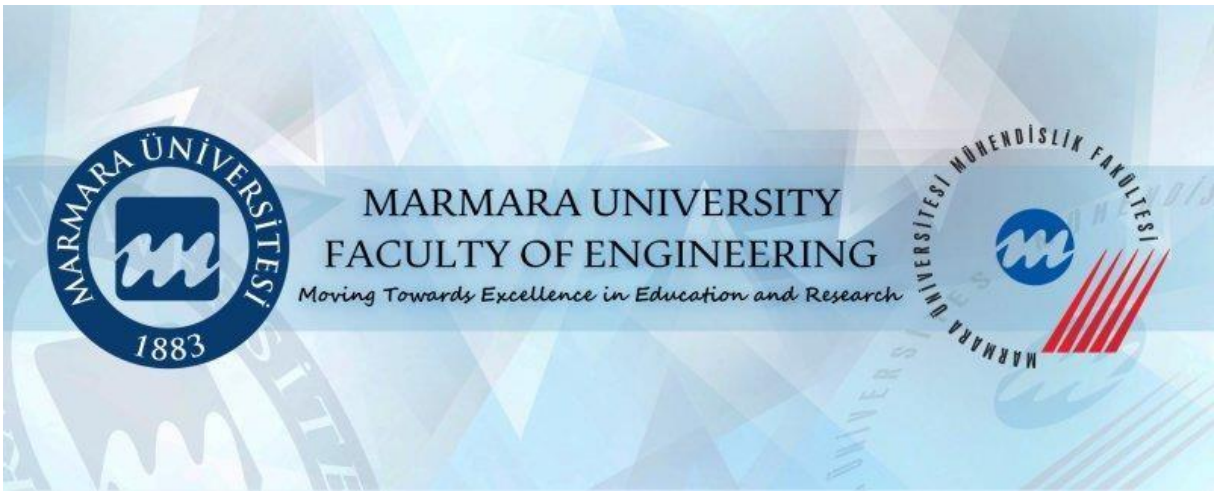

CSE4062 2020 SPRING PROJECT

ANALYSIS OF ELECTRIC PRODUCTION OF A WIND TURBINE

Delivery 2: Exploring Your Data

Group: 8

Name	ID Number	Mail Addresses
Furkan Can Ercan	150316044	furkanercan98@gmail.com
Muhammed Avcı	150414007	muhammedavci96@gmail.com
Yasin Gök	150115058	ygok.96@gmail.com

2) Exploratory Data Analysis

In this part of our project, we tried to analyze our dataset in a way, which will improve our understanding of the data and the general situation of our process that our dataset depicts. To do this we utilized python programming language and its very useful libraries such as Pandas,Seaborn and Scipy. We used Jupyter notebook platform to arrange our coding and analyses systematically.

Our main dataset is composed of measurements of a Wind Turbine located in Yalova Turkey. It explains a Wind Turbine's electric production process. It consists measurements from January 1st 00:00 to December 31st 2018 23:50 with 10 minutes gaps. Simply it includes measurements for every 10 minutes and these measurements are about electric production,therotical electric production as kilowatt/hour, Wind Speed as meter/second, wind direction angle as degree and date,time,month and day/night categories.

Even though it is a detailed dataset it does not include too many attributes to make our further purposes –regression analysis and building predictive models- easier. Therefore, we made some research to gather more data relevant with our processes.

After doing some researches we obtained data about weather conditions of the Yalova which is the location of our Wind Turbine. However, this weather dataset did not include measurements for every 10 minutes but it included measurements for a day as day and night. To merge these 2 datasets and compose our final dataset we assumed weather measurements to be consistent and same for every 10 minutes of measurements of our main data. After composing our final dataset we started to analyse it in a comprehensive manner.

