London Weather

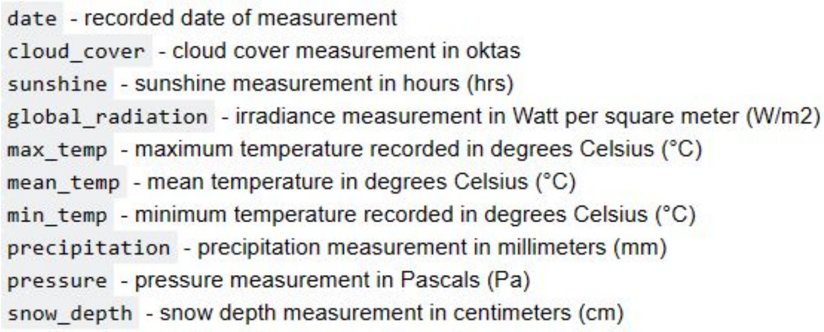
Author: Mathew Thomas

Purpose

- Using ML techniques, attain best prediction for mean temperature in London.
- Provide tangible insight that impacts decision making.
- Increase optimization & efficiency for current systems.

Dataset

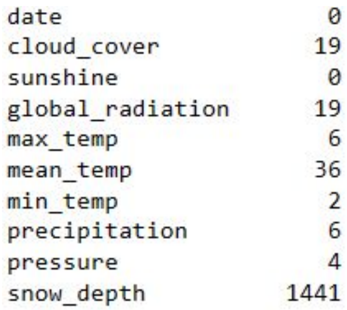
- Attained from Kaggle and is an aggregate of weather attributes extracted from the European Climate Assessment & Dataset (ECA & D).
- Measurements were reported from a weather station near London's Heathrow airport.
- Data from January 1st 1979 to December 31st 2020.
- Contains 15341 rows and 10 columns.



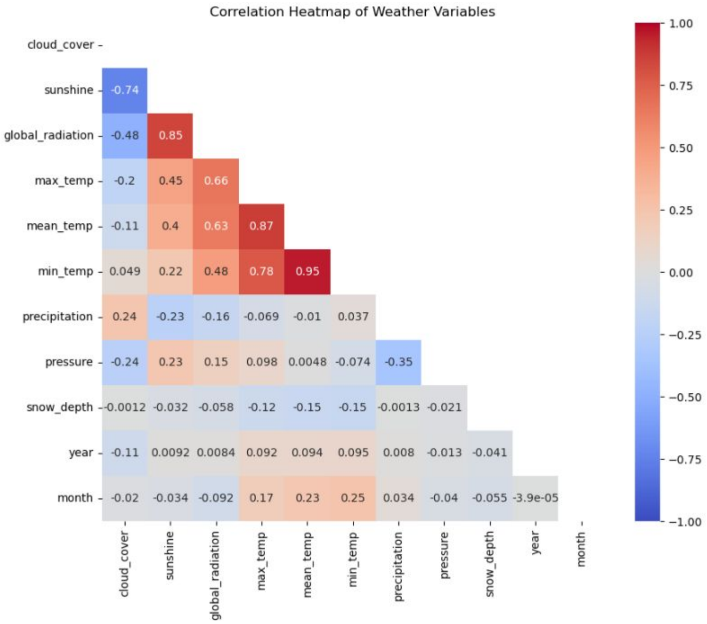
`date` - recorded date of measurement
`cloud_cover` - cloud cover measurement in oktas
`sunshine` - sunshine measurement in hours (hrs)
`global_radiation` - irradiance measurement in Watt per square meter (W/m2)
`max_temp` - maximum temperature recorded in degrees Celsius (°C)
`mean_temp` - mean temperature in degrees Celsius (°C)
`min_temp` - minimum temperature recorded in degrees Celsius (°C)
`precipitation` - precipitation measurement in millimeters (mm)
`pressure` - pressure measurement in Pascals (Pa)
`snow_depth` - snow depth measurement in centimeters (cm)

EDA

- Converted ‘date’ into a datetime format
- Created a new dataframe and extracted ‘year’ & ‘month’ then dropped ‘date’
- Checked for null values and noticed 9.39% for values in ‘snow_depth’ were missing
- Filled all null values for each column with their respective mean
- Correlations between the different variables, eg: ‘mean_temp’ & ‘min_temp’ strongest relationship
- Outliers within the dataset, eg: ‘max_temp’ value of 120°C



```
date                   0
cloud_cover           19
sunshine               0
global_radiation      19
max_temp               6
mean_temp             36
min_temp               2
precipitation          6
pressure               4
snow_depth          1441
```



Correlation Heatmap of Weather Variables

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth	year	month
cloud_cover											
sunshine	-0.74										
global_radiation	-0.48	0.85									
max_temp	-0.2	0.45	0.66								
mean_temp	-0.11	0.4	0.63	0.87							
min_temp	0.049	0.22	0.48	0.78	0.95						
precipitation	0.24	-0.23	-0.16	-0.069	-0.01	0.037					
pressure	-0.24	0.23	0.15	0.098	0.0048	-0.074	-0.35				
snow_depth	-0.0012	-0.032	-0.058	-0.12	-0.15	-0.15	-0.0013	-0.021			
year	-0.11	0.0092	0.0084	0.092	0.094	0.095	0.008	-0.013	-0.041		
month	-0.02	-0.034	-0.092	0.17	0.23	0.25	0.034	-0.04	-0.055	-3.9e-05	

Model

- Modelling relationships between variables for a continuous dataset.
- Implemented a Linear regression.
- Using stepwise regression, concluded with ‘max_temp’ & ‘min_temp’ provided the best r-squared with dependant variable ‘mean_temp’.

Next Steps

- Address outliers
- Trying substituting median values instead of mean for the null values
- Check to see other methods to expand my understanding of features in dataset
- Perhaps look into cross validation like Time Series