Solar Energy Analysis

Group 3, Section 2, MBD

25 December 2020

Objective

The objective of this project is to predict the incoming solar energy of 98 Oklahoma Mesonet sites. Before running the actual model, this report aims to give a general overview of the data and which variables might be of interest or should not be taken into consideration.

Renewable energy market

With an increasing trend to use less fossil resources and reduce emissions, the renewable energy industry is naturally gaining in importance and shows strong efficiency gains in recent years. Despite all this, energy production is essential for regions and this is mostly fluctuating in renewable energies, as they are highly dependent on weather conditions. For this reason, good forecasting models are becoming more and more important and this will also be our final outcome of this analysis. As already mentioned, we will use exploratory analysis to get a better overview of the data and following that build a model that will allow us to predict the energy production of different solar stations.

Data set

The data used in this report is a set of observations obtained from the American Meteorological Society 2013-2014 which can be found here (https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/data).

In total we have three different data sets:

- `solar_dataset.RData` : Information from the stations and additional features which had undergone a principal component analysis. Each observation corresponds to a day.
- `station_info.csv` : Information on each of the stations like latitude, longitudes, and elevation (meters)
- `additional_variables.RData` : Information on numerical weather predictions (NWP) which might be relevant for the solar energy production.

As mentioned, the solar data set was pre-processed by means of principal component analysis. This analysis is a dimensionality-reduction method which is often used when handling big data sets as the ones here. Since we do not have any information on the exact pre-processing steps we will not explain this further.

Exploratory Data Analysis

First, we would like to have an understanding of what our data actually looks like. The data set we will look at in the following steps is the merge of the solar data set and the additional variables. Below is a short summary of the dimensions of the overall data set:

General data overview

rows	columns	discrete_columns	continuous_columns	all_missing_columns	total_missing_values	complete_rows	total_observations	memory_usag
6909	557	2	555	0	202741	90	3848313	2906626

Next we decided to explore the descriptive statistics of each of our columns, in order to find out more about what the data looks like.

Descriptive statistics independent Variables

variable_names	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ACME	1	5113	16877462	7869606	16946400	17133839	9133113	12000	31347900	31335900	-0.1718741	-0.9826492	110056.3
ADAX	2	5113	16237534	7905850	16299300	16416948	9233188	510000	31227000	30717000	-0.1243033	-1.0087226	110563.2
ALTU	3	5113	17119189	7702989	17073600	17387346	8968989	900	31411500	31410600	-0.1726472	-0.9447221	107726.2
APAC	4	5113	17010565	7883455	17062500	17263585	9102867	3300	31616100	31612800	-0.1716810	-0.9937528	110250.0
ARNE	5	5113	17560173	7917965	17578500	17798731	9356392	477300	32645700	32168400	-0.1442343	-1.0087025	110732.7
BEAV	6	5113	17612143	7911267	17520900	17840528	9553874	300	32884800	32884500	-0.1321769	-1.0417475	110639.0
BESS	7	5113	17304074	7868140	17176500	17559084	9142382	510600	31887900	31377300	-0.1559424	-0.9903115	110035.8
BIXB	8	5113	15969634	8062981	15725700	16083165	9800283	74700	31702200	31627500	-0.0541698	-1.0801112	112760.7
BLAC	9	5113	16061707	7770310	15873000	16198770	9128665	36000	30771900	30735900	-0.0646583	-1.0631387	108667.7
BOIS	10	5113	18688943	7649349	18885900	18899424	9544534	467100	32922000	32454900	-0.1515437	-1.0027286	106976.1
BOWL	11	5113	16034081	7830831	16043700	16230524	9185152	497700	30781200	30283500	-0.1348628	-1.0108496	109514.1

Descriptive statistics dependent Variables

variable_names	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
PC1	1	6909	0.6759067	247.060366	4.3774061	6.5913130	334.182453	-620.60132	420.28774	1040.88906	-0.1197660	-1.190031
PC2	2	6909	-1.7116078	93.777836	14.7412304	7.0456390	80.582624	-535.51144	238.04418	773.55562	-1.0500313	1.719365
PC3	3	6909	1.0412696	76.072152	3.2427081	1.2492713	73.296788	-249.36432	364.22671	613.59103	0.0460135	0.311259
PC4	4	6909	-1.2291959	55.915254	-2.6565410	-2.1475618	46.458304	-305.06381	260.11485	565.17866	0.1526624	1.584053
PC5	5	6909	-2.2295472	50.428224	-7.2470006	-3.7923037	55.880753	-140.28042	225.03832	365.31874	0.3036154	-0.339266
PC6	6	6909	-0.4769296	45.912596	-0.2104490	-0.0220145	36.005313	-286.57653	255.54613	542.12265	-0.1289789	2.797711
PC7	7	6909	0.0884722	41.290797	-0.6104207	-0.5564135	35.325645	-182.73916	207.40843	390.14759	0.1725842	1.308365
PC8	8	6909	0.8147779	37.994583	0.8833097	0.7562666	34.303465	-175.06915	231.75532	406.82447	0.0354439	1.491789
PC9	9	6909	-1.3701773	34.341546	0.2496814	-0.6915616	27.539498	-189.86208	245.32364	435.18572	0.0038556	3.527361
PC10	10	6909	-0.2940384	29.988878	-1.4550747	-0.8763559	28.874672	-92.57566	201.21544	293.79111	0.3815078	1.217354
PC11	11	6909	0.8124266	29.718079	1.6277152	1.1312636	28.644126	-155.36289	149.00982	304.37271	-0.0318979	0.653638

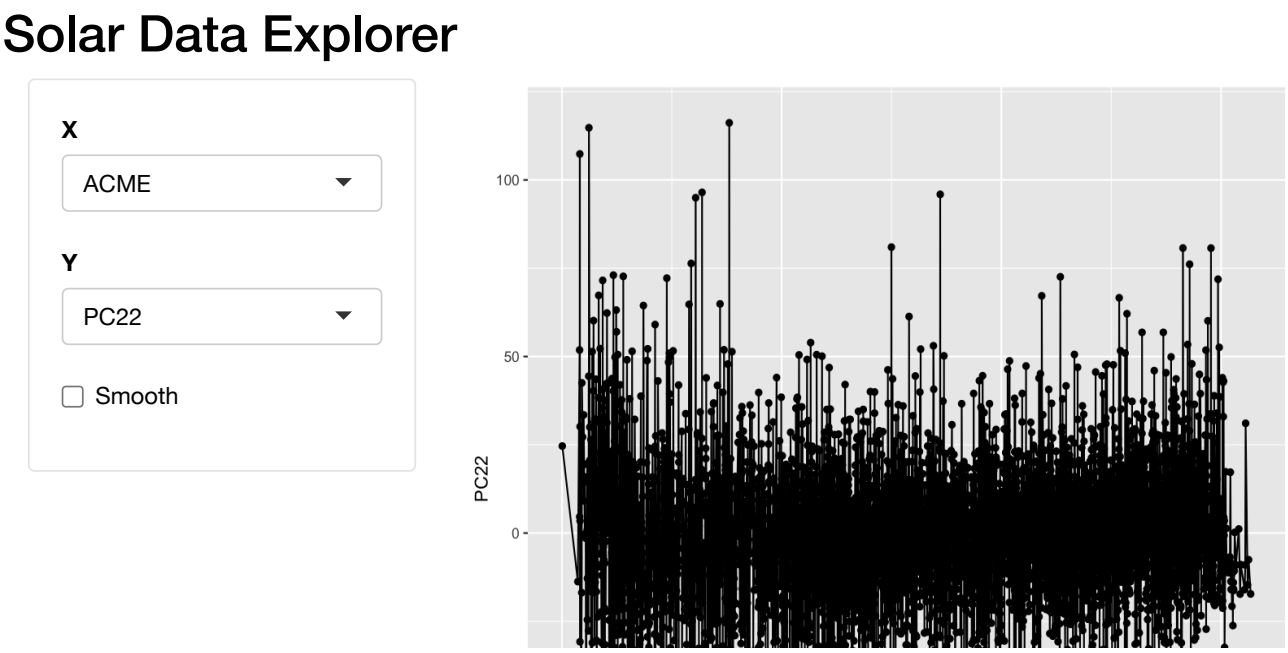
Solar data set overview

As the solar data set is the main component of our analysis we are going to give a short but more detailed overview what the main components of the data set are:

- column `Date` in `solar_data[,1]` : The level of our observations. For each day we have an observation per weather station
- column `ACME` , `ADAX` , `ALTU` ... in `solar_data[,2:99]` : the different weather stations
- column `PC1` , `PC2` , `PC3` ... in `solar_data[,100:456]` : the different features later used for our predictions

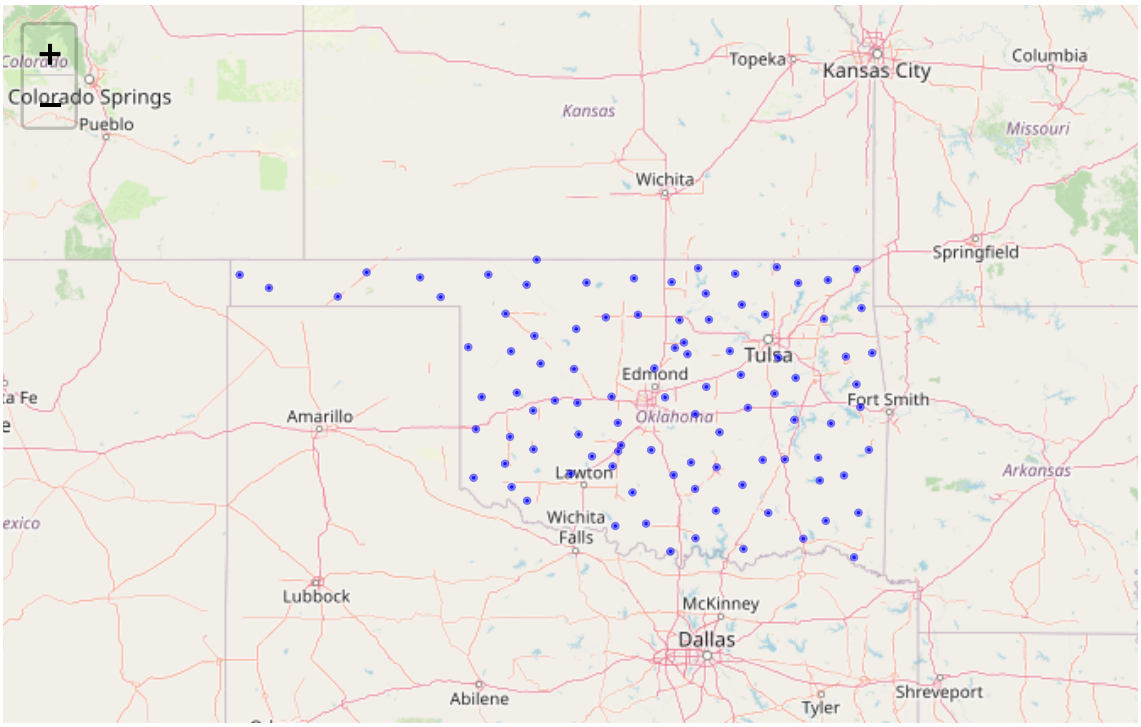
The period covered by the data set ranges from January 1st 1994 to 30th November of 2012 while we are missing the data of the weather stations in between January 1st 2008 to 30th November of 2012. These values will be later predicted by our model.

