Table of Contents

Overview and Motivation:

The goal of this project is to visualize WPI post office package data over a 30-day period to gain insights into package processing efficiency, storage patterns, and potential bottlenecks. By analyzing timestamps from package routing to storage and final delivery, we can track how long packages remain in different stages and identify areas where delays may occur. The dataset, sourced directly from the WPI post office, excludes recipient names for privacy protection while retaining key details such as locker assignments, carriers, and processing times. Understanding these trends will help improve package handling efficiency, optimize locker usage, and provide a clearer picture of how the campus mail system operates. Using visualizations, particularly a Sankey diagram, we can illustrate the flow of packages through the system, highlighting areas where processing times could be improved to enhance the overall experience for WPI students and staff.

Database Cleaning and Processing

The original dataset from the WPI post office contained package tracking information spanning 30 days, capturing key details such as tracking numbers, locker assignments, carrier names, and timestamps for different stages of processing. However, the raw data also included inconsistencies, missing values, and unstructured fields that required preprocessing before analysis. Some columns, such as "Notes", were unnecessary, while others, like "Routed Date Time" and "Stored Date Time", had missing entries that needed to be addressed. The "Tracking #" field contained trailing underscores, and the "Location 1" column mixed undergrad and grad student package assignments without a clear distinction. Additionally, the "Origin City" and "Origin State" fields were incomplete in some cases, affecting the accuracy of geographic analysis. To prepare the dataset for visualization, we performed several cleaning and processing steps to ensure consistency and reliability. Missing values in critical columns were filled, tracking

numbers were standardized, and a "Bank_Locker" identifier was created by combining locker bank and locker number to facilitate storage analysis. The timestamps for when packages were routed, stored, and delivered were converted into datetime format, allowing us to compute processing durations accurately. To enhance geographic tracking, the dataset was merged with an external city-to-county mapping file, filling in missing "Origin County" values where possible. There were many entries with proper city-state information that weren't given a county column entry after this process. To fix this, a new python script was created that would load in the CSV and use the “geopy” library to retrieve the county for a city-state tuple. These counties were saved to an output CSV as well as a temporarily stored dictionary. This dictionary sped up the process, bypassing the need to call to “geopy” if the county was already found. With numerous packages coming from the same cities in Massachusetts this optimization worked wonderfully. After these refinements, the cleaned dataset provided a structured and reliable foundation for analyzing package flow, processing times, and overall post office efficiency.

Meeting 1, 2/25/2025

Related Work:

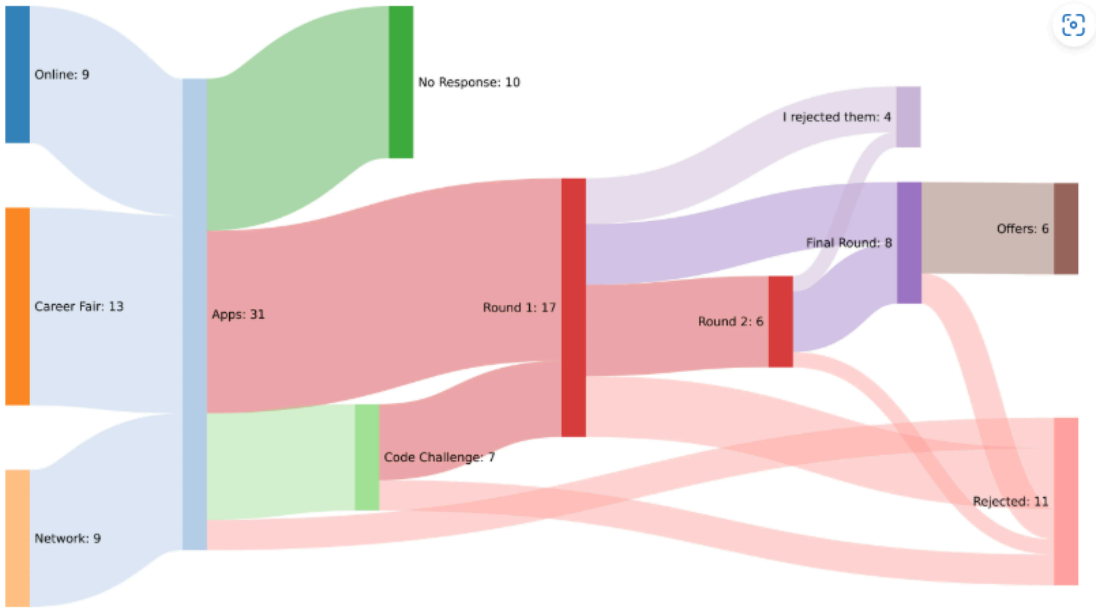


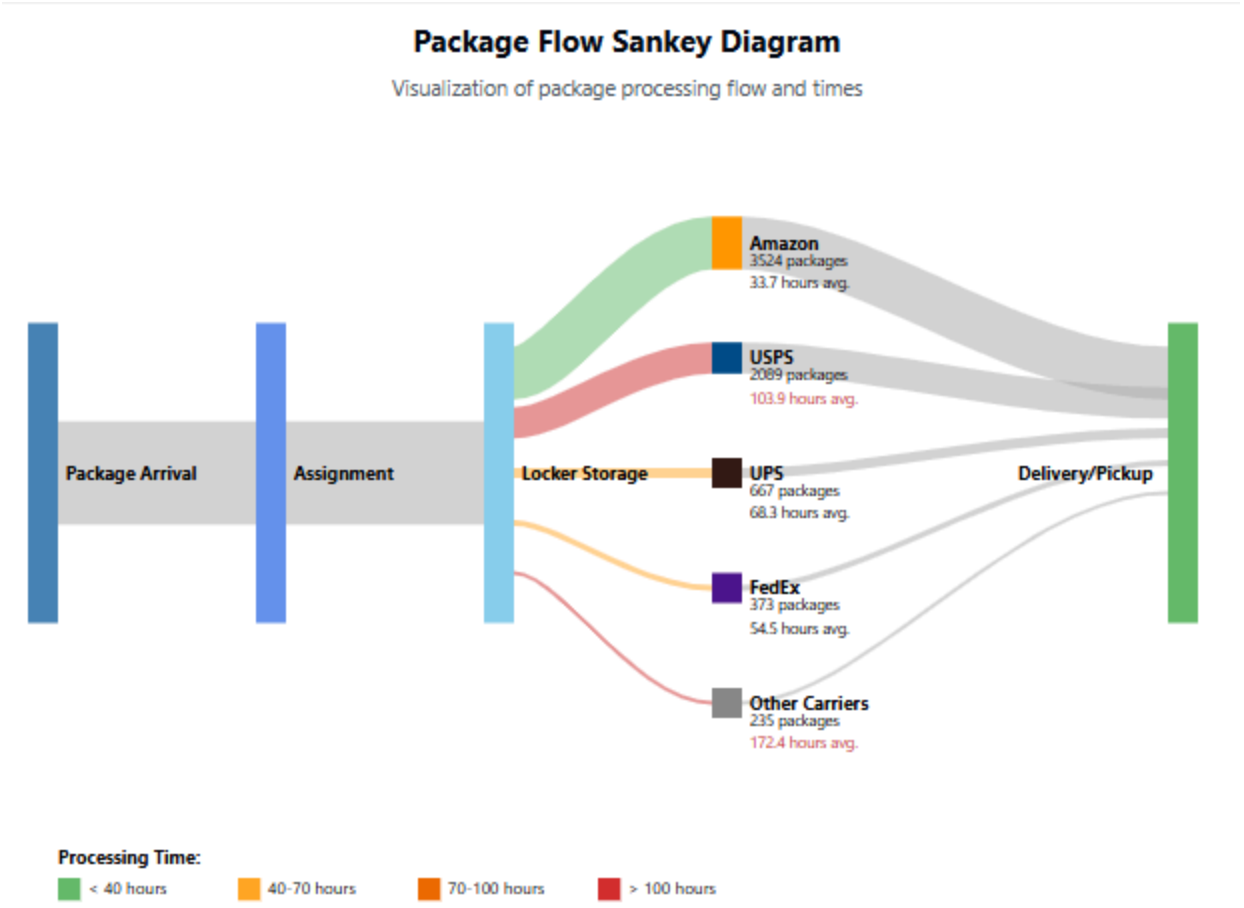
Figure Source: [Reddit - https://external-preview.redd.it/WczAQA-APQXp8f1cACNqYmQZZuR39_WF2ql458v4a14.png?auto=webp&s=d4f55d30c1cd034c31335da0e2ba3ffd7988ac6d]

To visualize the package flow and processing times, we will use a Sankey diagram, referencing the figure we posted above, to represent how long each package stays in different locations before being processed. Our cleaned dataset includes key timestamps: "Routed Date Time", "Stored Date Time", and "Delivered Date Time", which allow us to track the time spent at each stage of the process.

The flow will start from the package's arrival at the facility (Routed Date Time), moving to locker storage (Stored Date Time), and finally to delivery or pickup (Delivered Date Time). Each node in the Sankey diagram will represent a processing stage, including possible paths such as locker assignment, carrier pickup, and customer retrieval. The width of the connections will indicate the volume of packages transitioning between these stages, and the time spent at each stage can be color-coded to highlight delays or inefficiencies.

Additionally, we can incorporate locker bank utilization by breaking down processing times per "Bank_Locker", identifying which locker areas experience the most congestion. If "Carrier" information is available, we can analyze differences in processing times for different shipping companies, helping to optimize workflows and improve package management within the post office system. This visualization will provide actionable insights into bottlenecks and operational efficiency.

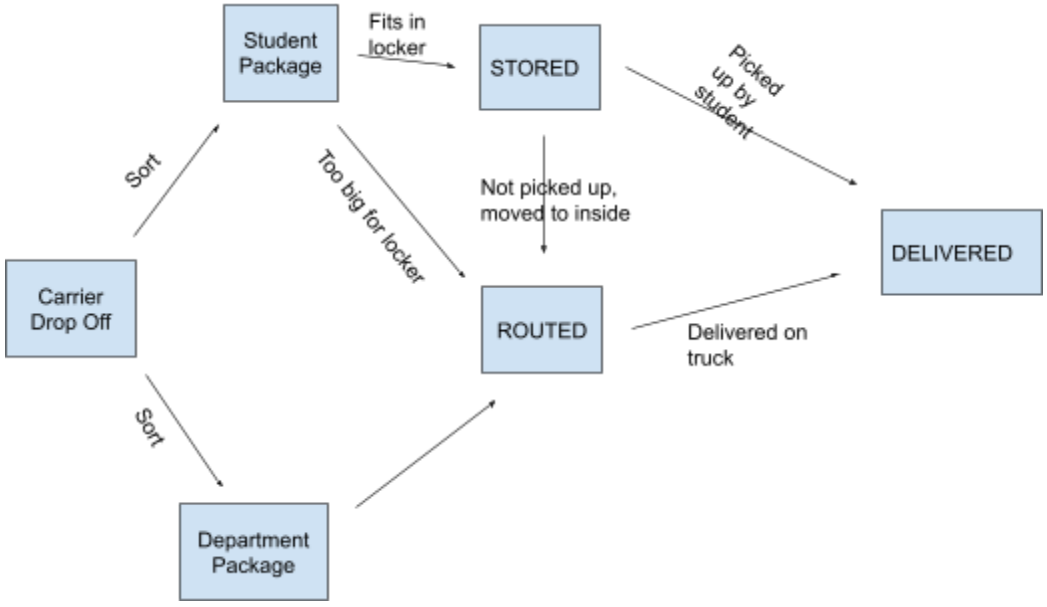
This will be a draft of how our Sankey diagram will look like:



- Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
    - Where do packages come from?
    - What is the breakdown of packages in different categories?
    - When do students pick up packages?
    - How long does it take students to pick up packages?

- Data: Source, scraping method, cleanup, etc.

- The source is the package management software (QTrak)



- Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

- Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?
  - Heatmap of package origins
  - Sankey diagram (all packages split by carrier, student or faculty, class year, etc)
  - Heatmap of pickup times by hour and day of week
  - Histogram of time it takes students to pick up packages

- Implementation: Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

- Evaluation: What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?