Visa Rejection and Tourism Analysis

This project explores the relationship between **visa rejection rates** and **tourism volume** using two datasets: one containing visa statistics and another detailing tourism metrics for various countries.

Project Steps

1. Data Cleaning and Preprocessing

- Merged two datasets on country after standardizing names (lowercased, trimmed).
- Converted categorical values for Safety and Cost of Living to numerical levels:
    - Low = 1
    - Medium = 2
    - Medium-High = 3
    - High = 4
- Converted Approximate Annual Tourists from strings like "14 million" to actual float values.
- Dropped rows with missing or invalid data in key columns.

2. Exploratory Data Analysis (EDA)

- **Visa Refusal Rate Histogram**:
    - Right-skewed distribution.
    - Most countries have **low refusal rates**.
    - Indicates lenient visa policies in general, with a few outliers.
- **Scatter Plot: Visa Refusal Rate vs Tourist Volume**:
    - Weak negative trend.
    - High refusal rates correlate with lower tourist counts.
    - Outliers exist, but overall direction aligns with hypothesis.
- **Correlation Heatmap**:
    - Weak correlation (~ -0.02) between visa refusal rate and tourist volume.
    - Stronger internal correlations among visa metrics (e.g., issued vs. not issued).

3. Hypothesis Testing

- Conducted a two-sample **t-test** comparing tourist volumes:
    - **Group 1:** Countries with high visa refusal rates (above median).
    - **Group 2:** Countries with low visa refusal rates (below median).

- **Results:**
  - T-statistic = -17.14
  - P-value = 1.03e-65
- **Conclusion:**
  - The difference in tourist volume between the two groups is **statistically significant**.
  - Lower visa refusal countries attract **significantly more tourists** on average.

Key Insights

- Visa policy is **not the sole factor**, but a **significant one** influencing tourist inflow.
- Countries aiming to boost tourism should consider more flexible visa procedures.
- Clean preprocessing and variable transformation are critical for valid analysis.

Technologies Used

- Python (Pandas, NumPy, Matplotlib, Seaborn, SciPy)
- Jupyter Notebook

Future Improvements

- Expand analysis with regression including more variables (Safety, Cost of Living).
- Introduce additional datasets (e.g., economic indicators, travel advisories).
- Deploy interactive visualizations via Dash or Streamlit.