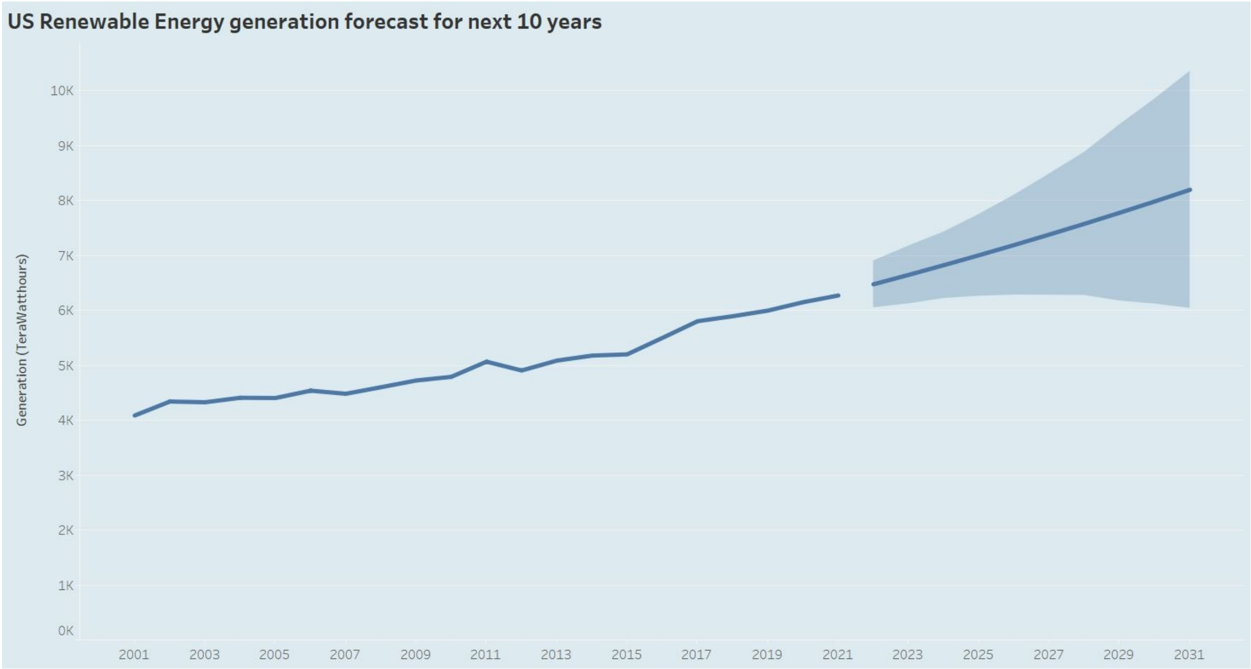
7) Analytics, Machine Learning

- We worked on machine learning that works with a linear regression model and a supervised learning model. We used AWS Sagemaker to work on the dataset and to train the machine learning model.

8) Evaluation and Optimization

- Using AWS will allow for a better evaluation and optimization of the machine learning model due the large processing power. We tried to work on creating a model that would create a 70% in the training dataset and 30% in the test set.

9) Results



-
- This graph above shows what the predicted energy consumption would be after the 2022 year. One of the questions we wanted to answer is “How much energy is required for the next 10 years?” and the graph above shows that the energy consumption will continue to increase close to 8 TeraWatt hours during the year 2031. Based on the prediction it looks like the energy consumption will continue to increase every year.

10) Future Work, Comments - students may want to consider the following questions What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

- For the most part the data was clean, this made it easier to work with. We did check using Sagemaker to see if the data needed to have any null values or adjustments made, but it was already in a clean format. There was one piece of data that could be considered an outlier, which is the energy data for the year 2022. Since data for that is incomplete it really can’t be utilized in our dataset. The dataset of US energy generation was unique since it was a really large dataset (big data) that comprises the energy usage in the US starting from the year 2001 to 2022 for all 50 states. It also includes attributes of energy source, the type of producer, and the generation of energy in terms of megawatt-hours.

Did you create any new additional features / variables?

- We did not need to create additional features or variables for this dataset. In this case the dataset was easy to work with, so no additional tasks were required to manipulate the dataset.

What was the process you used for evaluation? What was the best result?

Is there Bias in your work? What were the problems you faced? How did you solve them?

- There were multiple problems that we faced in terms of the machine learning model. The biggest issue was trying to create the model and actually run in AWS Sagemaker. Our data was clean and ready to use, but unfortunately we learned that we need to modify the dataset in order to work in AWS SageMaker. We made an attempt to solve it using AWS resources on how to operate and train a machine learning model in Sagemaker.

What future work would you like to do?

- For the future we want to make optimization to the machine learning model in order to make a more accurate representation of the dataset. Additionally, we want to see if there is a possibility of publishing this analysis of this US energy generation dataset. This includes the visualization that was created in the previous deliverable and a deeper analysis of the machine learning model.

Instructions for individuals that may want to use your work

- In order to use this work, first the dataset will need to be downloaded from Kaggle. Once the download is complete the upload the dataset to an AWS S3 bucket. In order to use the dataset it will need to be parsed using AWS Sagemaker to double check for any inconsistencies like null values. Once that is complete the visualizations can be created using AWS Sagemaker or Quicksight.
- This is the link to the dataset we used.
    - https://www.kaggle.com/datasets/kevinmorgado/us-energy-generation-2001-2022