

1. Group: Harshil Jain, Henry Warzecha, Steven Lymperis, Alexander Kahn
2. Data
 - a. Our total dataset consist of 4 CSV files with a combined size of 7.39 GB
 - b. The dataset contains a combined total of 46.2 million rows of data, with each row representing one taxi trip within New York.
3. Hypothesis Testing
 - a. One idea we are looking to test with this dataset is by looking at the payment method used during a taxi ride and using it to classify which regions are dominated by certain payment types. We plan to do this by looking at the longitude and latitude of the pickup and dropoff location of each taxi ride, and then dividing new york city into grid-blocks and looking at which payment type dominates each grid-block. We still need to decide the size of the grid-blocks we will use, but we are thinking of making the area the size of an average neighborhood in new york city.
 - b. Another idea we are looking to explore is looking at the payment for each taxi trip, and seeing if certain regions in new-york charge a higher amount per-mile than other regions, and thus categorizing new york into regions ranging from cheapest taxi rides to most expensive taxi rides.