Структура презентации

1. Данные
   - описание переменных
   - разведывательный анализ
   - обработка данных
2. Введение
   - мотивация
   - исследовательский вопрос
   - гипотеза
   - механизм
3. Модель
   - мат. модель
4. Результаты
   - основные результаты
   - проверка устойчивости
5. Выводы
   - применение результатов
   - ограничения и перспективы исследования

Описание переменных

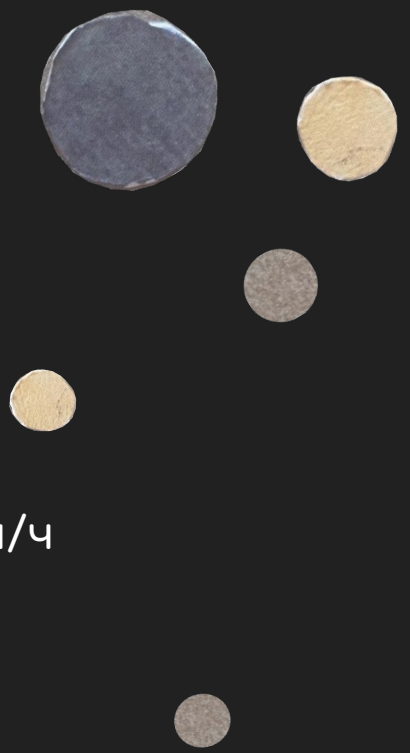
Данные
о штрафах ГИБДД, полученных клиентами Т-банка за 28.04.24 – 28.05.24 (около 0.5%)

97307 строки
×
22 столбца

Характеристики нарушителей по: полу, уровню образования, доходности, виду штрафа, региону, объему и мощности двигателя, данным по автомобилю и пр.

Описание переменных

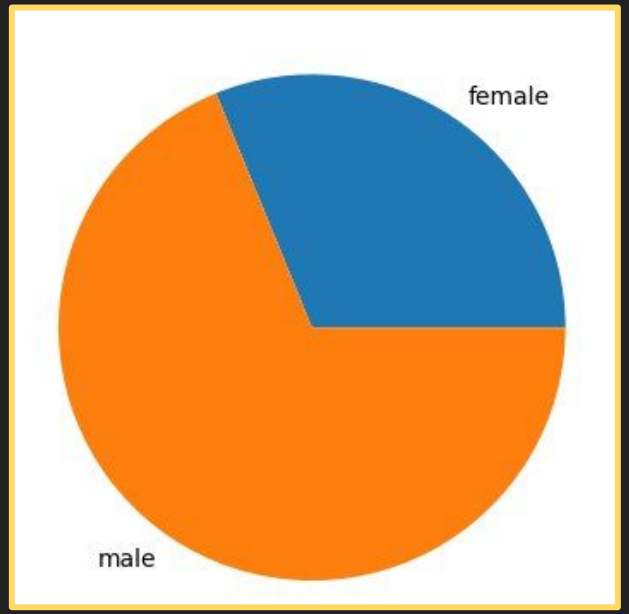
offenceshortstatement – самые частые статьи



1. Превышение скорости на 20-40 км/ч
2. Нарушение разметки
3. Не пристегнут ремень безопасности
4. Повторное превышение скорости на 40-60 км/ч
5. Превышение скорости на 40-60 км/ч

Описание переменных

gender_cd – пол



Примерно ⅔ водителей из датасета – мужчины (это примерно соответствует статистике* количества женщин и мужчин за рулем)

*https://carsweek.ru/news/News_in_the_world/1226945/

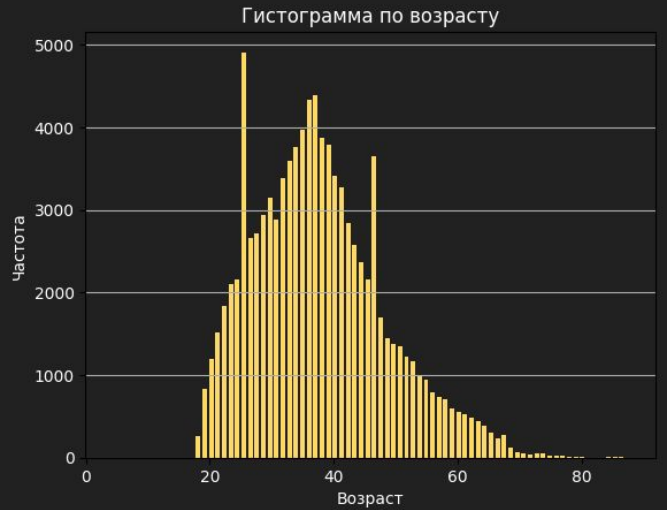
Описание переменных

public_holiday – выходные дни

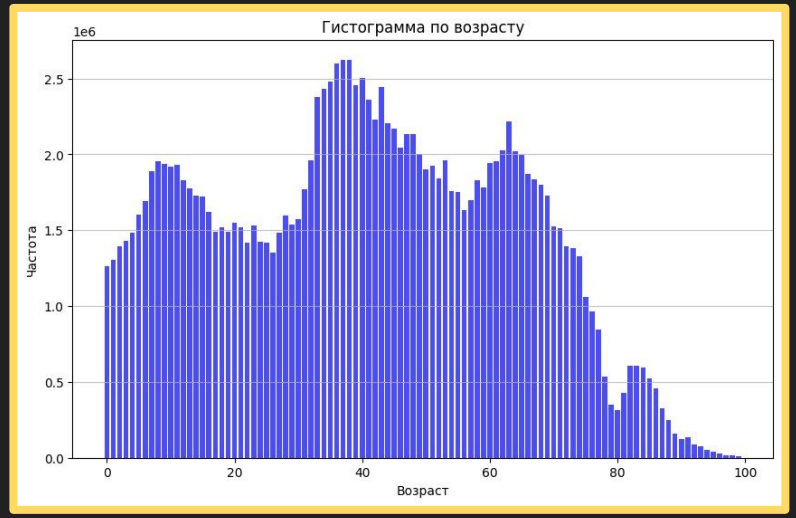
Примерно ¼ правонарушений происходит в праздники и выходные, а всего в рассматриваемом периоде 47% дней выходные или праздничные

Разведывательный анализ

Распределение по возрасту отличается, но по статистике* всего 7% водителей в год не получают штрафы => можем предположить, что выборка репрезентативна



