**ML Assignment 1:**

Supervised Model, Prediction (not recommendation)

Ideas:

Look for “beginner machine learning supervised projects”

1. Music Recommendation Challenge
   1. <https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data>
2. Social Media Sentiment Analysis?
3. Google Play Store categories
   1. <https://www.kaggle.com/lava18/google-play-store-apps/version/6>

Data Set: NYC Open Data – collisions over $1000 dollars since 2014.

<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data>

What can we predict?

* Location of crash (long./lat.)
* Cause of crash
* Severity/impact of crash

Business Problem: We are the NYPD – we have 100 police officers to investigate collisions across NYC. Where do we place them to collect the highest number of reports in their immediate area?

Analytics Problem: We have a data set of all collisions over $1000 over the last several years in NYC, with date and time. We need to run ML models to predict the point location of a collision given some input parameters (date, time, season, type of vehicle, etc.).

Discussion/group questions:

1. How do we determine if the assumptions inherent in linear regression modelling hold true (data in normal distribution, no collinearity, etc. )
2. Do we have to source data from the listed publicly available data sources?
3. Do we share with the rest of the class the data set we are working on? No duplication of projects?

Meeting Times:

Sunday – 1 PM, regular meeting time.

Deadline Dates:

Friday, Jan. 31 – 1:00 PM: Establish whether we will use the collision data set **and** identify and problems we need to solve, to use it.

Monday, Feb. 3 – 6:30 PM – **in person**: Complete data cleaning, prepare data fully ready to run ML models/algorithm, Flush out the business problem and work on explanatory text.

Wednesday, Feb. 6 – 6:30 PM – **virtual**: Assess results of running ML models