Individual portfolio

Michal Bartek

In this document I reflect on each Personal learning goal, provide the clear descriptions of the workflows and the results for each of the goal. In the reflection for each goal I add the percentage of the extent to which I feel I have met my expectations. 100% meaning the goal was fulfilled exactly how I imagined it to be after the course and 0% being absolutely unmatched to the goal set from the beginning. Additionally, I reflect on the whole project and the learning project (including the BCCs).

I also add the github page with all my scripts I have developed throughout the course to improve my data science skills.

Link to github with the scripts and formal submission of the Individual Portfolio:

https://github.com/MichalBartek-14/DSfSE_Portfolio

G1: Learn how to work with new python libraries for machine learning, for example: scikit learn. Improve and automate the skills with the pandas, geopandas python libraries.

Background

In the first goal, I outlined the most important theme I wanted to improve throughout the course Data Science for Smart Environments - *Programming in Python*. I had already used Python for simple tasks and scripting assignment, however, I have never applied it to a complex project with real-world datasets. Before the course, I recognized the opportunity to apply machine learning libraries to a project. Since I had limited experience with such Python libraries, I mentioned this in my first learning goal. Additionally, the content of the workshops was designed to boost the students' scripting skills, so I dedicated significant attention to this goal.

Process

I aimed to improve my coding skills throughout the course using tools such as Datacamp and available Jupyter notebooks to train myself in scripting outside the project. The process of implementing coding skills in the project began in the first week when we had to download and format much of our data. I gained new experience using an API to gather the required data and practiced basic data formatting with *geopandas*. Subsequently, I used the *xarray* library to read the .nc files resulting from the API download. To understand and work with the new libraries, I referred to external sources such as GitHub, Stack Overflow, and generative AI tools. The example code for the API request was sourced from the WIN50 website (Source 1).

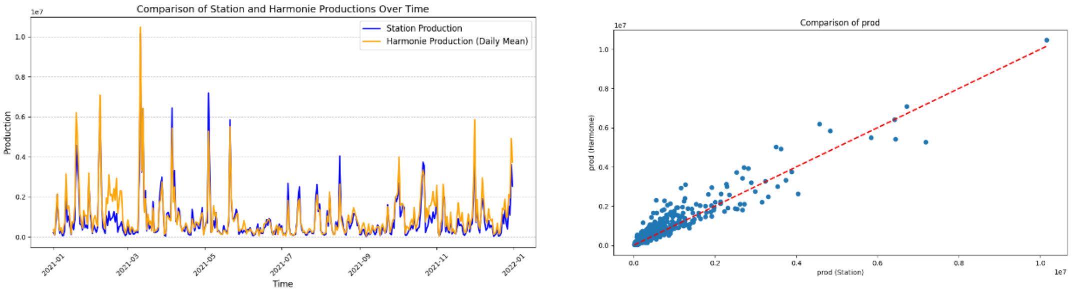
In the following days, my project team focused on further data wrangling, as we had to match multiple datasets spatially and contextually. This task proved more challenging than expected, requiring us to use GIS software instead of scripting. After GIS processing, I applied my skills in *geopandas*, *matplotlib*, and other libraries to visualize preliminary results.

Additionally, I attempted to implement object-oriented programming (OOP), which I had never used before, to automate some processes in my scripts. My motivation to explore OOP came from Jascha Grübel's lecture in Week 2.

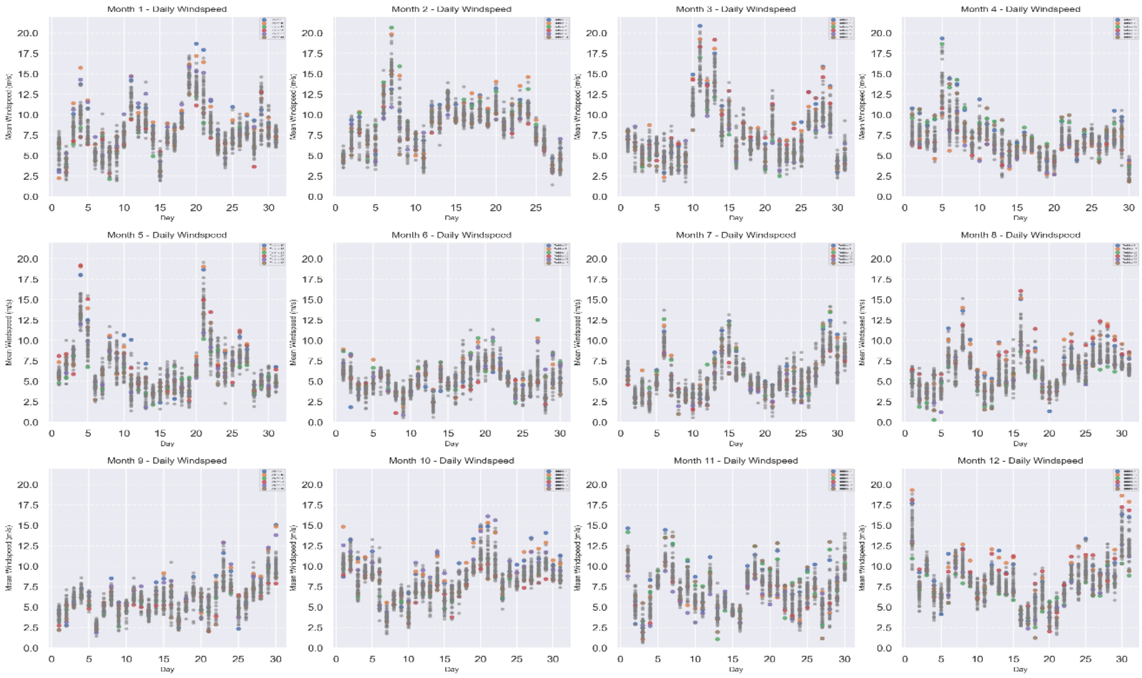
Lastly, I experimented with *scikit-learn* during the Week 2 tutorial and in a small test script for our project. However, the accuracy was underwhelming, and we did not proceed with using it for a prediction model due to certain limitations. I have learned about the available ML algortihms and dedicated time and effort to design an architecture for our project which we have eventually decided to not implement.

Results

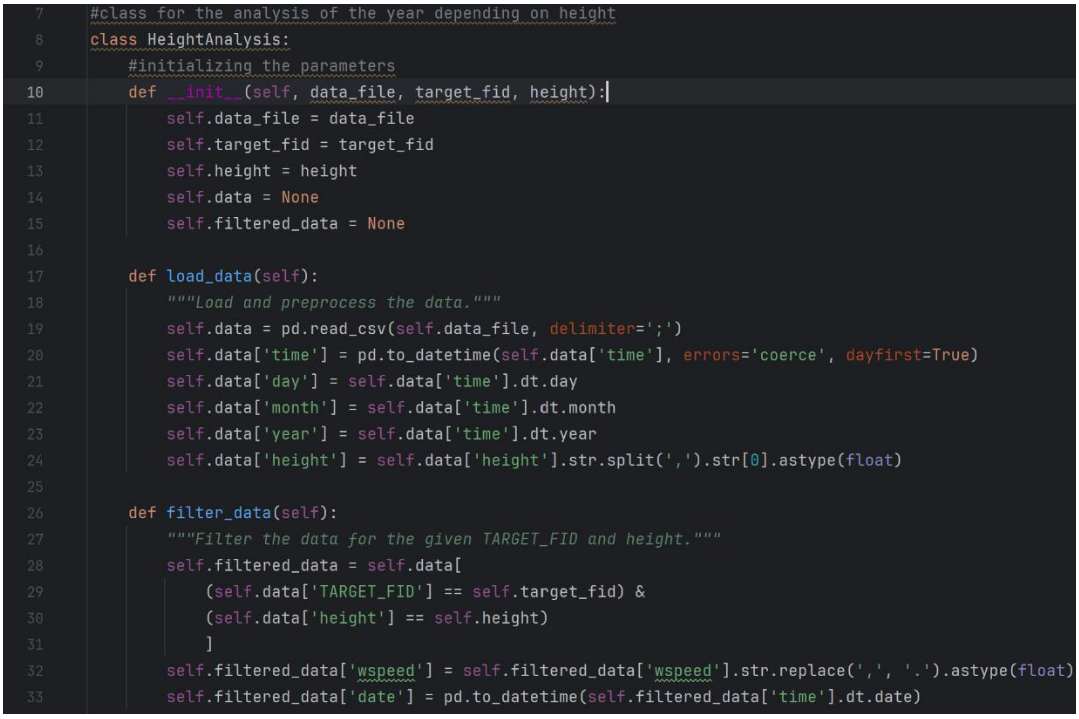
As a result, I produced multiple visualizations and core datasets that were used by other team members throughout the project. I was responsible for data acquisition, formatting the HARMONIE model, and merging and joining datasets. Furthermore, with my improved scripting skills, I created numerous visuals later incorporated into the final product. I provide a sample of the visuals that I have produced using the improved knowledge in this goal of my individual learning plan. Subsequently I add the sample of the python script using objected oriented programming (full python codes developed by myself are also available on github page) that was used to produce some of the visuals.



Comparison of the two datasets using pandas,numpy and matplotlib



Visual produced using Object oriented programming



```
#class for the analysis of the year depending on height
class HeightAnalysis:
    #initializing the parameters
    def __init__(self, data_file, target_fid, height):
        self.data_file = data_file
        self.target_fid = target_fid
        self.height = height
        self.data = None
        self.filtered_data = None

    def load_data(self):
        """Load and preprocess the data."""
        self.data = pd.read_csv(self.data_file, delimiter=';')
        self.data['time'] = pd.to_datetime(self.data['time'], errors='coerce', dayfirst=True)
        self.data['day'] = self.data['time'].dt.day
        self.data['month'] = self.data['time'].dt.month
        self.data['year'] = self.data['time'].dt.year
        self.data['height'] = self.data['height'].str.split(',').str[0].astype(float)

    def filter_data(self):
        """Filter the data for the given TARGET_FID and height."""
        self.filtered_data = self.data[
            (self.data['TARGET_FID'] == self.target_fid) &
            (self.data['height'] == self.height)
            ]
        self.filtered_data['wspeed'] = self.filtered_data['wspeed'].str.replace(',', '.').astype(float)
        self.filtered_data['date'] = pd.to_datetime(self.filtered_data['time'].dt.date)
```

Example of a script using Object oriented programming

**From the github page the script oriented towards the OOP is 'b04_HarmonieAnalysisOOP'*

Reflection

In summary, I improved my scripting skills and gained proficiency in standard Python libraries such as *pandas*, *numpy*, *matplotlib*, and *shapely*. I am satisfied with my implementation of object-oriented programming, which I successfully used on a small scale in our project.

However, I am less satisfied with my use of machine learning libraries. While the premade Jupyter notebooks in Week 2 helped my goal, I should have paired them with implementation in the project.

- **Traditional Data Science Libraries**: 100%
- **Object-Oriented Programming (Automation and Clarity of Scripts)**: 80%
- **Machine Learning Libraries**: 50%

Sources:

- WIN50 Website
- The youtube video on Learn Data Structures and Algorithms: YouTube Video

G2: Improve my Workflow. Improve the structuring of the workload and communicate it clearly with the team.

Background

As my second goal, I aimed to establish a clear workflow and communication within the team. This goal was set to define the boundaries of my competencies and responsibilities in the data project. Over the three-week span, it was essential for me to understand my role in the team and what was expected of me. I wanted to communicate these expectations clearly with the team.

I was assigned to a great team of students who provided me with opportunities to grow and learn during the process. Additionally, I was given the chance to lead the scripting part of the project, which became key toward the end. Communication played an important role in this goal, as all steps needed to be conveyed clearly.

Process

Since this goal was dedicated exclusively to the team project, I began by sharing my ideas and strengths with the team from the outset. During data acquisition, I took responsibility for acquiring the HARMONIE data via API. After successfully downloading this data, I was appointed to assist with the coding part of the project.

