Marta Koczerska
Marcelina Brodacka

Analiza danych statystycznych na temat szkół w Nowym Jorku

Poznań 2018

Spis treści

1. Wprowadzenie

1.1. Opis zbioru danych i wykorzystanych zmiennych

Zbiór danych, na których opiera się poniższa praca to: PASSNYC School Explorer (data dostępu: 10.12.2018r.), dostępna na stronie https://www.kaggle.com. Zawiera ona sumaryczne charakterystyki szkół, obejmujące demografię studentów i wystandaryzowane wyniki testów z publicznych źródeł danych dla 1272 nowojorskich szkół podstawowych i gimnazjów.

1.2. Cel pracy

Celem naszej pracy jest analiza danych dotyczących szkół w różnych okręgach Nowego Jorku. Ocenione zostaną związki między danymi dotyczącymi finansów, demografii, oceny zaufania do szkoły, średniej obecności oraz ocen z poszczególnych przedmiotów i egzaminów.

1.3. Opis wykorzystywanych zmiennych

1.3.1. Dane nominalne

Szkoła publiczna (Community School?)
Liczba danych: 1272
Liczba przyjmowanych wartości: 2
Dominanta: No
Liczba wystąpień dominanty: 1196

Okręg/miasto (City)
Liczba danych: 1272
Liczba przyjmowanych wartości: 45
Dominanta: BROOKLYN
Liczba wystąpień dominanty: 411

1.3.2. Dane ilościowe

Procent uczniów uczących się angielskiego (Percent ELL)
Liczba danych: 1272
Średnia: 12.48
Odchylenie standardowe: 11.36
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 4
Mediana: 9
Trzeci kwartyl: 17

Maksymalna wartość zmiennej to 99

Średni procent zaufania (Trust %)
Liczba danych: 1247
Średnia: 90.42
Odchylenie standardowe: 6.12
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 87
Mediana: 92
Trzeci kwartyl: 94
Maksymalna wartość zmiennej to 100

Średnia biegłość w matematyce (Average Math Proficiency)
Liczba danych: 1272
Średnia: 2.55
Odchylenie standardowe: 0.71
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 2.26
Mediana: 2.54
Trzeci kwartyl: 2.97
Maksymalna wartość zmiennej to 4.2

Średnia biegłość w języku angielskim (Average ELA Proficiency)
Liczba danych: 1272
Średnia: 2.42
Odchylenie standardowe: 0.63
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 2.23
Mediana: 2.43
Trzeci kwartyl: 2.74
Maksymalna wartość zmiennej to 3.93

Procent ludności azjatyckiej (Percent Asian)
Liczba danych: 1272
Średnia: 11.65
Odchylenie standardowe: 17.65
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 1
Mediana: 4
Trzeci kwartyl: 14
Maksymalna wartość zmiennej to 95

Wskaźnik zapotrzebowania na pomoc finansową (Economic Need Index)
Liczba danych: 1272
Średnia: 0.66
Odchylenie standardowe: 0.23
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 0.54
Mediana: 0.72
Trzeci kwartyl: 0.84
Maksymalna wartość zmiennej to 0.96

Szacowany dochód szkoły (School Income Estimate)
Liczba danych: 1272
Średnia: 33361.78
Odchylenie standardowe: 28576.41
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 0.00
Mediana: 34299.11
Trzeci kwartyl: 50769.81
Maksymalna wartość zmiennej to 181382.06

Wskaźnik obecności na zajęciach (Student Attendance Rate)
Liczba danych: 1272
Średnia: 90,9
Odchylenie standardowe: 15,51
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 92
Mediana: 94
Trzeci kwartyl: 95
Maksymalna wartość zmiennej to 100

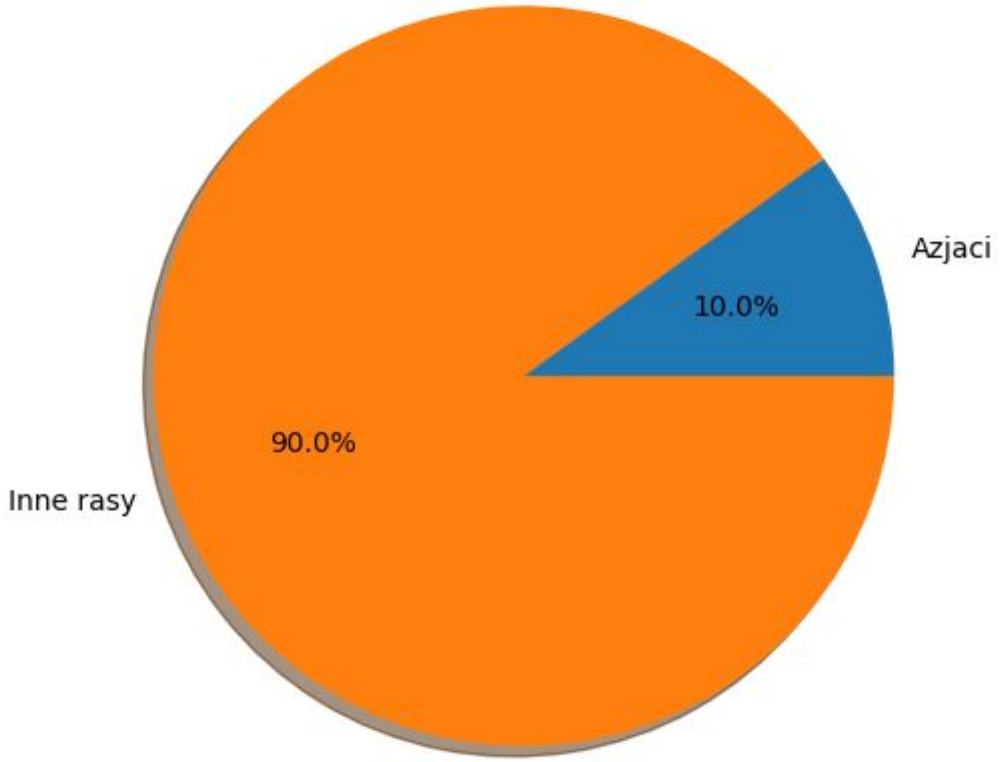
Procent ludności hiszpańskiej (Percent Hispanic)
Liczba danych: 1272
Średnia: 41.15
Odchylenie standardowe: 26.15
Minimalna wartość zmiennej to 2
Pierwszy kwartyl: 18
Mediana: 35.50
Trzeci kwartyl: 64
Maksymalna wartość zmiennej to 100

Procent ludności czarnoskórej (Percent Black)
Liczba danych: 1272
Średnia: 31.99
Odchylenie standardowe: 28.82
Minimalna wartość zmiennej to 0
Pierwszy kwartyl: 6
Mediana: 24
Trzeci kwartyl: 55.25
Maksymalna wartość zmiennej to 97

2. Część główna pracy

2.1. Jaki procent uczniów szkół w Brooklynie stanowią uczniowie rasy azjatyckiej?

Procent Azjatów w szkołach w Brooklynie wynosi: 9.94, zatem ok 10%.

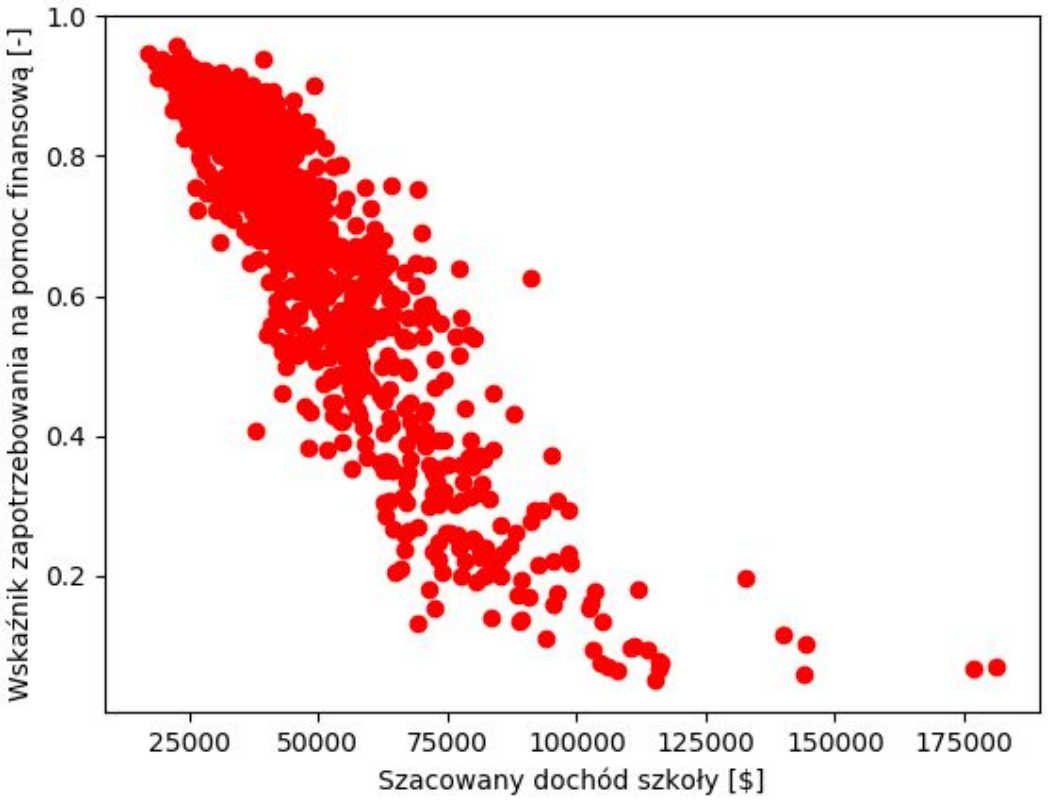


Ryc. 1. Wykres przedstawia procentowy udział Azjatów wśród uczniów szkół w Brooklynie.

2.2. Czy istnieje korelacja pomiędzy wskaźnikiem zapotrzebowania na pomoc finansową a szacowanymi dochodami danej szkoły?

Wyznaczona wartość współczynnika korelacji Pearsona: -0.48
Wyznaczony poziom istotności: $7.83*e^{-74}$

Poziom istotności mniejszy niż 0,05 oznacza, że związek jest istotny statystycznie. Ujemna wartość współczynnika korelacji wskazuje na przeciw proporcjonalną zależność pomiędzy zmiennymi. Co więcej, korelują one w sposób umiarkowany. Im wyższe szacowane dochody szkoły, tym mniejsze jej zapotrzebowanie na pomoc finansową.



Ryc .2. Wykres przedstawia zależność pomiędzy szacowanym dochodem szkoły wskaźnikiem zapotrzebowania na pomoc finansową.

2.3. Czy istnieje związek pomiędzy wskaźnikiem obecności na zajęciach a odsetkiem uczniów z Hiszpanii w danej szkole?

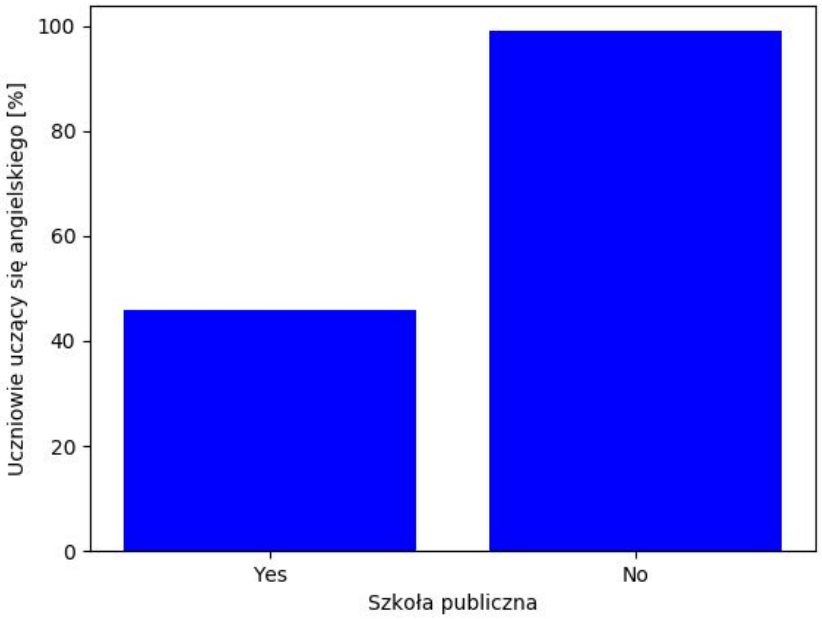
Wyznaczona wartość współczynnika korelacji Pearsona: 0.005
Wyznaczony poziom istotności: 0.86
Poziom istotności większy niż 0,05 oznacza, że korelacja pomiędzy zmiennymi jest nieistotna statystycznie. Nie ma zatem związku pomiędzy wskaźnikiem obecności na zajęciach a odsetkiem uczniów z Hiszpanii w danej szkole.

2.4. W jakim okręgu Nowego Jorku jest najwyższy odsetek uczniów ras innych niż biała?

Okręg w Nowym Jorku z najwyższym odsetkiem uczniów ras innych niż biała to: BROOKLYN. Wynik ten uzyskano poprzez utworzenie nowej zmiennej łączącej trzy zmienne: Procent ludności azjatyckiej, Procent ludności czarnoskórej i Procent ludności hiszpańskiej, a następnie wykorzystanie jej do porównania okręgów Nowego Jorku pod kątem odsetka uczniów wymienionych ras.

2.5. Czy istnieje związek pomiędzy ilością osób uczącą się angielskiego a tym, czy szkoła jest prywatna czy publiczna?

Wyznaczona wartość testu Shapiro-Wilka 0.85
Wyznaczona wartość istotności $1.8*e^{-33}$
Poziom istotności wyniku testu Shapiro-Wilka jest mniejszy niż 0,05, co oznacza, że rozkład zmiennej procent uczniów uczących się języka angielskiego nie jest zgodny z normalnym.
Wyznaczona wartość testu Levene'a 1120.12
Wyznaczona wartość istotności $8.53*e^{-204}$
Poziom istotności testu Levene'a jest mniejszy niż 0,05, co oznacza, że zmienne procent uczniów uczących się języka angielskiego i szkoła publiczna mają niejednorodne wariancje.
Wyznaczona wartość testu Manna-Whitneya 26060.0
Wyznaczona wartość istotności 0.0
Poziom istotności testu Manna-Whitneya jest mniejszy niż 0,05, co oznacza, że występują istotne różnice w średnim procencie uczniów uczących się języka angielskiego w szkołach publicznych i szkołach prywatnych, jest on wyższy w szkołach prywatnych.



