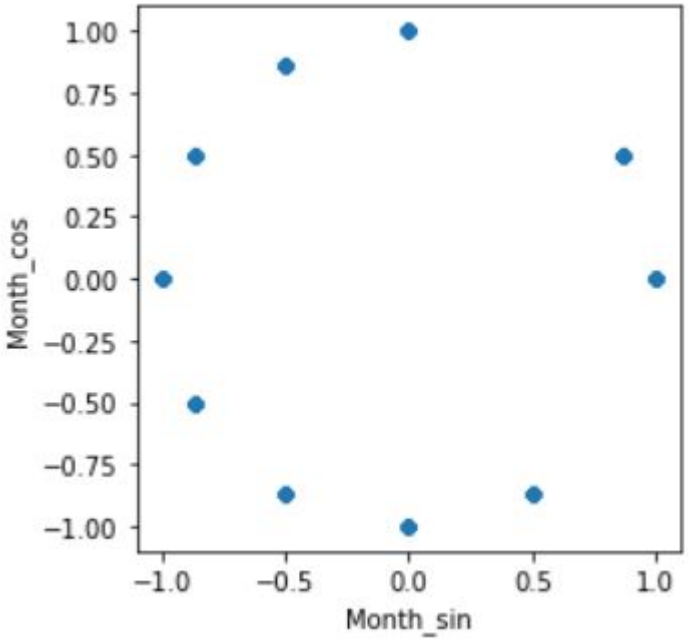
WUM - projekt 2
Online shoppers intentions

Mikołaj Spytek,
Artur Żółkowski

Co się zmieniło od ostatniej prezentacji?

1. Wszystkie z pokazywanych dziś klasteryzacji odbywały się bez zmiennej “Revenue”

Cykliczne kodowanie miesięcy



Ten sposób enkodowania pozwala na zachowanie prawdziwych odległości między miesiącami - grudzień blisko stycznia.

Poprawa wyznaczania metryk

```
stability_scores
```

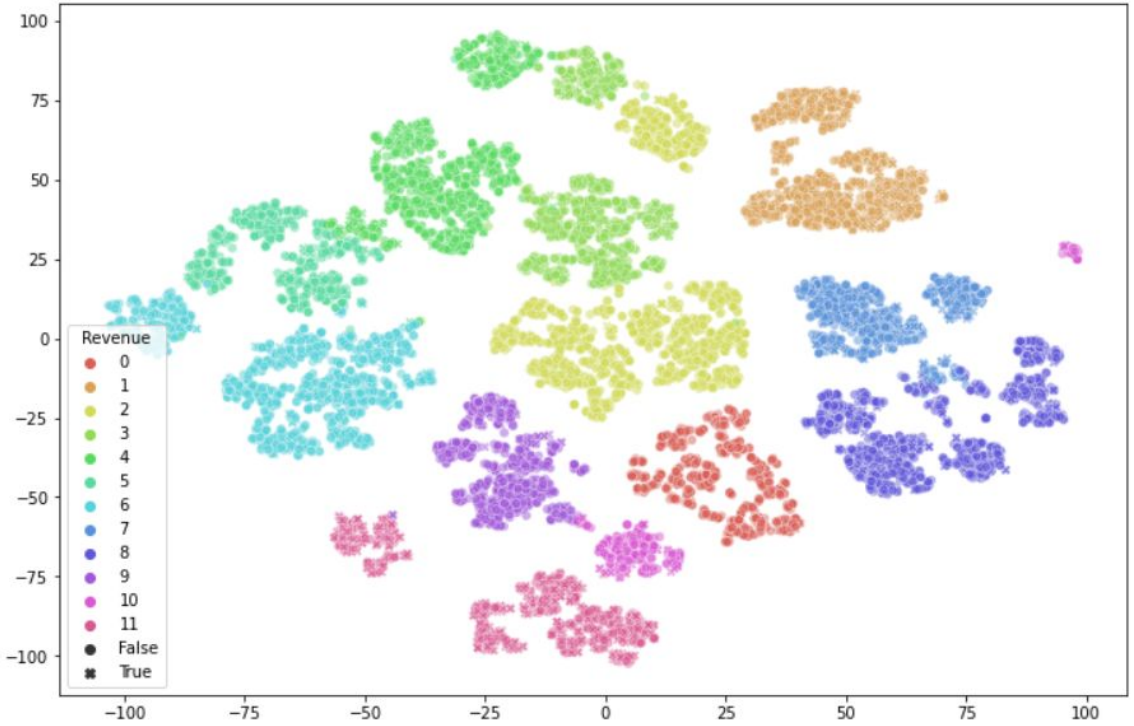
	2	3	4	5	6	7	8	9	10	11	12
KMeans	0.997010	0.998603	0.802918	0.916176	0.964812	0.935656	0.998195	0.992914	0.942443	0.932783	0.977070
Agglomerative - ward linkage	0.859431	0.887994	0.968292	0.923977	0.850819	0.885467	0.902903	0.942487	0.933957	0.930194	0.920582
Agglomerative - single linkage	0.600000	0.503843	0.859522	0.690673	0.644099	0.933815	0.942974	0.945501	0.954348	0.936235	0.906418
GMM - spherical covariance	0.741066	0.743817	0.611240	0.831221	0.730173	0.724903	0.845635	0.893364	0.923431	0.909229	0.900518

