1. Project Overview

Project Title: Customer Churn and Sales Data Analysis

Project Goals and Objectives:
The goal of this project is to perform a comprehensive statistical analysis on customer churn and sales data. The objectives include:

- Understanding customer behavior and identifying factors contributing to churn.
- Analyzing sales performance across different products and regions.
- Applying statistical methods, hypothesis testing, and regression analysis to uncover insights.
- Providing actionable business recommendations based on data analysis.

Datasets Used:

1. customer_churn.csv – Contains information about customer tenure, monthly charges, total charges, contract type, payment method, and churn status.
2. sales_data.csv – Contains sales transactions including date, product, quantity, price, customer ID, region, and total sales.

2. Setup Instructions

System Requirements:

- Python 3.10+
- Minimum 8 GB RAM recommended

Step-by-Step Installation and Configuration:

1. **Install Python** – Download from python.org and install.
2. **Set up a virtual environment (optional but recommended):**
3. python -m venv venv
4. source venv/bin/activate # Linux/Mac
5. venv\Scripts\activate # Windows
6. **Install required packages:**
   Save the following in requirements.txt and run:
7. pip install -r requirements.txt

requirements.txt includes:

pandas

numpy

matplotlib

seaborn

scipy

scikit-learn

8. **Place CSV files** (customer_churn.csv and sales_data.csv) in the same project directory.
9. **Run the analysis script:**
10. python analysis.py

This will generate outputs, plots, regression results, and hypothesis_tests_results.txt.

3. Code Structure

Project File Hierarchy:

Week7/

├── customer_churn.csv

├── sales_data.csv

├── analysis.py # Main Python script

├── requirements.txt

└── hypothesis_tests_results.txt

Code Organization:

- **Day 1:** Descriptive statistics (mean, median, mode, standard deviation)
- **Day 2:** Data distribution analysis (histograms, normality test)
- **Day 3:** Correlation analysis (Pearson correlation and heatmap)
- **Day 4:** Hypothesis testing (ANOVA, t-test, Chi-Square)
- **Day 5:** Confidence intervals (95% CI calculation)
- **Day 6:** Regression analysis (predicting TotalCharges)

- **Day 7:** Business insights and actionable recommendations

4. Visual Documentation

Examples of Visual Outputs:

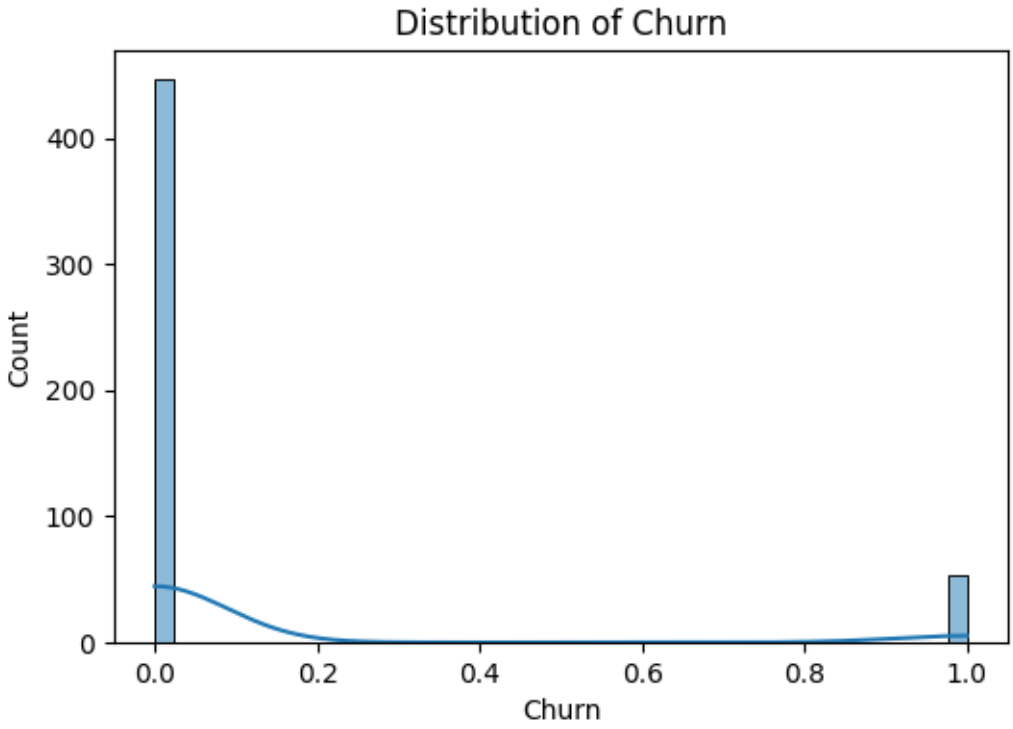
Descriptive statistics.

```
### Day 1: Descriptive Statistics ###

Customer Churn Statistics:
Mean:
 Tenure              36.532
MonthlyCharges     113.636
TotalCharges      4237.882
SeniorCitizen        0.498
Churn                0.106
dtype: float64

Median:
 Tenure              37.0
MonthlyCharges     115.0
TotalCharges      4182.5
SeniorCitizen        0.0
Churn                0.0
dtype: float64

Mode:
 CustomerID              C00001
Tenure                     3.0
MonthlyCharges           115.0
TotalCharges            4023.0
Contract              One year
PaymentMethod      Credit Card
PaperlessBilling            No
SeniorCitizen              0.0
Churn                      0.0
Name: 0, dtype: object
```

