Data Analysis Final Assignment Report

Team: Analog Avengers
Eingang Fabian & $KotschnigThomas$ & $KrennMatthias$

Note: This template provides a suggested structure aligned with the current task categories. You may adjust headings if needed and please ensure all required components are covered.

1 Contributions

Clearly state each team member's specific contributions. Be concrete.

- Eingang Fabian:
  - Dataset selection and acquisition
  - Data quality analysis and preprocessing pipeline
- Kotschnig Thomas:
  - Visualizations and EDA
  - Probability analysis tasks
- Krenn Matthias:
  - Regression modeling and interpretation
  - Report writing and figure polishing

2 Dataset Description

- "Bike sales in Europe" from https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe
- It has more than 100k entries of sales data from different countries. Streching from 2011 to 2016, with a daily sampling frequency.
- Key variables analyzed: Unit price, Unit cost, date, order quantity, customer age, product type
- Shape: 113036 rows x 18 columns
- No missing data, however, the entry of some dates is missing completely. This resolves in no missing data, but inconsistent time series. There is only one bigger gap, therefore we decided
- Any known limitations or caveats:

3 Task 1. Data Preprocessing and Basic Analysis

3.1 Basic statistical analysis using pandas

- Descriptive stats (mean, std, min, max, quantiles) for key variables:
- Grouped summaries where relevant (by day, device, category, test run):

3.2 Original data quality analysis including visualization

- Missingness patterns (counts, heatmap, timeline gaps):
- Outliers and suspicious values (plots and rule used):
- Consistency checks (timestamps order, duplicates, impossible values):

3.3 Data preprocessing

- Cleaning steps performed:
- Missing-value treatment (drop, impute, interpolate, forward fill, etc.):
- Outlier handling (range, threshold, IQR, percentile, justify choice):
- Feature engineering (e.g., scaling/normalization, log):
- Final dataset shape after preprocessing:

3.4 Preprocessed vs original data visual analysis

- Before vs after comparison plots (at least 2 to 3 key variables):
- What improved and what trade-offs exist:

4 Task 2. Visualization and Exploratory Analysis

4.1 Time series visualizations

- Plot of main variable(s) over time:
- Annotations for notable events or pattern shifts (if applicable):

4.2 Distribution analysis with histograms

- Histograms for key numeric variables:
- Notes on skewness, heavy tails, multi-modality:

4.3 Correlation analysis and heatmaps

- Correlation type used (Pearson or Spearman) and why:
- Heatmap and top correlated pairs with short interpretation:

4.4 Daily pattern analysis

- Aggregation method (hourly means, day-of-week, rolling averages):
- Plots showing daily cycles or weekday-weekend differences:
- What patterns are stable vs noisy:

4.5 Summary of observed patterns, similar to True/False questions

Write short, testable statements and answer them based on evidence. Example format below.

- Statement 1 (True or False): **....** Evidence: ...
- Statement 2 (True or False): **....** Evidence: ...
- Statement 3 (True or False): **....** Evidence: ...

5 Task 3. Probability Analysis

5.1 Threshold-based probability estimation

- Define threshold(s) and justify choice:
- Estimate probabilities of exceeding thresholds:
- Visual support (e.g., empirical CDF, bar plot, timeline highlights):

5.2 Cross tabulation analysis

- Define two categorical variables (or binned numeric variables):
- Present contingency table and interpret key cells:

5.3 Conditional probability analysis

- Define events A and B:
- Compute and interpret $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \mid A)$:
- Include at least one meaningful comparison and conclusion:

5.4 Summary of observations from each probability task

- Key takeaway from threshold probability:
- Key takeaway from crosstab:
- Key takeaway from conditional probability:

6 Task 4. Statistical Theory Applications

6.1 Law of Large Numbers (LLN) demonstration

- Variable chosen and why it makes sense:
- Experiment: show sample mean as n increases:
- Plot and short interpretation:

6.2 Central Limit Theorem (CLT) application

- Sampling procedure (sample size, number of trials, with or without replacement):
- Show distribution of sample means for increasing n:
- Plot(s): histogram(s) of sample means and comparison to normal shape:

6.3 Result interpretation

- What LLN showed in your data context:
- What CLT showed, and any deviations and why:

7 Task 5. Regression Analysis

7.1 Linear or Polynomial model selection

- Define target y and predictors X:
- Motivation for linear vs polynomial:
- Any train-test split rationale (time-aware split if relevant):

7.2 Model fitting and validation

- Fit procedure and preprocessing (scaling, feature selection):
- Validation method (holdout, time-series split, etc.):
- Metrics reported (RMSE, MAE, R^2) and why:
- Residual analysis (at least one plot recommended):

7.3 Result interpretation and analysis

- Main effects and practical meaning:
- Failure cases or where model performs poorly:

8 Bonus Tasks

- New dataset bonus (10): state why dataset is new and provide link:
- Q-Q plot with explanation (5):
  - Either for CLT sample means, or regression residuals:
  - Interpretation of deviations from normality:
- Interactive visualizations (up to 10): describe tool used and what interactivity adds:
- Cross-validation in regression (5): method used and how results compare to holdout:
- Additional exploration (up to 20): clearly state extra tasks and value gained:

9 Key Findings and Conclusions

- Main findings from preprocessing and EDA:
- Main findings from probability tasks:
- Main findings from LLN and CLT:
- Main findings from regression:
- Limitations:
- What you would do next if you had more time:

10 Reproducibility Notes

- Exact dataset source link and version or download date:
- Key libraries used and versions (optional but recommended):
- How to run the notebook end-to-end: