Summary of My Notebook Workflow (10 Key Points)

1. **Initial Exploration:** Started by loading the dataset and inspecting its structure to understand what each column represented. Confirmed that each row corresponded to a single customer rather than individual flights.

2. **Identified Key Variables:** Focused on `DistanceKM`, `NumFlights`, `Income`, `Customer Lifetime Value`, and `CancellationDate` as the main columns for analysis.

3. **Created Two Subsets:** Split the dataset into two groups — customers with cancellations (`subset_cancellations_only`) and those without cancellations (`subset_cancellations_na`). This separation allowed clear comparison between the two types of customers.

4. **Checked Missing Values:** Counted missing values properly using `.isna().sum()` to identify where data was incomplete, especially in the numeric columns.

5. **Median Imputation:** Replaced missing values for `Income` and `Customer Lifetime Value` with their respective medians in the non-cancellation subset. This ensured no customer was dropped and reduced skew from extreme values.

6. **Consistency Checks:** Verified that columns were already numeric and that there were no negative or unrealistic values for flight distance or number of flights.

7. **Outlier Filtering:** Applied the interquartile range (IQR) method to trim extreme values in key numeric fields such as `NumFlights` and `DistanceKM`, ensuring a cleaner dataset for comparisons.

8. **Visual Comparisons:** Produced simple, consistent visualizations (like boxplots and histograms) to compare distributions of income, lifetime value, and distance between customers who cancelled and those who didn't.

9. **Added `HasFlown` Flag:** At the end of the process, created a binary indicator showing whether each customer had ever flown, based on `DistanceKM` values greater than zero.

10. **Overall Workflow Logic:** The cleaning followed a clear order — inspect → split into subsets → handle missing values → trim outliers → visualize → add `HasFlown`. Each step built on the previous one to keep the process transparent, consistent, and faithful to the actual notebook sequence.