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

Звіт
з лабораторної роботи № 1: «Розвідковий аналіз даних»
із дисципліни «Аналіз даних»

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

Керівник:
*ст. викладач Тавров Д. Ю.*

# 1 Організація роботи в команді

В даному розділі описаний основний підхід роботи в команді та висвітлені інші особливості організаційного процесу.

## 1.1 Загальна ідея

В команді на кожну лабораторну роботу призначається свій тімлід, який проектуватиме поточну роботу, ставитиме вимоги і обов'язки підопічним та перевіряти звітність. Крім того частиною роботи тімліда є організація комунікації в команді та проведення зустрічей. На першу лабораторну роботу призначений тімлідом - Гармаш О.Є.

## 1.2 Розподіл обов'язків

Команда складається з чотирьох осіб, кожній з яких делеговано наступні задачі:

- Збір, мердж, обробка та підготовка даних до процесу EDA – Максим
- Розвідковий аналіз даних та візуалізація – Олексій
- Аналіз та пояснення отриманих результатів, «видобуток» цінностей/користі проведеного EDA – Олексій і Максим
- Структуризація файлового каталогу, контроль версій (*github*), створення документації (*codebook*) – Назарій
- Оформлення дизайну команди, створення презентацій, звітів – Олексій
- Конвертація коду та презентації в R Markdown, перевірка та редагування графіків, презентацій та інших звітів відповідно поставленим вимогам – Назар

У кожного студента на поставлені вимоги додатково є помічник, для того, щоб виключити можливість помилок, а також щоб кожен студент мав змогу розібратися в кожному кроці.

## 1.3 Вимоги до завдань

*Завдання 1*

- Визначити дослідницькі питання
- Відповідно до обраних дослідницьких питань знайти датасет для мерджу (якщо потребується)
- Приведення даних в охайний формат, приведення типів змінних, перевірка одиниць виміру
- Робота з пропущеними даними та викидами
- Документація виконаних дій для codebook та збір використаних джерел для завдання 6

*Завдання 2*

- Створення інформативно корисних графіків відповідно дослідницькому питанню (одновимірні, двовимірні, тощо)
- Обчислення та візуалізація кореляції між змінними
- Візуалізація на одному графіку декількох змінних за допомогою кольорів форм, розмірів
- Логаритмування для ліпшої візуалізації
- Оформлення графіків відповідно вимогам

*Завдання 3*

- Опис дослідницького питання, у чому суть та цікавість, мотивація дослідження
- Пояснення інформаційної користі графіків, у чому ідея, як вони пов'язані з дослідницьким питанням
- Опис кінцевих результатів EDA: чи було знайдено відповіді на питання, які саме; які додаткові питання виникли

*Завдання 4*

- Перевірка коду на незалежній машині (онлайн ресурс)
- Відокремлення очищення даних та мерджу в окремі файли
- Створення Codebook за стандартами: інформація про кожну змінну у т.ч. одиниці виміру, пояснення процесу очищення даних, додаткова інформація пов'язана з особливостями збору відповідних даних
- Створення єдиного каталогу для всіх лабораторних робіт на GitHub
- Завантаження кожного окремого файлу окремим коммітом відповідно стандартам контролю версій
- Збір використаних джерел для завдання 6

*Завдання 5*

- PDF звіт відповідно вимогам
- PDF презентація відповідно вимогам
- Створення індивідуального дизайну команди
- Коментування коду

*Завдання 6*

- Перевірка якості та конвертація коду в R Markdown
- Перевірка та редагування графіків відповідно стандартам та вимогам
- Перевірка codebook, звіту та презентації відповідно кожній з вимог
- Формування списку використаних джерел для кожного з кроків
- Формування посилання на запозичені картинки

# 2 Мотивація дослідження

## 2.1 Вступ

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.

## 2.2 Дослідницькі питання та їх користь

В даному дослідженні була спроба знайти відповідь на наступні **дослідницькі питання**:

- Чи існує залежність задоволеності рейсів від вікових категорій клієнтів?

**Відповідна гіпотеза**: *Існують вікові категорії що мають меншу частку задоволеності рейсом за інші*. Поділивши пасажирів на вікові групи і дослідивши рівні задоволеності можливо вдасться виявити деякі вікові категорії, які менше задоволені рейсом за інші, відповідно якщо ми будемо знати ці категорії, авіакомпанії слід буде звернути увагу на них та дізнатися причину цієї проблеми, щоб в подальшому розв'язати «вікове питання». Наприклад, можливо літаки авіакомпанії погано обладнані для дітей і тому частка задоволених людей юного віку на 10 % менша за всі інші вікові групи. В такому разі ми зможемо дати конкретні дії для компанії, аби покращити рівень задоволеності клієнтів.

Відповідно досліджуючи дане питання слід звернути увагу на залежність не тількі загального рівня задоволеності від вікових груп, а і конкретних факторів, з цього випливає наступне дослідницьке питання.

- Чи існує залежність конкретних факторів задоволеності рейсів від вікових категорій клієнтів?

**Відповідна гіпотеза**: *Існують вікові категорії, для яких вплив деяких факторів задоволеності менший за інші.* Дана гіпотеза пов'язана з попередньою і підхід як до дослідження так і до аналізу результатів ідентичний – можливо ми знайдемо деякі фактори, задоволеність якими в деяких вікових групах менша ніж в інших. Наприклад існує признак «Онлайн-бронювання» і можливо задоволеність цим фактором у людей похилого віку менша ніж у інших груп, це може бути в силу як неспроможності людей похилого віку вдало користуватися інтернетом – так і незручним або інтуітивно не зрозумілим інтерфейсом сайту, на якому клієнти бронюють квитки, в такому випадку компанія має дослідити це питання більш детально та визначити корені проблеми.

- Чи існує залежність між задоволеністю клієнтами рейсом та цілями перельотів?

**Відповідна гіпотеза**: *Пасажири, що подорожували в персональних цілях залишаються менш задоволеними перельотами*. Відповівши на дане питання, можна зробити висновки про загальний рівень задоволеності клієнтів, що мають різні цілі перельотів і у разі, якщо клієнти якоїсь конкретної цілі матимуть значно менший рівень задоволеності слід визначити причини цієї проблеми, можливо це пов'язано з тим, що компанія менш фінансує обслуговування літаків пасажирських рейсів і більше зконцентрована на бізнес рейсах, проте в такому випадку слід визначити наскільки сильно страждає одна група і «виграє» інша група людей. Наприклад якщо компанія надає більше зусиль на користь бізнес рейсів, але сумарний дохід в більшості залежить від пасажирськів рейсів (бо їхня кількість більша), то в такому разі їхні старання не виправдані і слід звернути увагу на менш фінансований тип рейсів.

- Чи існує різниця впливу комфортабельності сидіння в залежності від дистанції рейсів?

**Відповідна гіпотеза**: *Комфортабельністю можна нехтувати при невеликих дистанціях рейсів.* Користь від відповіді на дане дослідницьке запитання лежить у гіпотезі. Якщо виявиться, що вплив комфортабельності сидінь на дистанціях середньої та великої довжини більший за вплив на дистанціях короткої довжини рейсів, то в такому разі можна буде в якійсь мірі нехтувати, наприклад, якістю сидінь в нових літаках, якщо ми знаємо що вони будуть літати короткими дистанціями. Причиною може бути сукупність людських факторів і часу, наприклад якщо рейс триває лише годину-дві, то людині легше висідити цей час на поганому сидінні, на відміну від рейсів 9-10 годинної тривалості на тому ж поганому сидінні. В такому разі краще залишити якість сидінь на невеликих дистанціях минулу, а звернути увагу на інших більш тривалий тип рейсів.

- Які признаки мають найбільший/найменший вплив на задоволеність клієнтів бізнес/економ класу?

**Відповідна гіпотеза**: *Існують фактори, які по-різному впливають на людей, подорожуючих різними класами*. Користь для авіакомпанії може бути в зменшенні фінансування найменш корелюючих факторів для клієнтів економ класу і за рахунок цього збільшення фінансування найбільш корелюючих факторів для клієнтів бізнес класу. Наприклад, можна уявити ситуацію, що існує фактор «задоволеність TV» і даний фактор є найменш корельованим з цільовою змінною для клієнтів економ класу, це може бути пов'язано з тим, що люди надають перевагу користуванню власними гаджетами, аніж вбудованими телевізорами в літаках; у той же час нехай ми визначили що найбільш корельованим фактором з цільовою змінною для пасажирів бізнес класу є якість Wi-Fi, бо дані клієнти їдуть рейсом з робочих цілей і можливо їх більш важливо мати гарний зв'язок аніж людям, що подорожують в економ класі. Таким чином умовно прибравши телевізори з рейсів економ класу ми можемо встановити системи які забезпечуть краще з'єднання з інтернетом.

# 3 Дані

## 3.1 Опис та походження даних

Набір даних, який використовується в даному дослідженні, отримано з Kaggle та містить дані про результати опитування проведеного авіакомпаніями щодо рівня задоволеності пасажирів/клієнтів на основі різних факторів. Набір складається з 23 стовпців, таких як вік, стать, клас подорожі, затримки прибуття та відправлення, а також фактори, які впливають на рівень задоволеності клієнтів, наприклад обслуговування на борту, чистота, комфорт сидінь, обробка багажу тощо. Цільовою функцією в датасеті є стовпчик під назвою *Satisfaction,* який описує **загальний** рівень задоволеності клієнтів. Цільова функція має два класи – задоволений і незадоволений клієнт.

Дані є реальними результатами опитувань клієнтів деякої авіакомпанії США в 2015 році, всі стовпці є анонімними, тобто особисті дані опитуваних не використовуються, крім того невідомо яка саме компанія проводила це опитування, щоб не ризикувати репутацією.

## 3.2 Основні характеристики даних

**Розмір набору даних**: 129880 рядків (13.5 MB)

**Кількість змінних**: 23

**Кількість пропущених даних**: 393 рядки

**Кількість стовпців, що мають пропущені дані**: 1 стовпець

Датасет має 4 дійсні числові змінні, всі інші **категоріальні**.

**Дескриптивні статистики по кожній з дійсних змінни**х:

```
Age                          Min.: 7.00  Median: 40.00   Mean: 39.43   Max.: 85.00
Flight_Distance              Min.: 56    Median: 844     Mean: 1190    Max.: 4983
Departure_Delay_in_Minutes   Min.: 0     Median: 0       Mean: 14      Max.: 652.00
Arrival_Delay_in_Minutes     Min.: 0     Median: 0       Mean: 15.02   Max.: 638.00
Delay_overtake               Min.: -54   Median: 0       Mean: 0.4479  Max.: 234.0000
```

**Кількість викидів***: Departure Delay* - 20; *Flight Distance* – 11

Більш детальну інформацію про характеристики датасету можна знайти в codebook.

## 3.3 Підготовка та очищення даних

### 3.3.1 Конкатенація та кодування

Дані на Kaggle поділені на тренувальну та тестову вибірки *airline_satisfaction.csv* та *airline_satisfaction2.csv* відповідно. В цілях дослідження лише аналіз даних тому першим чином ці датасети конкатенуємо в один.

Наступним кроком приведемо назви колонок в зручний формат, замінивши пробіли на символ нижнього підкреслення. Крім того приведемо перші літери міток класів змінних *Type_of_Travel* та *Customer_Type* до верхнього регістру: *Business travel -> Business Travel; disloyal Customer -> Disloyal Customer.*

Цільова змінна *Satisfaction* має бінарний клас, тому закодуємо категорії *satisfied* та *neutral or dissatisfied* на 1 та 0 відповідно.

### 3.3.2 Обробка пропущених даних

Пропущені дані має лише змінна *Arrival_Delay_in_Minutes*, їхня кількість 393, а у відсотковому співвідношенні до розміру датасету дорівнює 0.3 %, що критично мало.

Дана дійсна змінна означає час затримки літака до прибуття в пункт призначення і в такому випадку ми маємо два очевидних варіанти обробки пропущених даних:

- Заповнити *N/A* статистичними характеристиками (модою/медіаною) відповідного стовпця
- Заповнити *N/A* значеннями зі змінної *Departure_Delay_in_Minutes*

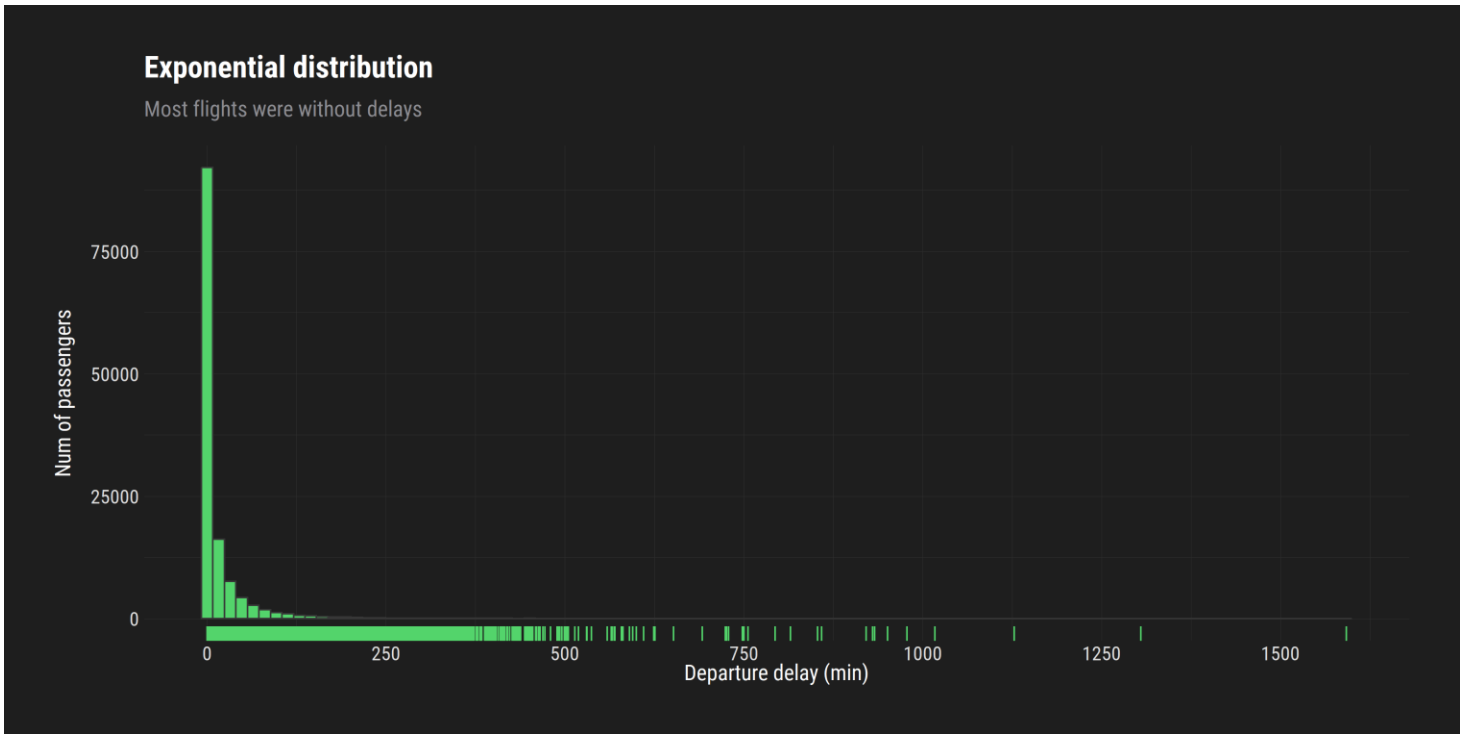
Медіана стовпця *Arrival_Delay_in_Minutes* дорівнює 7 хв, проте якщо ми заповнимо медіанним значенням пропущені дані, це може створити, не існуючі раніше, залежності, що приведе до неправильних результатих аналізу. Справді, якщо в нас затримка вильоту літака скажімо 120 хвилин, то складно повірити в те, що літак наздожене в повітрі 113 хв і запізниться до місця призначення лише на 7 хвилин, скоріш за все якщо літак спізнився при вильоті на 120 хвилин, то і прибуде він на 120 хвилин пізніше, тому доцільно використати **другий метод** та заповнити пропущенні значення *Arrival_Delay_in_Minutes* значеннями з *Departure_Delay_In_Minutes*.

Крім того, слід перевірити кореляцію між змінною затримки часу до вильоту та перед прибуттям, бо в разі кореляції рівної 1 матимемо справу з проблемою мультиколінеарності і в такому разі слід буде видалити один з рядків. Проте кореляція хоч і велика (0.965), але не дорівнює одиниці.

### 3.3.3 Виявлення викидів

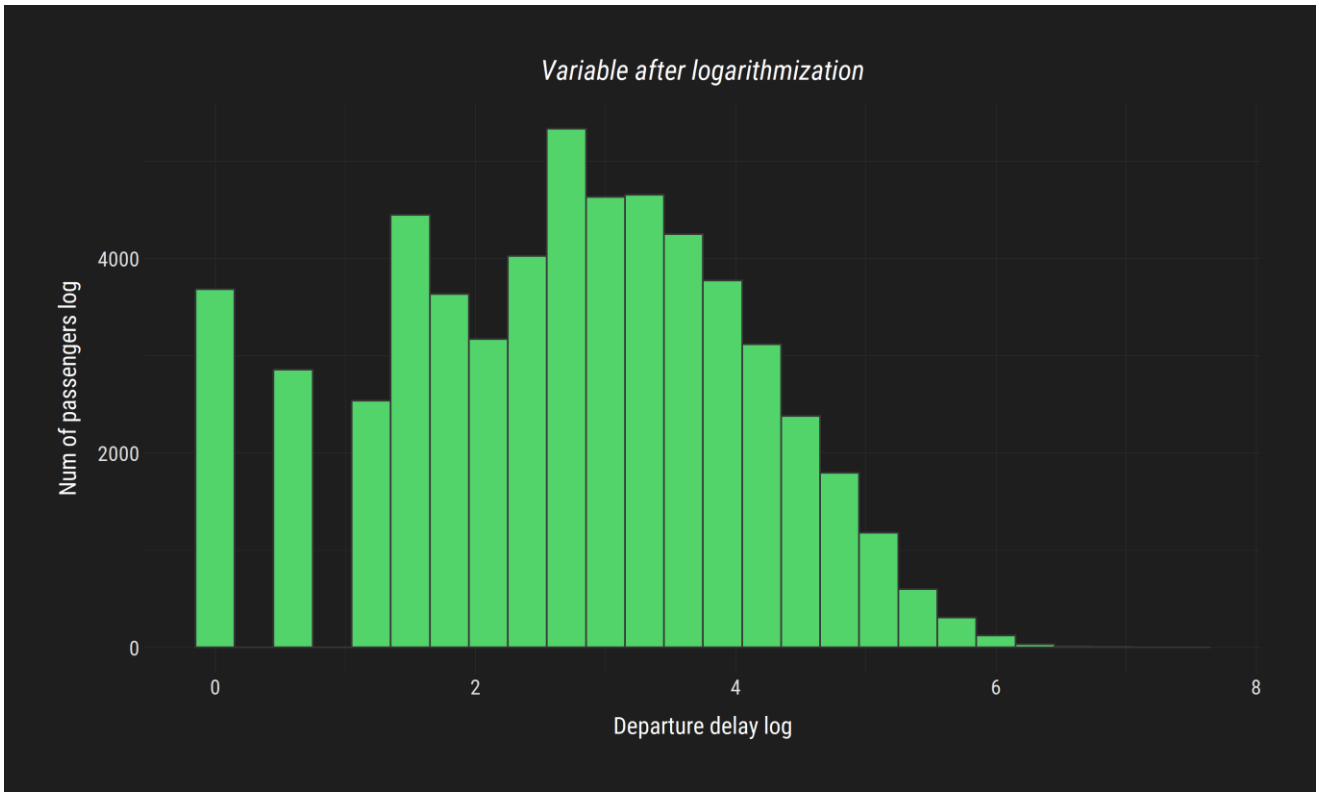
Категоріальні змінні представлені в даному датасеті фактично не мають можливості містити викидів, тому слід проаналізувати лише числові змінні.

*Departure_Delay_in_Minutes (Arrival_Delay_in_Minutes)*



На гістограмі ми чітко бачимо що більшість рейсів мають затримку в 0 хвилин, але ці значення заважають побачити картину для рейсів, що мали затримку. Крім того вісь *x* підказує, що є деякі значення (можливо викиди), що дуже звужують графік. Очевидно розподіл нагадує експоненційний, тому для того, щоб побачити повноцінну картину цієї змінної - **логаритмуємо** її.

**відомо що логаритм від нуля дорівнює нескінченності тому в нас є два варіанти - або відкинути значення рівні нулеві, або підняти вісь x на 1, виберемо другий*

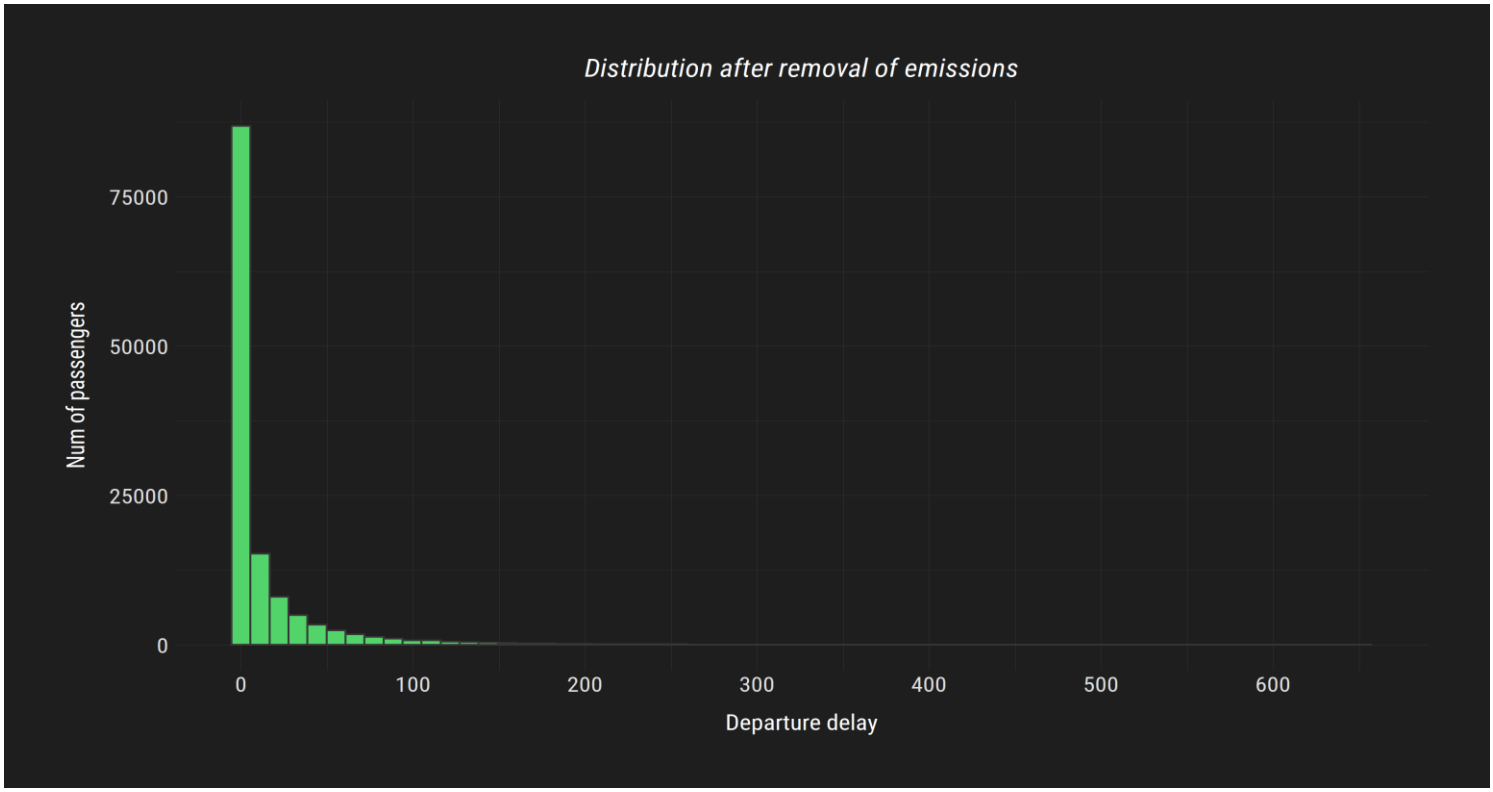



Дивлячись на графік логаритмованої змінної видно, що, незважаючи на нульові значення *(ln(1) == 0)*, розподіл виглядає нормальним. Але значення що більше 0 та менше 2 псують картину і здається що вони є викидами. Перевіримо це застосувавши IQR метод для визначення викидів.

```
Q1:  1.792 Q3:  3.689
Lower range:  -1.054 Upper range:  6.535
```

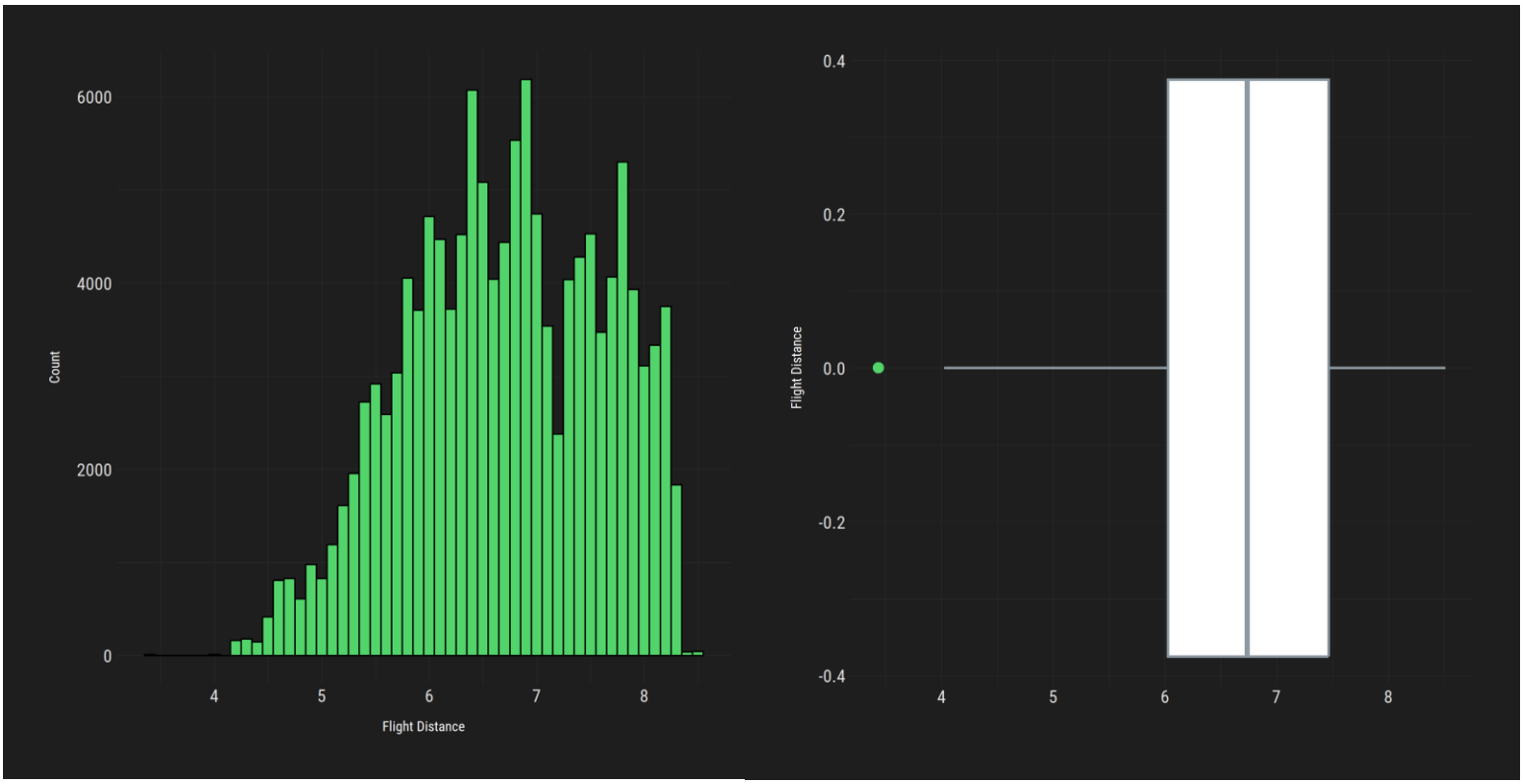
Аналізуючи отримані величини границь зрозуміло що значень затримок рейсів менше за 0 немає, тому сенсу обрізати датасет вище нижньої межи теж не існує. Проте верхня межа має значення 6.535. Всі значення що йдуть вище фактично псували нам початковий графік. Кількість значень що перевищують **верхню межу**: 20.

Хоча рейси і можуть затримуватися іноді на 1000 хвилин, але таке відбувається дуже рідко. Відносно величини датасету ця оцінка статистично незначуща, тому видалимо дані значення для коректності подальшого дослідження і повернемося до початкової системи координат.




Можна спостерігати, що по вісі *Departure delay* графік обрізався і через це став комфортнішим для візуального аналізу.

За аналогією до *Departure_Delay_in_Minutes* застосуємо IQR метод до *Flight_Distance:*




```
Q1:  6.026 Q3:  7.464
Lower range:  3.869 Upper range:  9.621
```

За результатами виявлено 11 викидів, кожний з яких менший за нижню границю. Продивимося ці дані:

| ...1 <dbl> | Gender <chr> | Customer_Type <chr> | Age <dbl> | Type_of_Travel <chr> | Class <chr> | Flight_Distance <dbl> |
|---|---|---|---|---|---|---|
| 29816 | Female | Loyal Customer | 38 | Business Travel | Eco | 3.433987 |
| 29824 | Female | Disloyal Customer | 23 | Business Travel | Eco | 3.433987 |
| 29863 | Female | Loyal Customer | 53 | Business Travel | Eco | 3.433987 |
| 29992 | Female | Disloyal Customer | 26 | Business Travel | Eco | 3.433987 |
| 30078 | Female | Loyal Customer | 22 | Personal Travel | Eco | 3.433987 |
| 30125 | Female | Loyal Customer | 54 | Personal Travel | Eco | 3.433987 |
| 30130 | Female | Loyal Customer | 12 | Personal Travel | Eco | 3.433987 |
| 30133 | Male | Loyal Customer | 70 | Personal Travel | Eco | 3.433987 |
| 30144 | Female | Loyal Customer | 17 | Personal Travel | Eco | 3.433987 |
| 30184 | Male | Loyal Customer | 43 | Business Travel | Eco Plus | 3.433987 |

1-10 of 11 rows | 1-10 of 25 columns

Викиди мають однакове *Flight_Distance* у всіх 11 випадках, це свідчить що такий рейс реально був, але слід припустити, що відповідна відстань могла бути записана неправильно, бо *exp(3.433987)* приблизно дорівнює 30 км, а це дуже мало для рейсу пасажирського літака. Тому видалимо дані викиди.

# 4 Розвідковий аналіз даних

## 4.1 Підхід та особливості аналізу

В даному дослідженні підійдемо до розвідкового аналізу з двох боків:

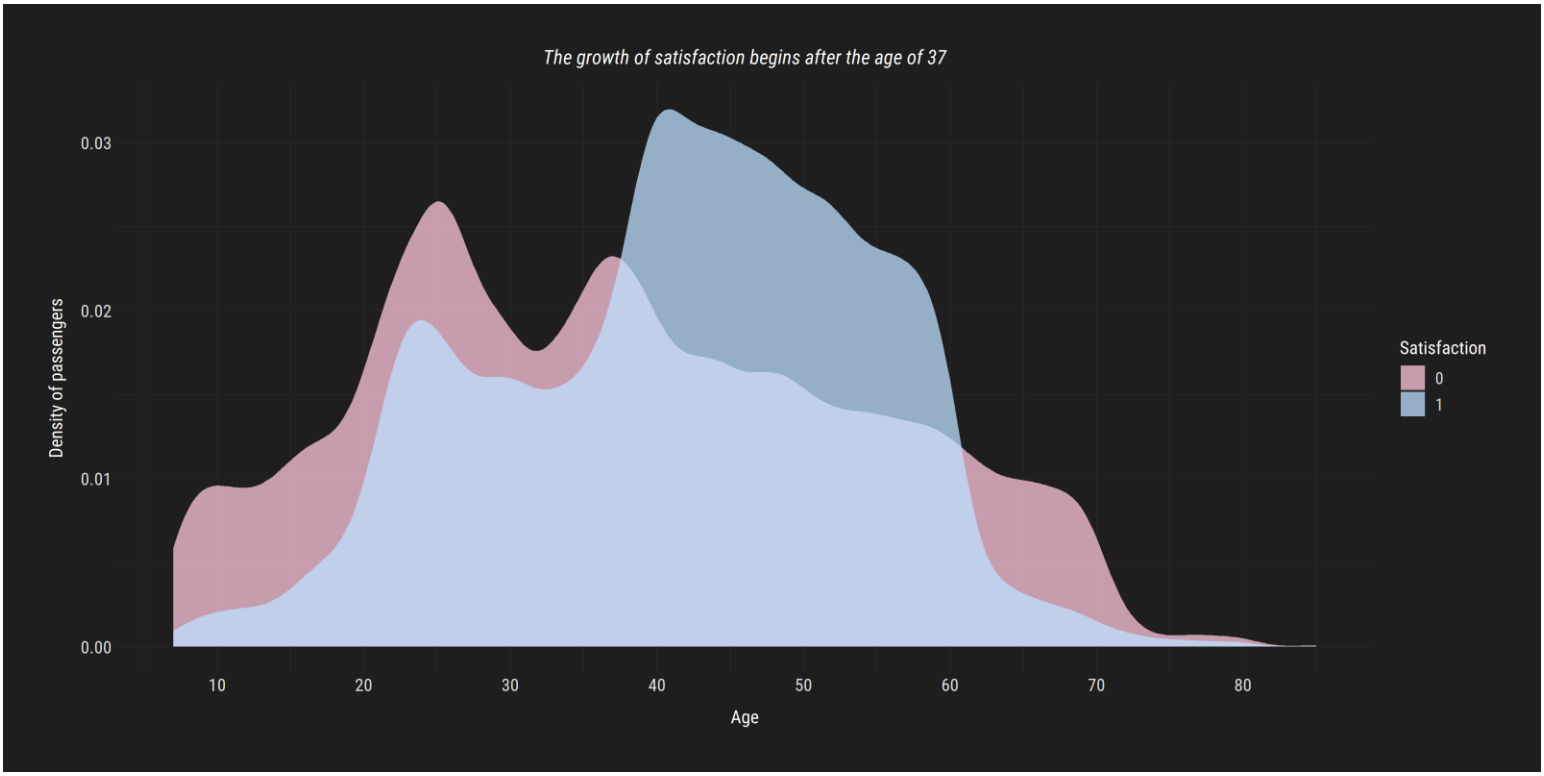
- Пройдемося по кожній змінній та спробуємо за допомогою візуалізації витягнути цікаву інформацію з даних або ж нові дослідницькі питання
- Опрацюємо та проаналізуємо дослідницькі питання, які були складені заздалегідь

## 4.2 Дослідження гіпотез

**Дослідницьке питання**: *Чи існує залежність задоволеності рейсів від вікових категорій клієнтів?*

**Гіпотеза**: *Існують вікові категорії, що мають меншу частку задоволеності рейсом за інші*

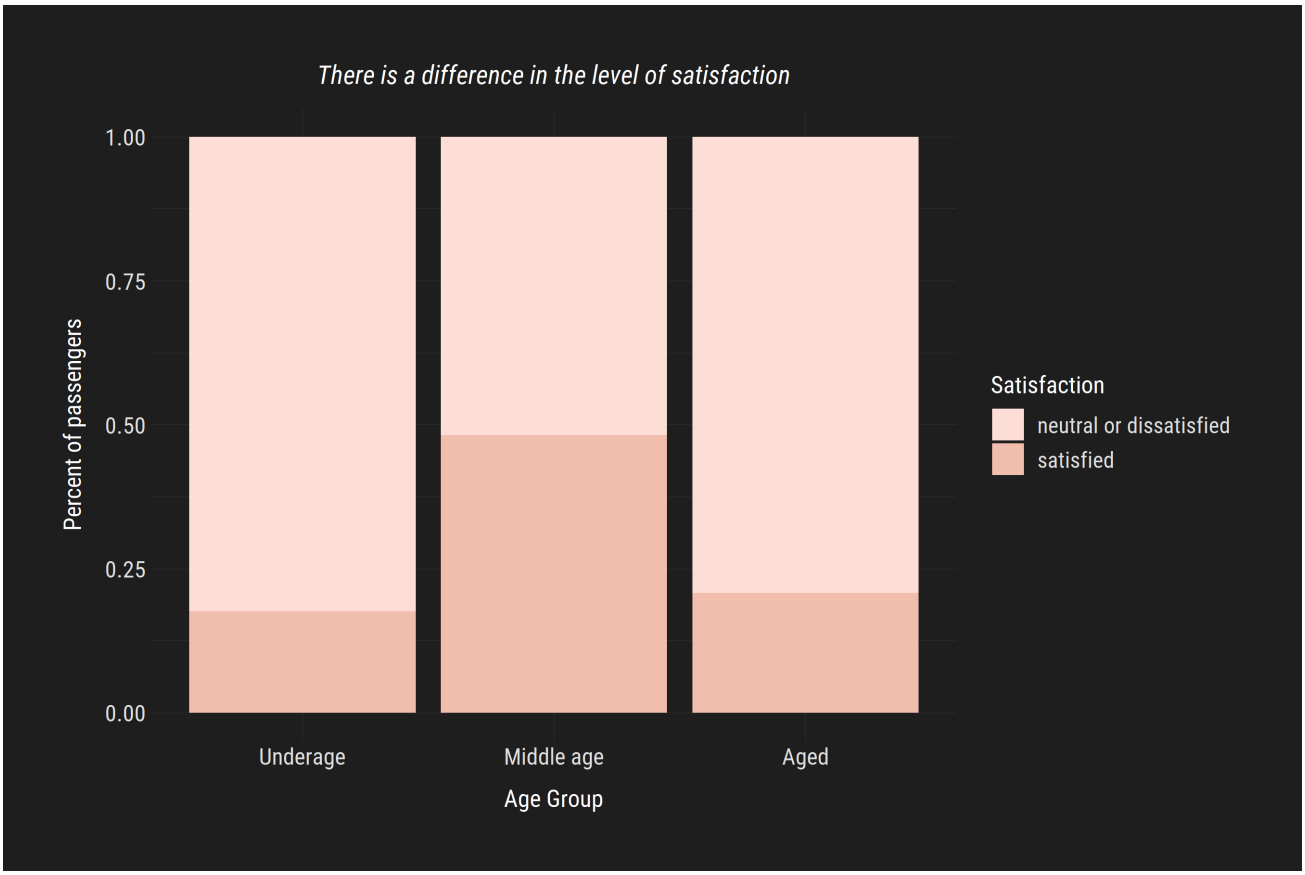
Побудуємо графік щільності розподілу віку пасажирів по двох категоріях: задоволений та незадоволений:




За графіком можна побачити, що частка задоволених пасажирів стрімко збільшується після умовної межі в 35 років.

Також пасажири в межах 18-60 років **сумарно** мають приблизно однаковий рівень задоволеності, в той же час можна побачити два «хвости» менше 18 років та більше 60 років де пасажири **сумарно** здебільшого незадоволені.

Поділимо датасет на вікові групи, створивши окрему змінну: *Underage (<18), Middle Age (18-60), Aged (>80)*, та побудуємо *barplot*, щоб переконатися у правильному **візуальному** висновку:




Подивимося на конкретні значення для пасажирів в межах 18-60 років:

```
Satisfied (18 < age < 60):  51163
Dissatisfied (18 < age < 60):  55407

Satisfied / amount of passengers:  48.01 %
```

Візуальні оцінки в першому випадку підтвердилися, пасажири від 18 до 60 років мають 48.01% задоволеності.

```
Satisfied (age < 18 & age > 60):  5256
Dissatisfied (age < 18 & age > 60):  18023

Satisfied / amount of passengers:  22.58 %
```

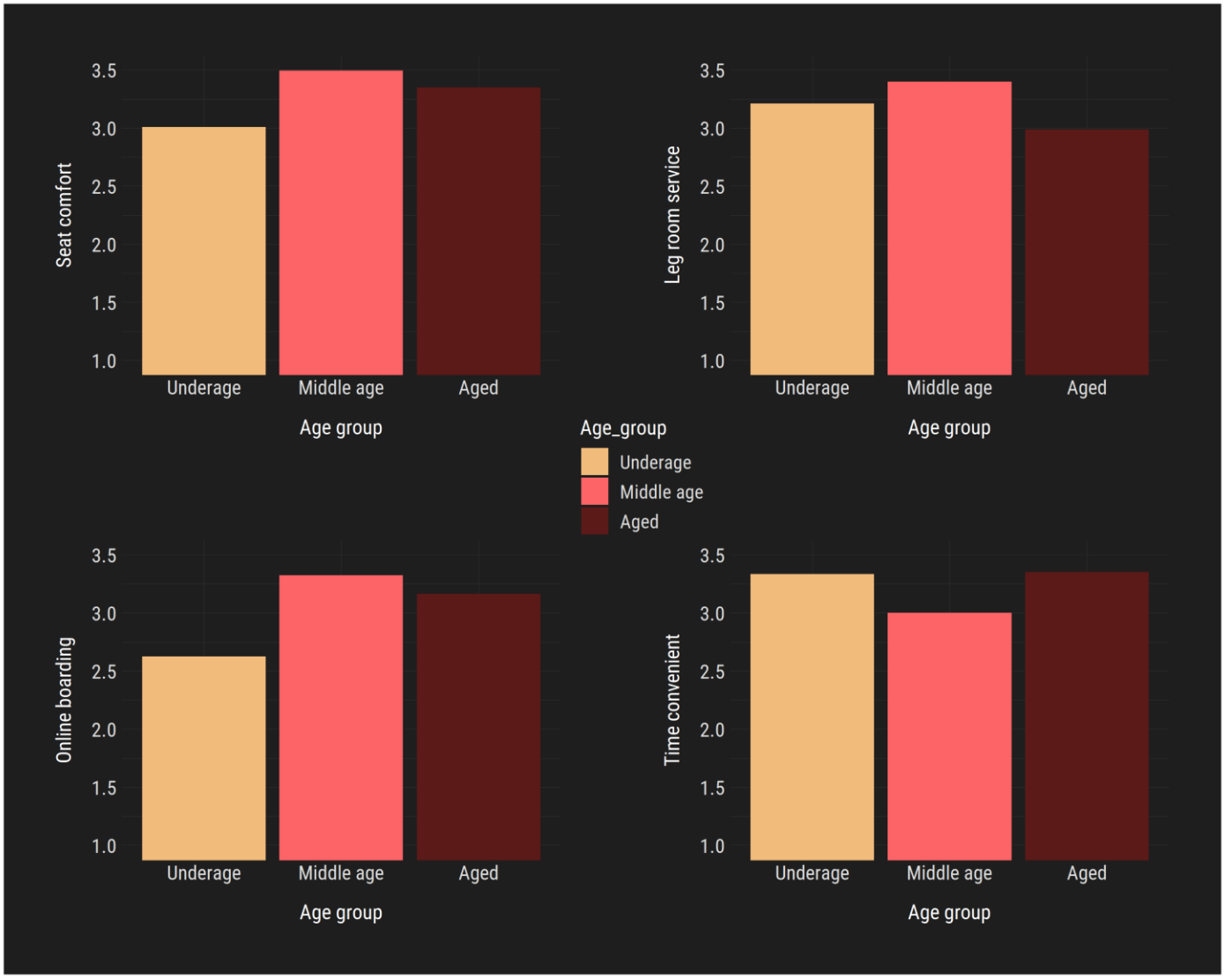
**Висновок:**

На відміну від минулої вікової групи, діти та люди похилого віку мають 22.58% задоволеності. Враховуючи попередній аналіз і висновки про націленість компанії на бізнес-мандрівки можна припустити, що компанія менше уваги приділяє звичайним пасажирським рейсам, якими літають діти та люди похилого віку, бо скоріш за все діти та люди похилого віку не літають на бізнес зустрічі. **Гіпотеза підтверджена**.

**Дослідницьке питання**: *Чи існує залежність конкретних факторів задоволеності рейсів від вікових категорій клієнтів?*

**Гіпотеза**: *Існують вікові категорії, для яких вплив деяких факторів задоволеності менший за інші.*

Виділимо ті змінні, де можливо наймовірніше помітна різниця між рівнями задоволеності для трьох вікових груп:

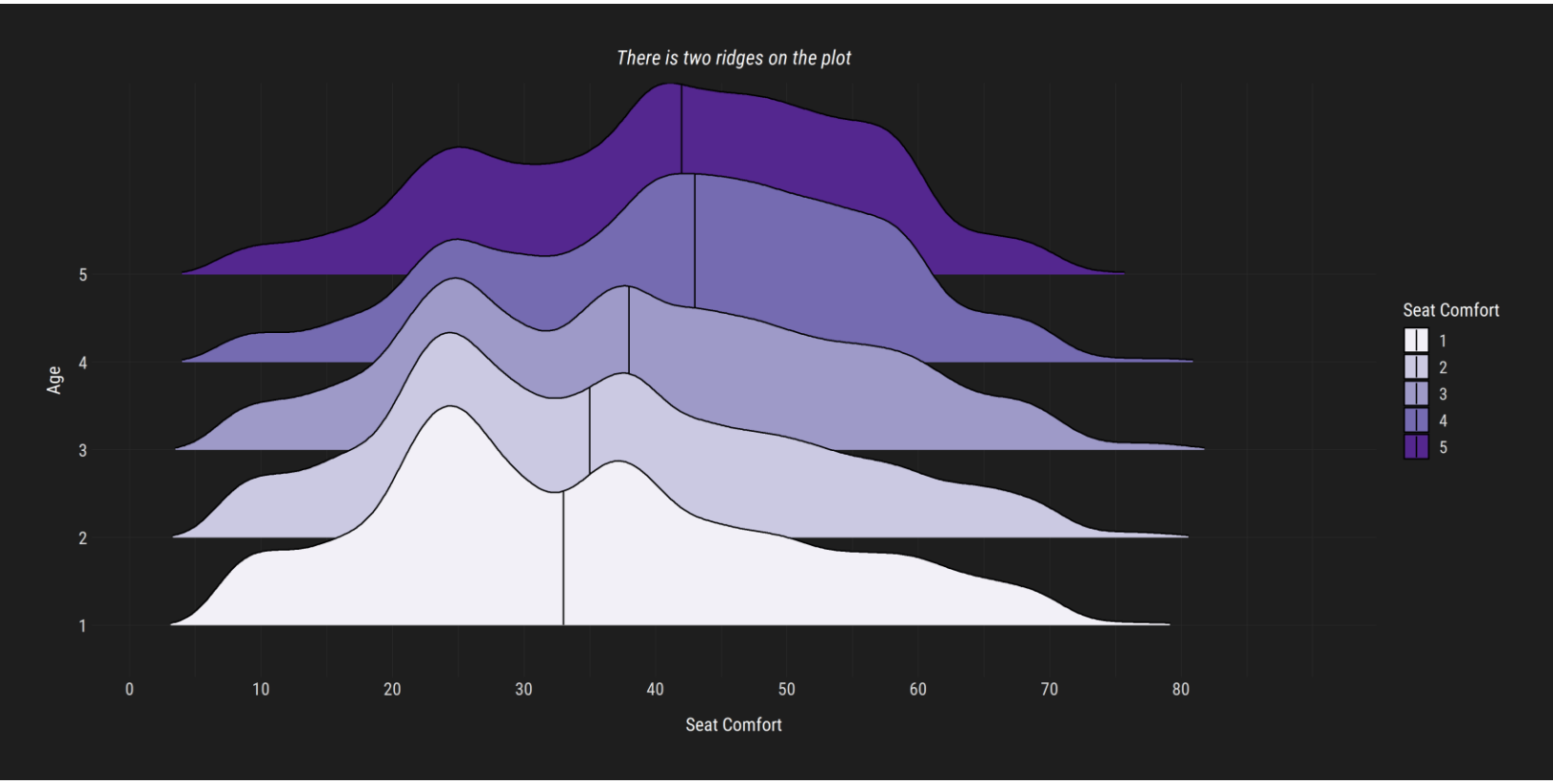


На графіку зображені чотири змінні, в яких присутня помітна різниця між рівнями задоволеності вікових груп, всі інші змінні мають приблизно однаковий рівень задоволеності серед усіх вікових груп.

**Seat comfort**

Середня задоволеність даним фактором найбільша у середньої вікової групи (3.5), трохи менша у людей похилого віку (3.4), а діти взагалі мають середню задоволеність комфортом сидінь на рівні 3. Така різниця може бути в силу непосидючості у дітей та обмежених фізичних можливостей у людей похилого віку, або ж загальною втомленістю, бо зрозуміло що людям 18-60 років легше витерпіти незручні сидіння або довготривалі рейси. Проте різниця значна, майже 0.5 і авіакомпанії слід детальніше вивчити це питання, можливо проблема лежить в тому, що літаки не достатньо добре облаштовані зручними сидіннями для дітей.

Побудуємо ridgeplot для цього фактору:




За графіком чітко видно два умовних горби, при чому лівий горб залишається незмінним відносно віку пасажирів, а правий горб зміщується зі збільшенням оцінки. Крім того, можна побачити як зміщується медіана в бік збільшення віку. Спостерігати можна також і збільшення частки задоволених пасажирів починаючи від 40 років. Також, що кількість незадоволених (оцінка 1-3) від 20 до 30 років значно більша, аніж кількість задоволених (оцінка 4-5), проте у людей 40-60 років очевидно, що ситуація зворотня.

**Leg room service**

Простором для ніг задоволена більше знову ж таки середня вікова група та **діти**, а середня задоволенім даним фактором людей похилого віку нижча за 3. Тут все не однозначно, можна згадати минулий фактор (Seat comfort) і там оцінка задоволеності у дітей була нижча за людей похилого віку, ми це списали в більшості на вплив віку, **але** відомо що дітям для зручності треба менше простору для ніг ніж людям похилого віку, звідси випливає (враховуючи минулий фактор) що таку низьку оцінку люди похилого віку поставили не через вікові проблеми, а реальну проблему в комфорту сидінь та простору для ніг в літаку.

**Online Boarding**

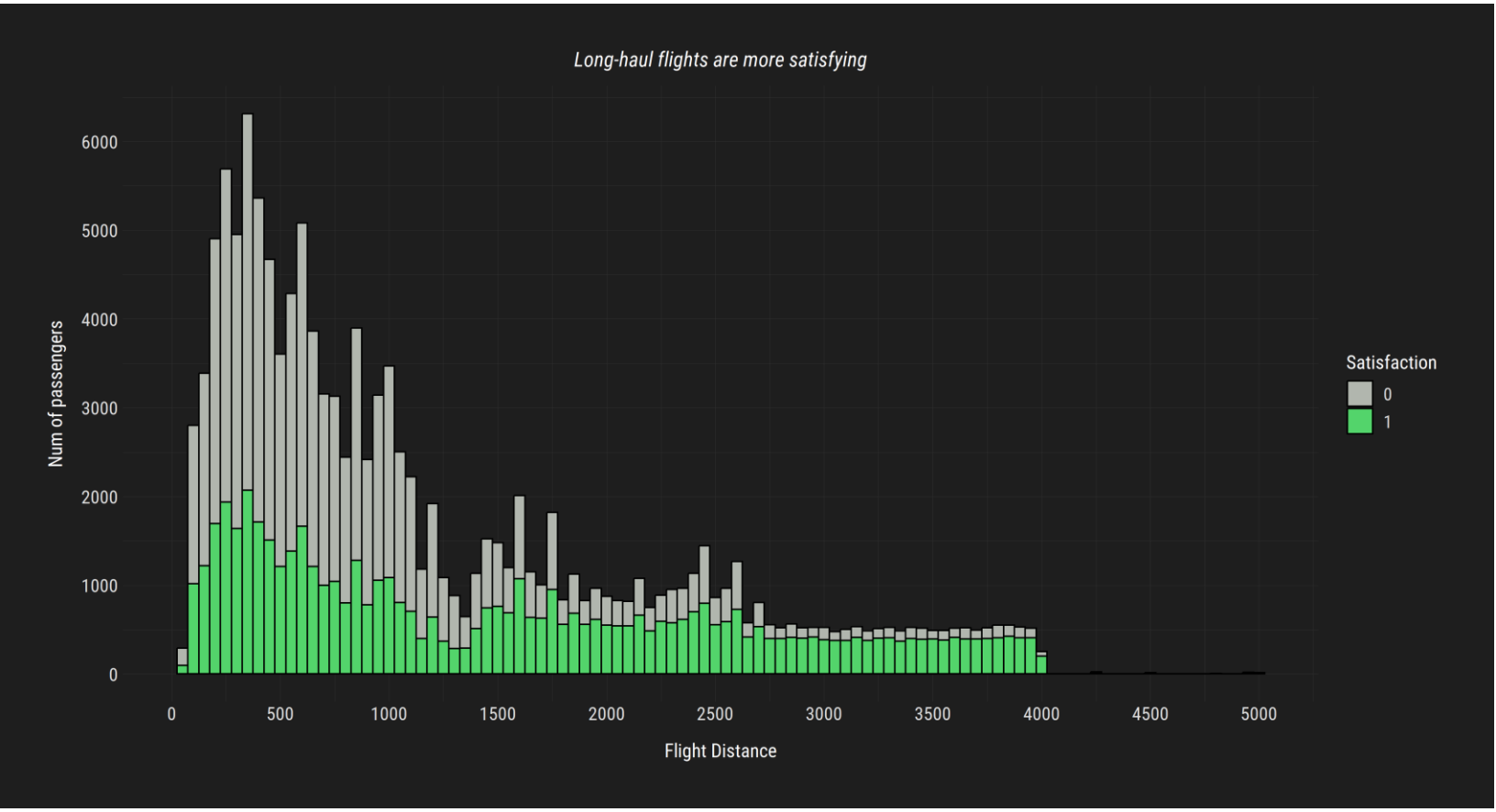
Середня задоволеність даним фактором найменша серед дітей, проте це можна пояснити тим, що діти здебільшого не мають можливості самостійно пройти реєстрацію на авіарейси, тому скоріш за все під час опитування діти ставили оцінки 0, якщо не знали що відповісти, або такі оцінки, які скажуть їм батьки. Проте є деяка різниця між задоволеністю онлайн-реєстрацією на рейс людьми похилого віку і середньої вікової групи, це може бути пов'язане з тим, що старші люди, як правило менш вдало користуються інтернетом, **або проблема лежить** в незручному або інтуітивно не зрозумілому (для людей 60+ років) інтерфейсі сайту авіакомпанії, на якому клієнти реєструються на рейси.