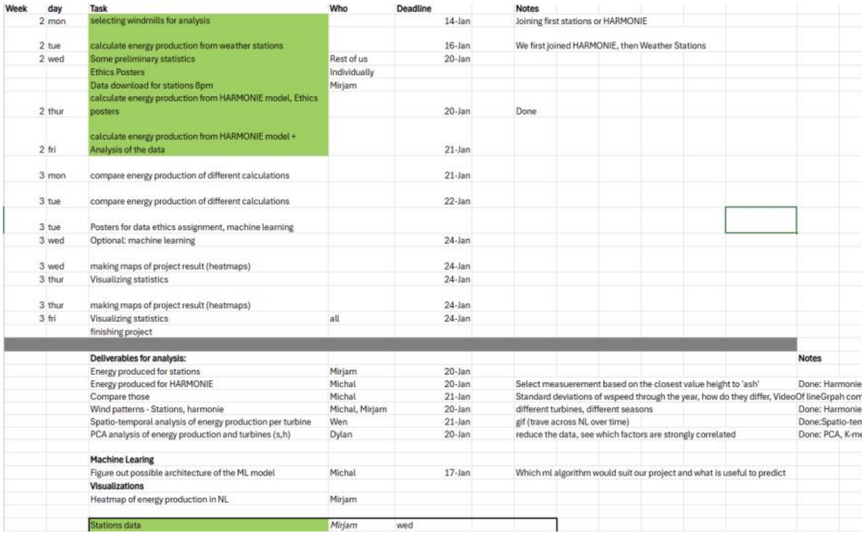
I clearly defined my workload, working during the afternoons and additional off-school hours to provide a foundation for other team members if needed. This primarily involved preparing core datasets by cleaning the data. This phase of clearly defined tasks was mostly limited to the first two weeks of the project.

In the latter stages, the boundaries became less clear, and my workload was not efficiently balanced due to issues in other areas of the team project. I stepped in to help other team members solve technical problems. As the deadline approached, team dynamics changed, and cross-functional roles emerged to achieve satisfactory results.

Results

The key results of this goal included daily meetings where tasks were divided, and workflows were communicated. A more tangible outcome was the planning sheet, where we set deadlines and documented notes from each team member.

I provide the example of the planning sheet, which was complimented by daily physical meetings with my teammates.



Week	day	Task	Who	Deadline	Notes	
2	mon	selecting windmills for analysis		14-Jan	Joining first stations or HARMONIE	
2	tue	calculate energy production from weather stations		16-Jan	We first joined HARMONIE, then Weather Stations	
2	wed	Some preliminary statistics	Rest of us	20-Jan		
		Ethics Posters	Individually			
		Data download for stations 8pm	Mirjam			
2	thur	calculate energy production from HARMONIE model, Ethics posters		20-Jan	Done	
2	fri	calculate energy production from HARMONIE model + Analysis of the data		21-Jan		
3	mon	compare energy production of different calculations		21-Jan		
3	tue	compare energy production of different calculations		22-Jan		
3	tue	Posters for data ethics assignment, machine learning				
3	wed	Optional: machine learning		24-Jan		
3	wed	making maps of project result (heatmaps)		24-Jan		
3	thur	Visualizing statistics		24-Jan		
3	thur	making maps of project result (heatmaps)		24-Jan		
3	fri	Visualizing statistics	all	24-Jan		
		finishing project				
		Deliverables for analysis:				**Notes**
		Energy produced for stations	Mirjam	20-Jan		
		Energy produced for HARMONIE	Michal	20-Jan	Select measuerement based on the closest value height to 'ash'	Done: Harmonie
		Compare those	Michal	21-Jan	Standard deviations of wspeed through the year, how do they differ, VideoOf lineGrpah com	
		Wind patterns - Stations, harmonie	Michal, Mirjam	20-Jan	different turbines, different seasons	Done: Harmonie
		Spatio-temporal analysis of energy production per turbine	Wen	21-Jan	gif (trave across NL over time)	Done:Spatio-ten
		PCA analysis of energy production and turbines (s,h)	Dylan	20-Jan	reduce the data, see which factors are strongly correlated	Done: PCA, K-me
		Machine Learing				
		Figure out possible architecture of the ML model	Michal	17-Jan	Which ml algorithm would suit our project and what is useful to predict	
		Visualizations				
		Heatmap of energy production in NL	Mirjam			
		Stations data	*Mirjam*	wed		

Example of the planning sheet in our project group

Reflection

In the initial stages, I successfully met my goal of clearly communicating my agenda and deadlines. However, in the final stages, I realized that it is impossible to fully anticipate the complexity of a project and the shifting dynamics of team workflows.