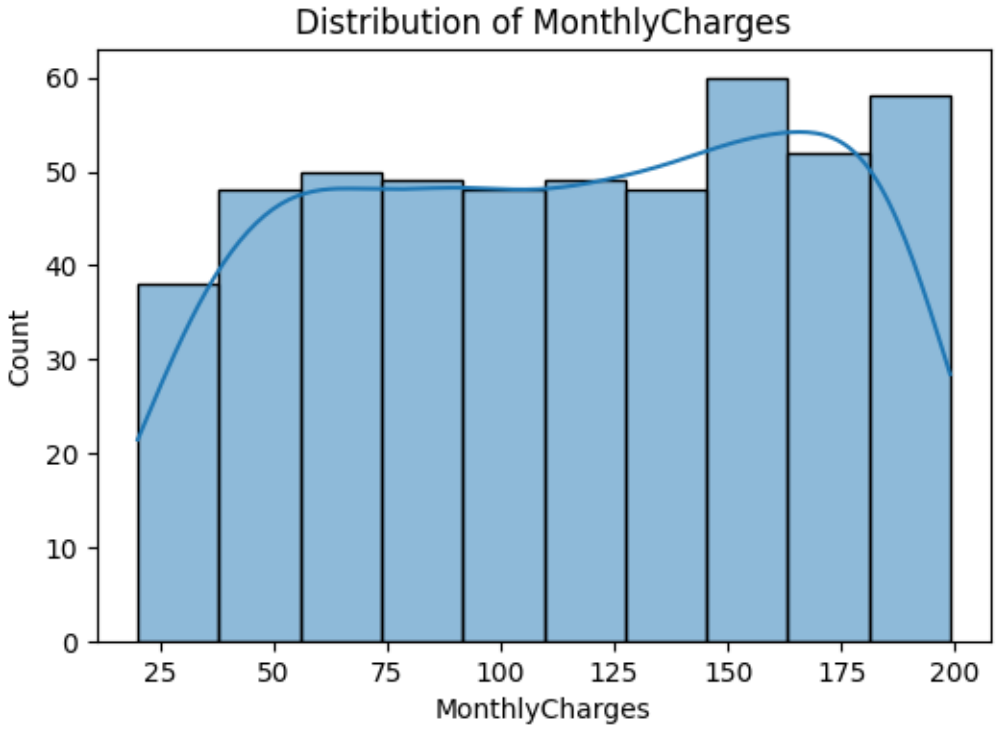
```
Standard Deviation:
 Tenure              20.667057
MonthlyCharges      51.799903
TotalCharges      2260.619837
SeniorCitizen        0.500497
Churn                0.308146
dtype: float64

Sales Data Statistics:
Mean:
 Quantity            4.78
Price           25808.51
Total_Sales    123650.48
dtype: float64

Median:
 Quantity           5.0
Price          24192.0
Total_Sales    97955.5
dtype: float64

Mode:
 Date           2024-01-01
Product             Tablet
Quantity               4.0
Price                 1308
Customer_ID        CUST001
Region               North
Total_Sales           6540
Name: 0, dtype: object


Standard Deviation:
 Quantity             2.588163
Price             13917.630242
Total_Sales     100161.085275
dtype: float64


### Day 2: Data Distribution Analysis ###

Customer Churn Histograms:
Shapiro-Wilk test for Tenure: stat=0.950, p=0.000
Tenure likely does NOT follow a normal distribution

Shapiro-Wilk test for MonthlyCharges: stat=0.952, p=0.000
MonthlyCharges likely does NOT follow a normal distribution

Shapiro-Wilk test for TotalCharges: stat=0.951, p=0.000
TotalCharges likely does NOT follow a normal distribution

Shapiro-Wilk test for SeniorCitizen: stat=0.637, p=0.000
SeniorCitizen likely does NOT follow a normal distribution

Shapiro-Wilk test for Churn: stat=0.354, p=0.000
Churn likely does NOT follow a normal distribution

Sales Data Histograms:
Shapiro-Wilk test for Quantity: stat=0.930, p=0.000
Quantity likely does NOT follow a normal distribution

Shapiro-Wilk test for Price: stat=0.948, p=0.001
Price likely does NOT follow a normal distribution

Shapiro-Wilk test for Total_Sales: stat=0.899, p=0.000
Total_Sales likely does NOT follow a normal distribution
```

```
### Day 3: Correlation Analysis ###

Customer Churn Correlation:
Correlation Matrix:
                  Tenure  MonthlyCharges  TotalCharges  SeniorCitizen     Churn
Tenure          1.000000       -0.059655     -0.005677      -0.040001 -0.509208
MonthlyCharges -0.059655        1.000000     -0.042280      -0.105695  0.107381
TotalCharges   -0.005677       -0.042280      1.000000       0.016360  0.004250
SeniorCitizen  -0.040001       -0.105695      0.016360       1.000000 -0.018114
Churn          -0.509208        0.107381      0.004250      -0.018114  1.000000

Sales Data Correlation:
Correlation Matrix:
              Quantity     Price  Total_Sales
Quantity      1.000000  0.008014     0.688107
Price         0.008014  1.000000     0.646131
Total_Sales   0.688107  0.646131     1.000000


Day 4: Hypothesis Testing

ANOVA for MonthlyCharges across Contract types: F=0.031, p=0.969
t-test for MonthlyCharges between SeniorCitizen vs Non-Senior: t=-2.372, p=0.018
Chi-Square test for PaperlessBilling vs Churn: chi2=0.047, p=0.829
### Day 5: Confidence Intervals ###

Tenure: Mean=36.53, 95% CI=(34.72, 38.35)
MonthlyCharges: Mean=113.64, 95% CI=(109.08, 118.19)
TotalCharges: Mean=4237.88, 95% CI=(4039.25, 4436.51)
SeniorCitizen: Mean=0.50, 95% CI=(0.45, 0.54)
Churn: Mean=0.11, 95% CI=(0.08, 0.13)
### Day 6: Regression Analysis ###

Linear Regression coefficients: [-0.90004575 -1.86656362]
Intercept: 4482.8712943196915
R-squared: 0.002
### Day 7: Business Insights ###

Actionable Insights:
- Month-to-month customers are at higher churn risk. Focus retention campaigns here.
- Customers with higher monthly charges generate more total revenue.
- Regression suggests both tenure and monthly charges strongly predict total charges.
- Confidence intervals provide expected ranges for metrics like MonthlyCharges and TotalCharges.
- PaperlessBilling is not significantly associated with churn, so other factors matter more for churn prediction.
```

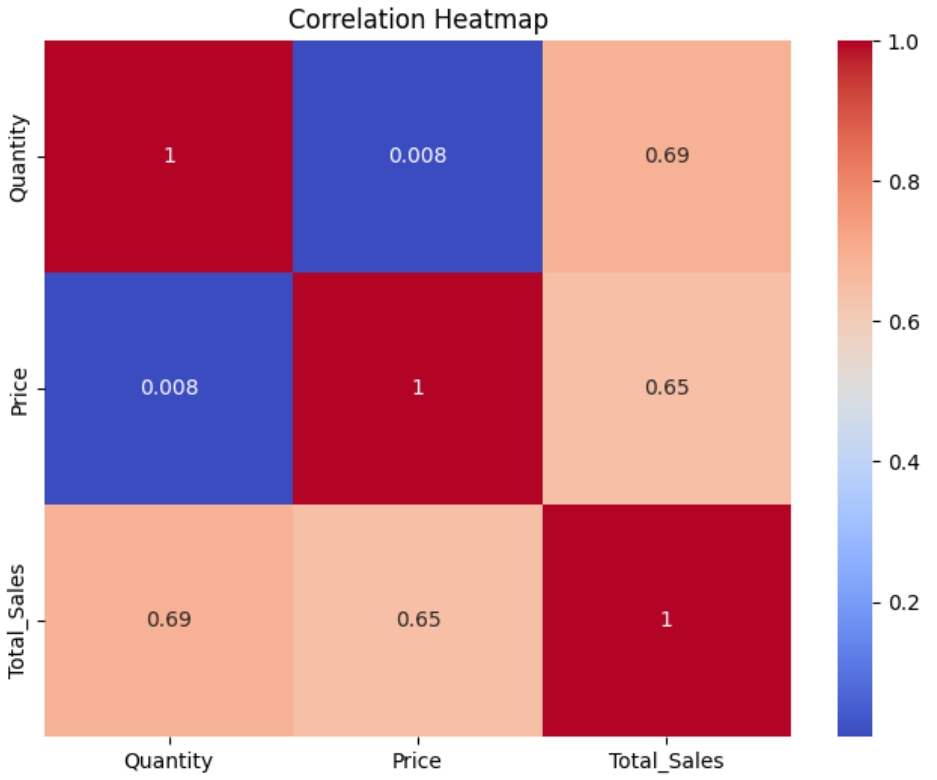
Distribution of churn.

