Инжиниринг управления данными

Мытрова Марина

Область проекта

Домен: авиация

Бизнес-проблема: анализ количества авиационных происшествий

Сбор данных

- Источник данных: сайт https://www.airdisaster.ru/
- Процесс сбора данных https://github.com/serpuhovichok/data_hw/blob/hw/flow.py#L8

Предобработка данных

https://github.com/serpuhovichok/data_hw/blob/hw/flow.py#L27

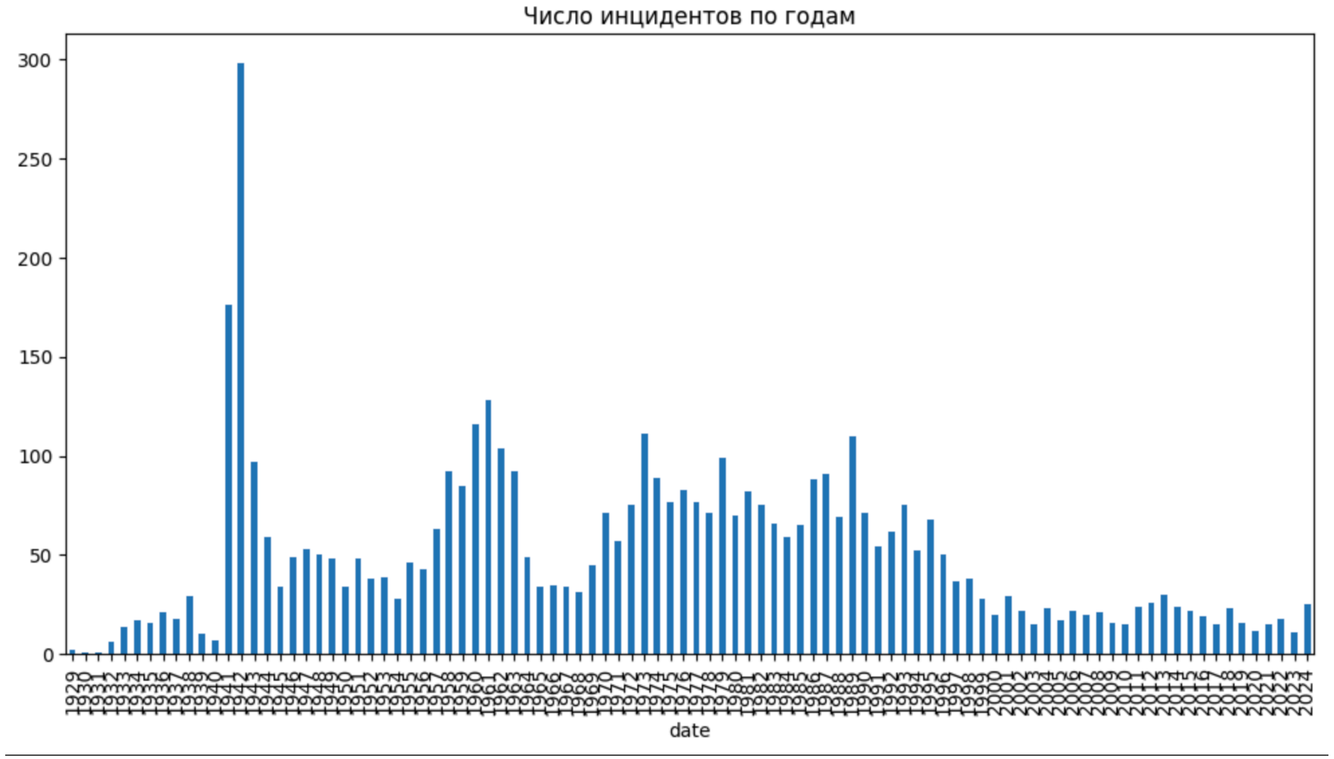
Процент пропущенных значений: 2.11

Полнота: 97.88

Число дубликатов: 230

EDA

https://github.com/serpuhovichok/data_hw/blob/hw/eda.ipynb



Метрики качества данных

Метрики полноты:
Процент пропущенных значений - 2.11%
Полнота - 97.88%

Метрики точности:
Число уникальных записей - 4585
Всего записей - 4815
Число дубликатов - 230

База данных

База данных выбрана SQLite, так как данных немного

https://github.com/serpuhovichok/data_hw/blob/hw/db_model.py

Пайплайн

В работе используется prefect

https://github.com/serpuhovichok/data_hw/blob/hw/flow.py#L51

Дашборд в Metabase