Szymon Rećko
Patryk Słowakiewicz

Zakupy Online

Eksploracja i analiza danych

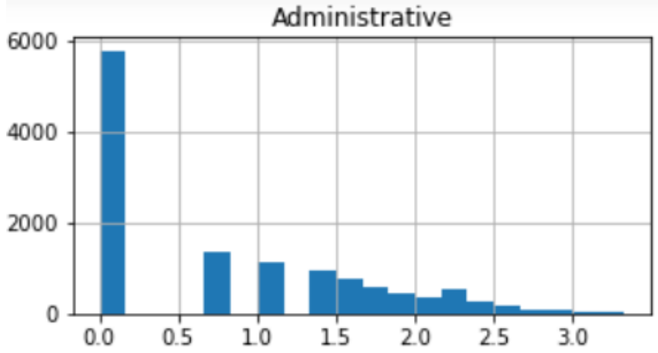
Opis zmiennych

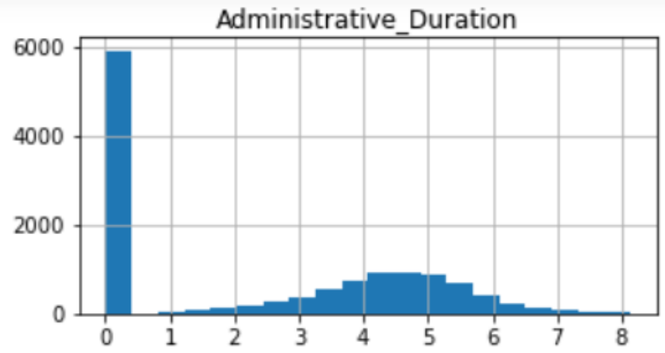
Administrative/Informational/Product Related (duration) - liczba odwiedzonych stron danego typu oraz sumerycznych czas spedzony na nich w czasie jednej sesji
Bounce Rate - współczynnik opuszczeń strony bez wykonania dalszych interakcji
Exit Rate - współczynnik odsłon strony w ostatniej sesji
Page Value - średnia liczba stron odwiedzonych przed dokonaniem zakupu
Special Day - odległość od najbliższego święta
Browser - przeglądarka użytkownika
Region - regiony
Traffic type - natężenie ruchu na stronie
Visitor type - nowy lub stały klient
Weekend - dzień weekendowy
Month - Miesiąc
Revenue - Dokonana płatność

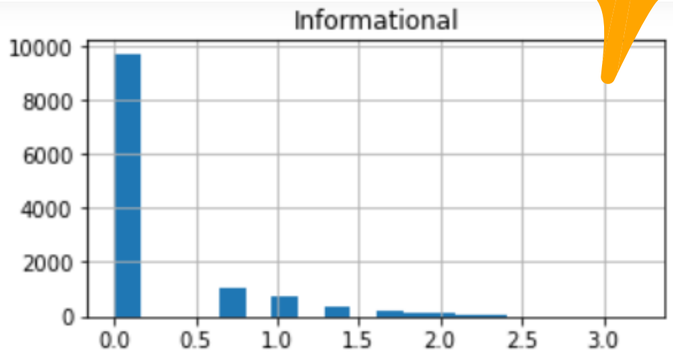
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157213	0.016813	0.050000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000

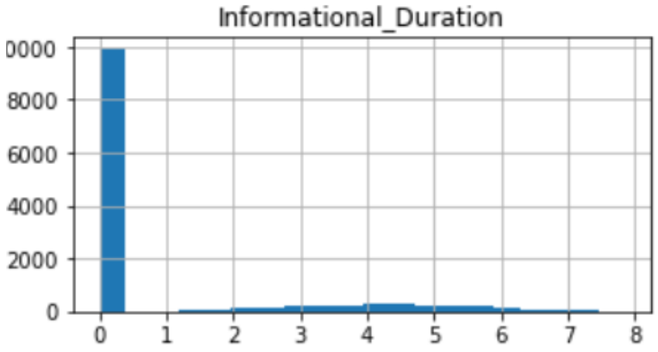
Rozkłady ciągłe

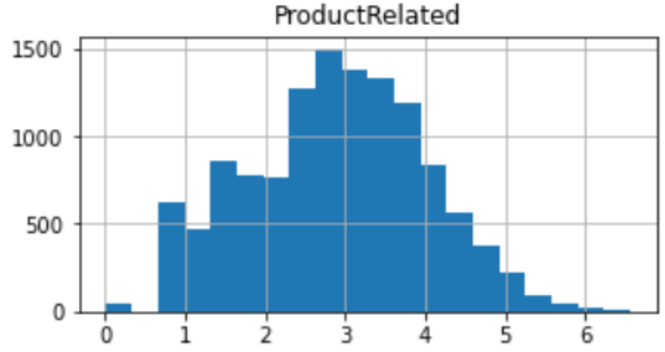
Po zlogarytmowaniu niektóre zminne zaczęły przypominać rozkład normalny przy czym nadal jest dużo wartości 0