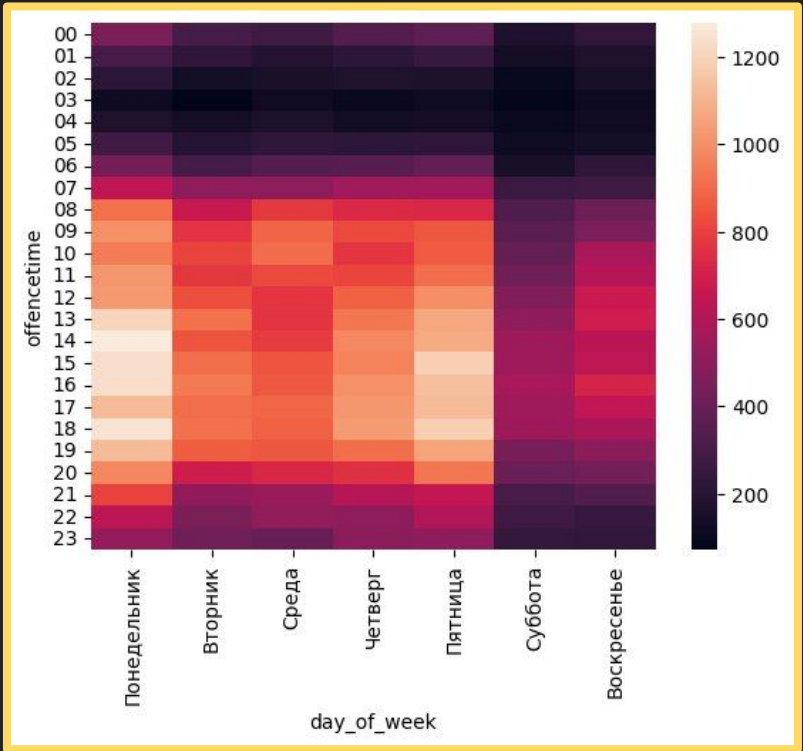
в датасете



в реальности

*https://kp-ru.turbopages.org/turbo/kp.ru/s/daily/27284.5/4420323/

Разведывательный анализ



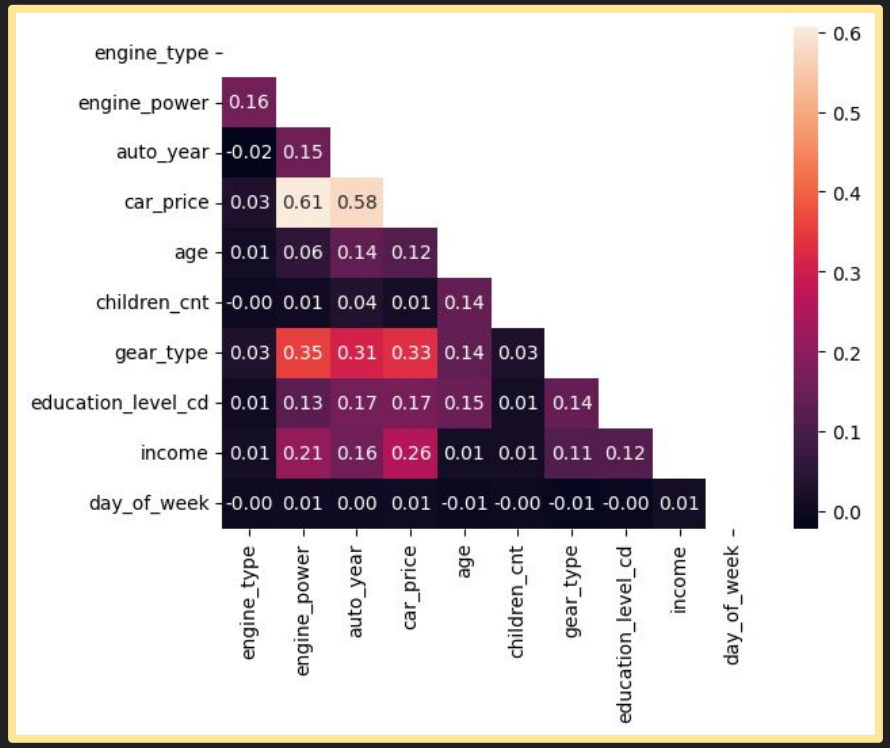
Из тепловой карты мы видим, что большинство правонарушений совершается в дневное время по будням

Разведывательный анализ

корреляционный анализ

Большинство корреляций очень слабые, но:

- цена машины коррелирует с мощностью двигателя и годом выпуска
- тип коробки передач коррелирует с мощностью двигателя, годом выпуска автомобиля и его ценой



Использованные переменные:

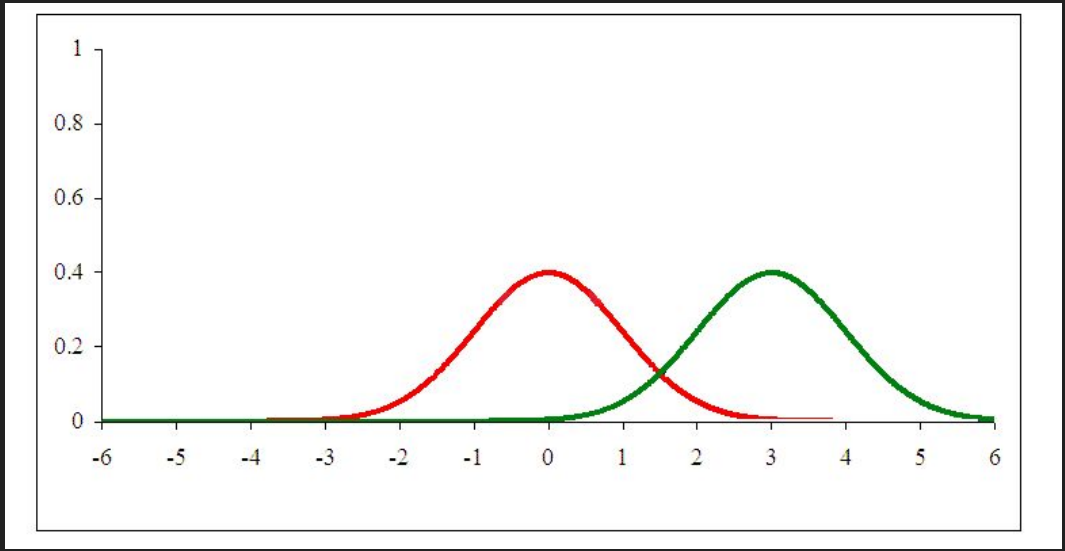
- age – возраст
- gender_cd – код пола водителя
- car_price – цена автомобиля
- engine_power – мощность двигателя
- auto_year – год выпуска автомобиля
- children_cnt – количество детей у водителя
- public_holiday – является ли данный день недели выходным
- education_level_cd – уровень образования водителя
- person_monthly_income_amt – размер месячного дохода водителя, у.е.
- gear_type – тип коробки передач

Обработка данных

- удалили str в числовых столбцах
- удалили несовершеннолетних и старше 100 лет
- удалили выбросы в доходах и ценах машин
- удалили большинство значений NaN или логически их заменили
- удалили неиспользуемые столбцы

