AI-AUGMENTED WORKFLOW FOR DATA ANALYSTS – PART 1

Workflow Breakdown & Cursor-based Automation Examples

Prepared by - Seung-Mi Jeon, Hrishikesh Bhatt

PROJECT OBJECTIVE

Goal

1. Decompose the end-to-end workflow of a data analyst
2. Identify repetitive tasks vs. human judgment points
3. Explore how AI tools like Cursor can boost productivity

Key Deliverables

1. Workflow framework
2. Pain points and decision points (human vs. automation)
3. Example prompts & execution screenshots

(ex: automated code for missing value handling, outlier removal, natural language → SQL transformation)

WORKFLOW OVERVIEW

Data Loading

Inspection
& Cleaning

Exploration

Modeling

Decision
Support

DETAILED WORKFLOW BREAKDOWN

Workflow	Data Loading	Inspection & cleaning	Exploration	Modeling	Decision Support
Tasks	• Structured Data – Querying SQL and NoSQL databases • Structured data – loading csv, xml, json files • Unstructured data – open source and collect from the web through APIs	• Verify dataset size and structure • Duplicate value removal • Handle missing values • Outlier detection and removal • Standardization of formats	• Basic statistical Analysis(Distribution, Unique value analysis, Correlations, VIF) • Feature Engineering (Select features, Create new, Apply transformations)	• Hypothesis Testing • ANOVA • Supervised Learning • Unsupervised Learning • Optimization • A/B Testing, Causal Inference • Hyperparameter Tuning and Model Selection • Interpretability	• Interpreting the data and analytical outputs • Contextualizing insights for business case • Actionable strategies based on insights • Presenting insights to stakeholders
Rule-based steps	• Syntax and logic for writing SQL queries • Loading data into a DataFrame	• Verifying size and structure • Removing duplicate rows • Standardization of formats (datetime, numbers, etc.)	• EDA - visualize distributions of all key variables • Examine unique values in categorical or ordinal data • Generate heatmaps for all variables • Calculate VIF for all independent variables	• Calculate test statistic and p-value • Initialize and run model • K-elbow visualizer to select number of clusters • Find the optimal points	• Reviewing statistical findings, modelling results – identifying significant trends / anomalies / risks • Generate dashboards
Human Judgement	• Validate Query outputs • Query optimization given database schema	• Deciding rule(s) for outlier detection and removal given distributions • Deciding rules for handling missing values given counts and distributions	• Specify whether unique value analysis is required • Select which features to keep / drop • Engineer new interaction terms or derive new features • Apply transformations	• Formulate the hypothesis and confidence level • Define independent and dependent variables • Select appropriate model • Specify objective and constraints • Define hyperparameter range • Select the formula of error calculation and criteria • Define the criteria	• Contextualize insights for business case • Present insights to stakeholders - decide format (report, graphs, design, etc.) • Translate insights into plain-language and next steps for business stakeholders

METHODOLOGY

1. Under each major Data Analysis Workflow:
   - Develop detailed use cases and process flow charts
   - Outline rule – based and Human Judgement – based steps
2. For each rule – based and Human Judgement – based step:
   - Write prompts for AI augmentation (automate steps or give analyst options where necessary)
   - Create detailed prompt frames for different data and model types
3. Test prompts using Cursor

EXAMPLE : EXPANDED PROCESS FOR DATA EXPLORATION

Exploration

Task	Summary Statistics	Distribution Analysis	Correlation and Relationships	Visual Summaries
Sub-Tasks	• Calculate mean, median, mode, standard deviation, percentile	• Plot histograms, density plots, line charts to identify skewness and kurtosis	• Compute pairwise correlations • Identify multicollinearity • Visualize with heatmaps or scatter matrices	• Segmentation / Group-wise analysis • Generate plots to summarize insights from segmentation / group-wise analysis
Rule-based steps	• Identify variable type based on patterns • Calculate mean, median, standard deviation for numeric variables • Calculate mode for categorical variables	• Auto - generating plots and matrices • Calculating skewness and kurtosis in the variables • Suggest transformations to normalize skewed data	• Correlation computation • Auto – generate correlation matrices with heatmaps for all variables • Calculate VIF for all independent variables • Explain which relationships are statistically significant	• Generate plots to summarize insights (stacked bar charts, column charts, differentiation using color, etc.)
Human Judgement	• Identify unusual variables (e.g. categorical variables with levels labelled numerically • Apply standardization techniques	• Visually inspect distributions and decide if transformations are needed • Interpreting implications on modelling • Apply transformation	• Select which features to keep / drop • Engineer new interaction terms or derive new features	• Based on dataset description, deciding dimensions along which group-wise analysis to be done • Design of outputs

CONSTRUCT PROMPT FOR EDA

You are a highly experienced data analyst conducting exploratory data analysis. The dataset is 'Customer_cleaned.csv', which has already been completely preprocessed.

…

Each file you generate should reflect the user's decision. Ensure that each generated file is clearly named and output files from each step are located on the folder named 'eda_step()_output'. Please include descriptive comments above each code block explaining what it does.

…

There are 4 steps for EDA. At each stage, always ask for human input in the chat before proceeding with tasks that require judgment, and generate separate EDA files based on the user's decisions. To use a prompt-chaining method, I will give you prompts for those 4 steps one by one. Wait for the user's natural language response in the chat, and then proceed to generate the corresponding code in a separate EDA analysis file.

Step 1: **Summary Statistics:**

- Calculate mean, median, mode, standard deviation, and percentiles for numerical variables.

- *` Rule-based:` * Automatically identify variable types based on patterns and compute these metrics for numeric variables; for categorical variables, compute the mode.

- *` Human judgement:` * Identify any unusual variables (e.g., categorical variables with numerically labeled levels) and decide if standardization is needed.

Split EDA Process with 4 steps and distinguish rule-based and human judgement ones with backtick!

FURTHER STEPS

1. Under each major Data Analysis Workflow:
   - Develop detailed use cases and process flow charts for AI – enabled deployment
   - Outline rule – based and Human Judgement – based steps
2. For each rule – based and Human Judgement – based step:
   - Write prompts for AI augmentation (automate steps or give analyst options where necessary)
   - Create detailed prompt frames for different data and model types
3. Develop metrics to test performance
4. Test prompts using Cursor