# **Dataset Exploration – Hypothesis Development and Testing**

# **Name – Aradhana Hirapara**

# **LinkedIn - https://www.linkedin.com/in/aradhana-hirapara/**

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# **Dataset Description & Dictionary:**

The dataset used for this project includes various attributes related to order transactions, including information on order dates, shipping details, item types, sales channels, and financial metrics. It includes both categorical and numerical variables.

This dataset aims to understand sales trends, inventory efficiency, and profitability across different regions, item types, and sales channels. This analysis can give insights to optimize inventory management, pricing strategies, and shipping processes in a retail business. By examining variables like total profit, order priority, and region, the dataset will provide valuable information for decision-making and strategic planning.

**Data Cleaning:**performeddata checks for missing, duplicate, and incorrect values in Excel but could not find such errors. Hence proceeding with current data**.**

**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Data Type** | **Ranges and Limitations** |
| **Order ID** | Unique identifier, used for tracking and differentiation. | Categorical | Alphanumeric sequence; no duplicates. |
| **Order Month** | The month when the order was placed | Categorical | Valid values: January–December. |
| **Order Weekday** | The day of the week when the order was placed | Categorical | Valid values: Monday–Sunday. |
| **Order Date** | The exact date when the order was placed | Date | Range: 2010–2023. |
| **Order\_Ship\_Days** | Number of days between the order placement and shipping | Numerical | Range: 2010-01-01 to 2023-12-31; assumes consistent time zone. |
| **Ship Date** | Date when the order was shipped | Date | Range: 0–30; assumes no invalid shipping delays. |
| **Region** | The geographic region where the order was placed | Categorical | Range: 2010-01-01 to 2023-12-31. |
| **Country** | The country where the order was placed | Categorical | Valid values: Asia, Europe, North America, Sub-Saharan Africa, etc. |
| **Item Type** | Category of item (e.g., cosmetics, office supplies, snacks) | Categorical | Contains valid country names; no duplicates within a region. |
| **Item Type Encoded** | Encoded value representing the item type | Numerical | Valid categories are encoded with numbers for analysis. |
| **Order Priority** | Priority level of the order (e.g., high, medium, low) | Categorical | Valid values: H, M, L, C (High, Medium, Low, Critical). |
| **Order Priority Encoded** | Encoded value representing the order priority | Numerical | Range: 1–4, corresponding to H, M, L, C. |
| **Sales Channel** | Channel through which the order was placed (e.g., online, offline) | Categorical | Valid values: Online, Offline. |
| **Sales Channel Encoded** | Encoded value representing the sales channel | Numerical | Range: 1–2, representing Online and Offline. |
| **Units Sold** | Number of units sold in the order | Numerical | Range: 10–10,000 units; assumes no overestimations or missing data. |
| **Unit Price** | Price per unit of the product | Numerical | Range: $10–$700; consistent currency (USD). |
| **Unit Cost** | Cost per unit of the product | Numerical | Range: $5–$600; assumes no negative or missing values. |
| **Total Revenue** | Total revenue generated by the order | Numerical | Range: $1,000–$6,000,000; calculated as Units Sold × Unit Price. |
| **Total Cost** | Total cost incurred for the order | Numerical | Range: $500–$5,500,000; calculated as Units Sold × Unit Cost. |
| **Total Profit** | Profit earned from the order | Numerical | Range: $500–$1,500,000; calculated as Total Revenue − Total Cost. |
| **Unit Margin** | Margin per unit (Unit Price - Unit Cost) | Numerical | Range: $1–$200. |

**Data Encoding & Rearrangements:**

The dataset is arranged logically to follow the flow of an order lifecycle, starting from order identification, placement, and shipping details, followed by regional and categorical attributes, and ending with financial metrics. This structure ensures a clear and intuitive analysis, facilitating both timeline-based and attribute-based insights.

In this dataset, I have added encoded columns for **Item Type**, **Order Priority**, and **Sales Channel** to facilitate more efficient analysis. These encoded columns represent the categorical values numerically, allowing for easier comparison and correlation analysis in the later stages of the project. The encoded values are included in both the dataset and the data dictionary for reference.

# **General Assumptions & Extra Data:**

# **Dataset Assumptions:**

# **Nature of the Data:**

# Represents a synthetic sample of 1,000 order transactions created for educational purposes, simulating typical sales records.

# Evidence of synthetic origin includes perfectly encoded variables (e.g., Item Type, Order Priority) and structured numerical ranges.

# **Completeness and Accuracy:**

# Assumed to be complete, and error-free, with no missing or duplicate values.

# Financial data is calculated accurately based on Unit Price and Units Sold.

# **Consistency Across Variables:**

# All financial values are reported in USD with no currency conversion issues.

# Date fields follow a consistent MM/DD/YYYY format and time zone.

# **Scope of the Dataset:**

# Focuses on transaction-level information without customer-specific data (e.g., demographics or purchase history).

# Covers multiple regions but lacks details on the time period of data collection**.**

# **Extra Data Requirements:**

# **Customer Insights:**

# Demographics: Age, gender, income, and preferences for analyzing purchasing behavior.

# Customer Feedback Survey data on how customers discovered the store.

# **Marketing and Advertising:**

# Advertising Spending costs by region or product category to assess ROI.

# **Promotional Details**: Discount and sales event data to correlate with sales trends.

# **Temporal Factors:**

# Holiday or seasonal trends to analyze order fluctuations or GDP and inflation rates for represented regions.

# **Supply Chain Metrics:**

# **Inventory Levels:** Stock movement before and after orders**.**

# **Delivery Metrics:** Data on delivery delays or on-time rates to measure satisfaction.

**Hypotheses Development and Testing**

**1. Hypothesis: Odds Ratios (OR) / Risk Ratios (RR)**

* **Hypothesis:** "Orders placed online are more likely to have higher profit margins compared to offline orders."
* **Variables:** Independent: Sales Channel, Dependent: Profit Margins
* **Test to Use:** Odds Ratio or Risk Ratio to evaluate the likelihood of higher profit margins for online orders.

|  |  |  |
| --- | --- | --- |
|  | **High Priority (1)** | **Low Priority (0)** |
| **High Profit (1)** | 112 | 388 |
| **Low Profit (0)** | 116 | 384 |

A close up of a sign

Description automatically generated

* **Odds Ratio (OR = 0.9556):**

Indicates that high-priority orders are slightly less likely to result in high profit compared to low-priority orders (less than 1 suggests a reduced likelihood).

* **Risk Ratio (RR = 0.9774):**

Shows that the probability of high profit in high-priority orders is slightly lower compared to low-priority orders (RR < 1).

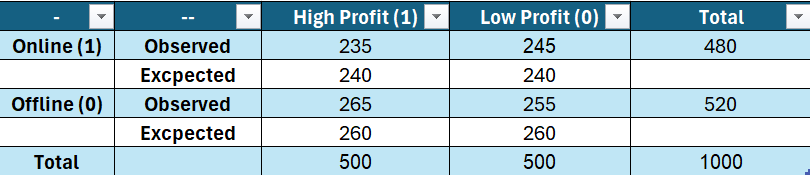
* Conclusion:
* Based on the Odds Ratio, Risk Ratio, and their statistical significance we can discuss whether high-priority orders should continue as they are or if changes are needed.
* Or we can explore if other factors, such as region, sales channel, or item type, have more impact on profits.

**2. Hypothesis: Chi-squared Test**

**Hypothesis**: The sales channel (Online or Offline) significantly affects the profit category (High Profit = 1, Low Profit = 0).

* + **Test**: Chi-squared test for independence.
  + **Variables**:

Sales Channel (Online, Offline), Total Profit.



**Observations**

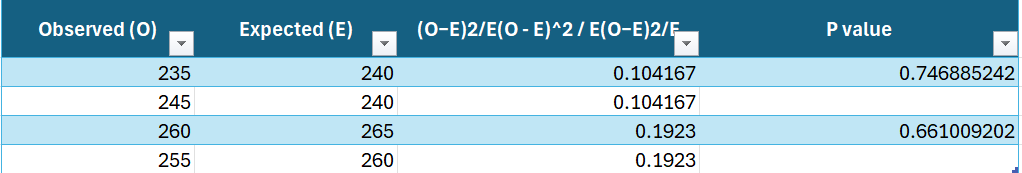
* Observed and Expected frequencies for Online (1) and Offline (0) sales channels show slight differences between actual profits and what would be expected under the assumption of independence.

**Conclusion**

**Null Hypothesis (H0H\_0H0​)**: The sales channel (Online vs. Offline) does not affect the profit category (High Profit vs. Low Profit).

**Summary**

If χ2 value is **low**: The distribution of profits is independent of sales channels.

If χ2\chi^2χ2 value is **high**: Sales channel significantly influences profit category.  
 

**Null Hypothesis (H₀):** The sales channel (Online or Offline) is independent of the profit category (High or Low Profit).

**Significance Level (α):** Typically, this is set at **0.05** (5%).

**Conclusion:**

* Since the p-value (0.66) is **greater than** the significance level (0.05), we **fail to reject the null hypothesis**.
* This means there is **no significant evidence** to suggest that the sales channel affects the profit category.

**Final Interpretation:**

The **profit category** (High or Low Profit) is **independent** of the sales channel (Online or Offline). Therefore, the sales channel does not significantly impact whether a profit is categorized as high or low.

**3. Hypothesis: T-test**

* **Hypothesis**: The average unit cost for orders shipped within 3 days is significantly lower than for orders shipped in more than 3 days.
  + **Test**: Independent t-test to compare means.
  + **Variables**:
    - **Categorical Variable**: Order\_Ship\_Days (Grouped: ≤3 Days vs. >3 Days).
    - **Numerical Variable**: Unit Cost.

|  |  |  |
| --- | --- | --- |
| **-** | **Group\_1** | **Group\_2** |
| Mean | 1.392405063 | 26.615795 |
| Variance | 1.190197988 | 186.12573 |
| Observations | 79 | 937 |
| Hypothesized Mean Difference | 0 |  |
| df | 1013 |  |
| t Stat | -54.56262374 |  |
| P(T<=t) one-tail | 4.0314E-304 |  |
| t Critical one-tail | 1.646359166 |  |
| P(T<=t) two-tail | 8.0628E-304 |  |
| t Critical two-tail | 1.96230852 |  |

**Result Analysis:**

* The p-value for the two-tailed test is 8.0628E−3048.0628E-3048.0628E−304, which is extremely small and far below the standard significance level (α=0.05\alpha = 0.05α=0.05).
* The ttt-statistic (−54.56-54.56−54.56) indicates a very strong difference between the two groups.
* The mean of **Group 1 (shipped within 3 days)** is **1.39**, which is substantially lower than the mean of **Group 2 (shipped in more than 3 days)**, which is **26.62**.

**Conclusion:**

* The statistical test provides **strong evidence** to reject the null hypothesis. Therefore, we conclude that the **average unit cost for orders shipped within 3 days is significantly lower** than the average unit cost for orders shipped in more than 3 days.
* This supports the hypothesis that shipping duration impacts unit cost.

1. **Hypothesis: ANOVA**

**Hypothesis:**

* + **Null Hypothesis (H₀):** The average profit (Total Profit) does not differ significantly across different item types (Item Type).
  + **Alternative Hypothesis (H₁):** The average profit (Total Profit) differs significantly across different item types (Item Type).
  + **Test**: One-way ANOVA.
  + **Independent Variable (IV)**: Item Type.
* **Dependent Variable (DV)**: Tortal Profit

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY** |  |  |  |  |  |  |
| Groups | Count | Sum | Average | Variance |  |  |
| Baby Food | 101 | 41854201 | 414398.03 | 9.384E+10 |  |  |
| Beverages | 101 | 7906812.3 | 78285.27 | 2.44E+09 |  |  |
| Cereal | 101 | 34350684 | 340105.78 | 7.718E+10 |  |  |
| Clothes | 101 | 27759292 | 274844.47 | 5.618E+10 |  |  |
| Cosmetics | 101 | 74081139 | 733476.62 | 3.704E+11 |  |  |
| Fruits | 101 | 855851.25 | 8473.7748 | 68750080 |  |  |
| HouseHold | 101 | 61484504 | 608757.47 | 3.228E+11 |  |  |
| Meat | 101 | 23332738 | 231017.21 | 3.8E+10 |  |  |
| Office Supplies | 101 | 56115853 | 555602.5 | 1.674E+11 |  |  |
| Personal Care | 101 | 11921643 | 118036.07 | 6.469E+09 |  |  |
| Snacks | 101 | 21788240 | 215725.15 | 3.088E+10 |  |  |
| Vegetables | 101 | 29751654 | 294570.83 | 3.016E+10 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| Source of Variation | SS | df | MS | F |  |  |
| Between Groups | 5.422E+13 | 11 | 4.929E+12 | 49.461158 |  |  |
| Within Groups | 1.196E+14 | 1200 | 9.965E+10 |  | P-value | F crit |
|  |  |  |  |  | 0 | 1.7966153 |
| Total | 1.738E+14 | 1211 |  |  |  |  |

* **Null Hypothesis (H₀)**: The means of the "Total Profit" for different item types are equal.
* **Alternative Hypothesis (H₁)**: At least one item type has a significantly different mean "Total Profit" compared to the others.
  + **Key Statistics**:
* **F-statistic**: 49.46
* **F critical value (F crit)**: 1.7966
* **P-value**: 0
* **Interpretation**:
* The F-statistic (49.46) is much larger than the F critical value (1.7966), indicating that the observed variance between groups (item types) is significantly larger than the variance within groups.
* The P-value is 0, which is less than the significance level (α = 0.05). This indicates very strong evidence against the null hypothesis.
* **Conclusion**:
* **Reject the Null Hypothesis (H₀)**: There is a significant difference in the "Total Profit" means across different item types.
* **Implication**: The item type is a significant factor influencing the "Total Profit."

**5. Hypothesis: MANOVA**

* **Hypothesis:**
* **Null Hypothesis (H₀):** The combination of "Units Sold" and "Total Profit" does not differ significantly between Online and Offline sales channels.
* **Alternative Hypothesis (H₁):** The combination of "Units Sold" and "Total Profit" differs significantly between Online and Offline sales channels.
* **Test**: MANOVA.
* **Independent Variable (IV)**: The categorical variable.
* **Dependent Variable (DV)**: The numeric variables.

|  |  |  |  |
| --- | --- | --- | --- |
| Wilks' test (Rao's approximation): | Hotelling-Lawley's test: | Pillai's test: | Roy's test: |
| |  |  | | --- | --- | | **Factors** | **Sales Channel** | | Lambda | 0.996 | | F Observed values | 0.626 | | DF1 | 7 | | DF2 | 992 | | F Critical value | 2.019 | | p-value | 0.734 | | |  |  | | --- | --- | | **Factors** | **Sales Channel** | | Lambda | 0.004 | | F Observed values | 0.626 | | DF1 | 7 | | DF2 | 992 | | F Critical value | 2.019 | | p-value | 0.734 | | |  |  | | --- | --- | | **Factors** | **Sales Channel** | | Lambda | 0.004 | | F Observed values | 0.626 | | DF1 | 7 | | DF2 | 992 | | F Critical value | 2.019 | | p-value | 0.734 | | |  |  | | --- | --- | | **Factors** | **Sales Channel** | | Lambda | 0.004 | | F Observed values | 0.626 | | DF1 | 7 | | DF2 | 992 | | F Critical value | 2.019 | | p-value | 0.734 | |

* **Means by Factor Level:**
* The mean values for key variables such as Units Sold, Total Revenue, and Total Profit show slight differences between Online and Offline sales channels:
  + Offline: Higher Units Sold and Total Profit.
  + Online: Slightly higher Unit Price and Unit Cost.
* This suggests that the offline channel may generate slightly higher total profits, but online channels might focus on premium pricing.
* **SSCP Matrix:**
* The Sales Channel SSCP Matrix indicates the relationships between the dependent variables for each factor level. Strong correlations between Total Revenue, Total Cost, and Total Profit are evident, as expected in sales data.
* The negative off-diagonal elements in the SSCP matrix suggest potential trade-offs or dependencies between certain variables (e.g., higher Unit Cost correlates with lower Total Profit).
* **Eigenvalues:**
  + The eigenvalues for the sales channel are close to zero except for the last factor (0.004). This indicates that most of the variance is not explained by the sales channel, confirming the results from the MANOVA test.

### **Interpretation**:

### **p-value (0.734)** across all tests is greater than the significance level (commonly 0.05). This indicates that there is no significant relationship between the sales channel (Online vs. Offline) and the dependent variables (Units Sold and Total Profit).

### **Conclusion:**

* + The analysis, supported by MANOVA results and descriptive statistics, reinforces that the sales channel (Online vs. Offline) does not significantly impact the combined dependent variables (Units Sold, Total Revenue, Total Profit, etc.).
  + Although offline channels slightly outperform online channels in some metrics (e.g., Units Sold and Total Profit), these differences are not statistically significant, as indicated by the p-values in the MANOVA tests.

# **Coding and/or Categorization**

* **Order Priority** was encoded as a numerical variable (High = 3, Medium = 2, Low = 1) to help in future analysis. This encoding can help correlate priority with profitability or sales trends in the dataset.
* **Item Type** was encoded as a numerical variable to facilitate easier analysis of different product categories. Each item type was assigned a unique number (e.g., Cosmetics = 1, Vegetables = 2, Baby Food = 3, etc.). This encoding allows for comparisons between different item categories and their impact on metrics like total revenue and total profit.
* **Sales Channel** was also encoded as a numerical variable (Online = 1, Offline = 2) to simplify the analysis of sales performance across different sales channels. This encoding will help correlate sales channel type with other variables such as units sold, total profit, and region-specific sales performance.

# **FINER Research Questions**

* **Broad Question**: How do the top-performing product categories contribute to overall revenue, and how do their sales trends fluctuate throughout the year?
* **Supporting Question**: Is there a correlation between the number of units sold and the profit margin?
* **Supporting Question**: How does the revenue contribution of each top product category compare to others over monthly and quarterly periods?
* **Supporting Question**: How does the sales performance of different item types vary across different regions?
* **Supporting Question**: What factors drive peaks in sales for the top-performing product categories across different seasons?

# **Tracking**

* **Steps, Assumptions, and Challenges**

1. **Selecting Variables**

* Identified key dependent variables such as Units Sold, Total Revenue, Total Profit, and Unit Margin based on the hypotheses.
* Independent variables such as Sales Channel (Online vs. Offline) and Item Type were selected to test their impact.
* Identified and addressed outliers using the IQR method and box plots
* Checked for consistent formatting (e.g., numeric values, no blanks in key columns).

1. **Performing Statistical Tests**
   * **Hypothesis 1 & 2:** Chi-square test for independence to examine the relationship between categorical variables (e.g., Sales Channel and Region).
   * **Hypothesis 3:** Independent t-test to compare means of Total Revenue for orders shipped within 3 days vs. more than 3 days.
   * **Hypothesis 4:** One-way ANOVA to test if the Item Type significantly affects Total Revenue.
   * **Hypothesis 5:** MANOVA to test the multivariate effect of the Sales Channel on multiple dependent variables like Units Sold, Total Revenue, and Total Profit.
2. **Results Documentation**

