

образовательные_проекты:

Этот репозиторий предназначен для проектов, которые я делала в ходе обучения на курсе [«Аналитик данных»](#) от **Яндекс Практикум**

Программа курса направлена на плавное изучение необходимых в работе аналитика навыков, последовательное наращивание сложности изучаемых тем и используемых инструментов. Данный курс подготовил меня к вступлению на путь анализа данных, помог мне овладеть необходимыми для работы навыками: выгружать, преобразовывать и очищать данные с помощью SQL-запросов, рассчитывать ключевые метрики работы компании и оценивать их значимость, запускать A/B-тестирования для проверки гипотез, помогать бизнесу принимать решения на основе данных, создавать дашборды и презентации.

В проектах использовались такие библиотеки как:

```
numpy pandas scipy sqlalchemy matplotlib warnings plotly seaborn datetime sys io requests  
math sklearn itertools
```

Часть 1. Введение в профессию «Аналитик данных». Основы Python и анализа данных

Охватываемые темы:

1. Переменные и типы данных. Вывод данных и арифметические операции.
2. Строки.
3. Списки.
4. Цикл for.
5. Вложенные списки.
6. Условный оператор. Цикл while.
7. Функции.
8. Словари.
9. Библиотека pandas.
10. Предобработка данных.
11. Анализ данных и оформление результатов.
12. Jupyter Notebook — тетрадь в ячейку.

Проект: музыка больших городов:

Исследование проведено на реальных данных Яндекс.Музыки. Проверены данные и проведено сравнение поведения пользователей двух столиц.

В ходе проекта были:

-
- определены и обработаны пропуски;
 - заменены типы данных на соответствующие;
 - удалены дубликаты;
 - проведены логическая индексация, группировка, сортировка.

Часть 2. Предобработка данных

Охватываемые темы:

1. Работа с пропусками.
2. Определение аномальных значений.
3. Преобразование типов данных.
4. Основные методы поиска дубликатов.
5. Работа с несовершенными реальными наборами данных.

Проект: исследование надёжности заёмщиков — анализ банковских данных:

На основе данных кредитного отдела банка исследовано влияние семейного положения, количества детей, уровня дохода и цели кредита на факт погашения в срок.

В ходе проекта были:

-
- определены и обработаны пропуски;
 - заменены типы данных на соответствующие;
 - удалены дубликаты
 - категоризованы данные;
 - один датафрейм декомпозирован на три.

Часть 3. Исследовательский анализ данных

Охватываемые темы:

1. Визуализация данных с помощью гистограмм и ящиков с усами.
2. Изучение срезов данных.
3. Нахождение взаимосвязей разных параметров в данных.
4. Объединение таблиц.
5. Получение выводов по сгруппированным данным.

Проект: продажа квартир в Санкт-Петербурге — анализ рынка недвижимости:

На основе данных сервиса Яндекс.Недвижимость определена рыночная стоимость объектов недвижимости разного типа, типичные параметры квартир, в зависимости от удаленности от центра.

В ходе проекта были:

- проведена предобработка данных;
- добавлены новые данные;
- построены гистограммы, боксплоты, диаграммы рассеивания.

Часть 4: Статистический анализ данных

Охватываемые темы:

1. Изучение объектов и их взаимосвязей методами статистики.
2. Выборки и статистическая значимость.
3. Выявление и обработка аномалий.

Проект: определение выгодного тарифа для федерального оператора сотовой связи:

На основе данных клиентов оператора сотовой связи проанализировано поведение клиентов и проведен поиск оптимального тарифа (проведен предварительный анализ использования тарифов на выборке клиентов, проанализировано поведение клиентов при использовании услуг оператора, рекомендованы оптимальные наборы услуг для пользователей).

В ходе проекта были:

- проведена предобработка данных, их анализ;
- проверены гипотезы о различии выручки абонентов разных тарифов и различии выручки абонентов из Москвы и других регионов.

Часть 5: Сборный Проект 1: Изучение закономерностей, определяющих успешность игр:

Выявлены параметры, определяющие успешность игры в разных регионах мира. На основании этого подготовлен отчет для магазина компьютерных игр для планирования рекламных кампаний.