For future projects, I understand the importance of being flexible with competencies and skill sets, as this is a crucial attribute of a data scientist in a team. Additionally, I now fully appreciate the importance of team communication and the conceptualization of each team member’s agenda.

- **Communication with the Team**: 90%
- **Structuring Workflow/Workload**: 80%

G3 Information: How to setup an analysis

Background

For this goal, I aimed to gain more insights into statistical analysis of data. Before the project, I hoped to conduct complex analyses; however, I underestimated the complexity of the data project, where analysis represented only a small part of the work. Since I had not yet encountered the dataset, it was difficult to define this goal more clearly than simply aiming to set up statistical analyses. After successfully formatting the data, I worked on this goal, though much later than initially anticipated.

Process

Due to the project schedule and limited implementation time, the analysis phase only began in the latter part of Week 2 and Week 3. At this point, the team and I discussed meaningful analysis ideas using the available data. We concluded that we would compare the HARMONIE model with real-time data from 2021.

I approached this problem by first familiarizing myself with the datasets, then plotting correlations and relationships within the data, and finally creating a regression model for comparison.

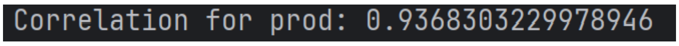
To practice analysis techniques outside the project, I watched publicly available video content on statistical analysis in data science and used generative AI tools to explain different techniques, including machine learning algorithms.

From the course content, I particularly enjoyed and learned from the clustering lecture in Week 3 and the Week 2 machine learning tutorials with their respective Jupyter notebooks.

Results

For our project, I created correlation matrices, regression models, and other graphs that best represented the data and relationships within it. I also developed a regression model comparing the HARMONIE data and weather station data, which was ultimately used as the final model for our presentation.

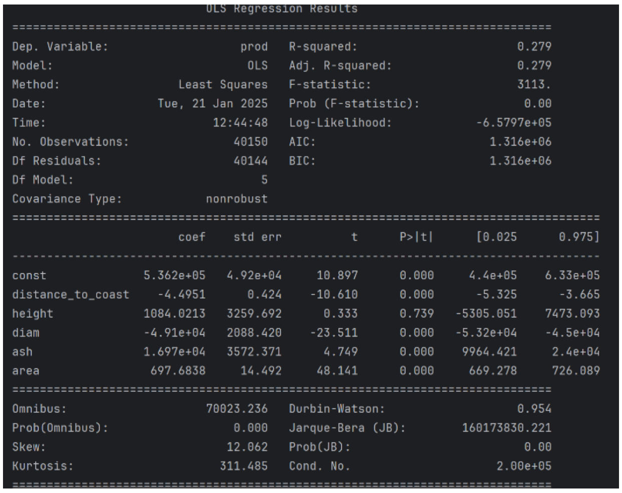
Here I present the correlation between the model and station data which showed high correlation.



```
Correlation for prod: 0.9368303229978946
```

Example of the python output for the correlation analysis

Additionally, I add the screenshot of one of the attempts to produce machine learning algorithm to predict the energy production with low overall accuracy and achieved R2



