Тинькофф.Город

Сервис «Игры»

Команда хихи-квадрат

Структура данных

В датасете дана информация о покупках в сервисе «Игры» Тинькофф.Город, а также о клиентах, совершивших заказ.

Количественные:

- дата заказа
- цена товара
- количество товара в заказе
- позиция в топе популярных игр
- возраст клиента
- доход клиента

Качественные:

Информация о клиенте –

- id
- пол
- уровень образования
- город

Информация о заказе –

- id заказа
- id категории
- название категории
- id товара
- название товара

Структура данных

88 699 записей о покупках
41 298 уникальных пользователей

Из них:

- **36 778** мужчин *(89,06%)*
- **4 520** женщин *(10,04%)*

Пользователи от **1** до **86** лет
Медиана возраста – **26** лет

Данные с **01.01.2022** до **29.10.2023**

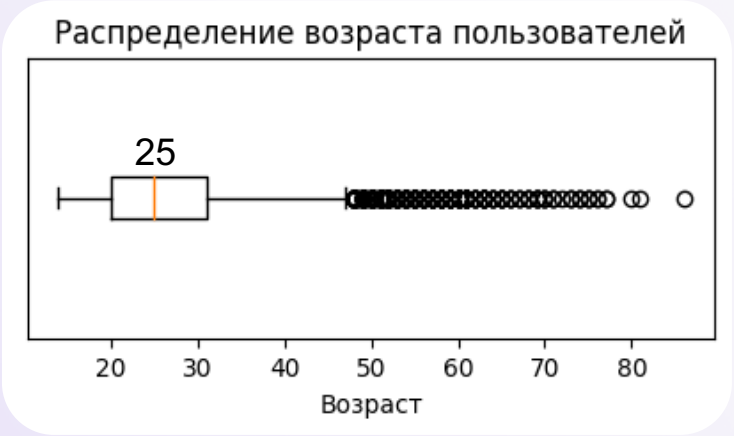
Топ-3 города:

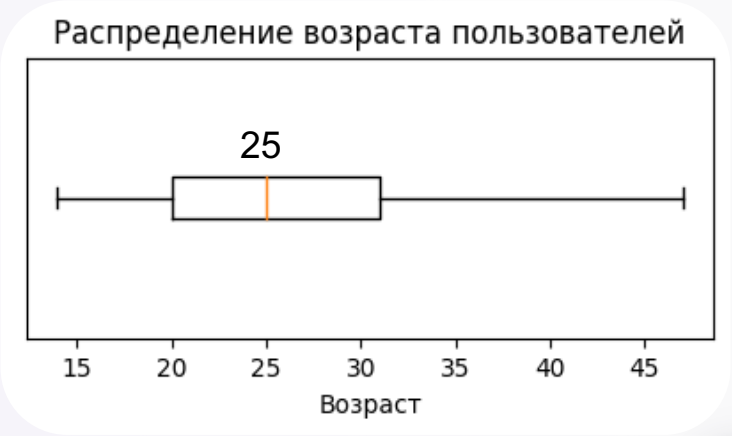
Москва, Санкт-Петербург, Екатеринбург

Обработка возраста

Убираем пользователей, чей возраст <14 лет, так как Тинькофф.Город – сервис для носителей дебетовой карты, минимальный возраст использования которой - 14 лет.

Также убираем пользователей с возрастом >47, они выбиваются из общего тренда.





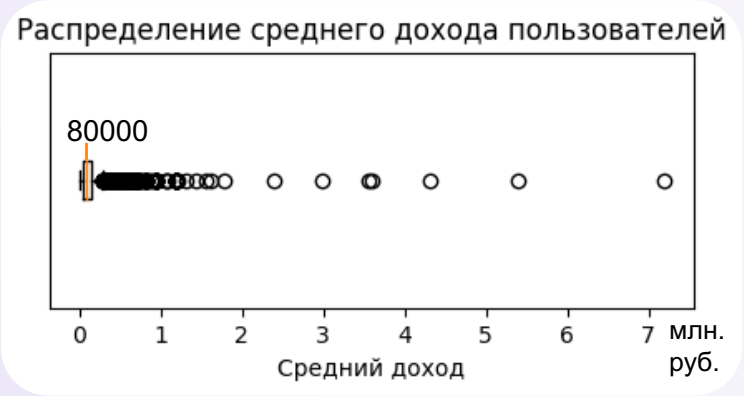
Всего при обработке возраста убрали **1850** пользователей *(4.48%)*

