

Instytut Informatyki

Studia Podyplomowe

Wizualna analityka danych

praca KOŃCOWA

Sebastian Gumula

Amazon w Liczbach: Analiza Trendów Sprzedaży i Profilu Klienta w Branży E-commerce

Opiekun pracy

Tytuł, imię i nazwisko opiekuna

Warszawa, 2024

STRESZCZENIE

Praca ta koncentruje się na wszechstronnej analizie danych w kontekście branży e-commerce, ze szczególnym uwzględnieniem klientów, popytu na konkretne typy produktów, analizy sentymentu, dekompozycji szeregów czasowych oraz tworzenia na jej podstawie modelu prognostycznego ARIMA.

Dogłębna analiza zachowań klientów oraz trendów sprzedażowych, a także badanie popytu na konkretne produkty umożliwiają lepsze zrozumienie preferencji konsumentów.   
Analiza sentymentu pozwala na ocenę opinii klientów, podczas gdy dekompozycja szeregów czasowych dostarcza wglądu w sezonowe i trendowe zmiany występujące w danych.  
Wreszcie, model prognostyczny umożliwia prognozowanie przyszłych wyników sprzedaży, co jest kluczowe dla opracowywania skutecznych strategii biznesowych.

Praca ta skupia się na kompleksowym podejściu do analizy danych e-commerce, kładąc nacisk na zrozumienie klientów, dynamikę popytu oraz efektywne prognozowanie przyszłych trendów.

Słowa kluczowe: Analiza Danych, Analiza Sentymentu, Prognozowanie, ARIMA

THESIS TITLE IN ENGLISH

Amazon in Numbers: Analyzing Sales Trends and Customer Profile in the E-commerce Industry.

Summary in English.

This thesis focuses on comprehensive data analysis in context of e-commerce industry, with a particular focus on customers, demand for specific types of products, sentiment analysis, time series decomposition and the creation of an ARIMA predictive model.

An in-depth analysis of customer behavior and sales trends, as well as the study of demand for specific products, enables a better understanding of consumer preferences.

Sentiment analysis allows assessment of customer opinions, while time series decomposition provides insight into seasonal and trend changes occurring in the data.

Finally, a predictive model enables forecasting of future sales performance, which is crucial for developing effective business strategies.

This work focuses on a comprehensive approach to analyzing e-commerce data, emphasizing understanding customers, demand dynamics and effectively forecasting future trends.

Keywords: Data Analysis, Sentiment Analysis, Time series Forecasting, ARIMA

SPIS TREŚCI

1. Sformułowanie problemu biznesowego ………………………………………… 4

2. Eksploracyjna analiza danych ………………………………………………….. 5

3. Przetwarzanie danych ……………………………..…………….………………. 8

4. Przetwarzanie danych do budowy modelu prognostycznego ……..…..….….. 10

5. Wizualizacja danych, tworzenie narracji i wnioski końcowe ..…..…..….….. 16

Literatura ……………………………………………………………………………. 19

# 1. Sformułowanie problemu biznezowego

Firma Amazon, będąca liderem w branży e-commerce, stawia sobie za cel zwiększenie efektywności operacyjnej i wzrostu przychodów.

W tym celu proponuję przeprowadzenie zaawansowanej analizy danych, skupiając się na analizie klienta, popycie na konkretne produkty, analizie sentymentu, dekompozycji szeregów czasowych oraz implementacji modelu prognostycznego ARIMA.

Cele:

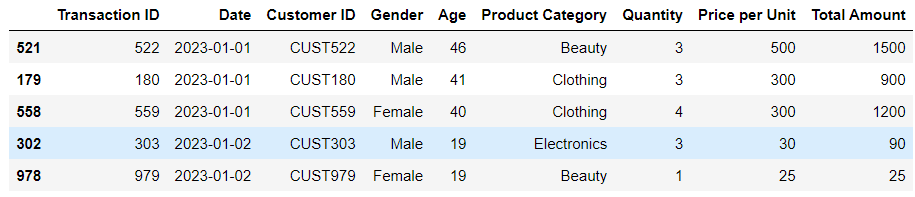
1. **Zrozumienie klientów:** Analiza zachowań klientów pozwoli nam lepiej zrozumieć ich preferencje, co umożliwi dostosowanie oferty do oczekiwań konsumentów.
2. **Optymalizacja popytu**: Badanie popytu na konkretne produkty pozwoli na identyfikację najbardziej atrakcyjnych produktów, co skutkować będzie lepszym zarządzaniem dostawą i zapasami.
3. **Ewaluacja sentymentu**: Analiza sentymentu opinii klientów pomoże w identyfikacji obszarów wymagających poprawy oraz w budowaniu pozytywnego wizerunku marki.
4. **Dekompozycja szeregów czasowych**: Zastosowanie dekompozycji szeregów czasowych pozwoli na zidentyfikowanie sezonowych i trendowych zmian, co usprawni planowanie marketingowe.
5. **Prognozowanie przyszłych trendów:** Implementacja modelu prognostycznego ARIMA umożliwi precyzyjne prognozowanie przyszłych trendów sprzedażowych, co jest kluczowe dla skutecznej strategii biznesowej.

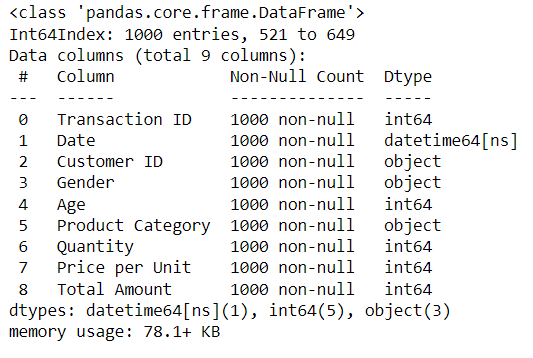
# 2. Eksploracyjna analiza danych

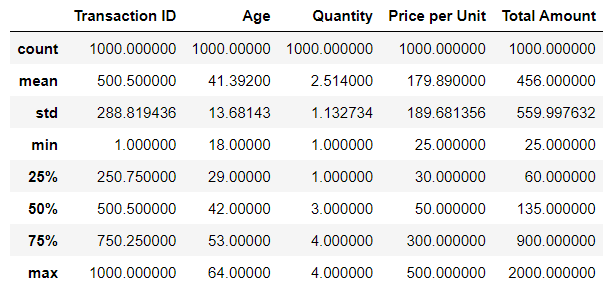
Celem realizacji przedstawionych założeń biznesowych, przeanalizowane zostały następujące zbiory danych:

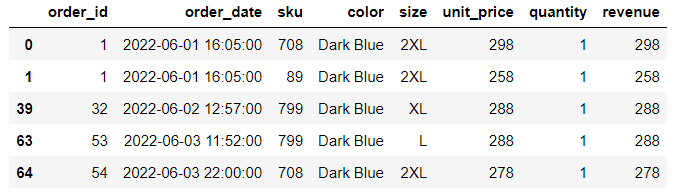
* retail\_sales\_dataset.csv
* women\_clothing\_ecommerce\_sales.csv
* reviews.csv

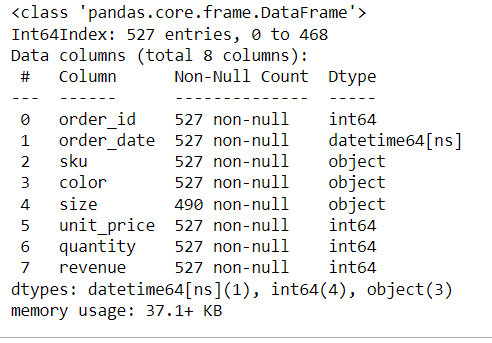
Poniżej przedstawiona została eksploracyjna analiza danych dla każdego datasetu:

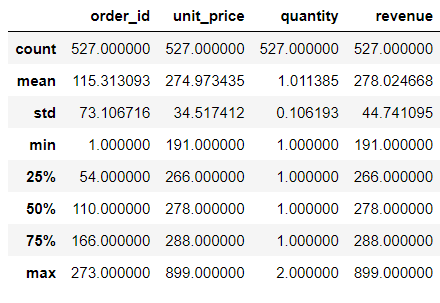


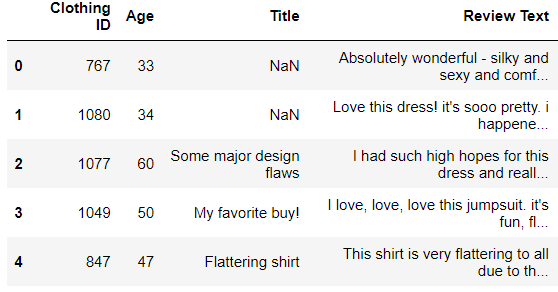


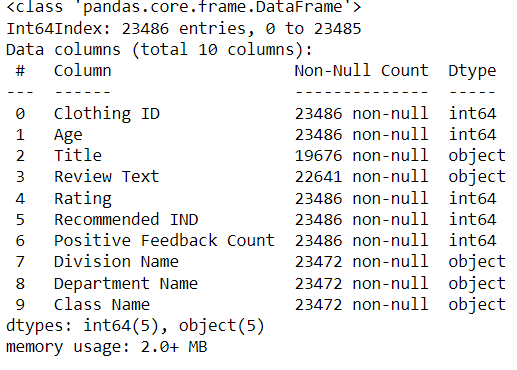


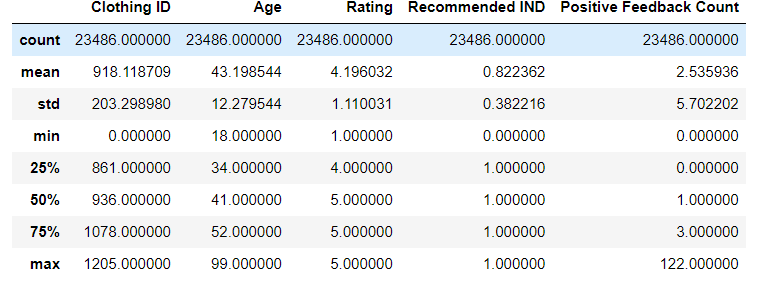






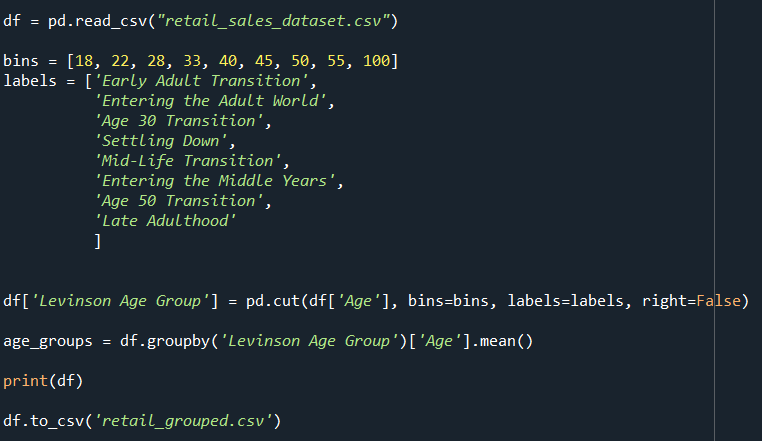






# 3. Przetwarzanie danych

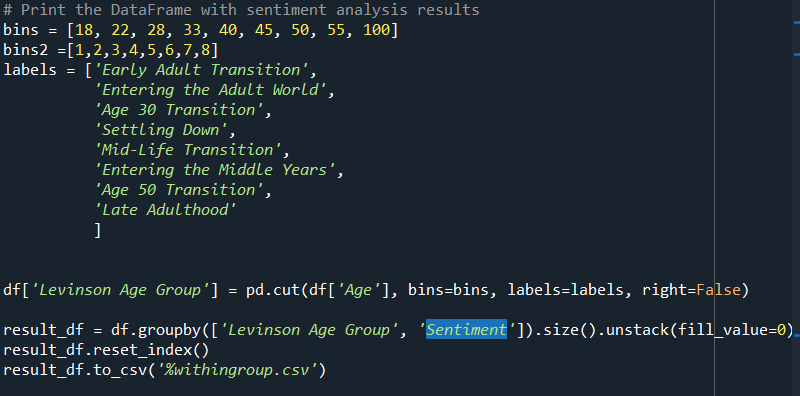
Celem stworzenia wartościowych wizualizacji oraz modeli, zbiory danych zostały odpowiednio przetworzone:



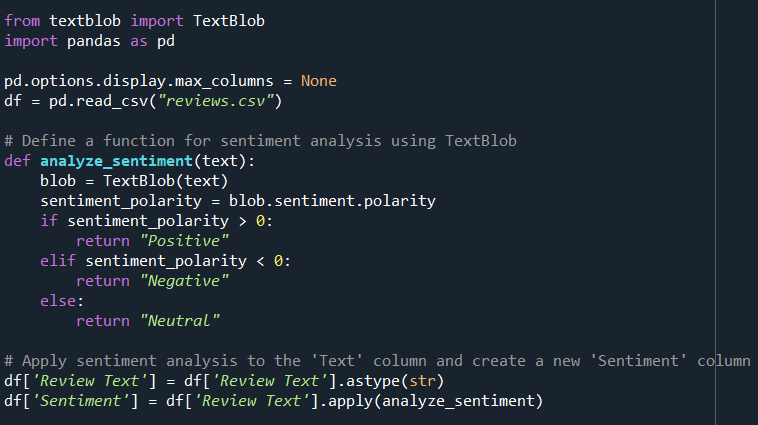
Do ‘retail\_sales\_dataset.csv’ dodana została nowa kolumna   
[‘Levinson Age Group’], celem przypisania klientów do określonych grup wiekowych, przedstawiających określone fazy życia człowieka [1].

Analiza sprzedaży w kontekście faz życiowych Levinsona może dostarczyć cennych informacji na temat preferencji zakupowych, potrzeb i oczekiwań klientów w różnych etapach ich życia dorosłego.

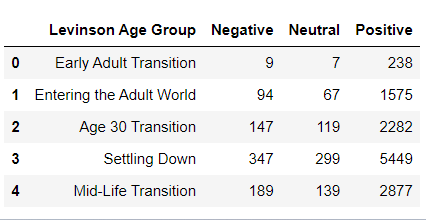
Działanie w zgodzie z tymi oczekiwaniami może przyczynić się do lepszego dostosowania strategii marketingowej oraz efektywniejszej sprzedaży.



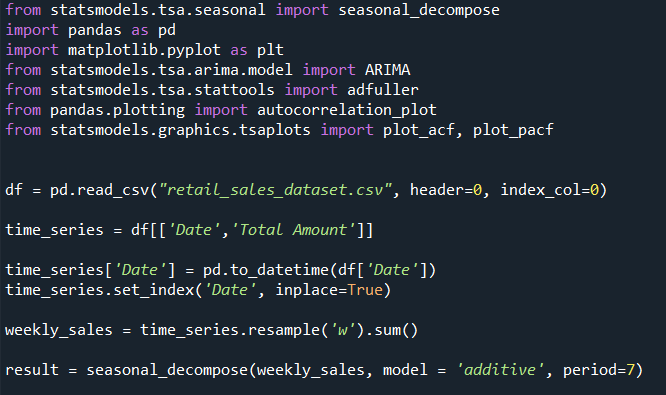
Podobnemu zabiegowi poddany został zestaw danych, przeznaczony do analizy sentymentu. Dodatkowo, wyznaczona została nowa kolumna [‘Sentiment’] zawierająca przetworzone za pomocą modelu NLP flagi *Positive, Neutral, Negative* uzyskane dzięki interpretacji tekstu opinii na temat określonego produktu:

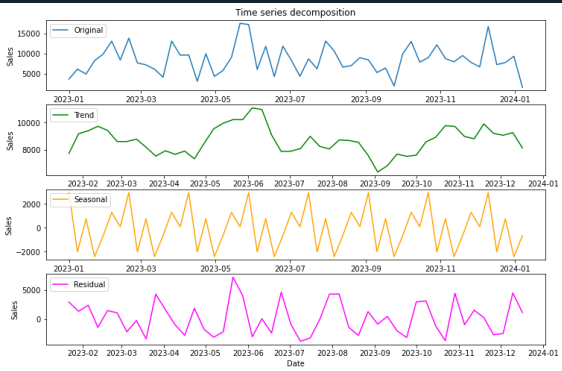


Na koniec wyznaczona została metryka, prezentująca udział poszczególnych sentymentów w każdej grupie wiekowej.



# 4. Przetwarzanie danych do budowy modelu prognostycznego



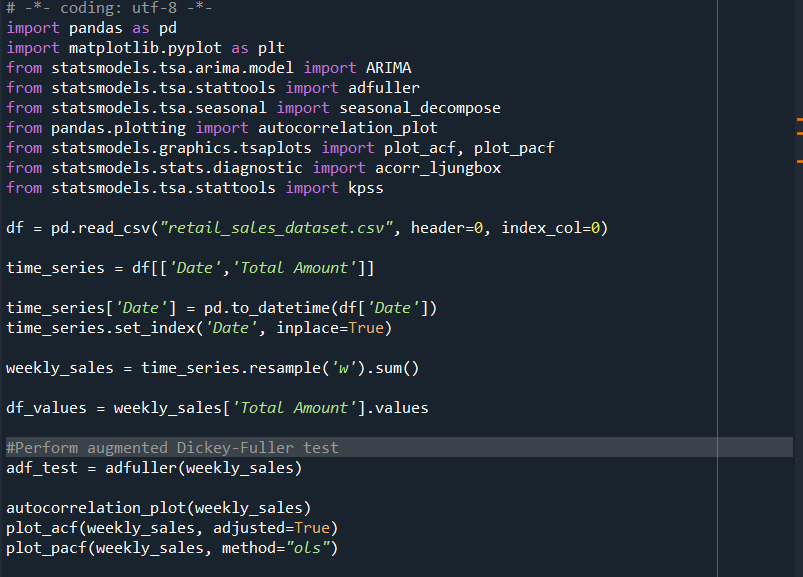


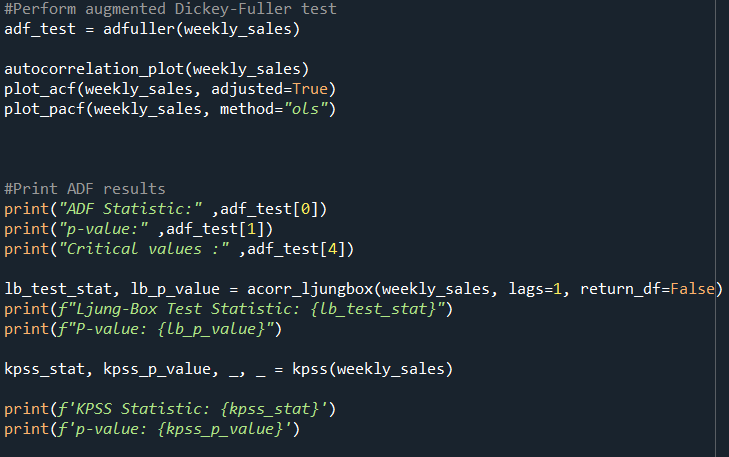
Aby stworzyć model prognostyczny, zgodnie z metodologią   
Boxa-Jenkinsa, badany szereg czasowy należy poddać procesom identyfikacji, estymacji oraz diagnostyki [2].

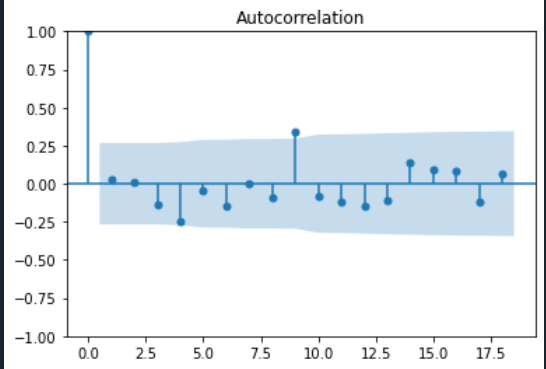
Proces identyfikacji ma na celu zbadanie zjawisk występujących w danym szeregu czasowym, tj. trend, sezonowość, zależności matematyczne między danymi (model addytywny i multiplikatywny) oraz stacjonarność.

