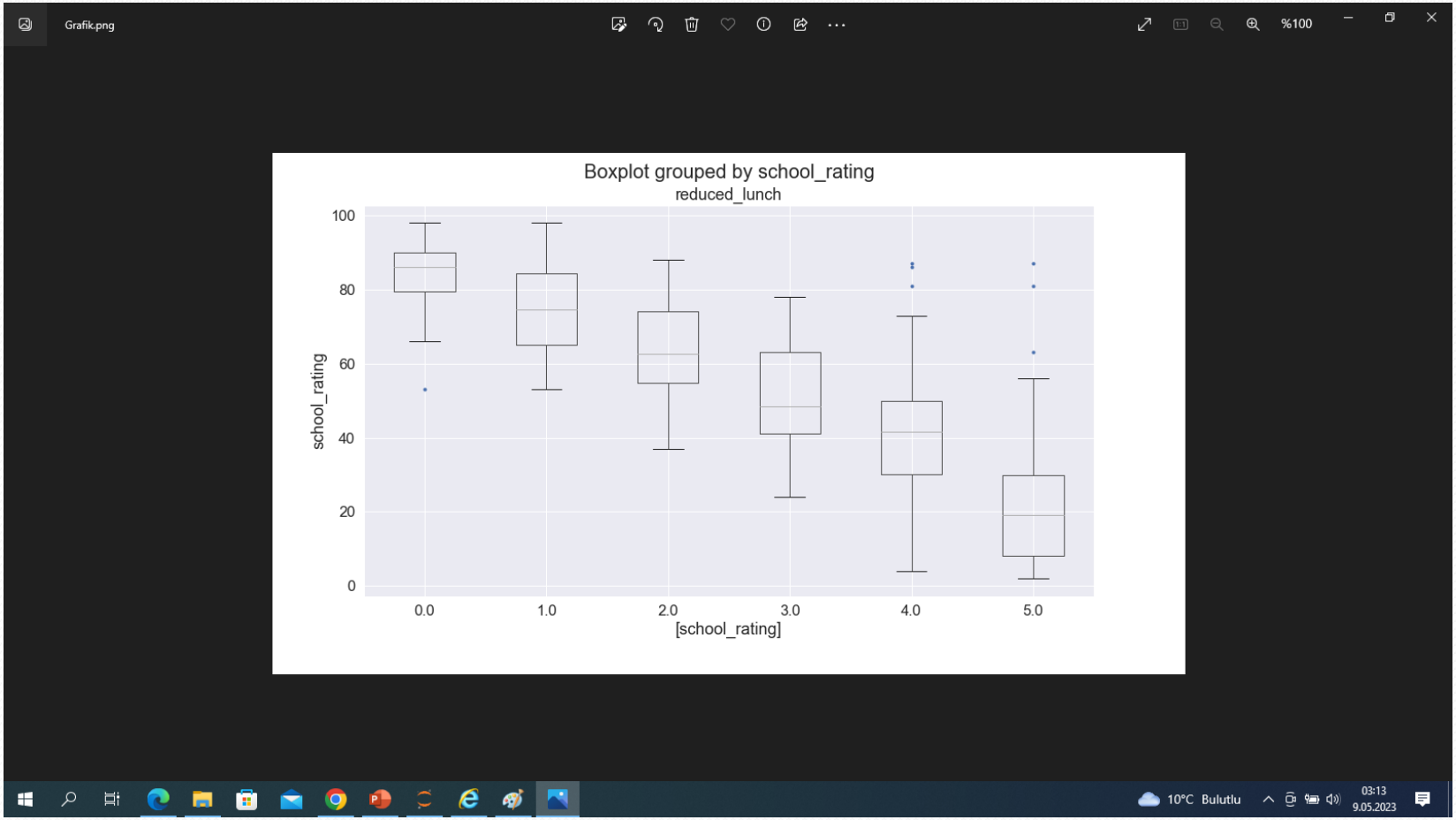
- # grafiğin kaydedilmesi

- fig.savefig("Grafik.png")

PYTHON PROGRAMLAMA



PYTHON PROGRAMLAMA



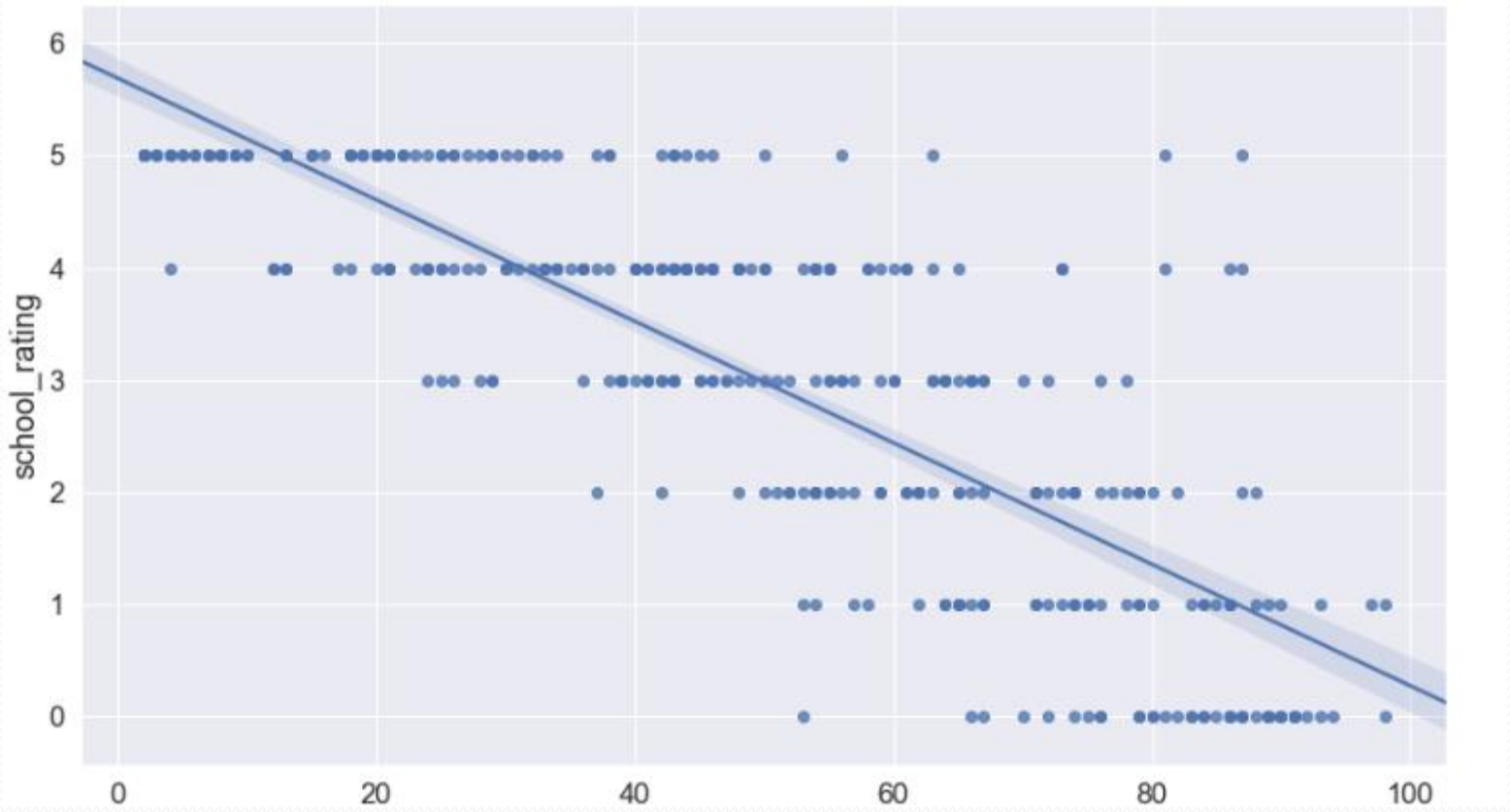
PYTHON PROGRAMLAMA

- Scatter Plot – serpilme veya saçılım diyagramı

PYTHON PROGRAMLAMA

- plt.figure(figsize=(14,8)) # grafik boyutunun belirlenmesi
- _ = sns.regplot(data=df, x='reduced_lunch', y='school_rating')

PYTHON PROGRAMLAMA



PYTHON PROGRAMLAMA

- Korelasyon Matrisi

PYTHON PROGRAMLAMA

- Kırmızı renk – positif korelasyon
- Mavi renk – negatif korelasyon
- Beyaz renk – korelasyon mevcut değil