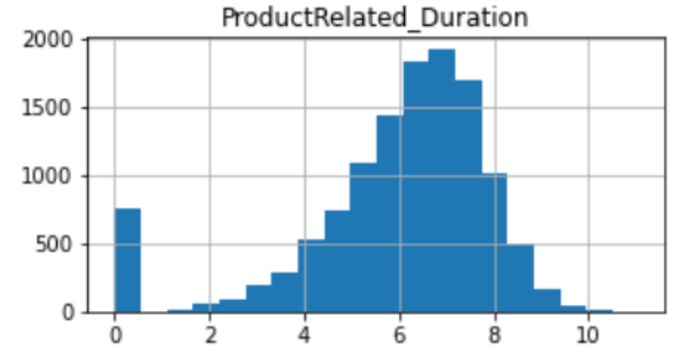


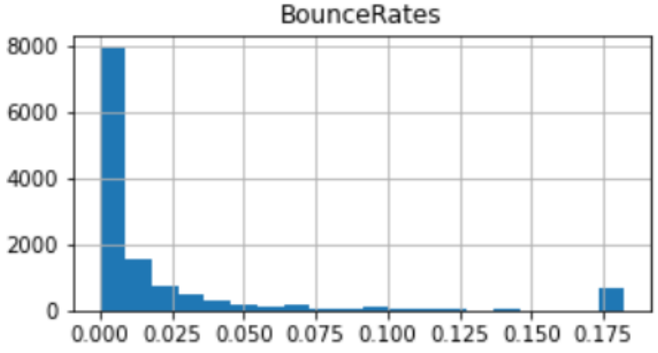


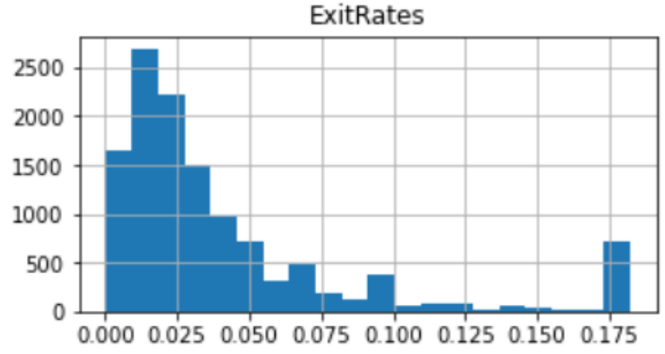


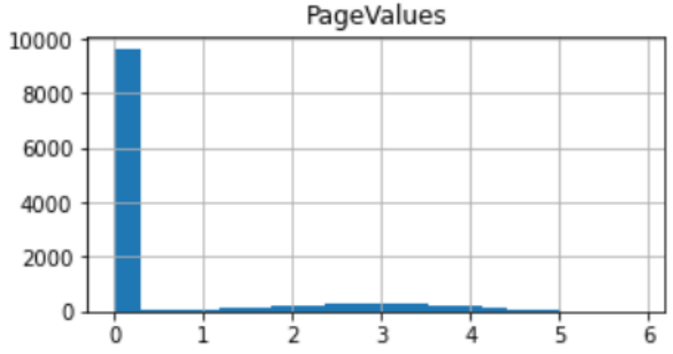








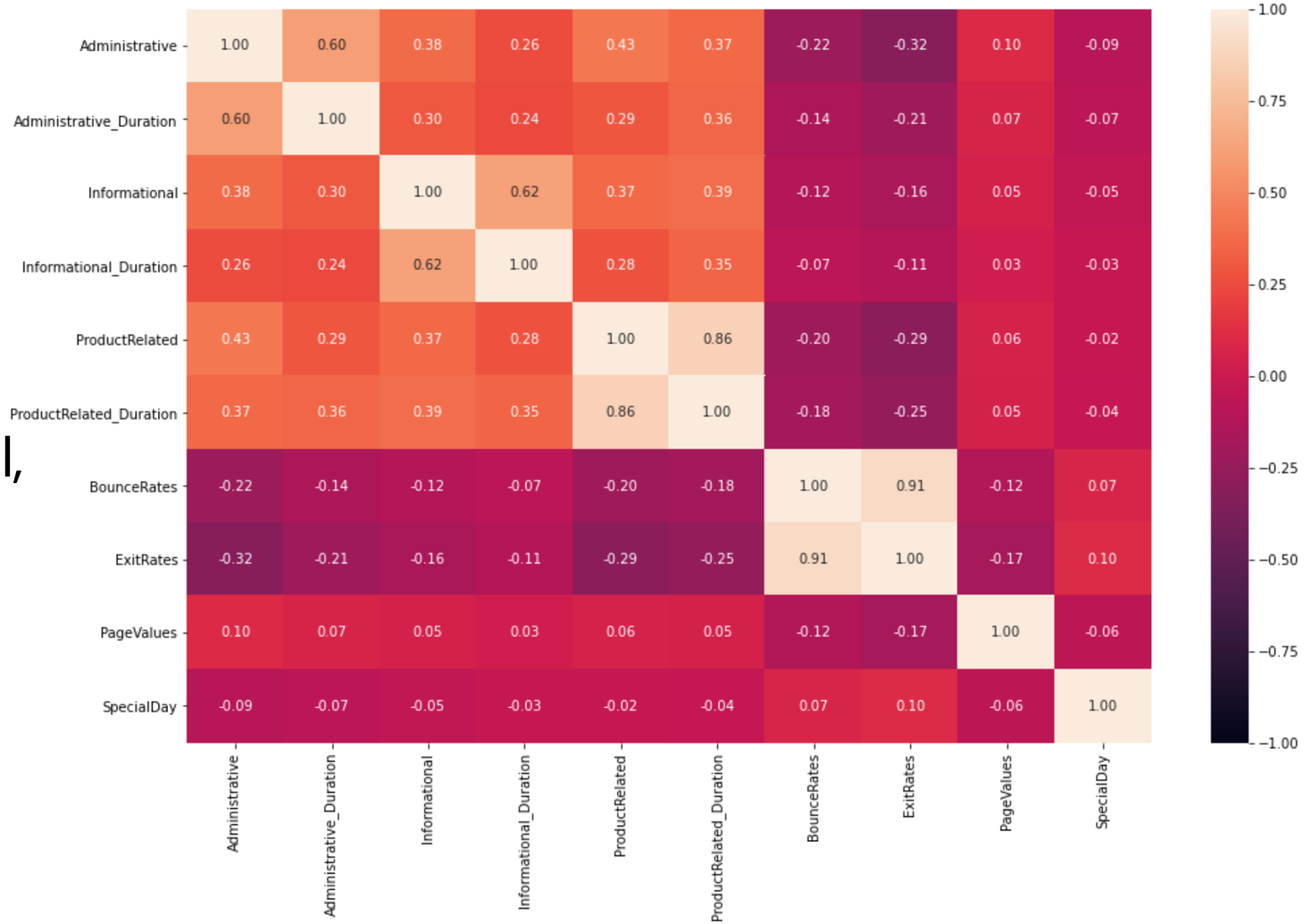




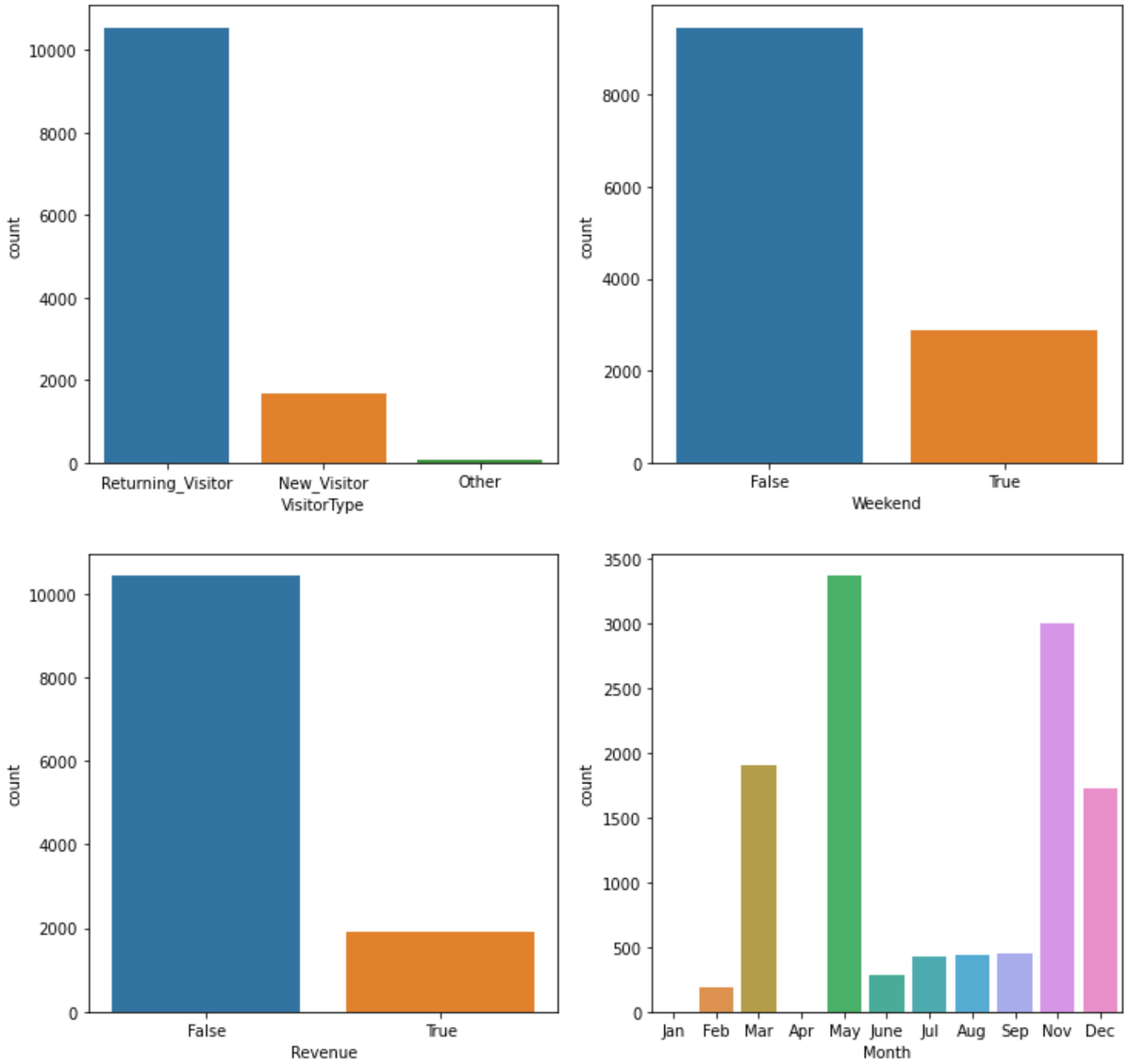
Znaleźliśmy też 720 obsetwacji dla których wszystkie Duration = 0

