Data Wrangling Project3

This project advances beyond basic data wrangling to execute a comprehensive **exploratory data analysis (EDA) aimed at uncovering deeper patterns, testing statistical hypotheses, and engineering new features for enhanced business intelligence**. The central thesis is that a clean dataset is not the endpoint but the starting point for a deeper investigation using statistical and visual methods, which can validate assumptions and reveal nuanced insights that a surface-level analysis would miss.

Execution

1. **Foundation (Wrangling & Cleaning):** The project starts with the same foundational data wrangling steps: loading, merging, and cleaning the four datasets. This initial phase establishes a reliable, unified dataset upon which the subsequent, more sophisticated analysis is built.
2. **Exploratory Data Analysis (EDA):** This is the core of the project. Instead of just reporting totals, it seeks to understand the *characteristics* and *relationships* within the data through multiple analytical techniques:
    - **Descriptive Statistics:** The analysis begins by generating descriptive statistics (.describe()) and a data summary (.info()), providing a foundational understanding of the data's central tendency, spread, and structure.
    - **Data Visualization:** Several plots are generated to visually explore the data:
        - **Histograms** of Purchase Price and Reimbursement Amount reveal the frequency distribution of these key financial variables.
        - A **Scatter Plot** of Purchase Price versus Reimbursement Amount is used to visually inspect the relationship between cost and revenue, with denied claims highlighted to check for visual patterns.
        - A **Correlation Heatmap** provides a quantitative summary of the linear relationships between all numerical features, quickly identifying which variables move in tandem.
3. **Statistical Inference:** The project moves from exploration to validation by conducting a formal hypothesis test. An **independent samples t-test** is performed to determine if there is a statistically significant difference between the reimbursement amounts for approved versus denied claims. The resulting high p-value (0.9831) leads to the conclusion that there is no significant difference, a crucial insight that might contradict initial assumptions.
4. **Feature Engineering:** The analysis is further enriched by creating a new, more insightful metric. **Profit Margin** is calculated as (Profit / Amount) * 100. This engineered feature

provides a normalized measure of profitability, which is more comparable across vials with different reimbursement amounts than the absolute profit alone. A histogram of this new feature is then plotted to understand its distribution.

Conclusion

This project exemplifies a mature data analysis workflow that progresses logically from cleaning to exploration, inference, and feature engineering. It demonstrates that by applying a combination of statistical summaries, data visualization, and hypothesis testing, a deeper, more reliable understanding of the business can be achieved. The findings—such as the lack of a significant difference in denied claim amounts and the distribution of profit margins—provide robust, evidence-backed insights that are far more powerful than simple summary reports.