В ходе проекта были:

- проведена предобработка данных, анализ;
- выбран актуальный период для анализа;
- составлены портреты пользователей каждого региона
- проверены гипотезы: средние пользовательские рейтинги платформ Xbox One и PC одинаковые; средние пользовательские рейтинги жанров Action и Sports разные (при анализе использовал критерий Стьюдента для независимых выборок).

Часть 6: Анализ бизнес-показателей

Охватываемые темы:

1. Метрики и воронки.
2. Когортный анализ.
3. Юнит-экономика.
4. Пользовательские метрики.

Проект: Анализ эффективности привлечения клиентов:

Задача для маркетингового аналитика развлекательного приложения Procrastinate Pro+. Несмотря на огромные вложения в рекламу, последние несколько месяцев компания терпит убытки. Задача — разобраться в причинах и помочь компании выйти в плюс.

В ходе проекта были:

-
- проведена предобработка данных;
 - проведен анализ данных от ProcrastinatePRO+;
 - рассчитаны различные метрики, использован когортный анализ: LTV, CAC, Retention rate и т.д. (использованы уже написанные ранее функции расчёта метрик);
 - сделаны выводы по полученным данным.

Часть 7: Принятие решений в бизнесе

Охватываемые темы:

1. Методы и инструменты проверки гипотез.
2. Проектирование экспериментов. Сезонность.
3. Когортный анализ.
4. A/B-тестирование.

Проект: Проверка гипотез по увеличению выручки в интернет-магазине — оценить результаты A/B теста:

Используя данные интернет-магазина приоритезированы гипотезы, произведена оценка результатов A/B-тестирования различными методами.

В ходе проекта были:

-
- проведена приоритизация гипотез по фреймворкам ICE и RICE;
 - проведен анализ результатов A/B-теста, построены графики кумулятивной выручки, среднего чека, конверсии по группам;

- посчитана статистическая значимость различий конверсий и средних чеков по сырым и очищенным данным;
- на основании анализа результатов теста принятие решения о целесообразности дальнейшего его проведения.

Часть 8: Как рассказать историю с помощью данных

Охватываемые темы:

1. Презентация результатов аналитического исследования.
2. Способы наглядного представления данных.
3. Создание отчётов, объясняющих выводы аналитика.
4. Библиотека Seaborn.

Проект: Исследования рынка общепита в Москве для принятия решения об открытии нового заведения:

Исследование рынка общественного питания на основе открытых данных, подготовка презентации для инвесторов. Будет ли успешным и популярным на долгое время кафе, в котором гостей обслуживают роботы-официанты.

В ходе проекта были:

-
- проведена предобработка данных;
 - по результатам анализа подготовлена презентация для инвесторов с рекомендациями.

Часть 9: Сборный Проект 2: Анализ пользовательского поведения в мобильном приложении:

На основе данных использования мобильного приложения для продажи продуктов питания проанализирована воронка продаж, а также произведена оценка результатов A/A/B-тестирования.

В ходе проекта были:

-
- проведена предобработка данных;
 - изучены принципы событийной аналитики, построена воронка продаж, исследован путь пользователей до покупки;
 - проанализированы результаты A/B-теста введения новых шрифтов;
 - проведено сравнение 2 контрольных групп между собой и с тестовой группой;
 - установлено, повлияет ли новый шрифт на поведение пользователей.

Часть 10: Базовый SQL

Охватываемые темы:

1. Введение в базы данных.
2. Срезы данных в SQL.
3. Агрегирующие функции. Группировка и сортировка данных.
4. Взаимоотношения между таблицами. Типы объединения таблиц.
5. Подзапросы и временные таблицы.

Проект: Исследование данных об инвестициях венчурных фондов в компании-стартапы:

В самостоятельном проекте этого курса работа идёт с базой данных, которая хранит информацию о венчурных фондах и инвестициях в компании-стартапы. Эта база данных основана на датасете Startup Investments, опубликованном на популярной платформе для соревнований по исследованию данных Kaggle.

В ходе проекта были:

-
- произведены различные выгрузки данных с помощью SQL-запросов;
 - произведены расчеты бизнес-показателей с помощью SQL-запросов.

