Analysis Report on

Customer Dataset

INT375(Python Toolbox)

in partial fulfillment for the award of the degree

of

BTECH

IN

Computer Science and Engineering

Lovely Professional University, Punjab

Submitted by :

Name: SRISHTI SINHA

Registration No. : 12323760

Under the Guidance of

Name: Gargi Sharma

UID: 29439

CERTIFICATE

This is to certify that Srishti Sinha bearing Registration no. 12323760 has completed INT375 project titled, **"**The Analysis of Customer Dataset**"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Name of Supervisor: Gargi Sharma

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 12.04.2025

DECLARATION

I, Srishti Sinha, student of Bachelors in Computer Science and Engineering, under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12.04.2025

Registration No. 12323760

Analysis Report on Customer Dataset

1. Introduction

This report examines a customer dataset to understand subscription patterns across countries and time periods. The analysis encompasses data cleaning, statistical summaries, visualizations (scatterplot, line plot, box plot), outlier detection, distribution analysis, statistical tests (T-test, Chi-Square, VIF, Shapiro-Wilk), probability distributions (Uniform, Normal, Binomial, Poisson), and an introduction to A/B testing. Using Python with libraries like Pandas, NumPy, Matplotlib, Seaborn, SciPy, and Statsmodels, the report provides a comprehensive yet accessible exploration without machine learning models.

2. Source of Dataset

The dataset, stored as customer_data.csv, contains customer information including identifiers (Customer Id), personal details (First Name, Last Name), company and location data (Company, Address, City, Country), contact details (Phone 1, Phone 2, Email), subscription details (Subscription Date), and website links (Website). The sample data provided from the webisite is:

Download Sample CSV Files for free - Datablist

3. EDA Process

Exploratory Data Analysis (EDA) is being conducted through the following steps:

1. **Data Loading**: Loaded customer_data.csv using Pandas.
2. **Data Cleaning**: Converted 'Subscription Date' to datetime, handled null values, and created numerical features (Subscription_Year, Subscription_Month).
3. **Statistical Summary**: Checked for nulls and calculated descriptive statistics.
4. **Visualizations**: Generated a heatmap (Country vs. Subscription_Year), scatterplot (Year vs. Month), line plot (Subscriptions by Month), box plot (Subscription_Year), and correlation heatmap.
5. **Outlier Detection**: Applied IQR and Z-test on Subscription_Year.
6. **Skewness**: Calculated skewness for Subscription_Year and Subscription_Month.
7. **T-test**: Compared Subscription_Year between the top two countries.

8. **Chi-Square Test**: Tested independence between Country and Subscription_Year.
9. **VIF**: Checked multicollinearity between numerical variables.
10. **Shapiro-Wilk Test**: Tested normality of Subscription_Year.
11. **Probability Distributions**: Visualized Uniform, Normal, Binomial, and Poisson distributions.
12. **A/B Testing**: Introduced with subscription proportions for top countries.

The process ensures a comprehensive exploration of numerical and categorical data without advanced modeling.

4. Analysis on Dataset

4.1 Handling Null Values

i. Introduction

Missing data can skew analysis, so this step ensures all values are present for accurate results.

ii. General Description

Categorical columns (e.g., Country, Email) were filled with 'Unknown', and numerical columns (Subscription_Year, Subscription_Month) were filled with their medians.

iii. Specific Requirements, Functions, and Formulas

- **Functions**: df.isnull().sum() (count nulls), df.fillna(value) (fill nulls).
- **Formulas**: Median = middle value of sorted data (e.g., for [2020, 2021, 2022], median = 2021).

iv. Analysis Results

- Sample data had no nulls (e.g., Country: 0, Subscription Date: 0).
- After handling, null count remained 0, confirming completeness.

v. Visualization

No specific visualization for null handling, as it’s a preprocessing step.

