**Project Report**

Group Members: David Jackson, Shi Liu, Cassy Myers

Project requirements

1-The sources of data that you will extract from.

2-The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

3-The type of final production database to load the data into (relational or non-relational).

4-The final tables or collections that will be used in the production database.

Project achievements

1. We found vehicle crash data for Cary from <https://data.world/townofcary/cpd-crash-incidents> and downloaded the csv file for the crash data (cpd-crash-incidents.csv). We also found weather data from <https://catalog.data.gov/dataset/local-weather-archive> and download csv file for the local area (rdu-weather-history.csv).

2. We used Jupyter notebook as a platform for data manipulations including reading csv data into Pandas dataframe, cleaned/filtered the two datasets, merging two datasets, and then loading a database server.

3. The database server that we used for receiving the data and for future final production service is PostgreSQL. We created a relational database “Cary\_crash” and three tables: weather, crash, weather\_crash. The records were loaded from respective dataframes created from original csv data files.

4. We have performed some initial analyses on the data using Pandas and plotted analysis results with Matplotlib. These initial analyses will be solidified into final collections for production use.

All of the above activities for the project are saved in a Jupyter Notebook file (data\_pipe10.ipynb). The database schema is saved in a text file (ETL\_Cary\_crash\_schema.txt).