Since it is hard to plot each of the features in a big data set like this we developed a small Shiny app to get a better understanding of the data:



Location of the stations

In conserdation of our goal to predict the solar energy production of 98 different solar panel sites, we thought it might be interesting to have a closer look where they are actually located:



Distribution Explanation

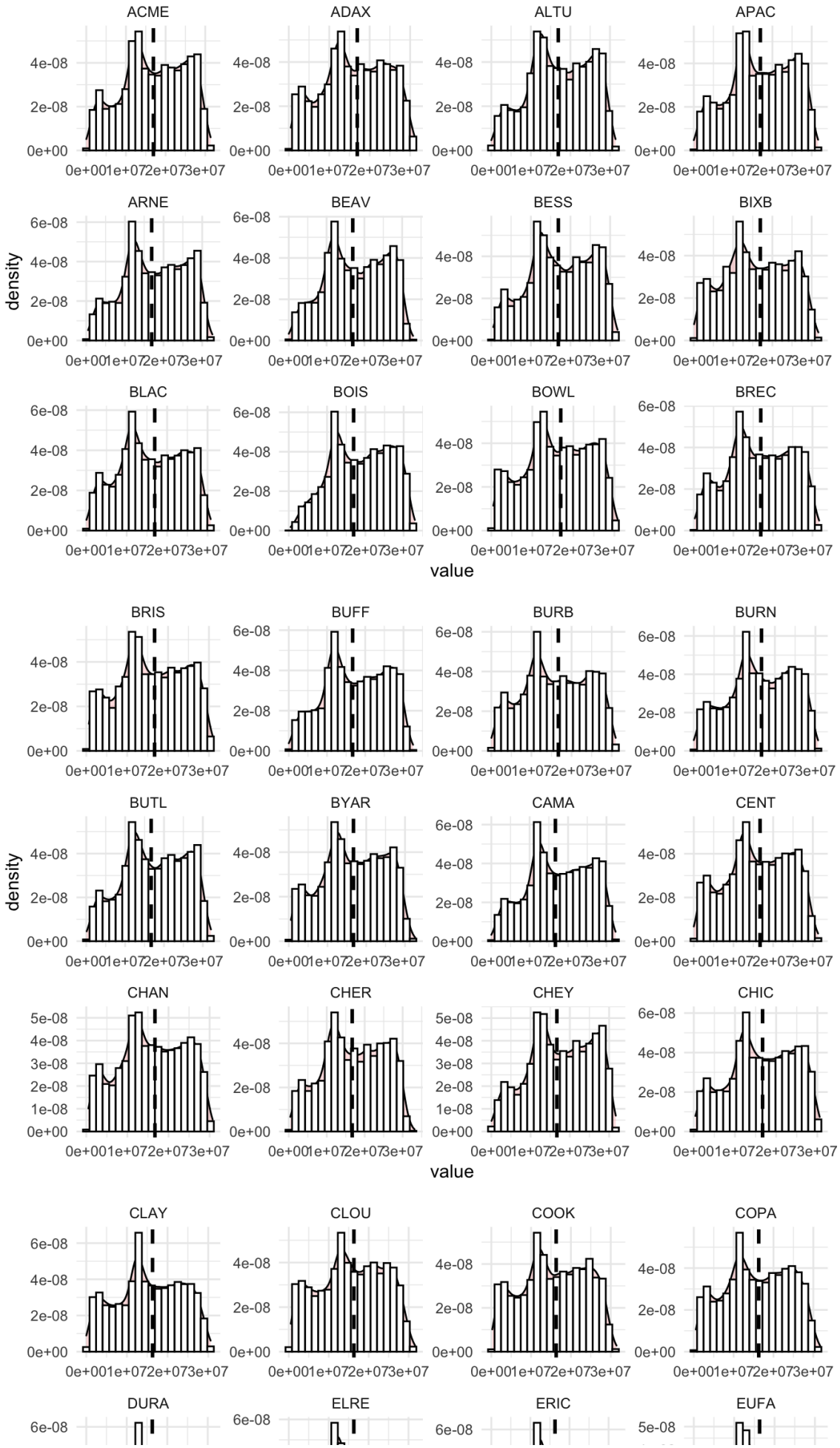
So far, we have mainly looked closer into the description of our data and the general statistics of it. Another important part, before running a machine learning model is the analysis of the data in terms of distribution.

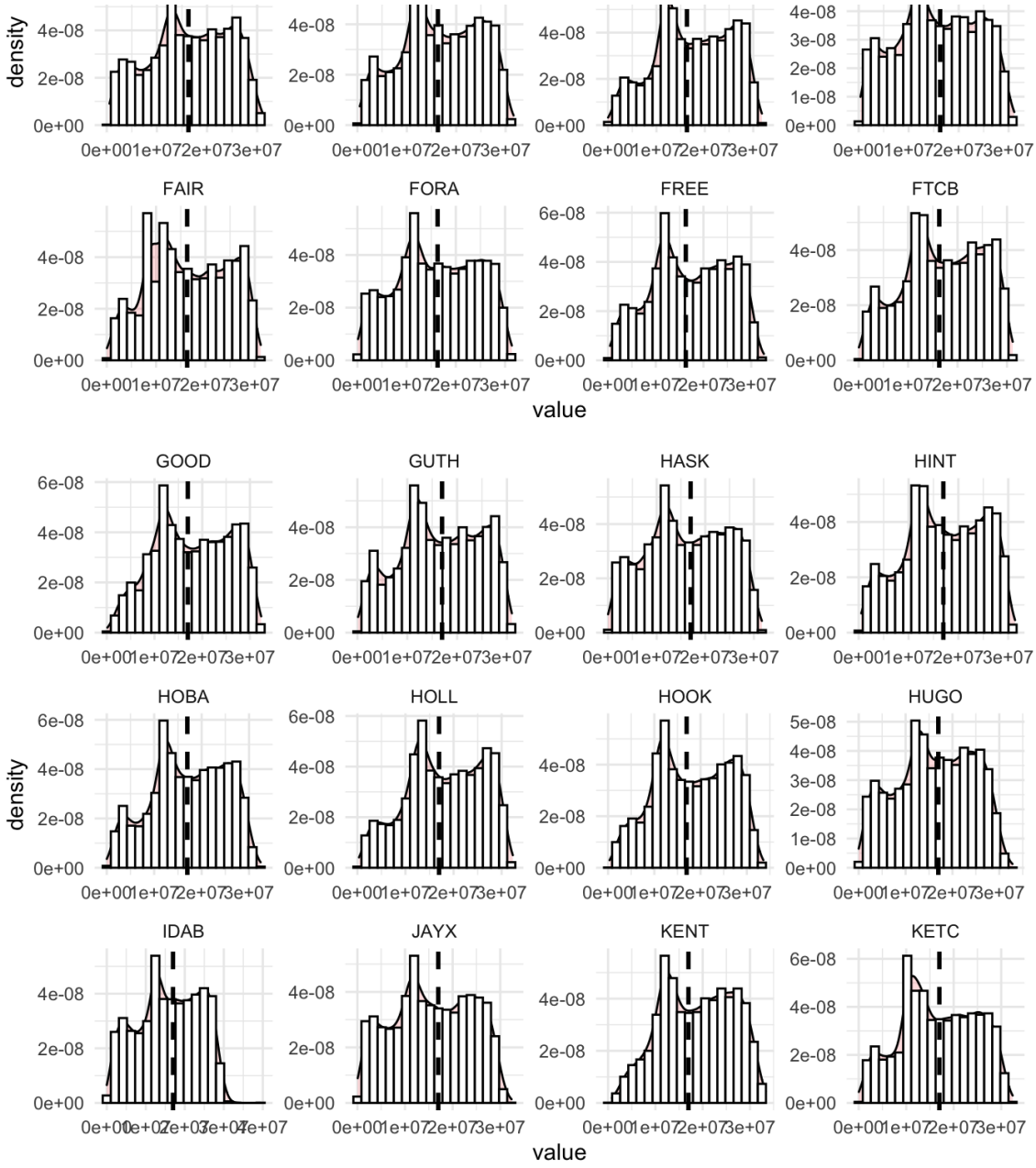
As mentioned earlier, our different features from the `solar_data` set went through a principal component analysis. When looking at some of the distributions of the features, it becomes obvious that almost all the data follows a near normal distribution. A normal distribution is something which is very important when analyzing data; it is the most important probability distribution because it describes many natural phenomena. Data with a normal distribution is applicable to a high variety of methods and conclusions of those data sets can be made more easily.

