НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
Факультет прикладної математики
Кафедра прикладної математики

Звіт
з лабораторної роботи № 4: «Регресійний аналіз»
із дисципліни «Аналіз даних»

Виконали:
Гармаш О. Є.
Хок М. Ш.
Куцалаба Н. В.
Маслов Н.Р.

Керівник:
ст. викладач Тавров Д. Ю.

1 Вступ

1.1 Мотивація дослідження

Літаки завжди були одним із перших варіантів для подорожей через їхню зручність і безпеку. З постійним підвищенням рівня життя людей, зростають групи клієнтів цивільної авіації, і люди висувають більш високі вимоги до якості авіаційних послуг. Прогнозування задоволеності пасажирів літака та визначення основних факторів, що на неї впливають, можуть допомогти авіакомпаніям покращити свої послуги та отримати переваги в складних ситуаціях і конкуренції. Таким чином, авіакомпанії повинні своєчасно досліджувати задоволеність пасажирів різними послугами та загальну задоволеність, щоб точно розуміти якість обслуговування існуючих послуг. Крім того, авіакомпанії повинні чітко розуміти основні фактори, що впливають на задоволеність пасажирів, і сформулювати відповідні стратегії для покращення якості обслуговування, щоб максимізувати загальну задоволеність пасажирів авіакомпанією та підвищити лояльність пасажирів.

Виходячи з вищезазначених проблем, у цьому дослідженні в якості об’єкта дослідження використовується повна інформація про пасажирів і результати опитування щодо задоволеності окремими факторами рейсу.

1.2 Завдання на роботу

У Лабораторній роботі 3 ми будували лінійну регресійну модель для встановлення принаймі статистично значущого зв’язку (якщо не впливу) між різними змінними. У поточній роботи стоїть задача розглянути непараметричні методи оцінки функції регресії та порівняти результати з лінійною моделлю. Також студенти повинні провести аналіз головних компонент для свого датасету.

У цій роботі ми продовжуємо працювати з дослідницькими питаннями із попередніх робіт. Це дасть змогу здійснити безпосереднє порівняння різних підходів до аналізу. Студенти повинні оцінити відповідну непараметричну модель за допомогою ядрової регресії та частково-лійної регресії.

2 Аналіз головних компонент

2.1 Особливості розбиття

Для проведення регресійного аналізу, слід згадати з минулих Лабораторних робіт, що ми визначили деяку різницю в оцінках сервісів між клієнтами, які подорожували бізнес та економ класами (нагадаємо, що існують два економ класи – *Eco* та *Eco Plus*, які ми об'єднали в один для спрощення дослідження). Саме на цьому підгрунті ми будуватимемо дві різні моделі, щоб дізнатися, які саме фактори впливають на клієнтів кожного класу.

Звісно, окрім розбиття по класах, існують і інші фактори, наприклад *Type of Travel* (ціль поїздки), *Gender*, *Customer Type* (лояльний або не лояльний пасажир), проте, досліджуючи розбиття по цих факторах ми не знайшли відмінностей у впливах сервісів на задоволеність пасажирів, а саме тому, в даному звіті про це не згадується.

2.2 Аналіз головних компонент

Перш ніж приступати до проведення аналізу головних компонент, потрібно перевірити, чи доцільно взагалі зменшувати розмірність даних. Звісно, навіть якщо метод, яким ми будемо визначати це скаже нам про недоцільність, ми застосуємо PCA для демонстрації знань.

Ми перевіряли доцільність застосування PCA за допомогою тесту сферичності Барлетта. Це статистичний тест, який перевіряє гіпотези про гомоскедатичність (тобто однорідність дисперсій) в регресійній моделі.
Нульова гіпотеза тесту Бартлетта стверджує, що дисперсії залежної змінної однакові для всіх груп. Якщо p-value менше заданого рівня значущості, то ми відхиляємо нульову гіпотезу і стверджуємо, що дисперсії неоднорідні.

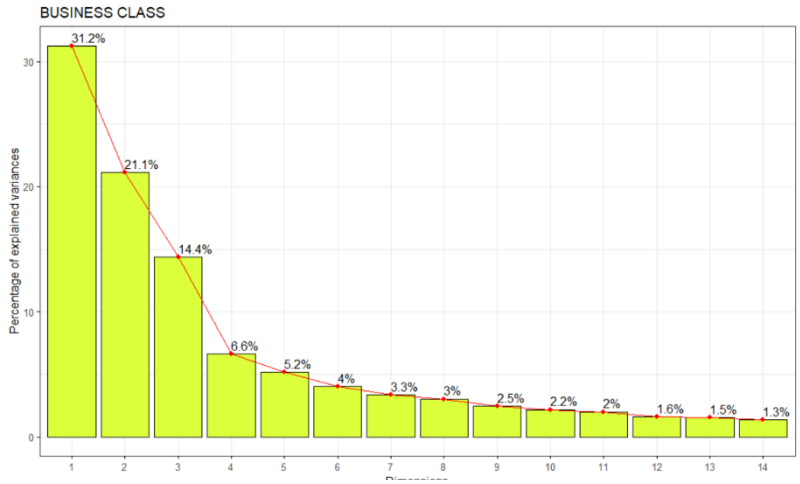
Фактично, якщо нульова гіпотеза не виконується, то це значить, що всі змінні є ортогональними і немає способу поєднати змінні як фактори або компоненти.

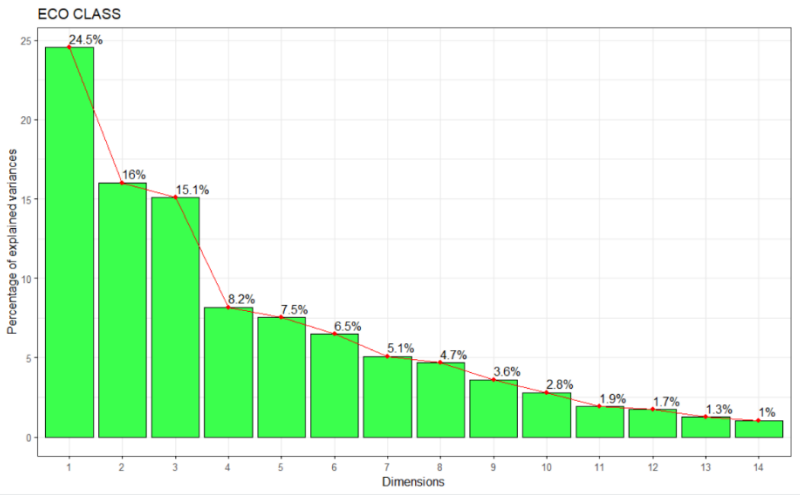
```
Bartlett's test: Xi^2(91) = 753659.7, p = 0
```

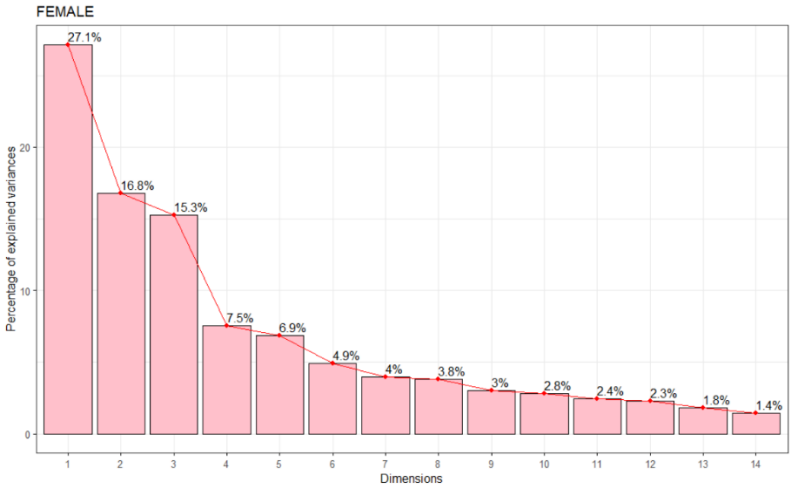
Результати показують, що значення $p < 0{,}001$ і є статистично значущим. PCA можна зробити. Цей тест не обов'язково виконувати, проте в деяких дослідженнях це доцільно робити. Звісно достатньо лише просто перевірити кореляційну матрицю і дізнатися чи корелюють змінні хоч в якійсь мірі.

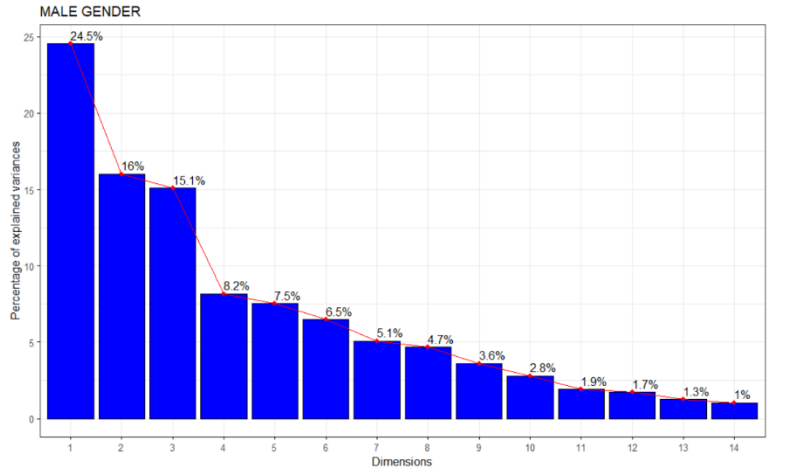
Слід також сказати, що PCA за все працює з неперервними змінними, проте в нас їхня кількість лише 4, тому ми були вимушені працювати лише з тим, що є. Також слід сказати, що для PCA дуже потрібна стандартизація даних, тому ми їх масштабуємо до стандартного відхилення 1 і середнього значення 0.

Також, через те, що задача полягає у виявленні нових залежностей, або підтвердженні старих гіпотез, а не в тому, щоб зменшити розмірність датасету, ми розбили датасет по фактору «класу» в якому пасажир подорожував та його статі.









Ми побудували scree plot, тобто графічне представлення власних значень головних компонент, за яким будемо визначати кількість головних компонент, які потрібно залишити для подальшого аналізу. Ми зробили два розбиття, які описані вище і побачили, що графіки для Male та Female майже не відрізняються, проте в економ та бізнес класах видно, що наприклад кумулятивну дисперсію на рівні 66% для бізнес класу описують три змінні, а в економ класі ту ж дисперсію описують 5 змінних. Саме тому, в цьому дослідженні ми виконували розбиття лише по **класу**.

Другим і важливим етапом є вибір оптимальної кількості компонент. Є велика кількість методів щоб визначити це число, деякі з них евристичні, наприклад elbow method (метод ліктя), інші ж базуються на серйозному підгрунті, проте в нашому дослідженні ми використаємо одразу декілька.

Звісно, в залежності від задач які ми переслідуємо, нам важливі різні показники, проте в даному дослідженні ми старалися якомога більше зберегти інформацію про змінні. Існує правило Кайзера, яке пропонує вибирати ті змінні, де власні числа перевищують одиницю, але це не завжди хороший спосіб.

Метод який ми вибрали – паралельний аналіз, один з найкращих способів емпірично оцінити кількість факторів, які треба зберегти. Спочатку ми генеруємо деяку кількість бутстрап вибірок даних на основі кількості початковог одатасету. Фактично це робиться для того, щоб позбавитися врахувати шум, який може з'явитися після генерації цих вибірок і впливати на кореляційні матриці. Далі обчислюємо середнє значення та 95% процентиль. В результаті виведемо таблицю, яка показує наскільки великими можуть бути власні значення в результаті простого використання випадково згенерованих наборів даних. Якщо наш початковий набір даних матиме власне значення, яке перевищує згенероване власне значення, то ми можемо взяти даний фактор для подальшого дослідження.

Parallel Analysis Results for Business

Method: pca

Number of variables: 14

Sample size: 62147

Number of correlation matrices: 300

Seed: 42

Percentile: 0.95

Component <dbl>	Mean <dbl>	0.95 <dbl>	num [1:14]
1	1.025	1.030	4.372
2	1.019	1.023	2.958
3	1.015	1.018	2.011
4	1.011	1.014	0.929
5	1.008	1.011	0.725
6	1.005	1.007	...
7	1.002	1.004	
8	0.998	1.001	
9	0.995	0.998	
10	0.992	0.995	

Parallel Analysis Results

Method: pca

Number of variables: 14

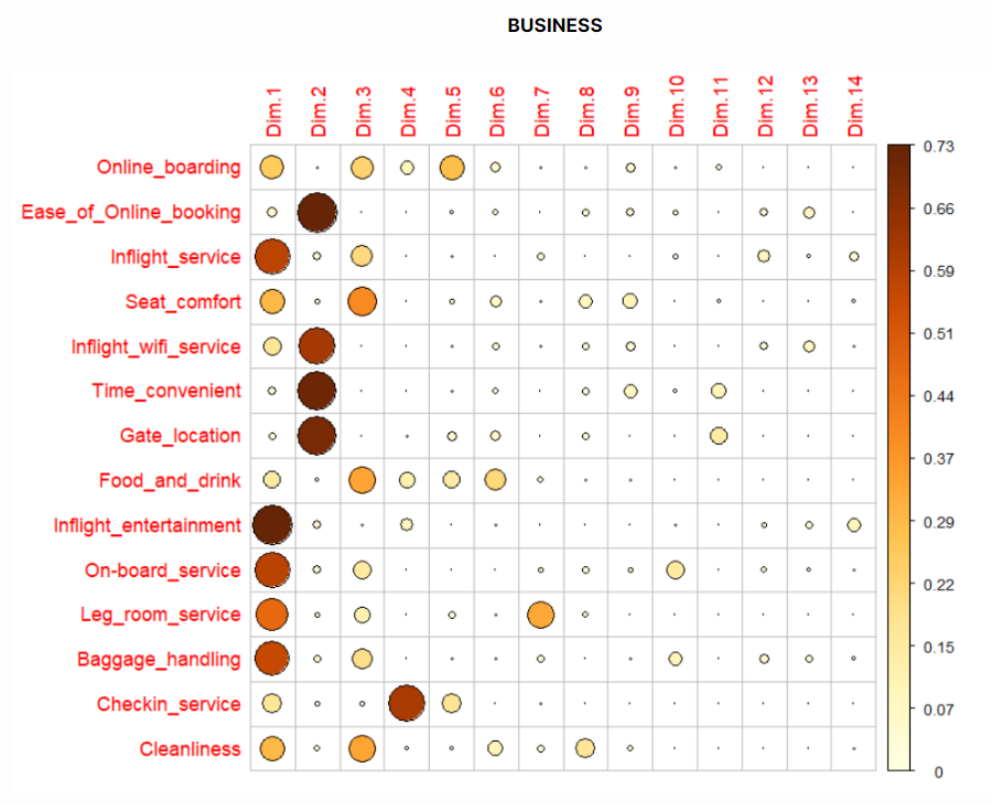
Sample size: 58293

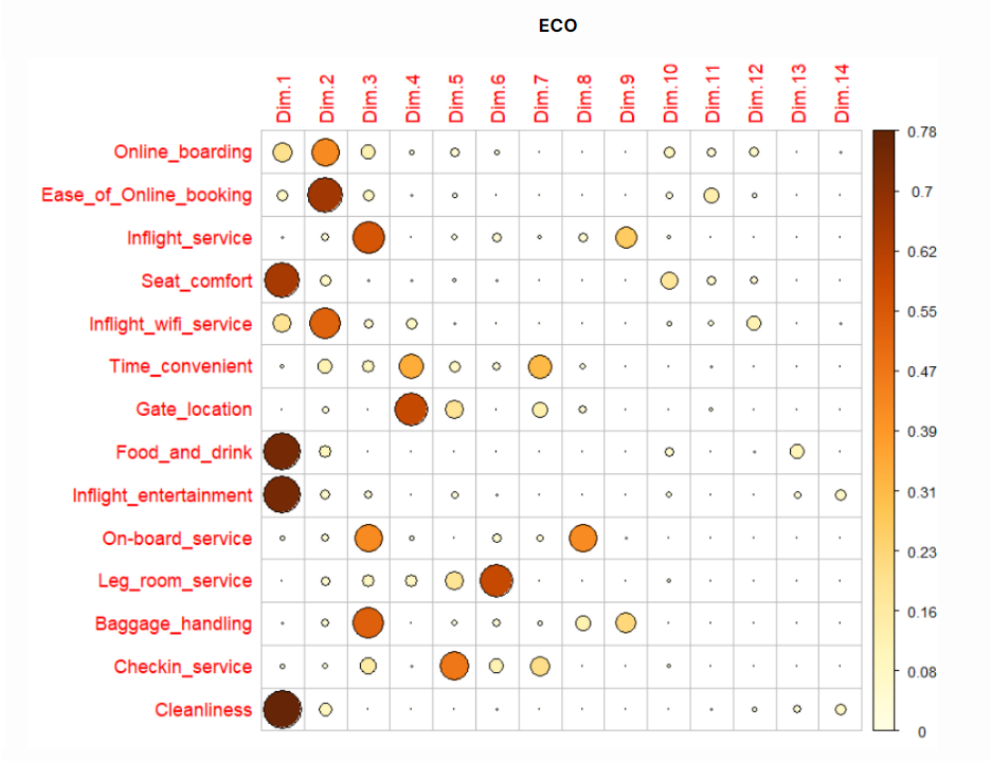
Number of correlation matrices: 300

Seed: 42

Percentile: 0.95

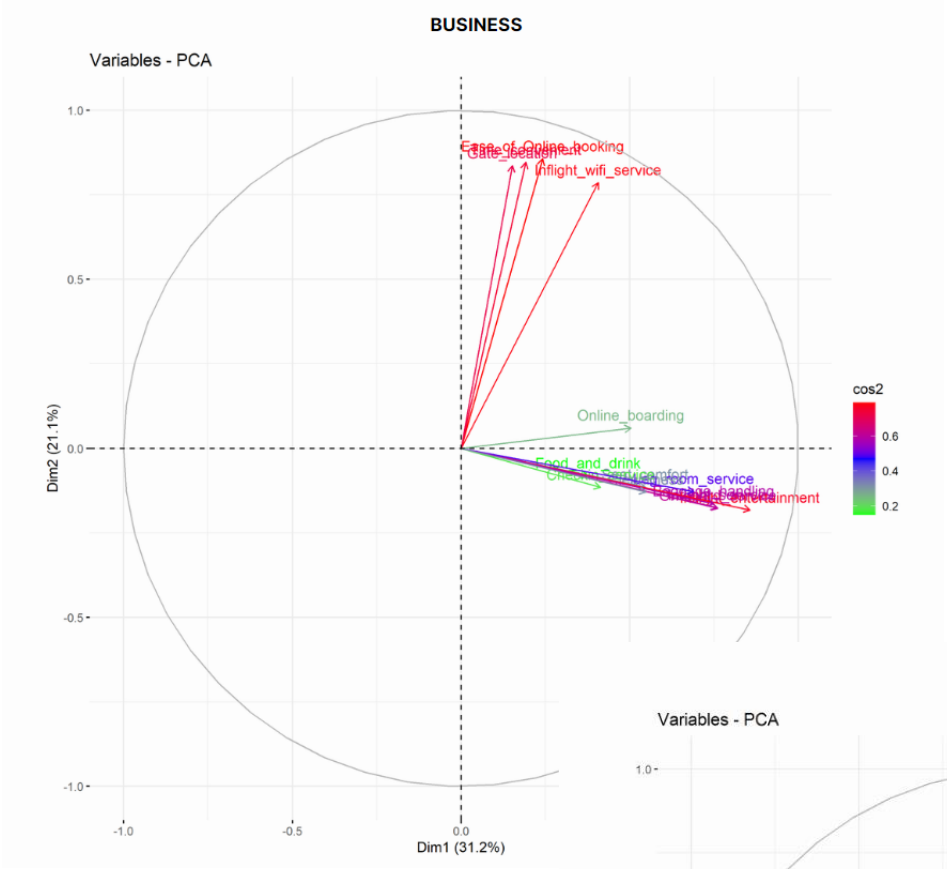
Component <dbl>	Mean <dbl>	0.95 <dbl>	num [1:14]
1	1.026	1.030	3.44
2	1.020	1.024	2.24
3	1.016	1.019	2.11
4	1.012	1.015	1.14
5	1.008	1.011	1.06
6	1.005	1.008	...
7	1.001	1.004	
8	0.998	1.001	
9	0.995	0.997	
10	0.992	0.994	

