Published using Google Docs Report abuse Learn more

Data Analytics with AI Individual Formative Assignment One: Python ETL and Visualisation

Duration: 2 Days

Participant activities: Data Extraction, Transformation, and Loading (ETL), Data Visualisation, and Agile Project Management.

Introduction

Welcome to the two-day individual formative assignment focused on Python ETL and Data Visualisation. You will create an individual project but will collaborate with your classmates as though in the workplace. The individual formative assignment aims to enhance participants' skills in data extraction, transformation, and loading, as well as data visualisation using Python tools.

Each learner will actively contribute throughout both days, following the principle of shared responsibility to ensure a balanced workload and a successful project outcome. The

Published using Google Docs Report abuse Learn more

Individual formative assignment Outcomes

- Design and implement an ETL pipeline using Python.
- Visualise data using Matplotlib, Seaborn, and Plotly.
- Adhere to the key milestones and deliverables.
- Maximise future maintainability through documentation, code structure, and organisation.
- Document and present the project process and outcomes.
- Demonstrate and document the development process through a version control system such as GitHub.

Sample Project Ideas:

Project Example Idea 1: Retail Sales Data Analysis

Data source:
https://www.kaggle.com/datasets/manjeetsingh/retaildataset

Data analysis goals:
Analyse retail sales data to identify trends, insights, and the impact of promotional markdowns on sales. Provide comprehensive, visually appealing sales reports and insights to assist in strategic decision-making.

Context
The challenge involves making decisions based on limited historical data, particularly around holidays and promotional events. The dataset includes historical sales data for 45 stores in different regions, with details about store types,

Published using Google Docs Report abuse Learn more

from the Excel sheets (Stores, Features, Sales).
    - Transform: Clean the data, handle missing values, and create new features such as sales differences between holiday and non-holiday weeks.
    - Load: Store the transformed data in a format suitable for analysis (e.g., a cleaned DataFrame).
  - Data Visualisation:
    - Descriptive Statistics: Display basic statistics such as average sales per store and department.
    - Trend Analysis: Plot sales trends over time for different stores and departments.
    - Impact Analysis: Visualise the impact of markdowns on sales during holidays versus non-holiday periods.
    - Comparative Analysis: Compare sales performance across different stores and regions.

Project Example Idea 2: Healthcare Insurance Cost Analysis

Data source: https://www.kaggle.com/datasets/willianoliveiragibin/healthca insurance

Data analysis goals: Analyse healthcare insurance data to understand how personal attributes and geographic factors influence insurance costs and develop predictive models for estimating healthcare

Published using Google Docs Report abuse Learn more

Hackathon 1: Python ETL and Visualization

Updated automatically every 5 minutes

Context
The dataset contains information on the relationship between personal attributes (age, gender, BMI, family size, smoking habits) and geographic factors and their impact on medical insurance charges. This data can be used to study how these features influence insurance costs and develop predictive reports for estimating healthcare expenses.

Potential features to include:

- ETL Pipeline:
  - Extract: Load data from the provided CSV file.
  - Transform: Clean the data, handle missing values, encode categorical variables, and create new features such as body mass index (BMI) categories.
  - Load: Store the transformed data in a format suitable for analysis (e.g., a cleaned DataFrame).
- Data Visualisation:
  - Descriptive Statistics: Display basic statistics such as average insurance charges by age, gender, and region.
  - Correlation Analysis: Visualise correlations between different attributes and insurance charges.
  - Predictive Analysis: Develop and visualise predictive reports for estimating insurance costs.
  - Geographic Analysis: Visualise the impact of geographic regions

Published using Google Docs Report abuse Learn more

Hackathon 1: Python ETL and Visualization

Updated automatically every 5 minutes

Card Customer Churn Analysis

Data source: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Data analysis goals: Assess which customers are likely to churn (when a customer ceases to be a customer) so that the bank can take proactive measures to retain them. Develop analytical reports to identify customers at risk of churning and provide actionable insights to reduce churn rates.

Context
A bank manager is concerned about increasing customer churn in their credit card services. The goal is to assess which customers are likely to churn using customer data and develop strategies to improve retention.

Potential features to include:

- ETL Pipeline:
  - Extract: Load data from the provided dataset.
  - Transform: Clean the data, handle missing values, encode categorical variables, and balance the dataset for churn analysis.
  - Load: Store the transformed data in a format suitable for analysis (e.g., a cleaned DataFrame).
- Data Visualisation:
  - Descriptive Statistics: Display basic statistics such as churn rates by age, salary, and marital status.
  - Correlation Analysis: Visualise correlations between different

Published using Google Docs Report abuse Learn more

probability.