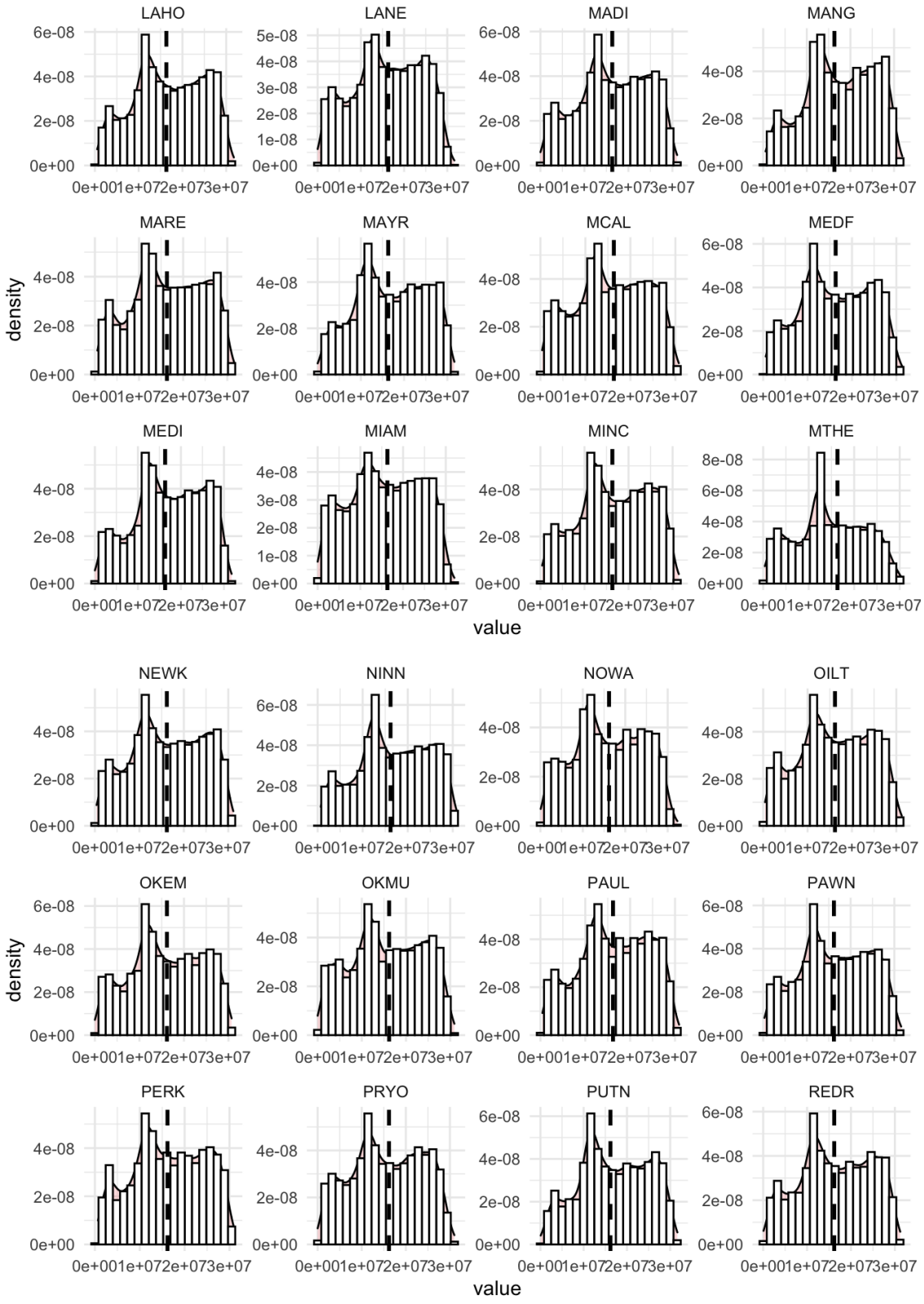
On the other hand, it becomes obvious that we cannot find a special pattern when looking into the distribution of the independent variables.

