

Program kursu Data Scientist



MODUŁ 0: Wstęp do Analizy Danych

Wprowadzenie do narzędzia Excel

- Instalacja i konfiguracja narzędzia Excel
- Weryfikacja poziomu posiadanych umiejętności
- Uzupelnienie braków i wyrównanie wiedzy do poziomu wymagań niezbędnych do rozpoczęcia zajęć
- Omówienie funkcjonalności programu Excel w procesie analizy danych
- Operacje arytmetyczne z wykorzystaniem narzędzia Excel

Wprowadzenie do Analizy Danych

- Wizualizacja i analiza danych – wstęp i inspiracja
- Proces analizy danych
- Przetwarzanie danych
- Eksploracja danych i narzędzia matematyczne
- Pogłębiona analiza danych i elementy statystyki
- Prezentacja danych i ich analizy
- Przykładowe zbiory danych i zagadnienia, których dotyczą

Proces Data Science: krok 1. Zdefiniowanie problemu

- Zdefiniowanie problemu

Proces Data Science: krok 2. Zebranie danych

- Zebranie danych
- Kluczowe wskaźniki efektywności

Proces Data Science: krok 3. Czyszczenie i przetwarzanie danych

- Czyszczenie danych
- Przetwarzanie danych

Statystyka w analizie danych

- Podstawy statystyki opisowej

Proces Data Science: krok 4. Eksploracja danych

- Eksploracja danych
- Analiza zmiennych w zbiorze danych
- Filtrowanie i sortowanie danych
- Tabele przestawne

Proces Data Science: krok 5. Pogłębiona analiza danych

- Wprowadzenie do pogłębionej analizy danych – machine learning
- Regresja liniowa
- Korelacja liniowa
- Inne krzywe dopasowania

Praca domowa

- Analiza problemu

Proces Data Science: krok 6. Komunikowanie wyników

- Data storytelling
- Wprowadzenie do wizualizacji danych
- Przykłady dobrych i złych wizualizacji
- Rodzaje wykresów
- Tworzenie wizualizacji
- Przeniesienie wizualizacji do Powerpoint

Wstęp do testów A/B

- Wprowadzenie do testów A/B
- Przykłady zastosowania testów A/B w biznesie
- Pierwszy test A/B

Egzamin cząstkowy

Projekt końcowy

- Wprowadzenie do projektu końcowego
- Zebranie danych
- Czyszczenie i przetwarzanie danych
- Eksploracyjna analiza danych
- Pogłębiona analiza danych
- Komunikowanie wyników
- Prezentacje projektów końcowych



MODUŁ 1: Python – Analiza Danych

Prework – Podstawy programowania

- Wprowadzenie do Pythona
- Typy danych w Pythonie
- Biblioteka standardowa w języku Python

Python

- Funkcje
- Listy
- Krotki
- Stringi
- Obiekty
- Pliki
- Wyjątki
- Biblioteki

SQL

- Podstawowe operacje
- PostgreSQL
- Relacje
- Funkcje

JSON i API

- JSON
- API
- Autentykacja

Pandas

- Filtrowanie
- Obróbka danych
- Grupowanie danych
- Merge
- Datetime
- Pivot
- OpenPyXL

Web scraping

- Wstęp do HTML-a
- Webscraping w Python
- Element BS

Wizualizacja danych

- Wizualizacja danych – tworzenie wykresów

Generowanie PDF

- Tworzenie PDF
- Tworzenie dokumentu przez story
- Tabele

Egzamin cząstkowy

Projekt końcowy





MODUŁ 2: SQL – Analiza Danych

Prework – Bazy danych

- Typy baz danych
- Podstawy UML
- Instalacja bazy danych i przygotowanie środowiska pracy
- Zapoznanie z bazą danych
- Przeglądanie danych (elementarny SQL)
- Podstawy logiki operatorów and, or
- Historia baz danych

Podstawy SQL

- Podstawy SQL
- Podstawy logiki oraz algebry Bool'a
- Dodatkowe klauzule
- Operacje na zbiorach
- Podzapytania
- Tabele
- Manipulacja rekordami w bazie danych

Relacje

- Relacje i ich typy
- Joiny i ich rodzaje
- Joiny w zastosowaniu
- Delete cascade
- Indexy
- Inne typy join, dobre praktyki join oraz podzapytań