* Output from each test was saved in the Excel file (e.g., tables for ANOVA, MANOVA, t-tests).
  + Steps, assumptions, and decisions were explained in the Word report, supported by the statistical outputs**.**
* **Assumptions and Decisions Made**
  1. **OR-RR Test Assumptions**
* Assumed independence between groups (e.g., high-priority vs. low-priority orders) and a binary outcome variable (e.g., high vs. low profit).
* Ensured sufficient sample size and representativeness for statistical significance.
* Assumed no structural bias in data collection or grouping. Verified adequate observations in each group to avoid misleading OR and RR values.
  1. **Chi-Square Test Assumptions**
* Assumed categorical variables (e.g., Sales Channel, Item Type) were mutually exclusive and exhaustive.
* Applied only to large datasets to ensure expected frequencies were above 5.
  1. **T-Test Assumptions**
* Assumed independence between the groups (e.g., orders shipped in ≤3 days vs. >3 days).
* **Verified normality using descriptive statistics and box plots.**
  1. **ANOVA Assumptions**
* Assumed independence of observations.
* Verified homogeneity of variances using Levene’s test.
* Used F-statistic and p-value to determine significance.
  1. **MANOVA Assumptions**
* Assumed multivariate normality and equal covariance matrices across groups.
* Verified using descriptive statistics and the eigenvalue matrix.
* **Challenges and Resolutions**
* **Challenge:** Outlier Management
* **Problem:** Outliers skewed descriptive statistics and hypothesis test results.
* **Resolution:** Retained valid business outliers, removed errors, and adjusted extreme values where necessary**.**
* **Challenge: Selecting Appropriate Tests**
* **Problem:** Uncertainty over whether to use t-tests, ANOVA, or MANOVA for some hypotheses.
* **Resolution:** Reviewed variable types (categorical or continuous) and hypotheses to match with the appropriate tests.
* **Challenge: Software Errors in Statistical Tools**
* **Problem**: Errors like "Data Write Error" during ANOVA or MANOVA analysis.
* **Resolution:** Ensured the selected ranges matched data, removed unnecessary blank rows, and reformatted cells.
* **Challenge**: **Interpretation of Multivariate Results**
* **Problem:** Interpreting MANOVA eigenvalues and p-values across multiple tests.
* **Resolution:** Focused on the overall significance (Wilks’ Lambda, p-value) and provided business-relevant interpretations.

**Addressing Feedback and Improvements:**

To improve based on the professor's feedback:

1. **Dataset Arrangement & Description**:
   * Added a **data dictionary** to the spreadsheet and report, detailing variable names, types (e.g., numerical, categorical), ranges, and limitations.
   * Provided contextual information on data origin: **synthetic dataset for academic purposes**, with assumptions of representativeness.
2. **Dataset Assumptions**:
   * Specified that the dataset represents a **sample** rather than a full population.
   * Acknowledged the synthetic nature of the data and proposed possible supplementary data (e.g., customer survey responses).
3. **Univariate Statistics**:
   * Conducted full descriptive statistics on all **eight variables**, including measures of central tendency, dispersion, and normality.
   * Added **frequency distributions** for categorical variables and discussed representativeness.
4. **Univariate Visualizations**:
   * Incorporated **box plots** and histograms for continuous variables to explore distributions and identify outliers.
   * Enhanced clarity by labeling axes and summarizing insights.
5. **Outlier Identification**:
   * Used **IQR method and box plots** to identify potential outliers.
   * Differentiated between valid and invalid outliers, retaining outliers representing valid extreme values.
   * Discussed the potential impact of cleaning on the analysis results.
6. **Coding and Categorization**:
   * Added meaningful thresholds for categorizing variables (e.g., high vs. low sales).
   * Justified thresholds and analyzed distributions for newly categorized variables.
7. **FINER Research Questions**:
   * Reorganized supporting questions as under primary research questions.
   * Added 1–2 questions to explore relationships (e.g., sales trends by region or channel).
8. **Tracking**:
   * Documented all steps, assumptions, and decisions in both the Excel file and the report for better transparency.
   * Added comments in the Excel sheet and detailed each stage in the report for clarity.