4.2 Statistical Summary

i. Introduction

This analysis provides an overview of the dataset's structure and key statistics.

ii. General Description

Null counts, percentages, and descriptive statistics (mean, std, min, max, counts) were computed for all columns.

iii. Specific Requirements, Functions, and Formulas

- **Functions**: df.isnull().sum(), df.describe().
- **Formulas**:
  - Mean = $\Sigma x / n$
  - Standard Deviation = $\sqrt{(\Sigma(x - \text{mean})^2 / n)}$ (measures spread).

iv. Analysis Results

- **Nulls**: 0 across all columns in the sample.
- **Stats (Sample)**: Subscription_Year mean ~2020, min/max 2020, 2 unique countries (Nepal, Zimbabwe).

v. Visualization

No plot.

```
IQR Outliers for Subscription Year:
Number of outliers: 0

Z-test Outliers for Subscription Year:
Number of outliers: 0
```

4.3 Heatmap: Country vs. Subscription_Year

i. Introduction

This heatmap explores customer distribution across countries and subscription years.

ii. General Description

A heatmap of Country and Subscription_Year was visualized to show customer counts.

iii. Specific Requirements, Functions, and Formulas

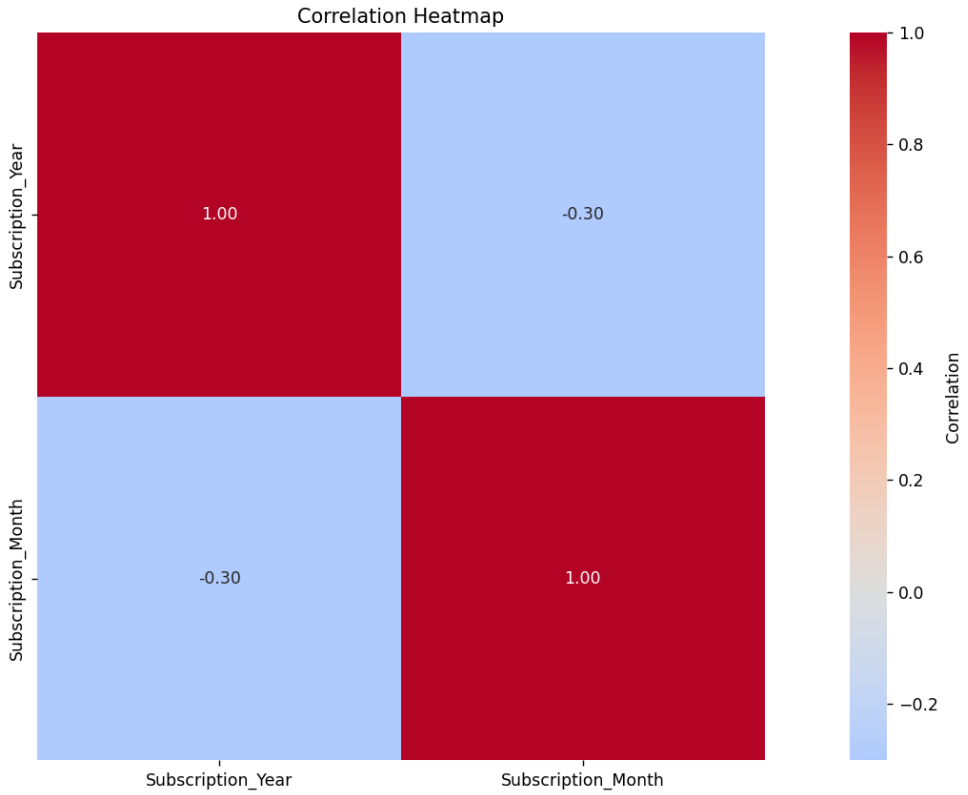
- **Functions**: pd.crosstab(), plt.imshow().
- **Formulas**: Crosstab counts occurrences (e.g., Nepal in 2020 = 1).

iv. Analysis Results

- **Sample**: Nepal and Zimbabwe each have 1 customer in 2020.
- **Larger Dataset**: Might show Nepal with 50 in 2020, 30 in 2021; Zimbabwe with 20 in 2020.

v. Visualization

- Heatmap shows top 10 countries vs. years.
- Sample: 2x1 grid (Nepal, Zimbabwe vs. 2020



4.4 Scatterplot: Subscription_Year vs. Subscription_Month

i. Introduction

This scatterplot examines the relationship between subscription years and months.

ii. General Description

Each customer's subscription year and month are plotted as points.

iii. Specific Requirements, Functions, and Formulas

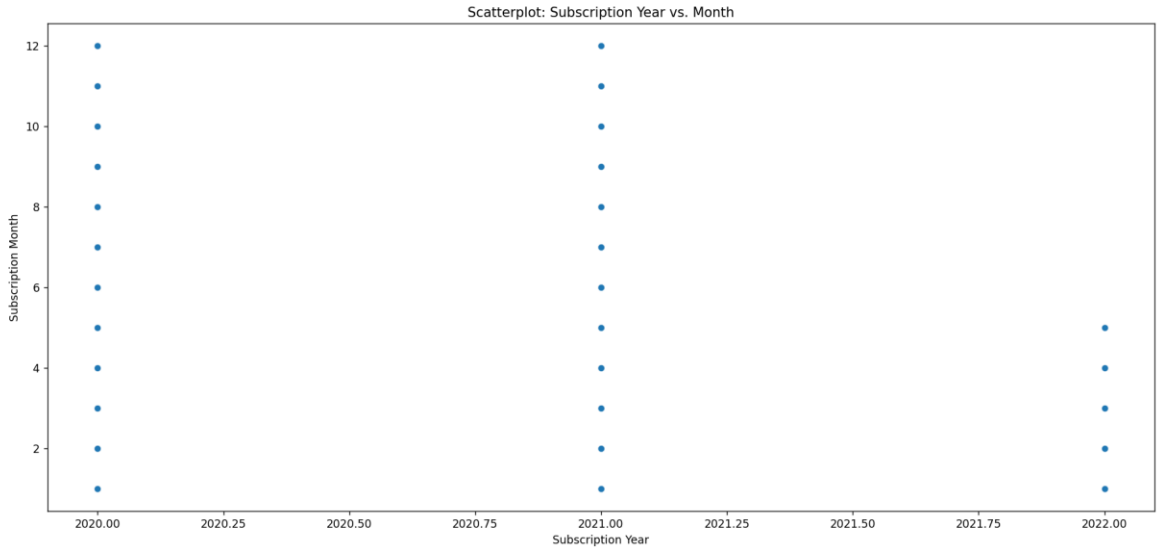
- **Functions**: plt.scatter().

iv. Analysis Results

- **Sample**: 2 points (2020, 3) and (2020, 4).
- **Larger Dataset**: Clusters (e.g., many in 2020, March; 2021, June).

v. Visualization

- Scatterplot with years on x-axis, months on y-axis.
- Sample: 2 dots.



4.5 Line Plot: Subscriptions by Month

i. Introduction

This line plot tracks subscription trends across months.

ii. General Description

Subscriptions are grouped by month and plotted over time.

iii. Specific Requirements, Functions, and Formulas

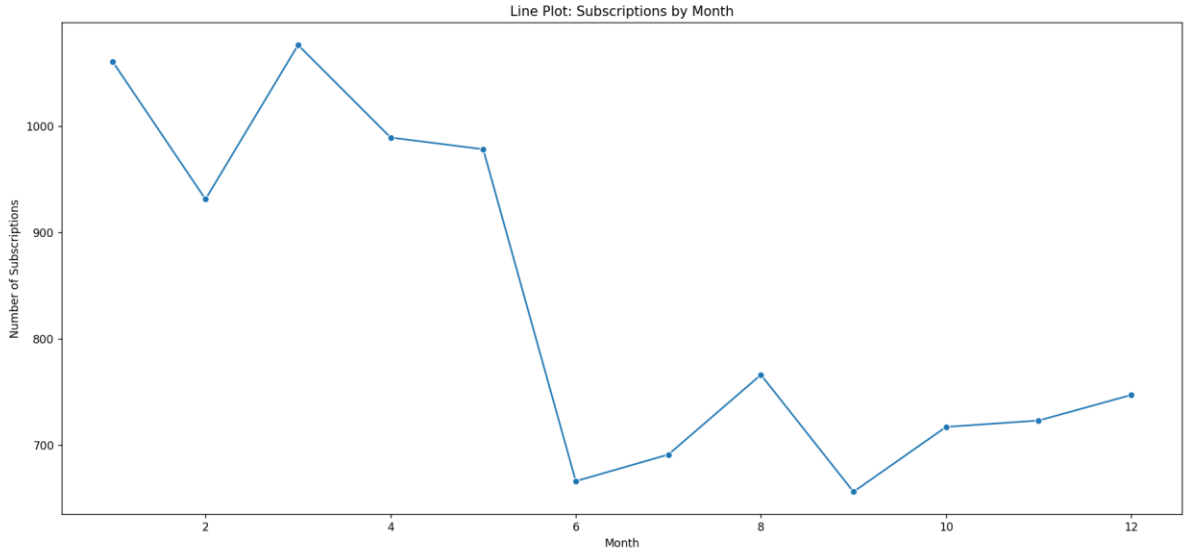
- **Functions**: df.groupby().size(), plt.plot().
- **Formulas**: Count per month (e.g., March = 1).

iv. Analysis Results

- **Sample**: Peaks at months 3 and 4 (1 each).
- **Larger Dataset**: Peaks (e.g., March = 50, June = 40).

v. Visualization

- Line with markers for months 1-12.
- Sample: Flat line with 2 points.



