We began with each of us looking at various datasets and Stephen found a study of LA Traffic Collisions on Kaggle that we agreed was interesting. I traced the datasets in that study to their source, the Los Angeles Data Catalogue: <https://data.lacity.org/>. From there, I downloaded a .csv covering 10 years of data, inclusively, using the Date Occurred for each incident. We reduced that dataset to five years – 2015 to 2019, inclusive.

I created various pivot tables to quickly review whether all the variables provided were necessary or compelling. The data included incidents from 21 LA precincts, and we decided to focus on North Hollywood, because it had many incidents. We filtered ‘Premise Description’ down to the most used option (‘Street’) and eliminated columns with superfluous (‘Crime Code’ and ‘Crime Code Description’) incomplete (‘Victim Age’, ‘Victim Sex’, ‘Victim Descent’) data. The ‘Address’ and ‘Cross Street’ columns contained spacing errors, which I corrected for later use and presentation.

The original .csv was 18 cols x 521,274 rows. The working file we used to begin was 9 cols x 13,079 rows.