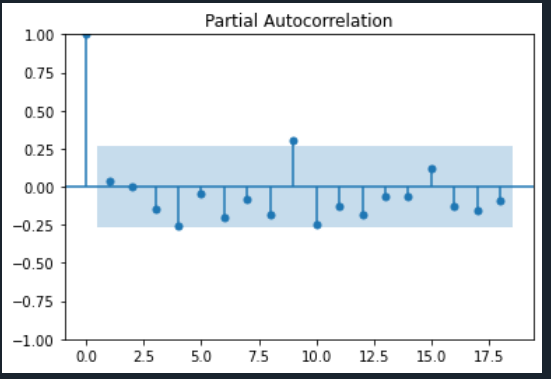
Do badania stacjonarności, można posłużyć się wykresami funkcji autokorelacji (ACF) oraz częściowej autokorelacji (PACF) , jak również testami statystycznymi, tj. rozszerzony test Dickeya – Fullera (ADF).

Polega on na badaniu hipotezy zerowej, według której każdy niestacjonarny proces zawiera pierwiastek jednostkowy, leżący w obrębie jego koła jednostkowego. W przypadku jego braku rozważana jest hipoteza alternatywna, według której wartość testu statystycznego szeregu stacjonarnego nie przekracza wartości krytycznych, znajdujących się w tablicach statystycznych testu ADF [3].



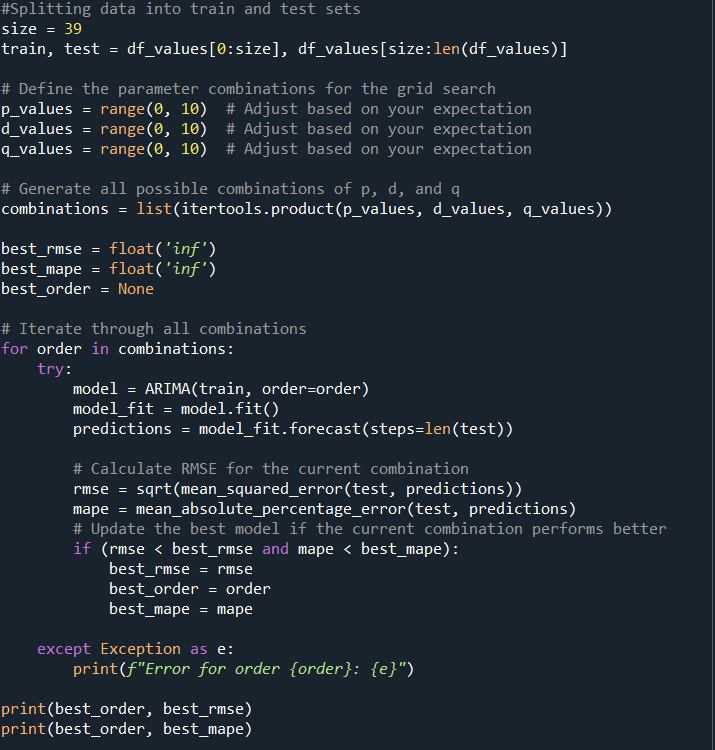


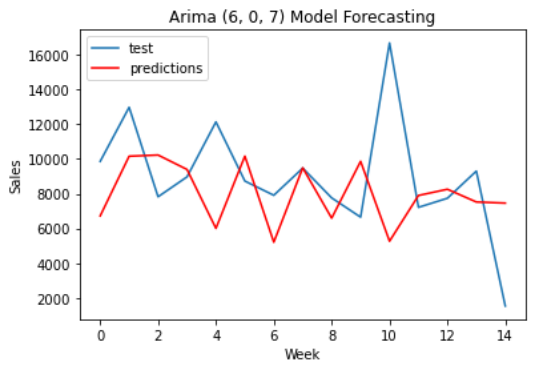


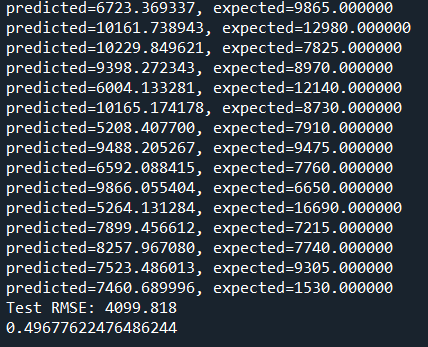


Na podstawie odczytu wartości p, q z wykresów ACF/PACF oraz optymalizacji hiperparametrów poprzez minimalizację metryk RSME oraz MAPE za pomocą algorytmu *grid search* wyznaczona została optymalna wartość parametrów p,d,q dla badanego szeregu czasowego. Ostatecznie zaproponowany został model ARIMA(6, 0, 7).





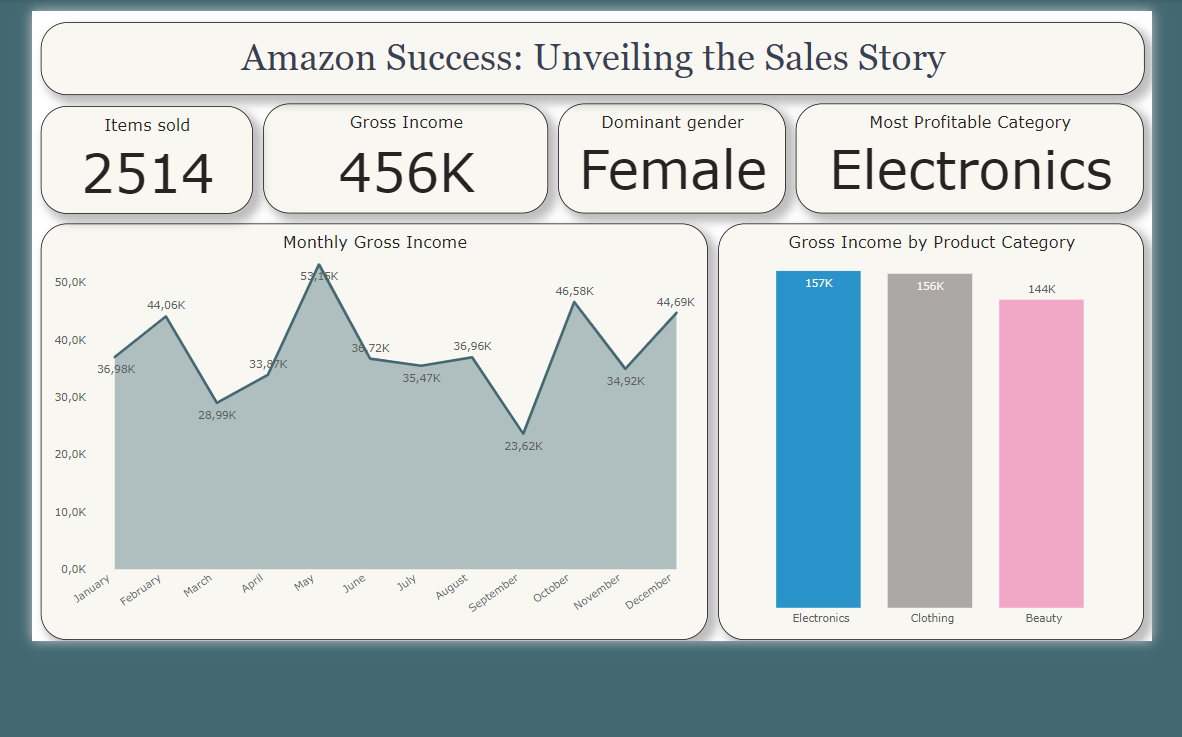




Pomimo zastosowania metodologii Boxa-Jenkinsa i dostrojenia hiperparametrów uzyskano model prognostyczny o stosunkowo wysokiej wartości błędu RMSE (4099.18) i MAPE (49.6%). Jest to wynik niezadowalający, dlatego model należy odrzucić.

# 5. Wizualizacja danych, tworzenie narracji i wnioski końcowe

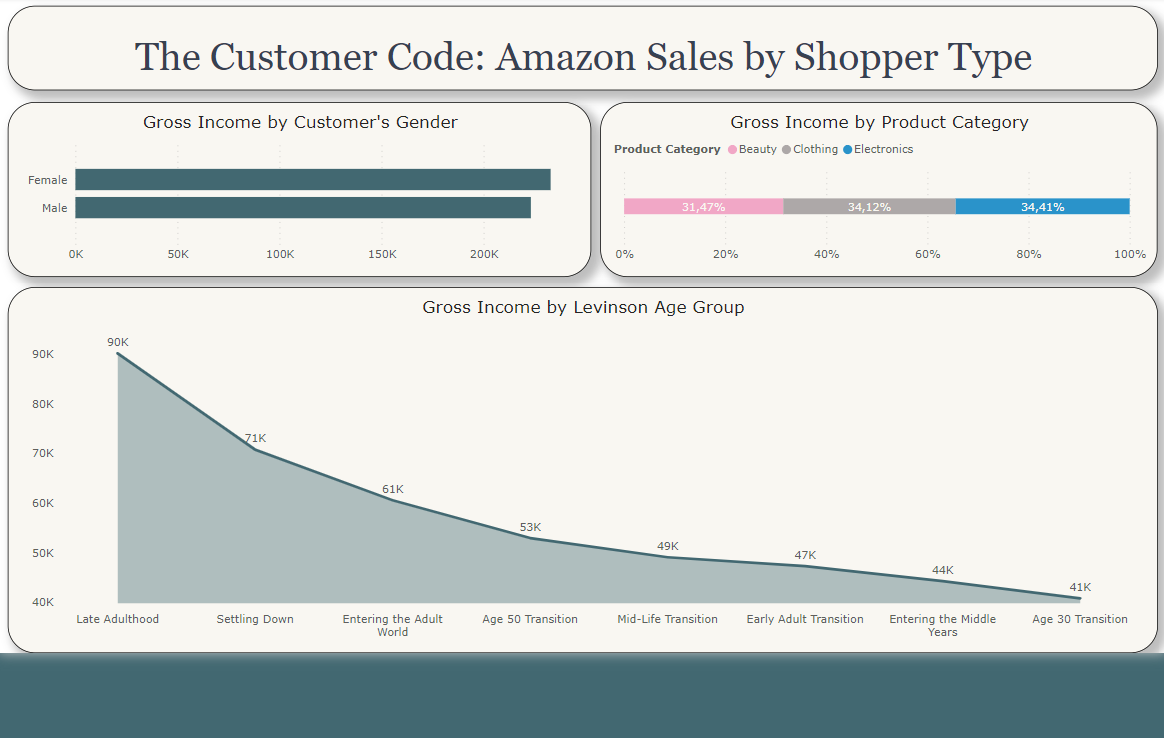
Po ukończonej analizie został stworzony raport, w celu prezentacji wniosków:

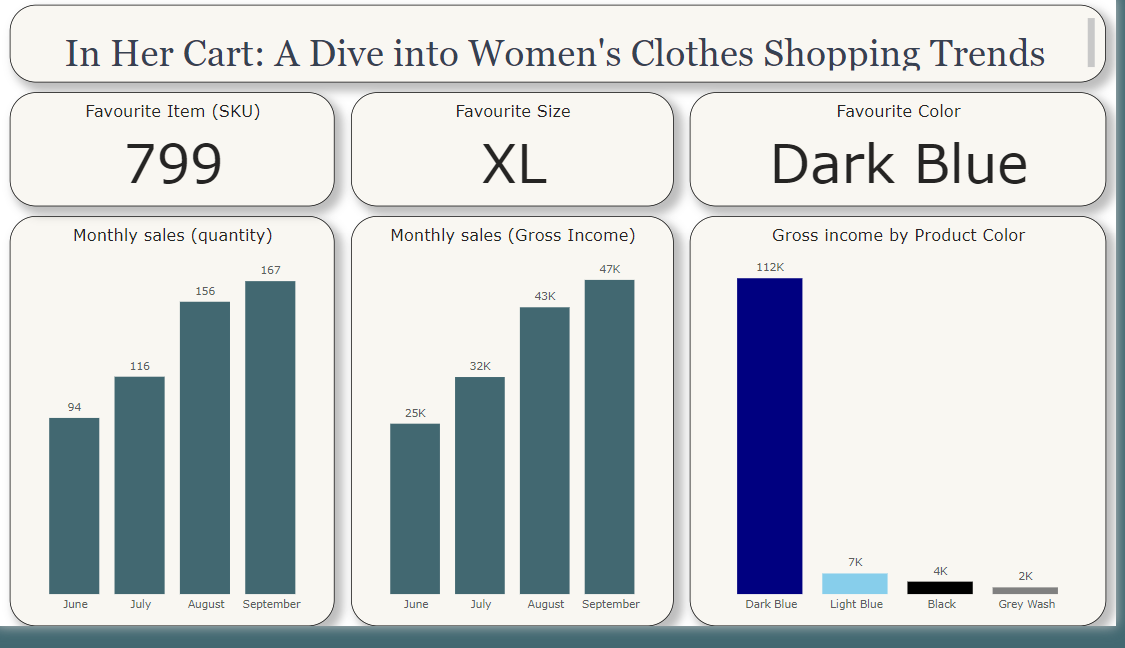


Pierwszy dashboard przedstawia obecny wynik sprzedażowy, wraz z uwzględnieniem najbardziej dochodowych grup produktowych oraz dominującej płci klienta. Celem narracji jest ekspozycja oraz skierowanie uwagi na potrzebę zdefiniowania profilu klienta, aby maksymalizować zyski.

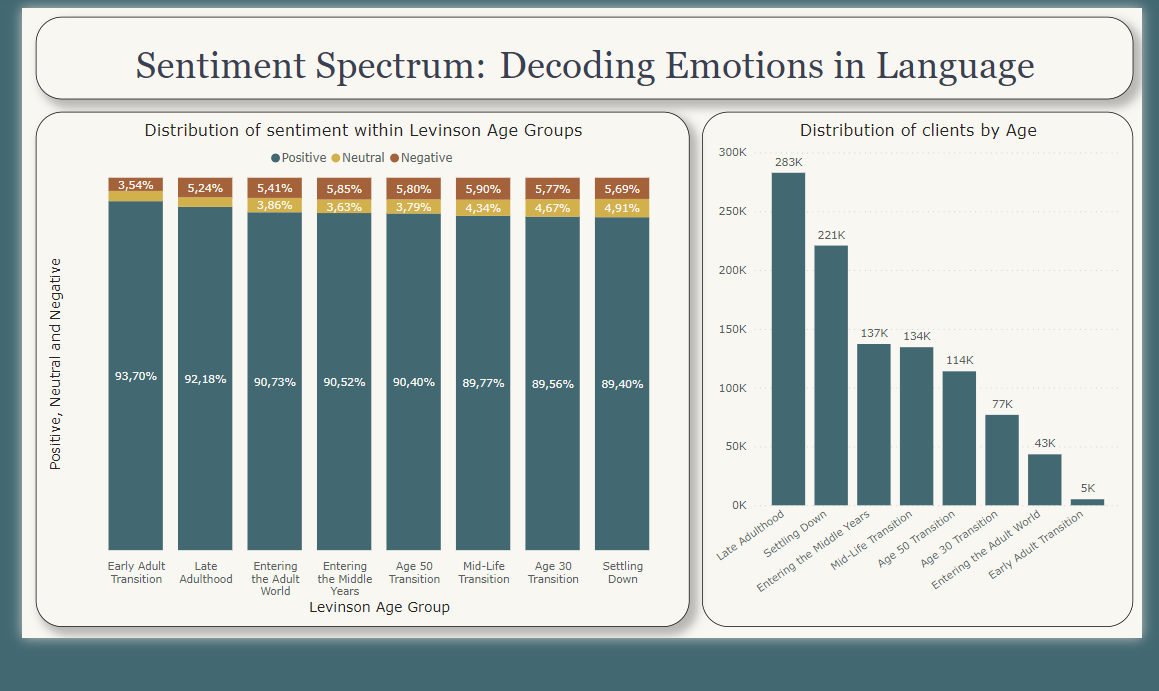
Drugi dashboard przedstawia podział klientów ze względu na ich płeć oraz grupę wiekową. Ma on na celu wstępne utworzenie profili klientów oraz przedstawienie jak bardzo dochodowe są poszczególne grupy.

