SPRAWOZDANIE

Zajęcia: Nauka o danych I
Prowadzący: prof. dr hab. Vasyl Martsenyuk

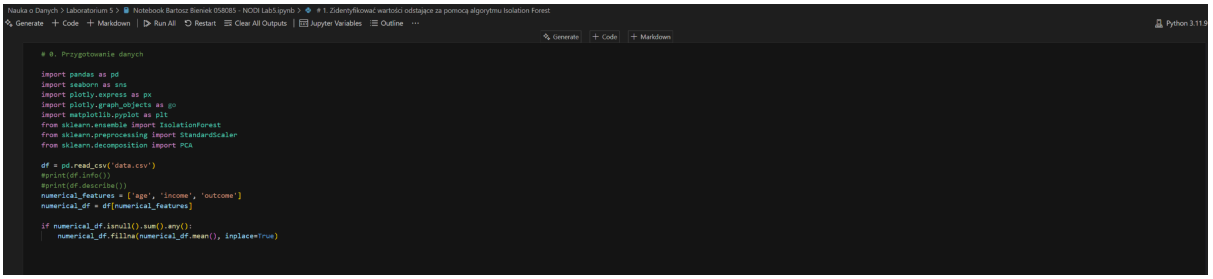
Laboratorium Nr 5 Data 23.11.2024 Temat: Wykorzystanie narzędzi do eksploracyjnej analizy danych (EDA) Wariant drugi (2)	Bartosz Bieniek Informatyka II stopień, stacjonarne, 1 semestr, gr.A

1. Polecenie

a. Zidentyfikować wartości odstające za pomocą algorytmu Isolation Forest,
b. Zredukować wymiarowość danych z użyciem PCA,
c. Stworzyć zaawansowane interaktywne wizualizacje danych,
d. Zwizualizować dane wielowymiarowe za pomocą zaawansowanych algorytmów (t-SNE, UMAP),
f. Stworzyć interaktywne wizualizacje danych w 2D i 3D,
g. Przeanalizować zależności między zmiennymi za pomocą macierzy korelacji.
h. Przeprowadzać testy statystyczne dla analizy różnic w grupach.

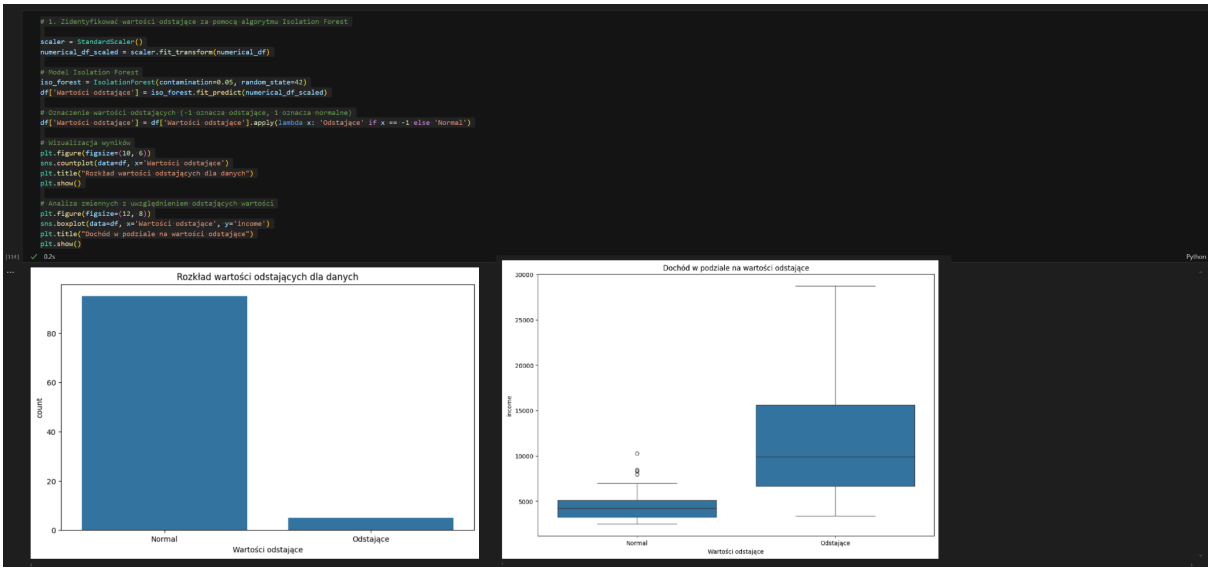
2. Opis programu opracowanego

https://github.com/mindgoner/Studia/tree/master/Nauka%20o%20Danych/Laboratorium%205