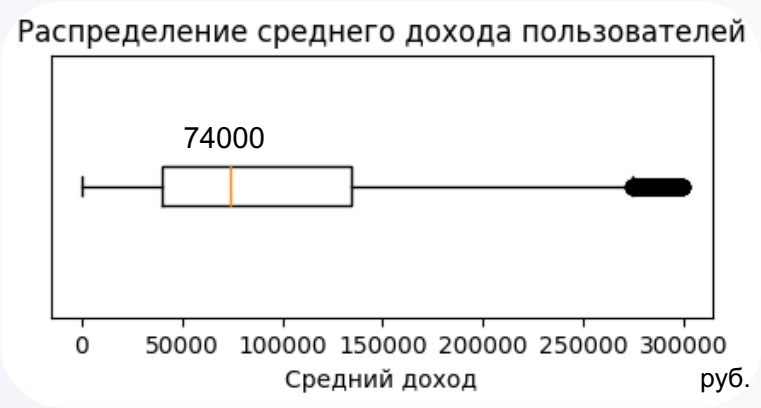
Обработка дохода

Записи с отрицательным месячным доходом - некорректные, удаляем их

Пользователей с отрицательным доходом - **7151** *(17.32%)*

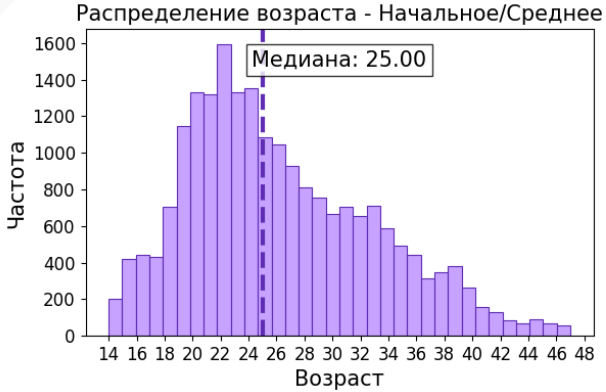
Удаляем пользователей с экстремальным значением среднего дохода

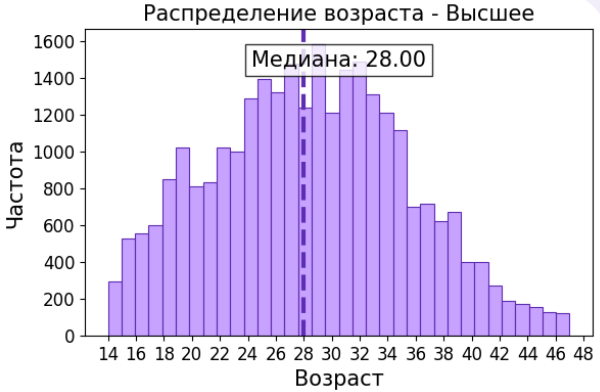


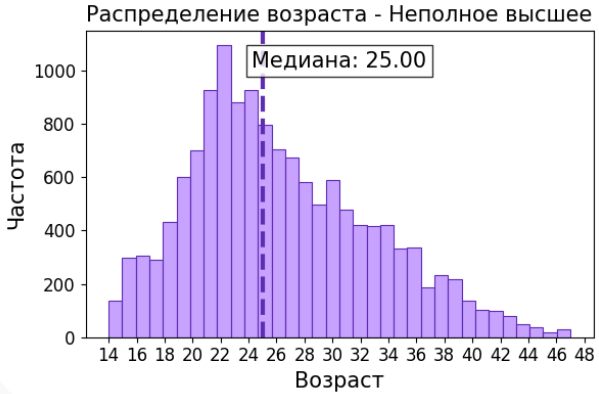


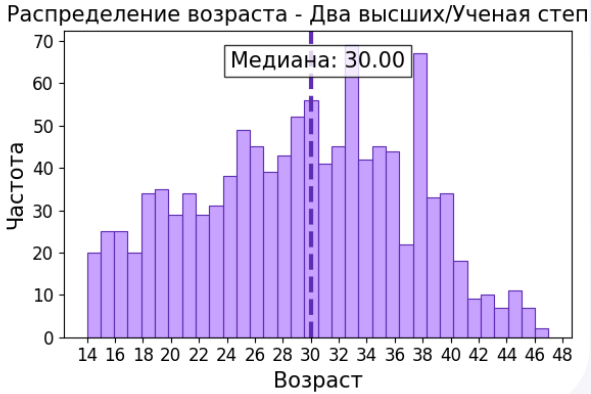
Верхняя граница: **298916** руб., выше нее удалено **1585** пользователей *(3.84%)*

Обработка образования









Заметно, что в графиках присутствуют клиенты с несопоставимым возрасту уровнем образования.

Пример – **87** людей в возрасте **<18 лет** имеют два высших образования или ученую степень.

Обработка образования

Так как данные для датасета собирались из других сервисов Тинькофф, то **вероятность** того, что в записи будет правильный возраст **выше**, чем правильный уровень образования.

Тогда, чтобы это исправить и оградить себя от неправильных выводов на основе некорректных данных, рассчитаем **минимальный** возраст получения каждого из уровней образований и присвоим клиентам подходящий уровень образования.

Начальное/среднее образование - **нет ограничений** (можно получить после 4 классов обучения)

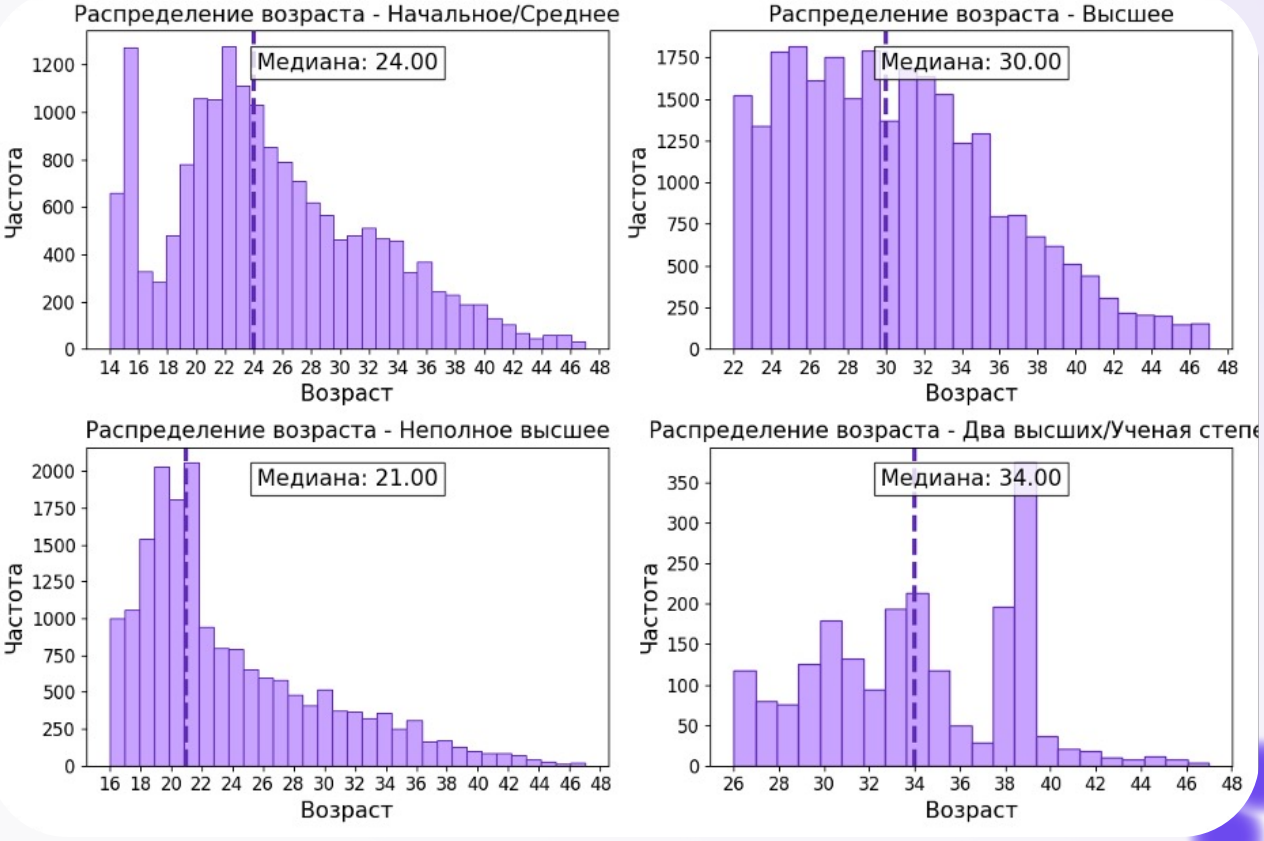
Неполное высшее - **16 лет** – поступление после 9 классов школы

Высшее образование - **22 год** - в 18 лет оканчивает школу + 4 года бакалавриата

2 высших - **26 года** - в 22 года первое высшее + 4 года бакалавриата

Ученая степень - **31 лет** – 18 лет + 4 года бакалавриата + 2 магистратуры + аспирантуры + 3 докторантуры.

