# Raport walidacyjny grupy zajmującej się klasteryzacją danych o obiektach Airbnb w Nowym Jorku, czyli Dominiki Gimzickiej i Bartosza Jezierskiego

Anna Ostrowska, Michał Iwaniuk

June 2024

## 1 Projekt

Projekt zrealizowany został przez Dominikę Gimzicką i Bartosza Jezierskiego na przedmiot „Wstęp do uczenia maszynowego" na 4 semestrze kierunku Inżynieria i Analiza Danych na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej.
Celem projektu było wybranie modelu uczenia maszynowego, który pozwalałby na jak najlepszy podział obiektów dostępnych do wynajęcia w Nowym Jorku przez stronę Airbnb na klastry. Ramka danych z tymi obiektami została wzięta ze strony insideairbnb.com (link do danych w ostatnim rozdziale: „Źródła"). W różnych etapach projektu użyte i wypróbowane zostały różne modele, mniej i bardziej zaawansowane i znalezione ich najlepsze do realizacji celu parametry.
Naszym celem, jako walidatorów, było przejrzenie i przeanalizowanie ich kodu i postępów po każdej częsci projektu oraz udzielenie informacji zwrotnej, co naszym zdaniem możnaby było poprawić lub zmienić.
Link do repozytorium na Githubie wraz z naszym feedbackiem można znaleźć w źródłach na końcu raportu walidacynego. W folderze validation na tym repozytorium znajdują się informacje zwrotne przekazane modelarzom.

# 2 Jaki feedback dawaliśmy? Czy wzięto go pod uwagę?

## 2.1 Kamień milowy 1 - EDA

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| "print(data["rating"][0]) contains_only_numbers_and_dots(data["rating"][0])" Tutaj ciezko stwierdzic na pierwszy rzut oka, co to miało zrobić/pokazać, warto wyjaśnić krótko chociaż jednym zdaniem (czemu akurat dla „rating" tylko?) | Tak (wyjaśnione) |
| Podoba mi się to sprawdzanie typów danych i co zawierają te z innymi typami itp. - fajnie zrobione | Tak |
| Fajnie, że jest dużo komentarzy i tekstu ospisującego, co robicie i dlaczego. | Tak |
| Może warto by było dopisać jakieś podsumowanie (lub chociaż słownie na prezentacji) do wykresów (niekoniecznie, bo nie jest ich dużo, więc łatwo samemu sobie zobaczyć, ale możecie pomysleć nad tym) | Tak |
| Fajnie, że już zaczęliście mapowanie | Tak |
| Na scatterplotach i pozniejszych heatmapach faktycznie niewiele widać, ale to już sami napisaliście - fajnie, że próbowaliście heatmapy, bo przynajmniej po przybliżeniu tam cokolwiek widać, a na scatterplotach nic. | Tak (na prezentacji pokazane zostały głównie heatmapy przybliżone na ważniejsze miejsca) |
| Do następnych wykresów (ceny od różnych kolumn - boxplot i histogramy): nie jest nic do nich napisane, ale prawdopodobnie planowaliście na prezentacji coś do nich powiedzieć, jeśli będziecie je pokazywać. Nie musicie dopisywać nic w kodzie, bo sporo widać, ale jeśli będą te wykresy pokazywane, to nie zapomnijcie coś sami do nich dopowiedzieć, a nie tylko pokazać:) | Tak - powiedziane na prezentacji |
| To samo do sekcji „wpływ zmiennych na rating", co w pkt. 7. | Tak |
| Podobają mi się mapy z dzielnicami i „poddzielnicami" - pomocne i fajnie widoczne. | Tak - używane i tworzone dalej w następnych etapach projektu. |
| Mapa z cenami i rating też bardzo fajna, ale jakiś komentarz dotyczący tego, że wielkość kropki to wysokość ceny by nie przydał (może być słownie tylko na prezentacji), bo nie ma w legendzie. | Tak |
| To samo co w pkt. 10 do ostatniego wykresu. | Tak |
| Ogólnie bardzo fajnie zrobione, dużo różnych rzeczy, dużo różnych wartości objaśnionych i zilustrowanych | Tak. |



Table 1: Realizacja zagadnień w ramach KM1

Tutaj trzeba przyznać, że modelarze zrealizowali każdy punkt naszego feedbacku i dokładnie objaśniali podczas prezentacji tego etapu projektu wszytskie elementy, które zasugerowaliśmy.

