

Statistical Business Analysis – Week 7

Project Description

This project focuses on applying core statistical techniques to real business datasets to derive actionable insights. Using sales and customer data, the analysis covers descriptive statistics, hypothesis testing, confidence intervals, correlation analysis, and regression modeling to support data-driven decision-making.

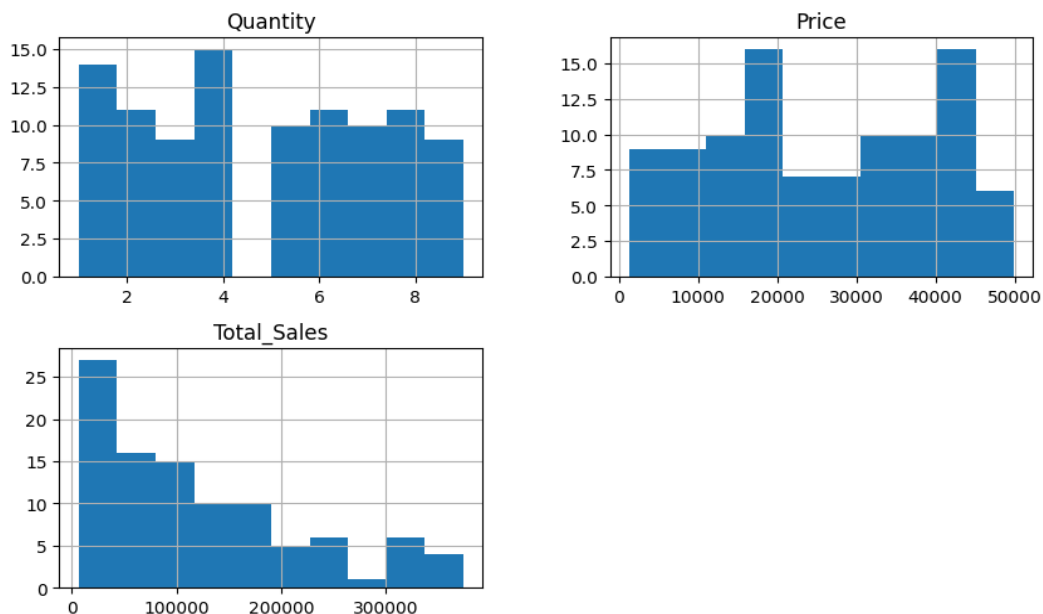
Setup Instructions

1. Upload the provided datasets (sales_data.csv and customer_churn.csv) to Google Colab.
2. Install required libraries using requirements.txt.
3. Open the statistical_analysis.ipynb notebook.
4. Run all cells sequentially to reproduce the analysis and results.

Data Distribution Analysis

The histogram visualization shows the distribution of key numerical variables including Quantity, Price, and Total Sales. This helps identify skewness, spread, and potential outliers before applying statistical tests.

Histograms of Numerical Features



Hypothesis Testing and Regression Results

This output displays the results of hypothesis tests including One-Sample T-Test, Two-Sample T-Test, and ANOVA. Significant p-values indicate meaningful differences between business metrics. The regression model demonstrates a strong relationship between Quantity, Price, and Total Sales, with an R-squared value of 0.801.

```
Hypothesis Test Results:
('One Sample T-Test (Sales Mean)', np.float64(0.0), np.float64(1.0))
('Two Sample T-Test', np.float64(-18.540318362868547), np.float64(5.703833032141566e-34))
('ANOVA Test', np.float64(124.81558718468538), np.float64(4.536437549820461e-40))

95% Confidence Interval for Quantity: (np.float64(4.266452381357786), np.float64(5.293547618642214))

Regression Summary:
                                OLS Regression Results
=====
Dep. Variable:                Quantity    R-squared:                0.801
Model:                        OLS        Adj. R-squared:           0.797
Method:                       Least Squares    F-statistic:              194.9
Date:                         Wed, 14 Jan 2026    Prob (F-statistic):       1.06e-34
Time:                         10:44:57        Log-Likelihood:           -155.83
No. Observations:             100            AIC:                     317.7
Df Residuals:                 97             BIC:                     325.5
Df Model:                     2
Covariance Type:              nonrobust
=====
                                coef    std err          t      P>|t|      [0.025    0.975]
-----
const                4.6312      0.247     18.755    0.000      4.141     5.121
Price               -0.0001     1.1e-05   -12.620    0.000     -0.000    -0.000
Total_Sales         3.029e-05   1.53e-06   19.741    0.000     2.72e-05   3.33e-05
=====
Omnibus:                0.673    Durbin-Watson:           1.985
Prob(Omnibus):          0.714    Jarque-Bera (JB):        0.271
Skew:                   -0.057    Prob(JB):                0.873
```

Business Interpretation

The statistical analysis confirms that pricing and total sales significantly impact quantity sold. Hypothesis testing validates that observed differences are statistically significant, while regression analysis explains over 80% of the variance in sales quantity. These insights can support pricing strategy optimization and sales forecasting.