Jakub Zieliński, 299041
Jakub Hebda, 283484

Sprawozdanie IUM

Zadanie 03 wariant 02:

"Wygląda na to, że nasze firmy kurierskie czasami nie radzą sobie z dostawami. Gdybyśmy wiedzieli, ile taka dostawa dla danego zamówienia potrwa – moglibyśmy przekazywać tę informację klientom."

Definicja problemu biznesowego:

Celem biznesowym jest zapewnienie niezawodnego modułu generującego predykcje dotyczące przewidywanego czasu dostawy produktów do użytkowników na podstawie całej historii transakcji zarejestrowanych w systemie. Informacja o przewidywanym terminie dostawy ma pomóc użytkownikom w świadomym zakupie. Chcemy przewidzieć ile dni od zakupu zajmie dostawa.

Analizujemy kontekst w jakim występuje potrzeba biznesowa – interesuje nas:

- jaka jest obecna sytuacja,
- co ma zostać wprowadzone/zmienione,
- jakie właściwości/cechy powinno mieć docelowe rozwiązanie,
- jakie są założenia, oczekiwania, ograniczenia, zasoby.

Kryterium sukcesu biznesowego:

Za sukces możemy uznać trafne predykcje czasu dostawy dla 90% zakupów. Za poprawną predykcję uznajemy dostarczenie produktu do klienta w przewidzianym okresie. Ciężko przewidzieć, czy poprawne predykcje wpłyną na podwyższenie ruchu na stronie. Dobry ruchem biznesowym ze strony klienta byłoby przeprowadzenie ankiety wśród użytkowników.

Zdefiniowanie zadania modelowania:

Po wstępnym przeanalizowaniu zadania zdecydowaliśmy, że najodpowiedniejszym typem zadania jest w naszym przypadku klasyfikacja. Proponowane kategorie dla predykcji modelu to 1 dzień, 2 dni, 3 dni, 4 dni, 5 dni i dłużej

Kryterium sukcesu analitycznego:

Kryterium zakłada osiągnięcie większy niż losowy odsetka trafnych predykcji kategorii czasu dostawy nazwane α. Stosunek liczby poprawnych predykcji do liczby wszystkich dostaw powinien być większy współczynnikowi α.

$\alpha = \sum_{i=1}^{5} P(X_i)^2$, gdzie $P(X_i)$ to: ilość dostaw w klasie “i dni” / wszystkie dostawy

Z analizy danych w pliku data_analysis:

$P(X_1) = 420 / 7031$

$P(X_2) = 2283 / 7031$

$P(X_3) = 3348 / 7031$

$P(X_4) = 923 / 7031$

$P(X_5) = 57 / 7031$

Wiec $\alpha = \sum_{i=1}^{5} P(X_i)^2 = 0,353$

Założenia:

Z braku innych danych, model tworzymy na tych, które otrzymaliśmy.

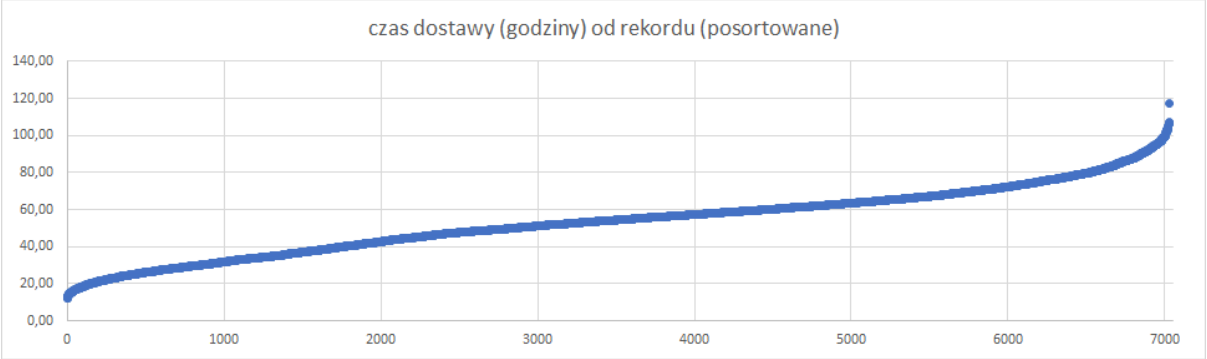
Produkt końcowy:

Wyniki predykcji będą przedstawiana na stronie koszyka przy wyborze firmy kurierskiej w postaci tabelki o nagłówkach: firma kurierska, przewidywany czas dostawy oraz jej koszt.

