MNLKuzmin / USEnergy_Generation Public

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security Insights Se

main

MNLKuzmin final touch ups ... 1 minute ago 43

View code

README.md

Phase 4 Project

USEnergy_Generation

Student name: Maria Kuzmin
Student pace: Flex
Scheduled project review date/time: Tuesday, March 21st 10AM
Instructor name: Morgan Jones

Time Series Model:

Business Problem:

Electrical Energy is constantly in more demand.
With technology advancing and broadening applications in different fields there are always more industries heavily depending on significant amounts of electricity.
A few examples of these industries are data centers, hospitals & healthcare facilities, warehouses & distribution centers, and the list goes on (for more details on this you can read this article).
Whether you want to invest in the energy sector and are looking into what type of source for power you should focus on, or if you are looking to start a massive Cloud service, and wondering where is the most efficient location to have your data servers built and run, or you want to invest in a hospital or a warehouse, this project aims at helping you understand what energy source and state within the United States, is best for you.
We are going to study the different sources of energy and how their production has been growing or decreasing in the past 20 years. Ultimately we will try to identify what would be the most reliable source of energy to focus on.
We will then try to predict the production via that source for the next 3 years.
Consequently we will look at what are the states that have been the greatest productors, and what is the seasonal pattern in those states.

Summary:

We studied the production of electrical energy in the United States, from 2001 until May 2022.
The data has information on the energy produced, with one record per month, divided by sources and by state.
The dataset consists of 496774 rows × 6 columns.
We decided to focus on the production of electricty via natural gas, as that showed a promising positive trend.

Data Preparation: we organized the data by date, studied it in terms of the production in all of the US, and also separated by state.

Modeling:
We decided to create a few ARMA models and later run a gridsearch to find the best parameters.
Consequently we used SARIMAX to add the seasonal aspect to our models, and ran two grid searches to find the best performing model.
To compare the models and pick the best ones we used a cross-validation on a split we made from the train set. The best model we picked had an AIC of 173.5.
With this model we made predictions on the test set, for a span of 51 months, and obtained predictions with a RMSE of 10.9 TWh.
We forecasted our data in the future for 3 years, and obtained results with a MSE between 5-9 TWh, finding a growth of up to 16.7%.
We studied also the production of natural nas by state and studied the seasonality of the time series.

How are we going to get there:

Here is a roadmap of the steps that we took:

The Data:

- Data Preparation:

    - Study of Energy Source
    - Natural gas, solar and wind
    - Checking for normality
- EDA of Natural gas:
    - Split Train Validation and Test Set
    - Subtracting Rolling Mean
    - Series Decomposition
    - Studying Autocorrelation: ACF PACF
- Modeling:
    - Baseline Model: Naive
    - ARMA Models
    - Grid Search for ARIMA models:
        - First Search
        - Best model from Grid Search
        - Cross Validation
        - Second Search
        - Best model from second search
        - Cross Validation
    - SARIMAX
        - First Grid Search SARIMAX
        - Best model after first search
        - Cross Validation
        - Second Grid Search SARIMAX
        - Best model after second search
        - Cross Validation
- Predicting on the test
- Forecasting in the future
- Study of Seasonality and States
- Results
- Limitations
- Recommendations

The Data

This CSV (organised_Gen.csv) is adapted from https://www.eia.gov/electricity/, where the main information related to energy generation in the United States is located. This dataset has the following columns:

YEAR
MONTH
STATE
TYPE OF PRODUCER
ENERGY SOURCE
GENERATION (Megawatt-hours)
Unnamed:0 (ID)

Source: U.S. Energy Information Administration (Sep 2021).
The dataset consists of 496774 rows that span the timeframe from January 1st 2001 until May 1st 2022, with one record per month.
The data is divided by energy source, type of producer and state.
The column we will focus on is the GENERATION one, which we will study in its evolution over time and which will also be the target for our study, since we want to predict the values of energy generated in the future.

Data Preparation

Data preparation included dropping columns, checkinf for null values, changing scale on the 'Generation (TWh)' column. Once we visualized the data by source we decided to focus on the sources that have shown to be more interesting, with an upward trend over time: natural gas, wind and solar (which includes thermal and photovoltaic). Next we checked the three samples of these sources for normality, skewness and kurtosis. We didn't find the data to be normally distributed, but we also got confirmation that the data was well suited for our project and no anomalies were present.