	Date	Time	LV ActivePower (kW	Wind Speed (m/s	Theoretical_Power_Curve (KWh	Wind Direction (°)	Month	Day/Night	Temp	Sun Hour	Moon Illumunation	Moonrise	Moonset	Sunrise	Sunset	DewPoint	WindChillC	WindGust	Humidity	RainMM	Pressure	Visibility
1	01 01 2018	0:00:00	380.0477905	5.31133604	416.3289078	259.9949036	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
2	01 01 2018	0:10:00	453.7691956	5.672166824	519.9175111	268.6411133	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
3	01 01 2018	0:20:00	306.3765869	5.216036797	390.9000158	272.5647888	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
4	01 01 2018	0:30:00	419.6459045	5.659674168	516.127569	271.2580872	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
5	01 01 2018	0:40:00	380.6506958	5.577940941	491.702972	265.6742859	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
6	01 01 2018	0:50:00	402.3919983	5.604052067	499.436385	264.5786133	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
7	01 01 2018	1:00:00	447.6057129	5.793007851	557.3723633	266.1636047	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
8	01 01 2018	1:10:00	387.2421875	5.306049824	414.8981788	257.9494934	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
9	01 01 2018	1:20:00	463.6512146	5.584629059	493.6776521	253.4806976	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
10	01 01 2018	1:30:00	439.725708	5.523228168	475.7067828	258.7237854	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
11	01 01 2018	1:40:00	498.1817017	5.724115849	535.841397	251.8509979	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
12	01 01 2018	1:50:00	526.8162231	5.934198856	603.0140765	265.5046997	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
13	01 01 2018	2:00:00	710.5872803	6.547413826	824.6625136	274.2329102	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
14	01 01 2018	2:10:00	655.1942749	6.199746132	693.4726411	266.7331848	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
15	01 01 2018	2:20:00	754.7625122	6.505383015	808.0981385	266.7604065	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
16	01 01 2018	2:30:00	790.1732788	6.634116173	859.4590208	270.4931946	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
17	01 01 2018	2:40:00	742.9852905	6.378912926	759.4345366	266.5932922	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
18	01 01 2018	2:50:00	748.2296143	6.446652889	785.2810099	265.5718079	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
19	01 01 2018	3:00:00	736.6478271	6.415082932	773.1728635	261.1586914	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
20	01 01 2018	3:10:00	787.2462158	6.437530994	781.7712157	257.5602112	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
21	01 01 2018	3:20:00	722.8640747	6.220024109	700.7646999	255.9264984	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
22	01 01 2018	3:30:00	935.0333862	6.89802599	970.7366269	250.0128937	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
23	01 01 2018	3:40:00	1220.609009	7.60971117	1315.048928	255.9857025	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
24	01 01 2018	3:50:00	1053.771973	7.288355827	1151.265744	255.4445953	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
25	01 01 2018	4:00:00	1493.807983	7.943101883	1497.583724	256.4074097	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
26	01 01 2018	4:10:00	1724.488037	8.376161575	1752.199662	252.4125977	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
27	01 01 2018	4:20:00	1636.935059	8.23695755	1668.470707	247.9794006	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
28	01 01 2018	4:30:00	1385.488037	7.879590988	1461.815791	238.6096039	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
29	01 01 2018	4:40:00	1098.932007	7.101376057	1062.285034	245.0955963	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
30	01 01 2018	4:50:00	1021.458008	6.955307007	995.9958546	245.410202	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
31	01 01 2018	5:00:00	1164.892944	7.098298073	1060.859712	235.2279053	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10
32	01 01 2018	5:10:00	1073.332031	6.953630924	995.2509608	242.8726959	1	0	4	8.7	97	17:23:00	7:20:00	8:27:00	17:46:00	3	5	6	89	0	1020	10

Figure 2.1: General demonstration of the Final Dataset

Our final dataset includes 50530 rows and 22 columns. It consists variables such as :

Date: Demonstrates the specific date of the measurement in dd/mm/yyyyy format. It is in text type.

Time: Demonstrates the specific time of the measurement in hh/mm/ss format. It is also in text type.

LV Active Power (kW) : It demonstrates the measured value of electricity production of the wind turbine in terms of kilowatts in corresponding time and date. It is a numeric attribute. It is also our target attribute. We will try to create a regression model to predict electricity production in the future.

Wind Speed: It demonstrates the speed of the wind in terms of meter/second in corresponding time and date. It is a numeric attribute.

Theoretical Power Curve: It demonstrates the theoretical calculation of the electricity production of the wind turbine in terms of kilowatt/hour, to do this calculation a theoretical formula is used which is based on wind speed and provided by the manufacturer firm of the Wind turbine. It is a numeric attribute.

Wind Direction: It demonstrates the angle of the wind in terms of degrees. It is a numeric attribute.

Month: It demonstrates the month of the corresponding measurement took place. It consists 12 different categories for each to represent 12 months. It is a categorical attribute.

Day/Night: It demonstrates the category of the corresponding measurement took place. It assumes from time 00:00 to 08:00 as night -0- and the rest of the day as day -1-. It is a categorical attribute.

Temp: It describes the temperature at the corresponding time being as a Celsius degree. It is a numerical attribute.

Sun Hour: It describes the time from sunrise to sun set as hours. It is a numerical attribute.

