# Page: W1P7 - Data Engineer

The Data Engineer—preprocessing and Feature Engineering role is critical in ensuring the integrity and quality of the dataset for subsequent analysis and modelling tasks.

This role contains several key responsibilities to effectively prepare the data for further analysis.

- Load and preprocess the dataset
- Handle missing data points and encode categorical variables for analysis.
- Perform feature scaling and normalisation to prepare the data for modelling.
- Select key features for clustering analysis.
- Collaborate with other team members to ensure data integrity and consistency.

See the basic activity breakdown for this role based on Module structure

Activity Breakdown for the Data Engineer role by project stages

| | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| Output | • Load the dataset into the analysis environment.<br>• Check for any missing or incomplete data points.<br>• Handle missing data by imputation or deletion as appropriate.<br>• Encode categorical variables for further analysis.<br>• Perform basic data cleaning and formatting to ensure data integrity. | • Identify relevant features for analysis based on project requirements.<br>• Engineer new features or transform existing ones to improve predictive performance.<br>• Handle feature scaling and normalisation to prepare data for modelling.<br>• Collaborate with team members to ensure consistency and integrity of features across modules. | • Validate c consister modules.<br>• Conduct checks tc of data tr preproce<br>• Documer issues er propose s resolutior<br>• Collabora members related cl smooth w |
| Skill set pathway | ACS WILDA Stage1 - Project Planning (https://acs-preview.percipio.com/track/ac39b3d8-e4e6-4184-b1b0-1d679e69d68b) | ACS WILDA Syage2 - Data Engineer (https://acs-preview.percipio.com/track/1d98039e-92a3-4266-b8ee-6a897377dbbe) | ACS WILDA Engineer preview.perci 72b4-4f3c-aa |
| Total internship hours | 60-70 Hours | 100 - 110 Hours | 100 - 110 H |