## Lab Assignment 4: Final project with free choice of topic

1. Please, choose one of the topics for the personal assignment from the list bellow, or choose the topic of your personal interest.
2. Use data from Kaggle or use web scraping for data collection.
3. Use `main.ipynb` file to write code for your project.
4. Visualize results of your personal project using matplotlib or seaborn.
5. Describe your personal project in detail in the markdown sections of `main.ipynb` file, what you did as well as the results of your findings.
6. List of topics:

   - **Data Collection Plan for a Real-World Problem**
     Identify a real-world problem (e.g., predicting air quality, analyzing traffic patterns) and design a data collection plan. Specify the data sources, collection methods (e.g., API, web scraping, manual entry), frequency of updates, and format.
   - **APIs for Data Collection**
     Select a public API (e.g., Twitter API, OpenWeather API, or a finance-related API) and write a script to gather data on a specific topic over a period. Demonstrate how the data collected can be used to gain insights.
   - **Web Scraping for Market Analysis**
     Using a Python library like BeautifulSoup or Scrapy, build a web scraper to collect data on a specific product category (e.g., laptops, mobile phones) from an e-commerce website. Collect details such as price, ratings, and product specifications.
   - **Data Collection Ethics Case Study**
     Analyze a case where data collection practices led to ethical or privacy concerns (e.g., Cambridge Analytica). Discuss what went wrong, how it could have been avoided, and propose ethical data collection guidelines for similar scenarios.
   - **Data Collection for Social Media Analysis**
     Select a social media platform (e.g., Twitter, Instagram) and collect data on a specific hashtag or topic using available APIs or a web scraping tool. Analyze data trends, engagement metrics, and sentiment.
   - **Assessing Data Quality from Different Sources**
     Choose two or more sources of data on a common topic (e.g., weather, economic indicators) and compare the data quality based on completeness, accuracy, timeliness, and consistency.
   - **Simulating Data for Machine Learning Models**
     Create a synthetic dataset that simulates real-world data (e.g., customer purchases, patient medical records) with different data types and structures. Use this data to test a machine learning model and discuss limitations of simulated data.
   - **Evaluating the Relevance of Collected Data**
     Choose a dataset or collect new data, and evaluate its relevance for a specific analytical objective (e.g., predicting customer churn). Discuss the factors affecting relevance, such as data recency, granularity, and representativeness.
   - **Advanced Prompt Engineering**
     Create a chain of prompts to complete a multi-step task (e.g., summarizing an article, then generating a list of questions based on the summary). Evaluate how each step's output impacts the quality of the final result.

- **Text Preprocessing and Cleaning**
  Select a dataset containing raw text (e.g., social media posts or product reviews) and perform a series of preprocessing steps including tokenization, lowercasing, stopword removal, stemming, and lemmatization. Evaluate how each step impacts the text for downstream tasks.
- **Text Classification with Word Embeddings**
  Use word embeddings (e.g., Word2Vec, GloVe, or BERT embeddings) to classify text samples into predefined categories (e.g., spam/ham, sentiment polarity). Evaluate the classification model's performance and discuss the impact of using embeddings.
- **Analyzing Bias in NLP Models**
  Select a pre-trained NLP model (e.g., a language generation model) and analyze it for potential biases in the outputs it generates on topics related to gender, race, or socioeconomics. Discuss ways to mitigate any observed biases.
- **Visualizing Geospatial Data with Heatmaps**
  Using a dataset with geographic information (e.g., earthquake occurrences or store locations), create a heatmap to visualize the density of occurrences. Experiment with different color schemes to represent intensity.
- **Network Graph Visualization for Social Connections**
  Using a social network dataset (e.g., friendships, collaborations), create a network graph to visualize relationships between entities. Experiment with node and edge properties to highlight central or highly connected nodes.
- **Regression Algorithms Comparison**
  Choose a publicly available dataset (e.g., energy consumption based on time of day and temperature, etc.) and perform multiple types of regression: polynomial, support vector regression, decision tree etc. Compare different algorithms.
- **Exploratory Data Analysis (EDA) and Visualization**
  Choose a publicly available dataset (e.g., from Kaggle) and perform a detailed Exploratory Data Analysis (EDA). Identify important patterns, trends, and relationships within the data.
- **Classification Model for Predicting Outcomes**
  Using a labeled dataset (e.g., from Kaggle), build and evaluate different classification models (e.g., Logistic Regression, SVM, etc.). Compare their accuracy and precision in order to identify the best model.
- **Building a Recommender System**
  Create a simple recommender system using collaborative filtering (e.g., user-based, item-based) or content-based filtering on a dataset (e.g., movie or product ratings). Evaluate the model's performance and discuss how to improve recommendations.
- **Natural Language Processing (NLP) for Sentiment Analysis**
  Use NLP techniques to perform sentiment analysis on a dataset of text reviews (e.g., product or movie reviews). Build and evaluate a sentiment classification model using algorithms such as Naive Bayes, Logistic Regression, etc.
- **T-Test for Difference in Means**
  Choose a dataset with at least two groups (e.g., male/female, control/experimental) and apply an independent t-test to check if there is a significant difference in a chosen continuous variable between the groups.
- **Chi-Square Test for Categorical Data**
  Use a dataset with categorical variables (e.g., customer demographics or purchase behaviors) and conduct a chi-square test to examine the association between two categorical variables.
- **ANOVA for Comparing Multiple Groups**
  Choose a dataset with a continuous variable and a categorical variable with more than two

groups (e.g., salary based on job level). Perform ANOVA to test if the means across the groups are significantly different.

- **Naive Bayes Classification**
  Use a labeled dataset with categorical labels (e.g. messages labeled as spam or not spam, tumors labeled as cancerous or not cancerous etc.) and build a Naive Bayes classifier to predict future data. Split your data into train and test data, evaluate the model's performance and discuss ways to improve it.
- **Decision Trees Classification**
  Use a labeled dataset with categorical labels (e.g. messages labeled as spam or not spam, tumors labeled as cancerous or not cancerous etc.) and build a decision trees classifier to predict future data. Split your data into train and test data, evaluate the model's performance and discuss ways to improve it.
- **Random Forest Classification**
  Use a labeled dataset with categorical labels (e.g. messages labeled as spam or not spam, tumors labeled as cancerous or not cancerous etc.) and build a random forest classifier to predict future data. Split your data into train and test data, evaluate the model's performance and discuss ways to improve it.
- **Clustering with K-Means on Customer Segmentation**
  Using a customer dataset (e.g., customer demographics, purchase history), apply K-Means clustering to identify distinct customer segments. Analyze the clusters and describe each segment's characteristics.
- **Dimensionality Reduction with PCA**
  Select a high-dimensional dataset (e.g., image data, gene expression data) and apply Principal Component Analysis (PCA) to reduce dimensionality. Visualize the data in the reduced space and discuss how much variance is explained by the principal components.
- **Cluster Visualization for High-Dimensional Data**
  Use a dimensionality reduction technique (e.g., PCA) on a high-dimensional dataset and visualize the clusters in 2D or 3D space. Experiment with color-coding clusters to enhance interpretability.
- **Clustering for Recommender Systems**
  Use clustering (e.g., K-Means) on a dataset of user preferences or product ratings to create user groups. Propose recommendations based on each group's preferences.
- **Ethics and Bias in Machine Learning**
  Choose a dataset and a machine learning model. Analyze the potential sources of bias in the dataset and discuss how the model's predictions could be influenced by these biases. Suggest techniques to mitigate these biases (e.g., data re-sampling, algorithm choice).