**ETL Project Report**

This project is a brief study of traffic collisions in New York City and what sort of events may cause higher or lower numbers of collisions.

I extracted information successfully from three sources and was unsuccessful with a fourth. The first source was catalog.data.gov, where a csv file contained all of the collision data I used. This data was collected by the state of New York and is a compilation of every accident that happened with at least $1000 of damage done. The file came in the form of a CSV and had many columns corresponding to different aspects of the crash. I decided I was mostly interested in the number of crashes happening on particular days, so I trimmed away most of the columns of extraneous information, leaving just the date and times of each crash. I converted the date to datetime and set it as the index. I also made a table representing the total number of crashes by day of the week so I could see which days have more crashes. Friday has the most while Sunday has the least, followed by Saturday.

The next source I used was the Alpha Vantage API. I used the api functionality to extract stock data for the New York Stock Exchange Composite, figuring this would be a good point of comparison to see if spikes or falls in stock data corresponded to highs or lows in collisions in the area. I edited the table to contain only the date and the closing value of the index. I found a very slight positive correlation between closing and number of crashes, meaning that to some very small extent, the market closing higher corresponded to more crashes.

The final source of data I successfully extracted from was retrosheet.org where I was asked to give the following disclaimer in my findings:

The information used here was obtained free of

charge from and is copyrighted by Retrosheet. Interested

parties may contact Retrosheet at "www.retrosheet.org".

This information came also in the form of a large, downloaded CSV file for each year I was looking at. These files contained info for every game so I first narrowed down to Yankees home games. I also added all of the info into one dataframe, so it was not separated by year. I converted the dates to datetime and set the date as the index in order to sort by date. I cut out many of the extraneous columns in order to simply have the date, day of the week, and visiting team. I also made a table for just games in which the Boston Red Sox were the visiting team, in order to see if the Yankees playing their rivals had an affect on collisions. I found that the average number of crashes was higher than normal on game days, but that the number of crashes on days when they played the Sox were lower than average game days.

I attempted to get the baseball information I wanted through HTML scraping on the website baseball-reference.com, but when I dug into their html script, their tables where impossible to find on the page, leading me to believe that they are trying to protect their website from scraping. That said, the info I did use from retrosheet.org is the info that this website bases their information on, so I didn’t manage the diversity in information retrieval, but I did get the same info.

In the end of my jupyter notebook I used sqlite3 to load my tables into a database. I chose SQLite3 for easiness and because it is a relational database. I figured I wanted to use a relational database because I had some tables with multiple columns in my project that I wanted to be able to view in an easy way, and because it’s so easy to port dataframes into the database here. I put the following tables in the database, all of which have been described (though some not by name) in the preceding paragraphs: YANKS\_TRIMMED, YANKS\_VS\_SOX, COLLISIONS\_FULL\_DF, COLLISIONS\_SUMMED, NYSE\_CLEAN, MERGED\_DF, YANKS\_OVERLAP, SOX\_OVERLAP.