Project Example Idea 4: Online Retail Transaction Analysis

Data source: https://www.kaggle.com/datasets/abhishekrp1517/online-retail-transactions-dataset

Data analysis goals: Analyse online retail transaction data to understand customer behaviour, identify popular products, and optimise pricing and marketing strategies. Provide insights into customer behaviour, popular products, and pricing strategies to improve sales and marketing efforts.

Context
The "Online Retail Transaction" dataset contains information on customer transactions made through an online retail platform. It includes data on products purchased, quantities, transaction dates and times, prices, customer identifiers, and customer locations. This dataset can be used to analyse customer behaviour and preferences, identify popular products, and optimise pricing and marketing strategies.

Potential features to include:

- ETL Pipeline:
    - Extract: Load data from the provided dataset.
    - Transform: Clean the data, handle missing values, encode categorical variables, and create new features such as total transaction value.
    - Load: Store the transformed data in a format suitable for

Published using Google Docs Report abuse Learn more

as total sales, average transaction value, and number of unique customers.
    - Trend Analysis: Visualise sales trends over time.
    - Customer Segmentation: Segment customers based on purchase behaviour and visualise their characteristics.
    - Product Analysis: Identify and visualise the most popular products.

Project Example Idea 5: Car Price Analysis

Data source: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

Data analysis goals:
Analyse the price of cars based on various independent variables to understand the factors affecting car pricing in the US market. Provide insights into car pricing dynamics to help the company design cars and develop business strategies that meet market demands.

Context
An automobile company aims to enter the US market and compete with local and European manufacturers. It needs to understand the factors influencing car prices in the US market to adjust its strategies and design accordingly. The dataset includes various attributes of cars, such as make, model, engine size, and price.

Potential features to include:

- ETL Pipeline:

Published using Google Docs Report abuse Learn more

variables, and create new features such as price-to-performance ratios.
    - Load: Store the transformed data in a format suitable for analysis (e.g., a cleaned DataFrame).
- Data Visualisation:
    - Descriptive Statistics: Display basic statistics such as average car prices by make and model.
    - Correlation Analysis: Visualise correlations between different attributes and car prices.
    - Predictive Analysis: Develop and visualise analytical reports for estimating car prices.

Project Example Idea 6: Custom Data Analysis Project

Data source: Participants are free to choose any dataset that interests them

Data analysis goals: Conduct an analysis on a dataset of their choice to derive meaningful insights and develop visualisations. Provide guidelines for conducting thorough data analysis, ensuring that all key steps are covered and the final presentation is comprehensive.

Guidelines for Custom Data Analysis Project:

- Project Ideation and Planning:
    - Choose a dataset that aligns with your interests or professional goals.
    - Define clear objectives for your

Published using Google Docs Report abuse Learn more

    - Extract: Load the dataset using Pandas or other relevant tools.
    - Transform:
        - Clean the data: Handle missing values, duplicates, and outliers.
        - Encode categorical variables if necessary.
        - Create new features if relevant to the analysis (e.g., calculating ratios, deriving new metrics).
    - Load: Ensure the transformed data is stored in a format suitable for analysis.
- Data Analysis:
    - Conduct descriptive statistics to understand the basic characteristics of the data (e.g., mean, median, standard deviation).
    - Visualise the distribution of key variables using histograms, box plots, or density plots.
    - Identify and visualise correlations between variables using correlation matrices and heatmaps.
- Data Visualisation:
    - Create visualisations to answer your research questions and provide insights:
        - Time series plots for trend analysis if the data is time-based.
        - Scatter plots to examine

Published using Google Docs Report abuse Learn more

data analysis.
        - Interactive dashboards using Plotly for more complex visualisations.
- Reports:
    - Develop and evaluate analytical reports.
    - Choose appropriate reports based on the data and objectives.
- Documentation and Presentation:
    - Document the entire analysis process, including:
        - Data sources and any preprocessing steps taken.
        - Key findings from data analysis and visualisations.
        - Report developing.
    - Prepare a presentation summarising:
        - Objectives and motivations for the project.
        - Key insights and findings from the analysis.
        - Visualisations and their interpretations.
        - Any recommendations or conclusions based on the analysis.

