🔍 Step 1: Data Exploration – Key Findings

🔍 Step 1: Data Exploration – Key Findings

- The `pages_visited` column in the last year's dataset was stored as a **decimal**, while in this year's data, it was an **integer**. We corrected this to ensure consistency across both.
- The campaign column appeared in two formats:
  - All uppercase ("FALSE") in one dataset.
  - All lowercase ("false") in another.
  - We assume the campaign flag indicates whether the user came from a marketing campaign.
- Most columns across datasets were originally of type string, even for numerical or boolean values.
- We standardized the data types as follows:
  - pages_visited: string → integer
  - campaign: string → boolean

📊 Quick Stats:

- New visitors (`web_new_visitors`):
  - 26.4% (1,216) were exposed to the campaign
- Last year's visitors (`web_last_year`):
  - 26.1% (2,612) were campaign-influenced
- New customers (`new_customers`):
  - 67.7% are women
  - 32.3% are men
  - The most frequent birthdate: 12/30/1899 (14.1%) — likely a placeholder/default

✅ Step 2: Data Cleaning & Merging

✅ Step 2: Data Cleaning & Merging

- Geolocation data was derived from `IP` addresses, creating new columns (e.g., `ip_country`, `ip_geopoint`).
- Excess or irrelevant columns were removed from both `web_last_year` and `web_new_visitors`.
- Standardized the spelling and case of the column `Campain` → `campaign`.
- Unified data types:
  - `pages_visited` converted to numeric data type
  - `campaign` cleaned to consistent boolean format
- Stacked `web_new_visitors` with `web_last_year` to create a complete visitor dataset.
- In `new_customers`, the first item price was used to create a revenue column (since it's their only known transaction). -- this was later changed in next steps
- Stacked `new_customers` with `customers_last_year` into a unified dataset:
  - `all_customers_prepared_with_revenue`.

Step 3: Feature Engineering & Enrichment

Step 3: Feature Engineering & Enrichment

- Created a new column age based on birthdate.
- Birthdates were parsed and transformed into age using parsing.
- Unrealistic ages (e.g., 125 years old) were replaced with the median age: 48.
- Joined datasets as follows:
- Left join: `all_customers_prepared` ⬅ `web_visitors_prepared on customer_id`
- Inner join: Resulting dataset ⬅ `SETUPDATA.country_gdp on ip_country` = `Country`
- Had to rename `ip_country` to `Country` to make the join functional.
- Capitalization was ignored during country matching.
- Finally, removed the manually created `revenue` for `new_customers`, since it's what we aim to predict (target variable).

⚙️ Step 4: Predictive Modeling – Customer Lifetime Value (CLV)

To predict CLV (1-year revenue), we followed these steps:

- Split the dataset into two subsets:
  - revenue_known: Customers with historical revenue (training set)
  - revenue_unknown: New customers without revenue (scoring set)
- Trained a regression model on revenue_known with revenue as the target.
- We experimented with multiple algorithms; Random Forest yielded strong results.
- No data quality issues flagged by Dataiku during diagnostics.
- The scatter plot for the Random Forest model showed strong predictive accuracy, with predicted values closely aligned to actuals along the regression line.

□ Step 5: Dashboard Creation

Our dashboard visualizes both customer insights and predictive model results. Key visual elements include:

□ 1. Customer Insights Dashboard

This dashboard provides a high-level overview of customer characteristics, geographic distribution, and revenue-related patterns derived from our combined dataset.

□ Key Performance Indicators (KPIs):

```
•    Average Customer Revenue
•    Average Customer Age
•    Total Number of Customers
•    Number of Countries Represented
```

□ Charts Included:

```
•    Purchase Count by Gender: Displays the total number of purchases segmented by gender.

•    Average Revenue by Region: Highlights how average customer revenue varies across
different countries.

•    Average Revenue by Age and Gender: Shows a dual-segmented distribution to identify trends
across age groups and gender.

•    Treemap – Revenue by Country: Visually emphasizes which countries contribute most
significantly to total revenue.
```

□

□ 2. Machine Learning Dashboard – Revenue Prediction (Regression Model)

This dashboard focuses on the performance of our regression model trained to predict customer lifetime revenue.

□ Charts Included:

```
•    Histogram: Absolute Feature Importance: Based on SHAP (Shapley) values, this chart
illustrates the average influence of each feature on the model's predictions. The values are
calculated on the test dataset, using the absolute mean SHAP values across all features.

•    Scatter Plot: Predicted vs Actual Revenue: Visualizes the model's predictive accuracy by
comparing predicted values against actual revenue. Points closer to the diagonal line
```

represent better predictions.

- Histogram: Error Distribution: Displays how the model's prediction errors are distributed.

Errors are clipped between the 2nd percentile (-83.673) and 98th percentile (79.522) to reduce the influence of outliers. The goal is a distribution centered around zero with limited spread — ideally forming a bell curve indicating normally distributed residuals.

 Metrics & Assertions:

- Full evaluation metrics (e.g., RMSE, MAE, R^2) are displayed to assess model performance.
- Assertions and interpretability tools such as SHAP explanations were used to verify consistency and fairness of the model.

Project Insight and Analysis

 Customer Insights

 Key Charts:

- Avg Revenue by Region: Highest in North America and Western Europe
- Revenue by Age & Gender: Peaks at age 61–70, with gender-based differences
- Purchase Count by Gender: More female users, but revenue varies
- Treemap – Revenue by Country: Top contributors: US, China, Japan

 ML Dashboard – Revenue Prediction

A regression model was built to predict customer revenue.

 Performance:

- R^2 Score: 0.794
- RMSE: 31.5
- MAE: 18.2
- Pearson Correlation: 0.89

 Insights:

- Top Features: First item price, GDP per capita, age, and campaign exposure
- Predicted vs Actual: Strong alignment, minimal errors
- Error Distribution: Bell-shaped, centered around 0, with low average bias

 Conclusion:

Revenue is strongly influenced by early purchase behavior and customer context (age, location, exposure). The model shows strong predictive accuracy and can support business decision-making for CLV targeting.