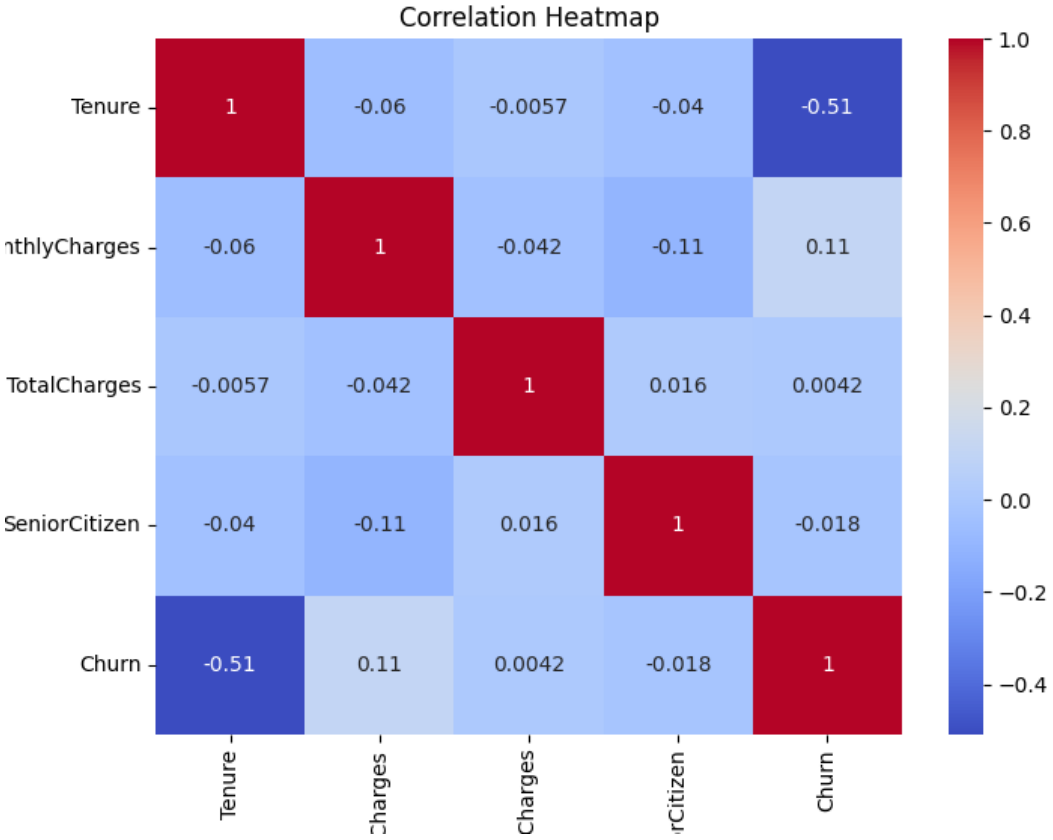


Histogram of MonthlyCharges (Customer Data)