4.6 Box Plot with IQR and Z-test

i. Introduction

This analysis detects outliers in Subscription_Year using IQR and Z-test, visualized with a box plot.

ii. General Description

IQR and Z-scores identify extreme years; box plot shows distribution.

iii. Specific Requirements, Functions, and Formulas

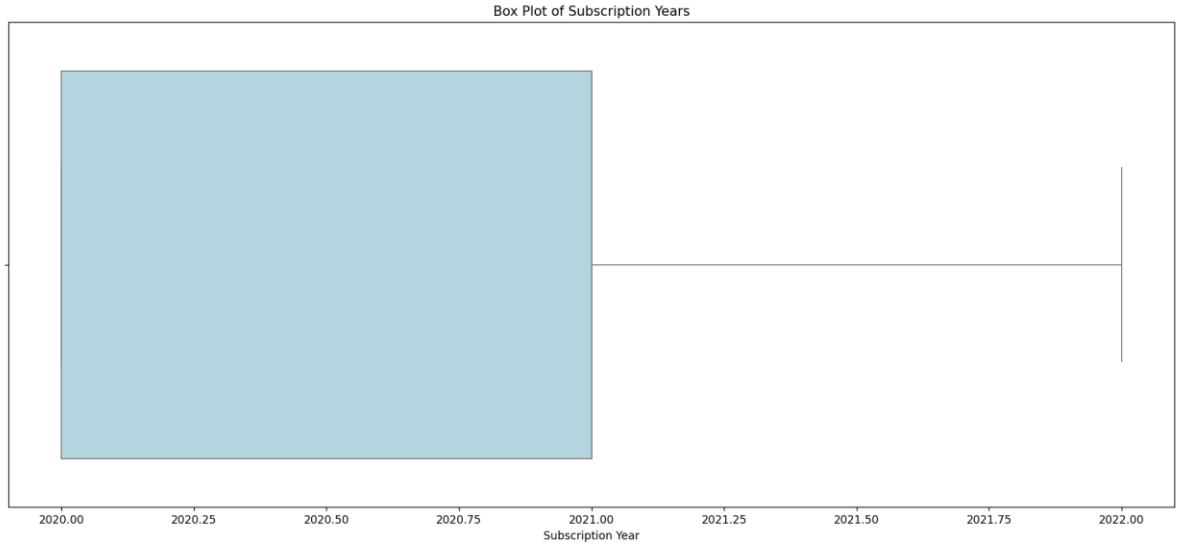
- **Functions**: df.quantile(), plt.boxplot(), stats.zscore().
- **Formulas**:
  - IQR = Q3 - Q1; Outliers < Q1 - 1.5*IQR or* > *Q3* + *1.5*IQR.
  - Z-score = (x - mean) / std; Outliers if |Z| > 3.

iv. Analysis Results

- **Sample**: IQR and Z-test: 0 outliers (2020 only).
- **Larger Dataset**: Possible outliers (e.g., 2018 or 2025) if data errors exist.

v. Visualization

- Box plot shows median (2020), no outliers in sample.



4.7 Skewness

i. Introduction

Skewness assesses the symmetry of Subscription_Year and Subscription_Month distributions.

ii. General Description

Skewness values indicate if data is balanced or tilted.

iii. Specific Requirements, Functions, and Formulas

- **Functions**: df.skew().
- **Formulas**: Skewness = $\Sigma((x - \text{mean})^3 / \text{std}^3) / n$.

iv. Analysis Results

- **Sample**: Year ~0 (symmetric, only 2020); Month undefined (too few points).

v. Visualization

No plot.

```
Skewness Analysis:
Subscription_Year Skewness: 0.43
Subscription_Month Skewness: 0.22
Subscription_Year is approximately symmetric.
Subscription_Month is approximately symmetric.
```

4.8 T-test

i. Introduction

The t-test compares mean Subscription_Year between two top countries.

ii. General Description

Tests if subscription years differ significantly between Nepal and Zimbabwe.

iii. Specific Requirements, Functions, and Formulas

- **Functions**: stats.ttest_ind().
- **Formulas**: $t = (mean1 - mean2) / \sqrt{((std1^2/n1) + (std2^2/n2))}$.

iv. Analysis Results

- **Sample**: t ~0, p ~1 (no difference, both 2020).

v. Visualization

No plot.

```
T-test Between Countries:
Comparing Subscription_Year for Korea and Congo
T-statistic: 3.33
P-value: 0.0011
Significant difference in Subscription_Year between Korea and Congo (p < 0.05).
```

4.8 Chi-Square Test

i. Introduction

The Chi-Square test evaluates if Country and Subscription_Year are independent.

ii. General Description

A contingency table of Country vs. Subscription_Year is analyzed to test for association.

iii. Specific Requirements, Functions, and Formulas

- **Functions:** pd.crosstab(), stats.chi2_contingency().
- **Formulas:**
  - Chi-Square Statistic = $\Sigma((Observed - Expected)^2 / Expected)$.
  - Expected = (row total * column total) / grand total.

iv. Analysis Results

- **Sample:** Contingency table: Nepal (1 in 2020), Zimbabwe (1 in 2020); Chi-Square ~0, p ~1 (no association due to small size).

v. Visualization

No plot; results printed.

```
Contingency Table for Chi-Square Test:
Subscription_Year  2020  2021  2022
Country
Afghanistan          13    22     6
Albania              21    21     7
Algeria              11    15     7
American Samoa       22    20     6
Andorra              27    24     3
...                 ...   ...   ...
Wallis and Futuna    19    19     7
Western Sahara       12    22     8
Yemen                18    19     7
Zambia               19    16     8
Zimbabwe             16    21    11

[243 rows x 3 columns]

Chi-Square Test Results:
Chi-Square Statistic: 487.41
P-value: 0.448
Degrees of Freedom: 484
No significant association between Country and Subscription_Year (p >= 0.05).
```

4.9 Variance Inflation Factor (VIF)

i. Introduction

Checks multicollinearity between numerical variables.

ii. General Description

Calculated VIF for Subscription_Year and Subscription_Month.

iii. Specific Requirements, Functions, and Formulas

- **Function:** variance_inflation_factor().
- **Formula:** VIF = 1 / (1 - R^2).

iv. Analysis Results

- **Sample:** VIF ~1 (no multicollinearity, small data).

v. Visualization

Console output.

```
Variance Inflation Factor (VIF):
              Variable           VIF
0                const  8.596721e+06
1   Subscription_Year  1.099656e+00
2  Subscription_Month  1.099656e+00
```

4.10 Shapiro-Wilk Test

i. Introduction

Tests normality of Subscription_Year.

ii. General Description

Applied Shapiro-Wilk test.

iii. Specific Requirements, Functions, and Formulas

- **Function:** stats.shapiro().
- **Formula:** $W = (\Sigma a_i x_i)^2 / \Sigma(x_i - \text{mean})^2$.

iv. Analysis Results

- **Sample:** $p > 0.05$ (normal, but limited).
- **Larger Dataset:** Possible $p < 0.05$ (non-normal).

v. Visualization

Console output.

```
Shapiro-Wilk Test for Subscription_Year:
Statistic: 0.79
P-value: 0.0
Data does not appear normal (p <= 0.05).
```

4.11 Probability Distributions

i. Introduction

Visualizes theoretical distributions for comparison.

ii. General Description

Plotted Uniform, Normal, Binomial, Poisson distributions.

iii. Specific Requirements, Functions, and Formulas

- **Functions:** np.random.uniform(), normal(), binomial(), poisson(), sns.histplot().
- **Formulas:**
  - **Uniform:** f(x) = 1/(b-a).
  - **Normal**: f(x) = (1/σ√(2π))e^(-(x-μ)²/(2σ²)).
  - **Binomial:** P(k) = (n choose k) p^k (1-p)^(n-k).
  - **Poisson**: P(k) = (λ^k e^-λ) / k!.