Полученное распределение



Результаты обработки

Были удалены дубликаты покупок одной и той же игры конкретным пользователем.
6445 записей *(7.27%)*

Также были убраны пользователи, возраст которых варьировался в течение наблюдения более, чем на **3** года.
10 клиентов *(0.02%)*

После чистки осталось:
Записей в датасете: **64479**, *72.69%* от исходного
Уникальных клиентов: **30702**, *74.34%* от исходного

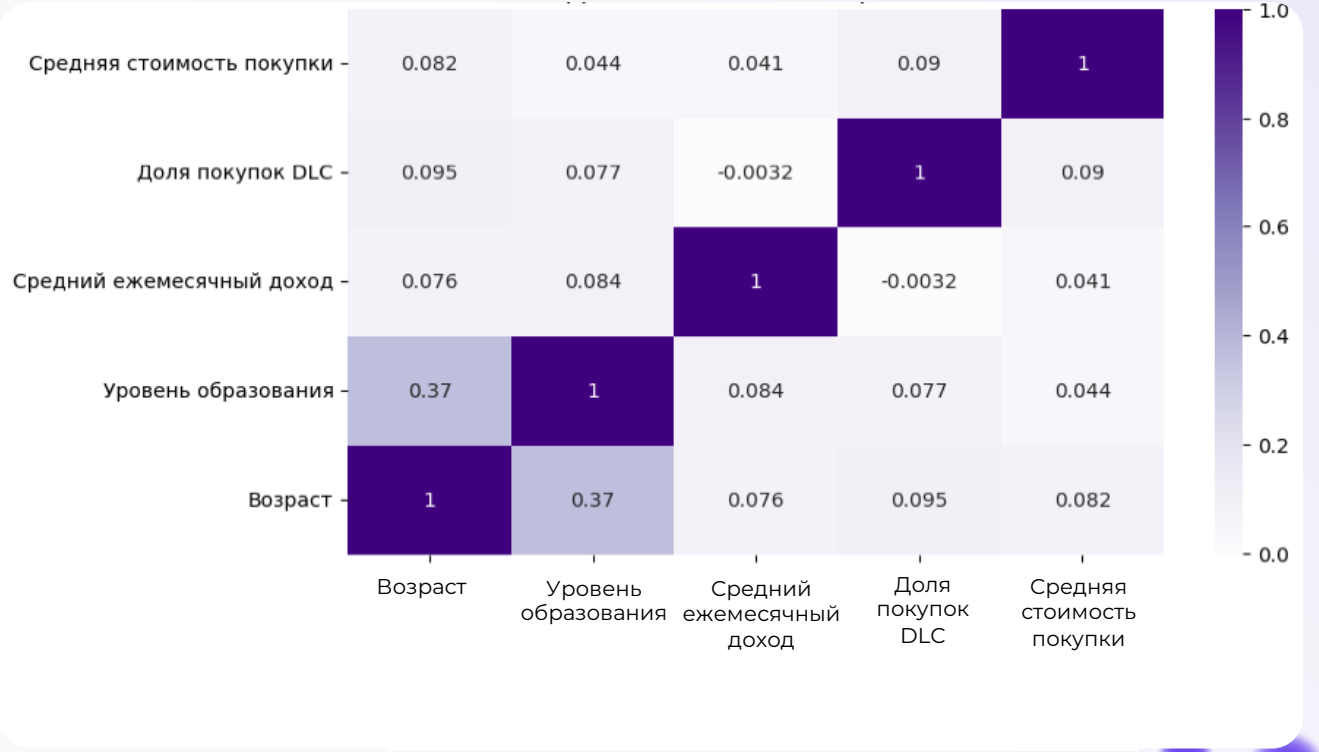
Вводим DLC

Смотрим на ключевые слова в названии игры и добавляем еще одно поле - тип товара**(DLC/игра).**

Ключевые слова для определения, что является DLC:
”DLC” - собственно само DLC
”edition” - издание
”pack” - пакет, набор
”collection” - коллекция
”bundle” - набор
”anthology” - антология

DLC (Downloadable Content) - дополнительный контент для компьютерных игр, который можно загрузить после выпуска основной версии игры.

Корреляция численных параметров



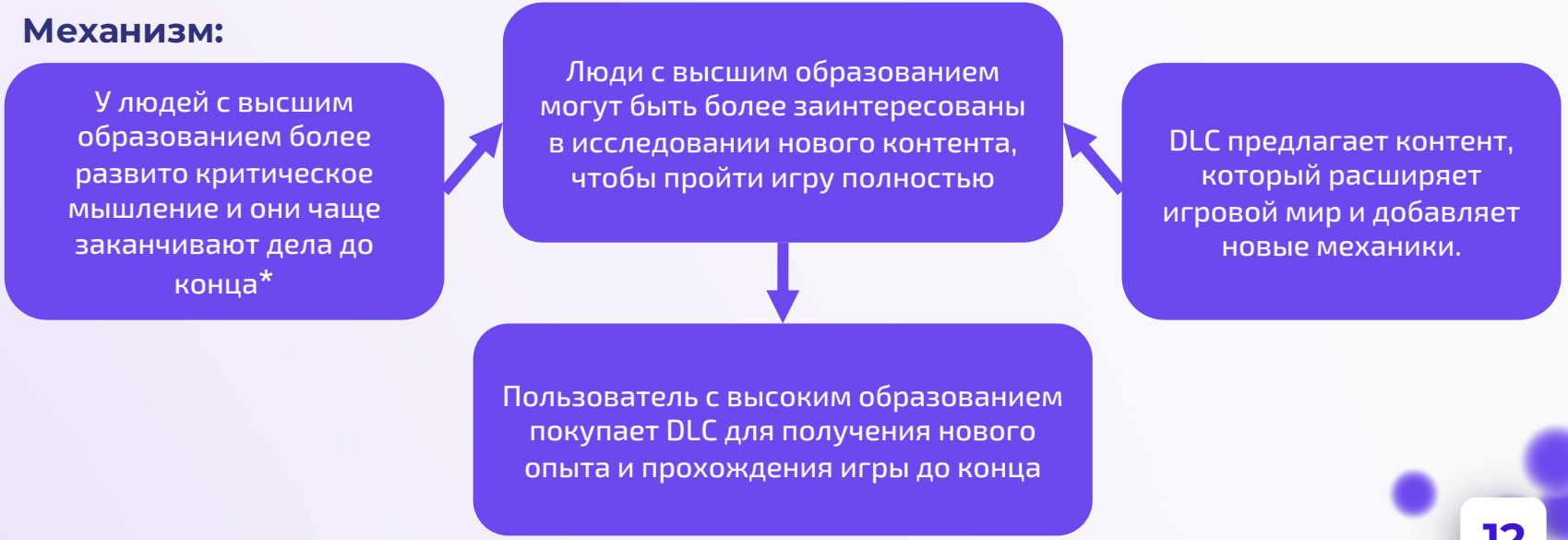
Исследовательский вопрос

Как демографические признаки человека влияют на то, насколько чаще он предпочитает купить дополнение к игре, чем новую игру?

Суть исследования

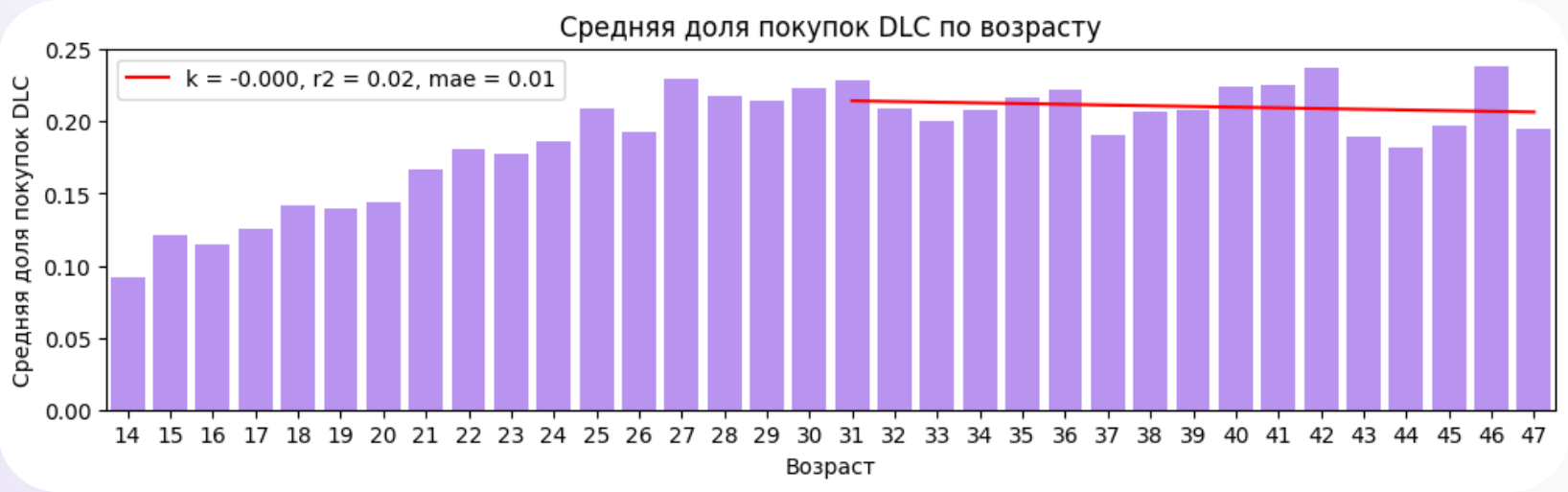
Гипотеза: Люди с высшим образованием чаще предпочтут купить DLC, чем люди без высшего образования.