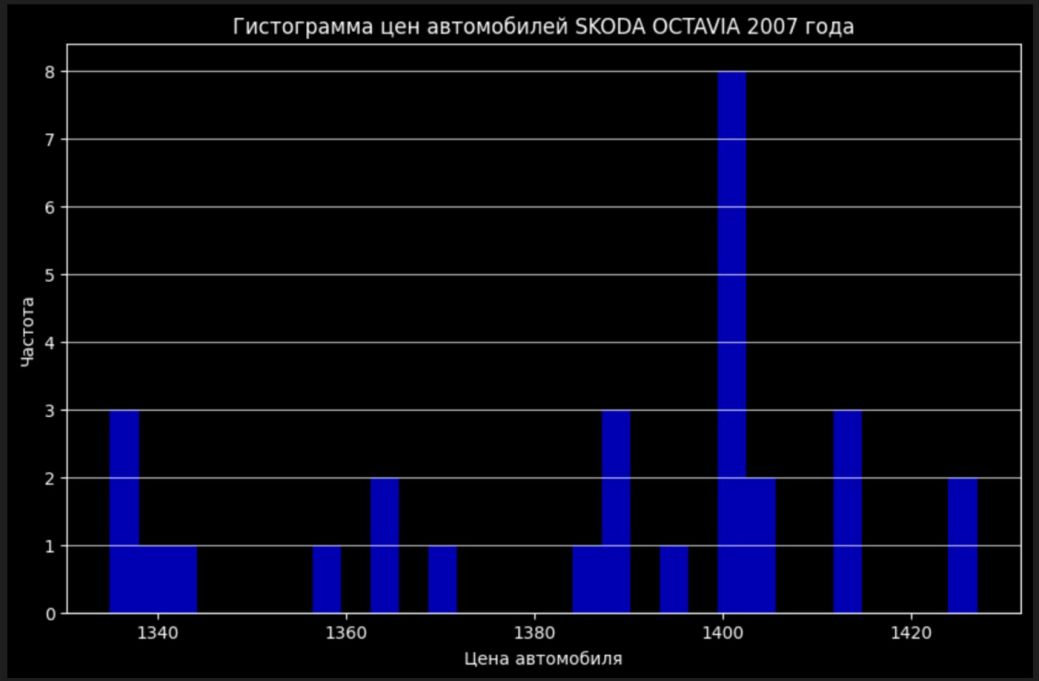
- перевели часть категориальных переменных в числовой вид
- почистили выбросы – неадекватные значения (например, “Oct-00” в данных о характеристиках машины)
- заменяли пропуски в данных на средние по возрастам (это корректно, потому что потом мы разбиваем данные на бакеты)

Условные единицы



Из-за изменения курса У. Е./рубли меняется распределение по доходу. Чтобы проверить репрезентативность выборки необходимо сравнить распределение по доходам в генеральной совокупности и в нашей выборке в одних единицах измерения

Условные единицы



Среднее значение: 1386.10
Стандартное отклонение: 27.13
Коэффициент вариации: 1.96%

Хорошей переменной для оценки конвертации У.Е. в рубли является цена машины, данная в У.Е.
В ней нет выбросов, и данные имеют очень маленькую дисперсию.

Условные единицы

Обоснование адекватности

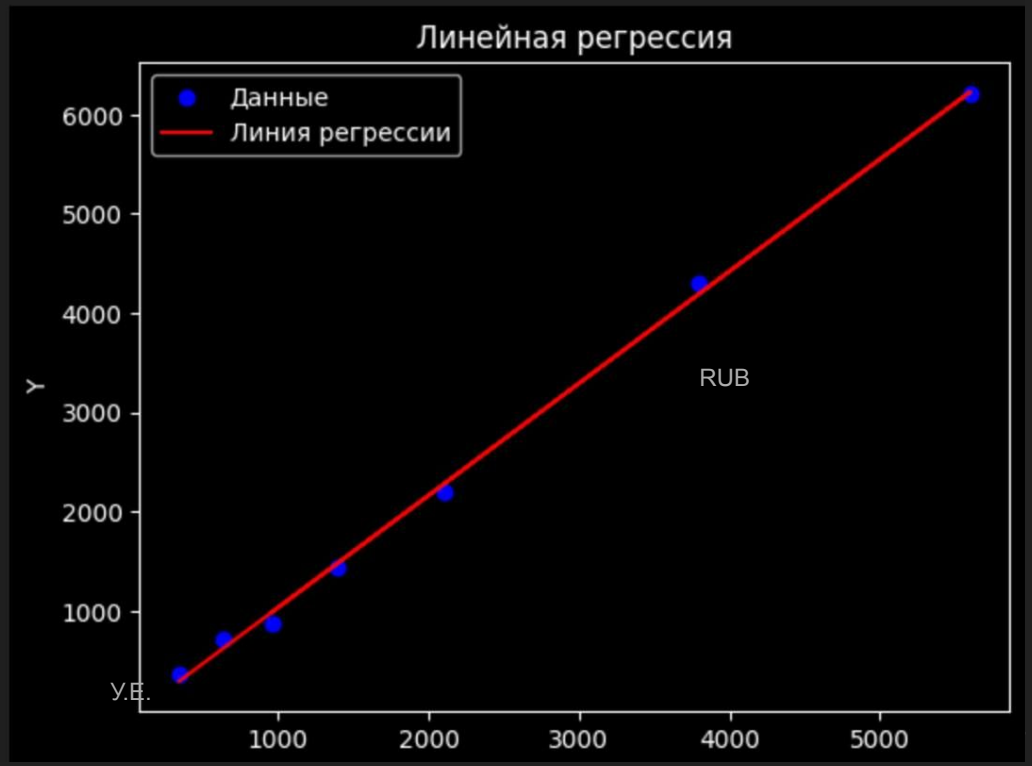
Мы выбрали 7 популярных моделей автомобилей в разных ценовых диапазонах и разных производителей и оценили их цены в у.е. в исходном Data Set и в рублях на реальном рынке

МОДЕЛЬ	У.Е.	RUB/1000
OCTAVIA 2007	344	360
KIA RIO 2012	640	720
AUDI A3 2015	1390	1450
HYUNDAI CRETA 2021	2100	2200
BMW 5 2021	5600	6200
MERCEDES E 2019	3800	4300
LADA VESTA 2018	960	870

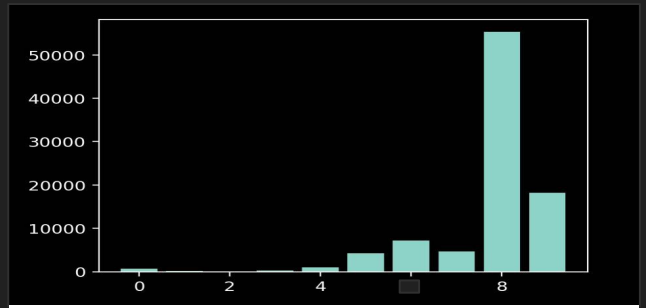
Условные единицы

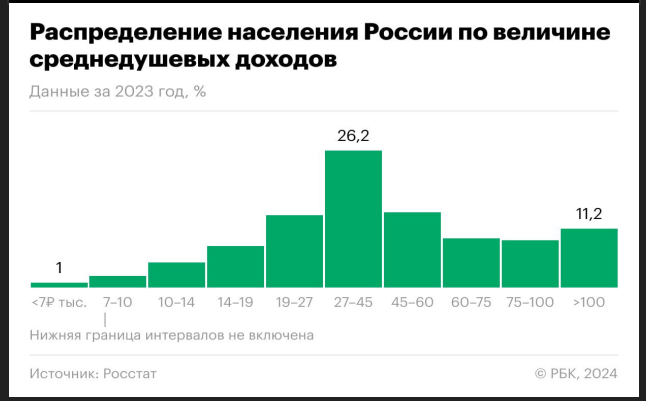
```
Коэффициент наклона (slope): 1.1282196431508815
Свободный член (intercept): -90.8585980714538
```

Цены в У.Е. очень похожи на цены в рублях, из чего можем сделать вывод, что можем конвертировать рубли в У.Е. по курсу 1/1



Условные единицы



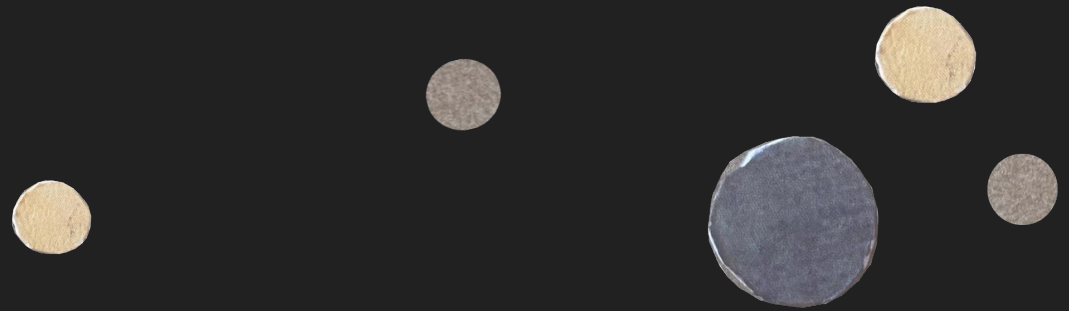


Распределение доходов в нашем датасете

Благодаря полученному выводу можно сравнить распределение людей по доходу в нашей выборке и в генеральной совокупности (Россия)

Мотивация

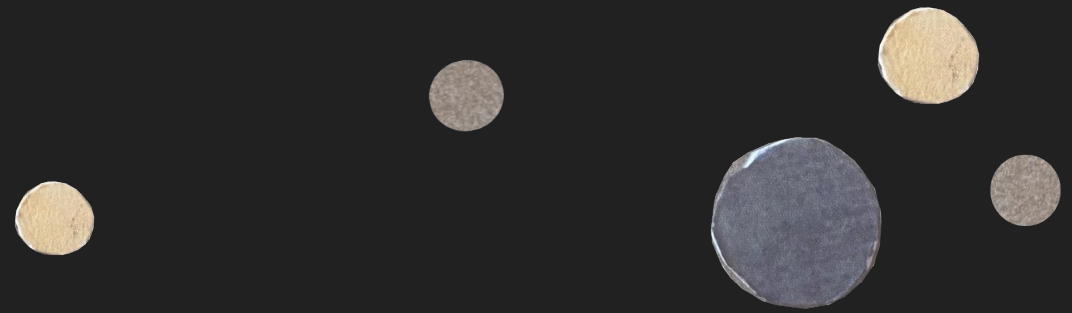
Исследование этих данных поможет выстраивать систему штрафов так, чтобы через систему мотивации снизить количество нарушений на дорогах



Исследовательский вопрос

?

Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений и если да, то как?



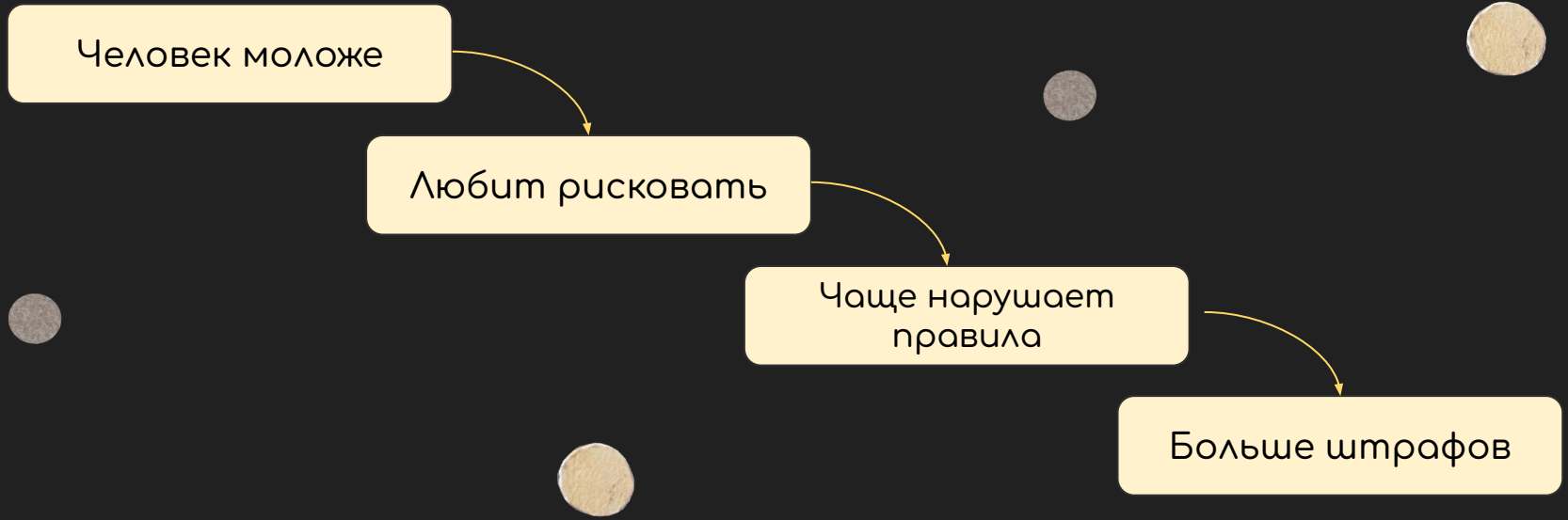
Гипотеза – 1

Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений и если да, то как?

