**Homepage**

* Predicting Restaurant Visitation in Japan
* Looking to open a restaurant in Japan?
  + Look no further! We have built a comprehensive tool to let you know where to build your restaurant. Utilizing machine learning, we are able to predict the visitation of your restaurant by:
    - Area Name
    - Day of week
    - Week of the year
  + Super Cool Edwin Graphs

**Analysis**

* The dataset we utilized was from a Kaggle competition in regards to predicting visitors of restaurants in Japan.
* Our data set included the following:
  + Specifically, we were given number of visitors by:
  + And number of reservations by:
* Because of the large amount of features, we decided a KNN algorithm would be best to predict visitor data.
  + (explain KNN)

**Pros**:

* No assumptions about data — useful, for example, for nonlinear data
* Simple algorithm — to explain and understand/interpret
* High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models
* Versatile — useful for classification or regression

**Cons**:

* Computationally expensive — because the algorithm stores all of the training data
* High memory requirement
* Stores all (or almost all) of the training data
* Prediction stage might be slow (with big N)
* Sensitive to irrelevant features and the scale of the data

**Quick summary of KNN**

* The algorithm can be summarized as:
* A positive integer k is specified, along with a new sample
* We select the k entries in our database which are closest to the new sample
* We find the most common classification of these entries
* This is the classification we give to the new sample

A few other features of KNN:

* KNN stores the entire training dataset which it uses as its representation.
* KNN does not learn any model.
* KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance.
* As KNN would be most accurate using a data set that was proportional in nature, our final decision was to only utilize the air visitors data set with the following features:
  + Area Name
  + Day of week
  + Week of the year
    - Excluded was:
      * Holidays
      * Genre Names.
    - Holidays and Genre Names was excluded because not every region had holidays or all genre types.
* **Data Clean**
  + The visitor data was given by day by hour. To make our data proportional we designated the 52 weeks we would measure and assigned each day its “week”.
  + This was done to help our algorithm understand seasons.
* **Results**
  + Finally, when we tested we got a R2 value of .667.
  + Below is the chart of how we decided on our N\_Neighbors:

**Sources**

* **Link Kaggle.**

**Who we are**