Moon Illumination: It describes the illumination degree of moon as a percentage. It is a numerical attribute.

Moon Rise: It describes the time of moonrise. It is a numerical attribute.

Moon Set: It describes time of moonset. It is a numerical attribute.

Sun Rise: It describes the time of sunrise. It is a numerical attribute.

Sun Set: It describes the time of sunset. It is a numerical attribute.

Dew Point Temp: It describes the temperature, which allows to creation of dew points as Celsius degree. It is a numerical attribute.

Wind Chill C: It describes the wind effect for the felt temperature as Celsius degree. It is a numerical attribute.

Wind Gust Kmph: It means sudden jumps at the wind speed which are shorter than 20 seconds in terms of kilometer/hour. It is a numerical attribute.

Humidity: It describes the humidity rate as a percentage. It is a numerical attribute.

Rain MM: It describes the raindrop fell as millimeter. It is a numerical attribute.

Pressure: It describes the pressure as milibars. It is a numerical attribute.

Visibility: It describes the visibility of an object in certain conditions. It is a meteorological measure. It is in terms of RVR runway visual range. It is a numerical attribute.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50530 entries, 0 to 50529
Data columns (total 22 columns):
Date                              50530 non-null object
Time                              50530 non-null object
LV ActivePower (kW)               50530 non-null float64
Wind Speed (m/s)                  50530 non-null float64
Theoretical_Power_Curve (KWh)     50530 non-null float64
Wind Direction (°)                50530 non-null float64
Month                             50530 non-null int64
Day/Night                         50530 non-null int64
Temp                              50530 non-null int64
Sun Hour                          50530 non-null float64
Moon Illimunation                 50530 non-null int64
Moonrise                          50530 non-null object
Moonset                           50530 non-null object
Sunrise                           50530 non-null object
Sunset                            50530 non-null object
DewPoint Temp                     50530 non-null int64
WindChillC                        50530 non-null int64
WindGustKmph                      50530 non-null int64
Humidity                          50530 non-null int64
RainMM                            50530 non-null float64
Pressure                          50530 non-null int64
Visibility                        50530 non-null int64
dtypes: float64(6), int64(10), object(6)
memory usage: 8.9+ MB
```

Figure 2.1 Decription of Dataset Attributes in Python

Our final dataset was a bit hard to demonstrate as a whole because it consists 50530 rows 22 columns so the first thing we did in our analysis was to divide it into 2 categories. We divide our 22 attributes as numerical and categorical attributes. After this division we made our deeper analysis in these 2 subsets. Of course our reason for this division was not only to divide our dataset into 2 subsets, but also to evaluate similar data together and to make a more focused analysis about

Attributes	Number of Zeros
LV ActivePower (kW)	10781
Wind Speed (m/s)	10
Theoretical_Power_Curve (KWh)	7749
Wind Direction (°)	75

Figure 2.2: Number of zero values

Attribute	Ought	Is
Date	365	344-355
Time	144	11-144

Figure 2.3 Number of necessary observations and the number of observations in the data

different types of data accordingly. At first we did not realize the missing values but we found numerous rows with 0 values. We assume that these 0 rows are missing values but changed with 0 or may be they are because of the some problem of the recording device. Also on some days and some hours there are less observations then it should be.

2.1 Analysis of Categorical Data

	Date	Time	Month	Day/Night	Moonrise	Moonset	Sunrise	Sunset
0	01 01 2018	00:00:00	1	0	17:23:00	07:20:00	08:27:00	17:46:00
1	01 01 2018	00:10:00	1	0	17:23:00	07:20:00	08:27:00	17:46:00
2	01 01 2018	00:20:00	1	0	17:23:00	07:20:00	08:27:00	17:46:00
3	01 01 2018	00:30:00	1	0	17:23:00	07:20:00	08:27:00	17:46:00
4	01 01 2018	00:40:00	1	0	17:23:00	07:20:00	08:27:00	17:46:00

Figure 2.1.1 Demonstration of Categorical Attributes

Our dataset consist of 8 different categorical data. Actually in a different data set date and time might not be seen as a category but since our data set includes nearly 50000 measurements every day itself can be identified as a category with approximately 144 values. Similarly our Time attribute includes 10 minutes of gaps and same for everyday and in our whole dataset every gap nearly includes 365 values. So we decided to consider time and date attributes as categorical attributes. Month and Day/Night attributes are clear categories for Months explains months from January to December and Day/Night attribute indicates a specific time is in the daytime or night-time. Again similar to the Time attribute we accepted the Moonrise Moonset Sunrise and Sunset attributes as categorical variables but different than the time this acceptance was mostly for differentiating them from the numerical values.