Analiza danych

- Analiza danych
- Widok
- Grupowanie danych
- Dodatkowe funkcje grupujące
- Funkcje okna
- Praca z datetime
- Kolejność operacji w SQL
- Rollup, query plan

Proceduralny SQL

- Proceduralny SQL
- Wprowadzenie do proceduralnego SQL-a
- Kontrola przepływu
- Transakcje
- Wbudowane funkcje w SQL-u
- Pętle
- Kursory
- JSON
- Normalizacja danych
- Notebook dla SQL-a

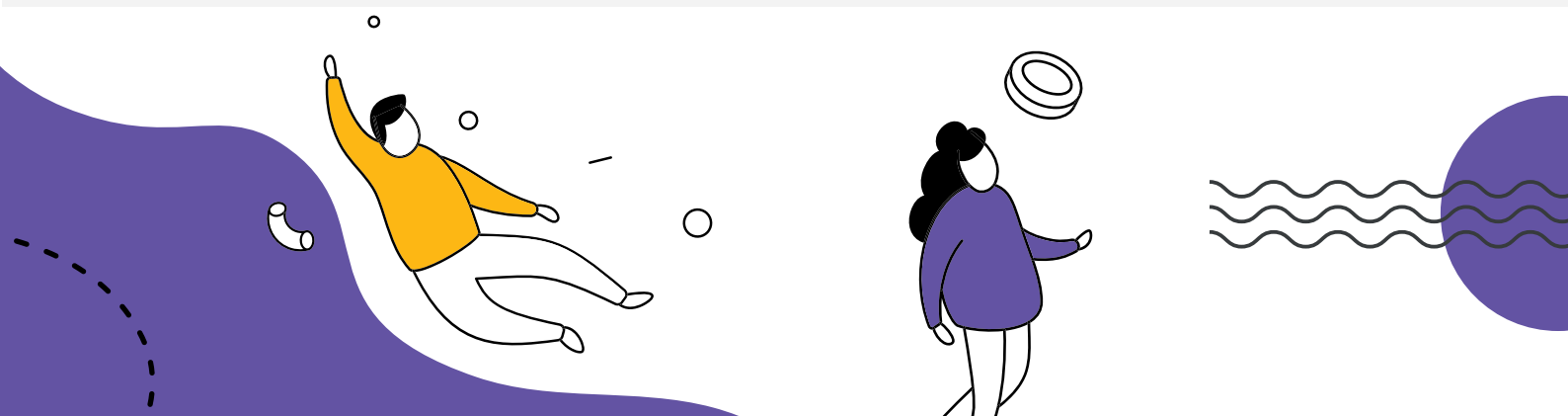
PREZENT — Rozdział AI: Użycie ChatGPT w pracy analityka (materiały dodatkowe, nieobowiązkowe)

- Wstęp do AI dla Analityków Danych
- ChatGPT i postgresSQL
- ChatGPT i Jupyter Notebook
- Wykorzystanie ChatGPT do optymalizacji procesów analitycznych



Egzamin cząstkowy

Projekt końcowy





MODUŁ 3: Wizualizacja Danych

Prework – Wstęp do wizualizacji danych

- Wstęp i przygotowanie
- Wstęp do wizualizacji danych
- Kodowania graficzne
- Kolor
- Datawrapper
- Typy danych i operacje

Seaborn/Plotly

- Teoria wizualizacji danych
- Mapy kolorów
- Seaborn
- Plotly
- Podstawy wizualizacji w poszczególnych bibliotekach
- Wykresy Plotly w Google Slides
- Podstawy Bokeh

Mapy

- Projekcje, współrzędne, kartogramy
- Wstęp do bibliotek
- Geopandas, Geometry, GeoSeries, GeoDataFrame osm
- Geojson, Shapefile
- Mapbox

Dash, grafy

- Networkx i Grafy
- Co to jest dashboard?
- Wstęp do dash
- Dash komponenty
- Dash datatable
- Dash callback i app lifecycle
- Edytowalna data table
- Interakcje z dashboardem
- Dash Cytoscape

Dashboardy, Data Storytelling

- Data storytelling
- Dashboard
- Tooltip
- Multipage app
- Style bootstrapowe
- Dash i SQL
- Dash i mapy

