7 Ethics in data-driven storytelling

Nigel Sjölin Grech

7.1 Ethical concerns regarding data acquisition

For this section I will consider the questions of provenance, in regards to intent and completeness and accuracy. I also consider quantification in regards to what is being measured and the ambiguity of the measurements.

When considering the wind power forecasting data set we can't speak much about the provenance of the data since it is not given. On Kaggle's usability score it is marked as not having any provenance. Since the data's origin is unknown we cannot be sure that it was originally collected for the purposes of prediction, we can say that was the intention of the kaggle member who posted the data.
We also know that the data is incomplete if this was by action of the uploader or how the data originally looked it is impossible to say since there is no metadata describing that. Certain columns such as WTG seem to suggest that the original data contained data on a number of turbines but only one was supplied.
Without knowing anything about the maintenance or care of the equipment and sensors that collected the data we can't say that it is all accurate. Is a gap in the data when a sensor broke down and can we trust the values leading up to that gap? The lack of information about the provenance of the data and the conditions it was collected under will lead to doubt in results obtained from it.

One failure of this data set is the ambiguity of what is being collected. The columns are not well described or given units of measurements. For the most part I had to take my best guess based on other works using the data and common scientific units used in relation to wind energy. It would be completely unethical to rely 100% on any statistical results or modelling from data that one cannot describe completely.

I believe that this data was not collected with enough care and would not recommend its use within publications. I also believe that it would be a breach of ethics to use the data (as it currently is) as the basis of some data science product.

7.2 Ethical concerns regarding data transformation

For this section I will discuss the topics of aggregation, analytic approach used and anonymisation.

In the course of my work I aggregated and resampled the data over arbitrary time periods such as days and months, I made use of the means of these aggregates to fill missing values. Given the large amount of missing data and the relatively small period of time that the data is consistent for, in this case a year, it can raise the question: did I create an artificial

skew in the data? There is also the question of resampling. The larger the period of time we reasample over the more variance will be eliminated from the data. On one hand we make it easier to interpret, on the other we can lose valuable features of the data. I believe that resampling on a daily level was appropriate for the purposes of EDA, however I may use a smaller window when building a model.
For the analysis carried out a more traditional EDA approach was used, however a more accurate and robust work could have been achieved using time series analysis techniques. While no information on the location of the wind turbine was given since there is information regarding wind speed and direction it could be possible to locate the turbine by cross referencing the data with historical weather data and known locations of wind turbines. If the location and origin of the data was abstracted for some security reasons this could represent a breach.

7.3 Ethical concerns regarding conveying and connecting insights

This data set did not have much in terms of story potential. It presented for the most part as expected, however I did take note of a plateau when plotting active power against wind speed. This gave me the opportunity to look at how we can be misled by plots and why it is important to confirm our assumptions with statistical testing.