Step 1:

Python
I have started out this project with python.

I approached this by cleaning each of the four data files individually before merging them and performing final table validations.

1. Cleaning the Source Files

- **For the checkouts.csv data**, I first implemented a custom function to handle a wide variety of messy date formats, including those with special characters or inconsistent separators. After standardizing the dates, I applied a series of validation rules, removing any records with future checkout/return dates, checkouts from before 2005(to ensure modern, reliable data,) or illogical transactions where the return occurred before the checkout. I made the strategic decision to fill missing return dates with the current date, operating under the logical assumption that these books are still on loan. Finally, I engineered two critical features for the analysis: loan_duration_days and a status to classify each loan as "On Time" or "Late".
- **For the customers.csv data**, my goal was to standardize demographic and geographic information while adopting a flexible approach to handling data quality issues.
    - A key strategic decision was to create a data_quality column rather than simply dropping records with errors. This flag identifies patrons with problematic birth dates (missing, in the future, or from before my business rule cutoff of **1930**).
    - I thoroughly cleaned all text fields, including implementing a custom function to correctly format street_address by capitalizing common abbreviations like "NE" and "SW".
    - The zipcode was validated to ensure it was a 5-digit number, with any invalid entries being clearly marked as "Unknown".
- **For the books.csv data**, I standardized all book-related attributes. I cleaned the id column by removing leading = signs to prevent data corruption in Excel. To handle inconsistent date formats in the publishedDate column, I created a clean publishedYear column, which preserved the original data while enabling accurate validation. I also cleaned the price and pages columns by reliably extracting numeric values from messy text, and standardized the categories by removing special characters.

- **For the libraries.csv data**, I performed similar text standardization on all location-related fields, including applying the custom street_address cleaning function to ensure consistency across the entire dataset.

2. The Final Merge and Validation

Once each file was individually cleaned, I merged them into a single master dataset. To ensure the final file was clean and easy to use in Power BI, I proactively renamed conflicting column names (e.g., name to customer_name and library_name).

This final, merged dataset allowed for two crucial cross-table validations:

1. I removed any records where a patron's checkout date occurred before their birth date.
2. I removed any records where a book was checked out before its publication year.

The result of this comprehensive process is a single, fully validated file, master_library_data.csv, which serves as a trustworthy foundation for all the analysis and visualizations in the Power BI dashboard.

Step 2:

PowerBI

Calculated Columns for Deeper Segmentation:

- **Age Group:** Patrons were grouped into logical age brackets ("Under 18", "18-29", "30-44", "45-59", "60+") to analyze behavioral patterns across different life stages.
- **Book Length & Price Range:** I binned books into clear categories like "Standard," "Long," and "Epic" for page count, and "Budget," "Standard," and "Expensive" for price. This transformed raw numbers into powerful analytical dimensions.
- **Category Group:** To clarify the analysis, I consolidated the overly granular book categories into broader, more meaningful themes.
- **Late Category:** To measure the *severity* of late returns, I created a performance metric that classified each late book from "1-2 Weeks Late" to "3+ Months Late." Based on the high probability of non-return, I strategically labeled any book over 90 days late as a **“Lost Book”**.

Key Findings: The Drivers of Late Returns

My analysis of the cleaned data revealed several distinct and actionable factors that are strongly correlated with late returns.

Geographic and Demographic Drivers:

- **Distance is a Key Factor:** There is a clear geographic pattern. The late return rate is significantly higher in the outskirts of the city, strongly indicating that the physical distance and inconvenience of traveling to a branch is a primary driver of late returns.
- **Gender Shows a Significant Divide:** Male patrons are **more than twice as likely** to return books late, with a late rate of **17.86%** compared to **9.09%** for female patrons. This is one of the most statistically significant demographic findings in the dataset.
- **Age Plays a Critical Role:** The **60+ age group** exhibits one of the highest late return rates at **11.43%**. Patrons **under 18** also show an elevated rate of **15.79%**, suggesting that both the youngest and oldest patrons face unique challenges affecting their return habits.

Book and Patron Profile Characteristics:

- **High-Investment Books are Returned Later:** Books that require a significant investment of time or money are more likely to be late. "Epic" length books (over 700 pages) and "Expensive" books (over $500) both have a late return rate approaching **19%**.
- **Education and Occupation Show Clear Trends:** Patrons with a **college education** have the highest late return rate at **11.65%**, which may suggest these books are used for extended learning or research. Similarly, patrons in **Business (12.66%)** and **Tech (10.53%)** are the most frequently late, reinforcing the hypothesis that books are being used for professional purposes.
- **Niche Categories Have High Rates:** While representing a smaller volume, specialized academic or professional book categories like **Agriculture (31.25%)** and **IT (20%)** have the highest late return rates of all, pointing to a mismatch between the standard loan period and the patrons' needs for these materials.

