DataScientest

Relevant training course:

Data Science/ML Engineering Project Proposal

Level of difficulty:

Intermediate (?)

Description of the project:

Help Desk Ticket Data

(Note: Table gives an overview of potential project steps/topics; general description and aim of project on page 2.

TIME (using issues.csv)	
Overall resolution time	Which ticket **characteristics** (e.g. priority, type, project) are associated with longer or shorter total **processing times**?
Bottlenecks in the workflow	In which **workflow states** do tickets spend most of their time, and which states look like **bottlenecks**?
Resolution outcomes	Which ticket and process **characteristics** are associated with non-ideal **resolutions** such as "Won't Do", "Cannot Reproduce", or "Duplicate"?
Prediction of time	**Prediction:** whether a ticket will take more than a given time threshold to resolve based on its attributes?
ASSIGNEES (using issues_snapshot.csv)	
Slow/fast assignees	How much **time** does a ticket spend with each **assignee**, and do some assignees systematically handle tickets faster or slower?
Resolution/processing patterns	How do **time spent, steps, and comments** per turn differ across assignees, projects, or issue types?
Multi-assignee flows	How often do tickets pass through **multiple assignees**, and what are the most typical assignment **sequences**?
Roles and responsibilities	Which **assignees** are mainly involved at the **beginning**/**end** of the ticket lifecycle?
PERFORMANCE (using issues_snapshot_sample.xlsx)	
Process-performance link	How are **manager performance ratings** (Q1–Q3) related to ticket and process **characteristics** such as priority, time spent, steps, and comments?
Performance differences	Do some **assignees or projects** show systematically higher or lower **performance** scores?
Prediction of performance	**Prediction:** distinguish low from high **performance ratings** using only ticket and process features?
Consistency and potential bias	Are **similar tickets** with **similar process** metrics sometimes **rated** very differently, and does this vary across projects or assignees?
PROCESSES (using issues_change_history.csv)	
Detailed process paths	What are the most common **status sequences** from creation to closure, and which paths are associated with very long **processing times** or non-ideal outcomes?
Reassignments and ownership changes	How often are **tickets reassigned**, and are there "ping-pong" patterns between certain statuses or assignees?
Process mining-style analysis	Which specific **status transitions** appear frequently in problematic or slow cases?
COMMUNICATION (using sample_utterances.csv)	
Communication patterns	How many **comments are exchanged** per ticket or turn, and what types of placeholders (e.g. ph_log, ph_code) appear most often?
Role-based communication	How do **messages from reporters vs. assignees** differ in length, structure, or use of placeholders?
Process-performance link with communication	Are certain **communication patterns** (e.g. many back-and-forth messages, many technical placeholders) associated with different performance scores or resolution times?
Basic NLP feature extraction	Which simple **text-based features** (counts, lengths, placeholders) are most **informative** when combined with **process features** from the other datasets?

DataScientest

Description:
The project analysis a real helpdesk dataset from an international software company, covering tickets reported between 2016 and 2023. It combines ticket-level information (e.g. priority, type, processing time, workflow steps), assignee-level information (who worked on the ticket and for how long), and a manager's performance ratings for a stratified sample of ticket-handling episodes. Using this data, the project explores how characteristics of tickets and their handling processes relate to perceived work quality and client relations and build simple predictive models that estimates performance ratings from ticket and process features.

Aim:
The aim is to understand which ticket and workflow characteristics are associated with higher or lower performance appraisals in a helpdesk team, and to show how the analyses—potentially supported by a dashboard—could help with monitoring quality, identifying problematic patterns in ticket handling, and designing improvements to the support process.

Resources to refer to:

- ### Data:
  The available data consists of several related helpdesk datasets from an international software company. At the core are ticket-level tables (issues.csv and issues_snapshot.csv) with information on issue ID, project, type, priority, creation and resolution times, final status, total processing time, number of workflow steps, time spent in each status, and who worked on each ticket and when. A stratified sample file (issues_snapshot_sample.xlsx) adds manager performance ratings on three dimensions (Q1–Q3) for selected ticket-handling episodes, and two additional files (issues_change_history.csv and sample_utterances.csv) provide detailed logs of status/assignee changes and the text of messages exchanged between customers and the helpdesk team.
  Rows: 66,691
  https://data.mendeley.com/datasets/btm76zndnt/2 (CC BY 4.0 license) ☺

- ### Bibliography:
  Abdellatif, Mohammad (2025), "Help Desk Tickets", Mendeley Data, V2, doi: 10.17632/btm76zndnt.2

Validation conditions for the project:

- an exploration, data visualization and data pre-processing **report** ;
- a final **report** and associated **code.**