Macierz korelacji

- Istnieje silna korelacja między "BounceRates" i "ExitRates"
- Widzimy również korelacje między ilością stron danego typu (Administrative, Informational, Product Related), a sumarycznym czasem na nich spędzonych.
- Żadna z powyższych nie jest specjalnie zaskakująca



Zmienne kategoryczne



- Jak widać istnieją znaczące różnicy w liczności kategorii, szczególnie widać to na przykładzie zmiennej "VisitorType" co trzeba będzie wziąć pod uwagę.
- Nie wytępują miesiące styczeń i kwiecień

Special Days

0.2	**Feb**	15
	May	163
0.4	**Feb**	21
	May	222
0.6	**Feb**	19
	May	332
0.8	**Feb**	19
	May	306
1.0	**Feb**	5
	May	149

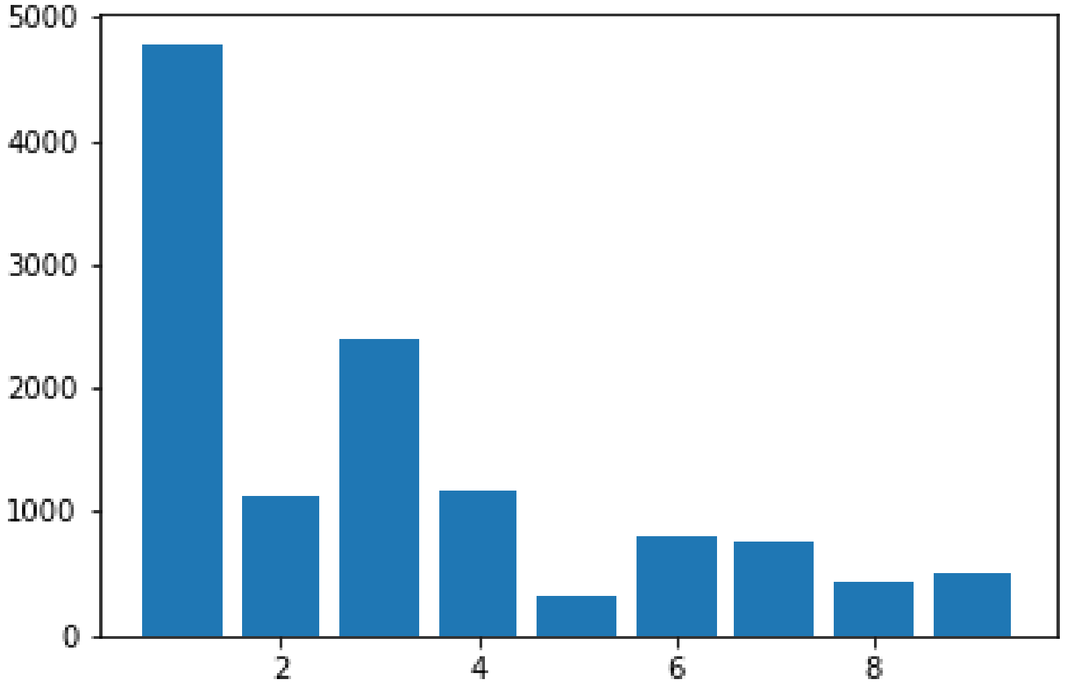
Interesująca wydaje się zmienna *Special Days* ponieważ wartości różne od 0 zystępują jedynie w Lutym i Marcu. Z czego przeważająca część w Marcu

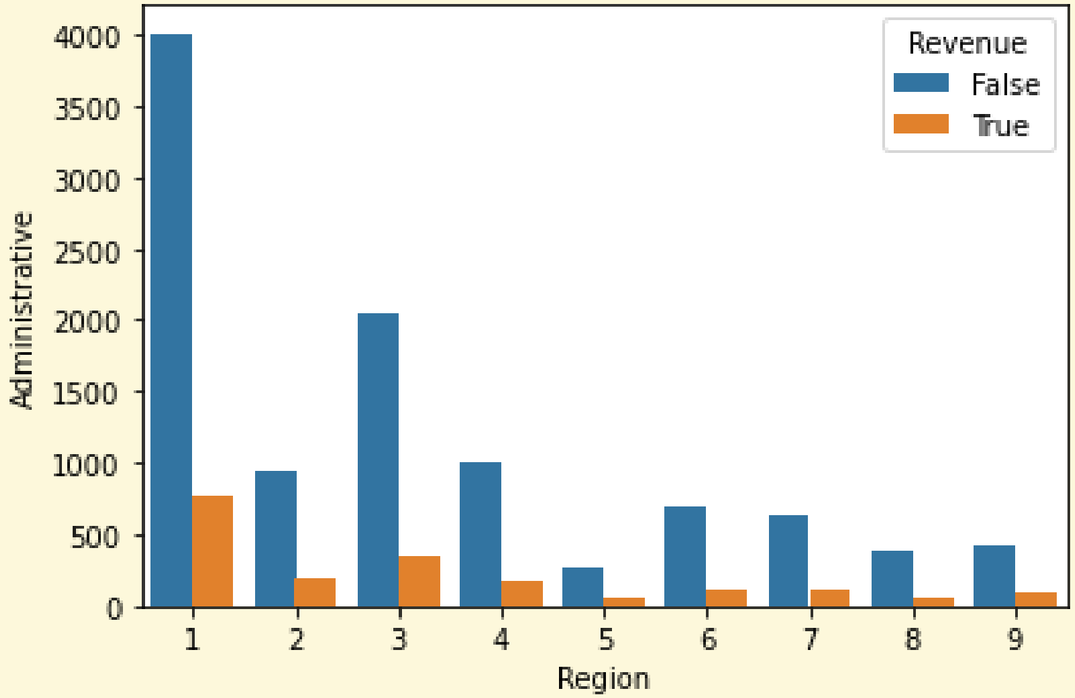
Weekend

Co ciekawe dni weekendowe wcale nie zwiększają znacznie ruchu na stronach ponieważ po przeskalowaniu liczba wyświetleń jest bardzo podobna

Region

Nie wszystkie regiony są zrównoważone w niektórych z nich mają marginalne liości sesji. Można też zauważyć dużą przewagę pierwszego regionu

