Cheat Sheet: Use of Generative AI for Data Engineering

Popular GenAI tools

Name of model	Usage	Link
Hal9	EDA tool to identify key insights on data	https://www.hal9.com/
Columns.ai	Data visualization tool to create useful charts	https://columns.ai/
Akkio	Data visualization tool to create data plots like regression plots, box plots, correlation heatmaps, and so on	https://www.akkio.com/
ChatGPT	AI language model	https://openai.com/chatgpt
sqlthroughAI	AI assistant for SQL queries	https://sqlthroughai.com/

Important prompts for data preparation

Task	Prompt
Data analysis and mining	Write Python code to analyze and mine a provided dataset (CSV file containing sales data)
Data pipelines and ETL workflows	Provide Python code for a data pipeline that accomplishes the following tasks: 1. Designs a data pipeline to extract data from a CSV file located at /content/CourseraDataset.csv. 2. Performs transformations to extract course details for courses rated 4.8 and above. 3. Loads the results into a CSV file at /content/HighRated_CourseraDataset.csv.

Important prompts for generating data insights and visualizations

Task	Prompt
Generate a statistical description of data	Write Python code to generate the statistical description of all the features used in the data set. Include "object" data types as well.
Create regression plots between a target and a continuous-valued source variable	Write Python code to generate a regression plot between a target variable and a source variable of a data frame.
Create box plots between a target and categorical source variable	Write Python code to generate a box plot between a target variable and a source variable of a data frame.
Evaluate parametric interdependence using correlation, p-value, and pearson coefficient	Write Python code to evaluate correlation, pearson coefficient, and p-values for all attributes of a data frame against the target attribute.
Group variables to create pivot tables, Create a p-color plot for the pivot table	Write Python code that performs the following actions: 1. Groups three attributes as available in a data frame df 2. Creates a pivot table for this group, using a target attribute and aggregation function as mean 3. Plots a pcolor plot for this pivot table

Important prompts for model development and refinement

Task	Prompt
Linear regression between a single source attribute and target attribute and evaluate it	Write Python code that performs the following tasks: 1. Develops and trains a linear regression model that uses one attribute of a data frame as the source variable and another as a target variable. 2. Calculates and displays the MSE and R^2 values for the trained model
Linear regression between multiple source attributes and target attributes and evaluate it	Write Python code that performs the following tasks: 1. Develops and trains a linear regression model that uses some attributes of a data frame as the source variables and one of the attributes as a target variable. 2. Calculates and displays the MSE and R^2 values for the trained model.
Polynomial regression model with a single source and target variable	Write Python code that performs the following tasks: 1. Develops and trains multiple polynomial regression models, with orders 2, 3, and 5, that use one attribute of a data frame as the source variable and another as a target variable. 2. Calculates and displays the MSE and R^2 values for the trained models. 3. Compares the performance of the models.
Pipeline creation for scaling, polynomial feature creation, and linear regression	Write Python code that performs the following tasks: 1. Create a pipeline for parameter scaling, polynomial feature generation, and linear regression. Use the set of multiple features as before to create this pipeline. 2. Calculate and display the MSE and R^2 values for the trained model.
Grid search with ridge regression and cross-validation	Write Python code that performs the following tasks: 1. Use polynomial features for some of the attributes of a data frame. 2. Perform a grid search on a ridge regression model for a set of values of hyperparameter alpha and polynomial features as input. 3. Use cross-validation in the grid search. 4. Evaluate the resulting model's MSE and R^2 values.

Author(s)

Abhishek Gagneja
Malini Sekar

Skills Network