Dataset insights and recommended ML use-cases

Dataset inventory and best ML use-cases

Inspected files:
- ai_job_market.csv — job listings with skills, titles, salary_range, company_size, posted_date, tools_preferred.
- customer_churn_dataset-testing-master.csv — labeled churn dataset with demographics, usage, support calls and `Churn` label.
- WA_Fn-UseC_-Telco-Customer-Churn.csv — Telco churn dataset (classic churn prediction problem).
- Google_stock_data.csv — historical OHLCV time series for Google (Date, Close, High, Low, Open, Volume).
- Mobile Reviews Sentiment.csv — large reviews dataset with `review_text`, `sentiment` labels and per-aspect ratings.
- Mental_Health_and_Social_Media_Balance_Dataset.csv — survey-style user data with Happiness_Index and behavioral features.
- Morning_Routine_Productivity_Dataset.csv — daily routine records with productivity score (1–10).
- products.csv — transactional basket data: lists of products per transaction.
- Global GDP Explorer 2025 (World Bank UN Data).csv — country-level GDP statistics (static table).

Recommended ML use-cases (high-level)

1) Churn prediction (customer_churn and Telco churn):
- Task: Supervised binary classification (predict churn).
- Models: Logistic Regression baseline; tree-based models (XGBoost/LightGBM/CatBoost).
- Metrics: AUC-ROC, Precision/Recall, F1, PR-AUC for imbalanced classes.
- Notes: handle categorical encoding, missing values, class imbalance, avoid time leakage.

2) Sentiment analysis (Mobile Reviews Sentiment):
- Task: Text classification (Positive/Negative/Neutral); aspect-based sentiment analysis.
- Models: TF-IDF + Logistic Regression or SVM baseline; fine-tune transformers (DistilBERT, XLM-R) for multilingual data.
- Metrics: Accuracy, Macro F1, confusion matrix; per-aspect F1 where applicable.

3) Time-series forecasting (Google_stock_data):
- Task: Forecast Close price or returns; anomaly detection.
- Models: ARIMA/Prophet baselines; LSTM/Temporal Transformer or gradient boosting on lag features.
- Metrics: MAE, RMSE, MAPE (careful near zero); use walk-forward validation.

4) Market-basket analysis and recommender (products.csv):
- Task: Association rules (Apriori/FP-Growth); recommenders (item2vec, collaborative filtering).
- Metrics: Support/confidence/lift for rules; Precision@K, Recall@K, NDCG for recommenders.

5) Salary prediction & skill extraction (ai_job_market.csv):
- Task: Regression for salary (use median of range) and NLP for skills extraction / job classification.
- Models: XGBoost for regression; spaCy/transformers for skill extraction.

6) Happiness/productivity studies (Mental Health, Morning Routine):
- Task: Regression (predict happiness or productivity), clustering for segmentation, explainability/feature importance.
- Models: Linear models and tree-based models; KMeans or hierarchical clustering.
- Notes: Great for interpretable models and causal-style exploration.

Top 3 starter projects (fast wins):
- Telco churn prediction (Telco dataset): high ROI, quick baseline and improvements.
- Sentiment classification (Mobile Reviews): good for NLP practice and transfer learning.
- Stock forecasting (Google_stock_data): learn time-series validation and forecasting models.

Next steps suggestions
- Choose one dataset and scaffold a notebook: EDA, preprocessing, baseline, model, evaluation, export.

- Provide a runnable notebook or scripts and a short README.