Data 607 5A Airplane Delays

Benjamin Bravo

2026-02-27

Approach

To complete this assignment, I will first recreate the airline delay data in a wide-format CSV file that mirrors the original table structure with two airlines, five cities, and counts for on-time and delayed flights. I will then import the dataset into R and use tidyr and dplyr to transform the data from wide format to long format so that airline, city, flight status, and counts are organized into clear variables. After tidying the data, I will calculate total flights and compute delay percentages both overall and for each city. I will present the results using tables and/or visualizations and include written explanations interpreting the findings. Finally, I will explain the difference between the overall airline comparison and the city-by-city comparison.

Anticipated Data Challenges

One challenge may be correctly restructuring the dataset from wide to long format without mislabeling airlines or cities. I may also encounter formatting issues, such as numeric values being read as characters (especially if they contain commas), which will require cleaning and conversion. Preserving any missing or empty cells from the original structure could also require careful handling. Additionally, calculating percentages accurately will be important, since comparing raw counts alone could lead to misleading conclusions. Finally, explaining the discrepancy between overall results and city-level results may require careful interpretation to clearly demonstrate how aggregated data can produce different conclusions than grouped data.