## 2.2 KM2 - inżynieria cech

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| Dlaczego używacie 2 scalerow (standard i maxmin)? W jakich modelach planujecie korzystac z jednego i z drugiego? | Tak - usunięto standard i pozostano przy MaxMin. |
| Trochę bez sensu, że dwa razy robicie transformacje i skalowanie? | Tak - poprawiono na robienie tego 1 raz. |
| Ja mam error „BracketError: The algorithm terminated without finding a valid bracket. Consider trying different initial points" przy transformacji danych, ale jak u was działa to fajnie. | Tak |
| Reszta jest okej, mogłoby być trochę więcej opisów, żeby było wiadomo, co dokładnie robicie i dlaczego lub gdzie i jak chcecie to wykorzystać. | Pół na pół - opisów w kodzie nie dodano, ale opisano ustnie npodczas prezentacji. |

Table 2: Realizacja zagadnień w ramach KM2

## 2.3 KM3 - ostateczne modele

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| Mapki to całkiem fajny pomysł | Tak |
| Skąd te wyższe liczby klastrów? 78 i 90 np, jakiej wyjaśnienie dlaczego akurat tyle? | Tak - ma być dodane wyjaśnienie. |
| Nie mogę nigdzie znaleźć waszego pomysłu biznesowego? Był gdzieś na poprzednich etapach zapisywany? | Tak - na prezentacji. |
| Poniższy kawałek kodu w ogóle mi się nie chce odpalić (podany kawałek kodu) | Tak - odpala się, ale długo. |
| Tutaj literówka: „smaples"<br>(drobne spostrzeżenie, ale w raporcie może źle wyglądać,<br>a pojawia się w kilku miejscach:))<br>print(f"Silhouette score dla eps=0.4 i min_smaples=8: score_4")<br>print(f"Silhouette score dla eps=0.8 i min_smaples=3: score_8") | Tak |
| Czemu akurat eps=0.1 przy mapce „istotne kolumny"?(po strojeniu parametrów) - wyjaśnić. | Tak |
| Przy komentarzu „Ponownie 3 wypada wyjątkowo dobrze. 17 i 19 również są dobre. 27 się nei wyróżnia, ale za to 34 już tak" chyba chodziło o 17 i 20: 19 ma wysoki score, a szukamy niskich wyników (w David Bouldin score przy GMM). | Tak |
| Cały czas powatrzacie, że ileś tam (3,17,19) klastrów to za mało - myślę, że przydałoby się jakieś wyjaśnienie, dlaczego. | Tak - ma być dodane wyjaśnienie chociaż słownie. |

Table 3: Realizacja zagadnień w ramach KM3

# 3 Sprawdzenie modeli na zbiorze dla walidatorów

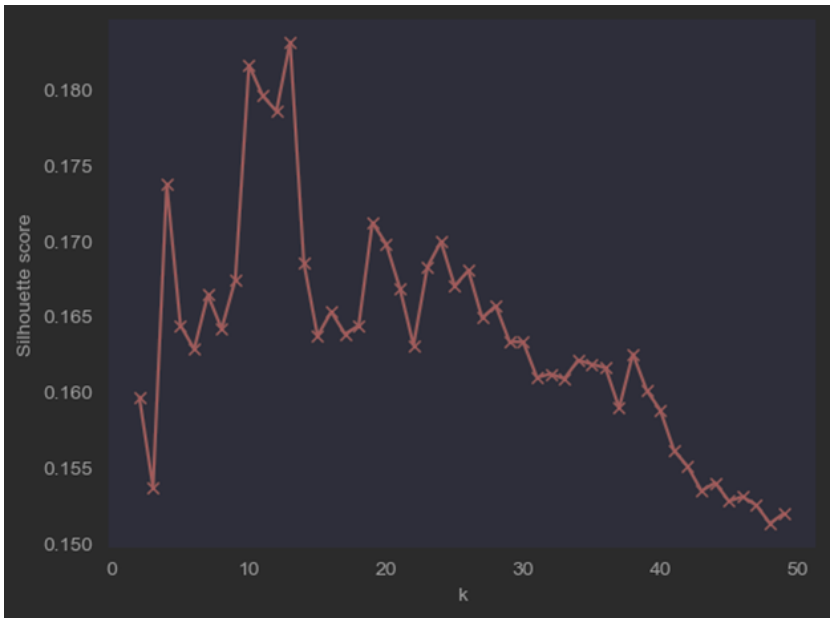


Figure 1: Silhouette score dla różnej liczby klastrów.

Trochę inaczej, niż u modelarzy (29 nie jest najlepsze u nas, a u nich było).

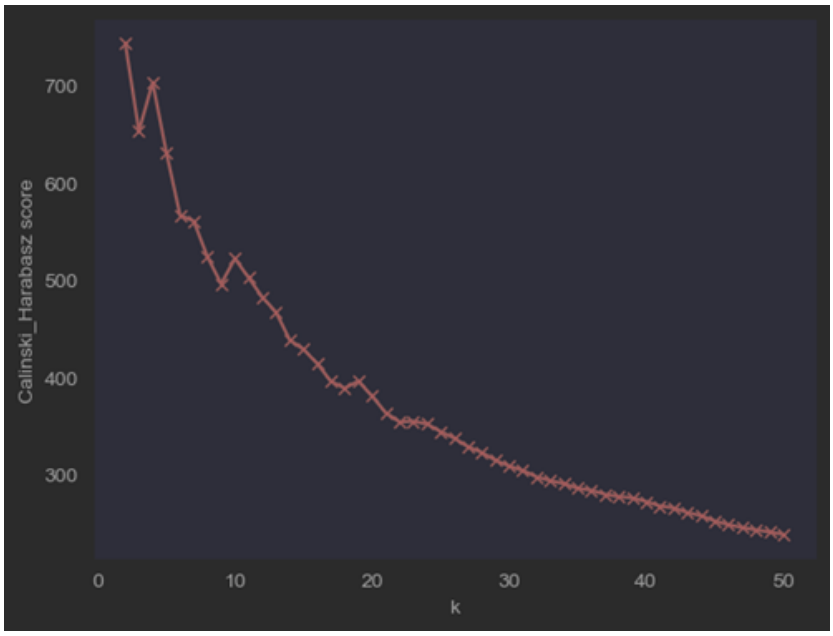


Figure 2: Calinski-Harabasz score dla różnej liczby klastrów.

Podobnie, jak u modelarzy.