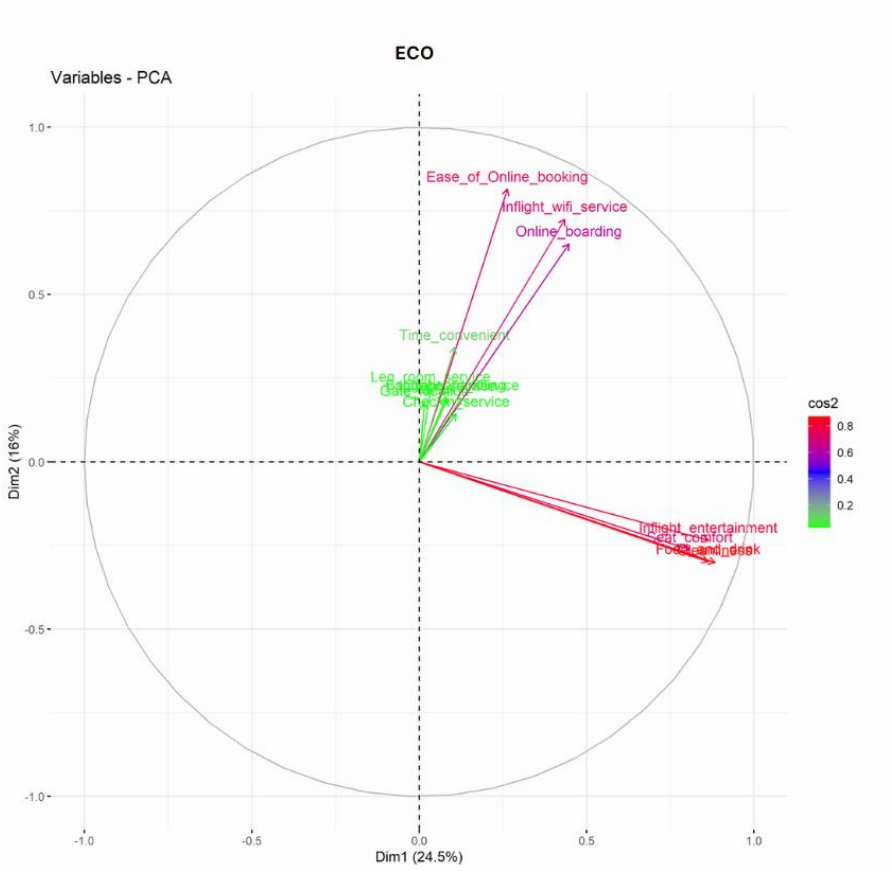




Подивимося на якість репрезентації. Фактично на графіку ми бачимо квадрати значень косинуса для кожної змінної, які показують наскільки змінна репрезентована кожним головним компонентом.

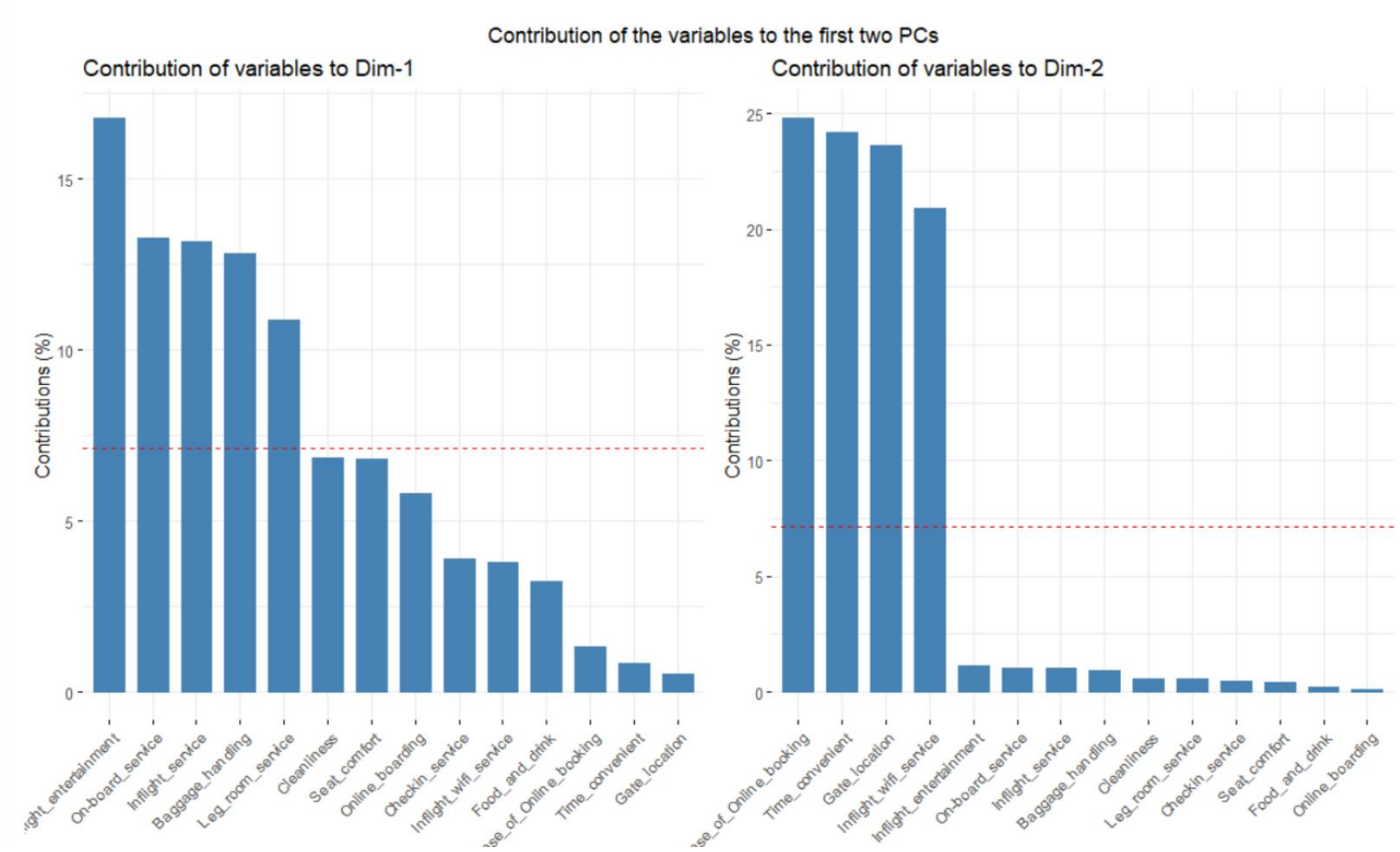
Тепер побудуємо графіки проекцій змінних на головні компоненти. Фактично, це графічне представлення, яке відображає внесок змінних у компоненти, їхню кореляцію один з одним та розподіл.

