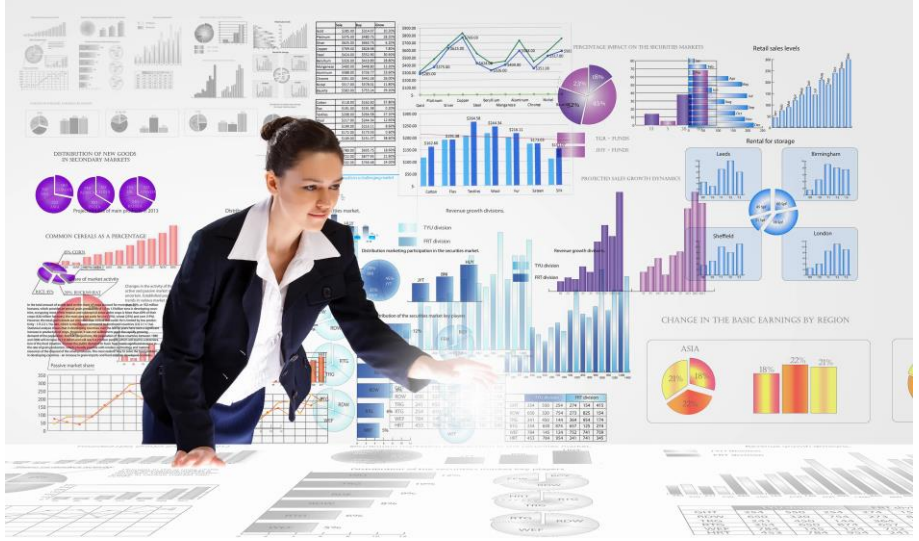
https://itstep.by/



Курсовой проект (Аналитик Данных) **Data Analyst**

Dr. Sergey Postnikov

Сергей Постников

Iteration 5, 03-Sep-2023, Sergey P.

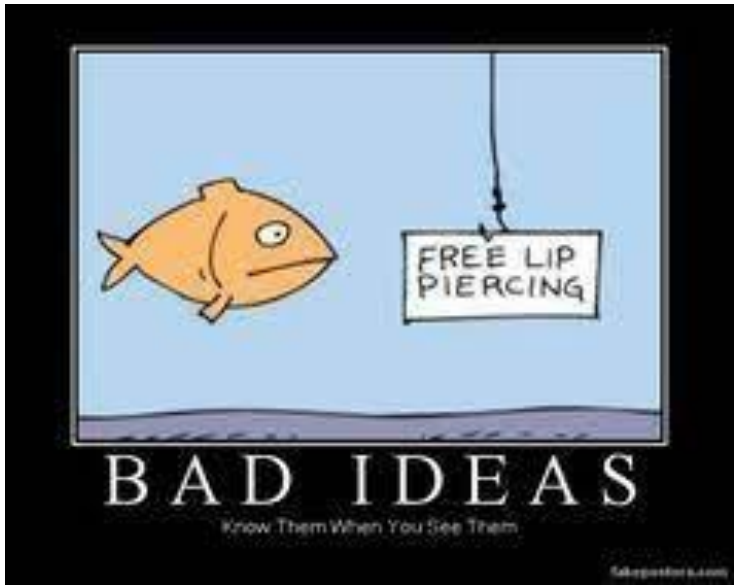
Сроки и формат, результаты и отчеты

- Доклад и «защита» - **9 ноября**
- **Презентация** ключевых результатов в PowerPoint (или PDF) для диалога (отобранные визуализации)
- Финальный Jupyter **Notebook**
  - с основными **этапами** анализа данных
- **Дашборд** отчета (BI)
- История проекта на **GitHub** (git commits)
- Набор изначальных, очищенных и полученных **данных**

(все на **GitHub**)

Предзащита идеи – 21 сентября требования

- Презентация **идеи** проекта PowerPoint (или PDF) для обратной связи
- Какой основной **вопрос**?
- Какая конечная **цель**?
- Какие **данные** нужны?
- **Название** проекта
- **Инструменты и методы**
- **Планирование (срок)**



Минимальные требования финала

- Присутствуют все 7 этапов анализа данных
- Каждый этап детализирован и акцентированы его моменты
- Какие из 4 типов аналитики использованы, как и где.
- Вычислены базовые статистические величины
- Указаны источники данных и их типы, описание начальных
- Использован Python и его библиотеки (pandas, matplotlib)
- Достигнута поставленная цель или ответ на вопрос.
- Лаконичная и понятная визуализация в BI и отчете
- Аргументация достигнутых выводов и рекомендации (??)

Web scraping

Хотя вы не найдете недостатка в отличных (и бесплатных) общедоступных наборах данных в Интернете, вы можете показать потенциальным работодателям, что вы также можете находить и очищать свои собственные данные. Кроме того, знание того, как очищать веб-данные, означает, что вы можете находить и использовать наборы данных, которые соответствуют вашим интересам, независимо от того, были ли они уже скомпилированы.

Если вы немного знакомы с Python, вы можете использовать такие инструменты, как Beautiful Soup или Scrapy, для поиска в Интернете интересных данных. Если вы не умеете кодировать, не волнуйтесь. Вы также найдете несколько инструментов для автоматизации процесса (многие предлагают бесплатную пробную версию), например Octoparse или ParseHub.