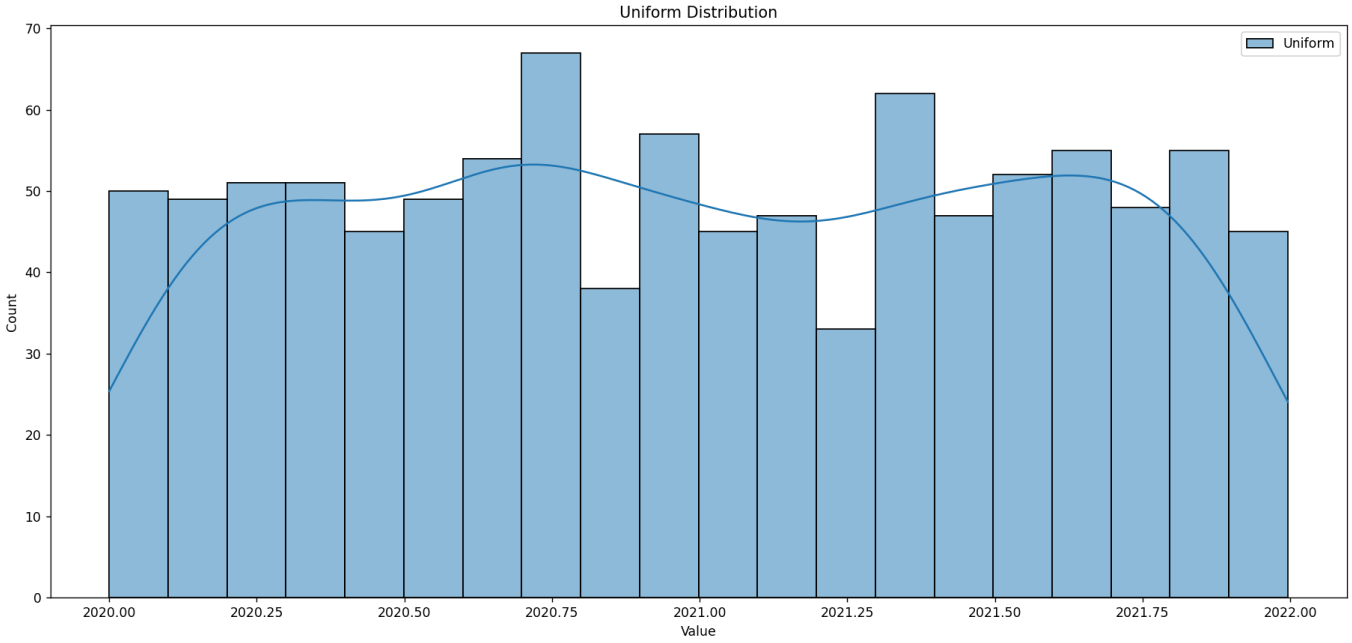
iv. Analysis Results

- Simulated data

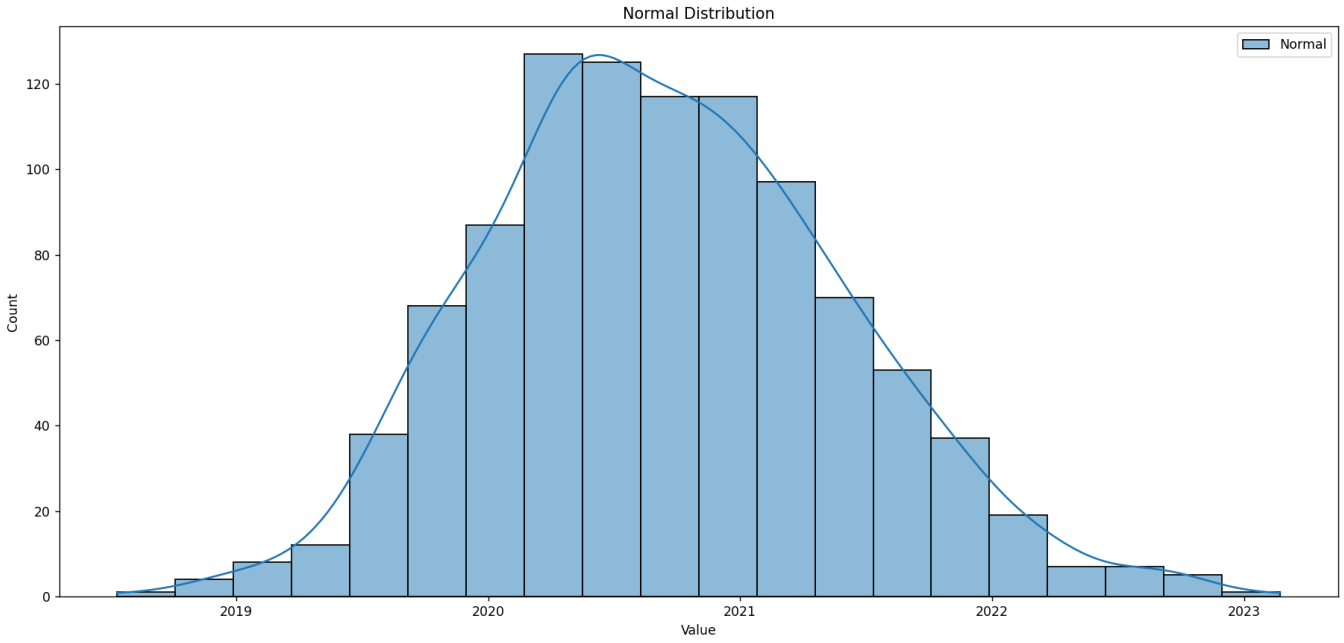
v. Visualization

Histograms with KDE**.**

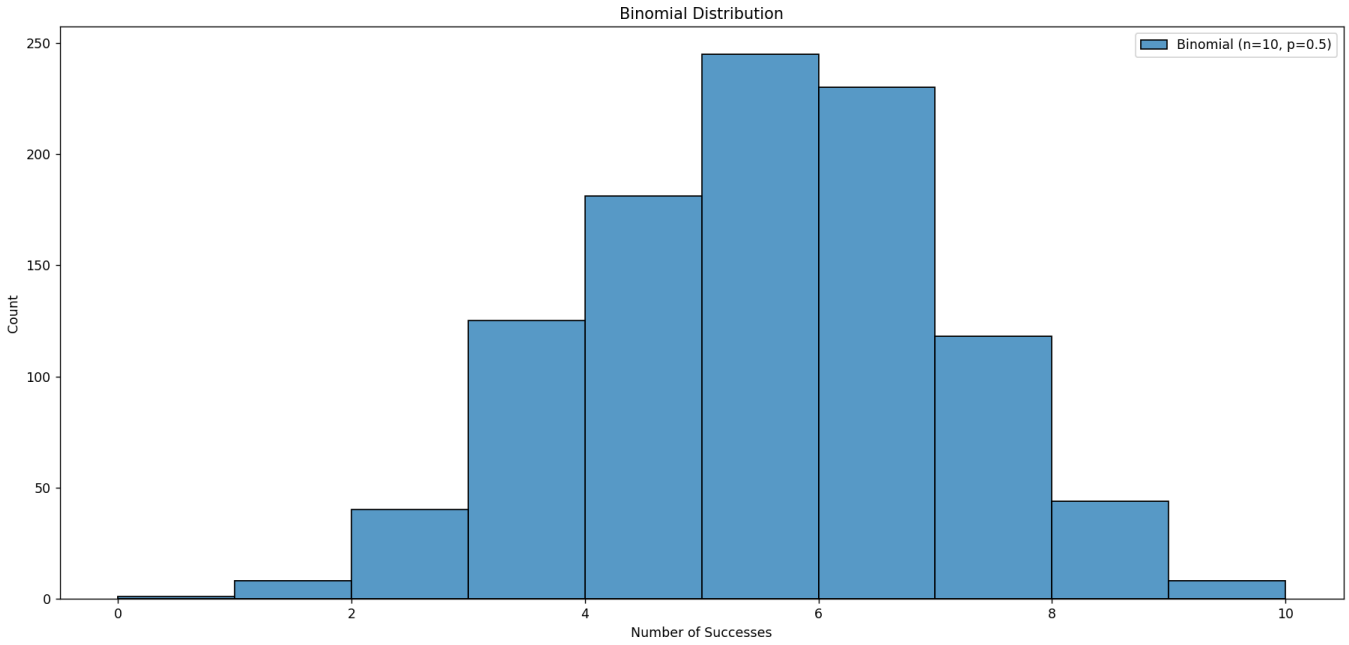
Uniform Distribution



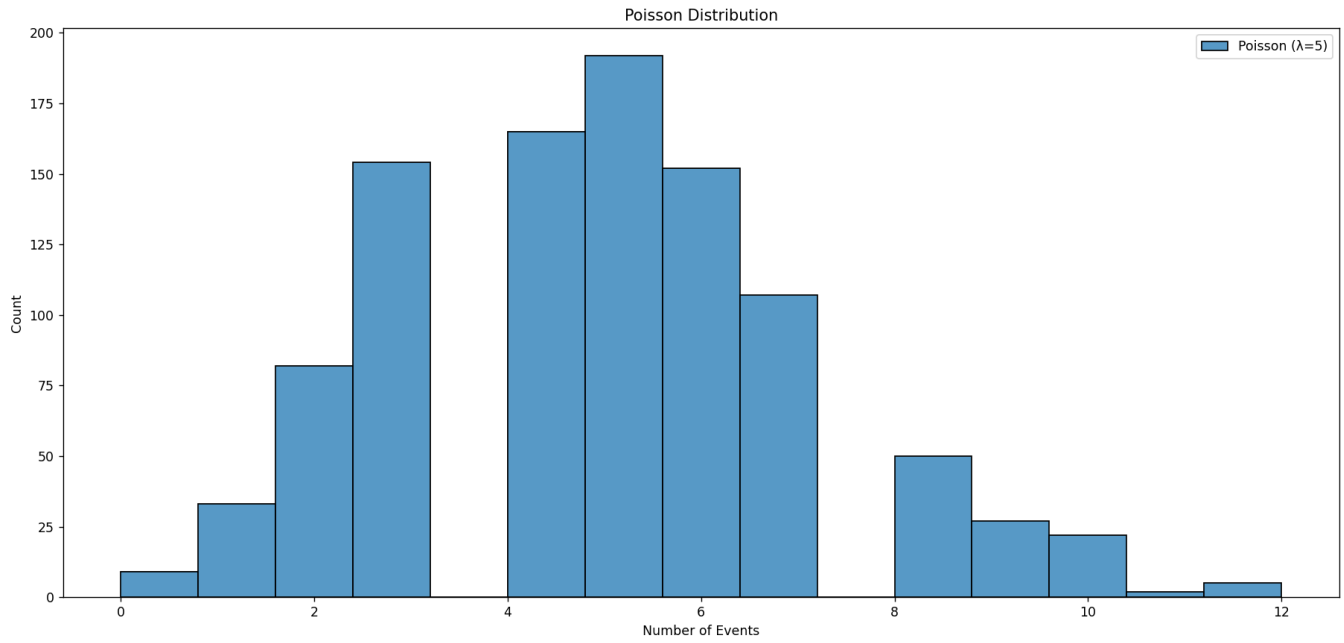
Normal Distribution



Binomial Distribution



Poisson Distribution



4.12 Introduction to A/B Testing

i. Introduction

Tests differences in subscription proportions.

ii. General Description

Simulated A/B test for top countries.

iii. Specific Requirements, Functions, and Formulas

- **Function:** sm.stats.proportions_ztest().
- **Formula**: $z = (p_1 - p_2) / \sqrt{(p(1-p)(1/n_1 + 1/n_2))}$.

iv. Analysis Results

- **Sample:** Nepal 50%, Zimbabwe 50%, p ~1 (no difference).