Analysis and Comments for Categorical Data

Category	Least Frequent	Frequency-Min	Most Frequent	Frequency-Max	Entropy	Number Of Categories
Date	02-10-18	11	31-12-18	144	8.47	356
Time	11:50	344	16:50	355	7.17	144
Month	11(November)	3800	7(July)	4464	3.58	12

Day/Night	0(Night)	25172	1(Day)	25358	0.99	2
Moonrise	13:30	39	No Moonrise	1726	8.13	306
Moonset	14:17	11	No Moonset	1728	8.14	309
Sunrise	17:49	127	17:36	1865	7.17	171
Sunset	7:01	61	5:32	2302	7.21	174

Figure 2.1.2: Statistics of Categorical Data

The Date attribute should be equal 144 since everyday is 1440 minutes and we take measurements in every 10 minutes but instead there are differences. There is not a single maximum which is but there are lots of days which does not have 144 measurements.

The time attribute should consist 365 observations for every 10 minute slice but instead it also differs from a minimum 344 observations to a max 355. This may be because of the lack of measurements in the missing days.

Month attribute is also changes and not stable but this may be explained with the different number of days for different months. It also may be because of the lack of observations to. It should be examined further.

Day/Night attribute is also should be equal for day and nights. There should be equal observations but there is not. However, still it is pretty close and entropy value is really close to 1 and it also shows an even distribution between day and night.

For the last 4 attributes the analysis do not value much for our processes ıt may value for a meteorological point of view but we did not consider them much.

2.2 Analysis of Numerical Data

	mean	std	count	min	max	median
LV ActivePower (kW)	1307.684332	1312.459242	50530.0	-2.471405	3618.732910	825.838074
Wind Speed (m/s)	7.557952	4.227166	50530.0	0.000000	25.206011	7.104594
Theoretical_Power_Curve (KWh)	1492.175463	1368.018238	50530.0	0.000000	3600.000000	1063.776282
Wind Direction (°)	123.687559	93.443736	50530.0	0.000000	359.997589	73.712978
Temp	15.956818	7.478934	50530.0	-1.000000	32.000000	16.000000
Sun Hour	10.394415	3.198427	50530.0	3.400000	14.500000	11.600000
Moon Illimunation	46.463131	31.548401	50530.0	0.000000	100.000000	46.000000
DewPoint Temp	10.533089	5.966731	50530.0	-5.000000	22.000000	11.000000
WindChillC	15.989234	8.341774	50530.0	-5.000000	32.000000	16.000000
WindGustKmph	14.597645	8.020789	50530.0	0.000000	52.000000	13.000000
CloudCover	35.742668	34.079858	50530.0	0.000000	100.000000	22.000000
Humidity	69.887592	16.452907	50530.0	31.000000	97.000000	72.000000
RainMM	0.004987	0.132024	50530.0	0.000000	3.500000	0.000000
Pressure	1014.676727	6.112598	50530.0	993.000000	1032.000000	1014.000000
Visibility	9.860400	0.987456	50530.0	0.000000	10.000000	10.000000

Figure 2.2.1 Descriptive Statistics for Numerical Data

For LV Active Power attribute mean and standard deviation is the same so we may say that standard deviation is relatively high. It also shows on the min and max values that the data of this attribute is very diverse. This diversity may be because of the extreme values or some outliers, because mean may be effected these values and the difference between median and mean is also says the same thing.

For Wind Speed variable standard deviation is not that high at least smaller than the mean but the min and max values also indicate a diverse data. However mean and median is relatively close so there may be not so much outliers.

Theoretical Power Curve is a guess for LV Active Power and it also shows on the statistics data is diverse statistics are very similar to LV Active Power attribute.

Win Direction data is also diverse and it means sudden direction changes and it may be related with Wind Gust variable because wind gust measures the speed of sudden wind bursts.

Pressure and visibility attributes are pretty stable has very small standard deviations they seem to be very consistent does not shows too much change.

The rest of the attributes are giving us ideas about more on the meteorological point of view than our process.

After making correlation calculations we may discover some relations but at the first glance they do not offer much information about our process.

For missing values that we mentioned above, most of the missing values are on the Active power and Theoretical Power Curve Attributes and they are not missing values but 0 values. Therefore, this effected our calculations and probably minimum values are automatically 0 for these attributes. So our comments about diversity should be revised. We decided to analyse our numerical one more time before model build and feature selection operations.