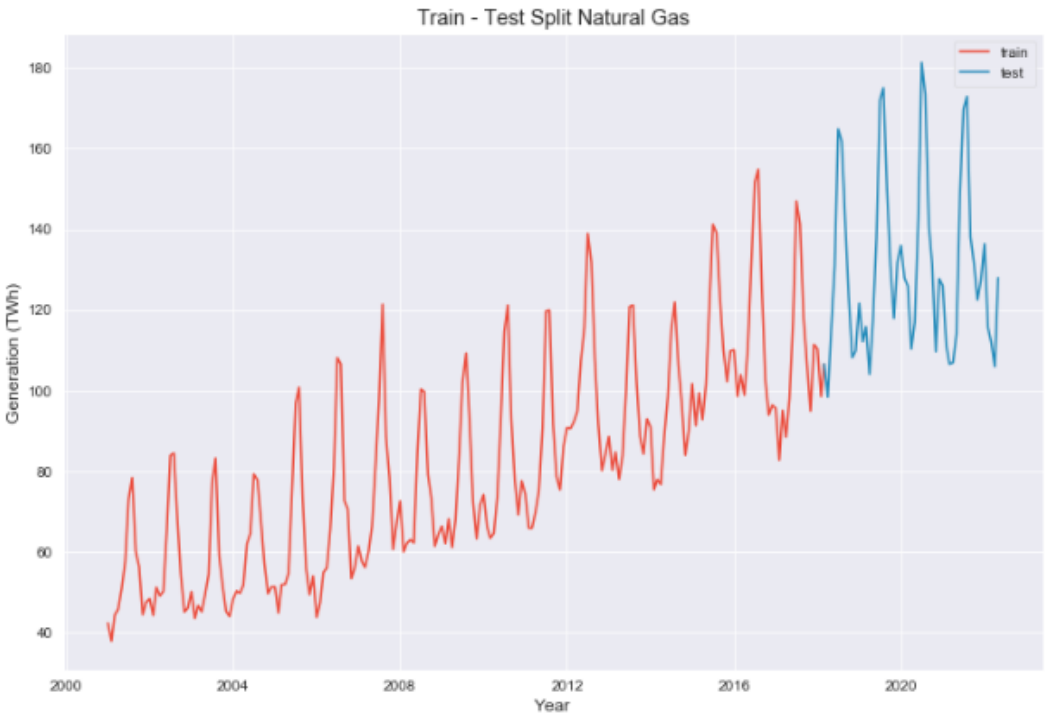
Focusing on one source:

Ideally we would have continued our whole analysis on all three sources, but for reasons of time we needed to focus on only one.
The exponential trend of solar and wind are very interesting, but they both carry risk due to their vulnerability toward factors out of our control: principally weather.
This is why we decided to continue our analysis focusing only on natural gas, whose production does not depend so heavily on uncontrollable factors.

EDA of Natural Gas

Split Train Validation and Test set

Once we have decided what source we were going to work on, we performed a train-validation-test split so that the models we are going to create, based only on the information we have from the train, do not suffer from data-leaking.
We did cross validation on our top models (that we will idenfity with grid searches), to be able to compare their performance and choose one best model.
We tested our best model on the test set, that in this way will represent data that was not seen by the model.
Lastly we used our model to forecast for the next 3 years.



Validation Set

Instead of taking only one validation set as part of the test set, we chose to use the `TimeSeriesSpit()` function to create splits of train and validation, that can be used for a cross validation.
We can see from the indeces below that the default number of splits is 5, and that the validation set always comes right after the train set, but the train set is increasing in size with every split.
The size of the validation set remains the same.
We used to do cross validation on our models.

Next we did some exploration on our data, subtracting the rolling mean, doing series decomposition, studying autocorrelation plotting ACF and PACF.

Modeling

First we built a baseline model, specifically the naive model, that simply predicts the data to be equal to the day before, and it is calculated by adding a shift to the original time series. The RMSE for the naive baseline model is: 12.32 TWh.

ARMA models

Next we started to create some ARMA models, first we created a few preliminary ones and then we did a Grid Search to find the best parameters to minimize the AIC of our model. The model we found with this first search had parameters (8,1,2) and AIC of 1401.174. The RMSE was 7.45 TWh on the train, and 13.13 TWh after cross validation. After this we did a second grid search to see if we could further improve the performance of our ARMA model. The model we found after the second grid search had parameters (12, 1, 4), and an AIC of 1351. The RMSE was 6.6TWh on the train and RMSE 13.91TWh after cross validation.

SARIMAX models

We decided to use SARIMAX to account for seasonality in our models. We did a first grid search and we found the best values to be (8, 2, 2) (8, 4, 1, 12) and had an AIC of only 390. The RMSE was 20.61TWh on the train and RMSE 602.88 TWh after cross validation. We performed a second grid search on SARIMAX and found best parameters (8, 1, 2)(12, 2, 1, 12) and AIC 173.51. This model had an RMSE on the train of TWh 7.53 and RMSE 24.47TWh after cross validation.

Now let us do a quick recap of the results thus far.

In terms of the RMSE:

The ARIMA model gave a better result, since it has a lower RMSE of 14.24 TWh after cross validation, while the RMSE of the SARIMAX model after cross validation is 24.47 TWh.

In terms of AIC:

The ARIMA model (before the SARIMAX searches) had an AIC of 1355.815. The model from the second SARIMAX grid search has an AIC of only 173.

Which metric:

There is a substantial difference between these two metrics.

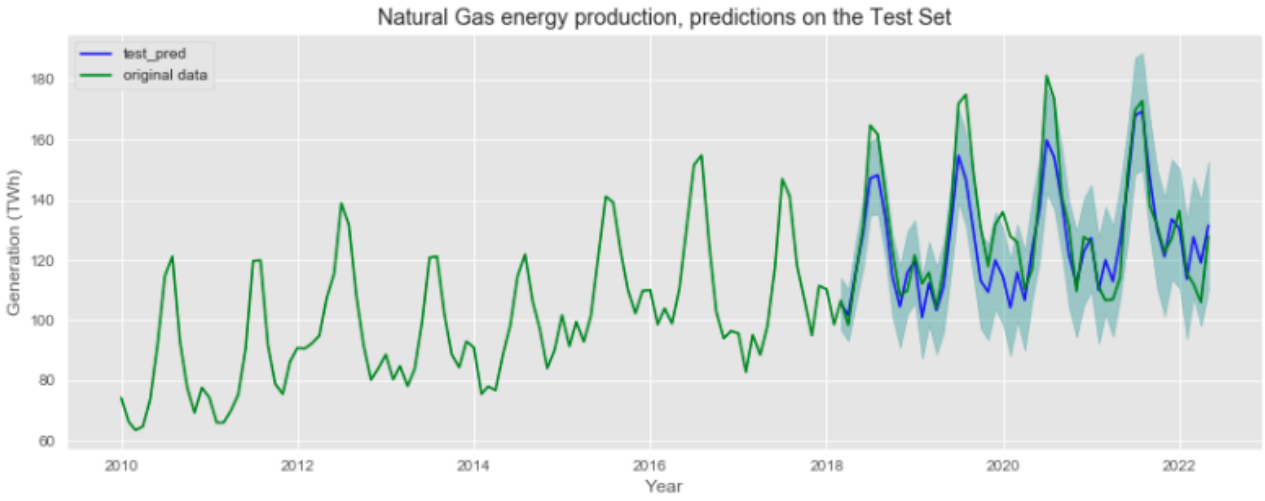
We want the AIC to be low because it gives us a sense of the overall goodness of the model and penalizes models that are too complex, so it should prevent overfitting. We also want the RMSE to be low because that is what ultimately tells us how well the model is performing, compared to the data.

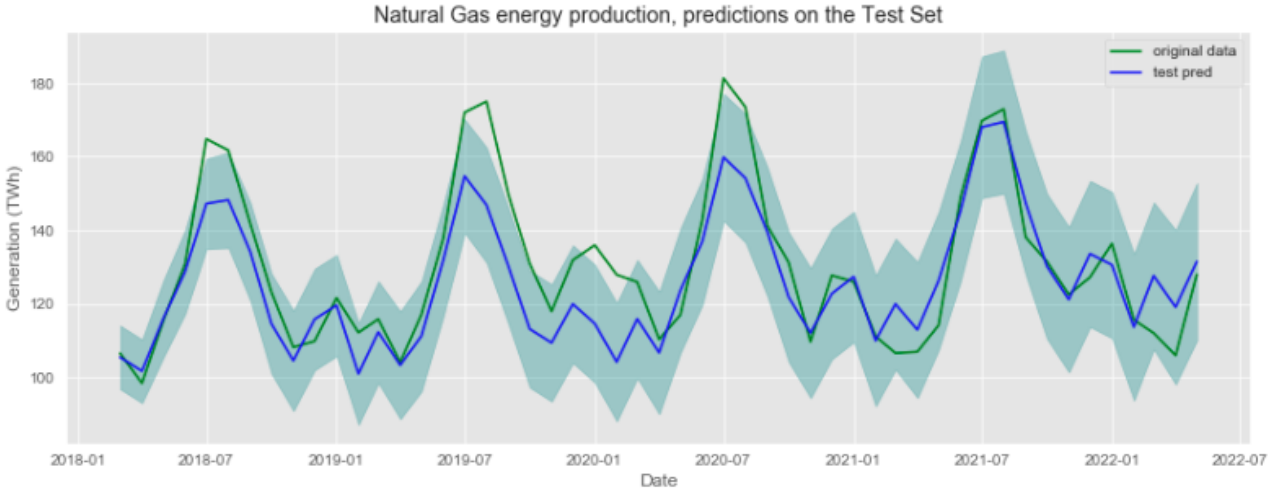
We decided to pick the model with the lowest AIC as our best model, since it prevents us from the risk of overfitting, and it is a more well rounded metric that evaluates the goodness of model overall, not as dependent on the data as RMSE is.

Once we have selected the SARIMAX(8, 1, 2)(12, 2, 1, 12) we proceeded to predict with this model on the test set, and we used again the RMSE as a metric to get a sense of how well this model is predicting on the data.

Predicting on the test

We used the model to do predictions on the test set and obtained predictions with an RMSE of 10.90 TWh.

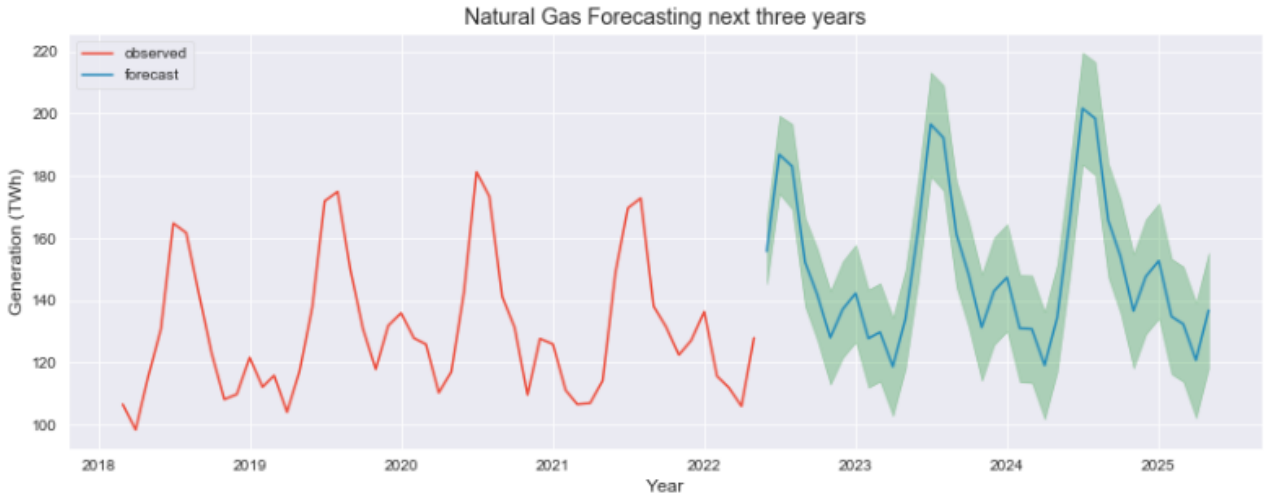




This is a great result and we were able to achieve a Root Mean Squared Error lower than the Baseline Naive Model!
The choice of the best AIC ended up being a wise one since we don't see overfitting in our model, as the performance didn't decrease once we used the model on the unseen test set.

Forecasting

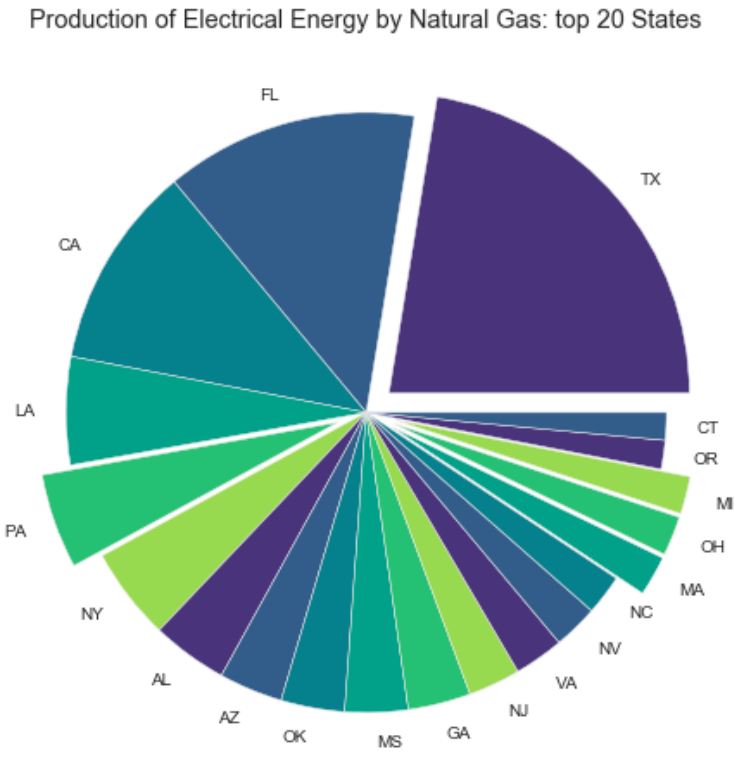
Then we used same model, but fitting it on the whole dataset we have available, to make the best possible predictions, for the next 3 years. This graphs shows our results:



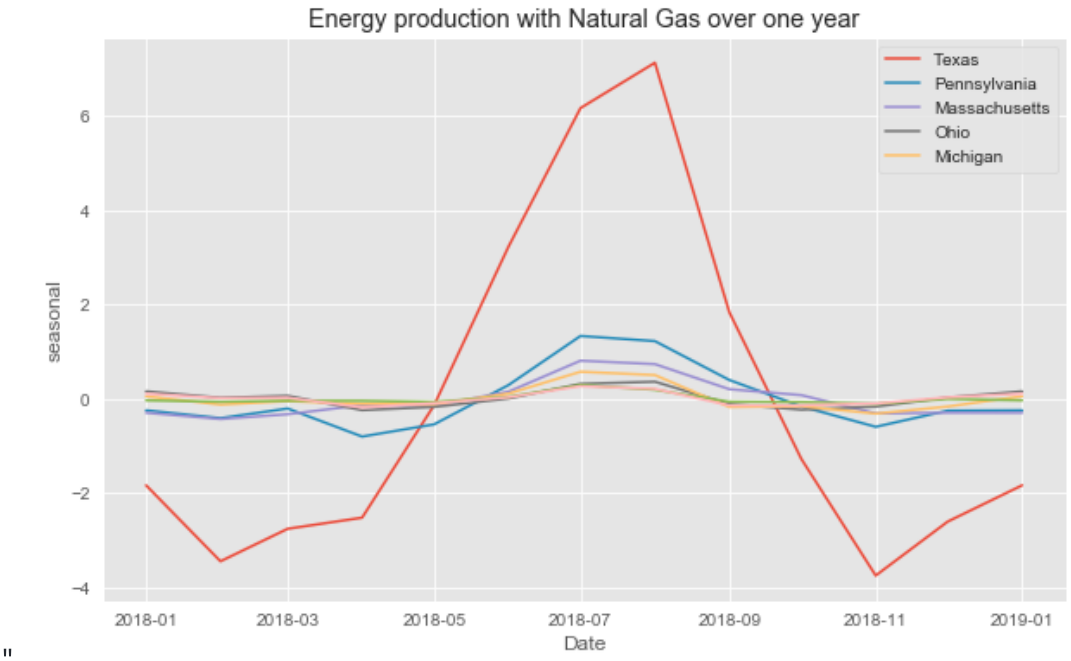
We see a steady growth in the production of natural gas, with a seasonality very similar to the one we have seen so far, and with a general growing positive trend. We calculated a year-over-year growth of 16.7% between 2021 and 2024 with a relative error of 4%.

Study of seasonality and states

We wanted to study the production of natural gas divided by State, to see which state produces the most of it. Focusing on the top 20 producers we found:



We can see that Texas has the lead in terms of production of Electrical Power via natural gas, followed by Florida and California. When we went to look at one year of production in Texas we found a very high spike in production in the summer months.
We can imagine that this is due to the hot climate in Texas during those months, that requires most houses and buildings to run constantly their air conditioner units.
We can expect to find a similar behaviour in the other highest producing states since they all happen to be in the South, where the heat is more severe in the summer.
This can be complicated and create shortages which can lead to several serious problems.
We can look into other states, not in the South, that don't suffer as much from this type of seasonality.
We took a few states in the North of the United States and studied their trend over a year, to see what that looks like in states that have a more mild climate.
Starting from the top producing states, we selected: Pennsylvania, Massachusetts, Ohio and Michigan.
This is what we found:



As we can notice the production of energy is much more stable in these other states, and we can imagine that is due to the much smaller consumption of energy related to air conditioners, since all these states have summers that are much more mild than Texas.
This can be a way to "beat the seasonality", if there is a possibility to invest in more than one state.
In this way we would be relying on energy in both Texas and another one of these states, that don't necessarily produce an incredible amount of energy, but given their stable production can be a valid backup option in case of shortages from the production in Texas.
Pennsylvania might be the best option in this sense since it was still pretty high in the order of producing states (it was number 5, looking at the graph above) but having a much less pronounced peak during the summer compared to Texas and presumably the other warmer states.

Results

With our analysis we wanted to obtain predictions of the production of natural gas.
We selected as our best model a SARIMA model with parameters (8, 1, 2)(12, 2, 1, 12) and with an AIC of 173.5.
With this model we made predictions on the test set, for a span of 51 months, and obtained predictions with a RMSE of 10.9 TWh.
We also forecasted our data in the future for 3 years, finding a growth of up to 16.7% between 2021 and 2024, with an MSE between 5-9 TWh.
We then studied the seasonality of our time series, identifying a recurring pattern every 12 months, and studying the different trends in a few different states, focusing on differences most likely based on geographical area and climate.

Limitations

Cleaning the data: we had almost no cleaning to do of our data. It was already in the vertical format that we wanted, and there was no missing data. If the data input was actually in the long format of dates we suggest the use of the `pd.melt( )` function to reformat the data before being able to run it through our model.
In case of missing data there are three possibilities for filling in the missing data, with either back fill, forward fill or interpolate.

Grid Search: the grid search would give different results for different data. Also the parameters chosen for the search were based on testing on some models that were not included in this notebook. With a different data sample, other models would turn out to be optimal. Therefore it would be necessary to choose other (p,d,q) values for the search and re-run the searches to find the best parameters for the models as described here. Moreover it is not to take for granted that the different models generated would necessarily perform as well as ours did, since they might not be able to pick up the correct trend and seasonality in a different time series and consequently not be able to predict the data with the same accuracy that we obtained.

Running Time:Note that all of the grid searches on SARIMAX had to be run over night or for an extended period of time due to the amount of calculations needed, which will increase as the parameters chosen for the search increase. Also, the fitting of SARIMA models and the cross validation calculations on their results took a considerable amount of running time. If running time is an issue the performance of the models will definitely decrease, as it would be much harder to find such well performing models just by trial and error of different parameters.

Recommendations

We recommend investing in natural gas, solar energy and wind energy, as those appear to be the most growing sources of energy across the states.
When taking into consideration the unpredictability of weather, solar and wind energy possess much more risk and thus less reliability. Ultimately we suggest focusing on states that have natural gas as a main source for generating electrical power.
We recommend Texas as the state that has been producing the most electrical power through natural gas. To smooth out the seasonality present in Texas due to the high temperatures in the summer and thus the higher load on generating electricity, we recommend also investing in another state like Pennsylavania, Massachusetts, Ohio or Michigan where there is more of a stable trend in the production of energy annually through natural gas.

Next Steps

To improve our model and for a more in depth study we could also:

- Study solar and wind generated energy trends.
- Collect more data for a better detailed analysis, scraping information both farther in the past and records by day instead of by month.
- Deepen our study of trend and seasonality looking into the possible most influential factors.
- Utilize more powerful tools like Prophet or AWS SageMaker's DEEPAR which were not available to us for this study.

For More Information

Please review my full analysis in my Jupyter Notebook or my presentation.

For any additional questions, please contact **Maria Kuzmin, marianlkuzmin@gmail.com**

Repository Structure

Description of the structure of the repository and its contents:

```
├── .ipynb_checkpoints
├── Graphs
    └── Frecasting.png
    └── PieStates.png
    └── Predsontest.png
    └── Predsontestzoom.png
    └── TrainTestSplit.png
    └── YearStates.png
├── .gitignore
├── Presentation.pdf
├── README.md
├── SARgrid1.pkl
├── SARgs.pkl
├── TimeSeriesNotebook.ipynb
```

```
└── organised_Gen.csv
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%