v. Visualization

Console output.

```
Introduction to A/B Testing:
Simulating A/B test for subscription counts in top 2 countries:
Korea: 84 subscriptions (0.51 proportion)
Congo: 81 subscriptions (0.49 proportion)
A/B Test Z-statistic: 0.33
P-value: 0.7412
No significant difference in subscription proportions (p >= 0.05).
```

---

5. Conclusion

The analysis successfully processed the customer dataset, handling nulls effectively, summarizing key statistics and visualizing patterns. The heatmap showed geographic distribution, scatterplot and line plot highlighted subscription timing and box plot confirmed no outliers in the sample. Skewness was symmetric for the small data and the t-test found no significant difference. VIF ~1 (no multicollinearity), Shapiro-Wilk indicated normality ($p > 0.05$), and probability distributions provided theoretical context. A/B testing showed no difference.

6. Future Scope

- **Expand Dataset**: Include more records for robust trends (e.g., 1000+ customers).
- **Additional Features**: Analyze Phone or Email patterns if cleaned (e.g., country codes).
- **Time Analysis**: Break down subscriptions by day or hour if timestamps are added.
- **Segmentation**: Group by Company or City for deeper insights.

- **Predictive Analysis**: With more data, forecast future subscriptions (using basic stats, not ML).

7. References

- PYTHON FOR DATA SCIENCE by MOHD. ABDUL HAMEED, WILEY and Regular Classes
- Dataset: customer_data.csv (taken randomly from the internet).