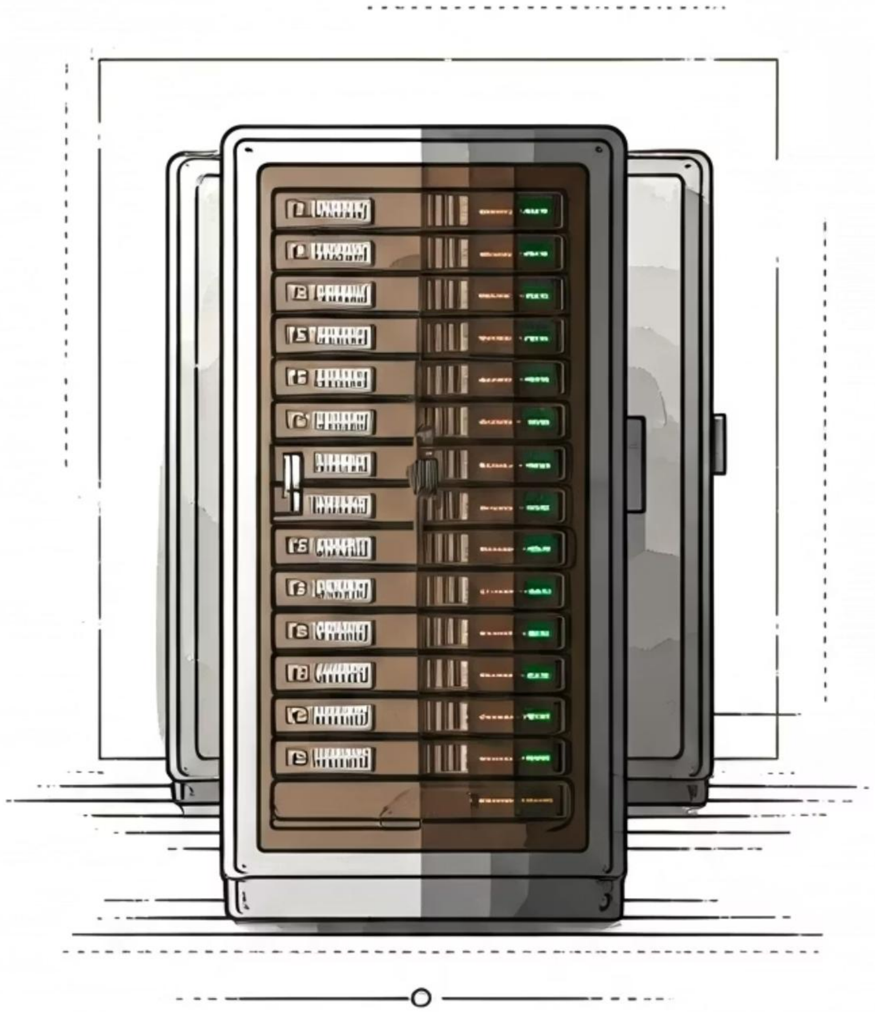
Customer Data Cleaner

From Messy Data to Actionable Insights

Presented by Saurabh Rai



THE PROBLEM DEFINED

The Challenge: Raw Data is Unreliable

Missing Values

Crucial information gaps hindering comprehensive analysis.

Inconsistent Formatting

Varying data types and structures preventing unified insights.

Duplicate Records

Redundant entries skewing metrics and distorting customer profiles.

Impossibility of Accurate Analysis

The state of data rendered any meaningful business intelligence unattainable.