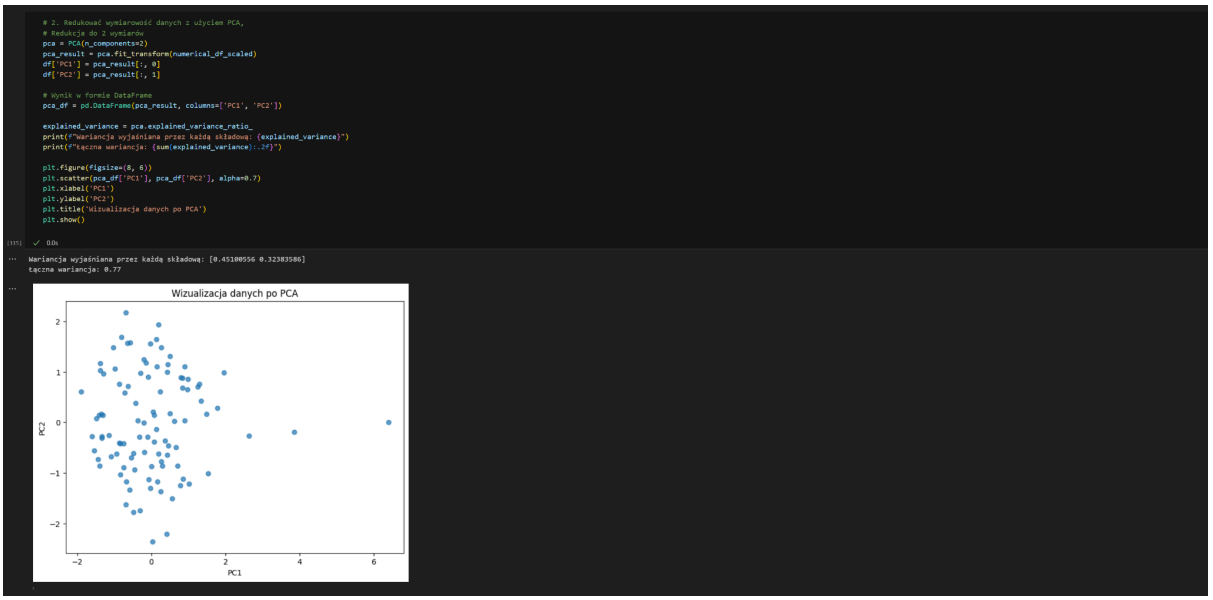
Rys. 1. Kod źródłowy.

Zaimportowano potrzebne biblioteki do analizy danych, wizualizacji i przetwarzania wstępnego. Wczytano dane z pliku CSV do DataFrame i wybrano kolumny numeryczne: age, income, outcome. Sprawdzono brakujące wartości i w razie potrzeby wypełniono je średnimi dla każdej kolumny.



Rys. 2. Kod źródłowy i efekt jego wykonania.

