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: customer age, order quantity, unit cost, unit price, profit, cost, revenue
- 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

There are no missing data in our dataset. This is why we did not add any visualization of this parameter. However, there is a timeline gap visible in the figure below.

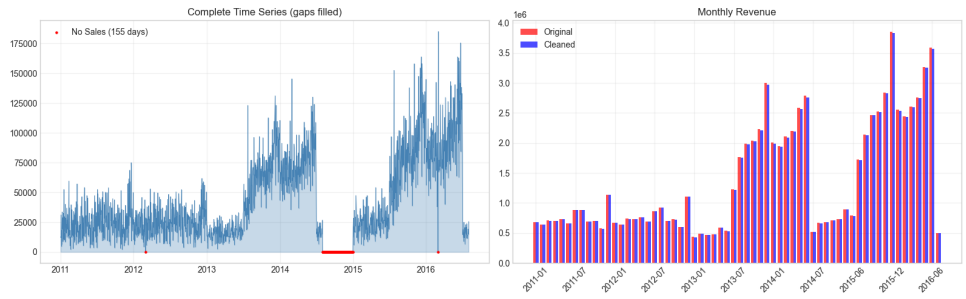

Figure 1: Consistency checks in time series and comparison of revenue before and after preprocessing

Outliers have been identified via IQR method seen in the figure below, but they have not been removed. This is because we wanted this data in the dataset for better and full analysis. Otherwise the dataset would have lost too much information.

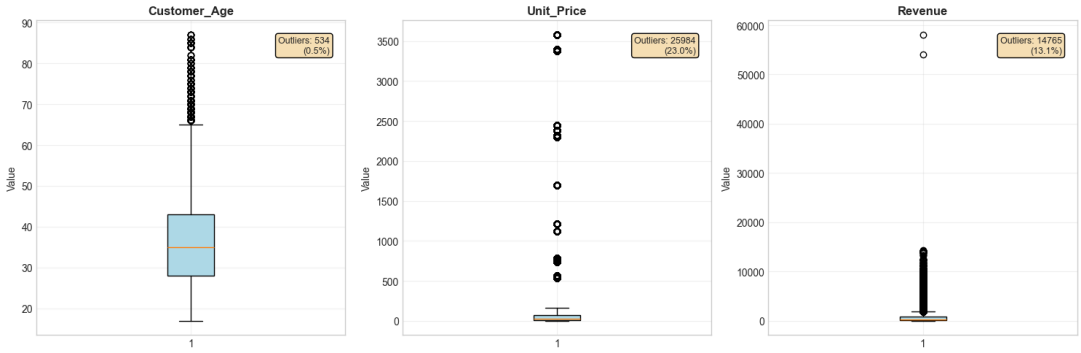

Figure 2: IQR applied on some key variables to identify outliers.

3.3 Data preprocessing

- Cleaning steps performed: Duplicates have been dropped.
- Missing-value treatment: No missing values were present in the dataset, therefore no treatment was necessary.
- Outlier handling: Outliers identified via IQR method (1.5 × IQR threshold) but intentionally retained. Justification: These extreme values contain valuable business insights (high-value transactions, unusual market events) that would be lost if removed.

- Feature engineering:
  Time-based features added: DayOfWeek, DayOfWeek_Name, WeekOfYear, Quarter, IsWeekend
  Financial features created: Profit_Margin, Avg_Unit_Profit, High-value flag

- Final dataset shape after preprocessing: (8 new features added, no rows removed)
  Original 113,036 rows × 18 columns → Cleaned 113,036 rows × 26 columns

3.4 Preprocessed vs original data visual analysis

The accuracy of the dataset improved by dropping the duplicates. A trade-off would be the possible removal of relevant data (no real duplicate).

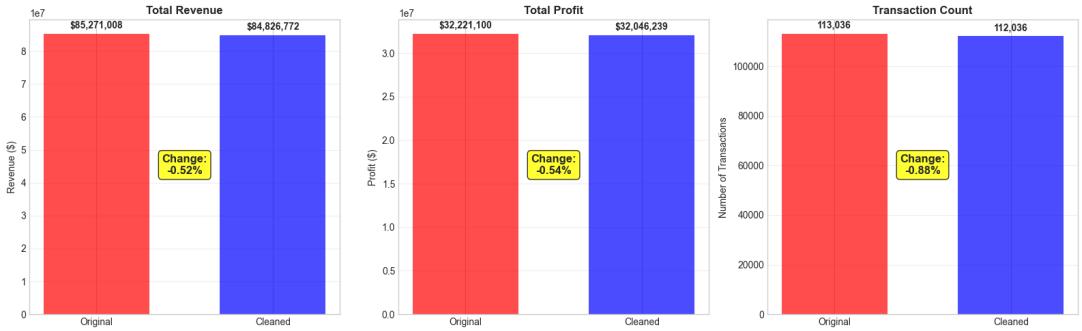

Figure 3: Comparison of key sales values due to the removal of 1000 duplicates.

4 Task 2. Visualization and Exploratory Analysis

4.1 Time series visualizations

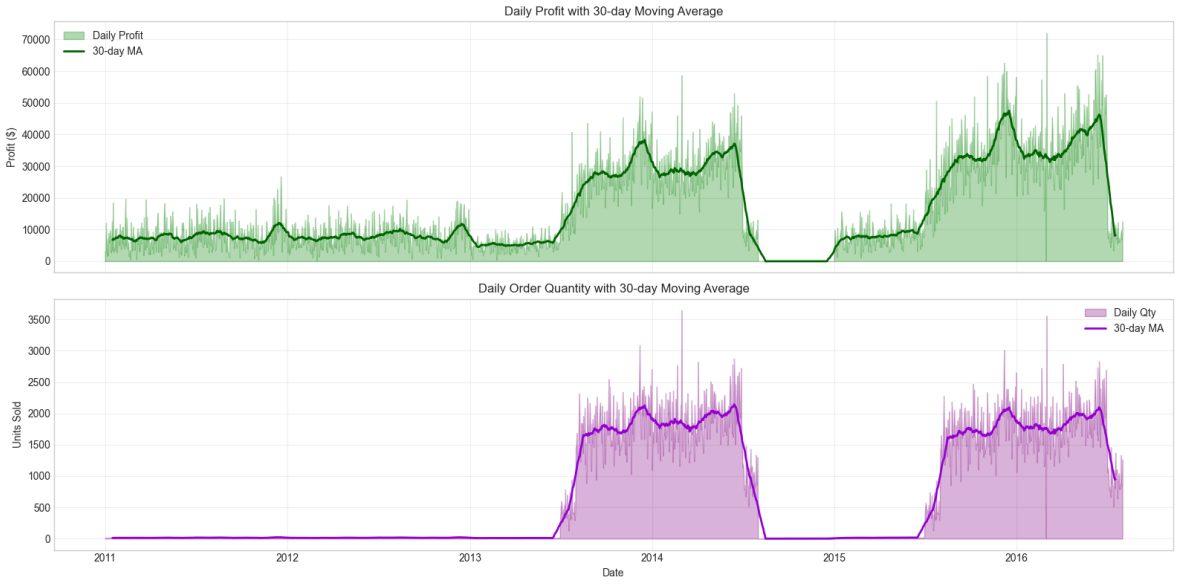

Figure 4: Time series visualization of sales over time.

4.2 Distribution analysis with histograms

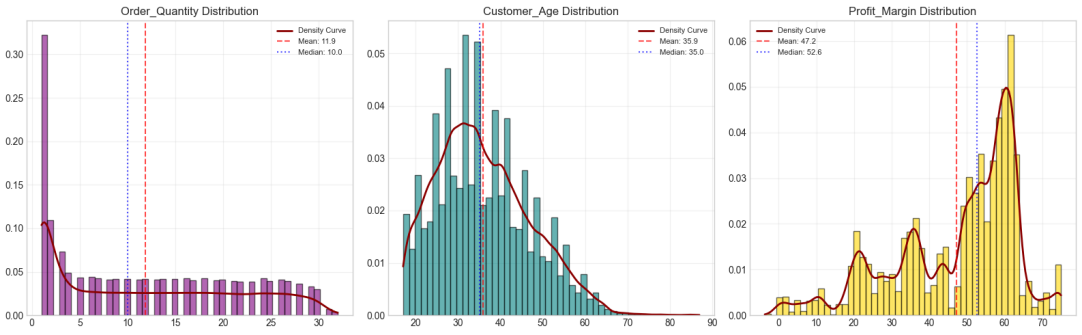

Figure 5: Histograms of key numeric variables showing distribution shapes.

Order Quantity: Nearly symmetric distribution (skewness: 0.378) with light tails, indicating values are concentrated near the center with fewer extreme outliers. The bimodal pattern suggests two distinct purchasing behaviors.

Customer Age: Right-skewed distribution (skewness: 0.524) with normal-like tails. The multimodal pattern reflects age clustering across different customer segments. Mean age of 35.92 years exceeds the median (35.00), confirming right skew with older customers in the tail.

Profit Margin: Left-skewed distribution (skewness: -0.856) with normal-like tails, indicating most transactions cluster at higher profit margins with a tail toward lower margins. The multimodal pattern suggests distinct profit tiers based on product categories.

4.3 Correlation analysis and heatmaps

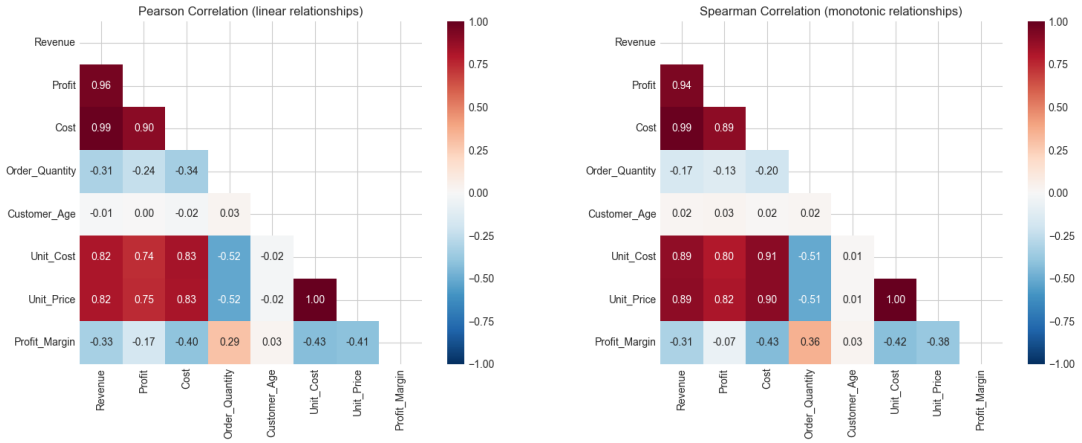

Figure 6: Correlation heatmap of key numeric variables.

Both correlation types have been calculated. Both types show the same strongest correlations. Those correlations are between the financial variables revenue, cost and profit. That makes sense, due to higher revenue leading to higher profit. Or higher costs also lead to higher revenue.

4.4 Monthly pattern analysis

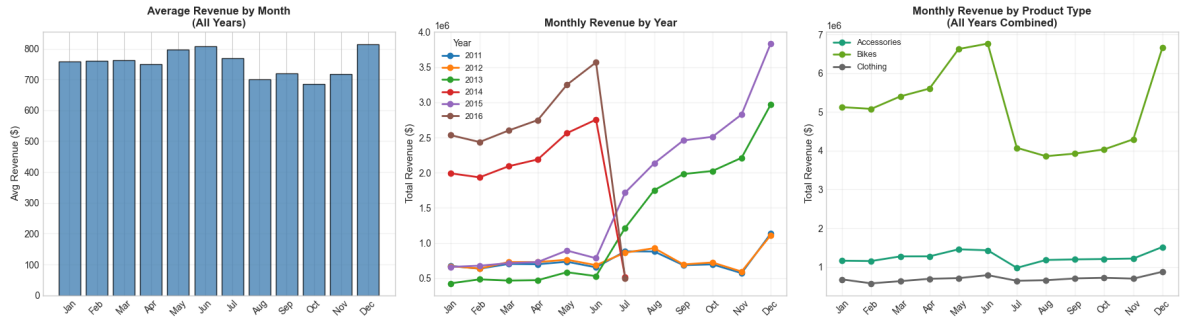

Figure 7: Monthly pattern analysis of sales.

Our dataset doesnt provide any timestamp data for the sales. Therefore, it was grouped into monthly data and reviewed over the whole year. The data of every month was then averaged over the years. It is very interesting to see the strong seasonal patterns of bike sales. Our dataset shows almost identical sales on the weekend compared to the weekdays. This shows that the shop was also open on weekends. Which is very unusual for euopean shops. Therefore, it has to be an online shop. It was interesting to see a deviation in customer age depending on the day of the week.

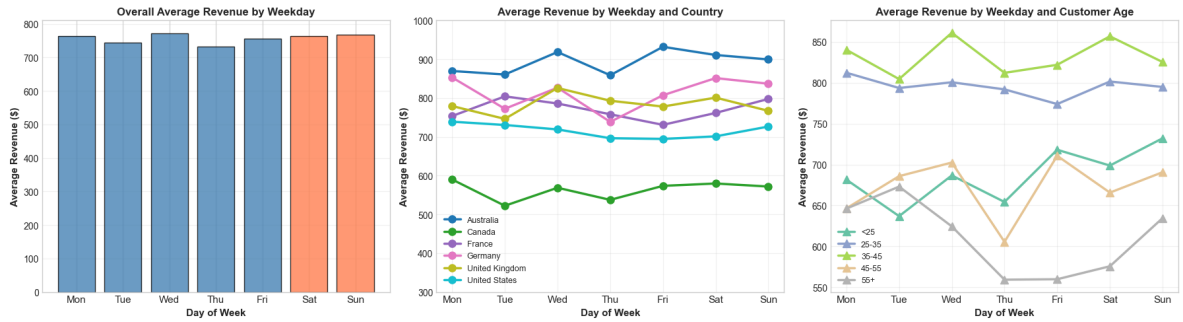

Figure 8: Weekday pattern analysis of sales.

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

1. Revenue shows a positive long-term trend over the dataset period **TRUE**
   - Evidence: 90-day MA grew 308.2% from start to end
2. Q4 (Oct-Dec) shows significantly higher sales than other quarters **FALSE**
   - Evidence: Q4 revenue is not higher than the average of other quarters
3. Revenue and Profit are strongly positively correlated (r ¿ 0.8) **TRUE**
   - Evidence: Pearson r = 0.957
4. Customer age has minimal impact on transaction revenue **TRUE**
   - Evidence: Age-Revenue correlation r = -0.009
5. Friday is the best performing day of the week **FALSE**
   - Evidence: Best day: Wednesday, worst: Thursday (5.2% difference)

6. December shows the highest monthly average revenue **TRUE**

   - Evidence: Best month: Dec, worst: Oct

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: