
От заказа до отправления: влияние возраста

На основе датасета «Т-Банк:
поездки на самокатах»

NeDanoJit



Структура данных



Качественные переменные:

- Название модели самоката
- Пол клиента
- Уровень образования клиента
- Регион проживания человека
- Семейный статус человека



Количественные переменные:

- Стоимость минуты
- Размер суммы, которая замораживается на счете в момент взятия самоката
- Километраж поездки
- Стоимость поездки
- Размеры выплаченного кэшбэка
- Возраст клиента

396750

строк

Данные за сезон 2024:

С апреля по
октябрь

71389

уникальных
пользователей

Используемые переменные

order_rk	Идентификатор заказа (поездки)
party_rk_id	Идентификатор клиента
transport_model	Название модели самоката
distance_km	Километраж поездки
created_dttm	Дата и время создания заказа
book_start_dttm	Дата и время начала поездки
local_book_start_dttm	Дата и время начала поездки в часовом поясе человека, который брал самокат
gender_cd	Пол клиента
age	Возраст клиента
education_level_cd	Уровень образования клиента

Вводим новые переменные

- **travel_time_h** – время, проведённое в поездке, ч
- **average_speed** = $\frac{\text{distance_km}}{\text{travel_time_h}}$ – средняя скорость во время поездки, км/ч
- **seconds_difference** = `book_start_dttm` - `created_dttm`
разница времени скана и времени начала поездки, с
- **hour** – час начала поездки с учётом часового пояса
- **trip_number** – количество поездок у пользователя, совершенных на момент начала этой поездки
- **day_time** – время суток, когда была совершена поездка

Статистика по основным количественным переменным

	age	seconds_difference	hour
Среднее	31.54	6.12	12.55
Среднекв. откл.	9.43	19.66	5.3
Минимальное	12	-1.74	0
25 процентиль	24	3.64	8
50 процентиль	31	4.21	13
75 процентиль	37	4.97	17
Максимальное	94	603.56	23