Механизм:



*** - Исследование(Влияние высшего образования, Индиана Университи, Блумингтон)**

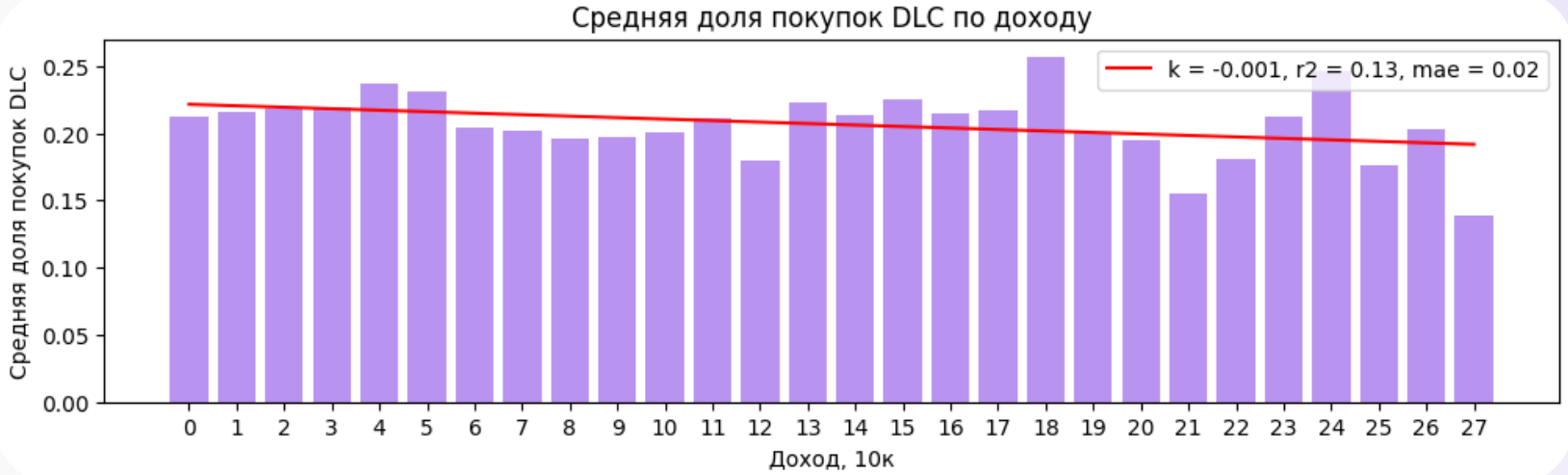
Зависимость от возраста



Среднее абсолютное отклонение(MAE) регрессии на промежутке **31-47 лет** составляет **0.01**, что довольно мало, поэтому регрессия хорошо описывает поведение данных.

t-тест Стьюдента показал p-value равное **0.203** (критерий значимости **0.05**). Значит, на этом промежутке средняя доля DLC не зависит от возраста.

Зависимость от дохода



Среднее абсолютное отклонение(MAE) регрессии на промежутке **31-47 лет** составляет **0.02**, что довольно мало, поэтому регрессия хорошо описывает поведение данных.