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   prod   R-squared:                       0.279
Model:                            OLS   Adj. R-squared:                  0.279
Method:                 Least Squares   F-statistic:                     3113.
Date:                Tue, 21 Jan 2025   Prob (F-statistic):               0.00
Time:                        12:44:48   Log-Likelihood:            -6.5797e+05
No. Observations:               40150   AIC:                         1.316e+06
Df Residuals:                   40144   BIC:                         1.316e+06
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const              5.362e+05   4.92e+04     10.897      0.000     4.4e+05    6.33e+05
distance_to_coast    -4.4951      0.424    -10.610      0.000      -5.325      -3.665
height             1084.0213   3259.692      0.333      0.739   -5305.051    7473.093
diam               -4.91e+04   2088.420    -23.511      0.000   -5.32e+04    -4.5e+04
ash                1.697e+04   3572.371      4.749      0.000    9964.421     2.4e+04
area                697.6838     14.492     48.141      0.000     669.278     726.089
==============================================================================
Omnibus:                    70023.236   Durbin-Watson:                   0.954
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        160173830.221
Skew:                          12.062   Prob(JB):                         0.00
Kurtosis:                     311.485   Cond. No.                     2.00e+05
==============================================================================
```

The output for ML algorithm to predict the energy production, which was not eventually used.

Reflection

I improved my skills in statistical analysis and learned to visualize and interpret data effectively. However, I recognize that my experience with advanced statistical techniques, such as machine learning algorithms, remains limited. While the tutorials and videos enhanced my understanding, applying these techniques in the project would have better reinforced my learning.

- **Skill in Statistical Analysis**: 70%
- **Advanced Statistical Techniques (e.g., ML Algorithms)**: 60%

Sources:

- Polynomial Regression Concept Video

G4 Knowledge: Data ethics, the consequences of my actions in data

Background

For my final personal goal, I aimed to expand my knowledge of data ethics, a topic I had not explored much before. I expected this course to help improve my understanding through interesting historical examples and reflections on ethics and morals within our project.

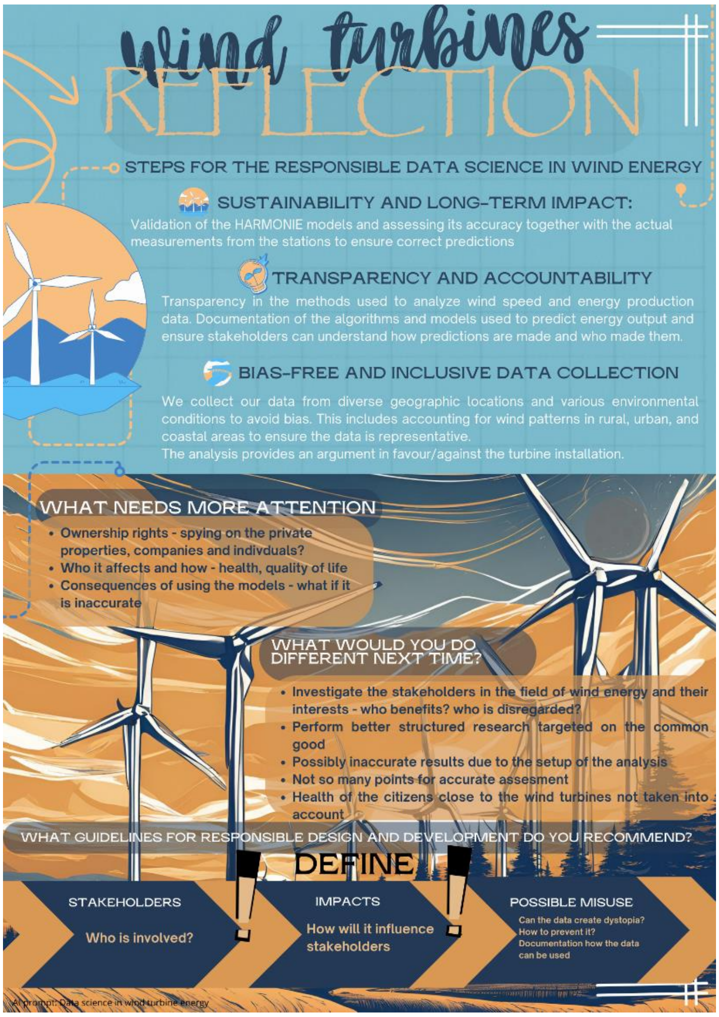
Process

The course's emphasis on this topic helped me achieve my goal, as multiple lectures and assignments directly addressed data ethics. I improved my understanding of good practices in data and algorithms, as well as the potential negative effects of bad practices.

I gained significant insights from the lecture on January 9, 2025, and a prerecorded session on epistemic and normative concerns. I applied this knowledge to create a poster reflecting the ethical considerations in our project, based on materials from the lectures. I also discussed the outcomes with my team and shared ideas during the Week 4 sharing session.

Results

The primary outcome of this goal was the poster, which reflected on the ethical practices followed in our project and areas for improvement. Additionally, I compiled notes from lectures and videos that helped me organize my understanding of data ethics in data science. I add the poster I have created for the data ethics discussion.



wind turbines

REFLECTION

STEPS FOR THE RESPONSIBLE DATA SCIENCE IN WIND ENERGY

SUSTAINABILITY AND LONG-TERM IMPACT:

Validation of the HARMONIE models and assessing its accuracy together with the actual measurements from the stations to ensure correct predictions

TRANSPARENCY AND ACCOUNTABILITY

Transparency in the methods used to analyze wind speed and energy production data. Documentation of the algorithms and models used to predict energy output and ensure stakeholders can understand how predictions are made and who made them.

BIAS-FREE AND INCLUSIVE DATA COLLECTION

