```r
# Load necessary libraries
library(tidyverse)
library(naniar)  # For missing value visualization
library(mice)    # For imputation

# Read the data
retail_data <- read.csv("F:\\UoB Study\\Marketing Analysis & Behaviour Science\\Assignment
1\\SmartFresh Retail.csv")
str(retail_data)


# Check for missing values
missing_summary <- sum(is.na(retail_data))
print(paste("Total NA values:", missing_summary))

# Prepare data for imputation
# Exclude ID and categorical variables that shouldn't be used for imputation
imputation_vars <- setdiff(names(retail_data),
                           c("Customer_ID", "Dt_Customer", "Last_Interaction"))

# Create imputation model
imputation_model <- mice(retail_data[, imputation_vars], m = 5, method = "pmm", seed =
123)

# Generate imputed dataset
imputed_data <- complete(imputation_model)

# Combine imputed data with original data
retail_data_clean <- retail_data
retail_data_clean[, imputation_vars] <- imputed_data

# Alternatively, for Annual_Income, you could use median by education level
income_by_education <- retail_data %>%
  group_by(Education_Level) %>%
  summarize(median_income = median(Annual_Income, na.rm = TRUE))

# Apply this imputation
for (i in 1:nrow(retail_data_clean)) {
  if (is.na(retail_data_clean$Annual_Income[i])) {
    edu <- retail_data_clean$Education_Level[i]
    retail_data_clean$Annual_Income[i] <- income_by_education$median_income[
      income_by_education$Education_Level == edu]
  }
}

# Identify outliers in spending columns
spending_cols <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
                   "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")

# Function to cap outliers using IQR method
cap_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  upper_bound <- q3 + 1.5 * iqr
  x[x > upper_bound] <- upper_bound
  return(x)
}

# Apply outlier capping
retail_data_clean[spending_cols] <- lapply(retail_data_clean[spending_cols], cap_outliers)

write.csv(retail_data_clean, "F:\\UoB Study\\Marketing Analysis & Behaviour Science
          \\Assignment 1\\SmartFresh-Retail-Clean.csv", row.names = FALSE)
```