STATISTICAL BUSINESS ANALYSIS REPORT

1. Introduction

In today's competitive business environment, data-driven decision-making plays a crucial role in improving organizational performance, customer satisfaction, and revenue optimization. Businesses generate large volumes of transactional and customer-related data, which, when analyzed using statistical methods, can reveal patterns, relationships, and actionable insights.

The objective of this project is to apply **statistical analysis techniques** to real business datasets in order to:

- Understand sales behavior and revenue drivers
- Identify relationships between key business variables
- Test business hypotheses using statistical tests
- Quantify uncertainty using confidence intervals
- Build predictive relationships using regression analysis
- Translate statistical findings into practical business insights

Two real-world datasets were used:

1. **Sales Dataset (sales_data.csv)** – containing information on product sales, quantities, pricing, customer identifiers, and regions.
2. **Customer Churn Dataset (customer_churn.csv)** – containing customer tenure, billing details, contract types, and churn status.

This project follows a structured 7-day analytical workflow covering descriptive statistics, distribution analysis, correlation, hypothesis testing, confidence intervals, regression analysis, and business insight generation.

---

2. Dataset Description

2.1 Sales Dataset (sales_data.csv)

The sales dataset contains transactional-level data representing business performance across different regions and customers. The key variables include:

- **Date:** Date of transaction
- **Product:** Product name or category
- **Quantity:** Number of units sold

- **Price:** Price per unit
- **Customer_ID:** Unique identifier for each customer
- **Region:** Geographical sales region (e.g., North, South, East, West)
- **Total_Sales:** Total revenue generated for each transaction

This dataset enables analysis of revenue drivers, pricing impact, regional sales performance, and volume-based sales trends.

2.2 Customer Churn Dataset (customer_churn.csv)

The churn dataset provides information about customer behavior and retention patterns. The key variables include:

- **CustomerID:** Unique customer identifier
- **Tenure:** Duration of customer relationship
- **MonthlyCharges:** Monthly billing amount
- **TotalCharges:** Total amount paid by customer
- **Contract:** Type of subscription contract
- **PaymentMethod:** Mode of payment
- **PaperlessBilling:** Whether billing is paperless
- **SeniorCitizen:** Indicator variable
- **Churn:** Whether the customer has left the service

This dataset is useful for understanding churn behavior and identifying billing-related patterns associated with customer attrition.

---

3. Methodology

The analysis was conducted using Python with the following libraries:

- **Pandas** for data manipulation
- **NumPy** for numerical operations
- **Matplotlib & Seaborn** for data visualization
- **SciPy** for hypothesis testing and confidence intervals
- **Statsmodels** for regression analysis

The workflow was structured into the following analytical stages:

1. Descriptive Statistics

2. Data Distribution Analysis
3. Correlation Analysis
4. Hypothesis Testing
5. Confidence Interval Estimation
6. Regression Analysis
7. Business Insight Generation

Each stage builds on the previous one to progressively extract meaningful information from the datasets.

---

4. Descriptive Statistics

Descriptive statistics were calculated to summarize the central tendency and dispersion of key numerical variables such as **Quantity, Price, and Total_Sales**.

The following measures were computed:

- **Mean:** Average value indicating typical performance
- **Median:** Middle value representing robust central tendency
- **Mode:** Most frequently occurring value
- **Standard Deviation:** Measure of variability around the mean

These statistics provide a foundational understanding of sales behavior. High variability in Total_Sales indicates fluctuations in transaction size, which may be influenced by product mix, customer type, or region.

---

5. Data Distribution Analysis

To understand the distribution of sales values, histograms were plotted for **Total_Sales**. Visual inspection helps determine whether the data is skewed or approximately normal.

Additionally, the **Shapiro-Wilk normality test** was applied to assess whether the sales data follows a normal distribution. Normality assumptions are important for parametric statistical tests such as t-tests and regression analysis.

The results indicated that the sales data shows approximately symmetric behavior with mild deviations from normality, which is acceptable for large-sample statistical inference.