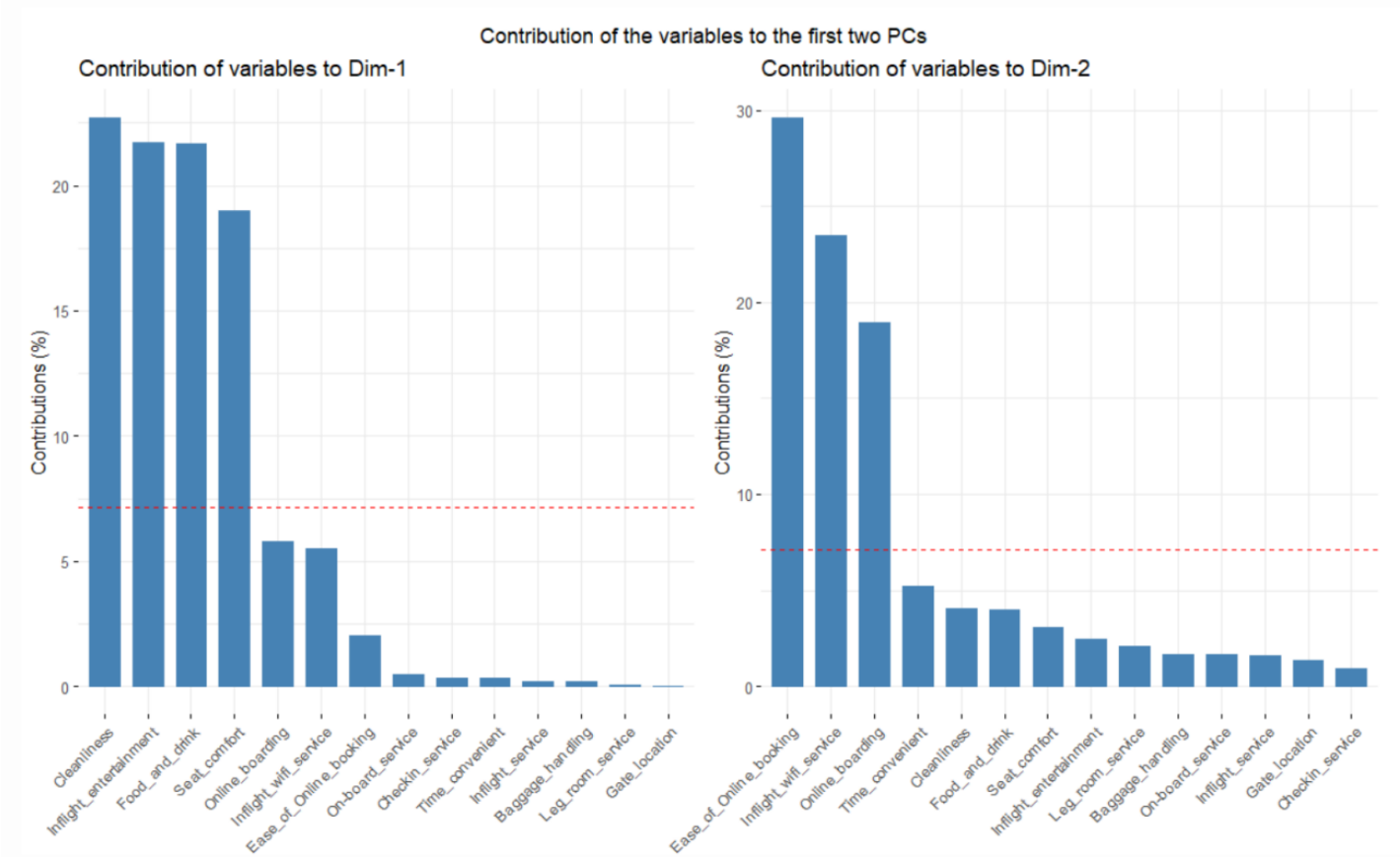




Позитивно корельовані змінні групуються разом, тоді як негативно корельовані змінні розташовані на протилежних сторонах початку початку графіка. Відстань між змінними та початком координат вимірює якість змінних на факторній карті. Змінні, які знаходяться далеко від початку координат, добре представлені на факторній карті. Бачимо, що Ease of online booking, inflight wifi service, Inflight entertainment розташовані дуже близько до окружності кореляційного кола, а це значить, що вони показують гарне представлення змінної на цих головних компонентах.

Далі графічно зобразимо тобто внесок змінних, або вплив (contribution) цих змінних у вигляді стовпчикових діаграм для кожних компонент. Червона пунктирна лінія на графіку вище вказує на очікуваний середній внесок (expected average contribution). Для певного компонента змінна з внеском, що перевищує цей контрольний показник, вважається важливою для внеску в компонент.





Щодо біплотів, то що ми не можемо зробити графік для всіх individuals, бо в нас дані містять інформацію про відгук конкретної людини і таких даних 120 тисяч, а групувати по якимось ознакам ми могли б по статі, або по цілі польоту (в нас є змінні Type of Travel), проте в такому випадку в нас буде лише дві групи і дві точки на графіку, тобто ми ніяк не можемо адекватно відобразити даний графік

Максимум це ми можемо взяти невелику вибірку, бо якщо ми візуалізуємо всі 120 тисяч даних то дуже одна велика купа точок. Фактично, ці графіки дуже добре себе показують на даних наприклад про гени, коли в нас є умовно 100 генів які ми досліджуємо і дуже багато змінних, які описують ці гени.

Список використаних джерел

1. «*Business Intelligence in Airline Passenger Satisfaction Study — A Fuzzy-Genetic Approach with Optimized Interpretability-Accuracy Trade-Off*» - Marian B. Gorzałczany, Filip Rudzi ´nsk, and Jakub Piekoszewski, Department of Electrical and Computer Engineering, Kielce University of Technology, Poland, 2021
2. «Investigating airline passenger satisfaction: Data mining method» - Tri Noviantoro, Jen-Peng Huang, College of Business, Southern Taiwan University of Science and Technology, Taiwan, 2022.
3. «*Feature Analysis on Airline Passenger Satisfaction using Orange Tool*» - Hannah Susan Mathew, Department of Computer Science, Rajagiri College of Social Sciences, Kochi, India, 2022.