PYTHON PROGRAMLAMA

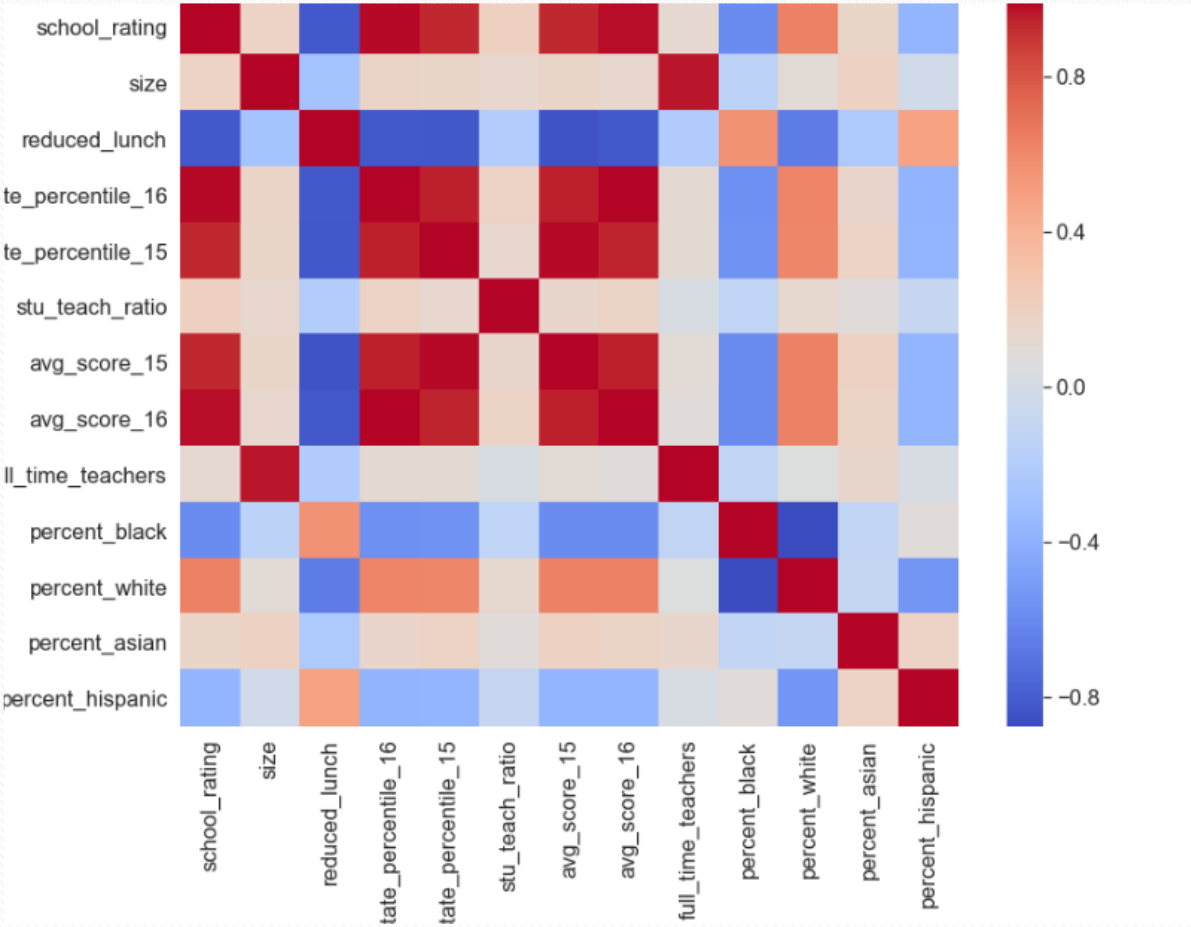
```
# Korelasyon matrisinin oluşturulması

corr = df.corr()
_, ax = plt.subplots(figsize=(13,10))

# Korelasyon matrisi grafiği

_ = sns.heatmap(corr, ax=ax,
                xticklabels=corr.columns.values,
                yticklabels=corr.columns.values,
                cmap='coolwarm')
```

PYTHON PROGRAMLAMA



PYTHON PROGRAMLAMA

Kaynaklar:

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

https://tr.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

https://www.pluralsight.com/guides/controlling-figure-aesthetics