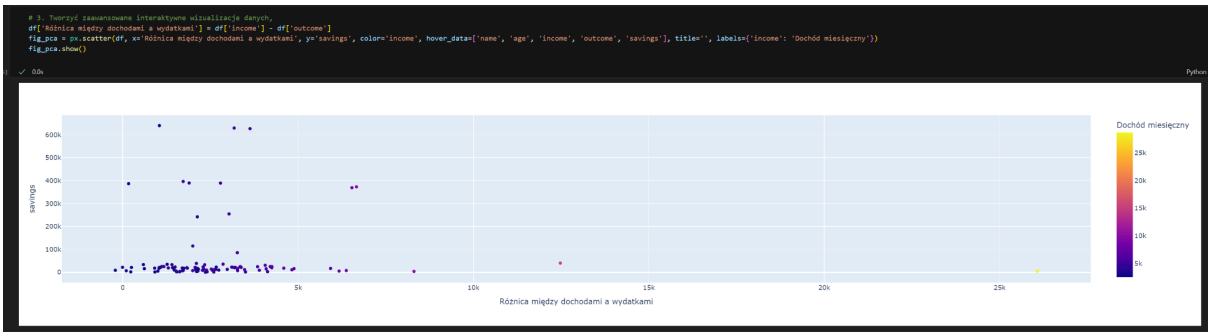
Przeskalowano dane numeryczne, aby algorytm Isolation Forest mógł poprawnie identyfikować wartości odstające. Wartości te oznaczono jako "Odstające" lub "Normal" i zapisano w nowej kolumnie DataFrame. Wyniki przedstawiono na wykresach: rozkład odstających oraz dochody w podziale na grupy.



Rys. 3. Kod źródłowy i efekt jego wykonania.

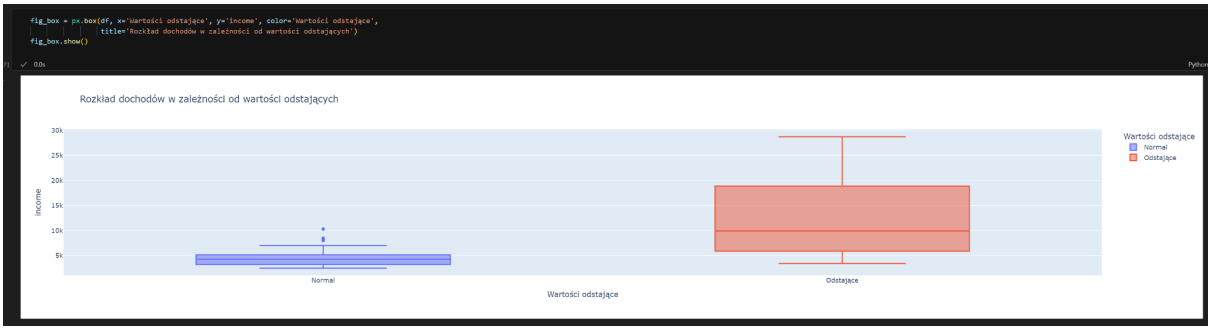
Zredukowano wymiarowość danych numerycznych do dwóch głównych składowych za pomocą PCA, zapisując wyniki jako PC1 i PC2 w DataFrame. Wyświetlono udział wariancji wyjaśnianej przez każdą składową oraz jej łączną

wartość. Wizualizację danych po redukcji wymiarowości wykonano za pomocą wykresu punktowego.



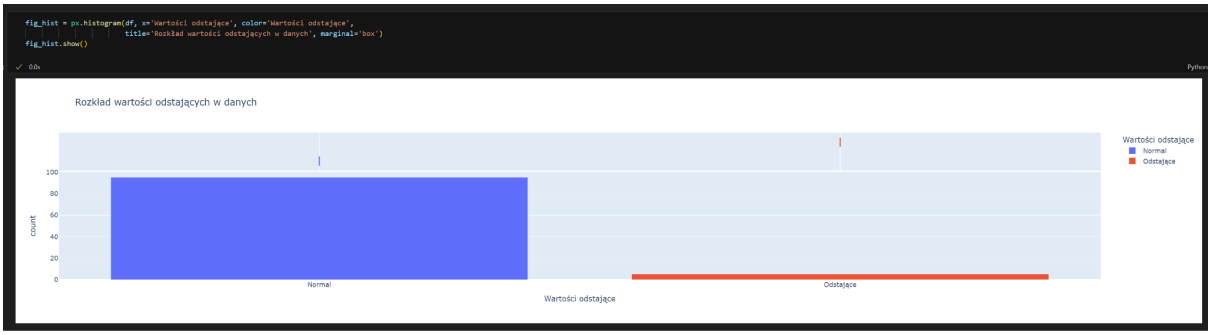
Rys. 4. Kod źródłowy i efekt jego wykonania.