Если вы не знаете, с чего начать, вот несколько веб-сайтов с интересными вариантами данных, которые могут вдохновить ваш проект:

Реддит

Википедия

Порталы вакансий

Совет: каждый раз, когда вы собираете данные из Интернета, не забывайте уважать и соблюдать условия обслуживания каждого веб-сайта. Ограничьте свои действия по парсингу, чтобы не перегружать серверы компании, и всегда указывайте свои источники, когда вы представляете свои данные в своем портфолио.

Пример веб-скрейпинга: Тодд В. Шнайдер из Wedding Crunchers изучил около 60 000 свадебных объявлений New York Times, опубликованных с 1981 по 2016 год, чтобы измерить частотность определенных фраз.

Data cleaning

Значительная часть вашей роли в качестве аналитика данных заключается в очистке данных, чтобы подготовить их к анализу. Очистка данных (также называемая очисткой данных) — это процесс удаления неправильных и дублирующихся данных, устранения любых пробелов в данных и обеспечения согласованности форматирования данных.

Когда вы ищете набор данных для практики очистки, ищите тот, который включает в себя несколько файлов, собранных из нескольких источников без особого контроля. Некоторые сайты, на которых вы можете найти «грязные» наборы данных для работы, включают:

CDC Wonder

Данные правительства

Всемирный банк

Data.world

/r/datasets

Пример проекта по очистке данных. В этой статье на Medium рассказывается, как аналитик данных Раахим Хан очистил набор ежедневно обновляемой статистики популярных видео на YouTube.

Exploratory data analysis (EDA)

Анализ данных — это ответы на вопросы с помощью данных. Исследовательский анализ данных, или сокращенно EDA, поможет вам понять, какие вопросы задавать. Это можно сделать отдельно от очистки данных или вместе с ней. В любом случае, во время этих ранних расследований вы захотите сделать следующее.

1. Задавайте много вопросов о данных.
2. Откройте для себя базовую структуру данных.
3. Ищите тенденции, закономерности и аномалии в данных.
4. Проверяйте гипотезы и проверяйте предположения о данных.
5. Подумайте, какие проблемы вы потенциально могли бы решить с помощью данных.

Пример исследовательского проекта по анализу данных: этот аналитик данных взял существующий набор данных об американских университетах в 2013 году из Kaggle и использовал его для изучения того, почему студенты предпочитают один университет другому.

10 free public datasets for EDA

- An EDA project is an excellent time to take advantage of the wealth of public datasets available online. Here are 10 fun and free datasets to get you started in your explorations.
- 1. National Centers for Environmental Information**:** Dig into the world's largest provider of weather and climate data.
- 2. World Happiness Report 2021**:** What makes the world's happiest countries so happy?
- 3. NASA**:** If you're interested in space and earth science, see what you can find among the tens of thousands of public datasets made available by NASA.
- 4. US Census**:** Learn more about the people and economy of the United States with the latest census data from 2020.
- 5. FBI Crime Data Explorer (CDE)**:** Explore crime data collected by more than 18,000 law enforcement agencies.
- 6. World Health Organization COVID-19 Dashboard**:** Track the latest coronavirus numbers by country or WHO region.
- 7. Latest Netflix Data**:** This Kaggle dataset (updated in April 2021) includes movie data broken down into 26 attributes.
- 8. Google Books Ngram**:** Download the raw data from the Google Books Ngram to explore phrase trends in books published from 1960 to 2015.
- 9. NYC Open Data**:** Discover New York City through its many publicly available datasets on topics like the Central Park squirrel population to motor vehicle collisions.
- 10. Yelp Open Dataset**:** See what you can find while exploring this collection of Yelp user reviews, check ins, and business attributes.

Sentiment analysis

Анализ настроений, обычно выполняемый на текстовых данных, представляет собой метод обработки естественного языка (NLP) для определения того, являются ли данные нейтральными, положительными или отрицательными. Его также можно использовать для обнаружения определенной эмоции на основе списка слов и соответствующих им эмоций (известных как лексикон).

Этот тип анализа хорошо работает с сайтами общедоступных обзоров и платформами социальных сетей, где люди могут высказывать свое мнение по различным вопросам.

Чтобы начать изучать, что люди думают об определенной теме, вы можете начать с таких сайтов, как:

- Amazon (обзоры продуктов)
- Гнилые помидоры (рецензии на фильмы)
- Фейсбук
- Твиттер
- Новостные сайты

Пример проекта по анализу настроений: в этом сообщении в блоге Towards Data Science исследуется использование лингвистических маркеров в твитах для диагностики депрессии.

Data visualization

Люди — визуальные существа. Это делает визуализацию данных мощным инструментом для преобразования данных в убедительную историю, побуждающую к действию. Отличные визуализации не только забавны в создании, но и способны сделать ваше портфолио красивым.

Пример проекта визуализации данных: Аналитик данных Ханна Ян Хан визуализирует уровень навыков, необходимый для 60 различных видов спорта, чтобы определить, какой из них самый сложный.

Там много данных, и вы можете многое с ними сделать. Попытка понять, с чего начать, может быть ошеломляющей. Если вам нужно небольшое направление для вашего следующего проекта, рассмотрите один из этих управляемых проектов по анализу данных:

1. Исследовательский анализ данных с помощью Python и Pandas. Применяйте методы EDA к любой таблице данных с помощью Python.

2. Учебное пособие по анализу настроений в Твиттере: очистите тысячи твитов и используйте их, чтобы предсказать, доволен ли клиент или нет.

3. Визуализация данных COVID19 с помощью Python: Визуализируйте глобальное распространение COVID-19 с помощью Python, Plotly и реального набора данных.