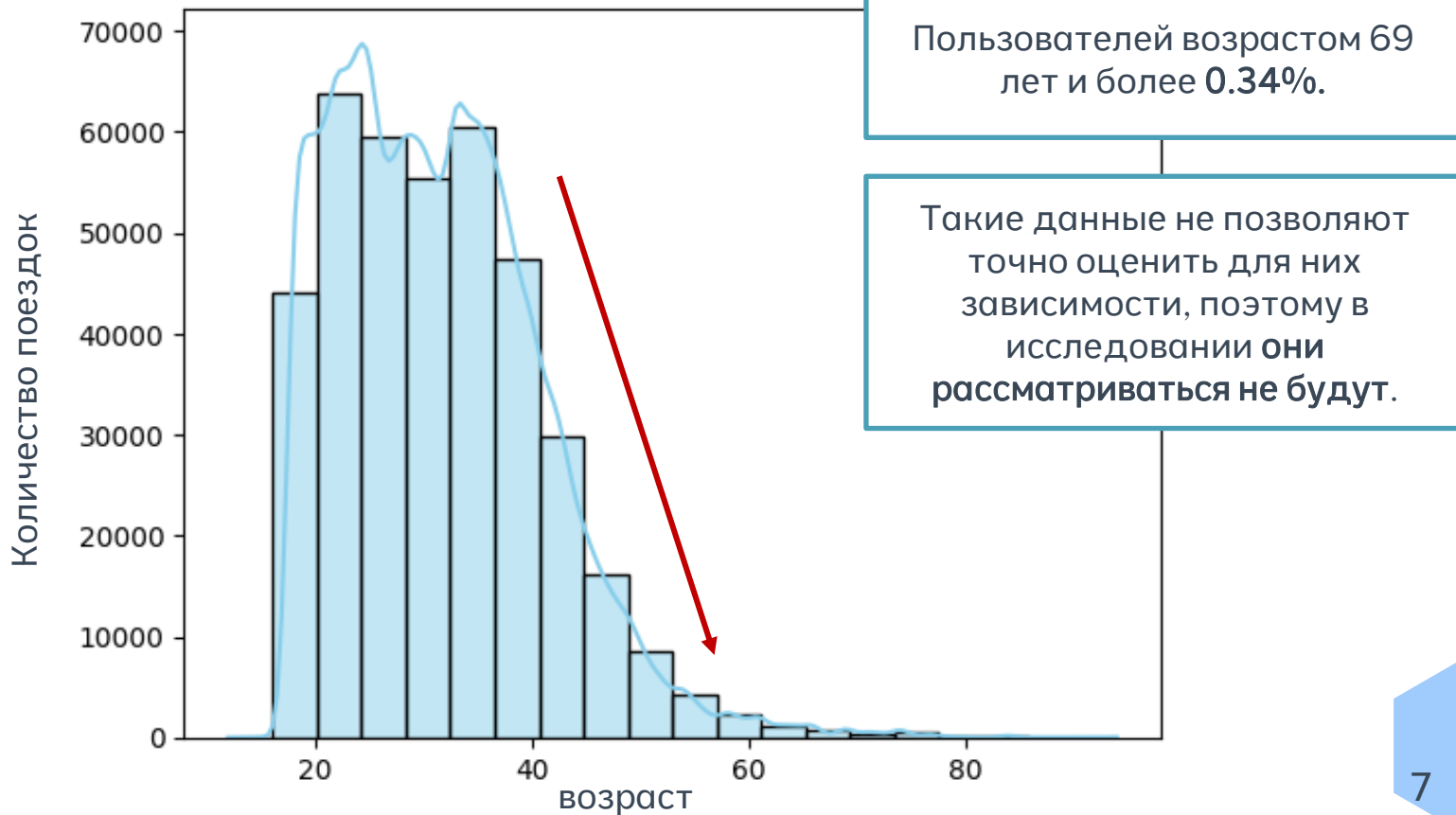
Наводим порядок

Удаляем выбросы (4540, 1,14%):

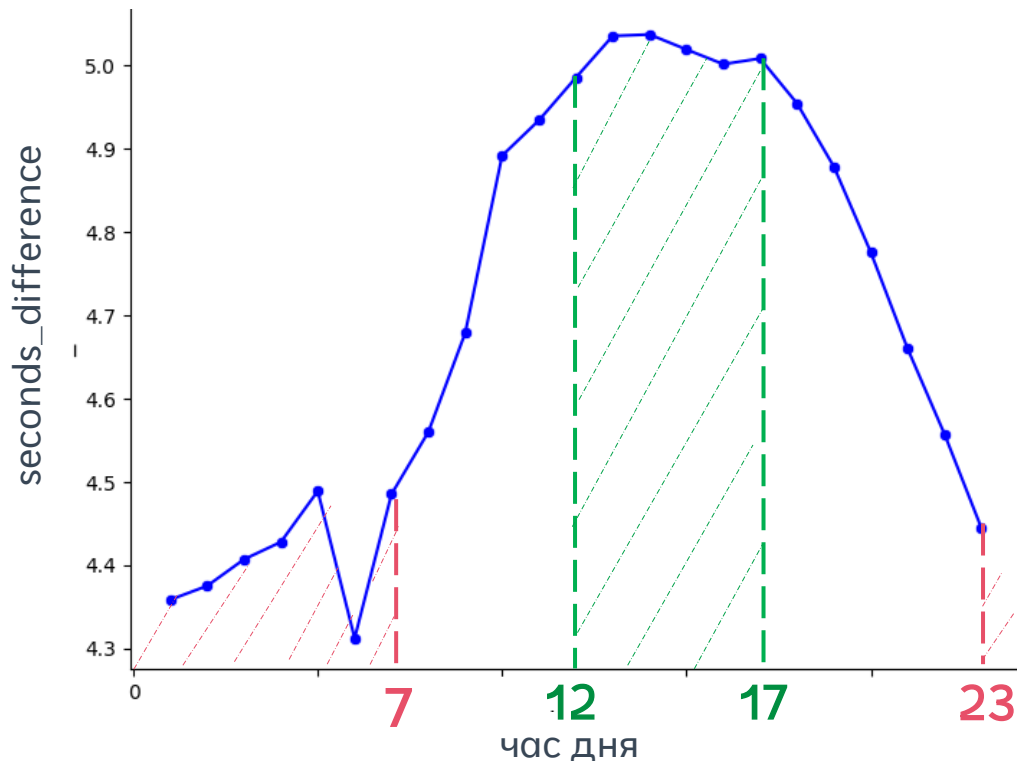
1. Убираем все строки со скоростью, превышающей 40 км/ч
2. Убираем отрицательные значения `seconds_difference`
3. Убираем оставшиеся выбросы* с помощью метода Z-оценки

*За выбросы считаем все значения переменной за пределами трех стандартных отклонений