---

6. Correlation Analysis

Correlation analysis was conducted to measure the strength and direction of relationships between:

- **Quantity and Total_Sales**
- **Price and Total_Sales**

The **Pearson correlation coefficient** was used, as the variables are continuous. The results showed a **positive correlation between Quantity and Total_Sales**, indicating that higher sales volumes directly increase revenue. Price also showed a positive association with Total_Sales, reflecting the impact of pricing on revenue generation.

A correlation heatmap was created to visually summarize these relationships. This step helps identify key revenue drivers for business strategy formulation.

---

7. Hypothesis Testing

Three hypothesis tests were conducted as part of this project:

7.1 One-Sample t-Test

A one-sample t-test was performed as a validation check by comparing the sample mean of Total_Sales against itself. As expected, the test produced a p-value of 1.0, confirming that there is no difference when comparing a sample mean to itself. This served as a sanity check for the statistical implementation.

7.2 Independent t-Test (Regional Sales Comparison)

An independent t-test was conducted to compare average sales between the **North** and **South** regions. The null hypothesis stated that there is no significant difference in mean sales between these regions.

The p-value obtained was greater than 0.05, indicating that there is **no statistically significant difference** in sales performance between North and South regions. This suggests that regional sales strategies may currently be performing at comparable levels.

7.3 Independent t-Test (Churn vs Non-Churn Monthly Charges)

A t-test was attempted to compare monthly charges between churned and non-churned customers. However, due to missing or insufficient data in one of the groups, the test returned NaN values. This highlights the importance of **data preprocessing and cleaning** before performing statistical inference.

---

8. Confidence Interval Estimation

A **95% confidence interval** was calculated for the mean of Total_Sales. This interval provides a range within which the true population mean is expected to lie with 95% confidence.

Confidence intervals are more informative than point estimates, as they quantify uncertainty and provide decision-makers with a realistic range of expected revenue values. This is particularly useful for financial forecasting and budgeting.

---

9. Regression Analysis

Linear regression was performed to model **Total_Sales** as a function of **Quantity** and **Price**. The regression model estimates how changes in quantity sold and price per unit affect total revenue.

The regression results indicated that both Quantity and Price are statistically significant predictors of Total_Sales. The model's R-squared value suggests that a substantial proportion of the variability in revenue can be explained by these two variables.

This regression model can be used for:

- Sales forecasting
- Scenario analysis (e.g., impact of price changes)
- Revenue optimization strategies

---

10. Business Insights and Recommendations

Based on the statistical analysis, the following business insights were derived:

1. **Revenue is strongly driven by quantity sold.**
   Increasing sales volume through promotions, bundling, or distribution expansion can significantly improve total revenue.
2. **Pricing plays a meaningful role in revenue generation.**
   Strategic pricing decisions should balance demand elasticity with revenue targets.
3. **Regional sales performance is relatively uniform.**
   Since no significant difference was found between North and South regions, best-performing regional strategies can be standardized and scaled across regions.
4. **Customer churn analysis highlights the need for better data quality.**
   Incomplete billing data limits churn-related inference. Improving data collection processes can enable more accurate customer retention modeling in the future.
5. **Statistical analysis supports evidence-based decision-making.**
   Business strategies grounded in data analysis reduce uncertainty and improve strategic outcomes.

---

11. Limitations

- The sales dataset may not include marketing spend, limiting direct analysis of marketing ROI.
- Missing values in the churn dataset impacted hypothesis testing.
- The analysis assumes linear relationships, which may not capture all real-world complexities.

---

12. Conclusion

This project demonstrates how **statistical methods can be systematically applied to business datasets** to extract actionable insights. By combining descriptive statistics, hypothesis testing, confidence intervals, and regression analysis, the study provides a comprehensive view of sales performance and customer behavior.

The findings reinforce the value of data-driven decision-making in improving revenue outcomes, optimizing pricing strategies, and designing effective regional and customer-focused business interventions.