**Time convenient**

Середня задоволеність даним фактором відрізняється від раніше перечислених, найменшу середню задоволеність мають люди середньої вікової групи, приблизно на рівні 3, в той час як діти та люди похилого віку в середньому ставлять оцінку задоволеності часовим затримкам рейсу на рівні 3.3. Тут все очевидно, людям середньої вікової групи частіше треба літати в бізнес/робочих цілях на відміну від дітей та людей похилого віку, які з більшою ймовірністю літають з туристичними/персональними цілями. Це можна пояснити тим, що якщо людина літає в бізнес цілях, то можливо вона летить на мовну зустріч і затримка рейсу більше шкодить її задоволеності даним фактором і навіть затримка в 20 хвилин може дуже вплинути на оцінку, на відміну від людей, що летять в туристичних цілях і не мають необхідності бути на місці в конкретно зазначений час.

**Flight Distance**

Побудуємо гістограму розподілу для даної змінної та виділимо на ній частки задоволених та незадоволених людей:




За графіком спостерігаємо, що авіаперельоти можна поділити на три категорії за рівнем задоволеності:

- 0-1500 km (візуально кількість незадоволених пасажирів майже вдвічі більша ніж задоволених)
- 1500-2500 km (спостерігається протилежні ситуація, кількість задоволених тепер приблизно вдвчі більша)
- 2500+ km (на цій ділянці можна помітити дуже велику різницю між задоволеними та незадоволеними)

Створимо окремо змінну Flight_haul, що буде відносити запис до однієї з категорій Short, Medium, Long, та подивимося на конкретні числа:

```
Short -haul flight
Satisfied:  30434
Neutral or dissatisfied:  60073
Satisfied / amount passenger of short haul flight:  0.34

Medium -haul flight
Satisfied:  13243
Neutral or dissatisfied:  8758
Satisfied / amount passenger of short haul flight:  0.6

Long -haul flight
Satisfied:  12742
Neutral or dissatisfied:  4599
Satisfied / amount passenger of short haul flight:  0.73
```

Візуальні оцінки підтвердилися, можемо спостерігати що короткотривалі рейси мають лише 34% задоволеності на відміну від довготривалих рейсів, де цей показник становить 73%. Це може означати про націленість компанії на авіаперельоти окремої дистанції, тобто авіакомпанія приділяє більше уваги рейсам, що мають велику відстань польоту. Поглянемо на кількість здійснених перельотів кожної з груп:

```
Num of passenger short haul: 90507
Num of passengers medium + long haul: 39342
```

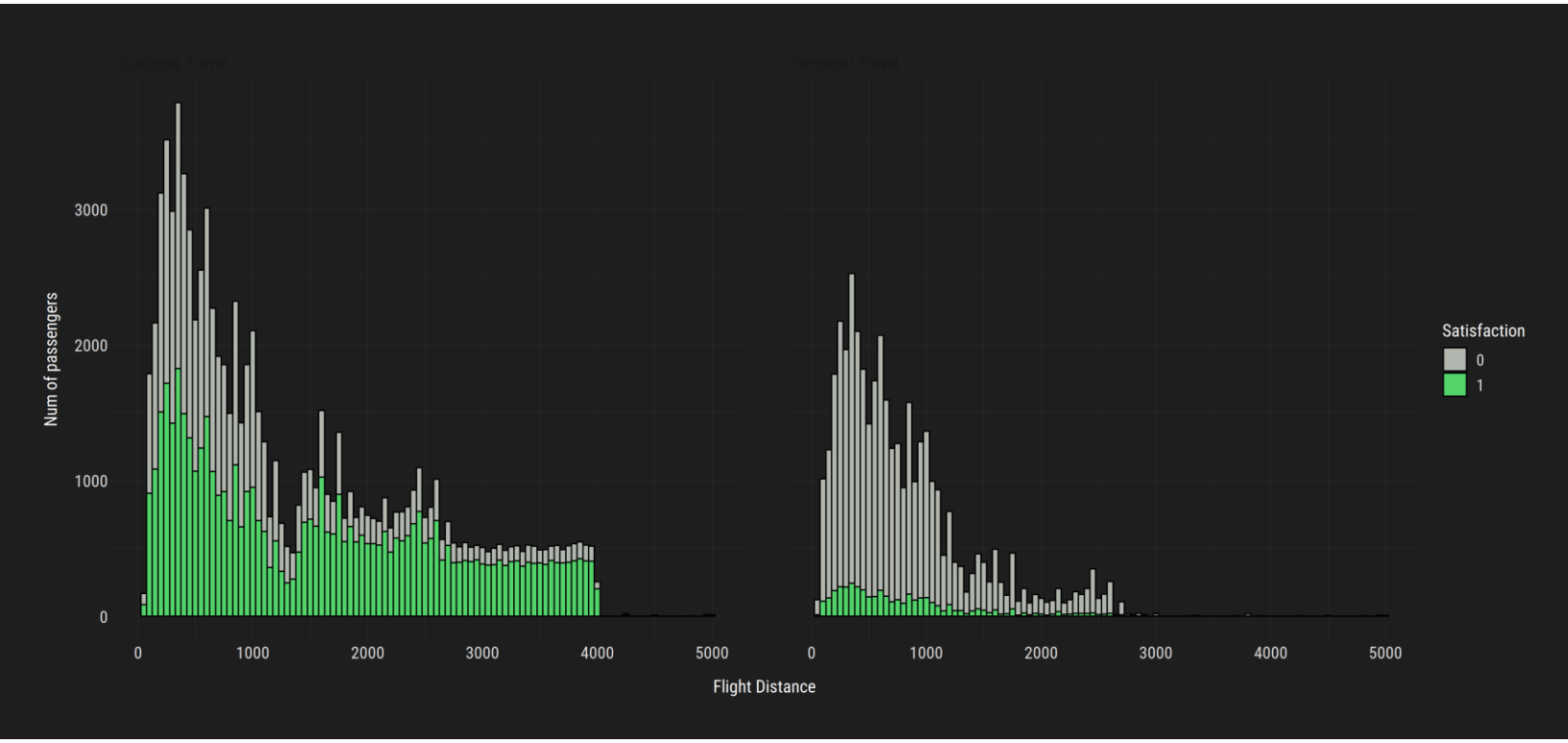
Пасажирів, що літали рейсами короткої дистанції приблизно вдвічі більше ніж пасажирів, що літали середніми та довгими дистанціями разом узяті, але це не означає, що співвідношення кількості рейсів таке ж. Часто перельоті на невеликі відстані (до 1500 км) можуть мати більшу кількість пасажирських місць ніж перельоти на велику відстань (3000+ км), або можливо рейси великих відстаней мають в середньому більшу кількість вільний (незайнятих) пасажирських місць ніж на малих дистанціях.

Проте, враховуючи всі вище досліджені показники, можна дати оцінку низькій зацікавленості авіакомпанії в підтримці і обслуговуванні рейсів коротких дистанцій, але можливо що компанія націлена не просто на перельоти великими дистанціями, а людьми якоїсь конкретної цілі подорожі – business/personal, з цього випливає наступне дослідницьке питання.

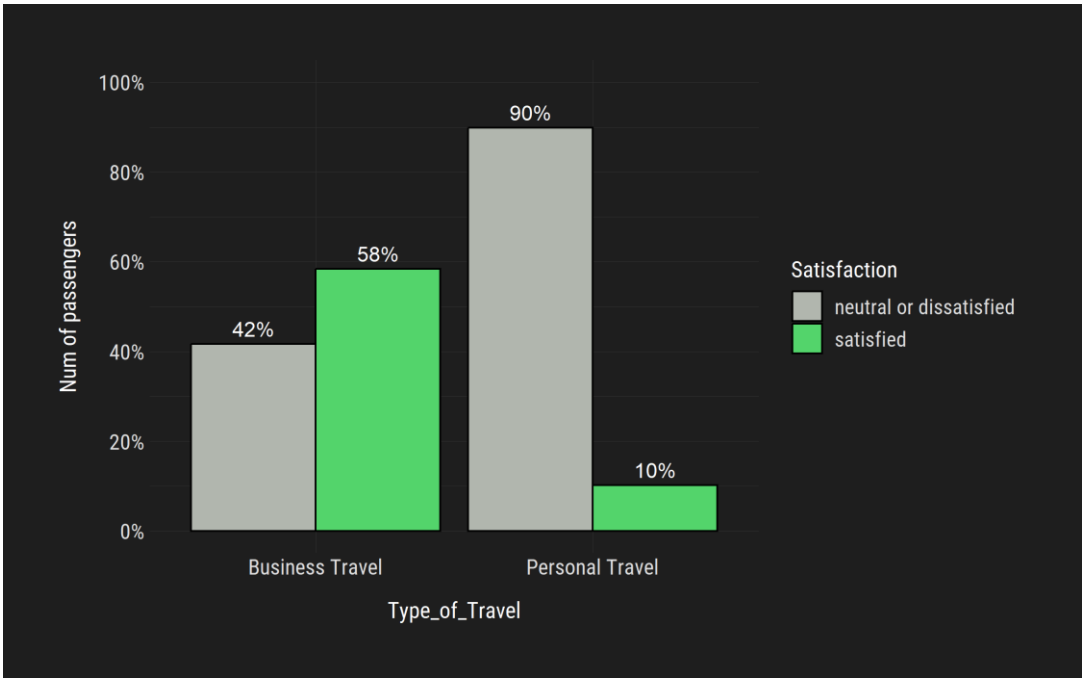
**Дослідницьке питання**: *Чи існує залежність між задоволеністю клієнтами рейсом та цілями перельотів?*

**Гіпотеза**: *Пасажири, що подорожують в персональним цілях залишаються менш задоволеними перельотом.*

Подивимося на наступний фацетований графік за типом подорожі (business/personal):




Візуально можна побачити **велику** різницю в задоволеності між людьми, що літають в бізнес та в персональних цілях. Якщо подивитися на лівий графік (бізнес тип подорожі), можна помітити, що починаючи з умовної межі в 1500 км кількість задоволених рейсами значно переважає незадоволених. Подивимося на наступний barplot:



```
Personal type % of dissatisfaction: 0.9
Business type % of satisfaction: 0.58
Business type % of satisfaction (medium and long flights): 0.74
```

