Project Proposal

## Project Title: California Energy Consumption Prediction

## Student Names: Frank Dong, Yang Weng, Nima Hendi, Sean Youngseon

### 1. Project Summary

This project aims to develop a predictive model to forecast energy consumption in the manufacturing industry. The approach I will take to address this problem is multi-faceted: (1) collect historical data on energy consumption and weather patterns, (2) preprocess and clean the data, (3) identify relevant features and engineer new ones, (4) develop and train machine learning models, and (5) evaluate the performance of the models and fine-tune them as necessary. Ultimately, the goal is to provide accurate predictions that can help make informed decisions and optimize energy usage.

### 2. Proposed Technical Approach

Combine the industry energy related data sets.(Gather the industry energy related data sets from multiple sources, clean and organize the data, and then combine them using a data integration tool to create a comprehensive dataset.) When combining industry energy-related datasets, it's important to first gather the data from multiple sources and clean and organize it to ensure that the data is consistent and accurate. This involves checking for duplicates, filling in missing values, and ensuring that the data is formatted consistently across all sources. Once the data has been cleaned and organized, a data integration tool such as Talend, Informatica, or Pentaho can be used to combine the data into a single comprehensive dataset.

To implement an MLP neural network model for industry energy consumption, we first need to preprocess the data by scaling and normalizing it to ensure that all data points are within the same range. The data can then be split into training and testing sets, with the training set used to train the MLP model and the testing set used to validate the model's accuracy.

Then visualizing the MLP model, plot the predicted values against the actual values to see how well the model is performing.

### 3. Data Sets

For this project, we will be working with dataset provided to us by Accenture.

U.S energy summary

U.S energy consumption summary

Industry energy consumption by category

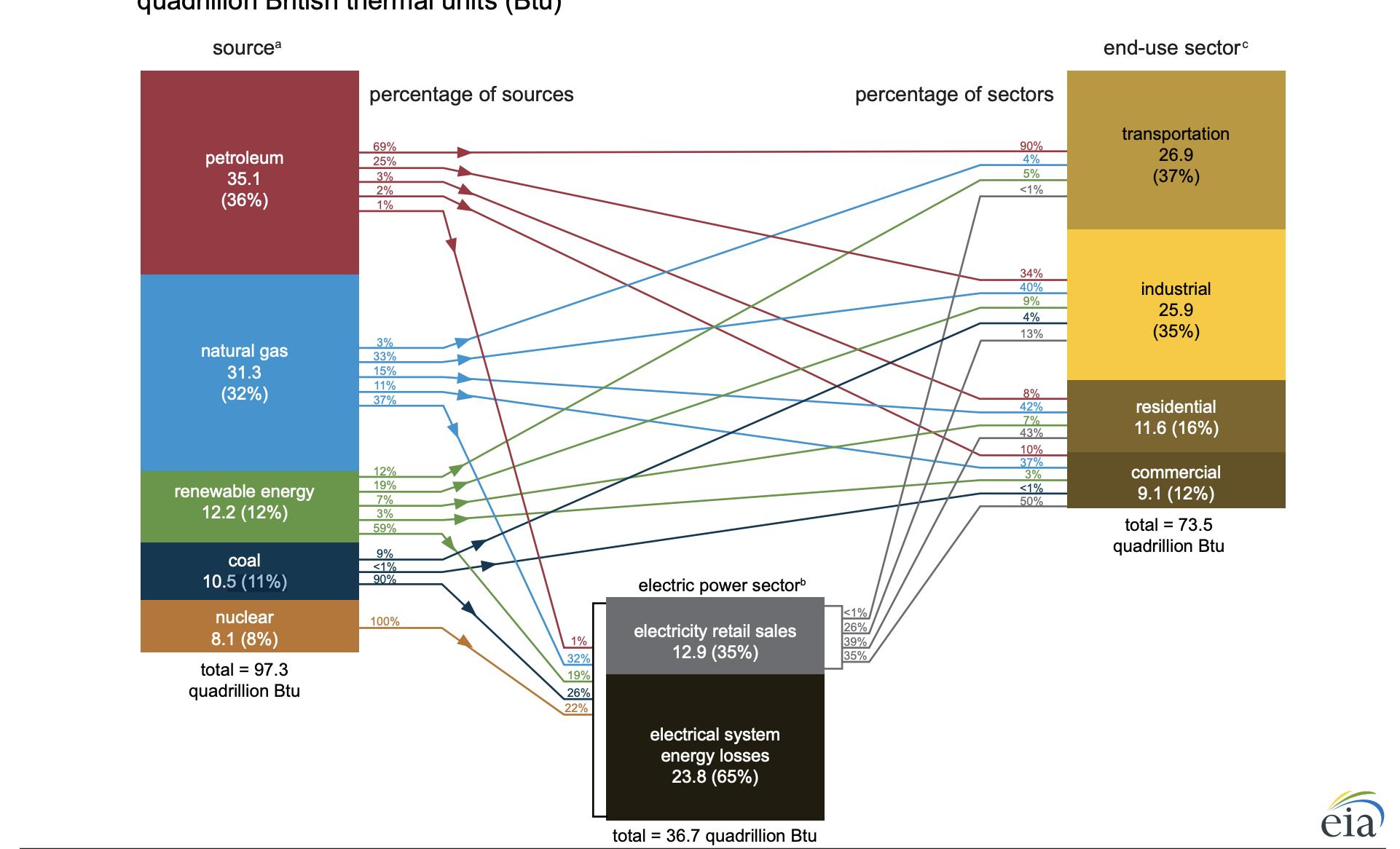
Energy trade by category (total 5 categories)(Petroleum, natural gas, coal, electric, nuclear)