Распределение поездок по возрасту



Зависимость времени, потраченного на активацию самоката от часа начала поездки

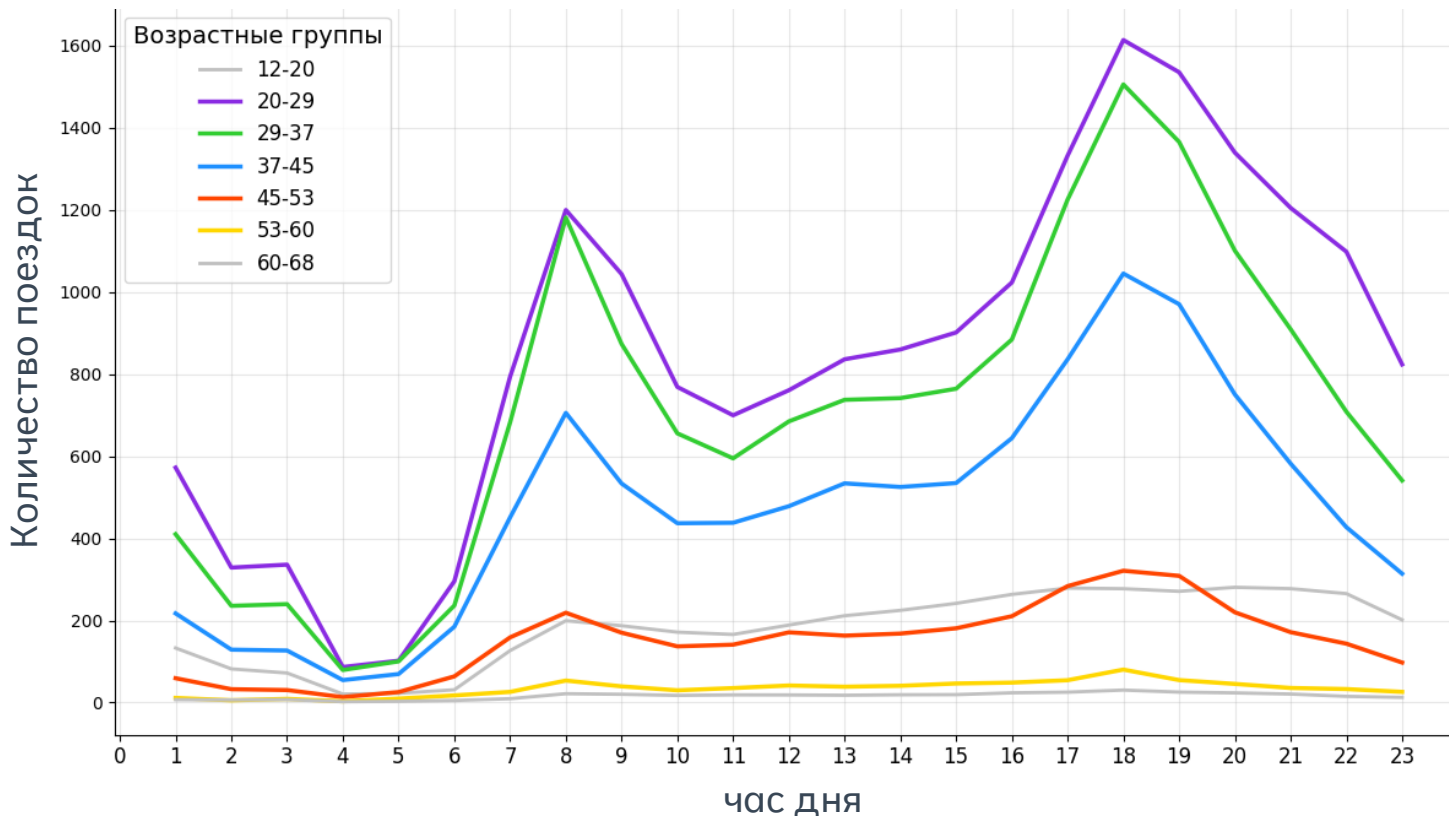


Люди тратят разное время на активацию самоката в разные часы дня:

Пик – промежуток 12:00 – 17:00
Максимальное значение seconds difference на графике = 5.037 с

Спад – промежуток 23:00 – 7:00
Минимальное значение seconds difference на графике = 4.312 с

Распределение поездок по возрастным группам и часам суток





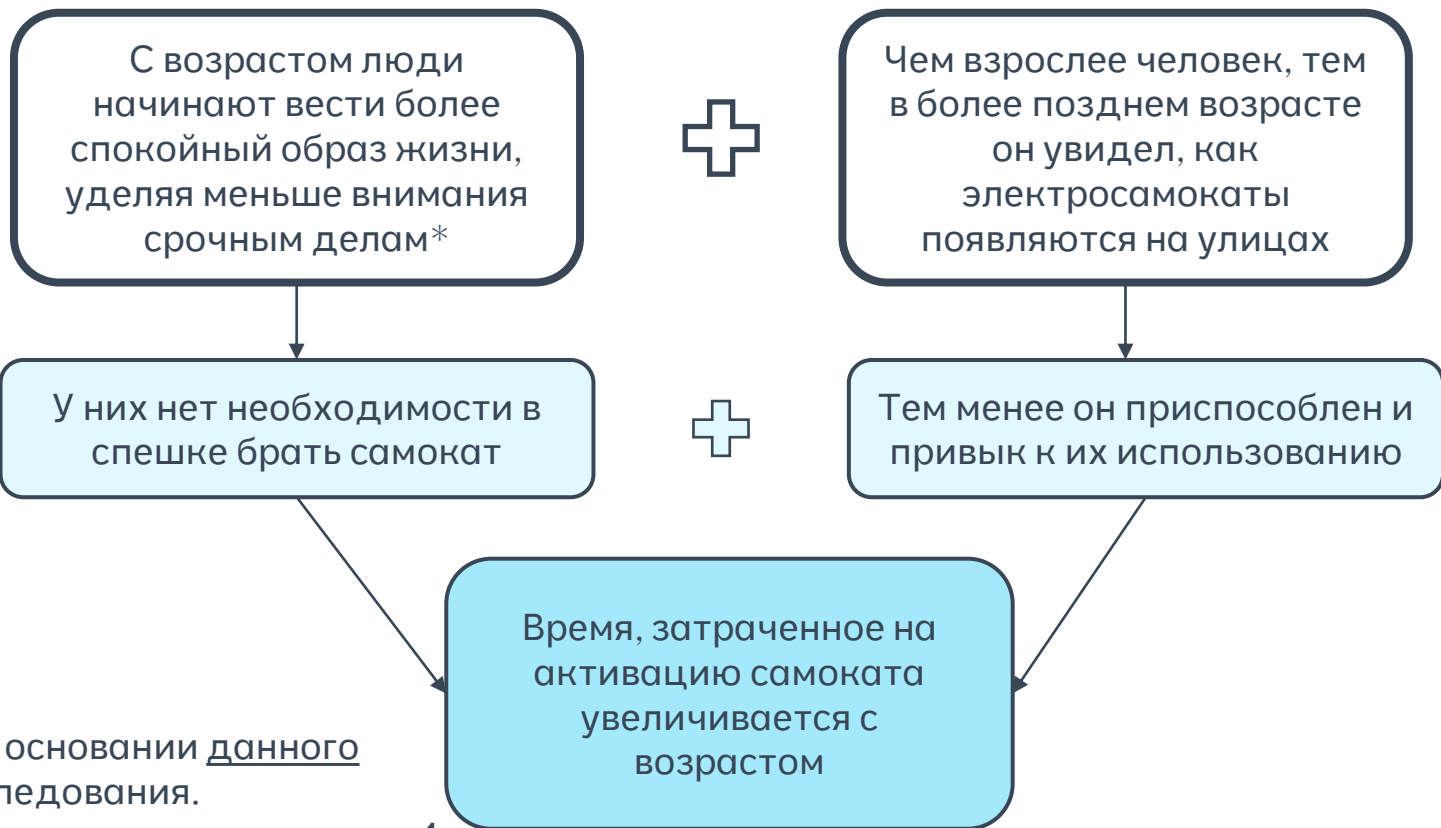
Исследовательский вопрос

**Как возраст
клиента влияет на
время между
созданием заказа и
началом поездки?**

Гипотеза

**Чем старше клиент,
тем больше времени
проходит между
созданием его
заказа и началом
поездки.**

Механизм



*на основании данного исследования.
ссылку см. в приложении 1

Альтернативный механизм

Вероятно, существуют сторонние факторы, влияющие на разницу времени старта поездки и сканирования QR-кода.

Например, человек **просто отвлётся** или вообще двигается расслабленно и просто не торопится.

➡ Потратил больше времени от сканирования QR-кода до начала поездки



Математическая модель

Используем метод **множественной регрессии**, чтобы рассмотреть влияния нескольких независимых переменных на одну зависимую

Используем в модели следующие переменные:

у – зависимая переменная	х – независимые переменные
seconds_difference – разница во времени между сканированием QR-кода и стартом поездки	<ul style="list-style-type: none">- время суток- возраст- квадрат возраста- количество заказов пользователя на момент поездки

Математическая модель

R-squared: 0.003

для подвыборки женщин

F-statistic: 24.84; значима на любом разумном уровне значимости

Объясняющая переменная	coef	std error	t	p-value
const	4.74	0.048	36.202	0
age_square	0.0003	<0.01	2.84	0.005
age	-0.0116	0.007	-1.6	0.11
trip_number	0.0048	0.001	9.18	0
day_afternoon	0.079	0.04	2	0.046
day_evening	-0.046	0.04	1.3	0.193
day_night	-0.115	0.05	2.14	0.03

Уровень значимости = 0.05

RSS1 = 609521.7

Математическая модель

R-squared: 0.002

для подвыборки мужчин

F-statistic: 113.3; значима на любом разумном уровне значимости

Объясняющая переменная	coef	std error	t	p-value
const	4.99	0.065	76.35	0
age_square	0.0003	<0.01	5.55	0
age	-0.0204	0.004	-5.14	0
trip_number	0.005	<0.01	25.4	0
day_afternoon	-0.006	0.016	-0.34	0.74
day_evening	-0.004	0.015	0.3	0.76
day_night	-0.005	0.02	-0.23	0.82

RSS2 = 3583543.5

Уровень значимости = 0.05

F-тест

$$\frac{(RSS - RSS_1 - RSS_2)/k}{(RSS_1 + RSS_2)/(n - 2k)} \sim F(k, n - 2k)$$

F = 568.385

P-value < 0.01

k – количество параметров модели

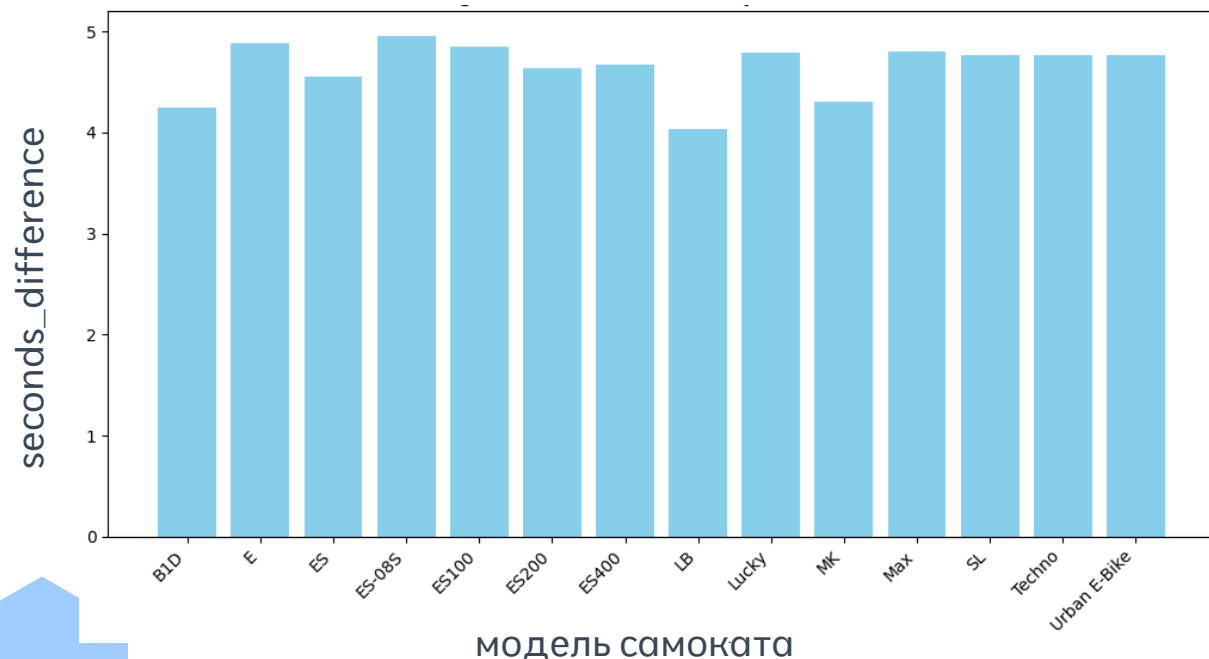
n – общее количество наблюдений

На основе F-теста было выявлено, что регрессии на подвыборках лучше объясняют данные, чем общая модель.

Таким образом выборки неоднородны из-за различных коэффициентов, и необходимо оценивать модели отдельно.

Значение RSS см. в приложении 3.

Различия в моделях самокатов



На уровне значимости 5% не было выявлено различий в средних значениях `seconds_difference` для разных моделей самокатов.*
В силу этого можно сделать вывод, что полученные оценки не подвержены смещению из-за различий в моделях.