Seasonal Trends:

- **Summer Slowdown:** The analysis confirmed a seasonal trend. Both the total number of checkouts and the late return rate are lower during the summer months, presenting a strategic opportunity for the library to reallocate resources or launch engagement campaigns during this quieter period.

My analysis of over **1,000 checkouts** reveals an overall late return rate of **9.82%**. While many of these are only a few days overdue, a significant portion are late by more than three months. I have strategically categorized these as **“Lost Books,”** representing a tangible loss to the library's collection and a key area for intervention.

From there, I would pivot to the core of the analysis—the "why" behind this number—by focusing on three key themes:

1. Geography and Convenience are the Strongest Predictors: My most significant finding is that late return rates are substantially higher in the city's outskirts. This creates a clear geographic "hotspot" on the map, strongly indicating that the physical distance and inconvenience of travel are primary factors influencing patron return behavior.

2. Patron Demographics Reveal Clear Patterns: The data shows that specific demographic groups have distinct return habits. The **60+ age group**, for example, has one of the highest late return rates. This suggests that this specific cohort could benefit from targeted outreach and more accessible return options.

3. The Type of Book Plays a Decisive Role: The data clearly shows that not all books are treated equally by patrons.

- **High-Investment Content:** Longer books (over 700 pages) and more expensive books (over $500) both have a late return rate approaching **19%**. This suggests patrons are more hesitant to part with items that require a greater investment of time or money.
- **Specialized and Academic Material:** Certain professional categories, like Agriculture and IT, have the highest late rates. This, combined with the finding that patrons with a college education are most likely to be late, supports the hypothesis that these books are being used for extended learning projects where the standard loan period is insufficient.

Based on these findings, we can confidently move from a one-size-fits-all approach to a more targeted, data-informed strategy that addresses the specific needs of different patrons and materials.

Step 3:

Actionable Recommendations

My recommendations are designed to be practical, measurable, and to directly address the key drivers of late returns identified in the analysis.

Recommendation 1: Implement a Proactive, Automated Reminder System (The Foundation)

- **The Finding: A significant portion of late returns across all demographics is likely due to simple forgetfulness. A timely reminder is the most effective way to mitigate this primary cause.**
- **The Solution: The first and most critical step is to implement an automated communication system that sends a friendly reminder to patrons before their books are due.**
    - **Primary Method (SMS): A text message reminder sent three days before the 28-day loan period ends. This is a direct, immediate, and highly visible notification.**
    - **Secondary Method (Email): For patrons who have an email on file, the SMS can be supplemented with a more detailed email reminder.**
    - **The Goal: To reduce the number of accidental late returns across the entire patron base with a low-cost, high-impact solution.**

Recommendation 2: Mitigate Geographic Barriers with a "Return to Any Branch" Policy

- **The Finding: The data clearly shows that patrons living in the city's outskirts have a significantly higher late return rate.**
- **The Solution: After establishing the reminder system, the next most impactful step is to remove the friction of travel. By allowing patrons to return a book to any library branch, regardless of where it was checked out, we directly address the inconvenience faced by our geographically distant patrons. This is a powerful, customer-centric policy that will reduce late returns and build significant goodwill.**

Recommendation 3: Pilot a "Library Concierge" Service for Senior Patrons

- **The Finding: The 60+ age group is a key at-risk demographic, and standard digital reminders may not be effective or address potential mobility challenges.**

- **The Solution:**
    - **Non-Digital Reminders: For senior patrons, offer an opt-in for an automated phone call reminder or a mailed postcard, ensuring the message reaches them effectively.**
    - **At-Home Service Pilot: To address potential accessibility issues, I recommend piloting a "Library at Home" service. This would offer a scheduled, at-home pickup and drop-off service, positioning the library as an essential community service that cares for its senior patrons. This could be tested in partnership with local senior centers to minimize operational costs.**

Recommendation 4: Introduce Dynamic, Content-Aware Policies

- **The Finding: Long, expensive, and specialized academic books are all returned late at a much higher rate, suggesting the standard loan period is insufficient for these materials.**
- **The Solution:**
    - **Smart Policies: For books over 700 pages or in high-risk categories like "Agriculture," pilot an automatic extended loan period of 35 days.**
    - **"Lost Book" Intervention: For books that fall into the “Lost Book” category (over 3 months late), trigger an automated message offering a clear path to resolution, such as returning the book with a partial fee waiver or paying the replacement cost. This turns a negative outcome into a manageable one.**