Homework Grading Report

Student Name:	Francisco Guadarrama
Assignment:	Assignment 1 - Introduction to R
Graded On:	October 03, 2025 at 10:37 PM
Final Score:	34.4 / 37.5 points (91.7%)

Score Summary

Overall Performance: Excellent (91.7%)

Instructor Assessment

Your work on this assignment demonstrates strong engagement and solid analytical skills. You successfully imported the messy dataset, performed a thorough initial assessment, and chose a well■justified outlier■capping strategy that balances statistical rigor with business relevance. Your reflection responses are thoughtful, especially the discussion of missing■value trade■offs and ethical considerations, showing you can connect technical decisions to real■world implications. To push your work to the next level, add quantitative impact analyses for each cleaning step and experiment with multiple imputation techniques. Overall, you are making excellent progress in mastering data cleaning and its business applications; keep challenging yourself with more complex datasets and advanced methods.

Reflection & Critical Thinking

Your answer to the question Which approach would you recommend for this dataset and why? is well■reasoned; you correctly identified that dropping rows with missing Sales_Amount preserves the integrity of revenue totals and explained why median imputation would be less appropriate for that key variable. By quantifying the loss (22 NAs out of 200, about 11%) you demonstrated a solid grasp of the trade■off between sample size and bias. In the trade■offs between removal and imputation discussion you highlighted the danger of bias when data are not MCAR and the tendency of single imputation to underestimate variance. Your explanation of how imputation can shrink variance and produce overly optimistic standard errors shows clear critical thinking about statistical consequences.

Analytical Strengths

Your R code correctly loads the tidyverse, reads the messy_sales CSV, and prints a concise overview of rows, columns, and column names, which demonstrates proper data■import hygiene. The use of `head()`, `str()`, and `summary()` provides a solid initial assessment of data structure and quality. The outlier handling section is well executed: you generated a boxplot for Sales_Amount, applied IQR■based winsorization, and then compared the capped dataset to the original. This shows you can move from visual diagnosis to a reproducible treatment method. Your final justification for selecting the capped dataset ties together statistical reasoning (preserving central tendency) with business considerations (maintaining spikes that reflect real promotions). This synthesis of technical and business perspectives is a strong indicator of analytical maturity.

Business Application

You linked the decision to keep outliers via capping to real■world scenarios such as promotional spikes, explaining how removing them could hide valuable revenue signals for KPI dashboards. This demonstrates an ability to translate data■cleaning choices into business impact. Your discussion of how missing high■value orders would bias average sales downward and how outliers inflate standard deviations directly connects data quality issues to downstream forecasting accuracy. This shows you understand the ripple effect of cleaning decisions on strategic planning.

Learning Demonstration

Your treatment of missing values—distinguishing between critical fields (Sales_Amount) and less critical ones (Quantity) and choosing different strategies for each—demonstrates a nuanced understanding of variable importance. This reflects solid progress in prioritizing data elements. The way you applied the IQR rule for outlier detection and then chose winsorization over outright deletion shows you have mastered a core data■cleaning technique and can justify it with both statistical and business arguments.

Areas for Development

While you explained the rationale for dropping rows with missing Sales_Amount, adding a quantitative comparison of key metrics (e.g., total revenue before and after removal) would strengthen your argument and make the impact more tangible. Consider exploring multiple imputation methods (e.g., mice) for variables like Quantity, and report the variance across imputations. This would address the limitation you noted about single imputation understating uncertainty. Your outlier visualization is a good start; you could enhance it by overlaying the capped values on the same plot or by providing a summary table of pre■ and post■capping statistics. This would give reviewers a clearer picture of how the treatment altered the distribution.

Recommendations for Future Work

Continue practicing robust statistical techniques such as trimmed means and robust regression on the capped dataset; these methods will further protect your models from residual outlier influence. Explore advanced imputation frameworks like predictive mean matching or Bayesian methods, and document the imputation model diagnostics. This will deepen your skill set for handling MAR scenarios. Build a reproducible data■cleaning pipeline using `drake` or `targets` so that each cleaning step (removal, imputation, capping) is versioned and can be rerun automatically. This will improve reproducibility and align with the governance principles you discussed.

Technical Analysis

Code Strengths:

- Proper implementation of R library loading and data import procedures
- Effective use of dplyr functions for data manipulation and filtering
- Appropriate application of ggplot2 for data visualization
- Systematic approach to data exploration and summary statistics

- Complete execution of all required analytical components

Code Improvement Suggestions:

- Consider using complete.cases() for more robust missing data handling

Example:

```
# Remove rows with missing values
clean_data <- sales_df[complete.cases(sales_df), ]
# Or check for missing values first
sum(is.na(sales_df))
```

- Explore the cut() function for creating categorical variables from continuous data

Example:

```
# Create categorical variables from continuous data
sales_df$amount_category <- cut(sales_df$amount,
breaks = c(0, 100, 500, 1000, Inf),
labels = c('Low', 'Medium', 'High', 'Very High'))
```

- Add correlation analysis using cor() to quantify relationships between variables

Example:

```
# Calculate correlation between numeric variables
cor(sales_df$amount, sales_df$rating, use = 'complete.obs')
# Or correlation matrix
cor(sales_df[, c('amount', 'rating', 'quantity')])
```

- Include additional summary statistics such as standard deviation and quartiles

Example:

```
# Additional summary statistics
sd(sales_df$amount, na.rm = TRUE) # Standard deviation
quantile(sales_df$amount, na.rm = TRUE) # Quartiles
IQR(sales_df$amount, na.rm = TRUE) # Interquartile range
```

Technical Observations:

- Demonstrates solid understanding of fundamental R programming concepts
- Code structure follows logical analytical workflow
- Shows appropriate selection of analytical tools for the business context

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