Obliczono różnicę między dochodami a wydatkami i zapisano ją w nowej kolumnie. Utworzono interaktywną wizualizację punktową, przedstawiającą związek między różnicą dochodów i wydatków a oszczędnościami, z użyciem koloru do reprezentacji poziomu dochodów. Włączono dodatkowe dane do wyświetlania w chmurkach, takie jak imię, wiek czy szczegóły finansowe.



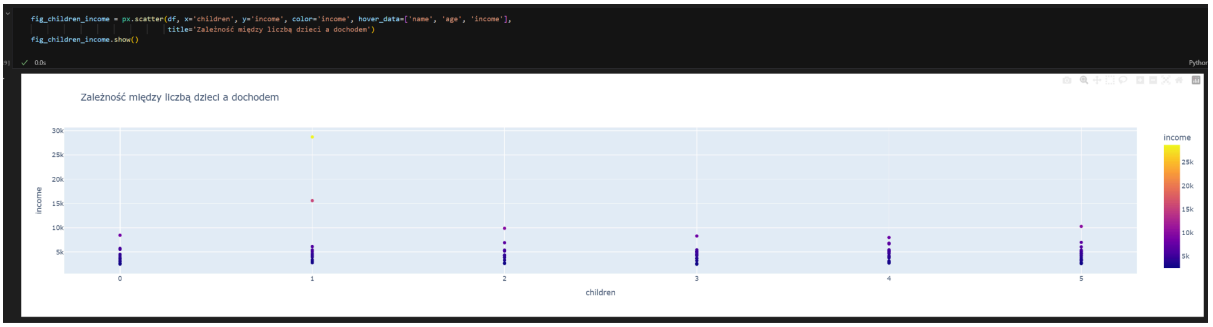
Rys. 5. Kod źródłowy i efekt jego wykonania.

Utworzono interaktywny wykres pudełkowy, który przedstawia rozkład dochodów w podziale na wartości oznaczone jako odstające i normalne. Kolorami oznaczono różne grupy, aby ułatwić porównanie. Wizualizacja pozwala zaobserwować wpływ wartości odstających na dochody.



Rys. 6. Kod źródłowy i efekt jego wykonania.

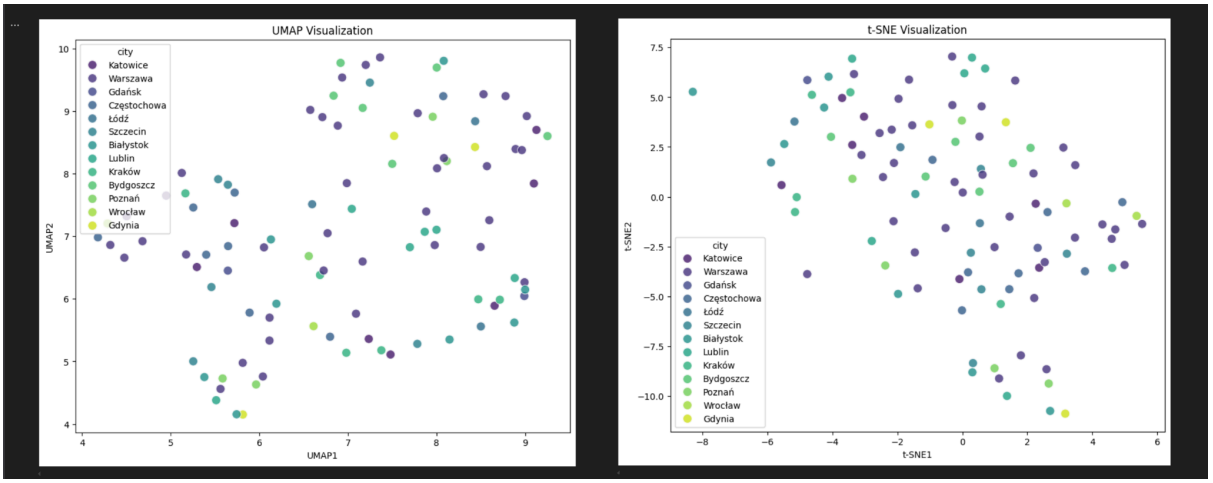
Stworzono interaktywny histogram przedstawiający rozkład wartości odstających w danych. Dodano margines w formie wykresu pudełkowego, aby uzupełnić analizę o dodatkowe informacje o rozkładzie. Kolory pozwalają łatwo odróżnić grupy wartości odstających i normalnych.



Rys. 7. Kod źródłowy i efekt jego wykonania.

Utworzono interaktywną wizualizację punktową, przedstawiającą zależność między liczbą dzieci a dochodem. Kolorami oznaczono różne poziomy dochodów, a dodatkowe dane wyświetlane w chmurkach pozwalają na bardziej szczegółową analizę.

Rys. 8. Kod źródłowy.



Rys. 9. Efekt wykonania kodu z poprzedniego rysunku.

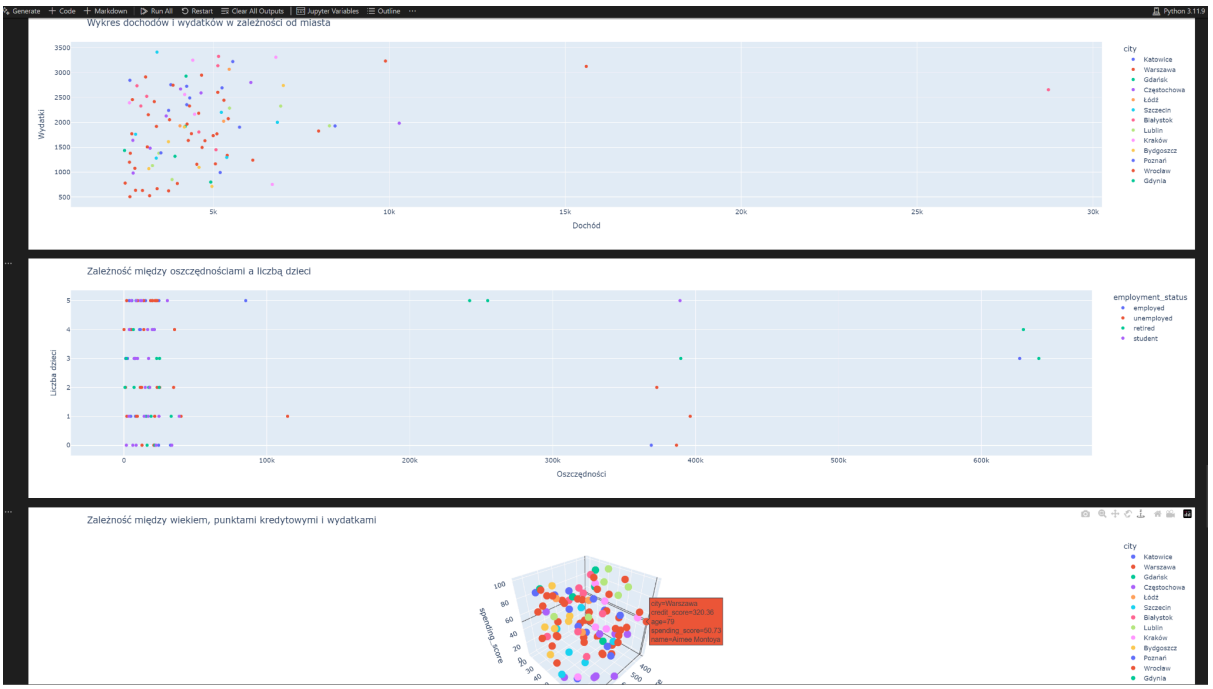
Przygotowano dane, wybierając kluczowe kolumny i skalując je za pomocą StandardScaler. Następnie zastosowano UMAP i t-SNE do redukcji wymiarowości do 2D. Wyniki obydwu algorytmów zwizualizowano, kolorując punkty według miasta, co pozwala lepiej zobrazować złożone relacje w danych. Wykorzystano wykresy punktowe z seaborn do przedstawienia wyników dla UMAP i t-SNE, co umożliwia analizę zróżnicowania między grupami.

```
# 5. tworzyć interaktywne wizualizacje danych w 2D i 3D,
fig = px.scatter(df, x='income', y='outcome', color='city', hover_data=['name', 'city'])
fig.update_layout(title="Wykres dochodów i wydatków w zależności od miasta",
                  xaxis_title="Dochód",
                  yaxis_title="Wydatki")
fig.show()


fig = px.scatter(df, x='savings', y='children', color='employment_status', hover_data=['name', 'employment_status'])
fig.update_layout(title="Zależność między oszczędnościami a liczbą dzieci",
                  xaxis_title="Oszczędności",
                  yaxis_title="Liczba dzieci")
fig.show()


fig = px.scatter_3d(df, x='credit_score', y='age', z='spending_score', color='city', hover_data=['name'])
fig.update_layout(title="Zależność między wiekiem, punktami kredytowymi i wydatkami",)
fig.show()
```