Celem narracji jest rozbudowa akcji oraz podkreślenie istotności wnikliwej analizy klienta.



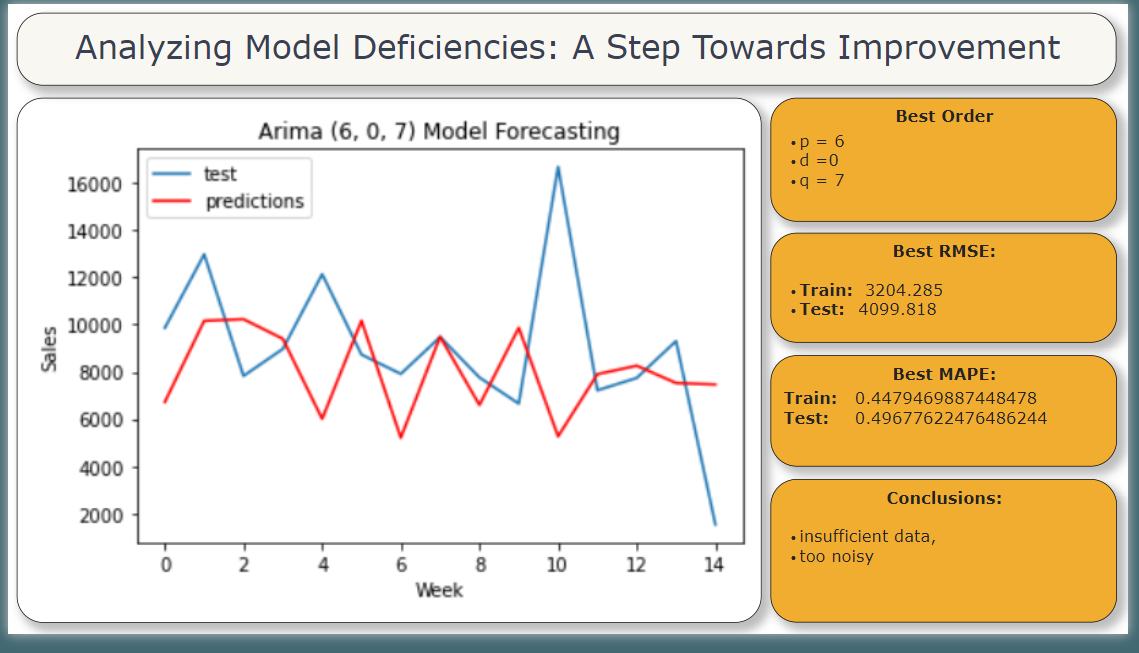


Trzeci dashboard bada głębiej preferencje dominującej grupy klientów (kobiety). Wskazuje na ulubione produkty, ich rozmiar oraz najczęściej wybierany kolor.



Czwarty dashboard pozwala na zbadanie rozkładu ilości klientów oraz procentowego udziału pozytywnych, negatywnych i neutralnych komentarzy na temat produktów, w każdej z rozważanych kategorii wiekowych.

Piąty dashboard przedstawia dekompozycję badanego szeregu czasowego, przedstawiając jego składowe trendu oraz sezonowości, a także wyniki testów ADF, KPSS i Ljung-Box, oraz interpretację wyników.



Ostatni dashboard zawiera walidację *walk-forward* modelu prognostycznego, wyznaczone optymalne hiperparametry oraz wnioski końcowe na temat jakości oraz ilości danych (dane pochodzą jedynie z 1 roku)

Pozytywne wyniki testów ADF oraz KPSS oraz widoczna składowa sezonowa na wykresie dekompozycji mogą stanowić informację o możliwości uzyskania wysokiej dokładności modelu prognostycznego w przypadku zwiększenia ilości i jakości danych.

# Literatura

1. https://www.psychologydiscussion.net/theory/levinsons-theory/levinsons-theory-stages-of-adult-life-human-development-psychology/13594
2. https://otexts.com/fpp2/.
3. <https://coin.wne.uw.edu.pl/pstrawinski/dane_fin/adf03.pdf>