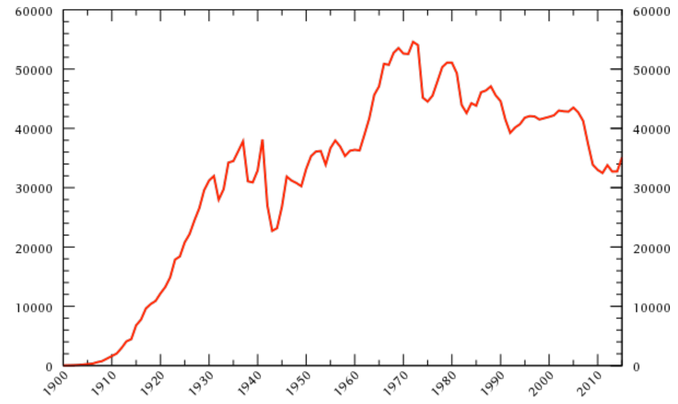
SAFEWAY
ROAD RISK CLASSIFICATION SYSTEM

TEAM 1
HYEMIN HWANG
SONAL JAIN
SCOTT LYNCH
NATANIEL RUIZ
DEVAN LABHA SHANKER

WHAT IS SAFEWAY?

According to NHTSA (National Highway Traffic Safety Administration), there has been a reported **35,092** deaths caused by motor vehicle accidents in 2015, averaging to about **95 DEATHS/DAY**.

With Safeway, we aim to reduce the number of deaths, by providing *local authorities and safety personnel* with data about the risk of a particular road i.e. the possibility of accidents happening on a road, and help them better utilize scarce resources to lower the overall accident rate. We aim to identify and classify high-risk traffic incident areas, enabling timely decision-making and preventative measures.



Motor Vehicle Deaths in the US (Wikipedia)

DATA

We used the Kaggle dataset, UK Car Accidents (2005 - 2015).

Data Fields
longitude, latitude, accident severity, number of vehicles, number of casualties, date, day of week, time, road number, pedestrian crossing, light conditions, weather conditions, road surface conditions, special conditions at site, urban or rural area.

Total Number of Records: 1.7 million
Total Data Size: 250 MB

APPROACH 1

Data Pre-processing
We wrote a preprocessing python script where we dropped columns that are not immediately useful.

Longitude	Latitude	Accident_Severity	Number_of_Vehicles	Number_of_Casualties	Day_of_Week	1st_Road_Class	1st_Road_Number
-0.191170	51.489096	2	1	1	3	3	3218
-0.211708	51.520075	3	1	1	4	4	450
-0.206458	51.525301	3	2	1	5	5	0
-0.173862	51.482442	3	1	1	6	3	3220
-0.156618	51.495752	3	1	1	2	6	0

We then implemented a binning function, where we first set the geographical limits of the data that we want to use. We filter out the data that does not lie in these bounds. We also filtered the data by year.

Pre-processed Training Data

`train_x_df`

	tod_Early-Morning	tod_Morning	tod_Afternoon	tod_Evening	road_type	speed_limit	1st_road_class	2nd_road_class
0	0	1	0	0	6.0	60.0	3.0	-1.0
1	1	0	0	0	6.0	60.0	3.0	-1.0
2	0	0	1	0	6.0	30.0	4.0	6.0
3	0	0	1	0	6.0	60.0	4.0	-1.0
4	0	0	1	0	6.0	30.0	6.0	6.0

`train_y_df`

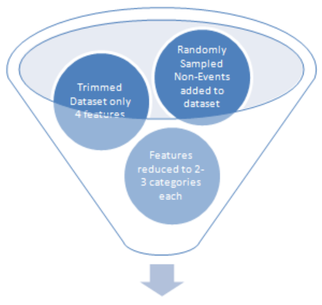
```
0    1.0
1    1.0
2    1.0
3    1.0
4    1.0
```

Predictive Model
We used Random Forest Regression to train on 9 years of data and predict the amount of accidents that will happen on January 2015 for each bin. We then compared the prediction with the actual data, achieving an acceptable mean absolute error.

```
MAE = metrics.mean_absolute_error(pred_y, test_y)
print MAE
```

```
0.15514027171
```