Rys. 10. Kod źródłowy.



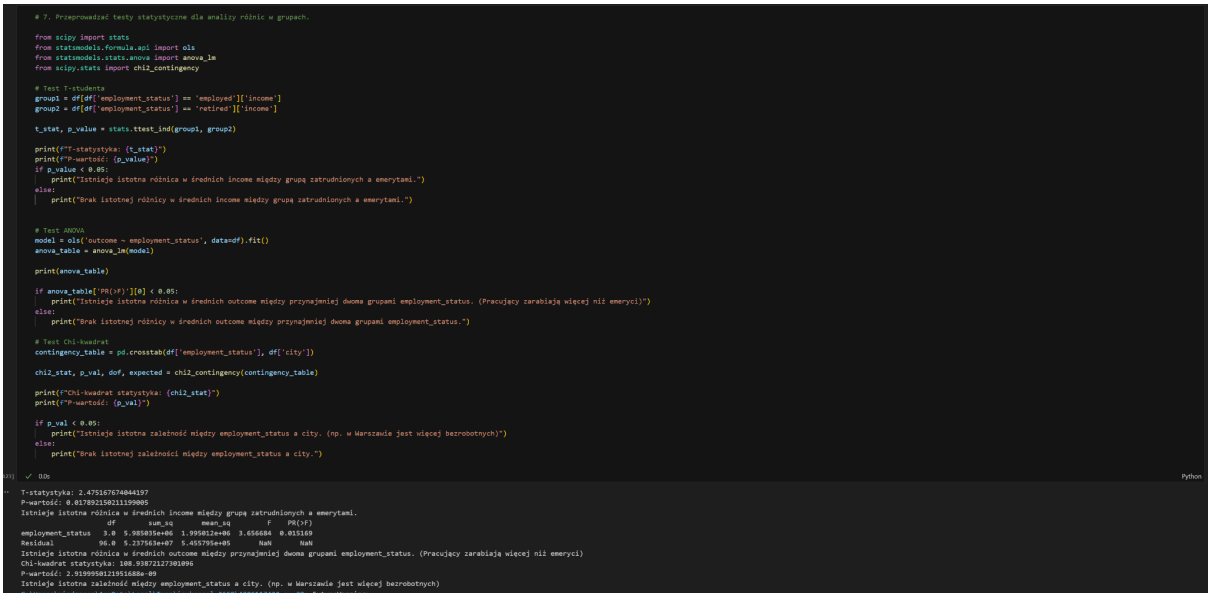
Rys. 11. Efekt wykonania kodu z poprzedniego rysunku.

W pierwszym kroku utworzono interaktywną wizualizację 2D, przedstawiającą zależność między dochodami a wydatkami, z uwzględnieniem miasta. Następnie stworzono drugi wykres 2D pokazujący związek między oszczędnościami a liczbą dzieci, z różnicowaniem na status zatrudnienia. Na końcu wykonano interaktywny wykres 3D, który ukazuje zależność między wiekiem, punktami kredytowymi i wydatkami, również z uwzględnieniem miasta.

Rys. 12. Kod źródłowy i efekt jego wykonania.