Wyznaczyliśmy metryki: silhouette, calinski-harabasz, davies-bouldin oraz stabilność klastrowania

Dla metod klastrowań oraz liczb klastrów przedstawionych w tabelce.

Na podstawie analizy tych czterech metryk wybraliśmy dwa klastrowania do przedstawienia

Klastrowanie 1 (wizualizacja TSNE)

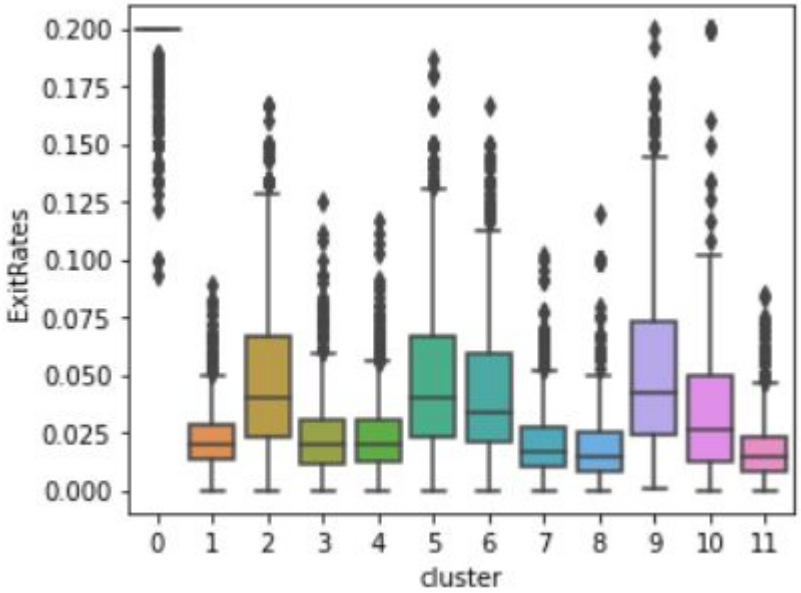


-12 klastrów, metodą K-means

- wszystkie osiągane metryki były wysokie

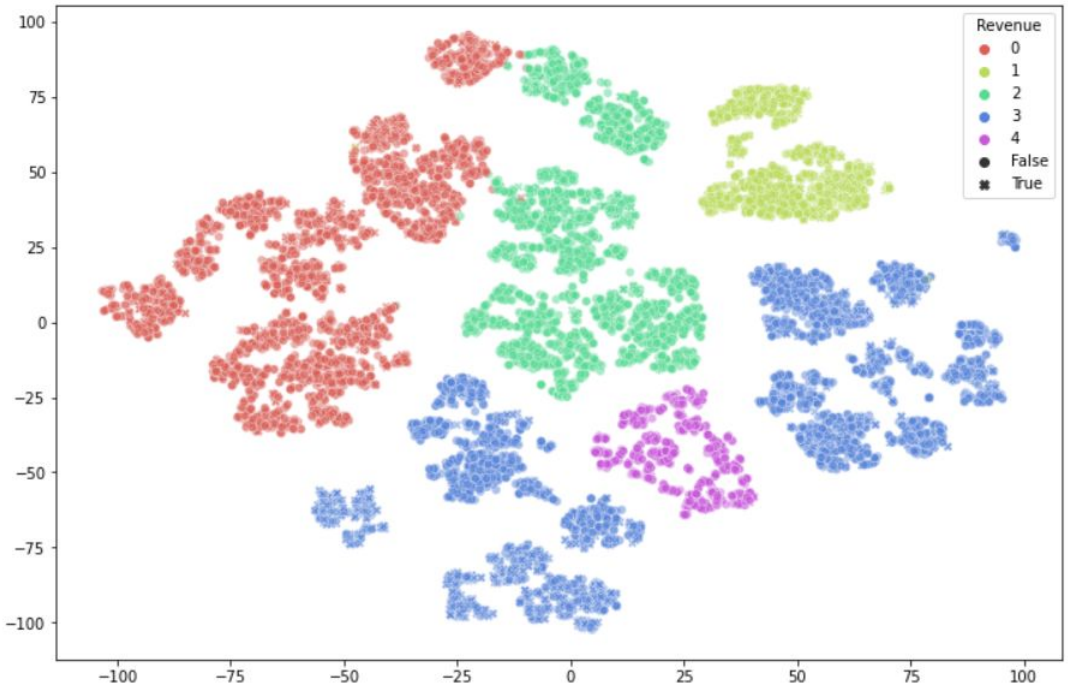
- na wizualizacji klastry widocznie się od siebie różnią

Próba rozpoznania klastrów

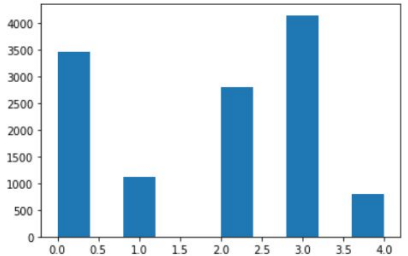


- za pomocą takich boxplotów patrzyliśmy na rozkłady zmiennych
- spróbowaliśmy nazwać poszczególne klastry, ale przy aż 12 nie byliśmy w stanie

Klastrowanie 2 (wizualizacja TSNE)

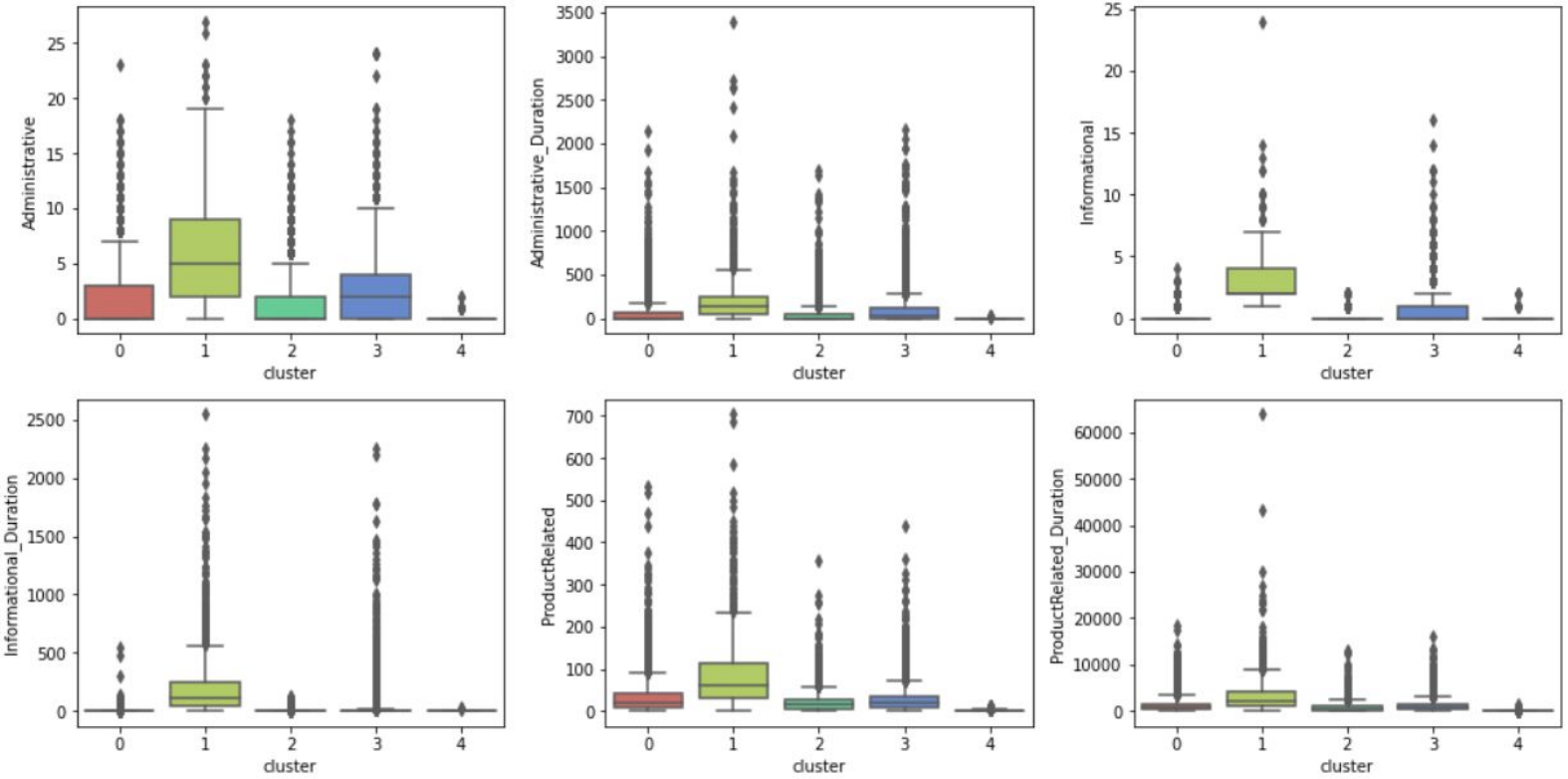


- KMeans, 5 klastrów
- również wysokie metryki (stabilność 0,96)
- klastry odseparowane od siebie na wizualizacji

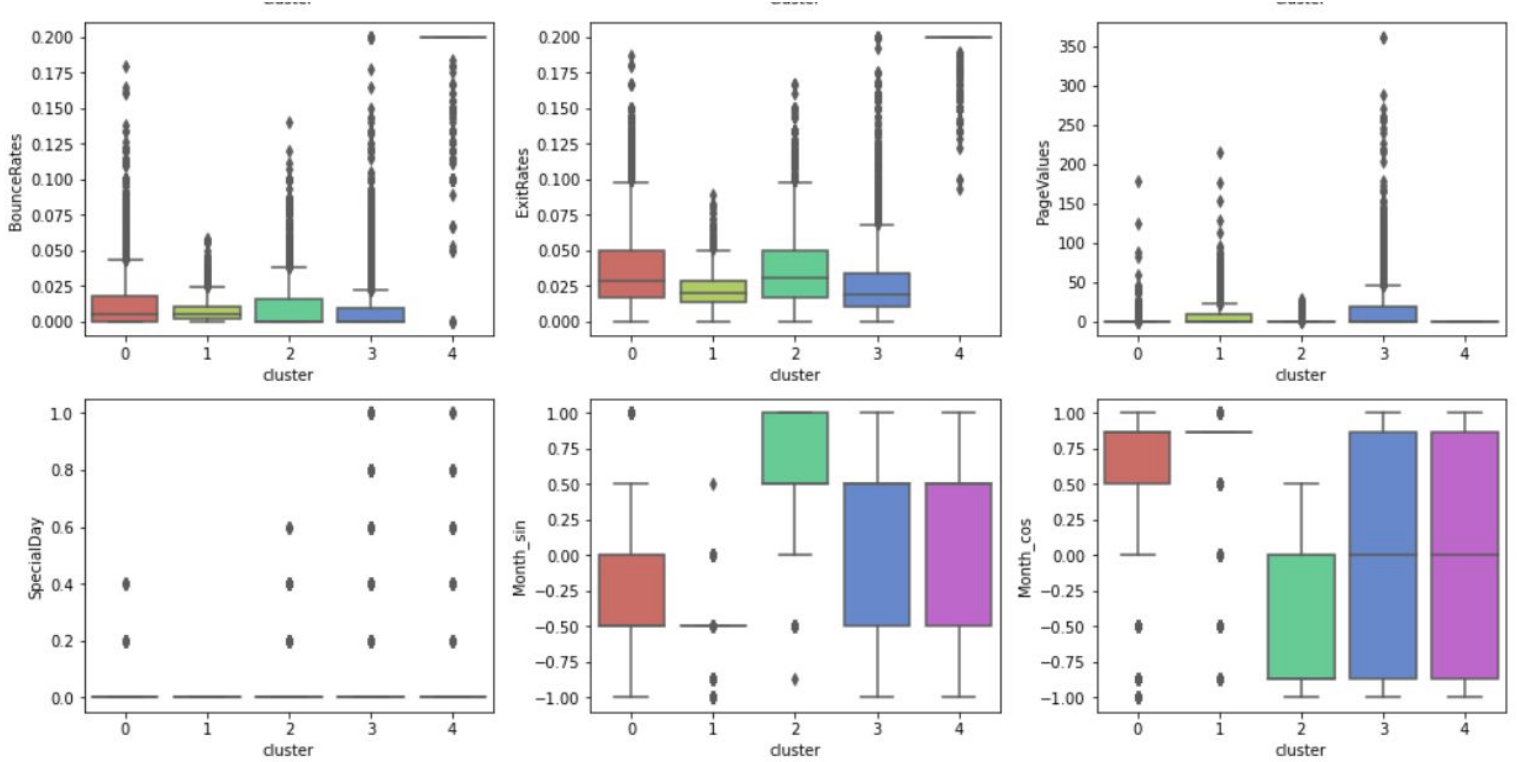


liczba obserwacji w poszczególnych klastrach

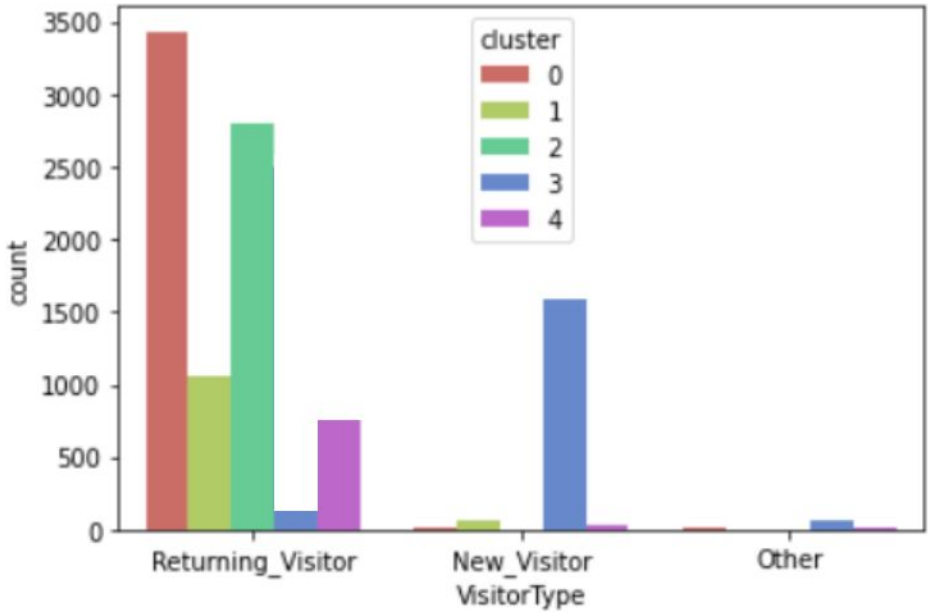
Rozkłady zmiennych w podziale na klastry (cz.1)



Rozkład zmiennych w podziale na klastry (cz.2)

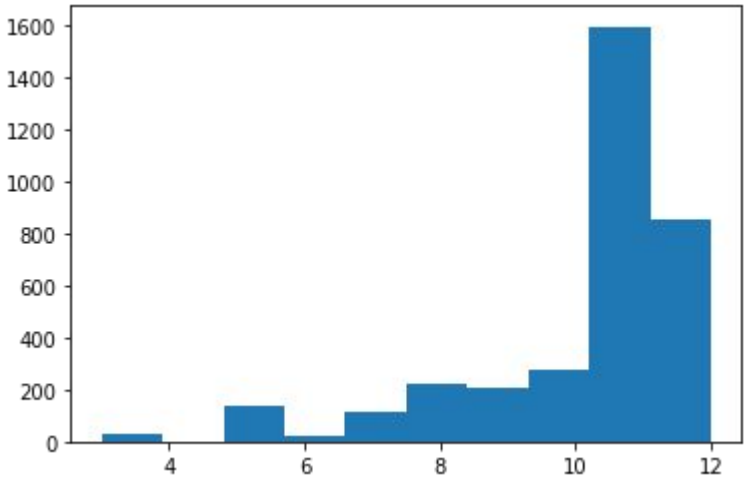


Zmienna “Visitor Type” ze względu na klaster

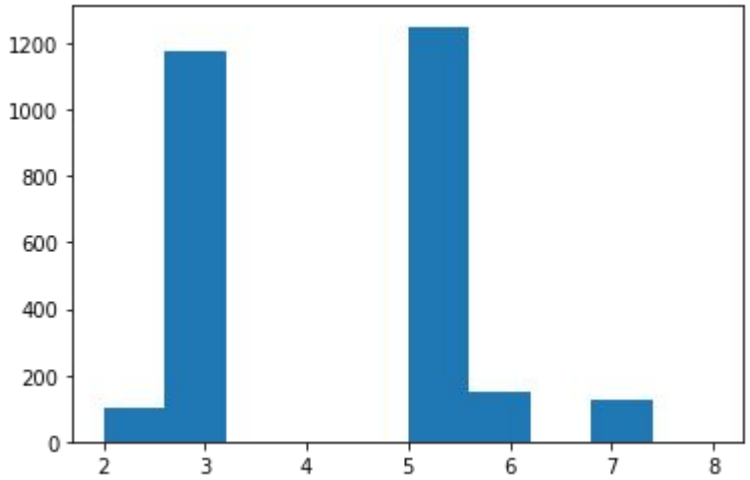


Zmienna miesiąc w klastrach 0 i 2

Miesiąc w klastrze 0



Miesiąc w klastrze 2



Klastry

- 0 - przeciętny użytkownik, zakupy w miesiącach zimowych
- 1 - ludzie świadomie robiący zakupy spędzają dużo czasu na stronach informacyjnych i dotyczących produktu, procentowo kupują najwięcej (revenue w ok. 30% obserwacji)
- 2 - przeciętny użytkownik strony, zakupy poza sezonem zimowym
- 3 - nowi użytkownicy - klaster został oddzielony praktycznie tylko względem tej zmiennej, nowi użytkownicy nie pojawiają się prawie w żadnym innym klastrze
- 4 - ludzie przekierowani z reklamy, np. z mediów społecznościowych, wysokie exit i bounce rates - od razu wychodzą, bardzo niskie revenue, tylko w 4/795 obserwacji

Metryki interpretowalne

Klastrowanie 2 (na 5 klastrów):

- minimum intercluster distance: 1,11
- mean intercluster distance: 4,56
- mean distance to center of cluster: 3,23

Klastrowanie 1 (na 12 klastrów):

- minimum intercluster distance: 0,50
- mean intercluster distance: 4,19
- mean distance to center of cluster: 2,96

Pozostałe (niekoniecznie udane) eksperymenty

- DBSCAN - ani metodą z zajęć ani innymi znalezionymi w internecie nie byliśmy w stanie dobrać hiperparametrów tak, aby klastrowanie było sensowne
- doszukiwanie się zakupowiczów sezonowych (np. przed Bożym Narodzeniem) w klastrach - zbiór nie jest zbalansowany pod względem miesięcy, dużo więcej obserwacji z zimy - miesiące będą naturalnie przekrzywione w tę stronę
- klastrowanie po PCA - próbowaliśmy zredukować wymiary przed klastrowaniem, jednak nie za każdym razem otrzymywaliśmy gorsze metryki oraz wizualizacje niż te opisane powyżej