APPROACH 2



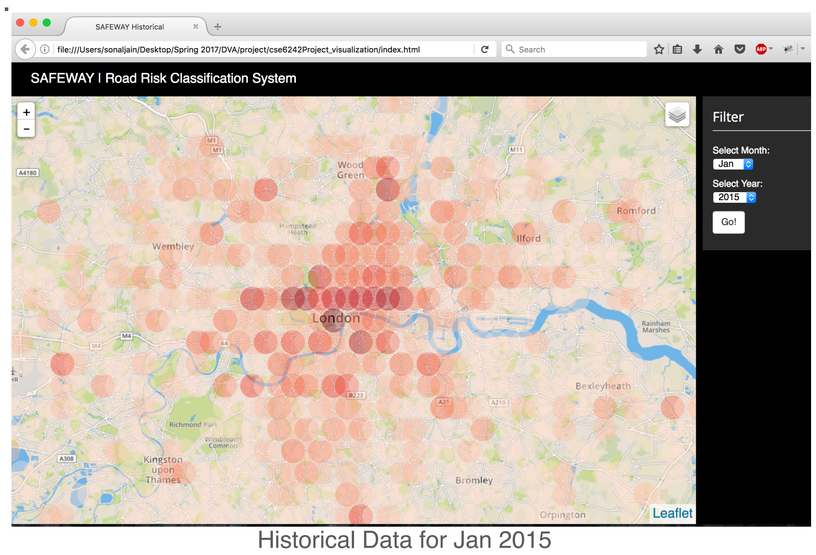
Regression output

$$predicted\ risk = \frac{e^{y'}}{(1+e^{y'})}$$

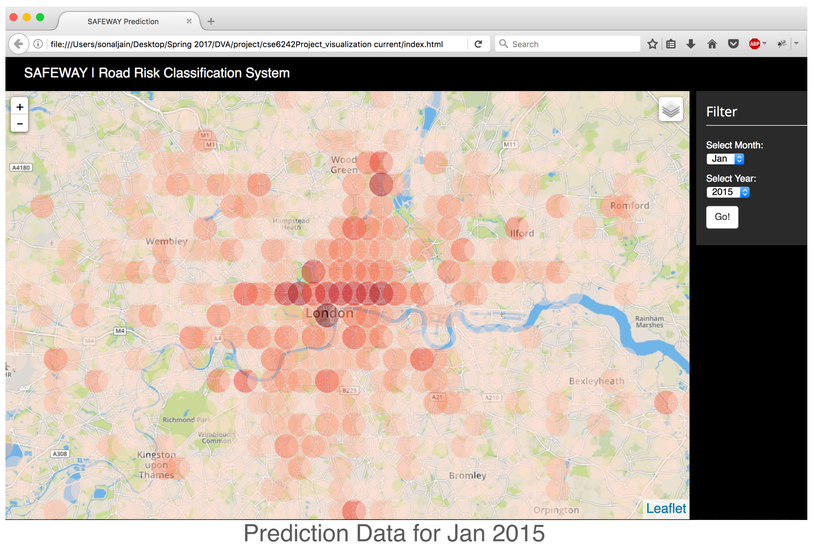
$$y' = -.1711 + .19852 * Congestion_{Rush\ Hour} - 1.02197 * Congestion_{Late\ Night} + 3.2193\ Road\ Type_{Divided\ Road} + 1.1980\ Road\ Type_{Local\ Road} + -2.92313 * Speed_{Medium} + .9501\ Speed_{High} - 1.22118 * Weather_{bad}$$

VISUALIZATION

Visualization
The visualization is divided into three sections: Historical, Prediction (Jan 2015 - with comparison to existing data), and prediction for each hour for the next week in London (still to be implemented). The user can filter according to a particular year and month for historical data. This historical data for Jan 2015 is compared in the figure below, with the prediction data for Jan 2015 by using Approach 1.



Historical Data for Jan 2015



Prediction Data for Jan 2015

EVALUATION

For our first model we obtain the mean absolute error between the ground truth for January 2015 and the January 2015 prediction of our Random Forest model. We obtain a score of 0.155 which means that on average we have this difference between the ground truth number of accidents in a bin and the prediction for that bin.

For the second approach, the first evaluation is a cross validation of the data set. With 2,000 points held out for testing and a decision criteria of .5 the model predicted 87% accidents with a 26% false positive rate. Real time evaluations are pending.