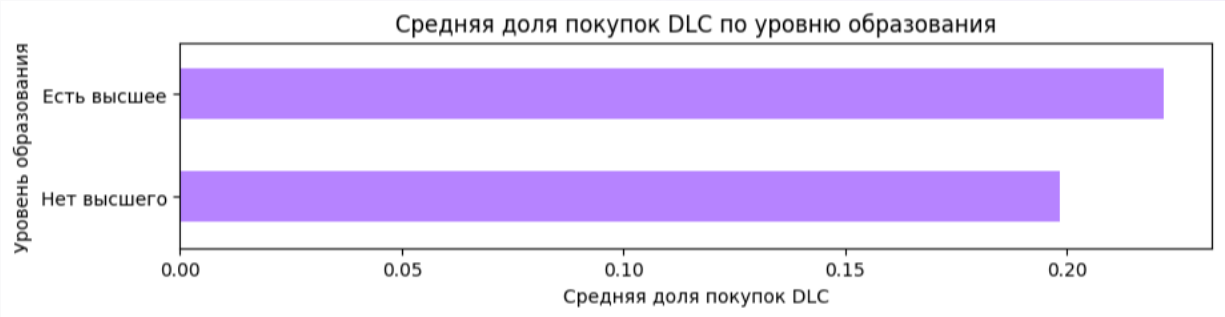
t-тест Стьюдента показал p-value равное **0.059** (критерий значимости **0.05**). Значит, на этом промежутке средняя доля DLC не зависит от дохода.

Проверка статистической значимости

Двухпропорционный односторонний z-тест

Уровень образования	Высшее	Нет высшего
Высшее	0.5	0.0056 z-value: 2.54
Нет высшего	0.0056 z-value: -2.54	0.5

Уровень образования	Высшее	Нет высшего
Высшее	-	>
Нет высшего	<	-



Проверка устойчивости модели

	2022	2023	Общее
P-value возраст	0.062	0.042	0.576
MAE возраст	0.02	0.01	0.01
P-value доход	0.630	0.290	0.059
MAE доход	0.02	0.04	0.02
Z-test ok?	✅	❌	✅

Итоги исследования

Гипотеза подтвердилась, но только для возрастной группы **31-47 лет.**

Ограничения:

- Данные охватывают ограниченный период времени (с **01-01-2022** по **29-10-2023**). Долгосрочные тенденции могут остаться незамеченными и результаты могут быть ограничены
- Много некорректных данных в базе, сделать более актуальными **доход и образование** клиентов

Перспективы исследования:

- Количество часов, проведенных в определенных играх у людей
- Платформа, на которой играет человек

Применение исследования

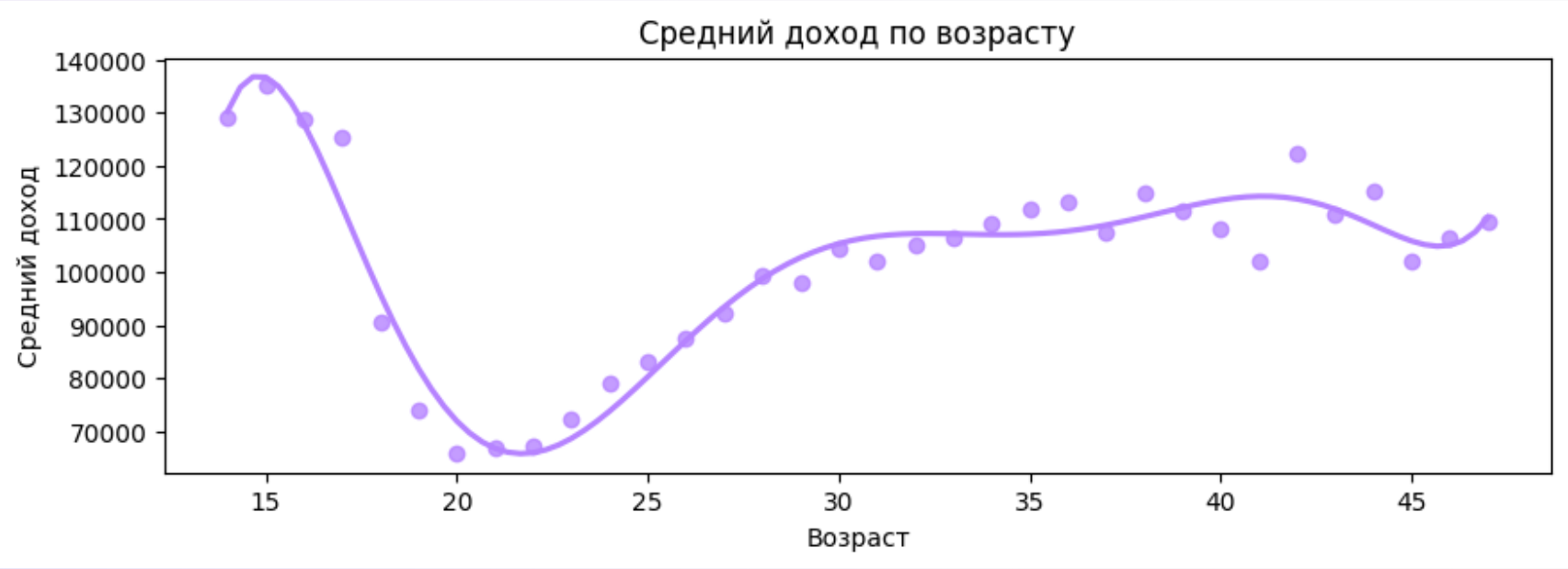
Практическая польза:
Найдена зависимость частоты покупки DLC с наличием у человека высшего образования, применимо для таргетированной маркетинговой компании.

