# 🧪 NYC Yellow Taxi Fare Statistical Analysis

This project performs statistical analysis on NYC Yellow Taxi data (January 2020) to identify whether there's a significant difference in average fare based on payment type (`Card` vs `Cash`). The analysis involved data cleaning, outlier removal, visualizations, and hypothesis testing.

---

## 📂 Dataset Used

- \*\*Source:\*\* NYC Taxi Trip Record Data

- \*\*File:\*\* `yellow\_tripdata\_2020-01.csv`

- \*\*Columns used:\*\* `passenger\_count`, `payment\_type`, `fare\_amount`, `trip\_distance`, `tpep\_pickup\_datetime`, `tpep\_dropoff\_datetime`

---

## 🔧 Steps Performed

### 1. \*\*Data Preprocessing\*\*

- Converted `tpep\_pickup\_datetime` and `tpep\_dropoff\_datetime` to datetime format.

- Calculated trip `Duration` in minutes.

- Selected relevant features for analysis.

- Removed nulls and duplicates.

- Filtered out invalid entries (e.g. zero/negative fares, distances, durations).

- Replaced payment types:

- `1` → `'Card'`

- `2` → `'Cash'`

- Restricted passenger count to 1–5 and valid payment types.

---

### 2. \*\*Outlier Treatment\*\*

- Applied IQR-based filtering on:

- `fare\_amount`

- `trip\_distance`

- `Duration`

---

### 3. \*\*Exploratory Data Analysis (EDA)\*\*

- \*\*Histograms:\*\* Compared `fare\_amount` and `trip\_distance` across `Card` vs `Cash` payments.

- Observed general trends and spread of fare distribution across payment types.

---

### 4. \*\*Statistical Hypothesis Testing\*\*

\*\*Objective:\*\* Determine if there's a statistically significant difference in average fare between card and cash payments.

- \*\*Null Hypothesis (H₀):\*\* No difference in average fare between customers using credit card and those using cash.

- \*\*Alternate Hypothesis (H₁):\*\* There is a difference in average fare between card and cash payments.

\*\*Test Used:\*\* Welch’s t-test (independent t-test with unequal variances)

t\_stats, p\_value = stats.ttest\_ind(a=card\_sample, b=cash\_sample, equal\_var=False)

- \*\*T-Statistic:\*\* 169.21

- \*\*P-Value:\*\* 0.0

✅ \*\*Conclusion:\*\*

Since the p-value is less than 0.05, \*\*we reject the null hypothesis\*\*.

There is a statistically significant difference in the average fare between customers who pay by card and those who pay by cash.

---

## 📊 Tools & Libraries

- `pandas`, `numpy`

- `matplotlib`, `seaborn`

- `scipy.stats`

- Jupyter Notebook (for analysis and documentation)

---

## 📁 Project Structure

Statistics\_Project/

│

├── yellow\_tripdata\_2020-01.csv # Raw dataset

├── Statistics.ipynb # Notebook with full analysis

├── README.md # Summary of the project