Zaimportowano niezbędne biblioteki i przygotowano dane numeryczne, uwzględniając kolumny income, outcome, savings. Następnie obliczono

macierz korelacji dla tych zmiennych, a wynik wizualizowano za pomocą interaktywnego heatmap. Wykres przedstawia związek między zmiennymi poprzez kolorystyczny rozkład wartości korelacji.



```python
# 7. Przeprowadzać testy statystyczne dla analizy różnic w grupach.

from scipy import stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from scipy.stats import chi2_contingency

# Test T-studenta
group1 = df[df['employment_status'] == 'employed']['income']
group2 = df[df['employment_status'] == 'retired']['income']

t_stat, p_value = stats.ttest_ind(group1, group2)

print(f"T-statystyka: {t_stat}")
print(f"P-wartość: {p_value}")
if p_value < 0.05:
    print("Istnieje istotna różnica w średnich income między grupą zatrudnionych a emerytami.")
else:
    print("Brak istotnej różnicy w średnich income między grupą zatrudnionych a emerytami.")


# Test ANOVA
model = ols('outcome ~ employment_status', data=df).fit()
anova_table = anova_lm(model)

print(anova_table)

if anova_table['PR(>F)'][0] < 0.05:
    print("Istnieje istotna różnica w średnich outcome między przynajmniej dwoma grupami employment_status. (Pracujący zarabiają więcej niż emeryci)")
else:
    print("Brak istotnej różnicy w średnich outcome między przynajmniej dwoma grupami employment_status.")

# Test Chi-kwadrat
contingency_table = pd.crosstab(df['employment_status'], df['city'])

chi2_stat, p_val, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-kwadrat statystyka: {chi2_stat}")
print(f"P-wartość: {p_val}")

if p_val < 0.05:
    print("Istnieje istotna zależność między employment_status a city. (np. w Warszawie jest więcej bezrobotnych)")
else:
    print("Brak istotnej zależności między employment_status a city.")
```

```
0.0s                                                                    Python
T-statystyka: 2.475167674044197
P-wartość: 0.017892150211199005
Istnieje istotna różnica w średnich income między grupą zatrudnionych a emerytami.
                     df         sum_sq       mean_sq          F    PR(>F)
employment_status   3.0  5.985035e+06  1.995012e+06   3.656684  0.015169
Residual           96.0  5.237563e+07  5.455795e+05        NaN       NaN
Istnieje istotna różnica w średnich outcome między przynajmniej dwoma grupami employment_status. (Pracujący zarabiają więcej niż emeryci)
Chi-kwadrat statystyka: 108.93872127301096
P-wartość: 2.9199958121951688e-09
Istnieje istotna zależność między employment_status a city. (np. w Warszawie jest więcej bezrobotnych)
```

Rys. 13. Kod źródłowy i efekt jego wykonania.

Ten kod przeprowadza różne testy statystyczne, aby zidentyfikować istotne różnice w danych:

Test T-studenta sprawdza, czy istnieje istotna różnica w średnich dochodach między grupą zatrudnionych a emerytami. Wynik pokazuje, że różnice są istotne, ponieważ p-wartość jest poniżej 0.05.

Test ANOVA analizuje, czy różnice w średnich wydatkach są istotne pomiędzy przynajmniej dwoma grupami zatrudnienia. W tym przypadku różnice są istotne, a Pracujący mają wyższe średnie dochody niż emeryci.

Test Chi-kwadrat sprawdza zależność między statusem zatrudnienia a miastem, ujawniając, że w Warszawie jest więcej bezrobotnych.

Każdy test dostarcza kluczowych informacji na temat struktury danych, co pomaga w głębszym zrozumieniu ich dystrybucji oraz związków.

3. Wnioski

Istotne różnice w dochodach i wydatkach między grupami zatrudnienia: Test T-studenta oraz ANOVA pokazują, że istnieje znacząca różnica w dochodach między pracującymi a emerytami, a także w średnich wydatkach pomiędzy tymi grupami. Pracujący mają wyższe dochody i większe wydatki niż emeryci.

Zależność między miastem a statusem zatrudnienia: Test Chi-kwadrat wskazuje na istotną zależność między employment_status a miastem, co może sugerować, że w większych miastach, takich jak Warszawa, więcej osób jest bezrobotnych.

Wpływ liczby dzieci i oszczędności na dochody: Analiza wykazała, że liczba dzieci i oszczędności wpływają na dochody, co może wskazywać na różnice w zarządzaniu finansami w zależności od demografii i sytuacji rodzinnej.