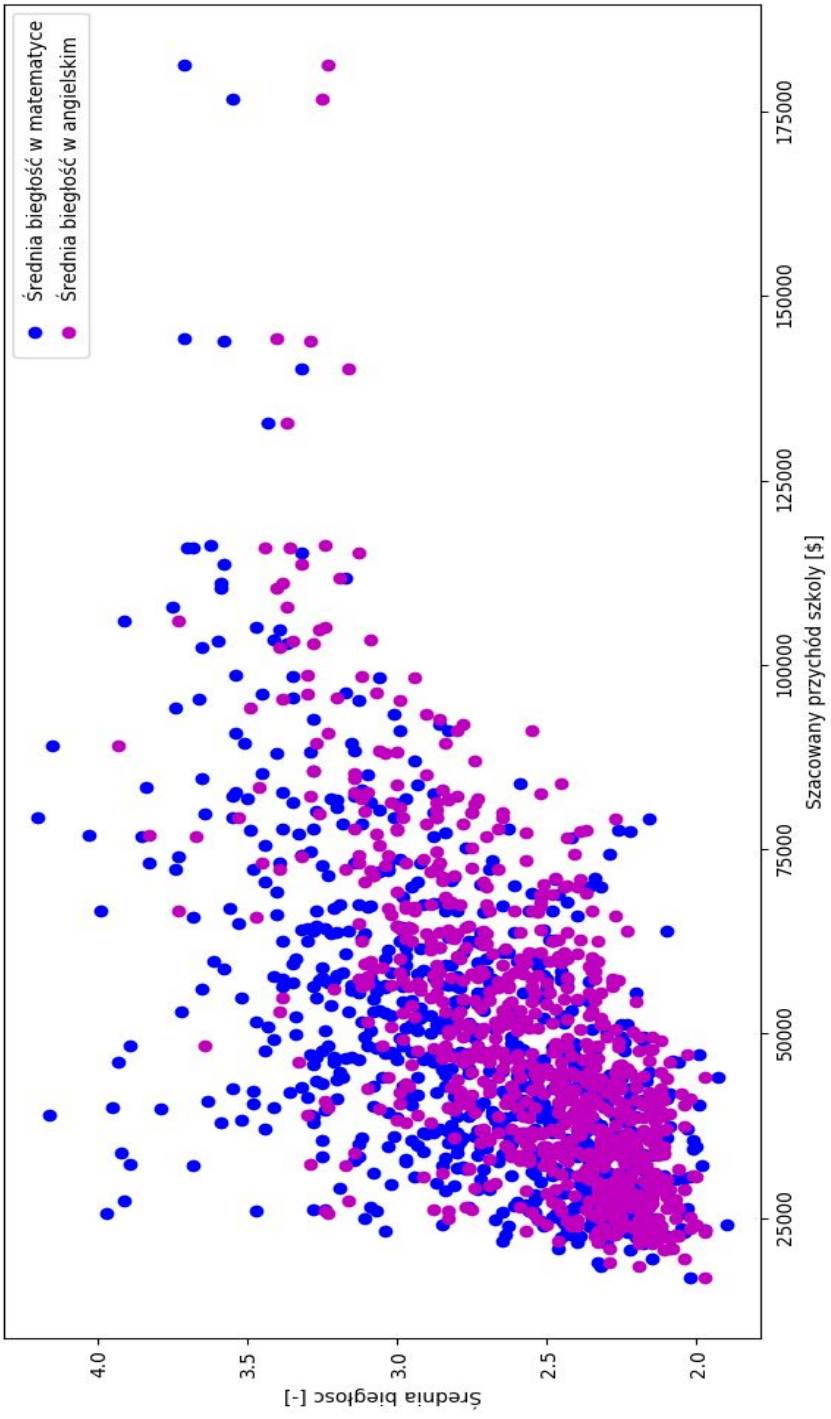
Ryc. 4. Wykres przedstawia zależność procentu uczniów uczących się języka angielskiego w zależności od tego, czy szkoła jest publiczna.

2.6. W jakim okręgu Nowego Jorku szkoły mają najwyższy średni procent zaufania?

Okręgi o największym procencie zaufania to: ['NEW YORK', 'NEW YORK', 'BRONX']
Maksymalny procent zaufania to: 100.0
Aby uzyskać wynik, znaleziono maksymalny procent zaufania i porównano z nim procent zaufania poszczególnych szkół i zapisano ich okręgi. Najwyższy maksymalny procent zaufania, równy 100%, uzyskały dwie szkoły w okręgu Nowy Jork oraz Bronx.

2.7. Jaki związek z przychodem szkoły ma średnia biegłość uczniów w posługiwaniu się językiem angielskim, a jaki średnia biegłość w matematyce?

Dla średniej biegłości w posługiwaniu się językiem angielskim: Wyznaczona wartość współczynnika korelacji Pearsona: 0.26. Wyznaczony poziom istotności: $7.48*e^{-22}$
Poziom istotności poniżej 0,05 oznacza, że związek między zmiennymi średnia biegłość uczniów w posługiwaniu się językiem angielskim i dochód szkoły jest istotny statystycznie. Jest to korelacja o słabej sile.
Dla średniej biegłości w matematyce: Wyznaczona wartość współczynnika korelacji Pearsona: 0.2. Wyznaczony poziom istotności: $2.12*e^{-12}$
Poziom istotności poniżej 0,05 oznacza, że związek między zmiennymi średnia biegłość uczniów w matematyce i dochód szkoły jest istotny statystycznie. Jest to korelacja o słabej sile.



Ryc. 5. Wykres przedstawia zależność średniej biegłości w matematyce i posługiwaniu się językiem angielskim w zależności od szacowanego przychodu szkoły.

