Data Analysis Final Assignment Report

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

1 Contributions

- 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, profit, cost, revenue, date, order quantity, customer age, product category, product
- 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 for it to be okay.

3 Task 1. Data Preprocessing and Basic Analysis

3.1 Basic statistical analysis using pandas

Statistical summary of key numeric variables was obtained using pandas `describe()` function:

Table 1: Desriptive statistics

	Customer Age	**Order Qty**	**Unit Cost**	**Unit Price**	**Profit**	**Cost**	**Revenue**
Count	113036	113036	113036	113036	113036	113036	113036
Mean	35.92	11.90	267.30	452.94	285.05	469.32	754.37
Std	11.02	9.56	549.84	922.07	453.89	884.87	1309.09
Min	17	1	1	2	-30	1	2
25%	28	2	2	5	29	28	63
50%	35	10	9	24	101	108	223
75%	43	20	42	70	358	432	800
Max	87	32	2171	3578	15096	42978	58074

Table 2: Grouped summary of Revenue, Profit, and Order Quantity by Country

Country	**Revenue**			**Profit**		**Order Quantity**	
	Sum	**Mean**	**Count**	**Sum**	**Mean**	**Sum**	**Mean**
United States	27975547	713.55	39206	11073644	282.45	477539	12.18
Australia	21302059	889.96	23936	6776030	283.09	263585	11.01
United Kingdom	10646196	781.66	13620	4413853	324.07	157218	11.54
Germany	8978596	809.03	11098	3359995	302.76	125720	11.33
France	8432872	766.76	10998	2880282	261.89	128995	11.73
Canada	7935738	559.72	14178	3717296	262.19	192259	13.56

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):

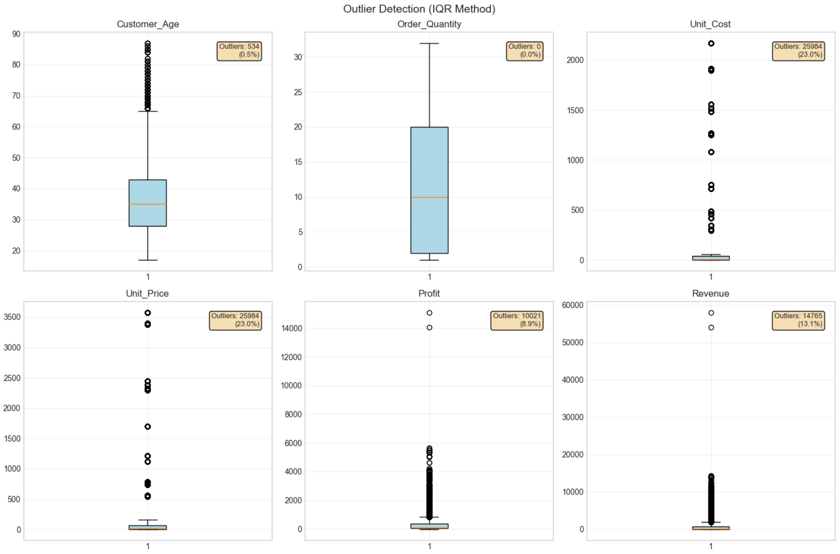

Figure 1: Your figure caption

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: