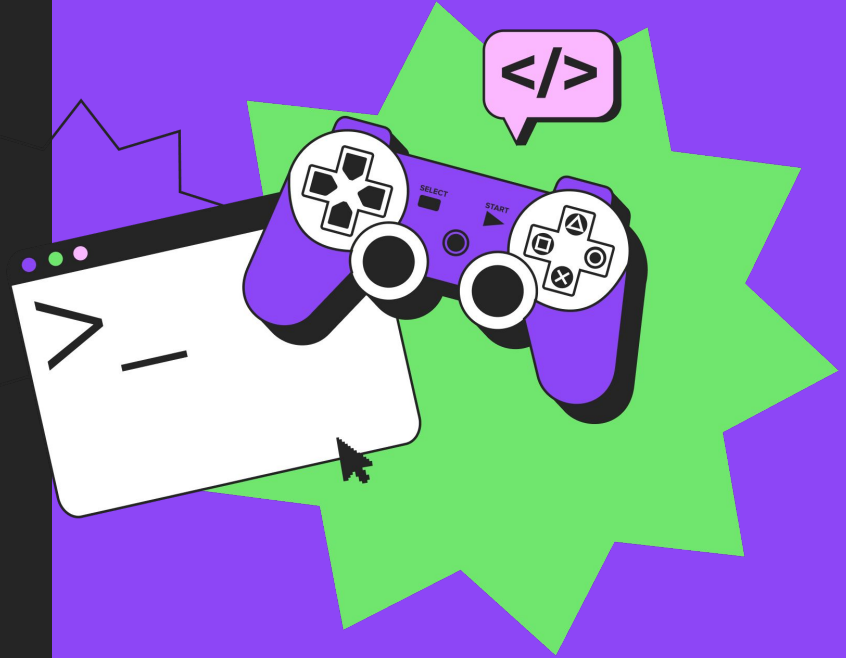





QnA





Что будет на уроке сегодня

-  Описание проекта
-  Этапы работы над проектом
-  Практическая часть





Вводные

Задача: провести анализ логов с сайта GB, рассчитать некоторые из метрик и выявить инсайты, которые могут быть полезны бизнесу.

Данные: логи сайта

session_id	идентификатор сессии
claim	признак наличия отправки формы заявки во время текущей сессии (0 – нет заявки, 1 – заявка отправлена)
60sec	признак длительности сессии 60 секунд и более (0 – сессия менее 60 сек, 1 – сессия 60 сек и более)
scroll_90	признак вертикальной прокрутки страницы глубиной 90% (0 – прокрутка менее 90%, 1 – прокрутка 90% и более)
hit_date	дата сессии
referer_url	адрес реферера страницы просмотра
url	адрес страницы просмотра
utm_source	utm-метка источника
utm_medium	utm-метка канала
gender	пол посетителя из куки (-1 – не определен, 0 – мужской, 1 – женский)
age	возраст посетителя из куки (-1 – не определен)
touch_screen	touch экран (0 – не определен, 1 – нет, 2 – есть)
has_vk_id	наличие идентификатора профиля пользователя Вконтакте (0 – нет, 1 – есть)
has_ok_id	наличие идентификатора профиля пользователя Одноклассников (0 – нет, 1 – есть)



Этапы работы над проектом

1. Загрузка данных. Проверка корректности типов данных
2. Исследовательский анализ данных
3. Статистический анализ данных
4. Предобработка данных (обработка пропусков, аномалий)
5. Построение воронки клиентов
6. Выявление инсайтов в данных



Загрузка данных. Проверка корректности типов данных

df.info() — выводит общую информацию о датафрейме.

`df['column_name'].astype('int')`

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values



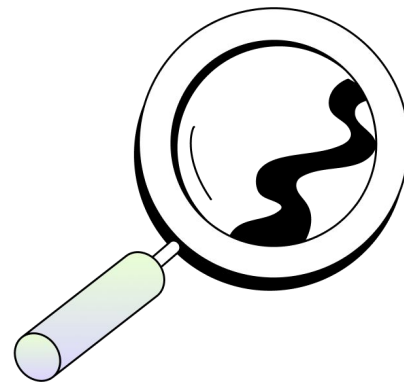
Исследовательский анализ данных

df.describe() — выводит статистические характеристики датафрейма.

`df['column_name'].value_counts()`

Построение графиков:

- Статичные графики — matplotlib, seaborn
- Интерактивные графики — plotly



Статистический анализ данных

- `df.describe()`
- `df.agg()`
- `df.groupby()`
- boxplots (“ящики с усами”)
- `.mean()`, `.median()`, `.mode()`, `.quantile()` и т. д.

`df[df['age'] > 0].groupby("touch_screen")['age'].mean()`





Предобработка данных (категоризация страниц)

re — модуль для работы с регулярными выражениями.

- re.match()
- re.search()
- re.findall()
- re.split()
- re.sub()
- re.compile()



Методы работы со строками

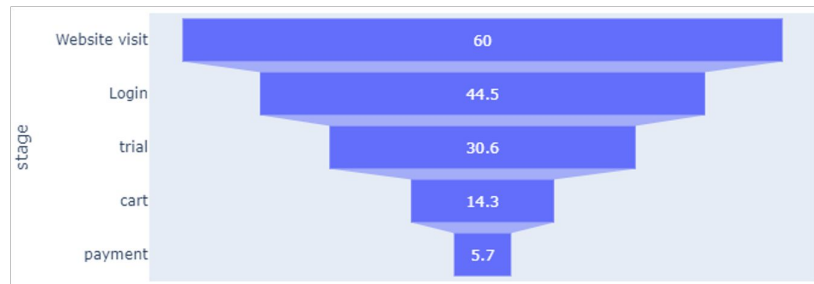


Построение воронки клиентов

```
import plotly.express as px

data = dict(
    number=[60, 44.5, 30.6, 14.3, 5.7],
    stage=["Website visit", "Login", "trial", "cart",
"payment"])

fig = px.funnel(data, x='number', y='stage')
fig.show()
```





На какие вопросы ищем ответы



Какие действия клиенты GB совершают чаще всего?



Есть ли аномалии в поведении пользователей: например, большое число коротких (менее 60 с.) сессий?
С чем эти аномалии могут быть связаны?



На каких страницах прокрутка 90% и более?



В какие даты было больше всего сессий? Рассчитайте DAU, MAU, WAU.



Какие изменения на сайте могут улучшить показатели: число посещений, длительность сессии и другие?



Полезные ссылки/дополнительные материалы



[Understanding Boxplots](#)



[Предварительная обработка данных](#)



[Что такое воронка продаж](#)



[Анализ данных с использованием Python](#)



Итоги урока



Описание проекта



Этапы работы над проектом



Практическая часть



Вопросы?

Вопросы?



Вопросы?