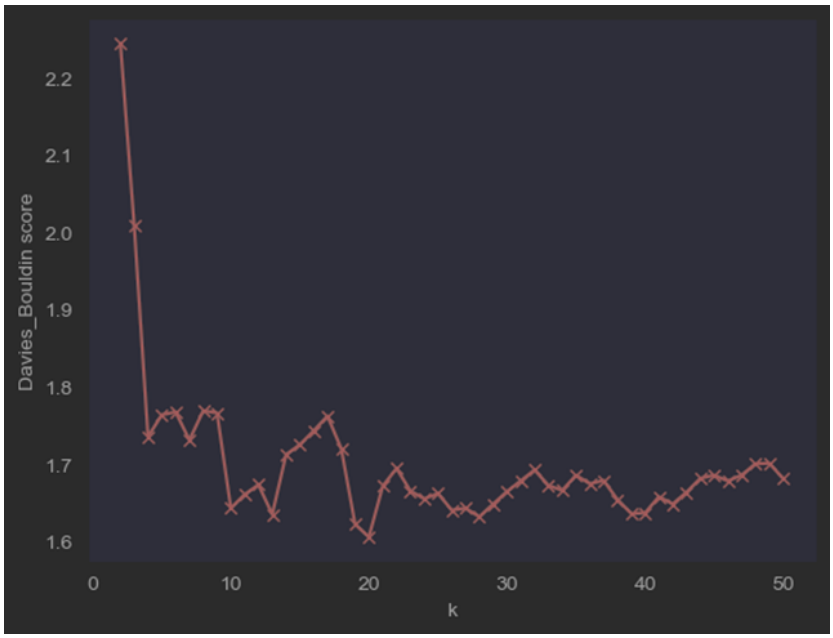


Figure 3: Davies-Bouldin score dla różnej liczby klastrów.

Trochę inaczej, niż u modelarzy, ale podobnie - na ich zbiorze 19 jest najlepsze, u nas nie, ale jest blisko.

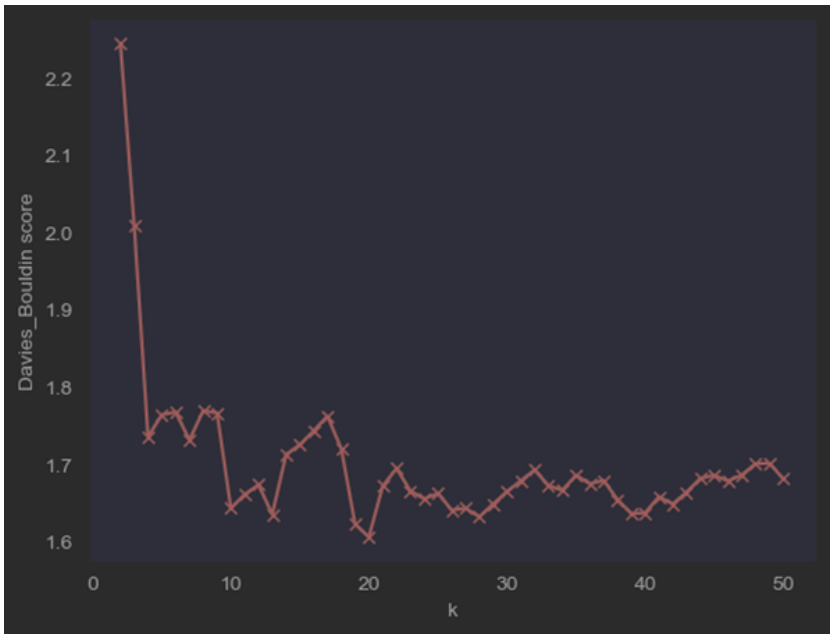


Figure 4: Davies-Bouldin score dla różnej liczby klastrów przy mzienionym parametrze algorithm na 'elkan'.

U nas wychodzi tak samo, jak przy poporzednich parametrach - tak jak na zbiorze modelrzy (u nich wychodziły te 2 tak samo, ale trochę inaczej niż u nas - niewielka różnica).

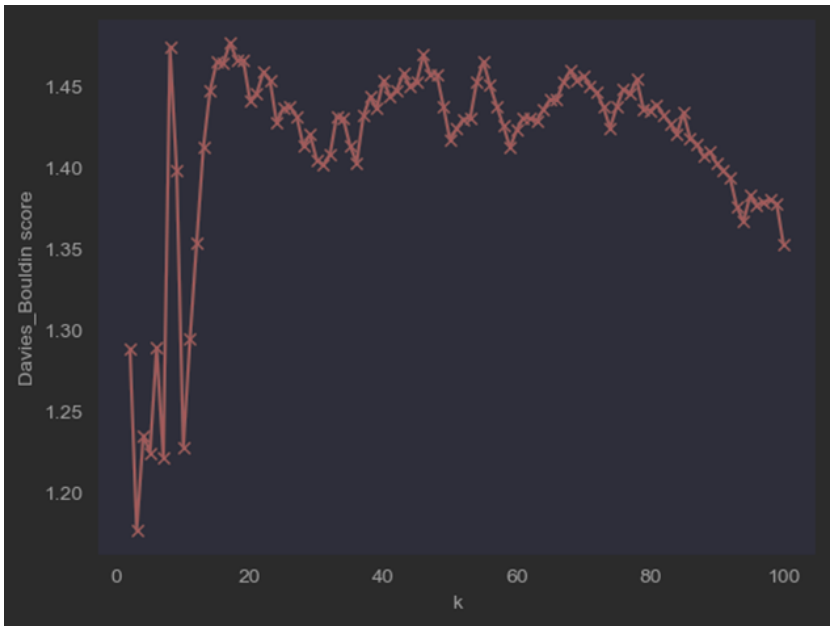


Figure 5: Davies-Bouldin score dla różnej liczby klastrów przy pozostawieniu tylko kilku najważniejszych kolumn.

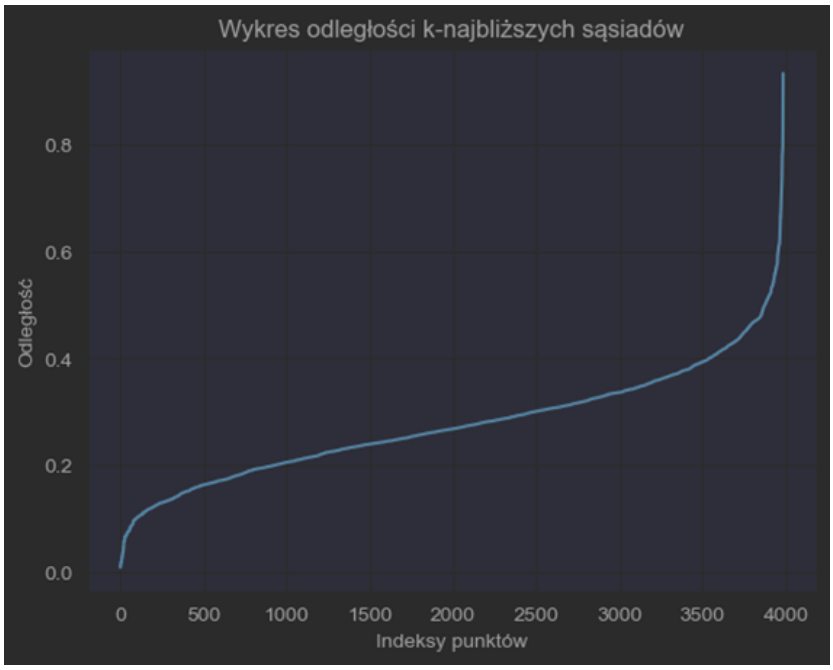


Figure 6: Wybieranie najlepszego parametru eps dla DBSCAN.

U nas wyniki trochę inne - najlepszy parametr eps wychodzi pomiędzy 0.4 a 0.6, a na zbiorze modelarzy pomiędzy 0.3 a 0.4.
W strojeniu parametrów, patrząc tylko na ep pomiędzy 0.3 a 0.4, u nas wygrało: Best score: -0.014048569661685752, eps: 0.4, min_samples: 19.
Pomiędzy 0.3 a 1 najlepsze wyniki na naszym zbiorze: Best score: 0.2342590145262561, eps: 0.95, min_samples: 15 (min_samples w obu przedziałach podobnie, na drugim przedziale eps się różni: u modelarzy był 0.8).

**Sprawdzanie, który zestaw parametrów jest lepszy:**
Silhouette score dla eps=0.4 i min_samples=8: -0.019458360793340727
Silhouette score dla eps=0.8 i min_samples=3: 0.17442343546044226
Drugi zestaw lepszy, tak jak u modelarzy.

Indeks Daviesa-Bouldina dla eps=0.4 i min_samples=20: 1.9749525444876865
Indeks Daviesa-Bouldina dla eps=0.8 i min_samples=15: 2.324629587513448
Tutaj również drugie lepsze jak u modelarzy.

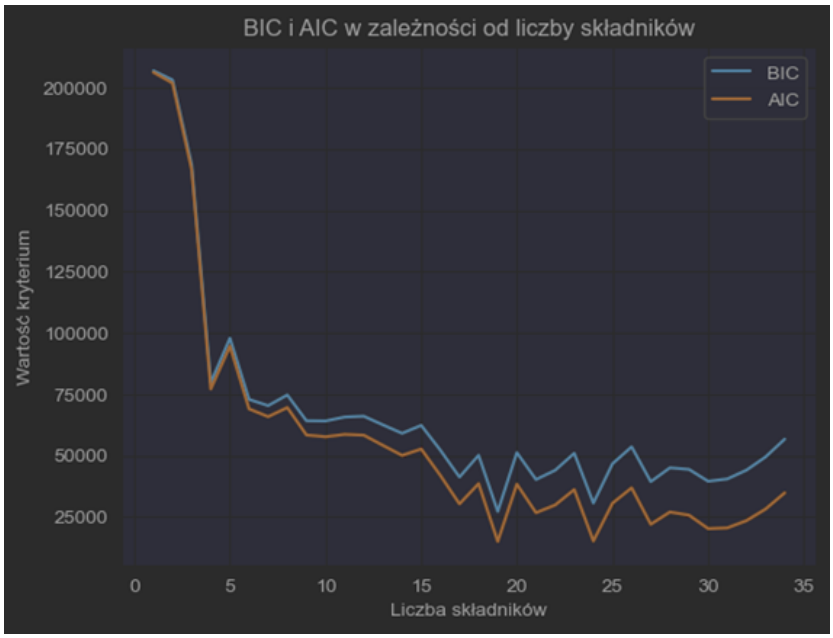


Figure 7: BIC vs AIC dla GMM.

Wychodzi u nas tak samo, jak u modelarzy - BIC i AIC wypadają podobnie.

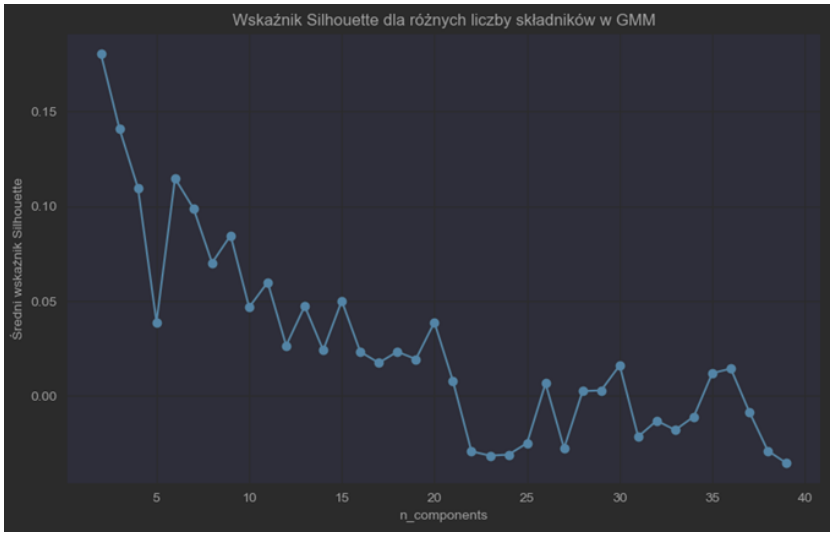


Figure 8: Wskaźnik Silhouette dla różnych liczby składników w GMM.

Tutaj mamy zupełne inne wyniki, niz na zbiorze modelarzy.