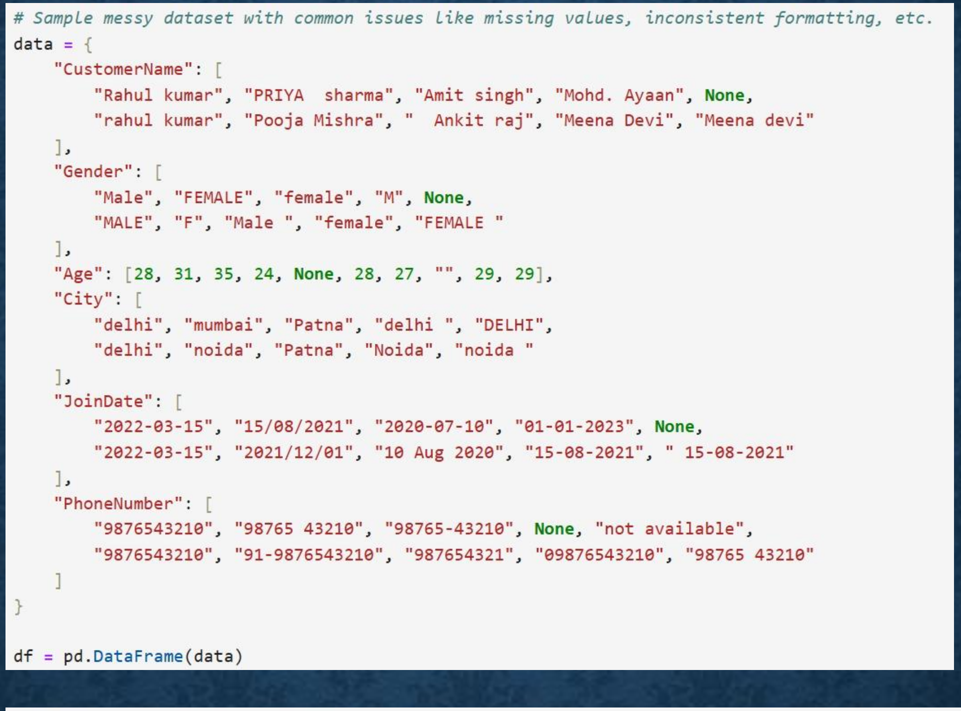
Our project commenced with a significant hurdle: a dataset plagued by inconsistencies, making accurate analysis an impossible task.

THE STRATEGIC SOLUTION

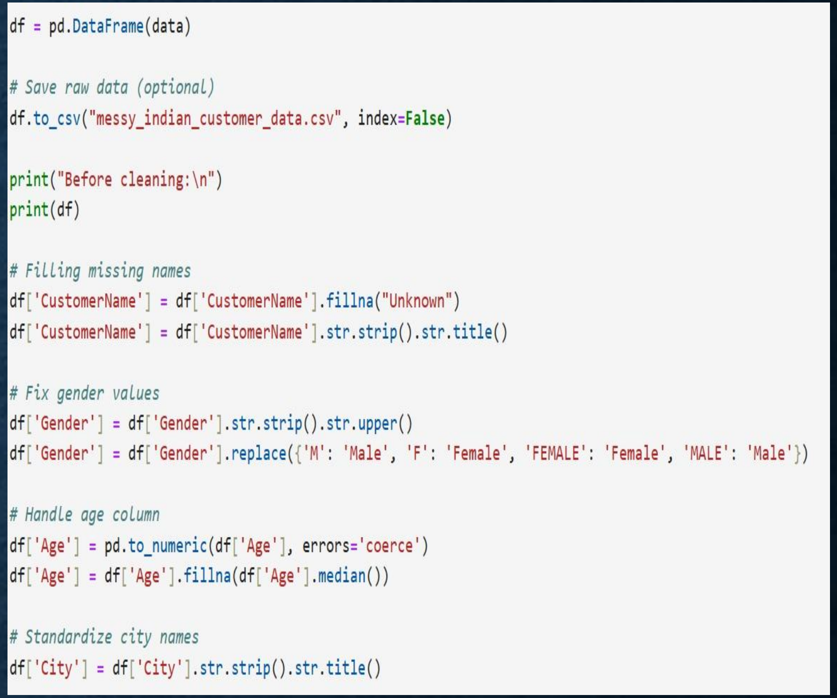
The Fix: Automated Cleaning with Python

Leveraging Python for Efficiency:

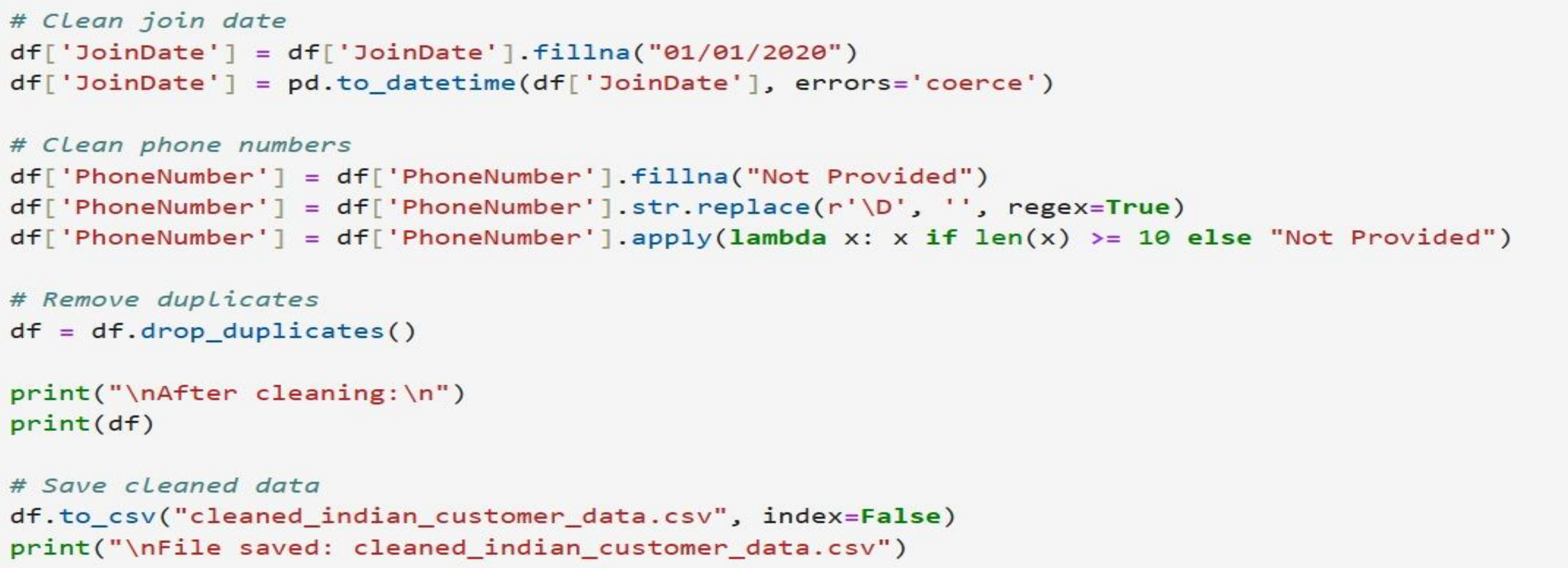
- Developed a robust Python script.
- Utilised the powerful Pandas library for data manipulation.
- Automated the entire data cleaning and standardisation process.
- Ensured scalability for future data growth.