High-Level Activities/Ceremonies

1. **Data Extraction, Transformation, and Loading (ETL)**

Published using Google Docs Report abuse Learn more

suitable format for analysis, and ensuring data quality and consistency.

2. **Data Visualisation Specialist (data analyst)**: Create visual representations of the data, develop static and interactive visualisations using Matplotlib, Seaborn, and Plotly, ensure visualisations effectively communicate insights and collaborate with the ETL Specialist to understand the data structure.
3. **Project Management milestones and deliverables (project manager)**: agile planning process, tracking progress, facilitating team collaboration and communication, ensuring the team meets deadlines, and documenting the project process outcomes.

Day 1: Ideation, ETL Implementation, Basic Visualisations:

Activities & Expected Outcomes:

1. Ideation Session: Brainstorm and discuss potential project ideas that address real or imagined user needs. You can use your classmates as proxy users to bounce ideas off.

Published using Google Docs Report abuse Learn more

matches your intended future data analyst role (or one that looks most enjoyable)
3. Initial data extraction and cleaning.
4. Milestone and Deliverable Planning: Set up a GitHub repository and project Kanban board, define user stories, and map them to the project goals.
5. Creation of basic visualisations
    a. Expected Outcomes:
        i. Clear project plan and defined roles.
        ii. Cleaned and transformed dataset ready for analysis.
        iii. Initial visualisations provide basic insights.

Deliverables & Milestones:

Activities Per Role Day 1

Project Management

Day 1:

- Come up with project ideas during the ideation session.
- Refine and select the final project idea, ensuring it aligns with the individual formative assignment's guidelines.
- Create a GitHub repository.
- Set up the project board using GitHub Projects or Trello, creating and assigning user stories and supporting tasks.

Published using Google Docs Report abuse Learn more

Day 1:

- Extract data from the dataset.
- Clean and transform data.
- Load data into a suitable format for analysis.
- Ensure data quality and consistency.

Data Visualisation

Day 1:

- Understand the data structure as an ETL Specialist.
- Create initial visualisations using Matplotlib and Seaborn.
- Develop basic charts and graphs to visualise key metrics.

Day 2: Advanced Visualisations, Testing, Presentation Preparation

Activities & Expected Outcomes:

1. ETL: Testing of ETL pipelines and visualisations for accuracy.
2. Regular Check-ins: Hold regular check-ins with data coaches or facilitators to monitor progress on the Kanban board and resolve any issues.
3. Code Review: Conduct code reviews to ensure clean code, proper indentation, and consistent naming conventions.
4. Visualisations: Development of advanced and interactive visualisations.
5. Preparation for final presentation and

Published using Google Docs Report abuse Learn more

audience. Expected Outcomes:

i. A comprehensive set of visualisations.
ii. Validated ETL and visualisation processes.
iii. A well-prepared presentation showcasing project outcomes.
iv. Jupyter Notebook (.ipynb) file with project code.

Activities Per Role Day 2

Project Management

Day 2:

- Conduct regular check-ins with data coaches or facilitators to monitor progress and address any challenges.
- Ensure that your activities are on track and resolve any roadblocks.
- Keep track of completed tasks and update the project board accordingly.
- Coordinate the preparation of the final presentation.
- Document any tests carried out in the README file

ETL

Day 2:

- Refine the ETL process based on initial feedback.
- Ensure the dataset is updated and accurate.
- Act as a Visualisation Specialist to provide necessary data.

Published using Google Docs Report abuse Learn more

- Create interactive dashboards and charts.
- Test visualisations for accuracy and usability.

Resilience Behaviours

Solution-Oriented Mindset:

Capture:

Remember to celebrate and share your achievements, progress, and challenges, focusing on maintaining a positive attitude.

Measure:

At the end of each day, reflect on your experiences and provide a brief summary of how you maintained a positive mindset throughout the individual formative assignment.

Task Estimation:

Capture:

Estimate the effort required for each task during the ideation phase. Remember you only have two days for the project.

Measure:

Track the actual time taken to complete each task and compare it with the initial estimates. This will assess the accuracy of task estimation.

Published using Google Docs Report abuse Learn more

Personal accountability means meeting expectations, deliverables, and milestones. Hold yourself accountable, and you prioritise the larger project goals. You will be measured on your personal contribution to achieving the project goal in addition to the project artefact itself.

Measure:

At the end of each day, review the planned tasks and compare them with the completed tasks. Personal accountability will be measured based on the consistency and accuracy of reporting progress and task completion.

