Hi classmate,

I use the wind turbine capacity dataset.

[wind\_turbines.csv](https://my.uclaextension.edu/users/189080/files/6424341?wrap=1&verifier=OlsaggoTBApeaMRu1pbo38CORs7ept8JHgiOESTA)[Download wind\_turbines.csv](https://my.uclaextension.edu/users/189080/files/6424341/download?verifier=OlsaggoTBApeaMRu1pbo38CORs7ept8JHgiOESTA&download_frd=1)

The data set is from the US government ([https://corgis-edu.github.io/corgis/csv/wind\_turbines/ (Links to an external site.)](https://corgis-edu.github.io/corgis/csv/wind_turbines/)). It investigates all the turbines in the USA to check how the capacity they have.

There are several variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Site.State | String | Two letter abbreviation of state where turbine is located (e.g., CA for California) | "IA" |
| Site.County | String | County where the turbine is located. | "Story County" |
| Year | Integer | Year when the turbine's project became operational | 2017 |
| Turbine.Capacity | Integer | Electrical generation capacity of the turbine measured in KW (kilo-watts) | 3000 |
| Turbine.Hub\_Height | Float | Height in meters of the turbine's hub | 87.5 |
| Turbine.Rotor\_Diameter | Float | Diameter in meters of the turbine's rotor | 125.0 |
| Turbine.Swept\_Area | Float | The area swept on each rotation of the turbine | 12271.85 |
| Turbine.Total\_Height | Float | Total height of the turbine, in meters | 150.0 |
| Project.Capacity | Float | Electrical generation capacity of the turbine measured in MW (mega-watts) | 30.0 |
| Project.Number\_Turbines | Integer | Number of turbines in this project | 10 |
| Site.Latitude | Float | Latitude (decimal degrees - NAD 83 datum) of where turbine is located | -93.518082 |
| Site.Longitude | Float | Longitude (decimal degrees - NAD 83 datum) |  |

However, mostly I will use Turbine.Hub\_Height/Turbine.Rotor\_Diameter/Turbine.Rotor\_Diameter/Turbine.Swept\_Area/Turbine.Total\_Height to complete my assignment.

Before digging into my dataset, I quickly glance at my data by using the **plot function**:

Diagram, engineering drawing

Description automatically generated

From the first look, the four variables, Turbine.Hub\_Height/Turbine.Rotor\_Diameter/Turbine.Swept\_Area/Turbine.Total\_Height may have some correlation.

* **Model 1**

I use the**Turbine.Hub\_Height** and **Turbine.Total\_Height** variable to complete my linear regression model.

Statistics analysis result:

The result tells me that the QQplot is OK, and no point beyond the cook's distance line.

Chart, scatter chart

Description automatically generated

The summary function illustrates that it can work.

The linear regression line will be:

**Turbine.Hub\_Height = 14.269581 + 0.517067 \* Turbine.Total\_Height** and all the variable has the **P-value <0.05**

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

* Model 2:

When I check the correlation between **Turbine.Hub\_Height** and **Turbine.Swept\_Area**, the linear regression is **not** good.

So I **sqrt the Turbine.Swept\_Area data** and plot a scatter plot again, it looks better than the previous one.

Chart, scatter chart

Description automatically generated

The statistics plot shows:

in model 2, the QQ plot is apparently better than the model 1, and the scale location looks ok.

Chart, scatter chart

Description automatically generated

The summary table:

The  linear regression will be:

**Turbine.Hub\_Height = 40.036397 + 0.479432\*sqrt(Turbine.Swept\_Area)**

All the variables P-value <0.05

Text

Description automatically generated

Chart, scatter chart

Description automatically generated