Jacob Bee Ho Brown
Sang Yun

**INFO 2950 Final Proposal**

**Proposal 1:** Wisconsin Teacher Info (from ILR stats)

- We will use an existing dataset of demographic and census-type data of teachers in Wisconsin. We will clean the dataset to drop unimportant values.
- Then with the data, we will be comparing salary with other variables such as ethnicity, gender, linguistic ability, experience
- With the information between different variables, we will run tests to see if there is a correlation between the variables analyzed
- From this point, we can implement another dataset which can be analyzed for more than just high school faculty but also previous education or higher education

**Proposal 2:** Analyze Uber traffic in NYC and its correlation to traffic issues

- We will use APIs to collect data from multiple datasets and put it into one neat, large dataset
- If we will use a dataset(s) from data files in non-csv format (i.e. JSON or SQLite), we will use APIs to convert it into a csv format so that we can easily integrate it into our one large dataset
- We will use datasets about Uber cars active during rush hour in NYC (fivethirtyeight) and compare it with datasets about the number of car crashes/accidents in NYC during the same times (google), emergency vehicle response times (google), and other possible datasets such as average Uber wait times to pickup.
- We will consider other factors that may also be contributing to the results we find such as information on taxi trip data during this time (google), Lyft traffic at this time, information on subway/public transit issues, and occurrences of major events in the city during this time.
- NYC Uber and taxi data is available, however, we need to make sure that we have data for each during the same (we also must define “rush hour” look data during other hours, not just rush hour). Lyft data is not as available. The City of New York had Lyft and Uber datasets, but they appear to have been deleted from the site. Overall, NYC traffic data, and public transportation as well as publicly operated emergency services data is also available. This idea is limited to New York City because of the availability of New York data (vs. other cities) and because this way we do not need to account for differences in environment/location/etc).

**Proposal 3:** Analyze NCAA men’s basketball player stats and information and its correlation to the number they are taken in the NBA draft

- We will use APIs to collect data from multiple datasets and put it into one neat, large dataset
- If we will use a dataset(s) from data files in non-csv format (i.e. JSON or SQLite), we will use APIs to convert it into a csv format so that we can easily integrate it into our one large dataset

Jacob Bee Ho Brown
Sang Yun

- We will look at previous NBA draft selections (google) and use data on the game stats players drafted (if drafted out of the NCAA) (google) to try to find correlations between performance in different categories and draft placement. We may also grouping players by position since different skills are more preferable at different positions.
- We may also consider the college the player played at and their NCAA tournament performance (if applicable)
- Something else that we may consider is the team that drafted the player and the quality of the team as well as the team's needs/current players or if the team did a draft and trade with the player.
- We may consider somehow considering data on the celebrity of the players and how that can also affect their draft position
- NCAA player and team and NBA draft data is widely available, however data on NCAA player celebrity value as well as data on qualitative aspects of players may not be as readily available.