Energy prices by category

Industry emissions by source

Link all these data by dates.

<https://www.eia.gov/totalenergy/data/monthly/>



### 4. Experiments and Evaluation

Use MSE to evaluate the model performance.

Use cross validation to avoid overfitting and underfitting. Divide training and testing in time series.

Modified training data along training to build better models, since our data cover a long period and old data might be noise for precise prediction.

Provide a brief and clear description of how you plan to evaluate the results of your project. For example, if you are doing classification, you should consider metrics such as classification accuracy and precision-recall, in addition to having access to training data. Will you use cross-validation, or does your data set(s) come with a fixed train-test partition? For unsupervised learning tasks like clustering or topic modeling, you may have to do some research to see how evaluation is done on these tasks. For some projects you may even have to do some user studies for evaluation, e.g., present users with results from Algorithm A and Algorithm B, using the same input data for each algorithm, without telling the user which algorithm is which, and have them select the one they prefer. Or your evaluation may be more qualitative in that you hope to generate insights about a particular problem.

### 5. Software

Provide a list of the major pieces of project software that you expect to use, divided into 2 sets: (1) publicly available code/libraries, and (2) code you will write yourself. The list of what public software you will use will probably be incomplete at this point (which is fine) since you may not know yet about all of the software that might be relevant to your project. You should plan to use Github to coordinate your code development on the project – if you have not used Github before, this will be an excellent opportunity to learn to use it.

Python, Visual Studio Code, Jupyter Notebook, Pytorch, R, R Studio, Pandas, numpy, matplotlib

### 6. Milestones

Provide a brief description of your Winter/Spring quarter work milestones for the time periods below:

* Winter
  + Weeks 8-10 Data sets Collection and Combination
* Spring
  + Weeks 1-2 Data sets EDA
  + Weeks 3-4 Statistic Model fit, and Machine learning Model fit （regression, decision tree
  + Weeks 5-6 Model validation and visualization
  + Weeks 7-8 Model analyze
  + Weeks 9-10 Final report

For example, much of the data gathering and preprocessing and coding (development and test) could happen in the earlier weeks, and much of the evaluation and writing in the later weeks. Note that you will be required to provide milestone updates and progress reports on a regular basis.