

Homework Grading Report

Student Name:	Francisco Guadarrama
Assignment:	Assignment 1 - Introduction to R
Graded On:	October 02, 2025 at 11:15 PM
Final Score:	34.2 / 37.5 points (91.2%)

Score Summary

Overall Performance: Excellent (91.2%)

Instructor Assessment

They also answered reflection questions in text. They did import, printed structure, summary, and wrote observations. Good. Part 2: missing value analysis and treatment. They computed total missing, missing per column, removed rows, imputed Customer_Name with mode, imputed Quantity with median, imputed Sales_Amount with median (though they used median for Sales_Amount, but they also discuss not imputing). They also compared summary stats after removal and imputation. So they completed. Part 3: outlier detection and treatment. They computed IQR thresholds, flagged outliers, created boxplot, removed outliers using Z-score >3, also performed capping (winsorization) for numeric columns, then specifically capped Sales_Amount. They printed summary, final dataset. Part 4: final dataset justification: they wrote justification in the markdown earlier (selected sales_outliers_capped). They also created comparison summary table. Part 5: reflection questions: They answered Q1 (missing value strategy), Q2 (outlier interpretation), Q3 (data quality impact), Q4 (ethical considerations). They gave thorough answers. Thus they completed all parts. Good depth. Good. But overall understandable. They used code to compute missing counts, outlier counts, etc. They gave some analysis of impact. They also used Z-score removal. They could have discussed why choose median for Sales_Amount (they earlier said not to impute). But they still imputed median for Sales_Amount. Slight inconsistency. But overall appropriate. Reflection assessment: mention thoroughness, examples, but could be deeper on trade-offs. Analytical strengths: code for missingness, imputation, outlier detection, capping, summary table. Business application: linking to revenue analysis, forecasting, compliance, etc. Learning demonstration: shows ability to use R, functions, summarise, etc. Areas for development: missing handling of negative Sales_Amount as returns, need to document decisions, avoid double imputation of Sales_Amount, could use more robust imputation methods, could show before/after visualizations, could include reproducibility (set seed), could comment code. Recommendations: next time, separate imputation and outlier steps, consider model-based imputation, use tidyverse pipelines, include more visualizations, document decisions.

```
assistantfinal{ "overall_score": 88, "business_understanding": 92, "communication_clarity": 85, "data_interpretation": 88, "methodology_appropriateness": 84, "reflection_quality": 90, "detailed_feedback": { "reflection_assessment": [ "Your reflection answers are thorough and well-structured. You clearly differentiate when removal versus imputation is appropriate, and you back each recommendation with concrete business examples.", "The discussion of outlier interpretation and ethical considerations shows a solid grasp of the broader impact of data-cleaning decisions on reporting, modeling, and compliance.", "A deeper dive into the potential bias introduced by median imputation of a key variable like Sales_Amount would strengthen the reflection, especially given your earlier note that imputation could distort totals." ], "analytical_strengths": [ "You successfully imported the dataset, inspected its structure, and produced a concise summary of missingness (total and per-column counts).", "The
```

mode function you wrote for categorical imputation works correctly, and you applied median imputation to numeric fields, demonstrating practical handling of different data types.", "Outlier detection using the IQR rule, Z-score filtering, and winsorization (capping) are all correctly implemented, and you visualized the results with clear boxplots.", "The final comparison table quantifies how row count, missing values, means, medians, and outlier counts changed from the raw to the cleaned data."], "business_application": ["You linked each cleaning step to a business outcome – e.g., preserving sample size for niche product forecasting, avoiding bias in revenue totals, and ensuring compliance-ready data for audit.", "Your justification for choosing the capped dataset references real-world concerns such as preserving legitimate sales spikes while preventing them from dominating KPIs.", "The ethical section highlights reproducibility, bias amplification, and governance, showing awareness of how data-cleaning decisions affect stakeholders."], "learning_demonstration": ["The notebook demonstrates competence with core R packages (tid

Reflection & Critical Thinking

- Good engagement with reflection components of the assignment
- Shows developing critical thinking about analytical processes

Analytical Strengths

- Demonstrates solid analytical approach and technical execution.

Business Application

- Shows awareness of business context in analytical work
- Appropriate framing of data analysis objectives

Learning Demonstration

- Developing understanding of data analysis methodology

Areas for Development

- Part 5: reflection questions: They answered Q1 (missing value strategy), Q2 (outlier interpretation), Q3 (data quality impact), Q4 (ethical considerations). They gave thorough answers.

Recommendations for Future Work

- Continue practicing with diverse datasets and analytical scenarios
- Strengthen connections between technical analysis and business implications

Technical Analysis

Code Strengths:

- Successfully implemented comprehensive data quality assessment with clear documentation of issues
- Demonstrated strong use of tidyverse package for data manipulation and visualization
- Implemented multiple missing value treatment strategies including removal and imputation
- Created effective outlier detection using IQR method with proper capping implementation
- Built a complete data cleaning pipeline with before/after comparisons
- Used ggplot2 for visualization of outliers with appropriate styling

Code Improvement Suggestions:

- Fix the incomplete imputation of Customer_Name - the assignment is missing the actual assignment statement
- Complete the outlier identification for Quantity column - the assignment is missing the actual assignment statement
- Remove the redundant duplicate outlier capping code that appears twice in the submission
- Add proper error handling for the mode function to handle edge cases better
- Consider using dplyr::mutate() instead of direct column assignment for cleaner code

Technical Observations:

- Excellent code organization with clear section headers and logical flow
- Strong documentation of data quality issues with specific examples
- Professional approach to data cleaning with multiple strategies compared
- Good use of R's functional programming capabilities with lapply and mapply
- Demonstrates solid understanding of data analysis workflow from import to final dataset
- Code shows good attention to detail in handling edge cases like NA values

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