Arnav Mehta (14132) Gyanesh Gupta (14259)

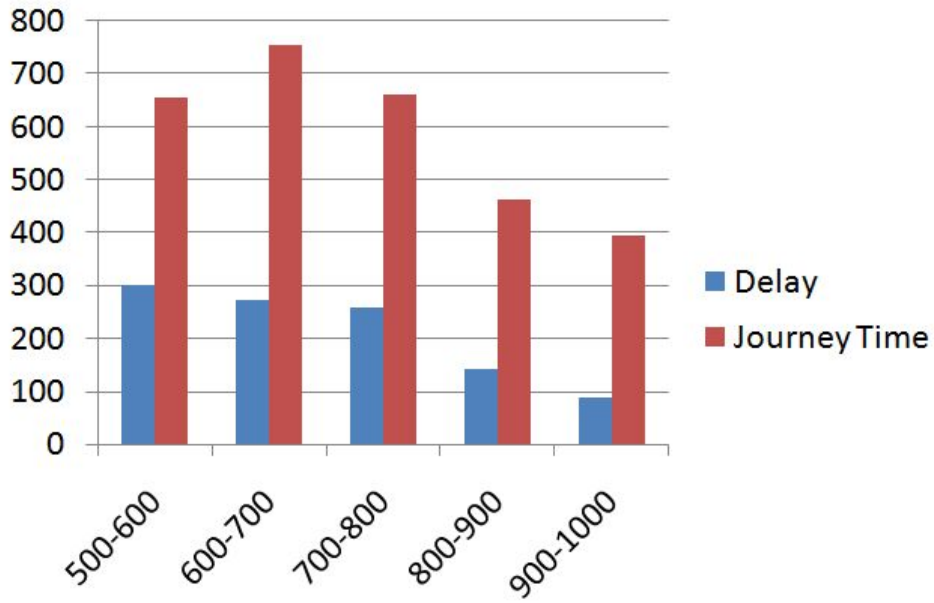
CSS698H
Train Analysis

Introduction

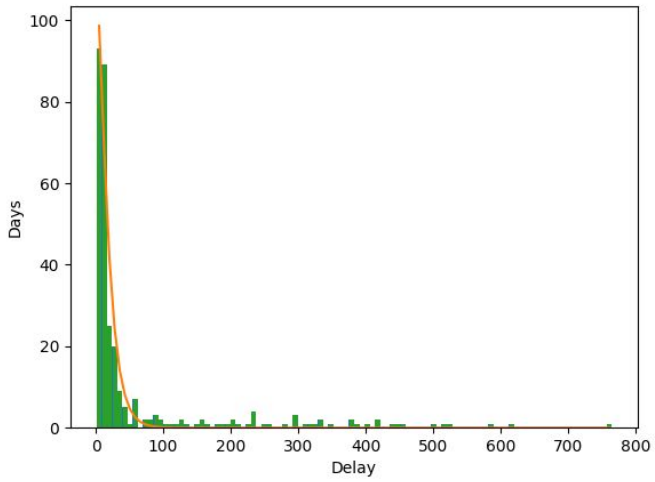
Using the delay data of trains for the year 2016 on the route of Kanpur to Delhi we ranked all trains by their timeliness and duration time. We calculated the confidence interval of our delay estimate. Using this information we the can recommend the optimum train (trains) then someone should book their ticket in once they have decided their ticket fare and arrival time where we have incorporated time of day.

Method and Plots

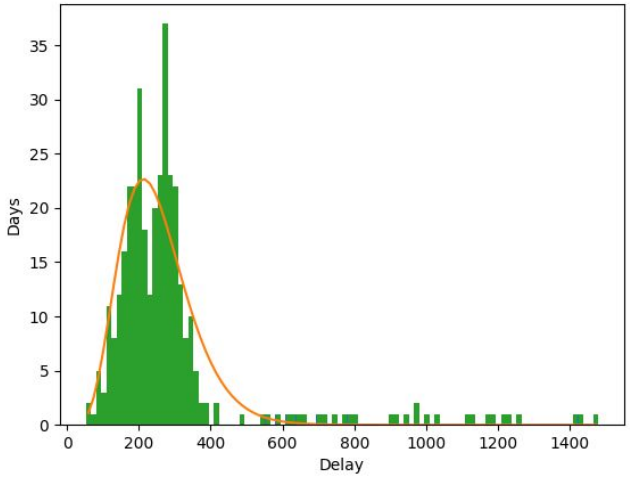
We first scraped data for all 70 trains which travelled between Kanpur and Delhi. We then found the cumulative distribution function for the histogram for each train. That was to be used to ascertain the confidence in our estimate of the delay and ranked on the basis of duration time (see Annexure). We used mean delay as parameter after that. We then fit histograms with curves. Trains that were very reliable showed behaviour similar to exponential distribution. Trains in the middle showed behaviour similar to a gamma function. Some trains showed bimodal distribution. This could be accounted for as often trains which run late get even more late as they are given lower preference for transit compared to trains which run around the right time. Also, based on the graph of day of the month vs. delay, we noticed the obvious trend of trains getting late during winters. Furthermore, trains getting late at other times can be attributed to accidents/track failure. We then made a recommendation program which took a time range of arrival and ticket cost range and recommended the most appropriate trains (which got you to Delhi the fastest) as output.



Fare and Delay Analysis



Train following exponential distribution



Train following gamma distribution

Recommender System

Our recommender system is mainly a python script, which runs another script. It then asks the user to input the range of fares and range of time when he wants to reach. The output is a list of top 5 trains, ranked by the total expected time of journey.

Possible extensions

The system recommends based on the mean delay. As an extension, it can answer based on a suitable confidence interval. Also, the query is day independent, hence we can include the factor of time of the year in it (By taking monthly samples, or data from previous years).