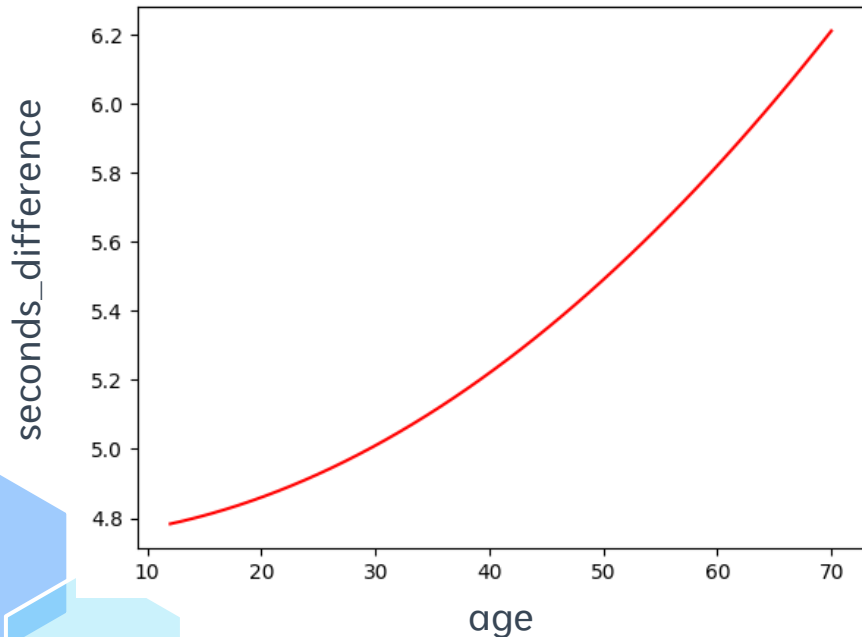
*рез-ты теста см. в приложении 6

**Наша гипотеза
частично
подтвердилась**

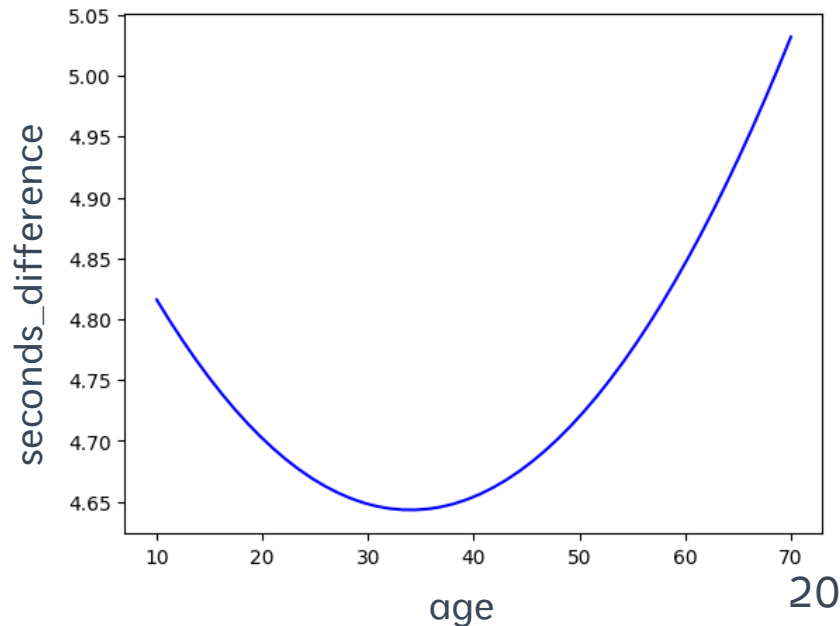


Интерпретация

Зависимость **для женщин** является **квадратичной** (с возрастом зависимая переменная возрастает)



Зависимость **для мужчин** также является **квадратичной** (зависимая переменная уменьшается до 34, затем возрастает)



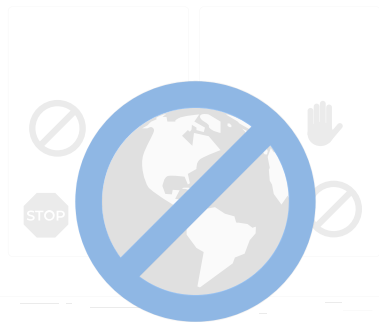
Policy implication



Введение тарифа, по которому у пользователя будет дополнительное время после скана QR-кода без списывания денег. Каждые несколько лет это время можно увеличивать, тем самым улучшая пользовательский опыт

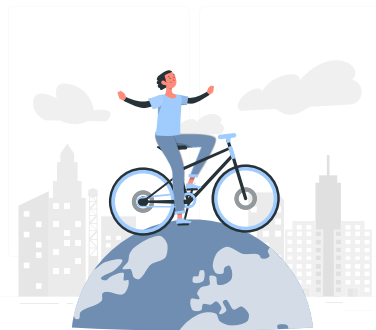


Оптимизация таргетированной рекламы: пользователям банка в более старшем возрасте, которые показали заинтересованность в тех или иных спортивных товарах и услугах, будет показана реклама сервиса, где **упомянутся «поездки без спешки и с комфортом».**



Ограничения

- Данные только за сезон 2024 года
- Невозможность обобщения на весь мир
- Возможно, ложные данные возраста (клиент соврал)
- Неполные данные (некоторые переменные просто отсутствуют)



Перспективы

- Долгосрочная динамика
- Больше регионов
- Повышение точности данных
- Углубление

**Спасибо за
внимание!!!**



Наш Github репозиторий



Наша команда:

- Косенкова Дарья
- Пеганова Виктория
- Миронов Максим
- Зайцев Роман
- Парфенцев Антон

ПРИЛОЖЕНИЕ

Ресурсы

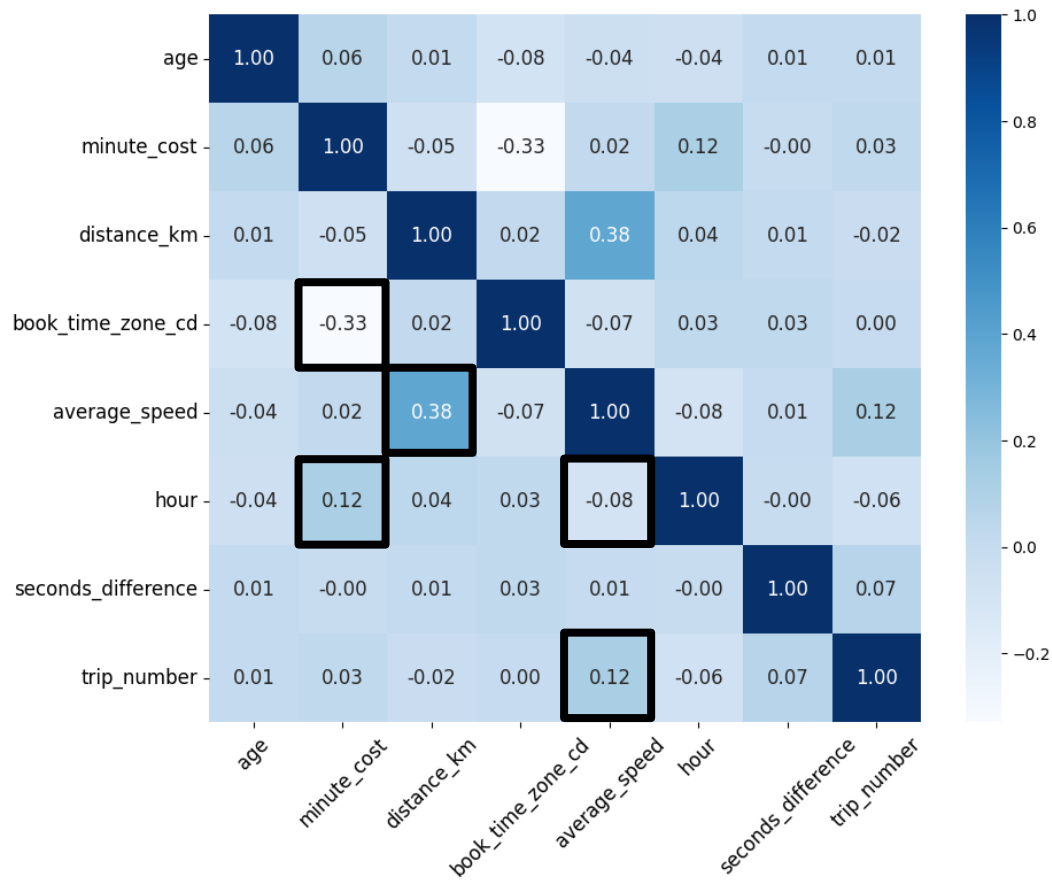
Для оформления:

- [International nurses day template](#)
- [StorySet](#)
- [Icon Pack: Research and Development | Lineal](#)
- [SlidesGo](#)
- [Freepik](#)

Исследование: смена целей и здоровое старение:

https://academic.oup.com/psychsocgerontology/article/76/Supplement_2/S105/6369279?login=false

Матрица корреляций



Математическая модель

для всех данных

R-squared: 0.003

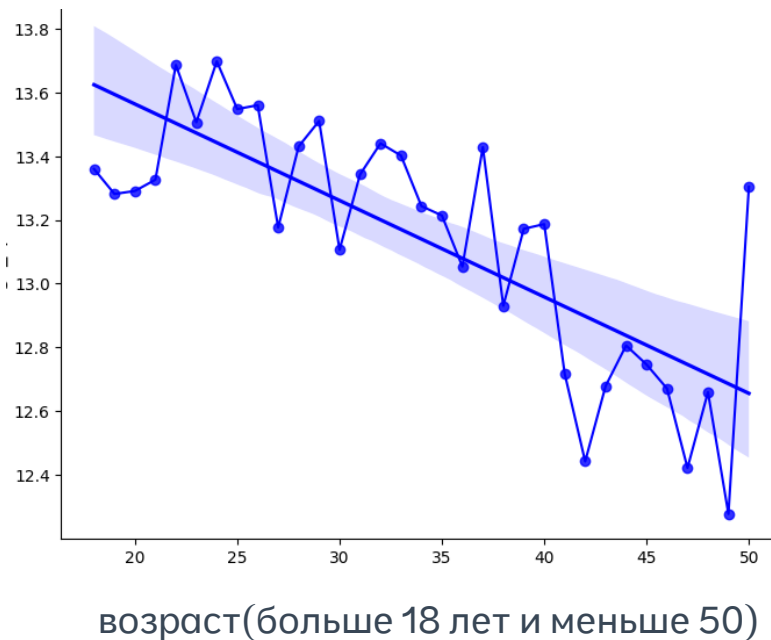
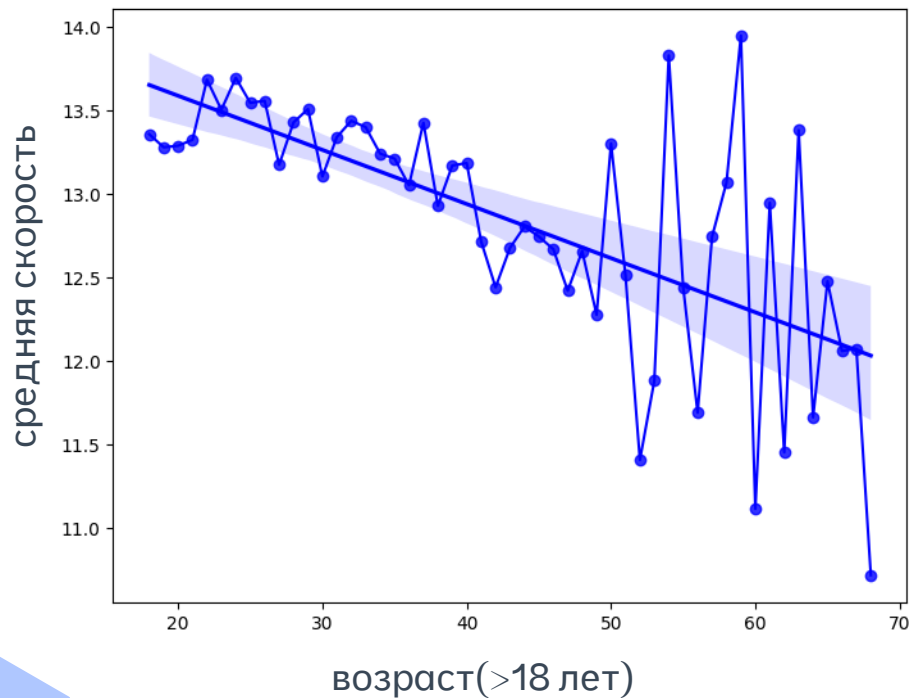
F-statistic: 24.84; значима на любом разумном уровне значимости

Объясняющая переменная	coef	std error	t	p-value
const	5	0.057	88.3	0
age_square	0.0003	<0.01	7	0
age	-0.0207	0.003	-6.16	0
trip_number	0.005	0	26.98	0
day_afternoon	0.03	0.015	2	0.045
day_evening	-0.019	0.013	-1.4	0.168
day_night	-0.039	0.02	-1.87	0.06

Уровень значимости = 0.05

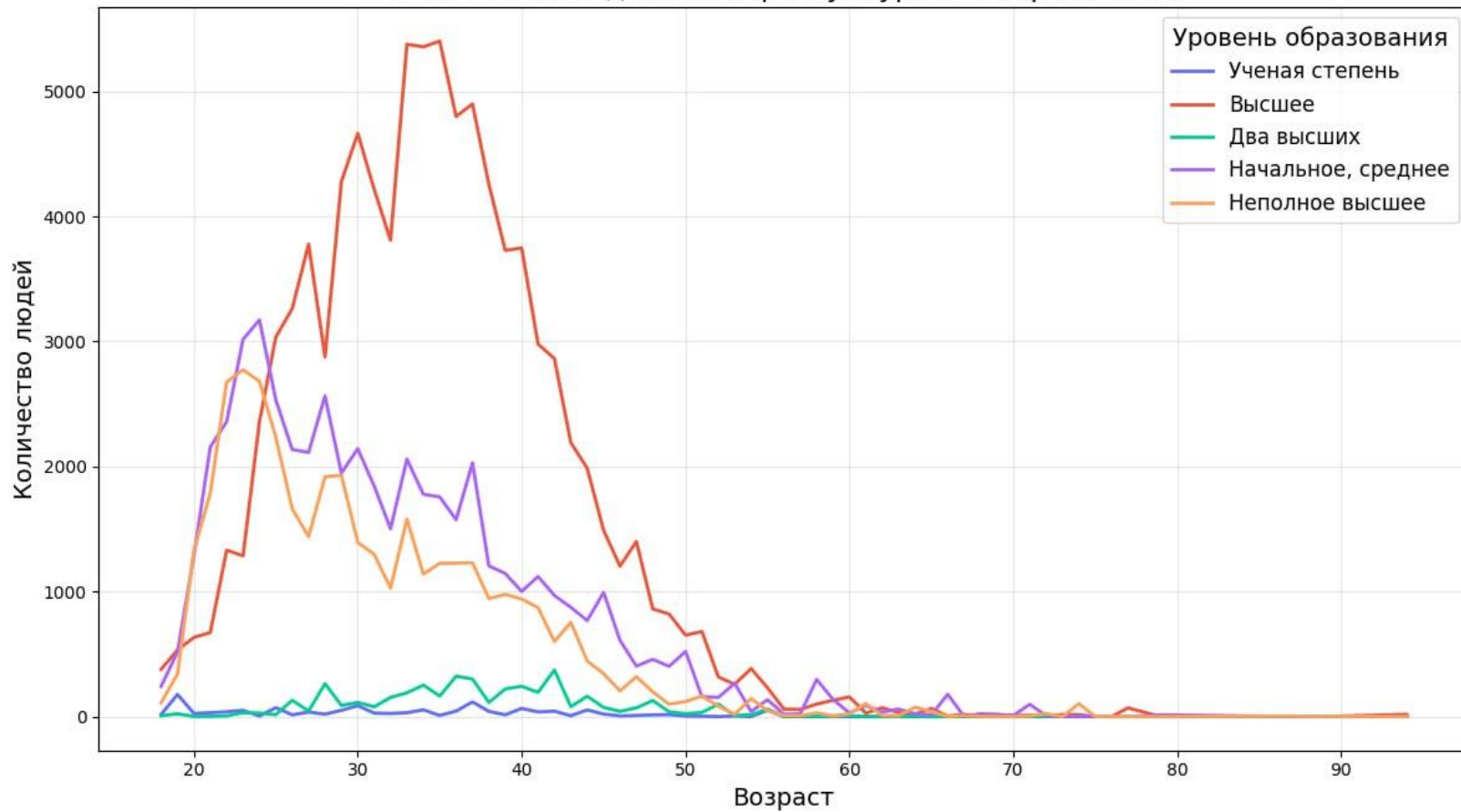
RSS = 4230286.7

Зависимость средней скорости поездок от возраста клиента



Коэффициент корреляции = -0.04

Количество людей по возрасту и уровню образования



ANOVA test

	f-statistic	p-value
E:ES200	10.99	0
E:MK	6.5	0.01
E:Max	9.75	0
E:Urban E-Bike	70.06	<0.01
ES-08S:LB	7.48	0
ES-08S:Urban E-Bike	10.02	0
ES100:ES200	5.74	0.02
ES100:MK	3.92	0.05
ES100:Urban E-Bike	13.28	0
ES200:ES400	8.77	0
ES200:Max	6.57	0.01
ES200:SL	10.45	0
ES200:Urban E-Bike	35	<0.01
ES400:MK	5.65	0.02
ES400:Urban E-Bike	17.67	<0.01
LB:Lucky	4.82	0.03
MK:Max	4.73	0.03
MK:SL	6	0.01
MK:Urban E-Bike	14.05	<0.01
Max:SL	17.57	<0.01
Max:Urban E-Bike	99.29	<0.01
SL:Urban E-Bike	67.88	<0.01