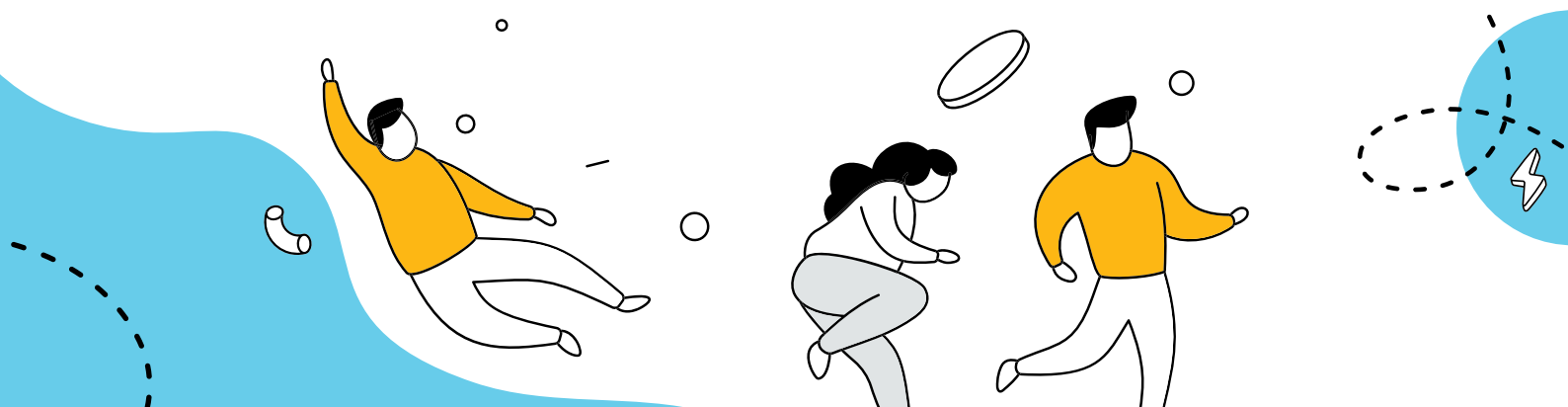
Tableau

- Co to jest Tableau?
- Przygotowanie środowiska pracy
- Pierwszy projekt
- Tworzenie Dashboardu
- Scenariusze wykorzystania
- Integracja Tableau i Mapbox

Egzamin cząstkowy

Projekt końcowy

Egzamin kończący kurs Analityk Danych





MODUŁ 4: Machine Learning

Wstęp do Machine Learning

- Wstęp do kursu
- Wymagania i konfiguracja środowiska
- Powtórka z Pythona (numpy, pandas, biblioteki do wizualizacji)
- Statystyka w uczeniu maszynowym
- Wstęp do uczenia maszynowego
- Jak odpowiednio przygotować dane do modelu?

Regresja

- Regresja
- Regresja liniowa
- Regularyzacja w modelu regresji liniowej
- Regresja wielomianowa
- Problem regresji z wykorzystaniem drzewa decyzyjnego
- Jak określić jakość modelu regresji? Metryki modeli regresyjnych

Klasyfikacja

- K najbliższych sąsiadów
- Regresja logistyczna
- Problem klasyfikacji z wykorzystaniem drzewa decyzyjnego
- SVM (maszyna wektorów nośnych)
- Jak określić jakość modelu klasyfikacji? Metryki modeli klasyfikacyjnych

Praca domowa

- Jak wrzucić swój notebook z Colaba na Kaggle?
- Regresja liniowa z regularyzacją
- Drzewo decyzyjne (regresja)
- Regresja logistyczna
- SVM

Podstawy sieci neuronowych i NLP

- Podstawy sieci neuronowych
- Podstawy NLP

Zaawansowane modele uczenia maszynowego

- Lasy losowe
- Boosting i bagging
- Sieci neuronowe
- Podsumowanie dnia 3

NLP – przetwarzanie języka naturalnego

- Jak pracować z danymi tekstowymi?
- Bag of words
- TF-IDF
- Word2Vec
- BERT

Praca domowa

- Jak poprawić działanie modelu, czyli optymalizacja hiperparametrów
- Lasy losowe
- XGBoost
- Sieci neuronowe

Uczenie nienadzorowane

- Uczenie nienadzorowane
- Redukcja wymiarowości
- Klasteryzacja
- Detekcja anomalii
- Podsumowanie dnia 5

Projekt końcowy, egzamin

- Projekt końcowy
- Prezentacja wyników końcowych
- Egzamin
- Podsumowanie kursu

Egzamin

Projekt końcowy