Analiza danych:

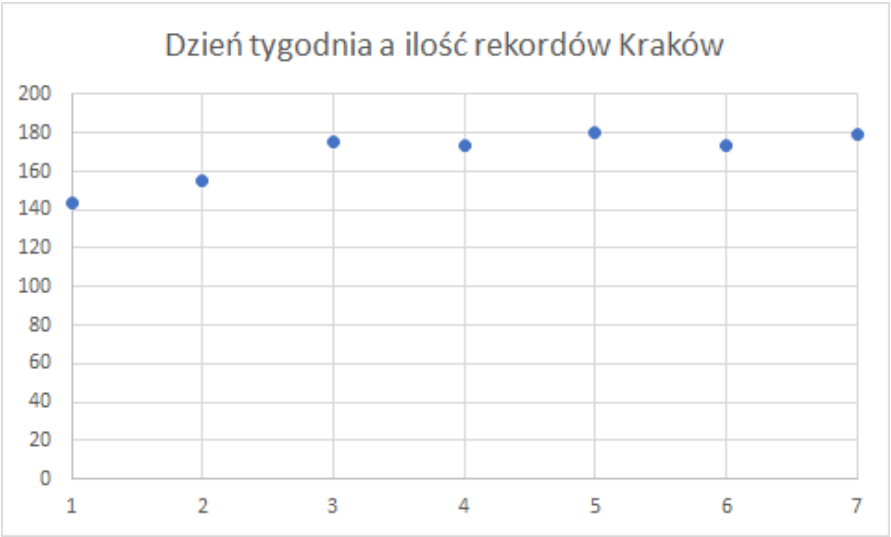
Z pliku sessions.jsonl zostały wybrane te rekordy, które kończą się zakupem. Połączyliśmy je z odpowiadającymi im rekordami z plików purchases.jsonl, products.jsonl i users.jsonl, obliczyliśmy czas dostawy oraz usunęliśmy kolumny, które naszym zdaniem nie noszą dodatkowej informacji. Powstałą tabelę zapisaliśmy do pliku data/processed/sessions.jsonl. Tak przygotowane dane sprawdziliśmy pod kątem anomalii w czasie dostaw, ilości zakupów na dany dzień tygodnia z podziałem na miasta oraz ogólnie:

Czasy dostaw są ciągłe.

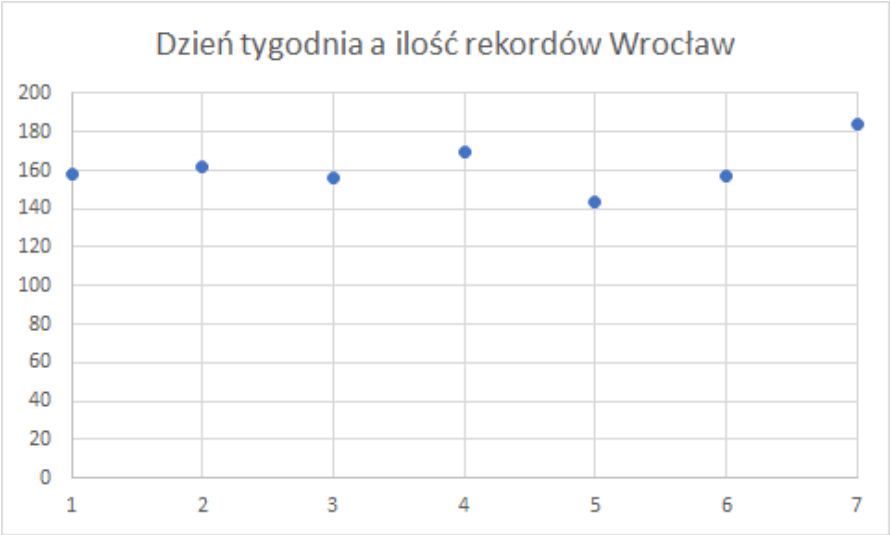


Rozkład ilości zakupów w tygodniu(1 - pon, 2 - wt...) dane są rozłożone równomiernie.

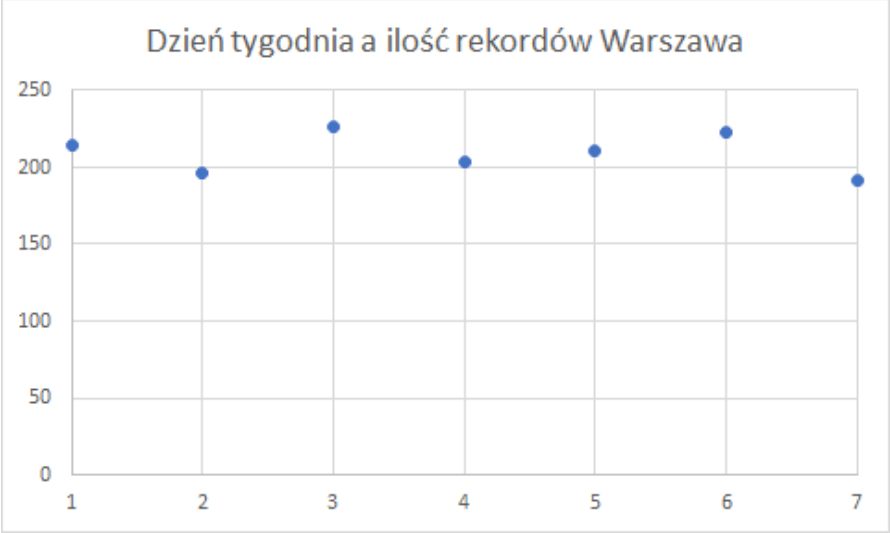
1	974	odchylenie
2	1012	standardowe
3	1042	32,8
4	949	
5	1001	
6	1025	
7	1028	



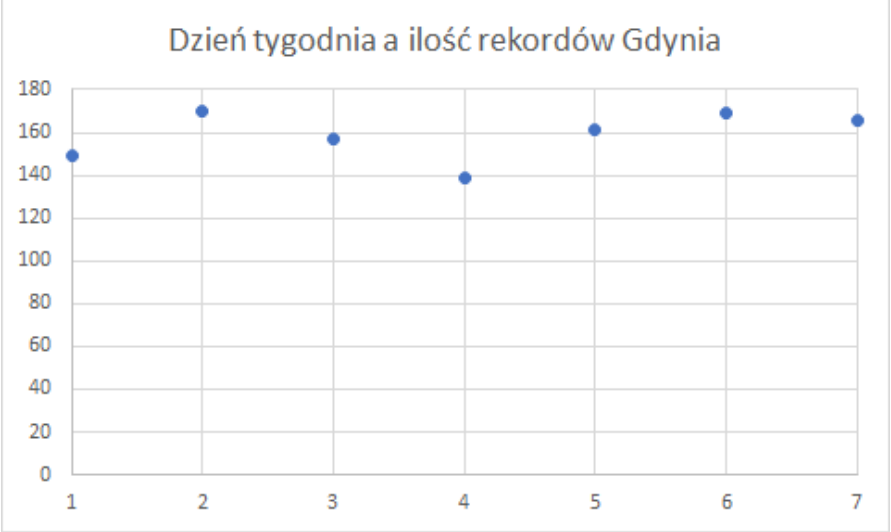
1	144	odchylenie
2	155	standardowe
3	175	13,6
4	173	
5	180	
6	173	
7	179	



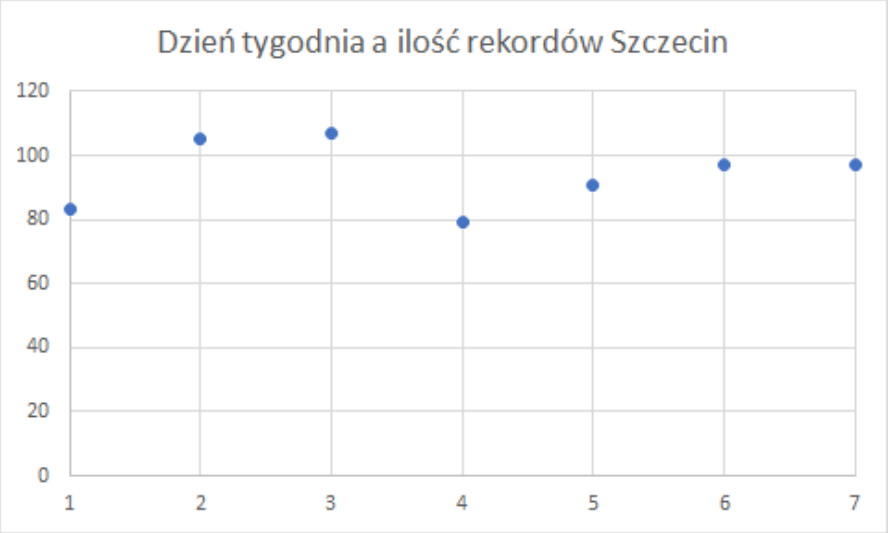
1	158	odchylenie
2	162	standardowe
3	156	12,6
4	170	
5	144	
6	157	
7	184	



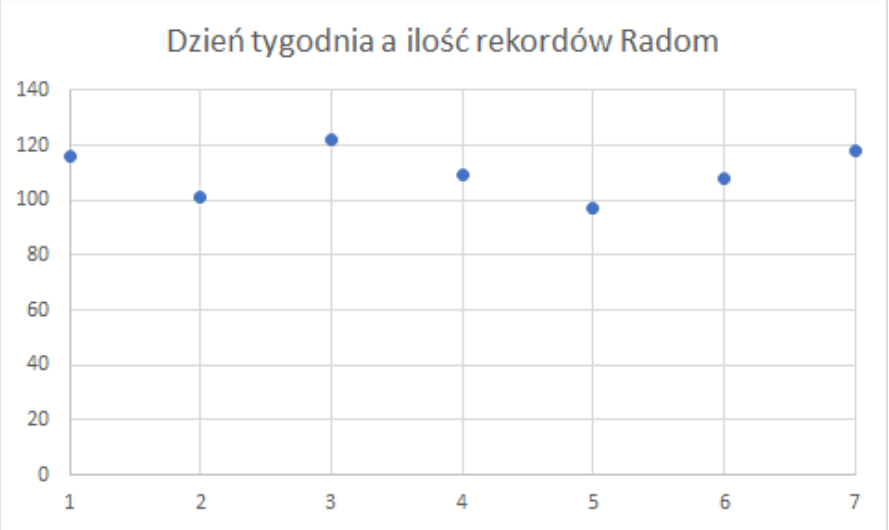
1	215	odchylenie
2	196	standardowe
3	226	13,3
4	203	
5	211	
6	223	
7	191	



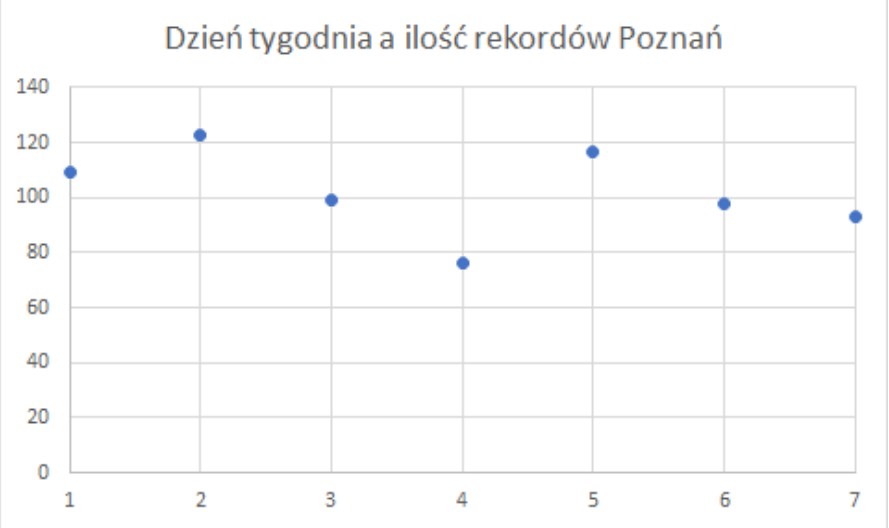
1	149	odchylenie
2	170	standardowe
3	157	11,4
4	139	
5	161	
6	169	
7	166	



1	83	odchylenie
2	105	standardowe
3	107	10,5
4	79	
5	91	
6	97	
7	97	



1	116	odchylenie
2	101	standardowe
3	122	9,1
4	109	
5	97	
6	108	
7	118	



1	109	odchylenie
2	123	standardowe
3	99	15,8
4	76	
5	117	
6	98	
7	93	

Dla żadnego z miast nie zauważyliśmy anomalii w ilości rekordów. Dane wyglądają na poprawne.

Dalsza analiza w pliku data_analysis.