Часть 10.1: Продвинутый SQL

Охватываемые темы:

1. Расчёт бизнес-показателей.
2. Основы оконных функций.
3. Когортный анализ.
4. Установка и настройка базы данных и клиента базы данных.
5. Рамки в оконных функциях. Дополнительные инструменты SQL.

Проект: Исследование данных о вопросах и ответах, постах и пр., размещенных на платформе StackOverflow:

В самостоятельном проекте этого курса работа идёт с базой данных StackOverflow — сервиса вопросов и ответов о программировании. StackOverflow похож на социальную сеть — пользователи сервиса задают вопросы, отвечают на посты, оставляют комментарии и ставят оценки другим ответам. Работа осуществлялась с версией базы, где хранятся данные о постах за 2008 год, но в таблицах есть информация и о более поздних оценках, которые эти посты получили.

В ходе проекта были:

- произведены различные выгрузки данных с помощью SQL-запросов и анализ результатов.

Часть 11: Автоматизация

Охватываемые темы:

1. Автоматизация процессов анализа данных.
2. Поточные аналитические решения.
3. Регистрация событий в логах, создание регулярных отчетов.
4. Дашборды. Мониторинг.

Проект: Создание дашборда по пользовательским событиям для агрегатора новостей:

Используя данные Яндекс.Дзена построен дашборд с метриками взаимодействия пользователей с карточками статей.

В ходе проекта были:

- произведена выгрузка данных с помощью SQL-запросов;
- проведен анализ данных и сохранение файла для дашборда в формате .csv;
- для визуализации данных создан дашборд с несколькими фильтрами на платформе Tableau public;
- по результатам подготовлена презентация с полученными графиками.

Часть 12: Основы машинного обучения

Охватываемые темы:

1. Задачи машинного обучения в бизнесе.
2. Алгоритмы машинного обучения.
3. Процесс решения задач машинного обучения.

Проект: Прогнозирование вероятности оттока пользователей для фитнес-центров:

На основе данных о посетителях сети фитнес-центров спрогнозирована вероятность оттока для каждого клиента в следующем месяце, сформированы с помощью кластеризации портреты пользователей.

В ходе проекта были:

- спрогнозирована вероятность оттока (на уровне следующего месяца) для каждого клиента;
- сформированы типичные портреты пользователей: выделены наиболее яркие группы, охарактеризованы их основные свойства;
- проанализированы основные признаки, наиболее сильно влияющие на отток;
- предложены рекомендации для стратегии взаимодействия с клиентами и их удержания.

Выпускной проект:

На основе всех полученных данных в курсе выполнен буткемп-проект в области "мобильные приложения":

Часть 1. **Определение неэффективных операторов колл-центров:**

На основе данных, содержащих информацию о колл-центрах и их операторах, найдены самые неэффективные/эффективные операторы, а также проверены статистические гипотезы.

В ходе проекта были:

- проведен обзор и предобработка данных;
- проведен исследовательский анализ данных;
- определены неэффективные операторы;
- проверены статистические гипотезы:
 - исходящие звонки колл-центров чаще пропускают чем отвечают на них;
 - при входящих обращениях из-за долгого времени ожидания часто не дожидаются назначения оператора.

Часть 2. **Тестирование изменений (А/В тест), связанных с внедрением улучшенной рекомендательной системы в интернет-магазине:**

На основе данных с действиями пользователей, технического задания и нескольких вспомогательных датасетов проведена оценка результатов A/B-теста изменений, связанных с внедрением улучшенной рекомендательной системы товаров интернет-магазина.

В ходе проекта были:

- проведен обзор данных и предобработка данных;
- проведена оценка корректности проведения теста и исследовательский анализ данных;
- проведен анализ A/B-теста.

Часть 3. **SQL. Анализ базы данных сервиса для чтения книг по подписке:**

Проанализирована база данных, содержащая информацию о книгах, издательствах, авторах, а также пользовательские обзоры книг. На основе этих данных сформулированы ценностные предложения для нового продукта.

В ходе проекта были:

- исследованы таблицы базы данных;
- произведены различные выгрузки данных с помощью SQL-запросов и анализ результатов.