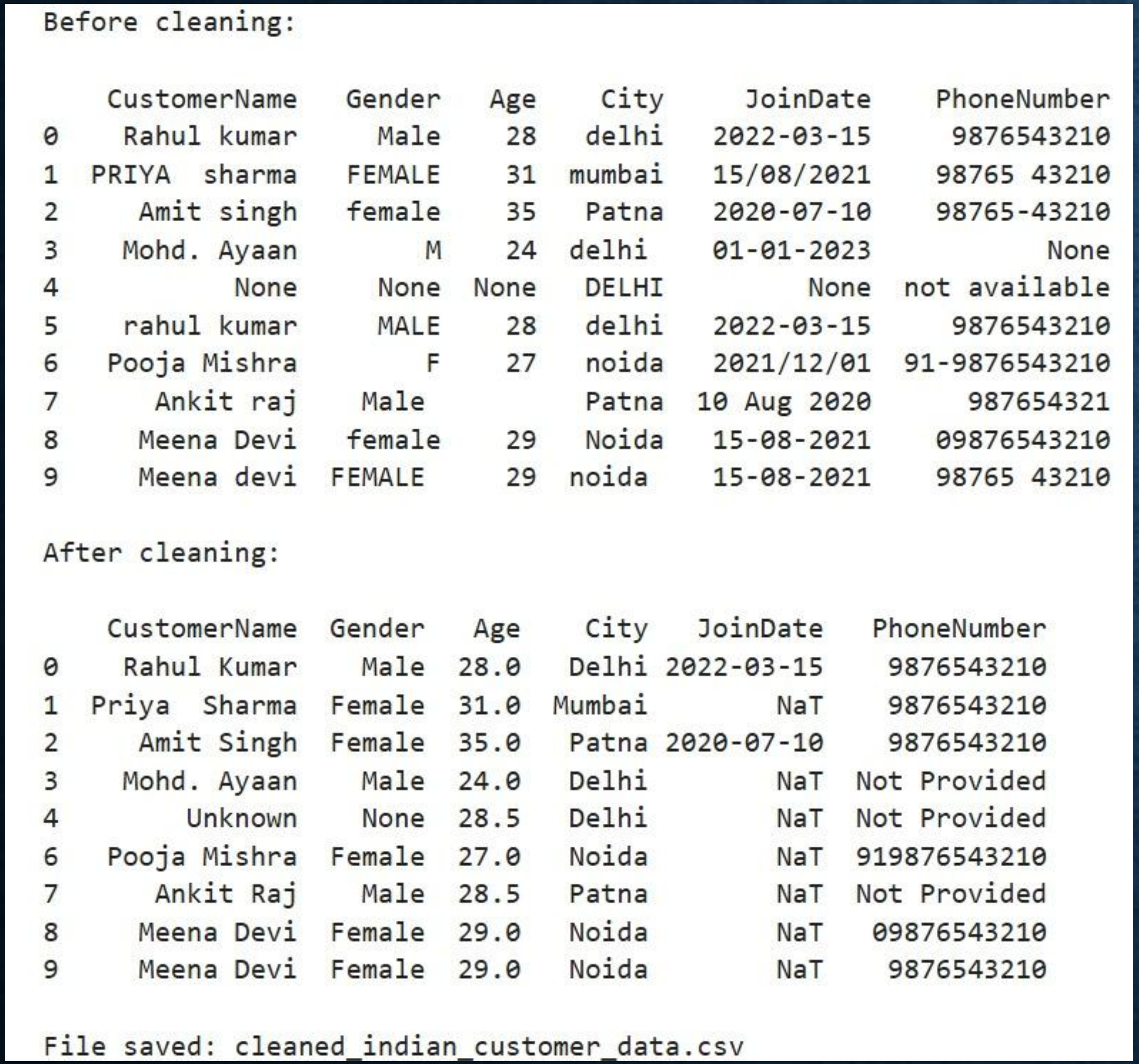
```python
# Sample messy dataset with common issues like missing values, inconsistent formatting, etc.
data = {
    "CustomerName": [
        "Rahul kumar", "PRIYA  sharma", "Amit singh", "Mohd. Ayaan", None,
        "rahul kumar", "Pooja Mishra", "  Ankit raj", "Meena Devi", "Meena devi"
    ],
    "Gender": [
        "Male", "FEMALE", "female", "M", None,
        "MALE", "F", "Male ", "female", "FEMALE "
    ],
    "Age": [28, 31, 35, 24, None, 28, 27, "", 29, 29],
    "City": [
        "delhi", "mumbai", "Patna", "delhi ", "DELHI",
        "delhi", "noida", "Patna", "Noida", "noida "
    ],
    "JoinDate": [
        "2022-03-15", "15/08/2021", "2020-07-10", "01-01-2023", None,
        "2022-03-15", "2021/12/01", "10 Aug 2020", "15-08-2021", " 15-08-2021"
    ],
    "PhoneNumber": [
        "9876543210", "98765 43210", "98765-43210", None, "not available",
        "9876543210", "91-9876543210", "987654321", "09876543210", "98765 43210"
    ]
}

df = pd.DataFrame(data)
```



```python
df = pd.DataFrame(data)

# Save raw data (optional)
df.to_csv("messy_indian_customer_data.csv", index=False)

print("Before cleaning:\n")
print(df)

# Filling missing names
df['CustomerName'] = df['CustomerName'].fillna("Unknown")
df['CustomerName'] = df['CustomerName'].str.strip().str.title()

# Fix gender values
df['Gender'] = df['Gender'].str.strip().str.upper()
df['Gender'] = df['Gender'].replace({'M': 'Male', 'F': 'Female', 'FEMALE': 'Female', 'MALE': 'Male'})

# Handle age column
df['Age'] = pd.to_numeric(df['Age'], errors='coerce')
df['Age'] = df['Age'].fillna(df['Age'].median())

# Standardize city names
df['City'] = df['City'].str.strip().str.title()
```



```python
# Clean join date
df['JoinDate'] = df['JoinDate'].fillna("01/01/2020")
df['JoinDate'] = pd.to_datetime(df['JoinDate'], errors='coerce')

# Clean phone numbers
df['PhoneNumber'] = df['PhoneNumber'].fillna("Not Provided")