3. Podsumowanie

W powyższej pracy wykonane zostały obliczenia, obejmujące zmienne dotyczące uczniów szkół w Nowym Jorku i podległych okręgach. Najwyższym odsetkiem uczniów rasy innej niż biała cechuje się szkoła na Brooklynie. Maksymalną wartość wskaźnika zaufania, 100%, otrzymały dwie szkoły w Nowym Jorku oraz na Bronksie. Otrzymano wynik, wskazujący na średni procent Azjatów w szkołach w Nowym Jorku, jest to 10%. Zbadano związki między: wskaźnikiem zapotrzebowania na pomoc finansową a szacowanymi dochodami danej szkoły, wskaźnikiem obecności na zajęciach a odsetkiem uczniów z Hiszpanii w danej szkole, ilością osób uczącą się angielskiego a tym, czy szkoła jest prywatna czy publiczna, przychodem szkoły a średnią biegłością uczniów w posługiwaniu się językiem angielskim oraz średnią biegłością w matematyce. Otrzymane wyniki świadczą o tym, że im wyższe szacowane dochody szkoły, tym mniejsze jej zapotrzebowanie na pomoc finansową. Średni procent uczniów uczących się języka angielskiego jest wyższy w szkołach prywatnych. Wykazano również, że istnieje słaby związek między dochodem szkoły oraz średnią biegłością w języku angielskim oraz matematyce.

4. Kod

2.1.

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("C:/Users/acer/Desktop/baza.csv.csv")
zmiennaMiasto = dane["City"]
zmiennaAzjaci = dane["Percent Asian"]
zmiennaMiasto = zmiennaMiasto.fillna(0)
zmiennaAzjaci = zmiennaAzjaci.fillna(0)
Brooklyn = dane[dane.City == "BROOKLYN"]
wynik = (Brooklyn["Percent Asian"])
srednia = wynik.mean()
print(“Procent Azjatów w szkołach w Brooklynie wynosi:”,  srednia)

rasy = ["Azjaci", "Inne rasy"]
procenty = pd.Series([10, 90])
plt.figure(1, figsize = (6,6))
plt.pie(procenty, labels = rasy, autopct = '%1.1f%%', shadow = True)
plt.show()
```

2.2.

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("C:/Users/acer/Desktop/baza.csv.csv")
zmiennaPomoc = dane["Economic Need Index"]
zmiennaDochod = dane["School Income Estimate"]
zmiennaDochod = zmiennaDochod.fillna(0)
zmiennaPomoc = zmiennaPomoc.fillna(0)
outper = st.pearsonr(zmiennaDochod, zmiennaPomoc)
print("Wyznaczona wartość współczynnika korelacji Pearsona:", outper[0])
print("Wyznaczony poziom istotności:", outper[1])
```

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("C:/Users/acer/Desktop/baza.csv.csv")
zmiennaPomoc = dane["Economic Need Index"]
zmiennaDochod = dane["School Income Estimate"]
plt.plot(zmiennaDochod,zmiennaPomoc,'ro')
plt.xlabel("Szacowany dochód szkoły [$]")
plt.ylabel("Wskaźnik zapotrzebowania na pomoc finansową [-]")
plt.show()
```

2.3.

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("C:/Users/acer/Desktop/baza.csv.csv")
zmiennaObecnosc = dane["Student Attendance Rate"]
zmiennaHiszpanie = dane["Percent Hispanic"]
zmiennaObecnosc = zmiennaObecnosc.fillna(0)
zmiennaHiszpanie = zmiennaHiszpanie.fillna(0)
out = st.pearsonr(zmiennaObecnosc, zmiennaHiszpanie)
print("Wyznaczona wartość współczynnika korelacji Pearsona:", out[0])
print("Wyznaczony poziom istotności:", out[1])
```

2.4.