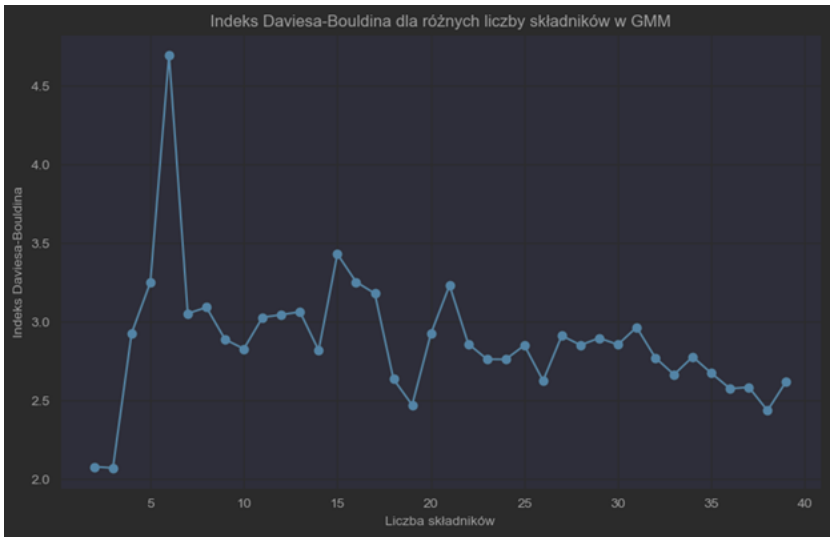


Figure 9: Indeks Daviesa-Bouldina dla różnych liczby składników w GMM

Dla indeksu Daviesa-Bouldina już bardziej podobnie do zbioru modelarzy.

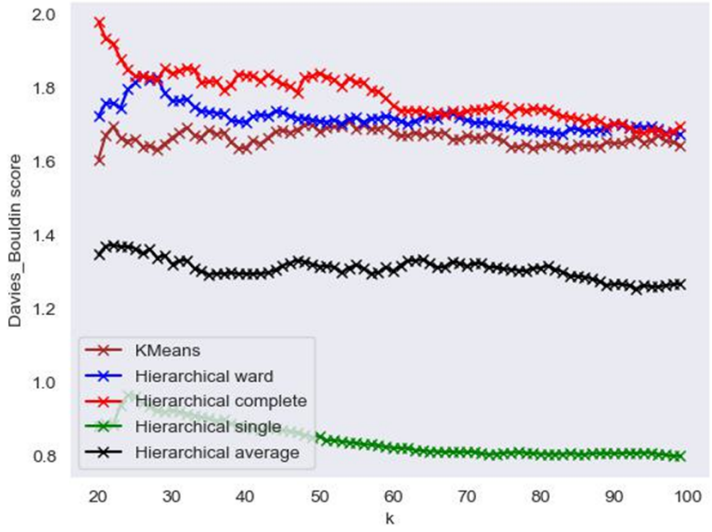


Figure 10: Porównanie różnych modeli i liczby klastrów za pomocą różnych metryk.

Tutaj wyniki prawie takie same, jak u modelarzy - bez większych różnic.

# 4 Podsumowanie

Dzięki regularnemu kontaktowi między nami a zespołem walidowanym, udało się nam na każdym etapie pracy nanosić komentarze i poprawki, które liczymy, że usprawniły i ulepszyły efekty ich pracy. Uważamy, że była to owocna współpraca, co pokazuje ilość wziętych pod uwagę naszych uwag. Uważamy, że zespołowi modelarzy udało się stworzyć bardzo dobry projekt i otrzymać dobry model.

# 5 Źródła

- link do zbioru danych, z których grupa modelarzy korzystała: https://insideairbnb.com/get-the-data/
- link do repozytorium na Githubie grupy modelarzy: https://github.com/AristocratesJ/New-York-Airbnb-Clustering-Task