We collect our data from diverse geographic locations and various environmental conditions to avoid bias. This includes accounting for wind patterns in rural, urban, and coastal areas to ensure the data is representative.
The analysis provides an argument in favour/against the turbine installation.

WHAT NEEDS MORE ATTENTION

- Ownership rights - spying on the private properties, companies and indivduals?
- Who it affects and how - health, quality of life
- Consequences of using the models - what if it is inaccurate

WHAT WOULD YOU DO DIFFERENT NEXT TIME?

- Investigate the stakeholders in the field of wind energy and their interests - who benefits? who is disregarded?
- Perform better structured research targeted on the common good
- Possibly inaccurate results due to the setup of the analysis
- Not so many points for accurate assesment
- Health of the citizens close to the wind turbines not taken into account

WHAT GUIDELINES FOR RESPONSIBLE DESIGN AND DEVELOPMENT DO YOU RECOMMEND?

DEFINE

STAKEHOLDERS
Who is involved?

IMPACTS
How will it influence stakeholders

POSSIBLE MISUSE
Can the data create dystopia?
How to prevent it?
Documentation how the data can be used

AI prompt: Data science in wind turbine energy

Poster reflecting the data ethics in our project.

Reflection
I acquired a substantial amount of knowledge about ethical practices in data science. I now grasp the importance of concepts such as consent and privacy in the context of data. I believe I fully met the expectations and objectives I set for this goal.

- **Understanding Data Ethics and Consequences of Practices**: 100%

Reflection on the learning process (with BCC):

The team I had the chance to work with during this project was a very good way to meet my learning goals. Everybody was helpful and enthusiastic about the project which allowed for a smooth process.

As one of the BCC I would highlight **different scientific backgrounds** we came from. I had more experience with the coding and acquiring the data which helped us during the initial stages of the project. On the other hand, my teammates had more experience with statistical analysis and physics regarding our project topic - wind turbines. Here we complimented each other well and were able to teach each other new skills. The perseverance and enthusiasm of all the team members meant we did not mind working more on certain parts of the project, since everybody showed clear commitment.

Another boundary-crossing challenge in the project was **cultural differences**. Two team members were Dutch, while the other two, including myself, were international students. These differences extended not only to our experiences with the topic of wind turbines but also to communication styles within the project. Fortunately, this did not cause significant issues, as we made an effort to communicate our expectations and areas of expertise beforehand. Through open and transparent communication, we were able to address any challenges and align our efforts effectively across our cultural backgrounds.

Reflection on the Project:

The project has provided a strong basis for me to train and practice my skills in a real world scenario. However, the usability and potential of the outcomes of our group project can be questioned. In the field of wind energy much of the research has already been done before. We treated our project mostly to replicate and confirm already known facts, rather than come up with groundbreaking research in this topic. The outcomes we have produced are already known and well embedded in the theme. We have confirmed the accuracy of the HARMONIE model on locations within the Netherlands and found relationships between the energy produced from wind turbines and influential predictor factors such as location of the turbine or height of the turbine. Overall I am satisfied with our project results since we have successfully confirmed known information and simultaneously have practiced on the real datasets within the data science team.

The daily log of the workload from the entire course:

*from week2 I have started to attribute the workload of each day to the specific learning goals.

Week1

Day 1.1
Getting myself familiar with the projects
Brainstorming possible data issues we can solve in a project during the course
Day 1.2:
Studying the documentation of the data featured in the description of the project.
