DIY Master's in Data Science (AI Era) — Study Session 1 Checklist

Today's goal: Begin your first hands-on data exploration session (Term 1: Foundations).

■ PREP (10–15 min)

- ■ Open WSL → run: conda activate ds
- ■ Launch Jupyter Lab → jupyter lab
- ■ Confirm kernel: Python (ds)
- ■ Open notebook: ds-zero-to-one/notebooks/01_intro_to_eda.ipynb

■ PART 1 — Load and Inspect Data (30–45 min)

- Import libraries: pandas, duckdb, polars
- Load the dataset (tips.csv) from seaborn-data
- Use df.info() to see columns and data types
- Use df.describe(include='all') for summary stats
- Check missing values with df.isna().sum()

■ PART 2 — Clean and Rename (30 min)

- Normalize column names → lowercase, underscores
- Rename key columns (e.g., total_bill → bill_total_usd, tip → tip_usd)
- Add a new column: df['tip_pct'] = df['tip_usd'] / df['bill_total_usd']
- Preview cleaned dataset with df.head()
- Save to data/processed/tips_cleaned.csv

■ PART 3 — Exploratory Analysis (45–60 min)

- Create Markdown section: ## Exploratory Analysis: Summary Statistics
- Run df.describe() and df.groupby('gender')['tip_pct'].mean()
- Write short notes on patterns you notice
- Use matplotlib → df['tip_pct'].hist()
- Try plotly express → px.scatter(df, x='bill_total_usd', y='tip_usd', color='gender')

■ PART 4 — Commit Your Work (10–15 min)

- git add notebooks/01_intro_to_eda.ipynb data/processed/tips_cleaned.csv
- git commit -m 'Session 1: cleaned + explored tips dataset'
- git push

- Check GitHub Actions → CI → ensure it prints 'CI env OK'

■■■■ REFLECT (Optional)

- What did you learn about your dataset today?
- What commands felt natural? Which need more practice?
- Set a reminder for tomorrow's session: visualization deep dive.

■ Tip: Print this checklist or keep it open beside your notebook for real-time tracking.