Raport z Analizy Danych – Trendy zakupowe

Hubert Kryściak

**1. Wprowadzenie**

Celem analizy było przeprowadzenie kompleksowego badania danych zakupowych klientów. Zadanie obejmowało wstępne przetwarzanie danych, analizę wizualną i statystyczną oraz budowę modelu predykcyjnego do przewidywania kwoty zakupu na podstawie dostępnych cech.

**2. Przegląd Danych**

Dane składają się z 3900 rekordów i 19 kolumn, w tym:

* **Customer ID**: Identyfikator klienta
* **Age**: Wiek klienta
* **Purchase Amount (USD)**: Kwota zakupu w USD
* **Category**: Kategoria produktu
* Inne kolumny, takie jak lokalizacja, rozmiar produktu, sezon i metoda płatności.

**3. Wstępne Przetwarzanie Danych**

Dane zostały załadowane i wstępnie przetworzone:

* Kolumny numeryczne, takie jak *Purchase Amount (USD)*, zostały skonwertowane na odpowiedni format liczbowy.
* Dane były kompletne, bez brakujących wartości, co uprościło analizę.
* Kolumna *Review Rating* została również przekształcona na format liczbowy.

**4. Analiza Wizualna**

Wykonano szereg wizualizacji, aby zrozumieć strukturę danych:

* Histogram kwot zakupu wykazał, że większość zakupów dotyczy niskich wartości z kilkoma wysokimi kwotami.
* Popularność kategorii produktów różniła się, wskazując na preferencje klientów.
* Średnie kwoty zakupu **delikatnie** różniły się w zależności od sezonu, co może mieć znaczenie dla strategii marketingowych.

**5. Analiza Statystyczna**

Przeprowadzono następujące testy statystyczne:

* **Test różnic w wydatkach między płciami**: Brak istotnych statystycznie różnic (*p = 0.378*).
* **Korelacja między wiekiem a kwotą zakupu**: Bardzo słaba korelacja (-0.01), wskazująca na brak związku.

**6. Model Predykcyjny**

Zbudowano model regresji liniowej do przewidywania kwoty zakupu na podstawie wieku i liczby wcześniejszych zakupów:

* **Średni Błąd Absolutny (MAE)**: 20.75 USD
* **R-squared**: -0.007, wskazujący na bardzo niską jakość modelu.

Wyniki sugerują, że wybrane cechy są niewystarczające do skutecznego modelowania.

**7. Wnioski**

Wyniki analizy wykazały, że dane nie zawierają wystarczającej ilości informacji umożliwiającej skuteczne przewidywanie kwot zakupu. Kluczowe obserwacje to:

* Brak znaczących różnic w wydatkach między płciami.
* Brak korelacji między wiekiem a kwotą zakupu.
* Model regresji liniowej nie jest odpowiedni do przewidywania kwot zakupu.

**Co można zrobić lepiej w przyszłości**:

* Rozważyć dodanie nowych cech, takich jak rodzaj produktu, promocje czy lokalizacja.
* Przetestować bardziej zaawansowane modele nieliniowe, takie jak drzewa decyzyjne, XGBoost czy modele sieci neuronowych.

KOD ŹRÓDŁOWY:

import pandas as pd  
import matplotlib.pyplot as plt  
from scipy.stats import ttest\_ind  
from sklearn.model\_selection import train\_test\_split  
from sklearn.linear\_model import LinearRegression  
from sklearn.metrics import mean\_absolute\_error, r2\_score

data = pd.read\_csv('shopping\_trends.csv')  
  
# zmiana na wartości liczbowe  
data['Purchase Amount (USD)'] = pd.to\_numeric(data['Purchase Amount (USD)'], errors='coerce')  
data['Review Rating'] = pd.to\_numeric(data['Review Rating'], errors='coerce')

plt.figure(figsize=(10, 6))  
plt.hist(data['Purchase Amount (USD)'], bins=30, edgecolor='k', alpha=0.7)  
plt.title('Rozkład kwot zakupu (USD)')  
plt.xlabel('Kwota zakupu (USD)')  
plt.ylabel('Częstotliwość')  
plt.grid(axis='y')  
plt.show()  
  
# Popularność kategorii produktów  
category\_counts = data['Category'].value\_counts()  
plt.figure(figsize=(10, 6))  
category\_counts.plot(kind='bar', edgecolor='k', alpha=0.7)  
plt.title('Popularność kategorii produktów')  
plt.xlabel('Kategoria')  
plt.ylabel('Liczba zakupów')  
plt.grid(axis='y')  
plt.show()  
  
# Średnia ilość zakupów per sezon  
avg\_purchase\_by\_season = data.groupby('Season')['Purchase Amount (USD)'].mean()  
plt.figure(figsize=(10, 6))  
avg\_purchase\_by\_season.plot(kind='bar', edgecolor='k', alpha=0.7)  
plt.title('Średnia kwota zakupu w zależności od sezonu')  
plt.xlabel('Sezon')  
plt.ylabel('Średnia kwota zakupu (USD)')  
plt.grid(axis='y')  
plt.show()

# Test różnicy w kwocie zakupu ze względu na płeć  
male\_purchases = data[data['Gender'] == 'Male']['Purchase Amount (USD)']  
female\_purchases = data[data['Gender'] == 'Female']['Purchase Amount (USD)']  
  
t\_stat\_gender, p\_value\_gender = ttest\_ind(male\_purchases, female\_purchases, equal\_var=False)  
t\_stat\_gender, p\_value\_gender  
  
# Korelacja pomiędzy wiekiem a kwotą zakupu  
correlation\_age\_purchase = data[['Age', 'Purchase Amount (USD)']].corr()  
correlation\_age\_purchase

# Przygotowanie danych do regresji  
features = ['Age', 'Previous Purchases']  
target = 'Purchase Amount (USD)'  
  
X = data[features]  
y = data[target]  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
model = LinearRegression()  
model.fit(X\_train, y\_train)  
y\_pred = model.predict(X\_test)  
  
mae = mean\_absolute\_error(y\_test, y\_pred)  
r2 = r2\_score(y\_test, y\_pred)  
  
mae, r2