df['PhoneNumber'] = df['PhoneNumber'].str.replace(r'\D', '', regex=True)
df['PhoneNumber'] = df['PhoneNumber'].apply(lambda x: x if len(x) >= 10 else "Not Provided")

# Remove duplicates
df = df.drop_duplicates()

print("\nAfter cleaning:\n")
print(df)

# Save cleaned data
df.to_csv("cleaned_indian_customer_data.csv", index=False)
print("\nFile saved: cleaned_indian_customer_data.csv")
```



```
Before cleaning:

    CustomerName   Gender   Age    City     JoinDate     PhoneNumber
0    Rahul kumar     Male    28   delhi   2022-03-15      9876543210
1  PRIYA  sharma   FEMALE    31  mumbai   15/08/2021     98765 43210
2     Amit singh   female    35   Patna   2020-07-10     98765-43210
3    Mohd. Ayaan        M    24  delhi    01-01-2023            None
4           None     None  None   DELHI         None   not available
5    rahul kumar     MALE    28   delhi   2022-03-15      9876543210
6   Pooja Mishra        F    27   noida   2021/12/01   91-9876543210
7      Ankit raj    Male          Patna  10 Aug 2020       987654321
8     Meena Devi   female    29   Noida   15-08-2021     09876543210
9     Meena devi  FEMALE     29  noida    15-08-2021     98765 43210

After cleaning:

    CustomerName  Gender   Age    City   JoinDate   PhoneNumber
0    Rahul Kumar    Male  28.0   Delhi 2022-03-15    9876543210
1  Priya  Sharma  Female  31.0  Mumbai        NaT    9876543210
2     Amit Singh  Female  35.0   Patna 2020-07-10    9876543210
3    Mohd. Ayaan    Male  24.0   Delhi        NaT  Not Provided
4        Unknown    None  28.5   Delhi        NaT  Not Provided
6   Pooja Mishra  Female  27.0   Noida        NaT  919876543210
7      Ankit Raj    Male  28.5   Patna        NaT  Not Provided
8     Meena Devi  Female  29.0   Noida        NaT   09876543210
9     Meena Devi  Female  29.0   Noida        NaT    9876543210

File saved: cleaned_indian_customer_data.csv
```