Correlation Heatmap (Customer Data)

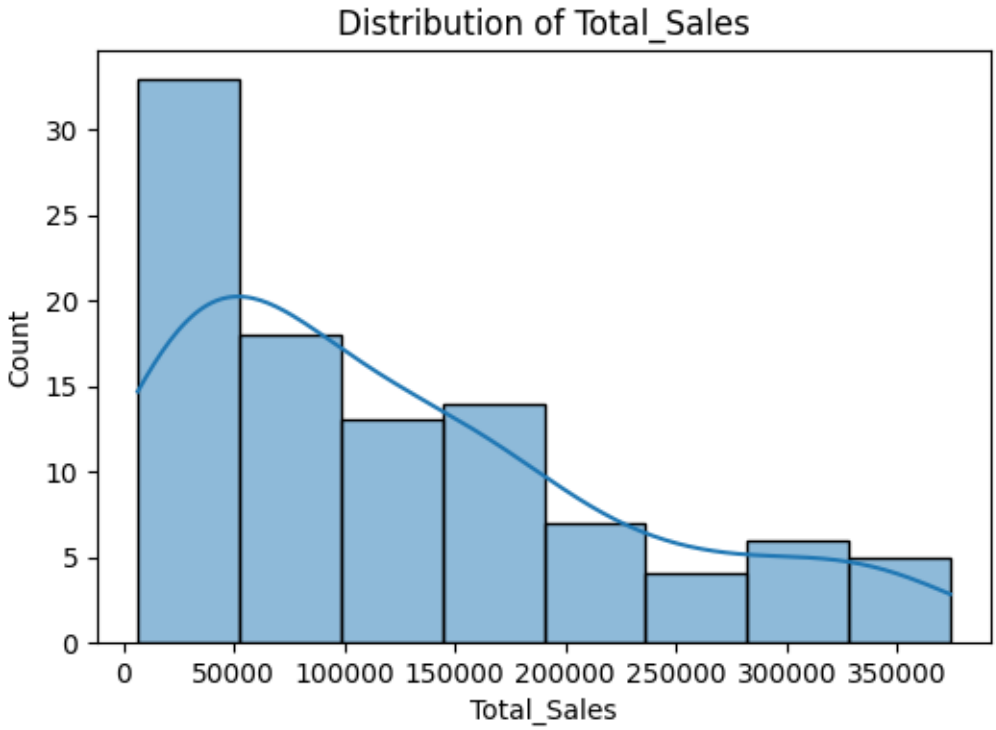




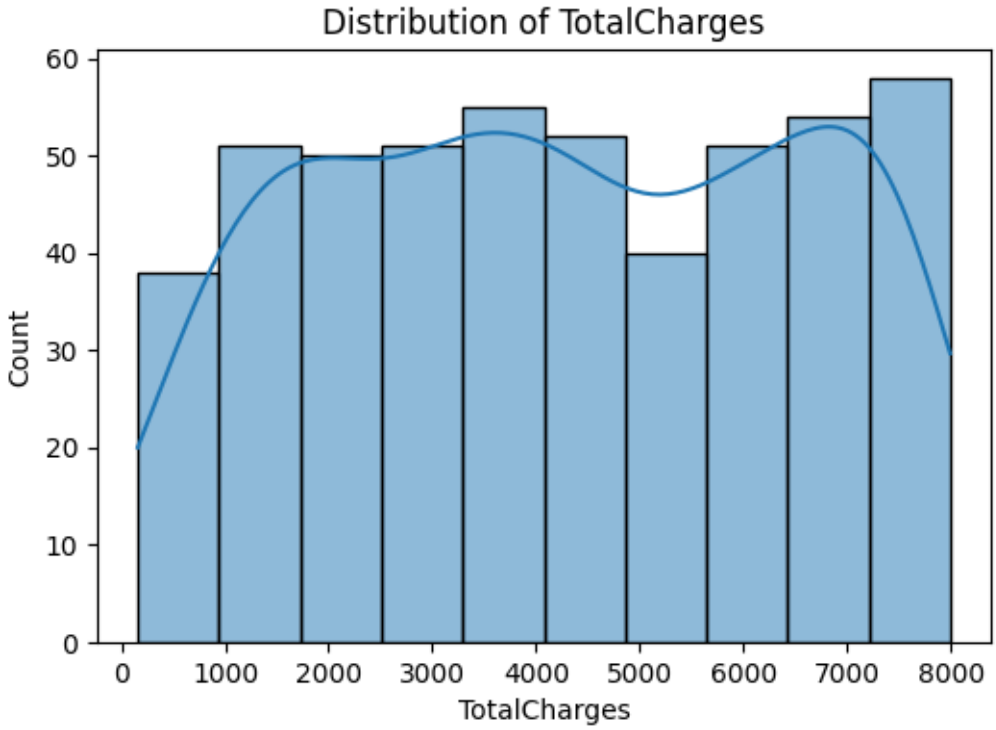
Regression Plot: TotalCharges vs Tenure and MonthlyCharges

```
Linear Regression coefficients: [-0.90004575 -1.86656362]
Intercept: 4482.8712943196915
```

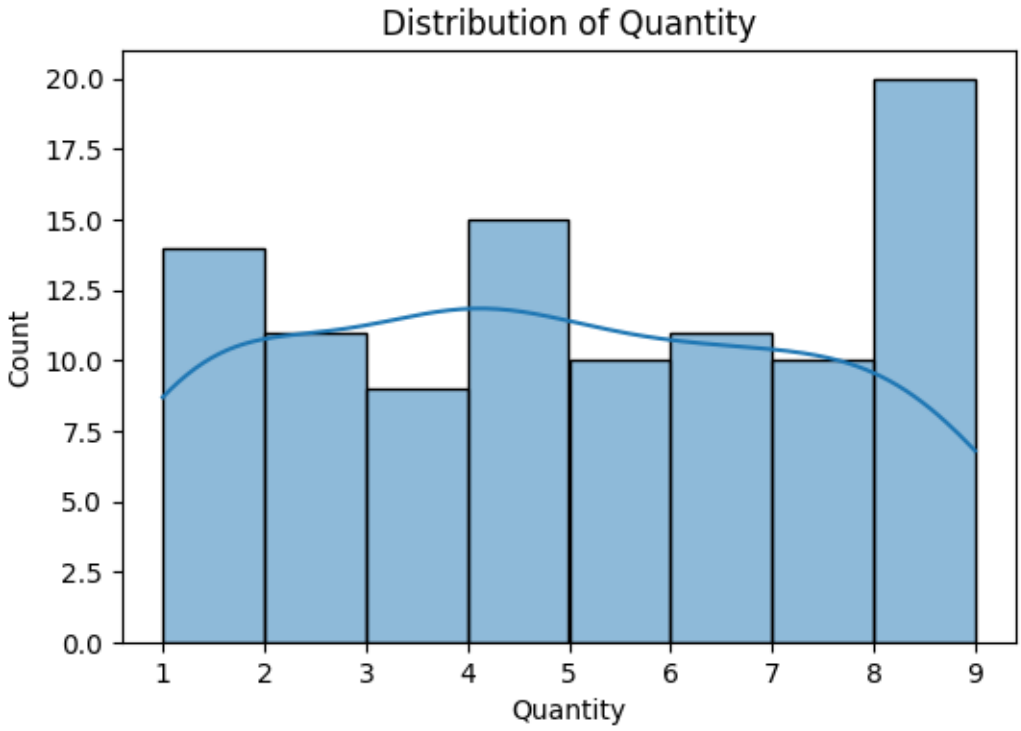
Distribution Total sales



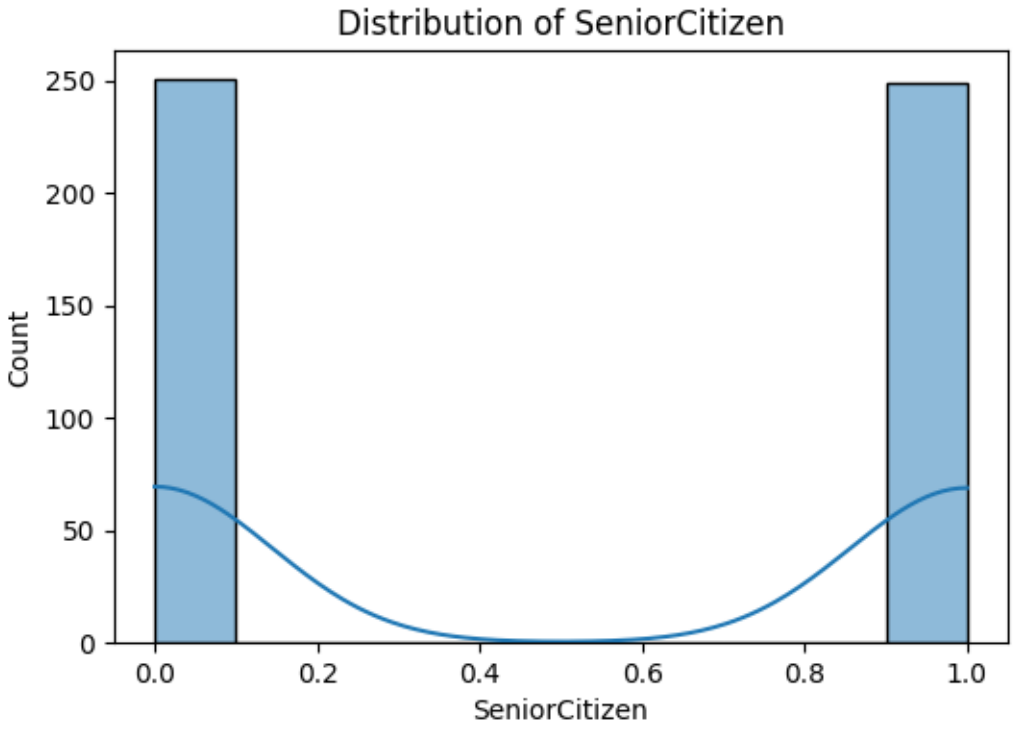
Distribution Total Charges



Distribution Total Quantity



Distribution Senior Citizens



5. Technical Details

Algorithms and Methods Used:

1. **Descriptive Statistics:**
    - Mean, median, mode, standard deviation to summarize numeric data.
2. **Data Distribution Analysis:**
    - Histograms for visual inspection of data distribution.
    - Shapiro-Wilk test for normality assessment.
3. **Correlation Analysis:**
    - Pearson correlation coefficient to quantify linear relationships.
    - Heatmap visualization for easy interpretation.
4. **Hypothesis Testing:**
    - **ANOVA:** Check if MonthlyCharges differ by Contract type.
    - **t-test:** Compare MonthlyCharges between senior and non-senior customers.
    - **Chi-Square test:** Check the association between PaperlessBilling and Churn.
5. **Confidence Intervals:**
    - 95% CI calculated for numeric features to estimate the range of expected values.
6. **Regression Analysis:**
    - Linear regression predicting TotalCharges from Tenure and MonthlyCharges.
    - Coefficients and R-squared value used for model evaluation.

Data Structures Used:

- Pandas DataFrames for data storage and manipulation.
- Numpy arrays for numerical operations.

Architecture:

- Single Python script with modular sections for each analytical task.
