Homework Grading Report

Student Name:	Mahreen Maknojia
Assignment:	Assignment 1 - Introduction to R
Graded On:	October 03, 2025 at 11:40 PM
Final Score:	34.2 / 37.5 points (91.2%)

Score Summary

Overall Performance: Excellent (91.2%)

Instructor Assessment

Your submission demonstrates strong engagement with the core concepts of data cleaning and a clear ability to connect technical choices to business outcomes. You effectively argued for imputation to preserve valuable sales information and chose capping to retain outlier observations without distorting overall trends. Your reflections on ethical considerations and the business relevance of missing data and outliers are thoughtful and well-articulated. To elevate your work further, focus on providing the exact imputation code, quantitative thresholds for outlier handling, and comparative summary statistics that illustrate the impact of each cleaning step. Keep building on this solid foundation, and you will continue to develop the analytical rigor and professional judgment essential for a successful career in business analytics.

Reflection & Critical Thinking

Your explanation of when to remove versus impute missing values shows solid business intuition; you correctly identified that a missing last name in a survey may be dispensable, while a missing location is critical for demographic analysis. This demonstrates an ability to weigh the analytical importance of each variable against data loss. Your answer on the ethical implications of data cleaning highlighted the need for transparency and maintaining the overall picture of the data, which is essential for building stakeholder trust. By emphasizing accurate reporting and openness about modifications, you show awareness of professional responsibility.

Analytical Strengths

Your R code correctly loads the tidyverse library, sets the working directory, and imports the messy_sales dataset, confirming that you can manage the environment and data ingestion steps reliably. The printed messages about row and column counts demonstrate good practice for verifying successful imports. The use of str() and summary() to inspect the dataset provides a comprehensive initial assessment, allowing you to quickly spot variable types and potential quality issues. This systematic approach shows you understand the importance of early data profiling. You created a boxplot to visualize outliers in the Sales_Amount variable and then applied both removal and capping strategies, which illustrates a thorough exploration of alternative treatments. Presenting both options indicates strong analytical thinking and a willingness to compare outcomes.

Business Application

Your recommendation to use imputation for this dataset is well-aligned with the business goal of preserving as much sales information as possible, especially when missing values are random. By keeping the dataset size stable, you enable more reliable downstream analyses such as forecasting. Choosing the capped version of the dataset as the final cleaned product reflects a balanced business decision: you retain all observations while mitigating the distortion caused by extreme values. This approach supports accurate average sales calculations without sacrificing sample size. Linking the removal of non-random missing rows to potential loss of valuable insights demonstrates that you can translate data-quality decisions into concrete business impacts, such as maintaining the integrity of customer segmentation analyses.

Learning Demonstration

Your discussion of the trade-offs between removal and imputation shows you grasp the statistical assumptions each method introduces, indicating a developing competency in data-quality theory. Recognizing that imputation introduces assumptions while removal reduces sample size is a key learning outcome. By articulating how outliers could stem from real business events versus data entry errors, you demonstrate an ability to differentiate between signal and noise-a critical skill for any analyst. This reflects growing maturity in interpreting data within its operational context. Your ethical reflection that analysts must be transparent about data modifications indicates an emerging understanding of professional standards, which will serve you well as you advance in analytics roles.

Areas for Development

While you correctly identified missing value patterns, providing the actual code used for imputation (e.g., median for numeric, mode for categorical) would strengthen the reproducibility of your work and allow reviewers to assess the appropriateness of the chosen methods. Your outlier treatment description would benefit from quantitative justification, such as reporting the IQR thresholds used for capping and the percentage of rows affected. Including these details would make your impact assessment more rigorous. Consider adding a concise summary table that compares key statistics (mean, median, standard deviation) before and after each cleaning step; this would give a clearer picture of how the data quality improvements affect analytical results.

Recommendations for Future Work

Practice implementing multiple imputation techniques (e.g., mice or missForest) on larger, more complex datasets to deepen your understanding of handling missing data under different missingness mechanisms. Explore robust outlier detection methods such as the MAD (median absolute deviation) or DBSCAN clustering, which can capture non-Gaussian outliers that the IQR method might miss. Applying these to sales data will broaden your toolkit. Continue documenting each cleaning decision in markdown cells with explicit rationale and code snippets; this habit will enhance the transparency of your workflow and prepare you for collaborative analytics projects.

Technical Analysis

Code Strengths:

• You correctly implemented the missing value detection and treatment approach using base R functions. Your use of `sum(is.na(messy_sales))` and `colSums(is.na(messy_sales))` accurately calculates total and column-wise missing values. The logic for identifying rows with missing values using `complete.cases(messy_sales)` is sound and produces the expected results.

• Your implementation of the mode function using `unique()`, `tabulate()`, `match()`, and `which.max()` demonstrates solid understanding of R's vector operations. You properly handled the categorical missing value imputation for Customer_Name by replacing NA values with the mode. This approach shows good technical competency in data cleaning techniques.

• You effectively applied the IQR method for outlier detection by correctly calculating quartiles and thresholds. Your use of `quantile()` with `na.rm = TRUE` ensures proper handling of missing values during calculations. The approach to identifying outliers using logical conditions shows good analytical thinking in data quality assessment.

Code Improvement Suggestions:

• While your approach to removing rows with missing values works correctly, you could simplify the code by using `drop_na()` from the tidyr package: `sales_removed_na <- drop_na(messy_sales)`. This alternative approach is more readable and commonly used in professional analytics workflows, making your code cleaner and more maintainable.

• Your outlier detection logic has a sign error in the threshold calculations. You wrote `upper_threshold <- Q1_sales - 1.5 IQR_sales` and `lower_threshold <- Q3_sales - 1.5 IQR_sales` which should be `upper_threshold <- Q3_sales + 1.5 IQR_sales` and `lower_threshold <- Q1_sales - 1.5 IQR_sales`. This would produce incorrect outlier identification and removal.

• For the outlier removal section, you left `outliers <- # YOUR CODE HERE` as a comment. You should replace this with `outliers <- sales_imputed[sales_imputed$Sales_Amount < lower_threshold | sales_imputed$Sales_Amount > upper_threshold, ]` to properly identify outlier rows. This would complete your outlier analysis and make your code fully functional.

Technical Observations:

• You demonstrate solid understanding of data cleaning fundamentals through your implementation of missing value treatment and outlier detection. Your ability to work with base R functions for data manipulation shows good technical competency in business analytics programming. This foundation will serve you well as you tackle more complex data analysis tasks.

• Your approach to handling missing data using both removal and imputation strategies shows good analytical thinking. You correctly applied different techniques for categorical (mode) and numeric (median) variables, which demonstrates understanding of appropriate data treatment methods. This shows developing expertise in data quality assessment.

• Your code organization supports reproducible analysis through clear sectioning and logical flow. You structured the missing value treatment section in a way that makes the logic clear and the results verifiable. This attention to code quality and documentation is important for professional analytics work and shows good programming practices.

Additional Code Enhancement Examples:

Data Exploration Enhancement:

```
# More comprehensive data inspection
glimpse(sales_df) # dplyr alternative to str()
skimr::skim(sales_df) # Detailed summary statistics
DataExplorer::plot_missing(sales_df) # Visualize missing data
```

Data Visualization:

```
# Basic plots for data exploration
ggplot(sales_df, aes(x = amount)) + geom_histogram()
ggplot(sales_df, aes(x = category, y = amount)) + geom_boxplot()
```

Data Cleaning:

```
# Handle missing values
sales_df <- sales_df %>%
filter(!is.na(amount)) %>%
mutate(amount = ifelse(amount < 0, 0, amount))
```

Performance by Category