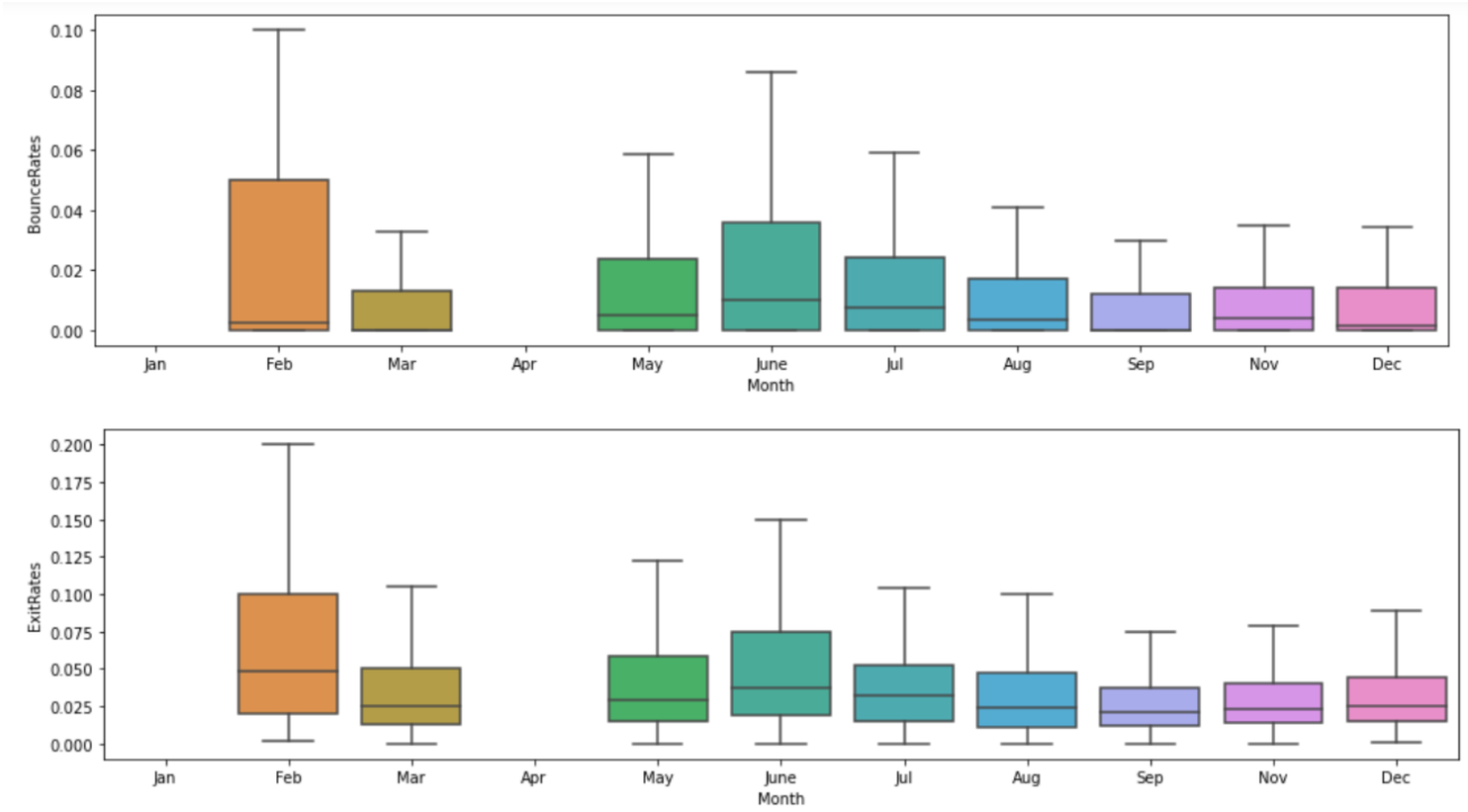




Co ciekawe nie niesiesie to za sobą dużych różnic procentowych w wartościach *Revenue*

Boxploty

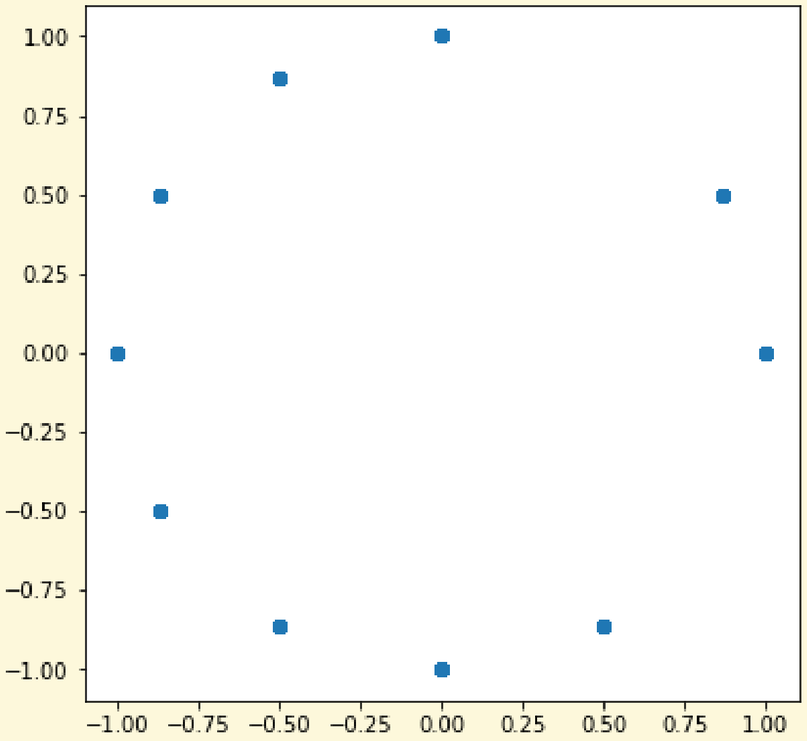
Stworzyliśmy jeszcze boxplot'y w zależności od miesiąca dla każdej zmiennej ciągłej. Można było na nich zaobserwować zmieniające się trendy, ale nic co by było na tyle znaczące by wziąć to pod uwagę w dalszej obróbce danych.



Przygotowanie danych

Usunęliśmy z danych Revenue

Kodowanie miesięcy



Aby uniknąć One-hot encodingu miesięcy postanowiliśmy wykorzystać Cyclical Encoding . Polega na zamianie danych na które występują w pewnym porządku cyklicznym na wartości x i y tak aby ostatecznie tworzyły okrąg.

$$x_{sin} = \sin\left(\frac{2*\pi*x}{\max(x)}\right)$$

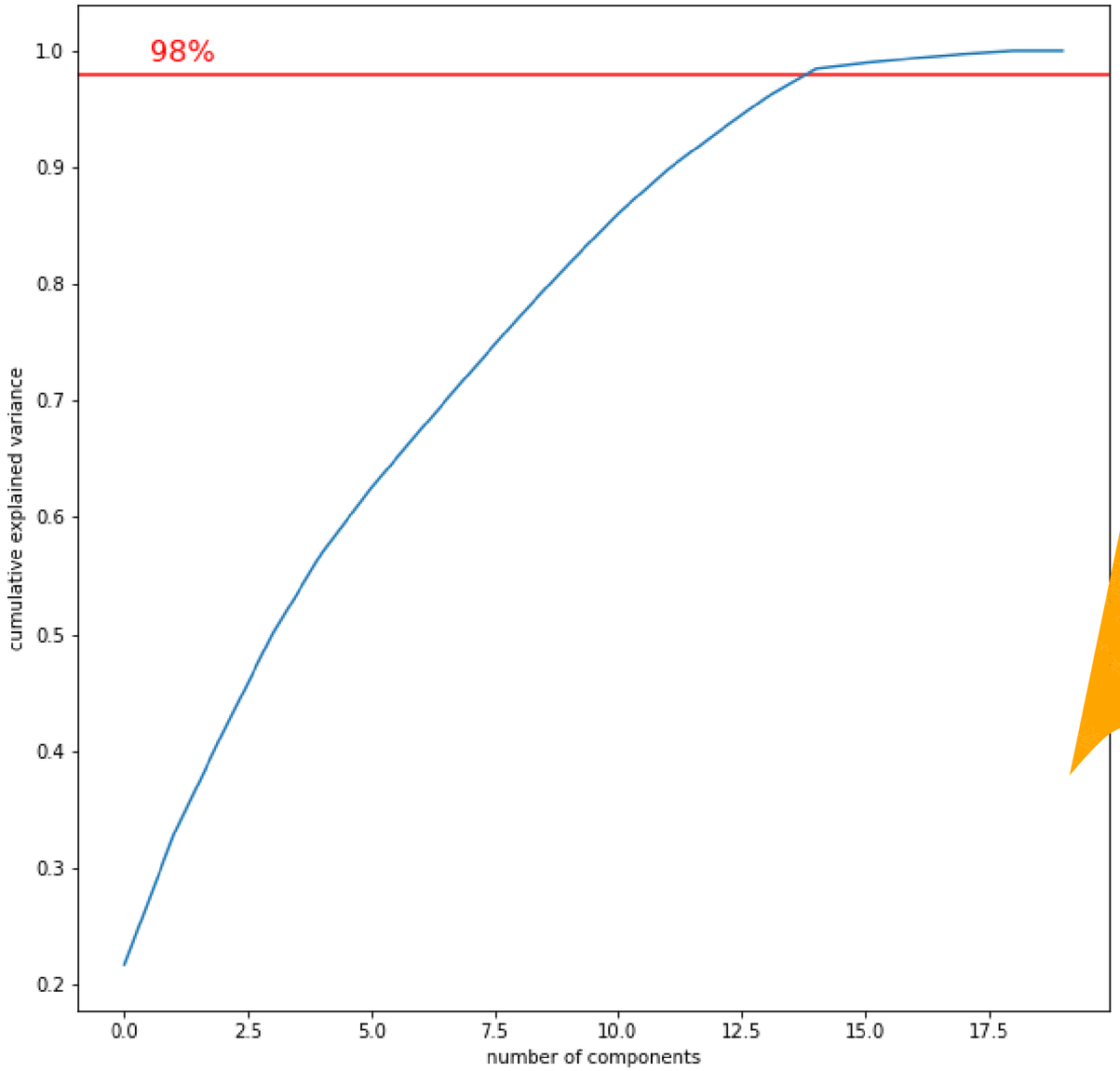
$$x_{cos} = \cos\left(\frac{2*\pi*x}{\max(x)}\right)$$

Visitor Type

Visitor Type jako jedyną zmienną zakodowaliśmy metodą One-Hot. Reszta zminnych posiadająca dużą liczbę wartości unikalnych pozostała danymi liczbowymi w formacie jaki otrzymaliśmy. Jeśli zmienne posiadały tylko dwie wartości unikalne zamieniliśmy je na 0 oraz 1.

PCA

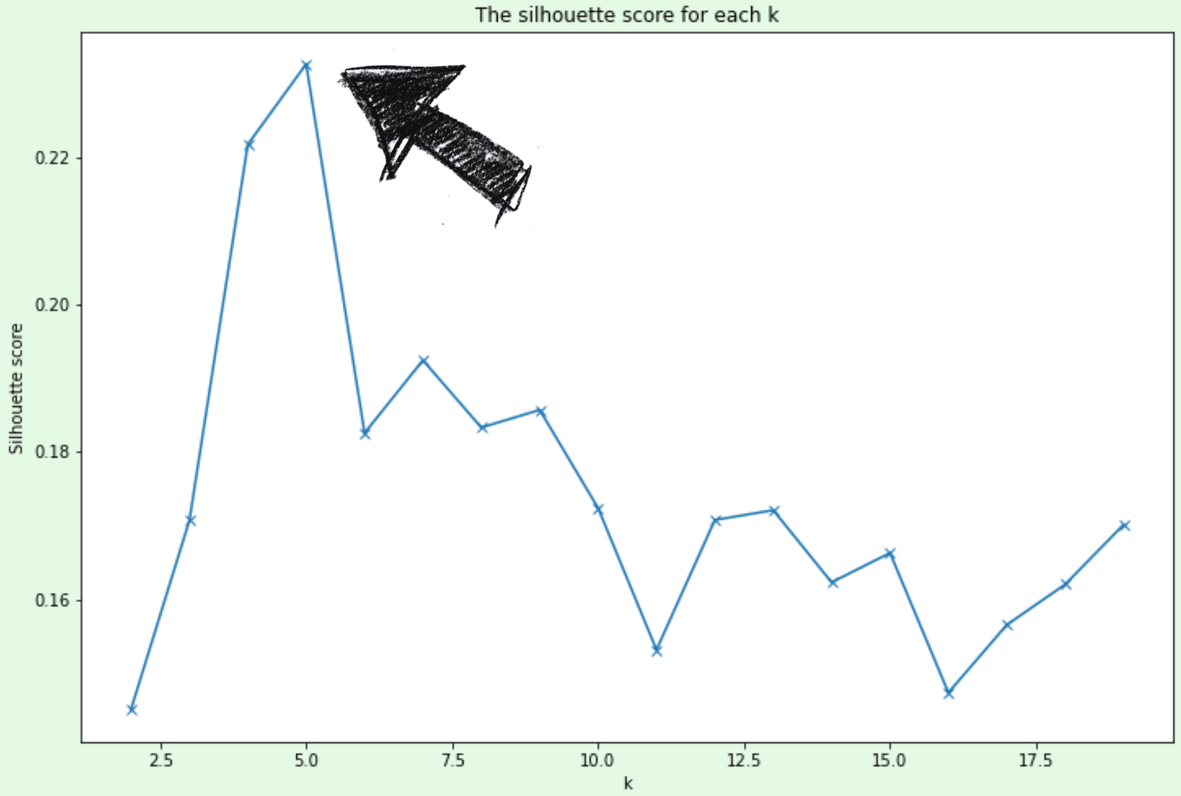
Przed użyciem PCA niektóre zmienne zostały zlogarytmowane. A następnie zeskalowane przy pomocy Standard Scalera.

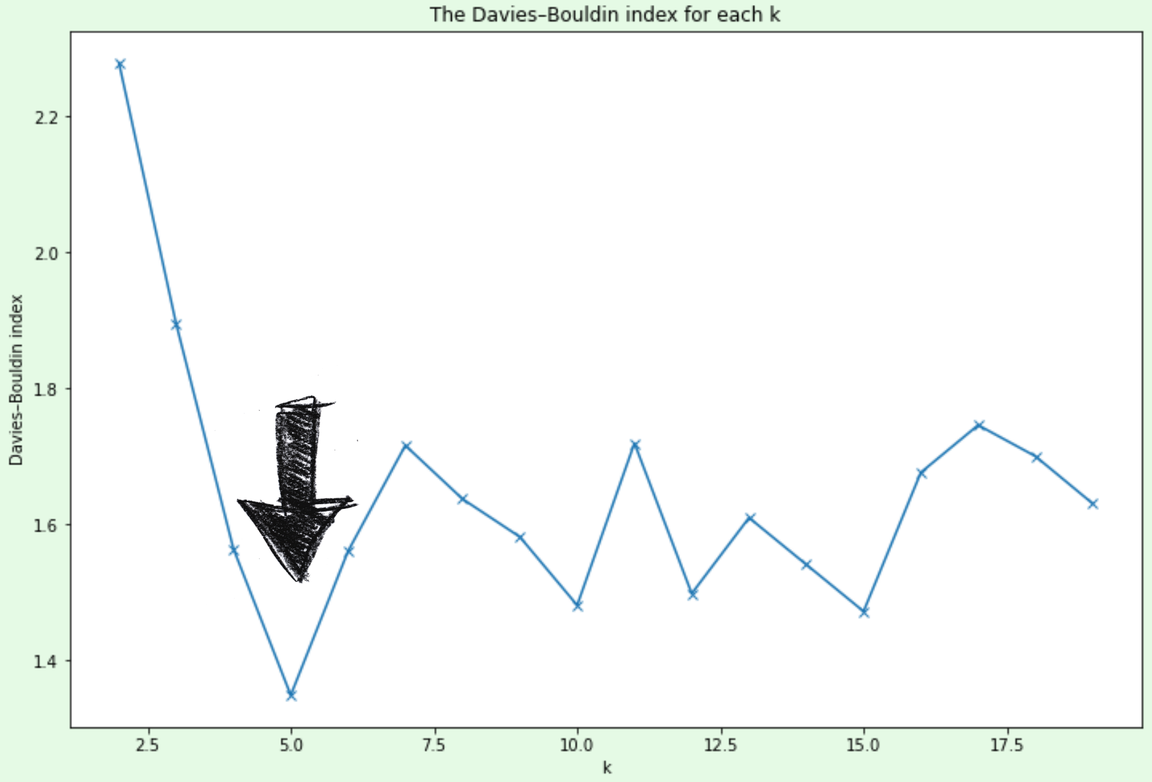


Wybraliśmy 14 z 20 zmiennych które wyjaśniają ponad 98% wariancji

BADANIE ILOŚCI KLASTRÓW

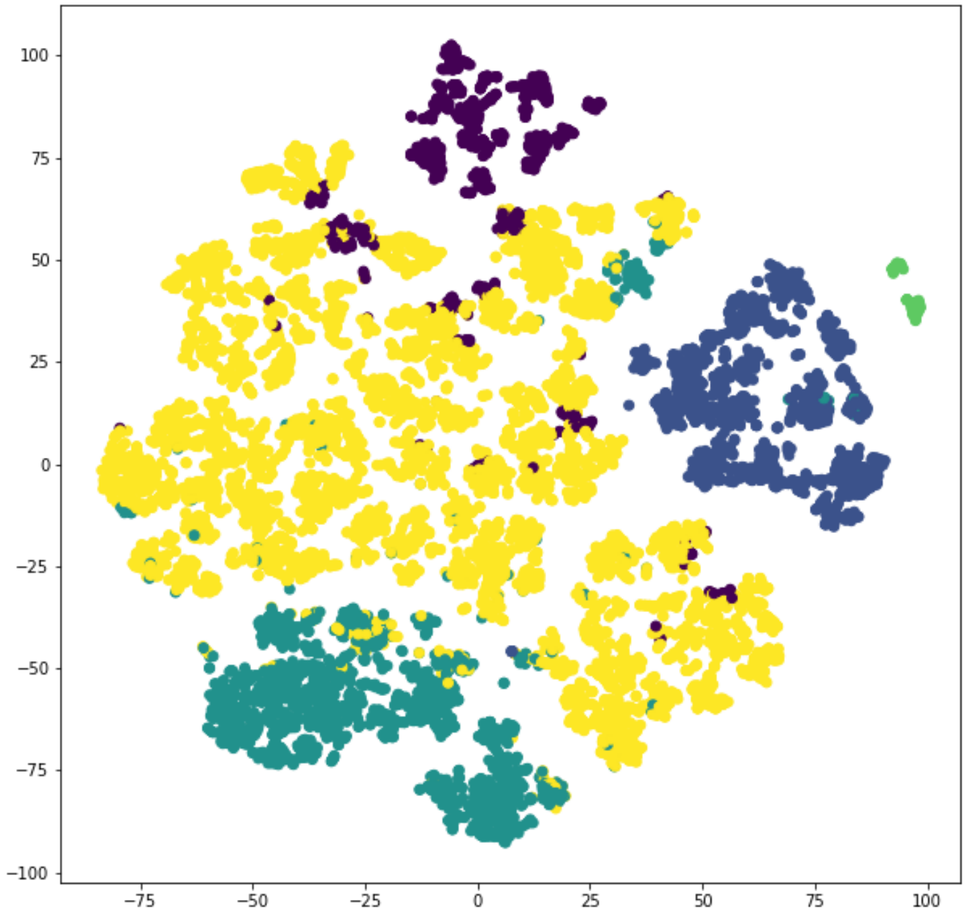
Stosunek odległości w klastrach do odległości między klastrami

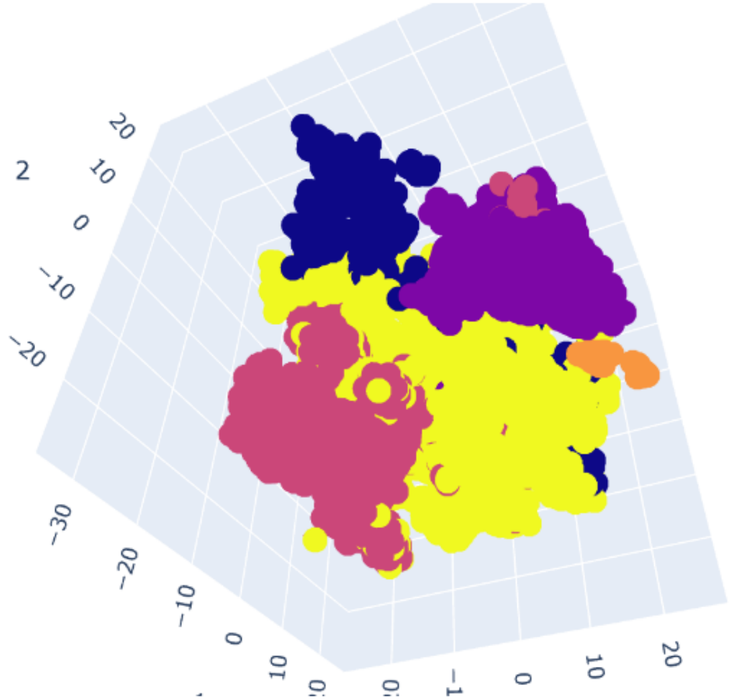




Modelowanie

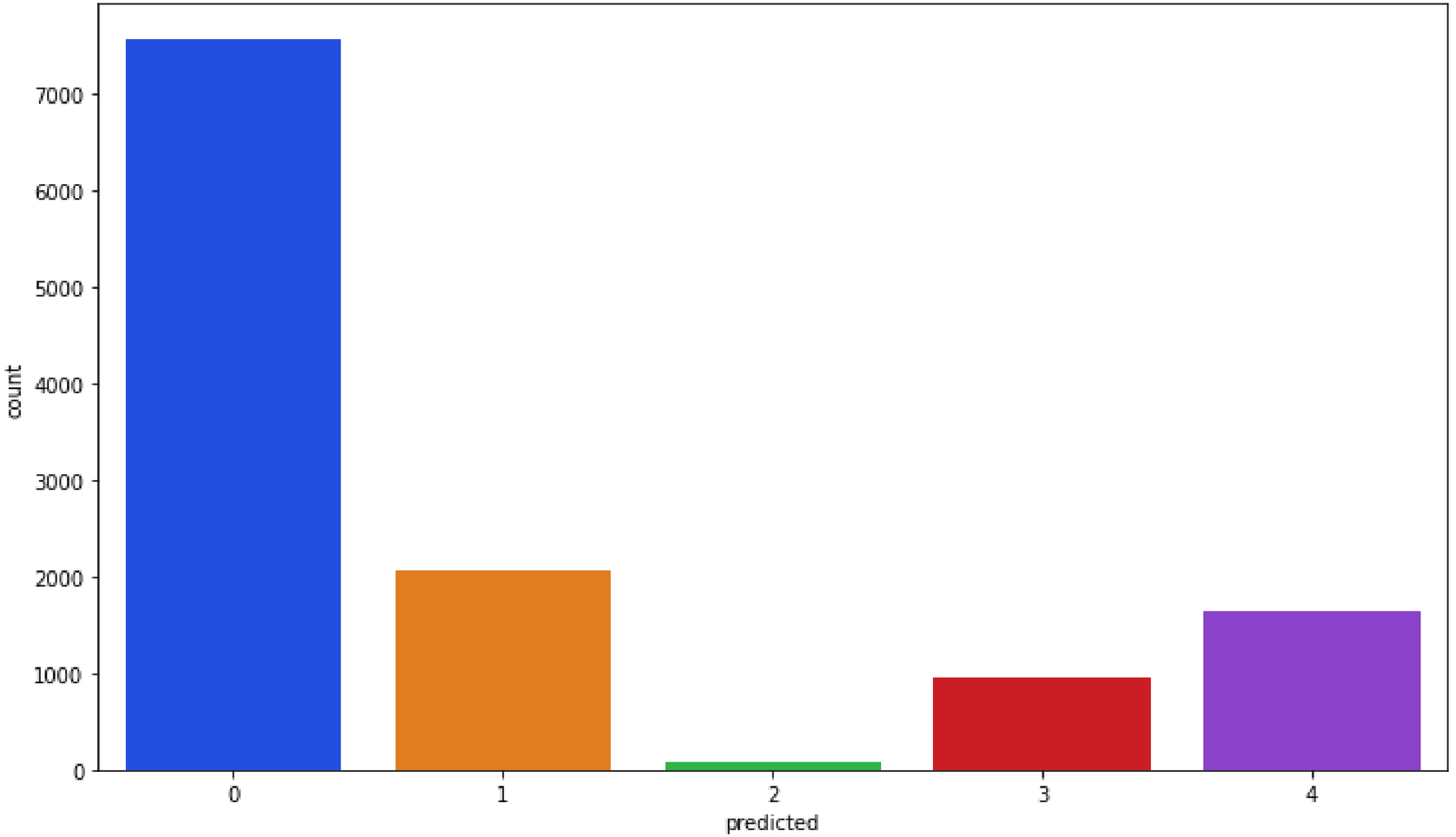
Do modelowania próbowaliśmy użyć K-means, K-medoids, DB Scan, Aglomeracja.
Ale najlepsze wyniki osiągnęliśmy dzięki K-mean z 5 klastrami






```
Minimal distance between clusters = 0.23.
Average distance between points in the same class = 5.33.
Standard deviation of distance between points in the same class = 1.578.
Average distance to cluster center = 3.83.
```

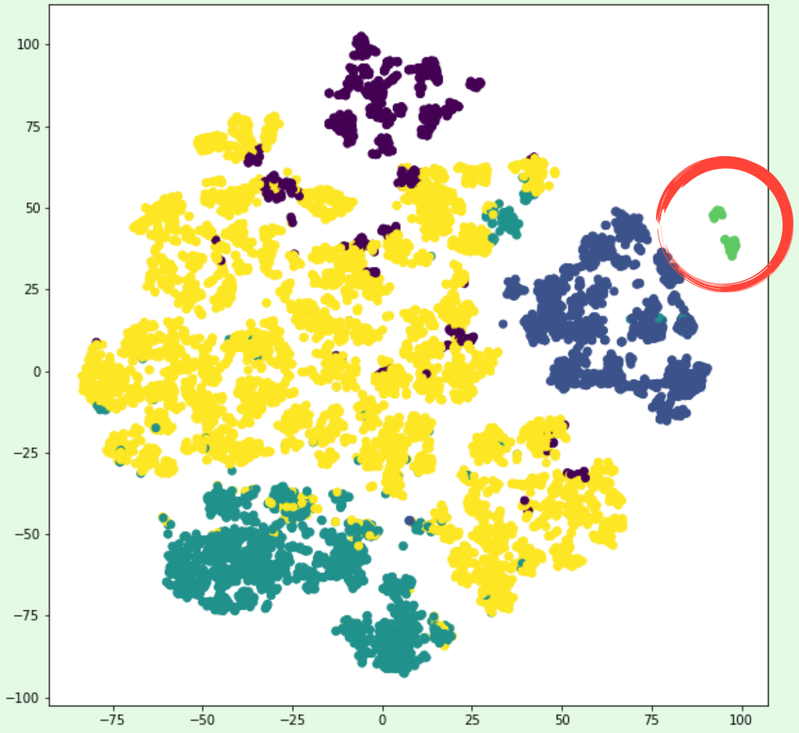
Analiza klastrowania

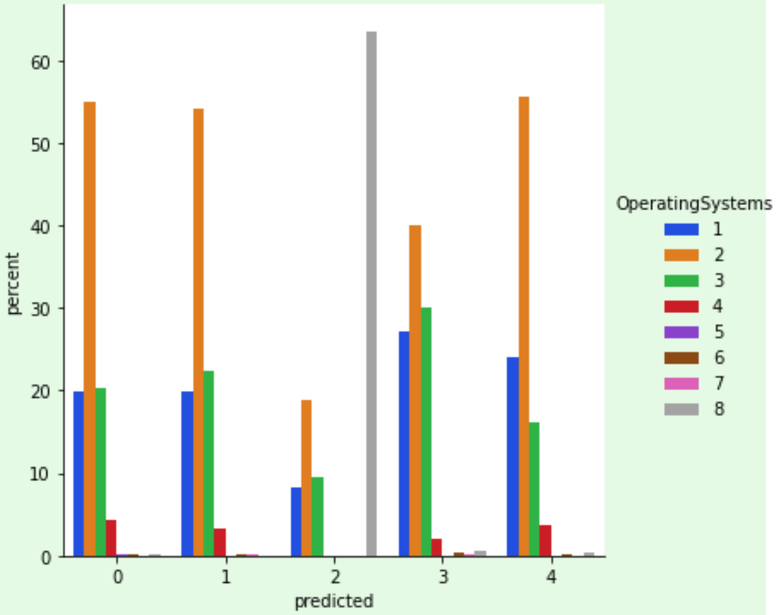


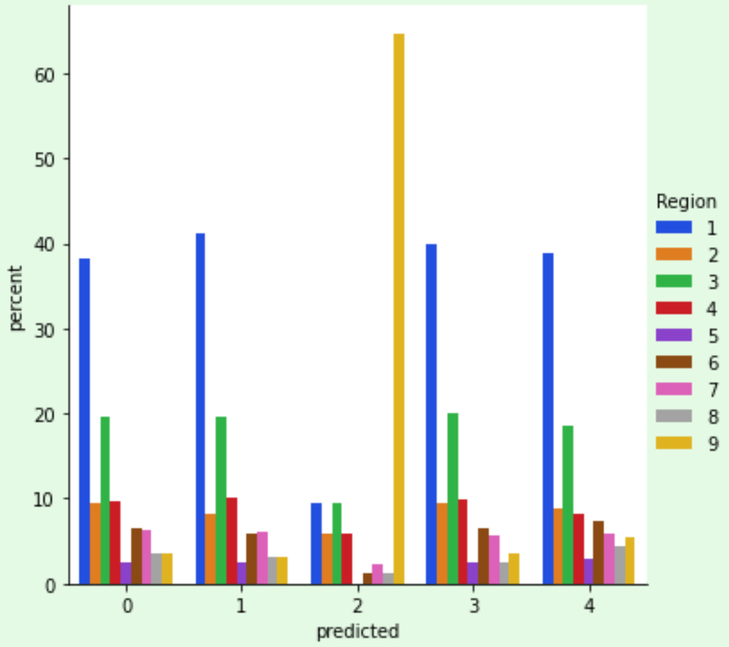
Wykres przedstawia liczność poszczególnych klastrów. Na następnych slajdach przedstawimy dwa najbardziej się wyróżniające, czyli 2 i 3.

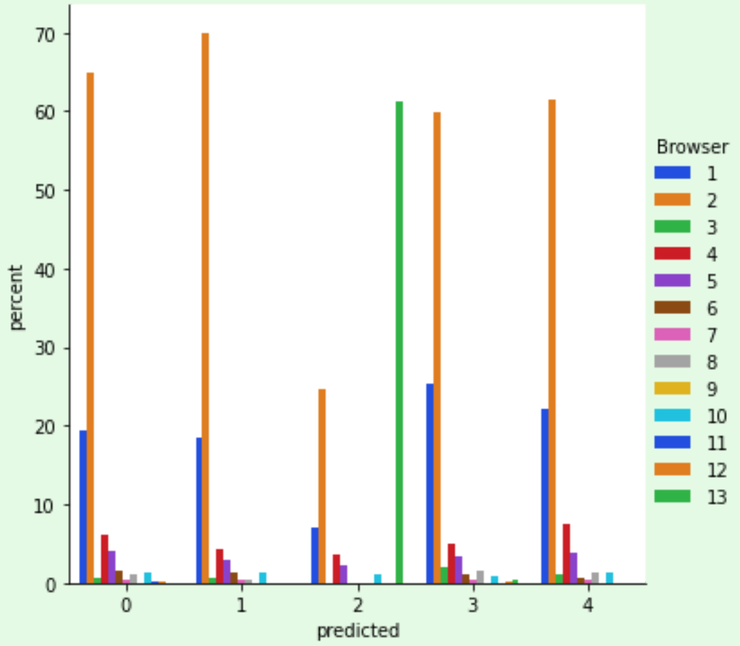
2

Nasz drugi klaster zawiera w sobie obserwacje które posiadają znaczną nadreprezentację w regionie 9. Poza tym jako jedyne w tak dużym procencie (ponad 60%) używają 'Operating System' równy 8 oraz Browser równe 13. Jest to bardzo nieliczna dlatego naszym zdaniem zgadzając się z wnioskami innej grupy mogą to być pracownicy/administratorzy strony którzy używają Linuxa.

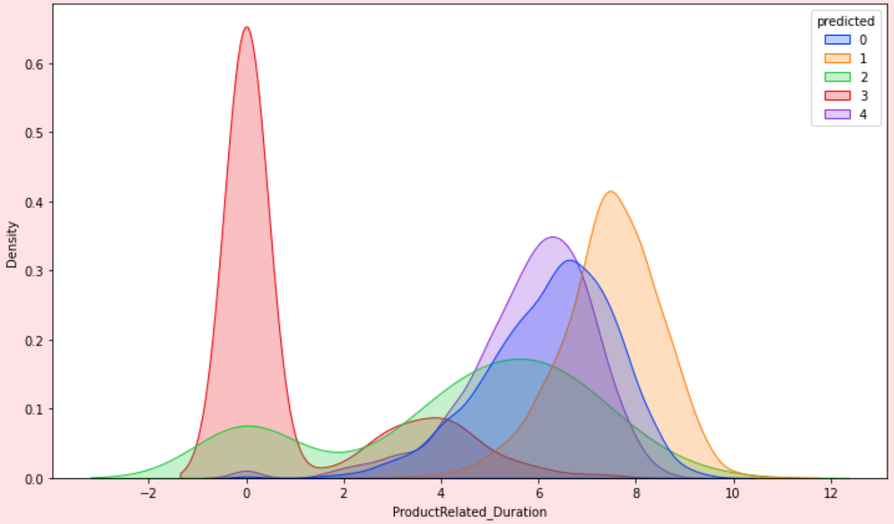


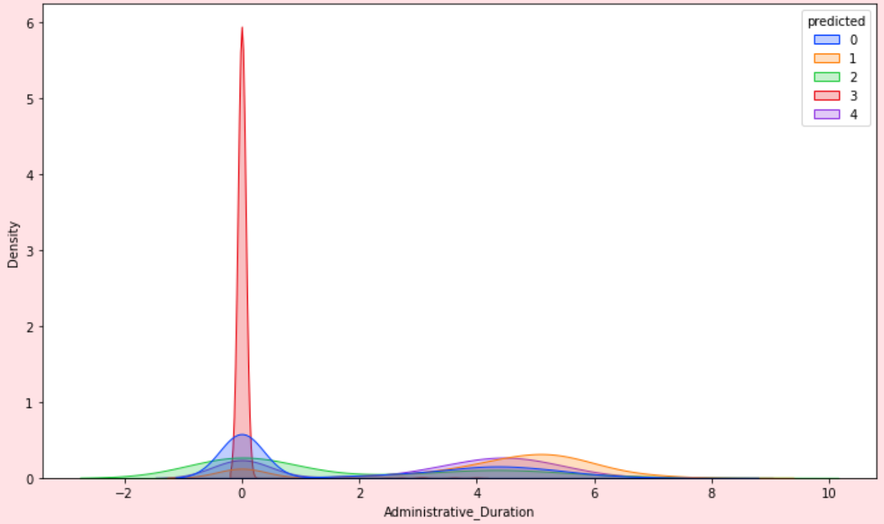






3





Osoby w trzecim klastrze cechują się tym, że nie spędzają wgl czasu na stronach, mają wysoki ExitRates i prawie wogóle nie dokonują zakupów. Sądzimy, że są to ludzie, którzy zostali przekierowani na strone przez przypadek.