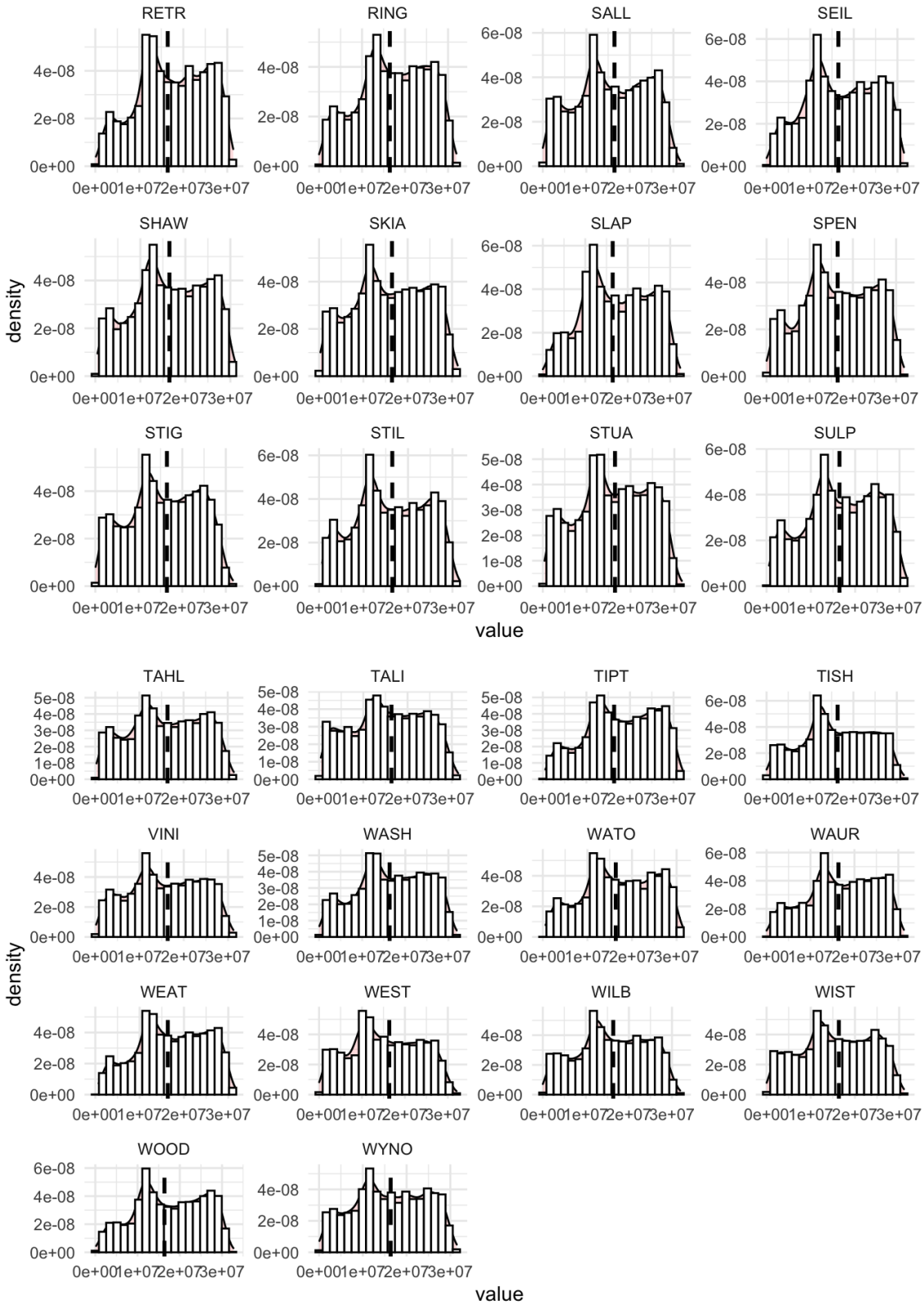
See below the distribution of both, independent and dependent variables:

Distribution of dependent variables

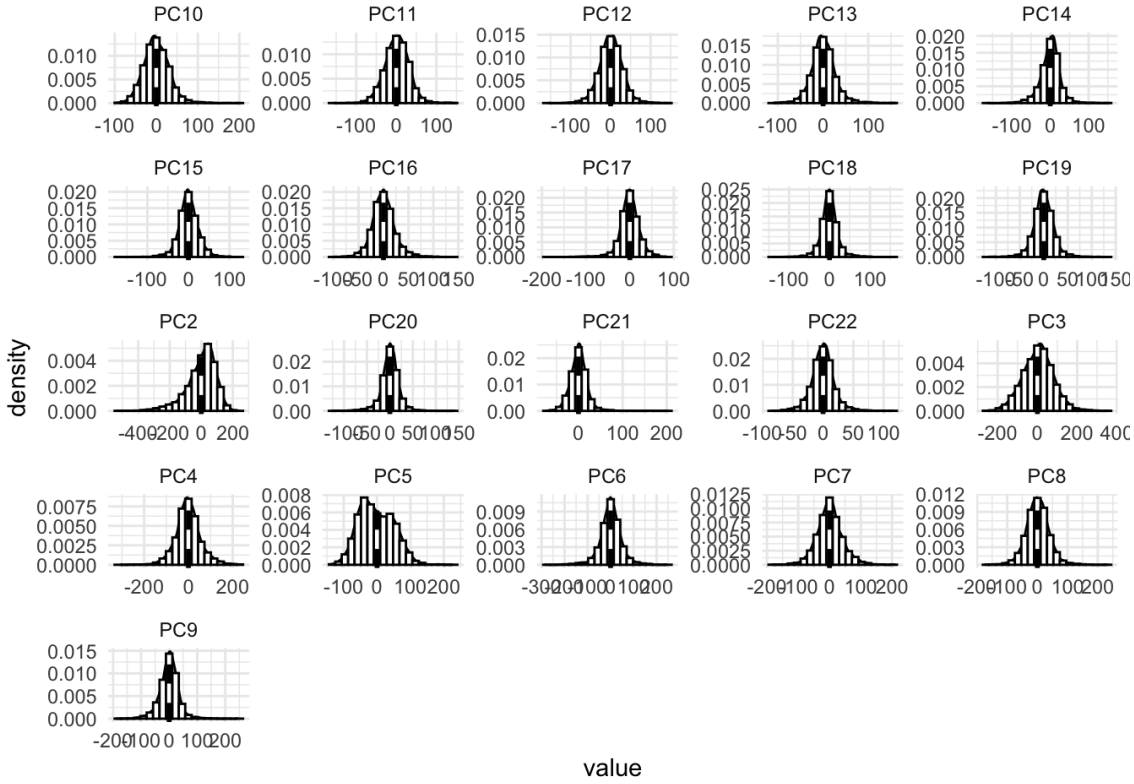

ACME
ADAX
ALTU
APAC
ARNE
BEAV
BESS
BIXB
BLAC
BOIS
BOWL
BREC
density
value
BRIS
BUFF
BURB
BURN
BUTL
BYAR
CAMA
CENT
CHAN
CHER
CHEY
CHIC
CLAY
CLOU
COOK
COPA
DURA
ELRE
ERIC
EUFA


density
FAIR
FORA
FREE
FTCB
value
GOOD
GUTH
HASK
HINT
HOBA
HOLL
HOOK
HUGO
IDAB
JAYX
KENT
KETC
value


LAHO
LANE
MADI
MANG
MARE
MAYR
MCAL
MEDF
MEDI
MIAM
MINC
MTHE
density
value
NEWK
NINN
NOWA
OILT
OKEM
OKMU
PAUL
PAWN
PERK
PRYO
PUTN
REDR
density
value

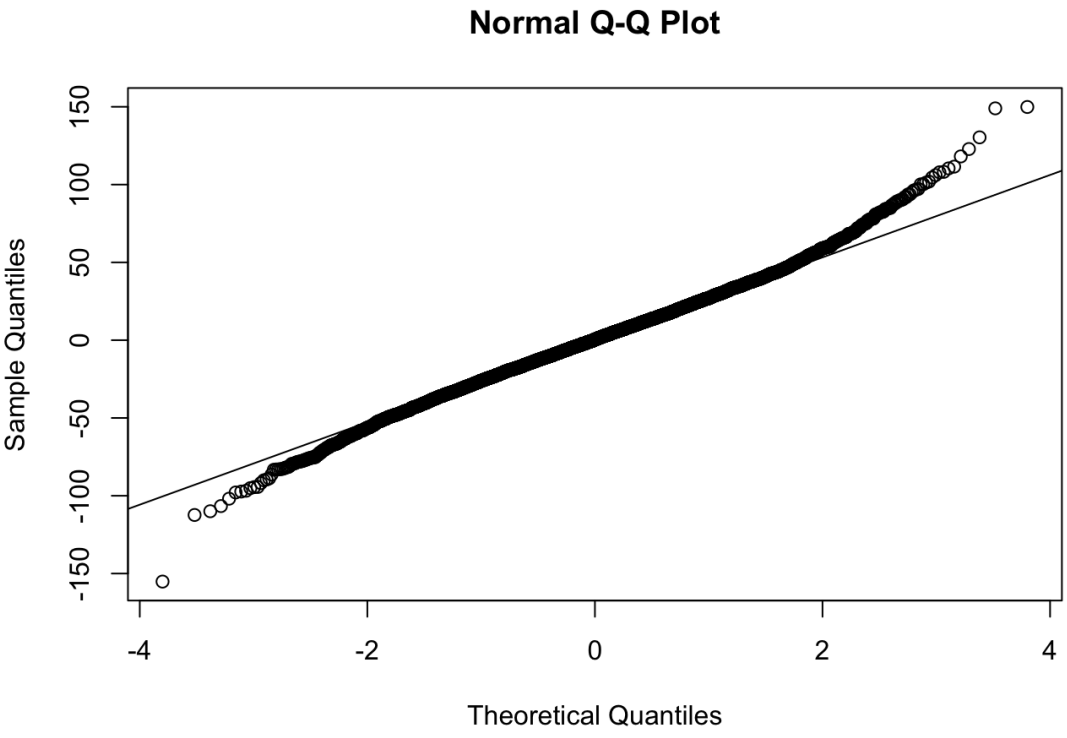


Distribution of independent variables



To further analyze the distribution of the pre processed features we could also have a look at the QQ plot which is a graph that can be used to test a variable for the presence of a normal distribution. If the observations lie close to the line, the distribution of the given data has the same shape as the normal distribution.

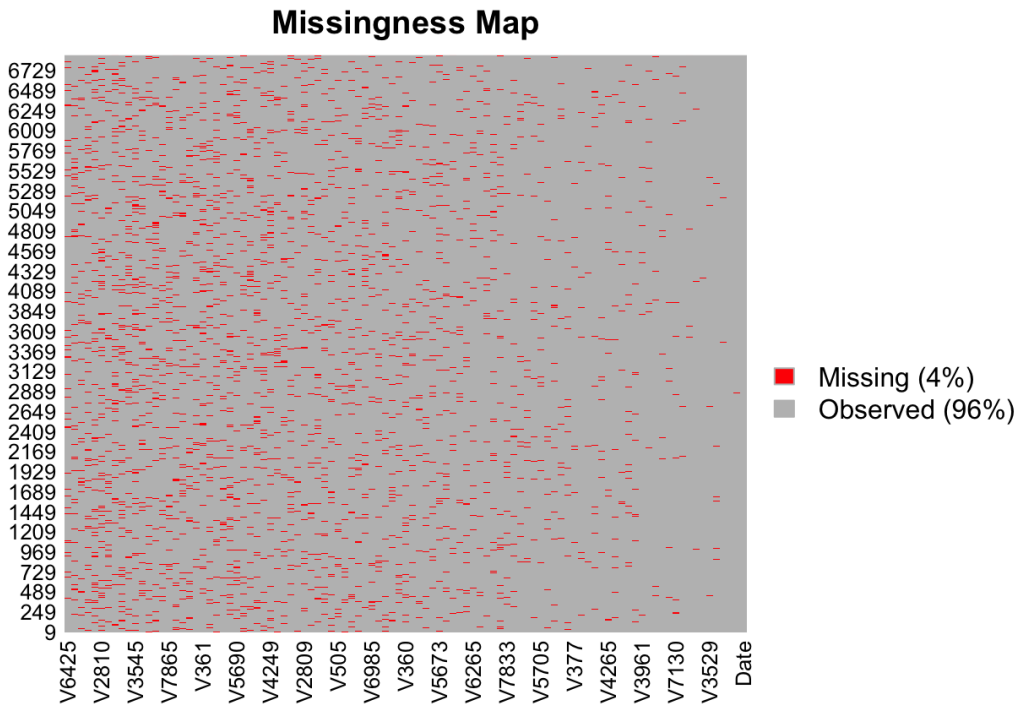
Here is an example for the feature `PC12`.



Data Handling

Throughout our exploratory analysis we also focused on handling outliers and missing values.