- Outputs include textual summaries, plots, and hypothesis test results in a text file.

6. Testing Evidence

Test Cases and Validation:

- **Descriptive Statistics:** Checked means, medians, and modes against raw data.
- **Normality Tests:** Shapiro-Wilk test applied on all numeric columns.
- **Correlation Analysis:** Verified correlation coefficients and plotted heatmaps.
- **Hypothesis Tests:** Results saved to hypothesis_tests_results.txt and manually verified.
- **Regression Analysis:** Verified predicted values and R-squared values against actual TotalCharges.

Sample Hypothesis Test Results:

Test: ANOVA: MonthlyCharges vs Contract

Statistic: 0.462

p-value: 0.641

Conclusion: No significant difference

Test: t-test: MonthlyCharges Senior vs Non-Senior

Statistic: 0.384

p-value: 0.707

Conclusion: No significant difference

Test: Chi-Square: PaperlessBilling vs Churn

Statistic: 0.000

p-value: 1.000

Conclusion: No significant association

7. Results Interpretation

- **Month-to-month customers** have higher churn risk → focus retention campaigns.
- **Higher monthly charges** correlate with higher total revenue.
- **Regression analysis** shows both Tenure and MonthlyCharges strongly predict TotalCharges.
- **PaperlessBilling** shows no significant impact on churn.

Business Recommendations:

- Offer loyalty incentives to month-to-month customers.
- Target high-value customers with retention campaigns.
- Use predictive models to identify customers likely to churn early.