Больше штрафов получаются молодёжью

Механизм – 1

Больше штрафов получаются молодёжью



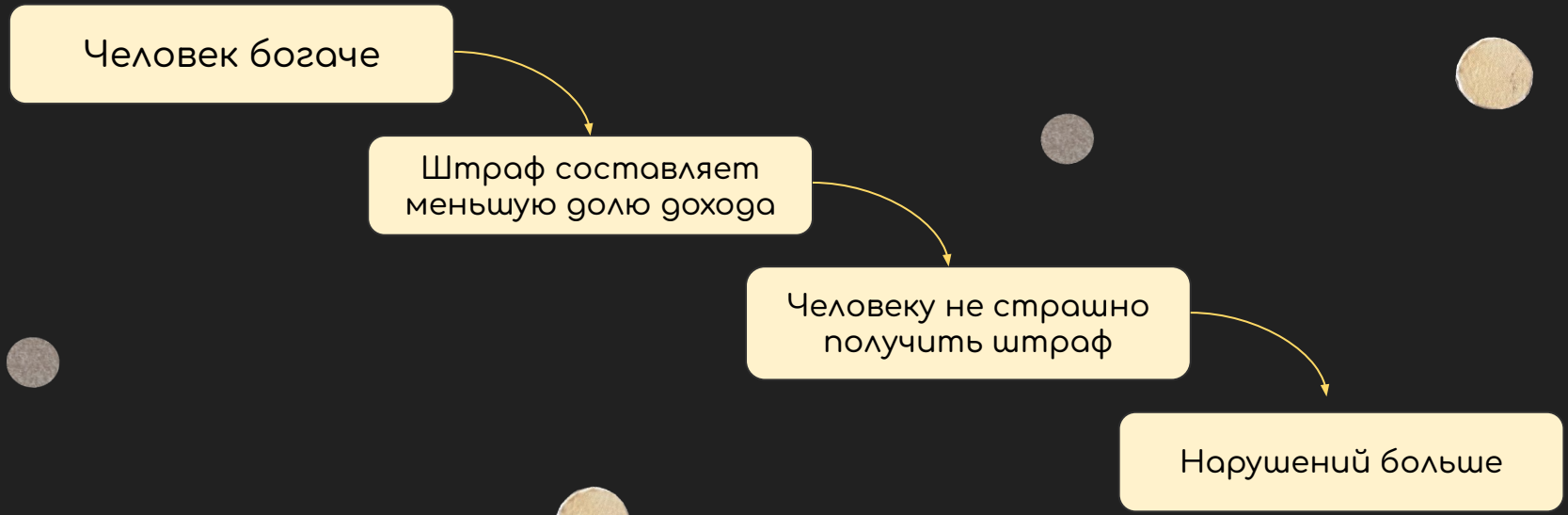
Гипотеза – 2

Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений и если да, то как?

Более богатые люди чаще получают штрафы

Механизм – 2

Более богатые люди чаще получают штрафы



Математическая модель

Больше штрафов получаются молодёжью

reg = 1000*(кол-во нарушений возраста age) / (кол-во людей возраста) =
= α*age + [$\gamma_1 \times \Delta_1$]

Δ_1 = {car_price_{avg}/1000; $engine\_power_{avg}$/100; ($auto\_year_{avg}$ – 2000); $gender\_cd_{avg, dummy}$; $children\_cnt_{avg}$; $education\_level\_cd_{avg, dummy}$; $public\_holiday_{avg, dummy}$; $gear\_type_{avg, dummy}$; $person\_monthly\_income\_amt_{avg}$}

γ - вектор коэффициентов контрольных переменных, Δ - вектор контрольных переменных данной регрессии

Математическая модель

Более богатые люди чаще получают штрафы

reg = (кол-во нарушений в i-ом персентиль/кол-во людей в i - ом персентиль) = $\alpha^{*}i - \beta^{*}car\_price_{avg}/1000 + [\gamma_2 \times \Delta_2]$

Δ_2 = { $engine\_power_{avg,\ f}/100$; $(auto\_year_{avg} - 2000)$; $gender\_cd_{avg,\ dummy}$; $children\_cnt_{avg}$; $education\_level\_cd_{avg,\ dummy}$; $public\_holiday_{avg,\ dummy}$; $gear\_type_{avg,\ dummy}$; $age_{avg}/100$}

γ - вектор коэффициентов контрольных переменных, Δ - вектор контрольных переменных данной регрессии

Результаты – 1

	coef	std err	t	P>|t|	[0.025	0.975]
x_age	-0.0193	0.004	-5.118	0.000	-0.027	-0.012
car_price	0.0010	0.158	0.006	0.995	-0.316	0.318
engine_power	0.0494	0.326	0.152	0.880	-0.602	0.701
auto_year	-0.3060	0.676	-0.452	0.653	-1.659	1.047
education_level_cd	-0.2675	0.104	-2.583	0.012	-0.475	-0.060
public_holiday	-0.2328	0.357	-0.652	0.517	-0.947	0.481
gear_type	1.1647	0.532	2.191	0.032	0.101	2.228
person_monthly_income_amt	0.4919	0.161	3.052	0.003	0.169	0.814
children_cnt	-0.7340	0.396	-1.853	0.069	-1.526	0.059
gender_cd	0.4893	0.360	1.357	0.180	-0.232	1.210
const	0.6753	0.714	0.946	0.348	-0.753	2.104

Выводы:

1. Гипотеза подтвердилась, т.е. социальная группа молодежи чаще игнорирует правила, поэтому регулярнее получает штрафы

Результаты – 1.1

	coef	std err	t	P>|t|	[0.025	0.975]
x_age	-0.0193	0.004	-5.118	0.000	-0.027	-0.012
car_price	0.0010	0.158	0.006	0.995	-0.316	0.318
engine_power	0.0494	0.326	0.152	0.880	-0.602	0.701
auto_year	-0.3060	0.676	-0.452	0.653	-1.659	1.047
education_level_cd	-0.2675	0.104	-2.583	0.012	-0.475	-0.060
public_holiday	-0.2328	0.357	-0.652	0.517	-0.947	0.481
gear_type	1.1647	0.532	2.191	0.032	0.101	2.228
person_monthly_income_amt	0.4919	0.161	3.052	0.003	0.169	0.814
children_cnt	-0.7340	0.396	-1.853	0.069	-1.526	0.059
gender_cd	0.4893	0.360	1.357	0.180	-0.232	1.210
const	0.6753	0.714	0.946	0.348	-0.753	2.104

Выводы:

1. Гипотеза подтвердилась, т.е. социальная группа молодежи чаще игнорирует правила, поэтому регулярнее получает штрафы

2.Получено свидетельство, что чем больше доход, тем больше штрафов получает человек

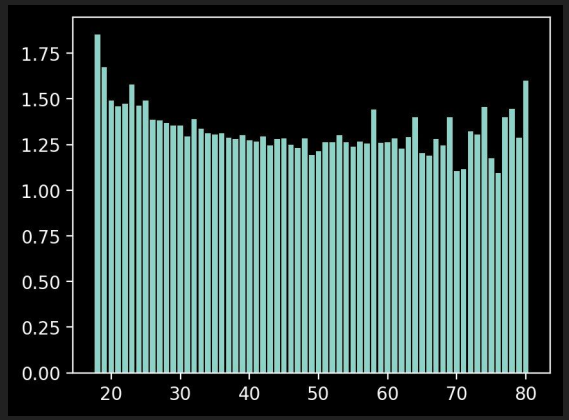
Результаты – 1.2

	coef	std err	t	P>|t|	[0.025	0.975]
x_age	−0.0065	0.001	−7.186	0.000	−0.008	−0.005
car_price	−0.0757	0.031	−2.403	0.019	−0.139	−0.013
public_holiday	0.5330	0.145	3.667	0.000	0.243	0.823
gear_type	−0.3332	0.114	−2.919	0.005	−0.561	−0.105
gender_cd	−0.8161	0.154	−5.309	0.000	−1.123	−0.509
const	2.4019	0.162	14.822	0.000	2.078	2.726

Omnibus:	19.576	Durbin-Watson:	1.728
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32.564
Skew:	1.005	Prob(JB):	8.49e-08
Kurtosis:	5.639	Cond. No.	837.

Изменение в формуле: делим на всех людей конкретного возраста в выборке, а не в генеральной совокупности и не умножаем на 1000

Зависимость очень слабая



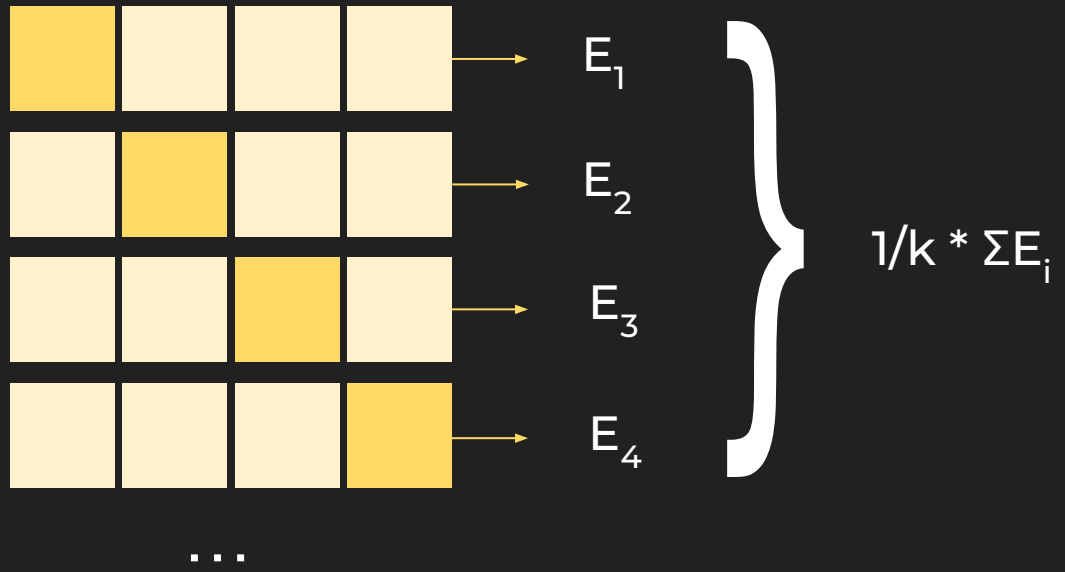
Переменная reg по возрастам

Интерпретация – 1

Линейная регрессия среднего числа нарушений по возрасту не показала значимой корреляции, при этом просто число нарушений показало значимую зависимость, соответственно по выборке нарушающих справедливо что каждый отдельный водитель, в среднем, нарушает одинаково часто независимо от возраста, при этом почти все водители получают штрафы. Соответственно люди определенного возраста не пере-представлены относительно выборки водителей в нашей выборке.

Проверка устойчивости – 1

K-fold cross validation



Проверка устойчивости – 1

K-fold cross validation

Мы провели K-fold для переменной reg по переменным age, gender_cd, gear_type, public_holiday
Получили для age. Средняя MSE: 0.03 Стандартное отклонение MSE: 0.03
для gender_cd. Средняя MSE: 0.03 Стандартное отклонение MSE: 0.03
для gear_type. Средняя MSE: 0.05 Стандартное отклонение MSE: 0.07
для public_holiday. Средняя MSE: 0.04 Стандартное отклонение MSE: 0.04

=> по всем переменным модель очень устойчива

Результаты – 2

	coef	std err	t	P>|t|	[0.025	0.975]
per	4.402e-05	0.001	0.051	0.959	-0.002	0.002
car_price	-0.0399	0.093	-0.427	0.671	-0.226	0.146
engine_power	0.5819	0.192	3.035	0.003	0.201	0.963
gender_cd	0.1601	0.200	0.799	0.426	-0.238	0.558
auto_year	-0.3416	0.273	-1.252	0.214	-0.884	0.201
education_level_cd	0.0720	0.120	0.602	0.549	-0.166	0.310
public_holiday	0.2200	0.428	0.514	0.608	-0.630	1.070
gear_type	-0.3487	0.312	-1.119	0.266	-0.968	0.270
children_cnt	-0.1494	0.178	-0.841	0.403	-0.502	0.204
const	0.9667	0.308	3.136	0.002	0.354	1.579

Omnibus:	3.941	Durbin-Watson:	1.569
Prob(Omnibus):	0.139	Jarque-Bera (JB):	3.279
Skew:	0.408	Prob(JB):	0.194
Kurtosis:	3.348	Cond. No.	3.68e+03

Корреляция частоты получения штрафов и относительного уровня дохода не является статистически значимой.

Интерпретация – 2

Линейная регрессия частоты штрафов по перцентилям дохода не показала статистически значимого результата. Соответственно люди с разными уровнями дохода нарушают с одинаковой частотой (аналогичный факт, что все они в той или иной форме представлены в данной выборке).Соответственно люди определенного уровня дохода не перепредставлены в нашей выборке.