THE TRANSFORMATION UNVEILED

The Impact: From Chaos to Clarity

Before Cleaning

- Disparate sources and formats.
- High error rate and data redundancy.
- Untrustworthy for decision-making.

After Cleaning

- Single, unified source of truth.
- Standardised and validated entries.
- Reliable for strategic insights.

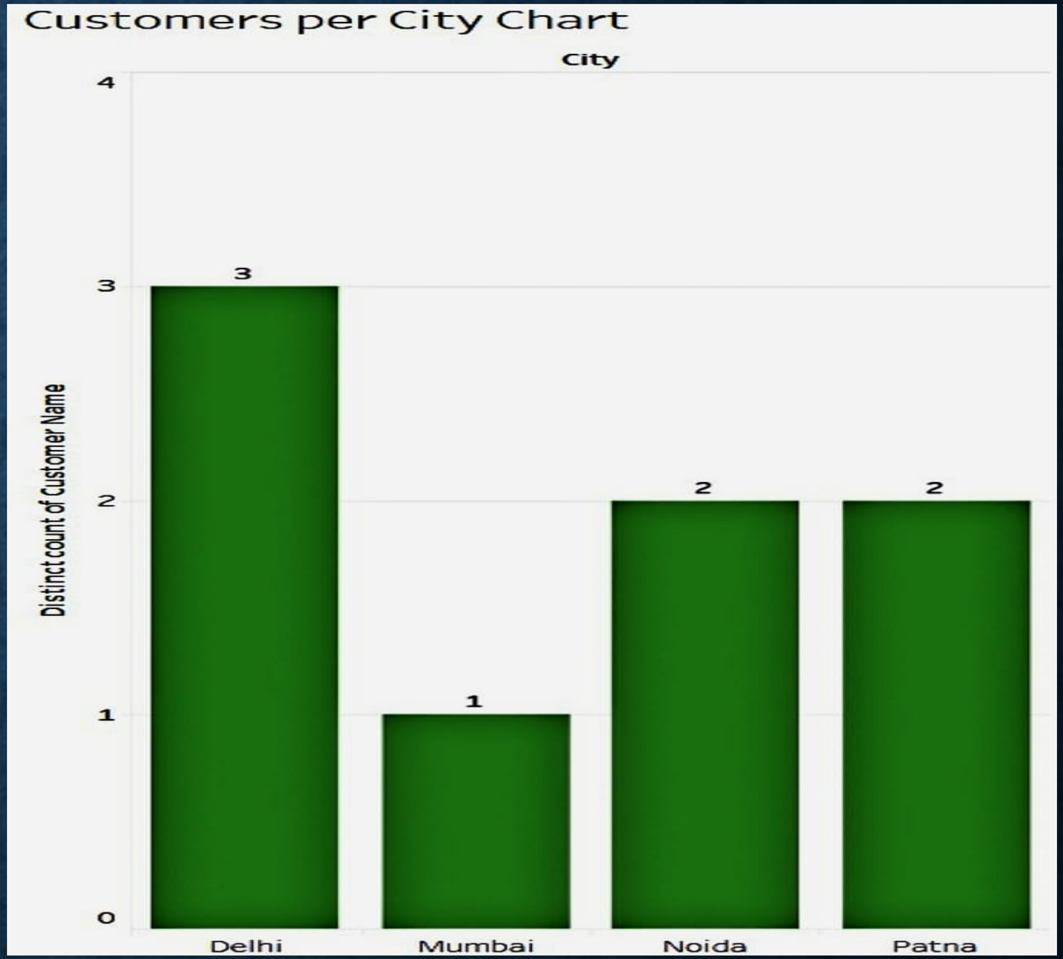
The script's successful execution established a single source of truth, making our data structured, reliable, and ready for accurate analysis.

KEY CUSTOMER INSIGHTS

Report 1: Where Are Our Customers?