Dealing with Distractions:

Capture:

During daily stand-up meetings, each member will observe and discuss any distractions they encounter.

Measure:

Learners must briefly describe the distractions they faced during the day and explain how they managed to refocus on their tasks. This will provide insights into their ability to handle distractions.

Problem Solving:

Capture:

Learners are required to document in the README any challenges or obstacles they encounter and the problem-solving approaches they use during the implementation phase.

Published using Google Docs Report abuse Learn more

assessment will be part of daily reflections.

Consistent Coding Practices:

Capture:

Code reviews of workbooks can be done using Copilot or linters to ensure coding practices are followed.

Measure:

The success of consistent coding practices is measured based on the feedback received during code reviews.

Project Management Methodology Adoption:

Capture:

The adoption of Agile methodologies is documented through the usage of a project management tool for managing tasks.

Measure:

The effectiveness of project management is assessed based on the completion of tasks and the project's final success as a minimal viable product.

Judging Criteria:

Overall Points:

- Teams choosing one of the proposed datasets (Project Examples 1-5) can score a

Published using Google Docs Report abuse Learn more

of 100 points. These additional points are awarded because this option requires additional effort in ETL, as the data may be dirty or not in an easily visualisable format.

Assessment Criteria	Minimum Score	Maximum Score	Max Grade
Scoring	0.9 (1 for custom project)	4.5 (5 for custom project)	90 (100 for custom project)
ETL extraction	No data	All CSV, Excel or API data was extracted correctly.	
ETL data load	No data loaded	Data loaded into a suitable format for Python.	
ETL transformation	No data transformation	All data is logically cleaned, preprocessed and transformed into the correct format for analysis	
ETL data quality and integrity	No improvement in data quality	Data is of high quality/integrity and has been validated	
ETL data documentation	ETL process not documented	Documentation for the ETL process, including data sources and transformation logic.	
Static visualisation	It is not clear what is visualised	Clear, easy-to-view basic visualisation such as histograms, bar charts, and line plots using Matplotlib.	

Published using Google Docs Report abuse Learn more

Hackathon 1: Python ETL and Visualization

Updated automatically every 5 minutes

		using Seaborn.	
Visualisation interactivity	No purpose for visualisation	Interactive Plotly visualisations	
Visualisation customisation	Only standard chart type	Customise plots with titles, labels, legends, and annotations for better clarity.	
Visualisation relevance	Visualisation not relevant to the report's purpose	Each visualisation is important to the report	
Project planning	No plan	Project task process documented using project management tool (Kanban board)	
Project documented	No documented tasks	Project tasks and user stories are documented within the project management tool (Kanban board)	
Project adaptation	No flexibility	Evidence of adjusting the project plan based on feedback and changing requirements.	
Project review	No review	Evidence of conducting reviews to assess progress and make necessary adjustments.	
Project reflection	No reflection on progress	Evidence of performing retrospectives to identify lessons learned and areas for improvement.	

Published using Google Docs Report abuse Learn more

		and how to use them. It also explains the project, its objectives, and how to reproduce it.	
Presentation of process	No mention of how the learner got from data to the visualisations	Clear presentation of the deliverables and milestones on the visualisation creation process. Evidence of version control.	
Project naming	Poorly named files and functions.	Name files consistently and descriptively, without spaces or capitalisation, to allow for cross-platform compatibility. Organise Jupyter Notebook code into well-defined and commented sections	
Code quality	Poorly written code	Write code that meets at least minimum standards for readability (consistent indentation, blank lines only appear individually or, at most, in pairs). Ensure all code from external sources is properly attributed and referenced.	
Presentation of findings	No conclusions on the data.	Present the story of what key findings were made with the data.	

ETL Pipeline (25%): Efficiency of the ETL process, handling of missing values, data

Published using Google Docs Report abuse Learn more

visualisations, use of Matplotlib, Seaborn, and Plotly for creating informative visualisations, and relevance of visualisations to the research questions.
Code Quality (25%): Clean code structure, proper documentation of code, use of comments, and use of version control systems like GitHub.
Documentation and Presentation (25%): Quality of README.md file, clear and concise presentation of the project process and outcomes, and effectiveness of the presentation in communicating insights and findings.

Winners Selection:

- Among the prepared projects, which are a minimal viable product, judges will select 3 winners (based on a result-oriented approach in the projects).
- The winning teams will receive the maximum grades: 100 points for the Custom Data Analysis Project and 90 points for the proposed datasets (examples 1 - 5).