Практическая польза

Мы можем сделать вывод о том, что выборка водителей, получивших штрафы, близка к тому, чтобы репрезентативно отражать совокупность всех клиентов Т-Банка, которые сравнительно **регулярно водят машину.**

Банк может использовать полученные результаты для дальнейшего анализа в других областях, так как эта база данных дает хорошее понимание о том, как устроена выборка всех водителей в стране.

Policy implication

Конкретные меры

Относительно полученных данных, банк может создать более опирающуюся на среднестатистических показателях экосистему в сервисе оплаты штрафов (Т-банк оплата), что будет впоследствии развивать клиентуру и предотвращать переход клиента к другому банку за счет привыкания пользователя к многофункциональному приложению.

Применение таргетированности рекламы - тем, кто не водит машину, сервис по уплате штрафов не рекламируется

Другие варианты механизма

Почему наши гипотезы не подтвердились?

Люди нарушают ПДД не из-за того, что склонны к этому, а из-за стимулов, и эти стимулы не меняются с возрастом. То же самое можно сказать про доход.

Это может быть связано с тем, что мы используем мелкие нарушения и люди их совершают скорее спонтанно.

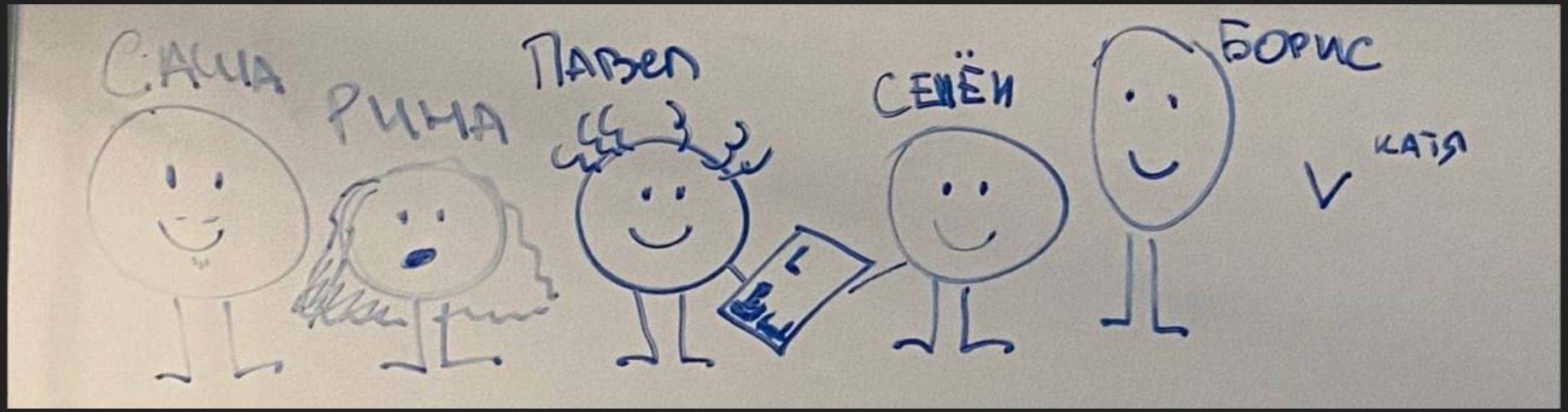
Ограничения и перспективы

1. Данные представлены за 1 месяц => не учитывается сезонность
2. Нет данных о возрастном распределении водителей вообще => не можем оценить, какая доля людей вообще не получает штрафы
3. Нет данных о более критичных правонарушениях (например, вождение в пьяном виде), потому что за них применяются другие санкции => нельзя увидеть полную картину нарушений
4. Выборка репрезентативна для клиентов Т-Банка, но не для всей России

↓

1. Рассмотреть данные за больший период времени (желательно за несколько лет, чтоб учесть тренд)
2. Собрать недостающие данные
3. Собрать данные о нарушениях с другими санкциями, чтоб получить более полное описание поведения водителей
4. Сделать поправку на смещенность выборки клиентов Т-Банка, чтобы экстраполировать результаты на всю Россию

Приложение

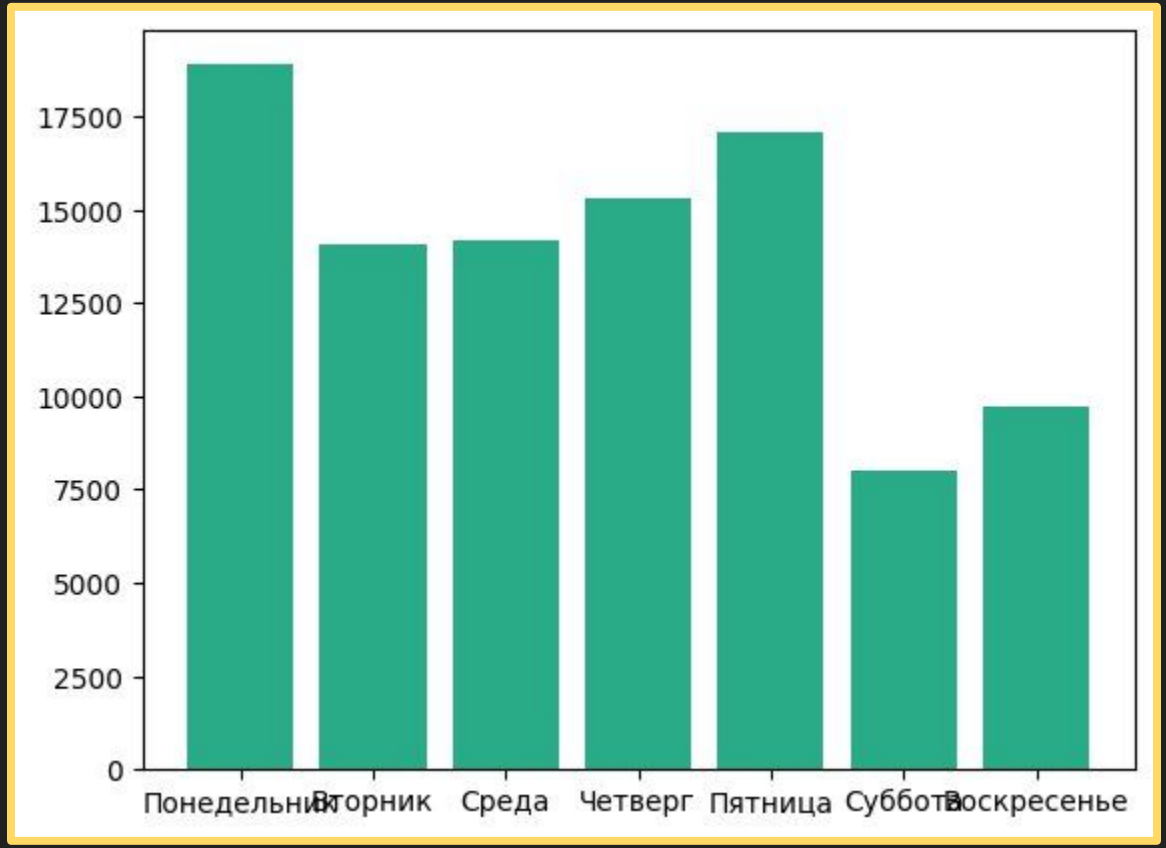


Команда 12:
Александр Штайн - тим лидер, аналитик и математик
Екатерина Столяренко - программист-аналитик и дизайнер
Павел Маляренко - программист-аналитик
Семен Киселев - программист-аналитик
Борис Шишаев - программист-аналитик
Екатерина Чичина - дизайнер