Мы выявили, что чем выше у пользователя образование, тем чаще он покупает DLC. На основе этого можно делать (Policy Implication):

- **Таргетированная Реклама** - таргетированные рекламные кампании, направленные на клиентов без высшего образования, подчеркивая выгоды покупки DLC для глубокого погружения в игровой мир.
- **Привлечение** в сервис большего количества клиентов без высшего образования для дальнейшего совершения ими покупок на основе предложений рекламы.

Приложения

Приложения


Средний доход по возрасту
Средний доход
140000
130000
120000
110000
100000
90000
80000
70000
15
20
25
30
35
40
45
Возраст

Приложения

OLS Regression Results

Dep. Variable:	y	R-squared:	0.084
Model:	OLS	Adj. R-squared:	0.036
Method:	Least Squares	F-statistic:	1.739
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	0.203
Time:	18:48:11	Log-Likelihood:	58.564
No. Observations:	21	AIC:	-113.1
Df Residuals:	19	BIC:	-111.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2399	0.021	11.351	0.000	0.196	0.284
x1	-0.0007	0.001	-1.319	0.203	-0.002	0.000

Omnibus:	0.803	Durbin-Watson:	1.855
Prob(Omnibus):	0.669	Jarque-Bera (JB):	0.618
Skew:	0.390	Prob(JB):	0.734
Kurtosis:	2.687	Cond. No.	232.

OLS Regression Results

Dep. Variable:	dlc_to_all_ratio	R-squared:	0.130
Model:	OLS	Adj. R-squared:	0.097
Method:	Least Squares	F-statistic:	3.892
Date:	Wed, 20 Dec 2023	Prob (F-statistic):	0.0593
Time:	00:45:44	Log-Likelihood:	65.843
No. Observations:	28	AIC:	-127.7
Df Residuals:	26	BIC:	-125.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2215	0.009	25.165	0.000	0.203	0.240
income	-0.0011	0.001	-1.973	0.059	-0.002	4.63e-05

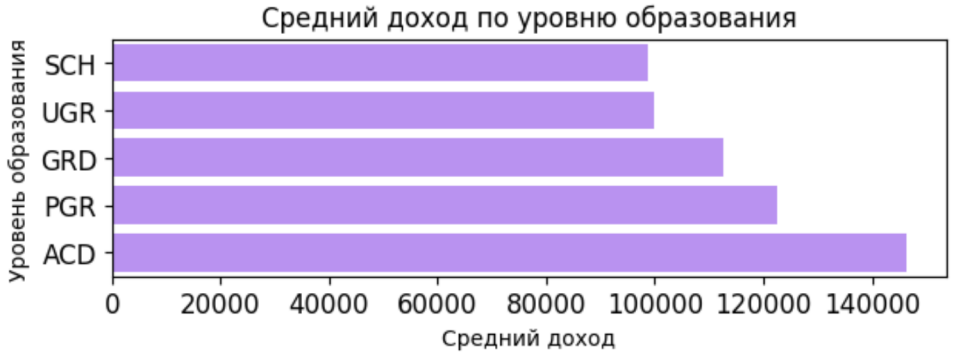
Omnibus:	2.052	Durbin-Watson:	1.575
Prob(Omnibus):	0.358	Jarque-Bera (JB):	0.841
Skew:	0.203	Prob(JB):	0.657
Kurtosis:	3.746	Cond. No.	30.7

Приложения


Распределение возраста клиентов
Количество клиентов
Возраст
1.32%
0.37%

Приложения


Средний возраст по уровню образования
Уровень образования
UGR
SCH
GRD
PGR
ACD
0
5
10
15
20
25
30
35
Средний возраст


Средний доход по уровню образования
Уровень образования
SCH
UGR
GRD
PGR
ACD
0
20000
40000
60000
80000
100000
120000
140000
Средний доход