Titanic Dataset Engineering Pipeline

The Titanic dataset can be accessed via: `https://bit.ly/titanic-data`

1. Ingest

> Stream raw data from the cloud repository.
> `df = pd.read_csv('https://bit.ly/titanic-data')`

2. Data Profiling

> Inspect raw signals to identify data quality issues like missing values and statistical skewness.
> `df.describe()`

3. Scrubbing

> **Drop:** Remove features with >70% missingness (e.g., 'Cabin').
> `df.drop(columns=['Cabin'], inplace=True)`
>
> **Numerical Imputation:** Fill 'Age' with the median to preserve central tendency.
> `df['Age'].fillna(df['Age'].median(), inplace=True)`
>
> **Categorical Imputation:** Fill 'Embarked' with the mode.
> `df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)`

4. Refining (IQR Filtering)

> Remove extreme outliers using the Interquartile Range (IQR) method.
> `Q1 = df['Fare'].quantile(0.25)`
> `Q3 = df['Fare'].quantile(0.75)`
> `IQR = Q3 - Q1`
> `df = df[(df['Fare'] >= Q1 - 1.5*IQR) & (df['Fare'] <= Q3 + 1.5*IQR)]`

5. Export

> Convert categorical text to numeric dummy variables for machine learning compatibility.
> `df_final = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=True)`