**Висновок**:

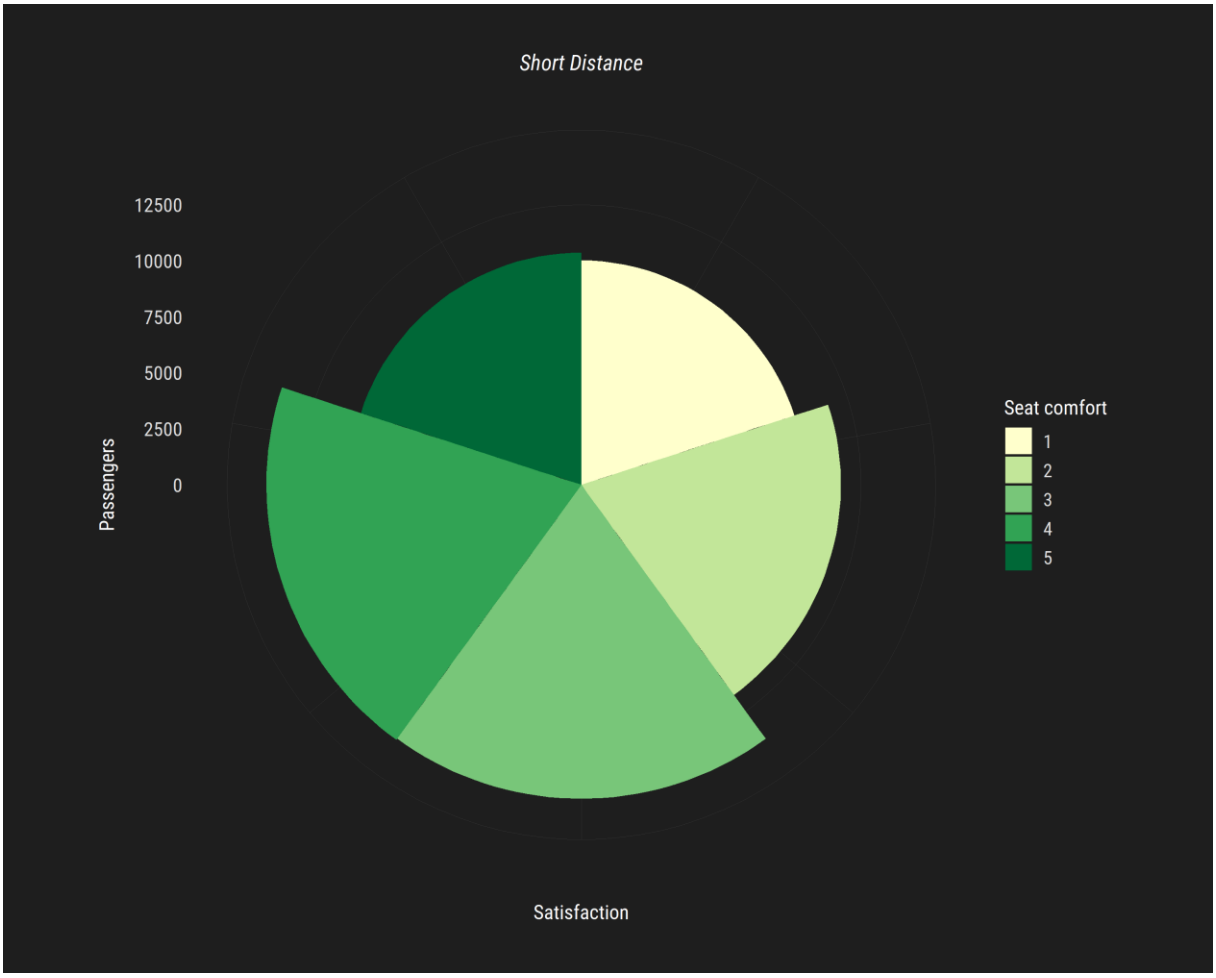
Дивлячись на два попередні графіки, можна переконатися в тому, що авіакомпанії слід звернути увагу на рейси з типом «персональний переліт», бо незалежно від відстані, ті, хто подорожував з особистих причин були майже на 90% (!) незадоволеними, можливо причина лежить у тому, що у людей, що подорожують в особистих цілях виникають проблеми в обслуговуванні до поїздки або під час.

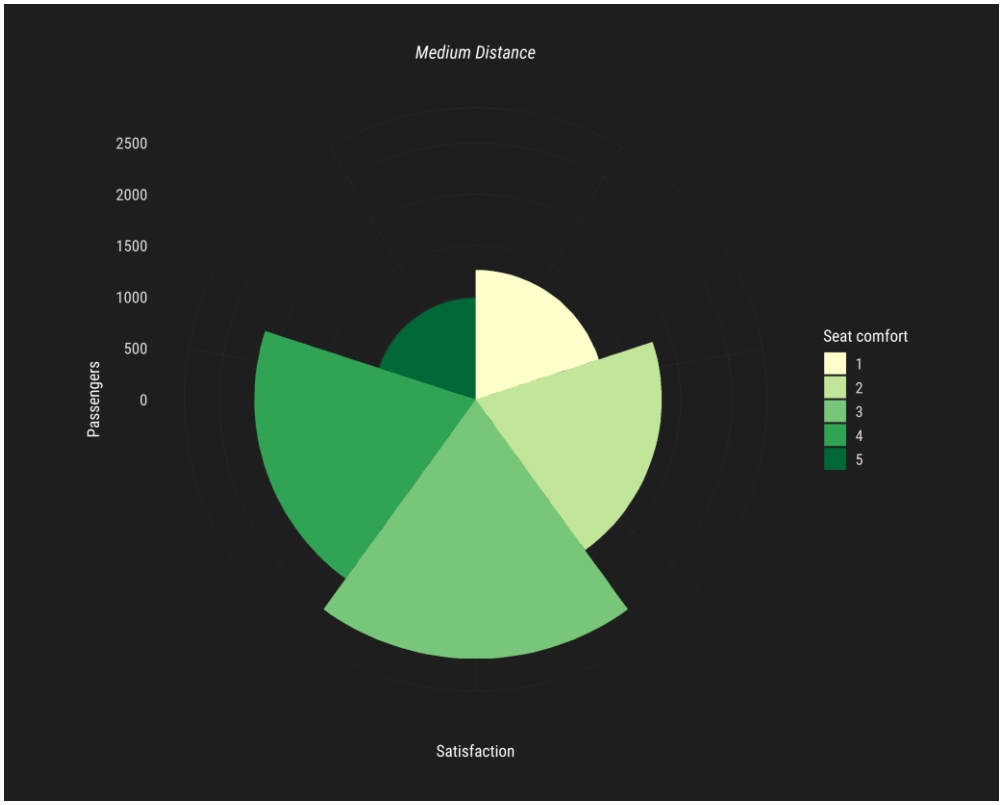
Проте пасажири, що літали з бізнес причин незалежно від відстані були на 58% задоволені, а якщо брати лише рейси більше за 1500 км, то пасажири задоволені аж на 74%.

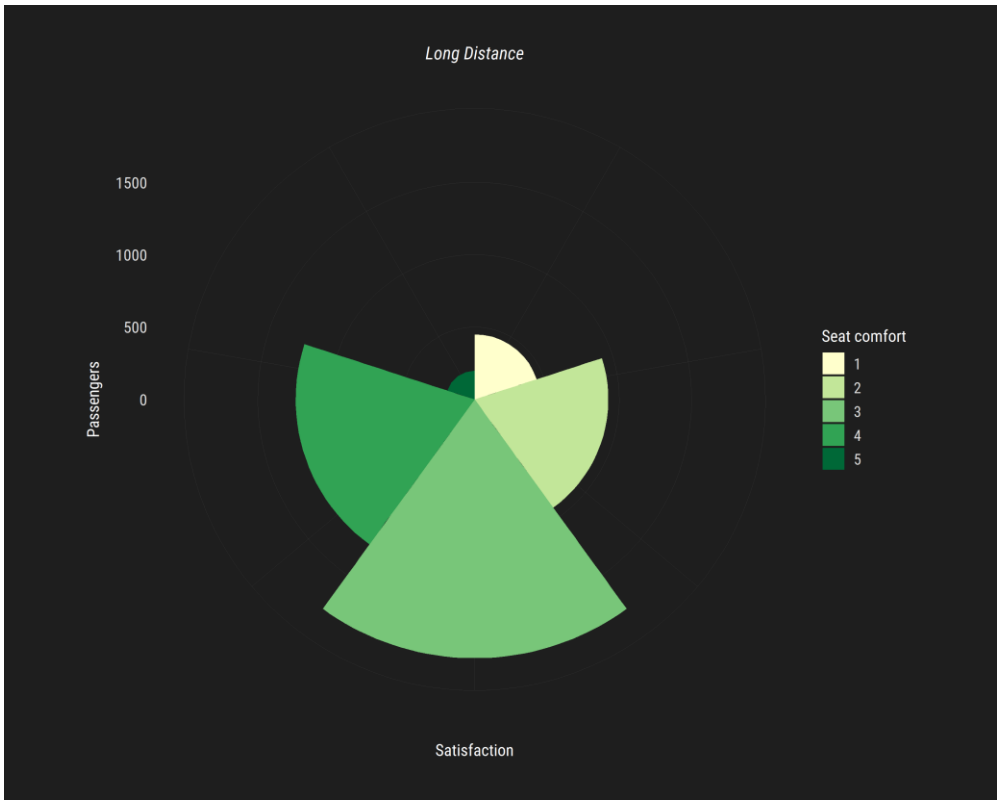
**Дослідницьке питання**: *Чи існує різниця впливу комфортабельності сидіння в залежності від дистанції рейсів?*

**Гіпотеза**: *Комфортабельністю можна нехтувати при невеликих дистанціях рейсів.*

Побудуємо polar barplot по кожній з категорій дистанції:




Medium Distance
2500
2000
1500
1000
500
0
Passengers
Satisfaction
Seat comfort
1
2
3
4
5


Long Distance
1500
1000
500
0
Passengers
Satisfaction
Seat comfort
1
2
3
4
5

Досліджуючи питання комфортабельності сидінь треба слід досліджувати лише ту групу людей, яка залишилася незадоволеною рейсом, щоб визначити вплив комфортабельності сидінь та в подальшому зменшити кількість незадоволених.

Якщо людина залишилася **незадоволеною рейсом**, то слід звернути увагу на рейси тієї дистанції (short, medium, long), на якій частка **задоволених сидіннями** людей **менша**, можливо саме через некомфортне сидіння людина залишилася **незадоволеною рейсом**.

Будемо вважати, що людина задоволена сидінням лише якщо вона поставила оцінку 4 або 5, у всіх інших випадках вона незадоволена сидінням.

```
Fraction of seat satisfaction on Short -haul:  0.41
Fraction of seat satisfaction on Medium -haul:  0.36
Fraction of seat satisfaction on Long -haul:  0.31
```

Аналізуючи отримані результат, можна прийти до **висновку**, що чим більша відстань рейсу, тим менше оцінка комфорту сидіння. Тобто на невеликих відстанях навіть ті люди, що залишилися незадоволеними рейсом більше задоволені сидіннями, ніж люди, що подорожували великими відстанями на 10%.

А отже комфорт сидіння має більший вплив на перельоти великих дістанцій.

**Дослідницьке питання**: *Які признаки мають найбільший/найменший вплив на задоволеність клієнтів бізнес/економ класу?*

**Гіпотеза**: *Існують фактори, які по-різному впливають на людей, подорожуючих різними класами.*

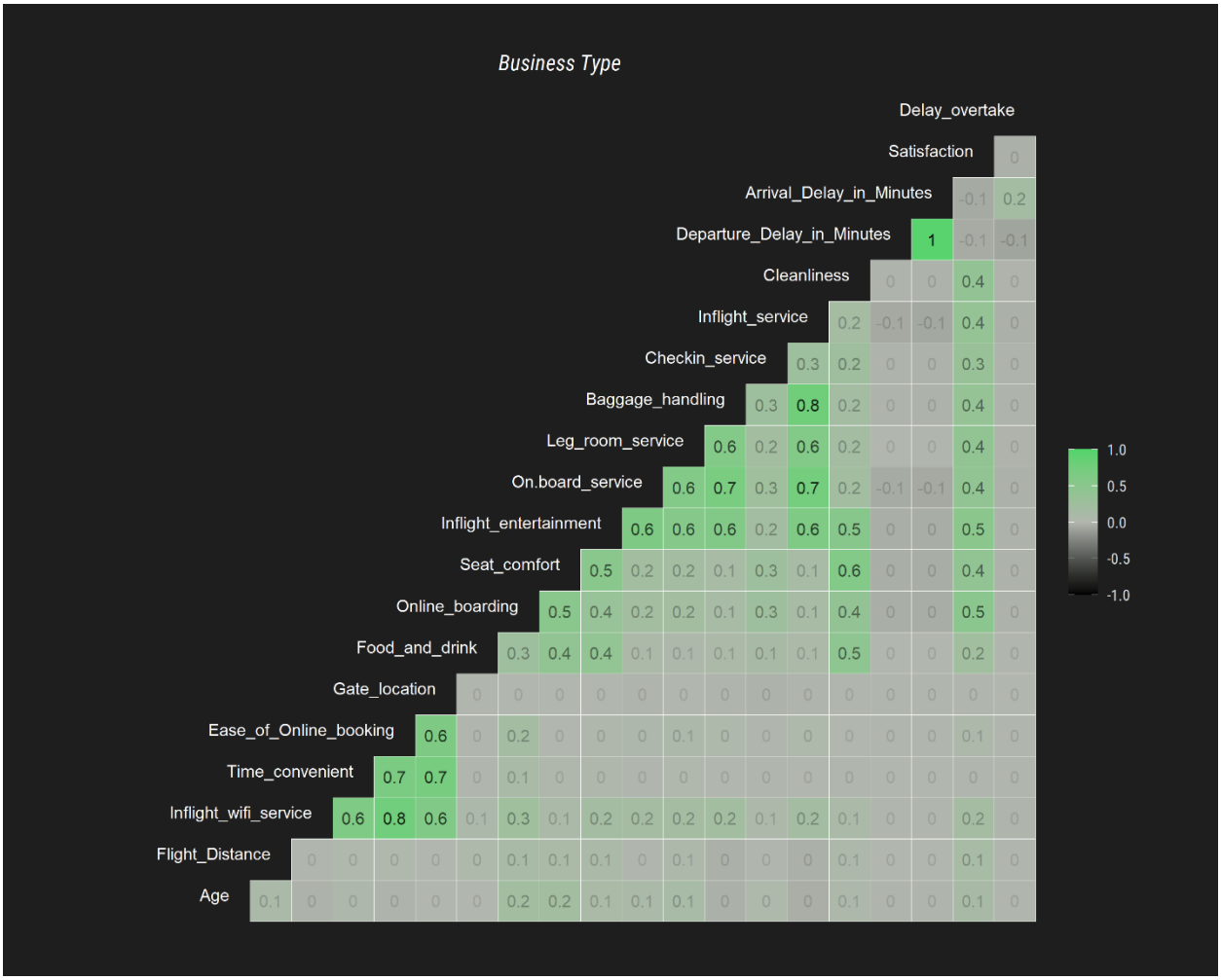
Як відомо, обслуговування бізнес класу виходить авіакомпаніям дорожче і можливість втратити клієнта бізнес класу більш негативна, ніж клієнта економ класу. Визначивши як фактори найбільше впливають на клієнтів бізнес класу ми зможемо їх поліпшити. (наприклад за рахунок зниження фінансування і так не важливих факторів для клієнтів економ класу)

Подивимося кількісний склад кожного класу:

```
Business      Eco Eco Plus
   62147    58293     9409
```

Через те, що виділяється лише два великих класи Business та Eco, додамо до Eco клас Eco Plus, клієнтів якого лише 7%.

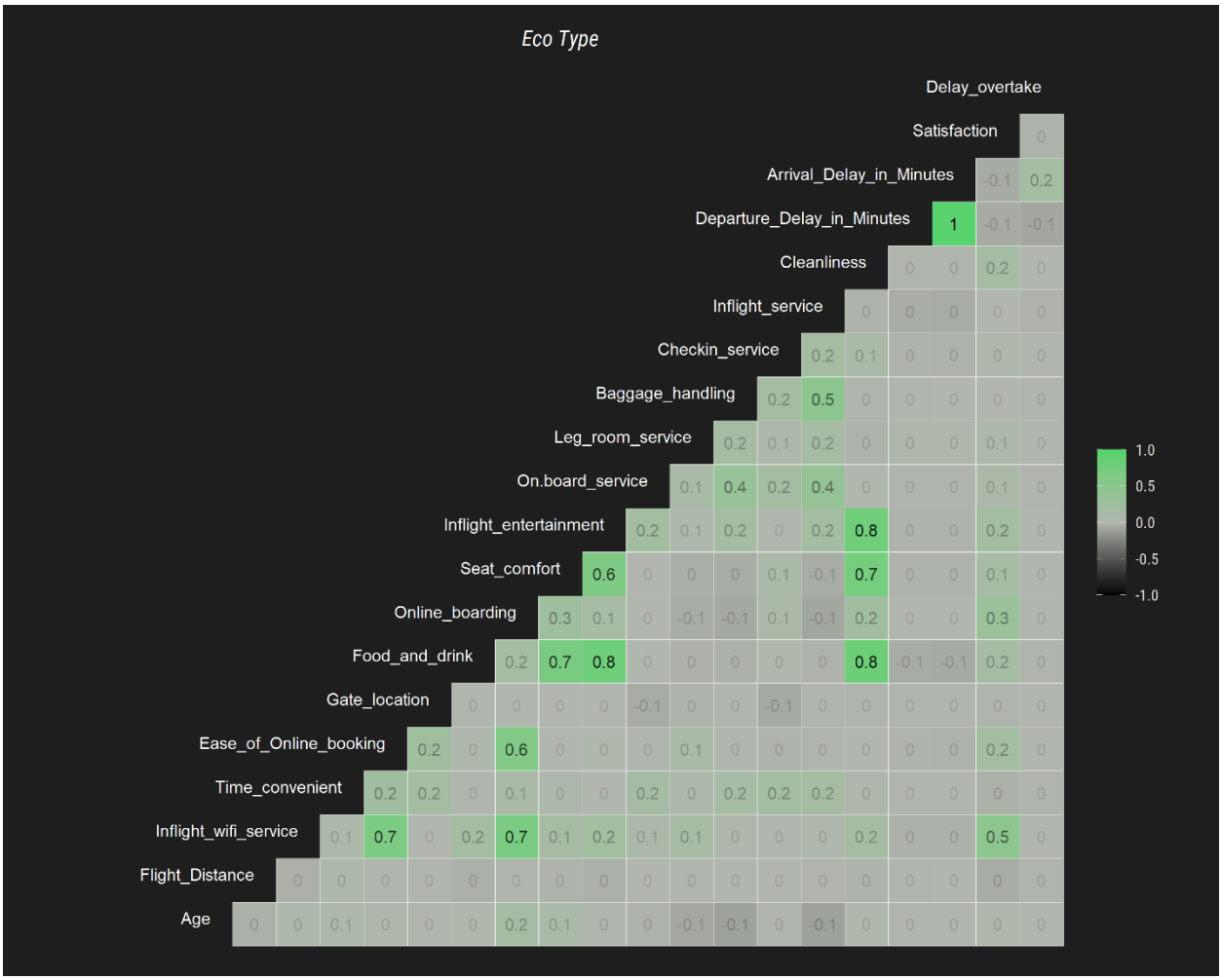
Побудуємо відповідно для двох класів трикутну матрицю кореляцій, та подивимося кореляцію із цільовою змінною:




В результаті можна спостерігати, що для клієнтів бізнес класу найбільш корелюючими з цільовою змінною факторами є Online boarding (0.51), Inflight Entertainment (0.5), On-board Service(0.43)

В той час найменшим корелюючими є Gate Location (-0.003), Departure/Arrival time convenient(0.013), Ease of Online booking(0.06)

Тепер побудуємо ту ж трикутну матрицю кореляцій для клієнтів економ класу:




Проте різниця впливу факторів все ж існує, а отже гіпотеза підтверджується, можна спостерігати, що для клієнтів економ класу найбільш корелюючими з цільовою змінною факторами є Inflight wifi service (0.47), Online_boarding (0.31), Ease_of_Online_booking (0.21), Food_and_drink

В той час найменшим корелюючими є Gate Location, Departure/Arrival time convenient, Ease of Online booking

# 5 Висновки

Ключовою метою дослідження є виявлення корисних інсайтів в даних за допомогою яких можна було б сформулювати деякі стратегії для покращення якості обслуговування існуючих послуг, щоб підвищити загальний рівень задоволеності пасажирів авіакомпанії. В результаті дослідження та поставлених дослідницьких питань можна зробити наступні висновки:

- На відміну від людей середнього віку, діти та люди похилого віку здебільшого **незадоволені** рейсами авіакомпанії

На основі цього висновку складно одразу скласти стратегію дій для авіакомпанії, щоб підвищити цей рівень, але попри те ми виявили велику проблему і знаємо в якому напрямку треба працювати для покращення задоволеності пасажирів.

- Існує проблема в комфортності сидінь в літаках для дітей

Як ми побачили, діти здебільшого незадоволені комфортністю сидінь у літаках, в такому разі компанії треба як слід зайнятися питання облаштування спеціальних сидінь, або прикріплень на сидіння для комфорту дітей.

- Низька задоволеність онлайн-реєстрацією серед людей похилого віку

В процесі дослідження ми визначили, що середня задоволеність онлайн-реєстрацією серед людей похилого віку нижча ніж від людей середнього віку, такий показник може бути пов’язаний з наприклад незручним інтерфейсом сайту авіакомпанії, або можливо існують деякі проблеми в процесі реєстрації на рейс, тощо. Так чи інакше авіакомпанії слід зайнятися цим питанням, та можливо провести опитування серед користувачів сайтом.

- Націленість компанії на бізнес мандрівки

Досліджуючи питання задоволеності клієнтів в залежності від цілей перельоту, ми визначили, що на відміну від бізнес цілей, де загальна задоволеність становить 58%, а для рейсів 1500км+ так взагалі 74%, пасажири які користувалися послугами авіакомпанії в персональних цілях мають рівень задоволеності 10% (!). Виходячи з логічних міркувань можна зробити висновок, що компанія просто націлена на бізнес мандрівки і тому рівень задоволеності в персональних цілях такий низький, бо якби авіакомпанія не була націлена на якусь конкретну групу, то складно уявити взагалі існування такої авіакомпанії, де задоволеність послугами 10%, враховуючи що цей бізнес дуже сильно залежить від клієнтів.

- Найбільш впливовіші фактори – online boarding, inflight entertainment, inflight wifi service

В результаті вивчення останнього дослідницького питання ми дійшли до висновку, що найбільш корелюючими факторами для клієнтів бізнес класу є Online Boarding та Inflight Entertainment в той же час для клієнтів економ класу – Inflight wifi service. Знаючи найвпливовіші фактори авіакомпанії слід детальніше дослідити їх та виявити кластери пасажирів, які невдоволені даними факторами, щоб надалі викорінити причину незадоволеності. Також можна запропонувати авіакомпанії ще одну стратегію, описану в пункті про мотивацію дослідження: ми визначили найменш корелюючі фактори, якістю яких можна знехтувати, щоб виділити більше коштів на покращення найбільш кореліюючих факторів.

Впливу викидів (31 запис) та пропущених значень (393 записи) не виявлено здебільшого в силу розміру датасету з яким була робота.

# 6 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.