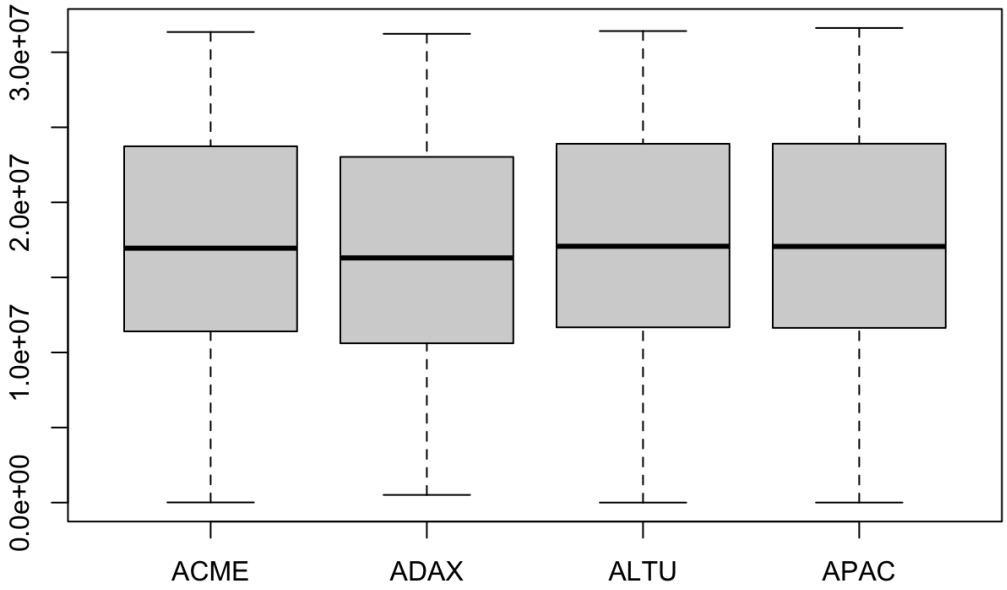
For the part of missing values, we found that there were only some in the additional data set given for the task. Below a visualization of the missing values in the data set:



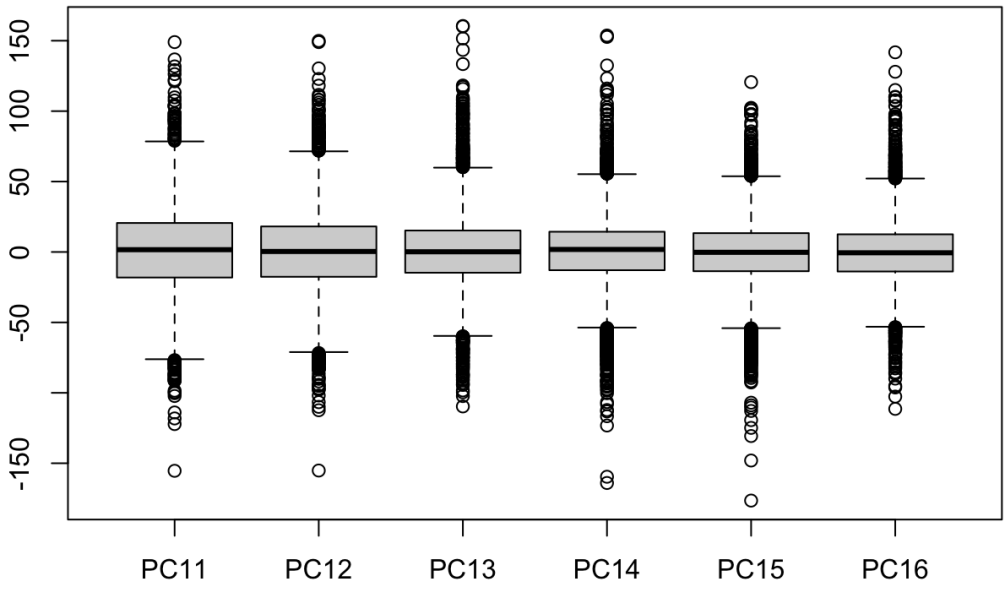
As can be seen in the **Missingness Map** we only have 4% missing data. Therefore, we decided to impute those missing values by means of the Miss forest package. This imputation algorithm fits a random forest in order to predict the missing values.

Furthermore, we analyzed how many outliers we obtain per column. All in all, we found that there were no outliers in the dependent variable and very few in the pre processed features. On the other hand, the additional data was characterized by a high number of outliers.