Published using Google Docs Report abuse Learn more

Hackathon 1: Python ETL and Visualization

Updated automatically every 5 minutes

Day 1: Ideation, ETL, Basic Visualisations

9:00 am - 10:30 am
- All members:
  - Kick-off meeting and ideation session start. Contribute project ideas that align with the individual formative assignment's guidelines.

10:30 am - 11:30 am
- Project Management:
  - Setting up the project GitHub repository. Organising the project board using GitHub Projects or Trello.

11:30 am - 12:30 pm
- ETL:
  - ETL pipeline setup.

12:30 pm - 1:30 pm: Lunch Break

1:30 pm - 3:00 pm
- ETL:
  - Continue ETL implementation and communicate about basic visualisations.

3:00 pm - 4:30 pm
- Data Visualisations:
  - Start and refine basic visualisations

4:30 pm - 5:00 pm
- Milestone check-in for a review of Day 1 work and Day 2 prep.

Day 2: Advanced Visualisations, Testing, Presentation

9:00 am - 9:15 am
- Morning stand-up meeting for progress discussion.

9:15 am - 12:00 am
- Data Visualisations:
  - Advanced visualisation development.

12:00 am - 12:30 pm
- ETL:
  - Testing and troubleshooting.

12:30 pm - 1:30 pm: Lunch Break

1:30 pm - 3:30 pm
- Data Visualisations:
  - Finalise visualisations and

Published using Google Docs Report abuse Learn more

Feedback gathering and update checks on the project board.
Progress check-ins

4:00 pm - 4:15 pm
Dry run for presentations.

4:15 pm - 5:00 pm
Final presentations and evaluations.

Individual formative assignment Outcomes: Detail

O1 Design and implement an ETL pipeline using Python. (Jupyter Notebook)

1.1 Extract: Load data from various sources (CSV, Excel, APIs).

1.2 Transform: Clean and preprocess data, handle missing values, and perform necessary transformations.

1.3 Load: Store the transformed data in a suitable format for analysis (e.g., DataFrame, database).

1.4 Automate: Develop scripts to automate the ETL process.

1.5 Validate: Ensure data quality and integrity through validation checks.

1.6 Document: Provide documentation for the ETL process, including data sources and transformation logic.

O2 Visualise data using Matplotlib, Seaborn, and Plotly.

2.1 Static Visualisations: Create basic plots such as histograms, bar charts, and line plots using Matplotlib.

2.2 Advanced Visualisations: Develop more complex visualisations like pair plots, heatmaps, and violin plots using Seaborn.

2.3 Interactive Dashboards: Build interactive visualisations and dashboards using Plotly.

2.4 Customise: Customise plots with titles, labels, legends, and annotations for better clarity.

Published using Google Docs Report abuse Learn more

O3 Employ a project management methodology.

- 3.1 Plan: Use tools like GitHub Projects or Trello to manage tasks and track progress.
- 3.2 Collaborate: Conduct regular stand-up meetings to discuss progress, challenges, and next steps.
- 3.3 Document: Document user stories, sprints, and tasks within the project management tool.
- 3.4 Adapt: Adjust the project plan based on feedback and changing requirements.
- 3.5 Review: Conduct sprint reviews to assess progress and make necessary adjustments.
- 3.6 Reflect: Perform retrospectives to identify lessons learned and areas for improvement.

O4 Document and present the project process and outcomes.

- 4.1 README.md: Write a detailed README.md file explaining the project, its objectives, and how to reproduce the results.
- 4.2 Process Documentation: Document each step of the ETL and visualisation process, including challenges faced and solutions implemented.
- 4.3 Screenshots: Include screenshots of key visualisations and the ETL pipeline.
- 4.4 Code Attribution: Ensure all code from external sources is properly attributed and referenced.
- 4.5 Organise Jupyter Notebook code into well-defined and commented sections
- 4.6 Presentation: Prepare a comprehensive presentation summarising the project, key findings, and insights.
- 4.7 Write code that meets at least minimum standards for readability (consistent indentation, blank lines only appear individually or, at most, in pairs)
- 4.8 Name files consistently and descriptively, without spaces or capitalisation, to allow for cross-platform compatibility.

Published using Google Docs Report abuse Learn more

Hackathon 1: Python ETL and Visualization

Updated automatically every 5 minutes

5.1 Use consistent and effective markdown formatting to produce a README file in English that is well-structured, easy to follow, and has few grammatical errors.