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("C:/Users/acer/Desktop/baza.csv.csv")
zmiennaAzjaci = dane["Percent Asian"]
zmiennaHiszpanie = dane["Percent Hispanic"]
zmiennaMurzyni = dane["Percent Black"]
zmiennaMiasto = dane["City"]
zmiennaAzjaci = zmiennaAzjaci.fillna(0)
zmiennaMurzyni = zmiennaMurzyni.fillna(0)
```

```
zmiennaHiszpanie = zmiennaHiszpanie.fillna(0)
zmiennaMiasto = zmiennaMiasto.fillna(0)

InneRasy = zmiennaAzjaci + zmiennaHiszpanie + zmiennaMurzyni
Naj = InneRasy[0]
for i in range(len(zmiennaMiasto)):
   if InneRasy[i] > Naj:
      Naj = InneRasy[i]
      miastoNaj = zmiennaMiasto[i]
print(“Okręg w Nowym Jorku z najwyższym odsetkiem uczniów ras innych niż biała to:”, miastoNaj)
```

2.5.

```
import pandas as pd
import scipy.stats as st
import numpy as np
dane = pd.read_csv("school.csv")
publiczna=dane["Community School?"]
publiczna = publiczna.str.lower().replace({'yes': 1, 'no': 0})
angielski=dane["Percent ELL"]
angielski=angielski.dropna(0)
outang=st.shapiro(angielski)
print("Wyznaczona wartość testu Shapiro-wilka", outang[0])
print("Wyznaczona wartość istotności", outang[1])
outleven=st.levene(angielski, publiczna)
print("Wyznaczona wartość testu Levene'a", outleven[0])
print("Wyznaczona wartość istotności", outleven[1])
outmannwhitneyu=st.mannwhitneyu(angielski,publiczna)
print("Wyznaczona wartość testu Manna-Whitneya", outmannwhitneyu[0])
print("Wyznaczona wartość istotności", outmannwhitneyu[1])

import pandas as pd
import scipy.stats as st
import numpy as np
import matplotlib.pyplot as plt
dane = pd.read_csv("school.csv")
publiczna=dane["Community School?"]
angielski=dane["Percent ELL"]
angielski=angielski.dropna(0)
```

```
plt.bar(publiczna,angielski, color='b')
plt.xlabel("Szkoła publiczna")
plt.ylabel("Uczniowie uczący się angielskiego [%]")
plt.show()
```

2.6.

```
import pandas as pd
import scipy.stats as st
dane = pd.read_csv("school.csv")
miasto=dane["City"]
zaufanie=dane["Trust "]
maksymalne=[]
max=zaufanie[0]
for i in range(len(zaufanie)):
    if zaufanie[i] > max:
        max = zaufanie[i]
        maksymalne.append(miasto[i])
print("Miasta o największym procencie zaufania to: ", maksymalne)
print("Maksymalny procent zaufania to: ", max)
```

2.7.

```
import pandas as pd
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("school.csv")
przychod=dane["School Income Estimate"]
przychod=przychod.fillna(0)
biegloscmatematyka=dane["Average Math Proficiency"]
biegloscmatematyka=biegloscmatematyka.fillna(0)
biegloscangielski=dane["Average ELA Proficiency"]
biegloscangielski=biegloscangielski.fillna(0)
outmat=st.pearsonr(przychod, biegloscmatematyka)
print("Wyznaczona wartość współczynnika korelacji Pearsona: ", outmat[0])
print("Wyznaczony poziom istotności: ", outmat[1])
outang=st.pearsonr(przychod, biegloscangielski)
print("Wyznaczona wartość współczynnika korelacji Pearsona: ", outang[0])
print("Wyznaczony poziom istotności: ", outang[1])

import pandas as pd
```

```
import scipy.stats as st
import matplotlib.pyplot as plt
dane = pd.read_csv("school.csv")
przychod=dane["School Income Estimate"]
biegloscmatematyka=dane["Average Math Proficiency"]
biegloscangielski=dane["Average ELA Proficiency"]
plt.plot(przychod, biegloscmatematyka,'ro', color='b',label="Średnia biegłość w matematyce")
plt.plot(przychod, biegloscangielski,'ro', color='m', label="Średnia biegłość w angielskim")
plt.legend()
plt.xlabel("Szacowany przychód szkoly [$]")
plt.ylabel("Średnia biegłosc [-]")
plt.show()
```