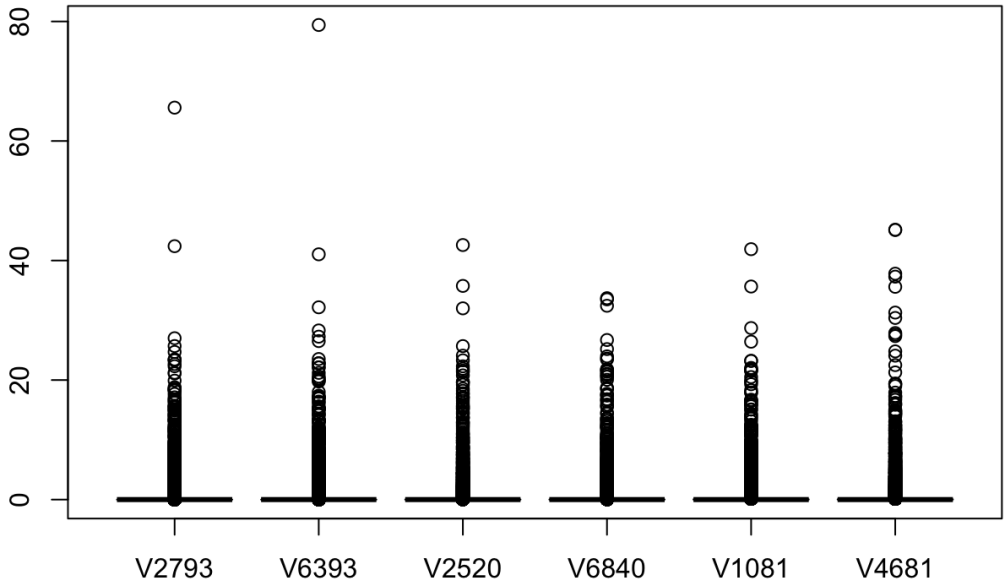
Example boxplot of dependent variables



Example boxplot of pre-processed features



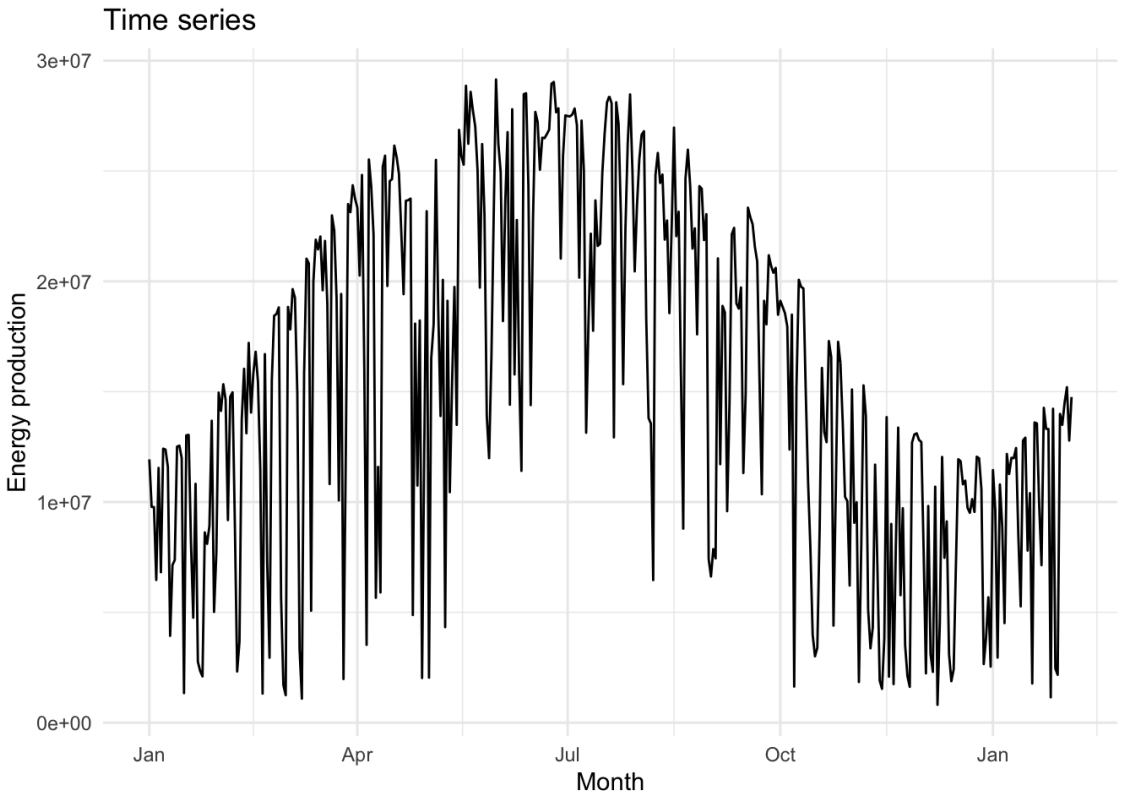
Example boxplot of additional features



Seasonality

As we have data for each day and season, we also wanted to look at whether energy production is constant throughout the year or fluctuates from time to time. To do this, we looked at a sample year and a sample solar panel site and plotted the production throughout that year. This graph clearly shows that production fluctuates over time which is understandably, as there is more sun in summer than in winter. This is a very insightful information because we can use it to generate a new

categorical feature indicating in which season or month the solar energy was generated.



The Model

After performing all this explanatory data analysis, it was time to fit a model. In order to predict the solar product at the 98 weather stations for the dates after 31/12/2007, we chose to implement a Gradient Boosting Machine (GBM) algorithm. We chose this algorithm mainly due to its ability to use Laplace in it's distribution argument, which uses absolute loss to calculate error, and is therefore perfect since MAE is the evaluation metric of the competition. Additionally, GBM is a good choice because it trains many models in a gradual manner, meaning it is able to come up with a good solution even with little hyperparameter tuning.

The data we used for our model consisted of all the PC columns from the solar data set, as well as an additional column that was made from the date column, but only keeping the month. We added this data column in order to include the seasonality factor explained earlier in the report. We chose to add all the principle components, as we found that adding all of them lead to the smallest MAE when we scored our model on the test data we kept when training the model. We also tried adding in some of the additional variables, but found that this worsened our model. Another aspect we tried in order to attempt our model was removing any of the outliers in the training data. This however again worsened our model. It makes sense that in a case like solar product data there will be many "outliers", since the amount of sunshine per day can vary greatly. In the end, we therefore chose to keep the outliers in the data.

Due to time constraints, we were only able to try a few different hyperparameter combinations, however had more time allowed, we would have definitely tried to tune these values even more. Additionally, we tried adding more trees over time (which increased our score significantly each time), however with the computational power of our computers and the time constraints the max number of trees we were able to include in the model was 1000. Had we had more time or computer power, we would have increased this number further. With the resources we had available, our best model was a GBM model with hyper parameters of 1000 trees, an interaction depth of 6 and a shrinkage of 0.05. With this model we achieved an overall MAE of 2410799.85720 when submitting our predictions in the Kaggle competition.

References

Built In. 2020. A Step-By-Step Explanation Of Principal Component Analysis. [online] Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (https://builtin.com/data-science/step-step-explanation-principal-component-analysis) [Accessed 20 December 2020].

En.wikipedia.org. 2020. Renewable Energy Industry. [online] Available at: https://en.wikipedia.org/wiki/Renewable_energy_industry (https://en.wikipedia.org/wiki/Renewable_energy_industry) [Accessed 23 December 2020].

Medium. 2020. Gaussian Distribution: Why Is It Important In Data Science And Machine Learning?. [online] Available at: https://medium.com/ai-techsystems/gaussian-distribution-why-is-it-important-in-data-science-and-machine-learning-9adbe0e5f8ac (https://medium.com/ai-techsystems/gaussian-distribution-why-is-it-important-in-data-science-and-machine-learning-9adbe0e5f8ac) [Accessed 22 December 2020].

Rpubs.com. 2020. Rpubs - Missforest - Missing Data Imputation. [online] Available at: https://rpubs.com/lmorgan95/MissForest (https://rpubs.com/lmorgan95/MissForest) [Accessed 24 December 2020].

Singh, H., 2020. Understanding Gradient Boosting Machines. [online] Medium. Available at: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab (https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab) [Accessed 25 December 2020].