Превышение скорости на 20-40 км/ч 68512
Нарушение разметки 7936
Не пристегнут ремень безопасности 6339
Повторное превышение скорости на 40-60 км/ч 2706
Превышение скорости на 40-60 км/ч 1781
Использование телефона за рулем 1324
Движение по обочине 1322
Движение по выделенной полосе (Москва и Санкт-Петербург) 1178
Повторный проезд на запрещ.сигнал светофора 998
Пересечение стоп-линии 947
Проезд на красный сигнал светофора 910
Поворот (разворот) в запрещенном месте 607
Остановка или стоянка в неположенном месте 550
Нарушение правил пользования световыми приборами, звуковыми сигналами 405
Движение по выделенной полосе 384
Поворот не из крайней полосы 359

Превышение скорости на 60-80 км/ч 311
Повторное превышение скорости более чем на 60 км/ч 271
Выезд на полосу встречного движения или на трамвайные пути встречного направления 147
Не пропустил пешехода 117
Превышение скорости более чем на 80 км/ч 57
Проезд грузовых ТС в запрещенном месте (Москва и Санкт-Петербург) 46
Повторный выезд на полосу встречного движения или на трамвайные пути встречного направления 34
Остановка на перекрестке 25
Остановка или стоянка на трамвайных путях либо далее первого ряда от края проезжей части 11
Выезд на велосипедную или пешеходную дорожку 8
Разворот или движение задним ходом в запрещенном месте 6
Остановка или стоянка на пешеходном переходе или тротуаре 4
Нарушение правил проезда через железнодорожные переезды 3

Остановка или стоянка на автобусной остановке 3
Остановка или стоянка на месте, отведенном для ТС инвалидов 3
Остановка или стоянка в неположенном месте (Москва и Санкт-Петербург) 1
Движение во встречном направлении по дороге с односторонним движением 1
Остановка на автомагистрали за пределами места под стоянку 1


17500
15000
12500
10000
7500
5000
2500
0
Понедельник
Вторник
Среда
Четверг
Пятница
Суббота
Воскресенье

Способы подсчёта переменных:

1. $gender\_cd_{avg,\ dummy}$; $children\_cnt_{avg}$; $education\_level\_cd_{avg,\ dummy}$; $public\_holiday_{avg,\ dummy}$; $person\_monthly\_income\_amt_{avg}$- подсчитываются как средние этих значений по людям, находящимся в каждом бакете.
2. car_price_{avg}/1000; $engine\_power_{avg\ f}$/100; ($auto\_year_{avg}$– 2000) ; $gear\_type_{avg,\ dummy}$ - считаются как средние по всем нарушениям всех людей из бакета

‘dummy’ переменные:

1. $gender\_cd_{dummy}$: male = 1, female = 0
2. $education\_level\_cd_{dummy}$ =
   SCH =0
   UGR = 1
   SPC = 1.5
   GRD = 2
   PGR = 3
   ACD = 4
   MGR = 5
   AGR = 2
3. $public\_holiday_{dummy}$: yes = 1, no = 0
4. $gear\_type_{avg,\ dummy}$

```
from sklearn.model_selection import KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import numpy as np
import pandas as pd


df = pd.read_csv('data_finaly2.csv')

kf = KFold(n_splits=10, shuffle=True, random_state=42)

mse_scores = []

for train_index, test_index in kf.split(df):
    train_set = df.iloc[train_index]
    test_set = df.iloc[test_index]


    X_train = train_set['podstavliaem_peremen'].values.reshape(-1, 1)
    y_train = train_set['reg'].values
    X_test = test_set['podstavliaem_peremen'].values.reshape(-1, 1)
    y_test = test_set['reg'].values


    model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    mse_scores.append(mse)

    print(f'MSE для текущего фолда: {mse}')

mean_mse = np.mean(mse_scores)
std_mse = np.std(mse_scores)

print(f'Средняя MSE: {mean_mse}')
print(f'Стандартное отклонение MSE: {std_mse}')
```

K-fold

Кандидатская на соискание ученой степени по социальной психологии “Сравнительная соц-псих характеристика “бедные”, “богатые” как соц групп”

Научная статья по педагогике и психологии “Физиологические предрасположенности к риску”

Научная статья по социологии “Проявления отклоняющегося поведения у россиян”

	coef	std err	t	P>|t|	[0.025	0.975]
x_age	54.7060	10.147	5.391	0.000	34.370	75.042
car_price	0.2137	0.201	1.065	0.292	−0.189	0.616
engine_power	−0.3956	0.408	−0.971	0.336	−1.212	0.421
auto_year	−0.9542	0.864	−1.104	0.275	−2.687	0.778
education_level_cd	−0.0857	0.139	−0.616	0.540	−0.364	0.193
public_holiday	0.0845	0.476	0.178	0.860	−0.869	1.038
gear_type	0.7602	0.696	1.092	0.280	−0.635	2.155
person_monthly_income_amt	0.5999	0.187	3.207	0.002	0.225	0.975
children_cnt	−0.5167	0.547	−0.944	0.349	−1.613	0.580
gender_cd	0.3997	0.471	0.849	0.400	−0.544	1.344
const	−0.3073	1.095	−0.281	0.780	−2.502	1.887

Omnibus:	10.869	Durbin-Watson:	0.337
Prob(Omnibus):	0.004	Jarque-Bera (JB):	11.249
Skew:	0.819	Prob(JB):	0.00361
Kurtosis:	4.187	Cond. No.	771.

1/age

	coef	std err	t	P>|t|	[0.025	0.975]
x_age	67.4685	6.738	10.013	0.000	54.003	80.934
person_monthly_income_amt	0.3578	0.095	3.786	0.000	0.169	0.547
const	−1.1185	0.126	−8.888	0.000	−1.370	−0.867

Omnibus:	10.942	Durbin-Watson:	0.218
Prob(Omnibus):	0.004	Jarque-Bera (JB):	11.432
Skew:	0.814	Prob(JB):	0.00329
Kurtosis:	4.227	Cond. No.	260.