Learning the downloading system of KNMI and Harmonie versions
Studying the API tokens
Copy/making a script that downloads the data from HARMONIE.
Problem with the size of the data.
Day 1.3:
Acquaintance with the jupyter notebooks
Continuous search for ideal data
Conversion of the nc file into the shp file
filtering the data
Day 1.4:
Exploring the capabilities of the API call for the wins50 data

The matching of the data from the windmills to the Netherland geography
Preparing the presentation for the project proposal
Day 1.5:
Personal learning plan 10:00 - 12:00
14:00 - 18:00- Working on the selection of the windmills only within the administrative boundaries of the Netherlands using the geojson file of the Netherlands.
matching the windmills to the measurements of the Harmonie model

Week2

Day 2.1:
The intro to the Databases
ex01 and ex03 on the databases
A look into the github on APIs. Interest found in looking into the publicly available apis
https://github.com/public-apis/public-apis
Day 2.2:
lecture on ML and data wrangling
Learned new ways how to deal with the NaNs
Object oriented programming use in the ML problem.
Struggling with the ArcGIS pro to match the turbines and stations/harmonie
Managed to match the harmonie and turbines into one dataset
Communication in team about further progress **G2**
Communication about the ethics posters and preliminary analyses
Day 2.3:
Work on the project. Preliminary analysis and statistics.
Analysing the windspeeds at 10m height. Developing python code to correctly read and plot the windspeeds throughout the year. Pandas library **G1**
Working on the ethics poster.
Day 2.4:
making a setup for the day: Watch the content of the day, make notes, review ML day, work on the project and explore the data further.
Identifying the potential to use OOP. Making the class of height wind turbine analysis.
Producing the graph for the analysis of wind turbine 1 and different windspeeds at different heights during the year.
Setting up an analysis to statistically see if there is relationship between heights and windspeeds in the example of WT1. **G3**
Resolving the issue of the missing wind turbines 13 and 38
Day 2.5:
Presentations
Figuring out which ML could suit our project. Structuring the further process in ML.G3
The idea to derive "distance to coast variable" as a predictor for the energy/wspeed.
The algorithm with the best fit for us "random forest" a s a regression problem.
First ideas of the code and library we need to use - scikit learn G1
Plotting additional analyses on different wind turbines to gain more data insights.

Week3

Day 3.1:
Lecture on ethics - making notes, connecting the concepts to our project. **G4**
Scripting the plot of all the measurements for all the turbines throughout the year. Playing around with the classes and practising the structure of Object-oriented programming. **G1**
Discussing some formatting issues with the team. **G2**

Working on the selecting the measurements closest to the ash of the windmills

Day 3.2:
The clustering lecture
Working on the column distance to coast which would be one of the predictors for the models.
Trunks out to be more difficult than thought. Problems with the coordinate system, defining the coast from the json file. Eventually got the variable distance to coast in meters **G1**
Working on the regression to investigate the preliminary relationship. **G3**

Day 3.3:
Visualisation lecture.
The correlation plot. **G3**
Struggling with the joins. Ethics posters finish. **G4**
Making decisions with the team regarding the station data (proceed with the limited amount of data or try more techniques to match it) .
Formatting the station data.
Trying to utilize the station data with the code made for HARMONIE.
Struggle -> different variable formats, names, structuring.

Day 3.4:
Stations data analysis extra quickly to catch up with the rest of the project.
The comparison analysis of the Stations and the HARMONIE. **G3**
The ethics debate with other groups. **G4**
Working on the presentation, setting the order, the people presenting. **G2**

Day 3.5:
Finishing the presentation.
Documenting the code and uploading it to Github.
Presentation.
Finishing the submission, zipping the final folder.