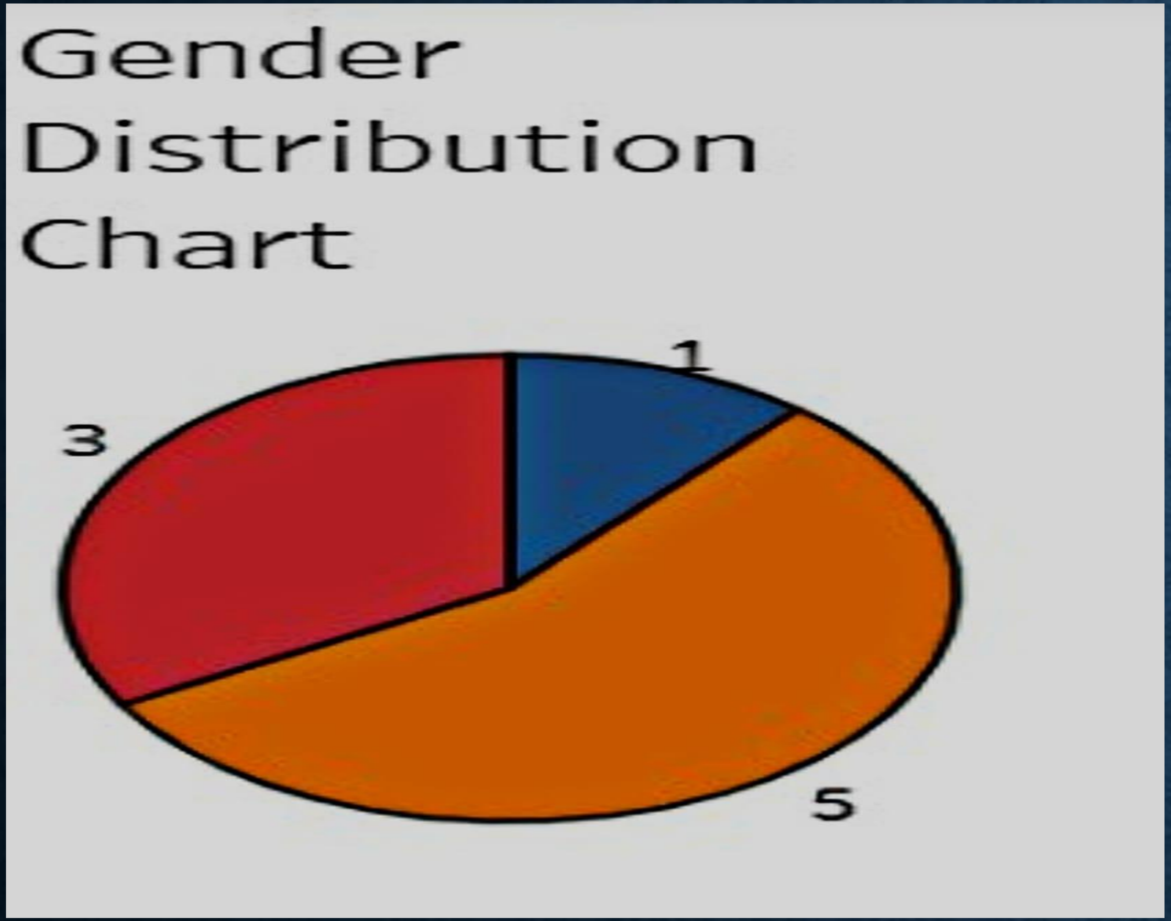
Geographic Distribution:

- Delhi emerges as the dominant customer hub.
- Noida and Patna identified as significant secondary markets.
- These insights guide targeted regional marketing and expansion efforts.



DEMOGRAPHIC INSIGHTS

Report 2: Gender Breakdown



Customer Gender Profile:

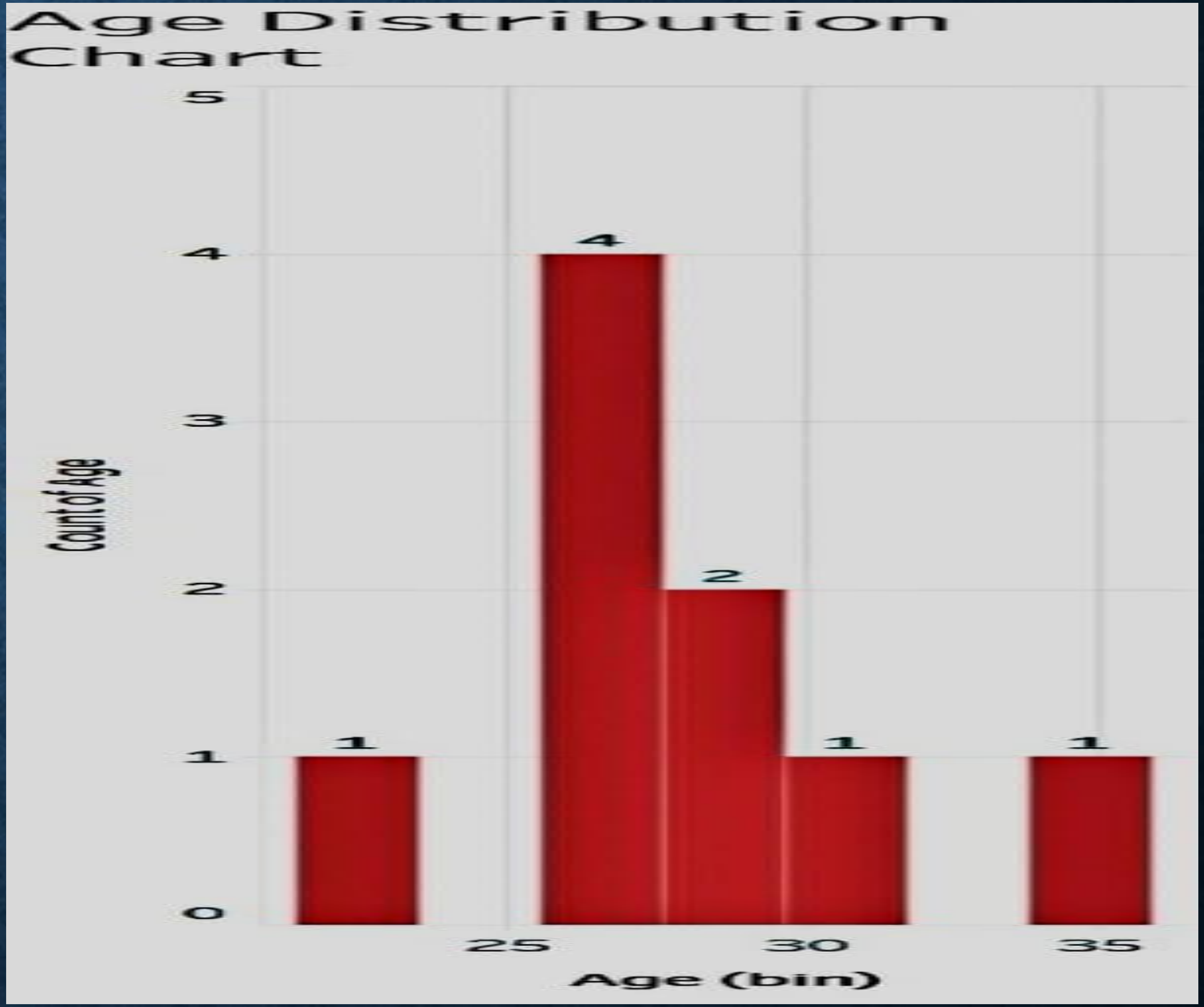
- Our customer base is predominantly female.
- This demographic represents a key opportunity for tailored marketing campaigns.
- Insights suggest focusing product development and communication strategies on this segment.

FURTHER DEMOGRAPHIC ANALYSIS

Report 3: Customer Age Profile

Age Distribution:

- The core of our customer base is young.
- Highest concentration observed between 26-28 years old.
- This insight is critical for developing age-appropriate products and marketing content.



PROJECT CONCLUSION

Project Summary & Key Learnings

Automated Data Cleaning

Successfully implemented Python for efficient and accurate data cleaning, reducing manual effort.

Visualized Key Metrics

Utilised Tableau to transform raw data into clear, actionable visual insights for stakeholders.

Reliable Data Asset

Converted inconsistent data into a trustworthy resource for informed business decision-making.

Accessible Insights

Created an interactive Tableau Public dashboard for easy exploration of cleaned data and reports.

Explore the Live Dashboard

Click here to interact with the full dashboard on Tableau Public