# Name: Taha Shahid
# Project: Capstone Project 1: Milestone Report

## Problem statement: Why it's a useful question to answer and for whom.

The data set selected is called "Weather History in Szeged 2006-2016" and as the name suggest we predict which machine learning technique and neural network system will give the best accuracy for the prediction of temperature.

The dataset is provided by the Kaggle repository and will be acquired through downloading csv format of the dataset from the website.

To analyze this problem, different machine learning algorithms such as Logistic Regression and Neural Networks (LSTM, MLP) will be used to see which method has the best accuracy for the prediction of temperature. Also, PCA (Principal component analysis) will be used to see if dimensionality can be reduced. Any or all methods learned in the machine learning algorithms will also be applied.

For the deliverables of the project, an IPYTHON notebook code will be provided along with the paper summarizing the findings as well as a PowerPoint presentation.

## Description of the dataset, how you obtained, cleaned, and wrangled it

### What kind of cleaning steps did you perform?

The weather of szeged 2006-2016 dataset was downloaded in excel(csv) format from the Kaggle repository. For cleaning the data, multiple steps had to be followed such as: fixing missing values, float to integer conversion, and string to integer conversion.

**How did you deal with missing values, if any?**

Missing values were found in both quantitative and categorical columns. To deal with the missing values in the quantitative column, Loud Cover consisted only of zeros and was removed from the dataset for analysis, and zeros in the Pressure column of the Dataset and was replaced with the medians of pressure close to them as we know that pressure never takes zero value in millibars. Also, there were some missing values in the categorical Precipitation Type column. Used the temperature to replace the zeros with possible precipitation type as it is dependent on the temperature.

**Were there outliers, and how did you handle them?**

No, most of the data did not have any outliers apart from the fact of missing data.  there were outliers and they were determined using the criteria that was mentioned with the dataset as follows

- Formatted Date – Datetime column
- Hourly Summary – As string object
- Precipitation Type – As string object
- Daily Summary – As string object
- Temperature (C) – float
- Apparent Temperature (C) – float
- Humidity – float
- Wind Speed (km/h) – float
- Wind bearing (degrees) – float
- Visibility (km) – float
- Loud Cover – float
- Pressure (millibars) – float

## Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)

### a. Summary of findings

Some of the questions that were investigated from the surface of the above-mentioned dataset are:

**Proportion of most common categories per hourly summary.**

- Partly Cloudy (33%)
- Mostly Cloudy(29%)
- Overcast (17%)
- Clear (11%)
- Foggy(7%)
- Others(3%)

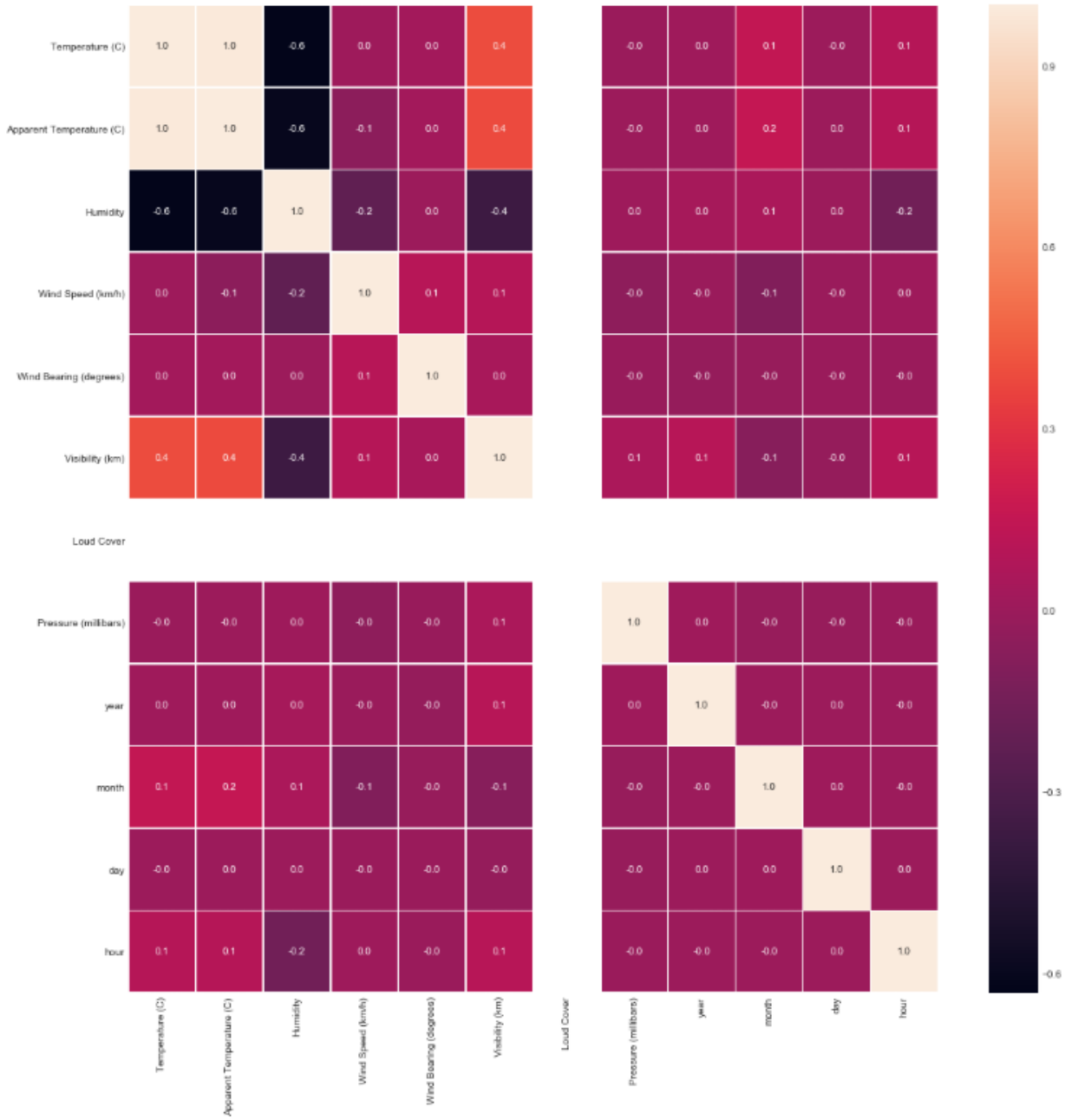
| col_o | count |
|---|---|
| Summary | |
| Partly Cloudy | 0.329000 |
| Mostly Cloudy | 0.291271 |
| Overcast | 0.172073 |
| Clear | 0.112905 |
| Foggy | 0.074109 |

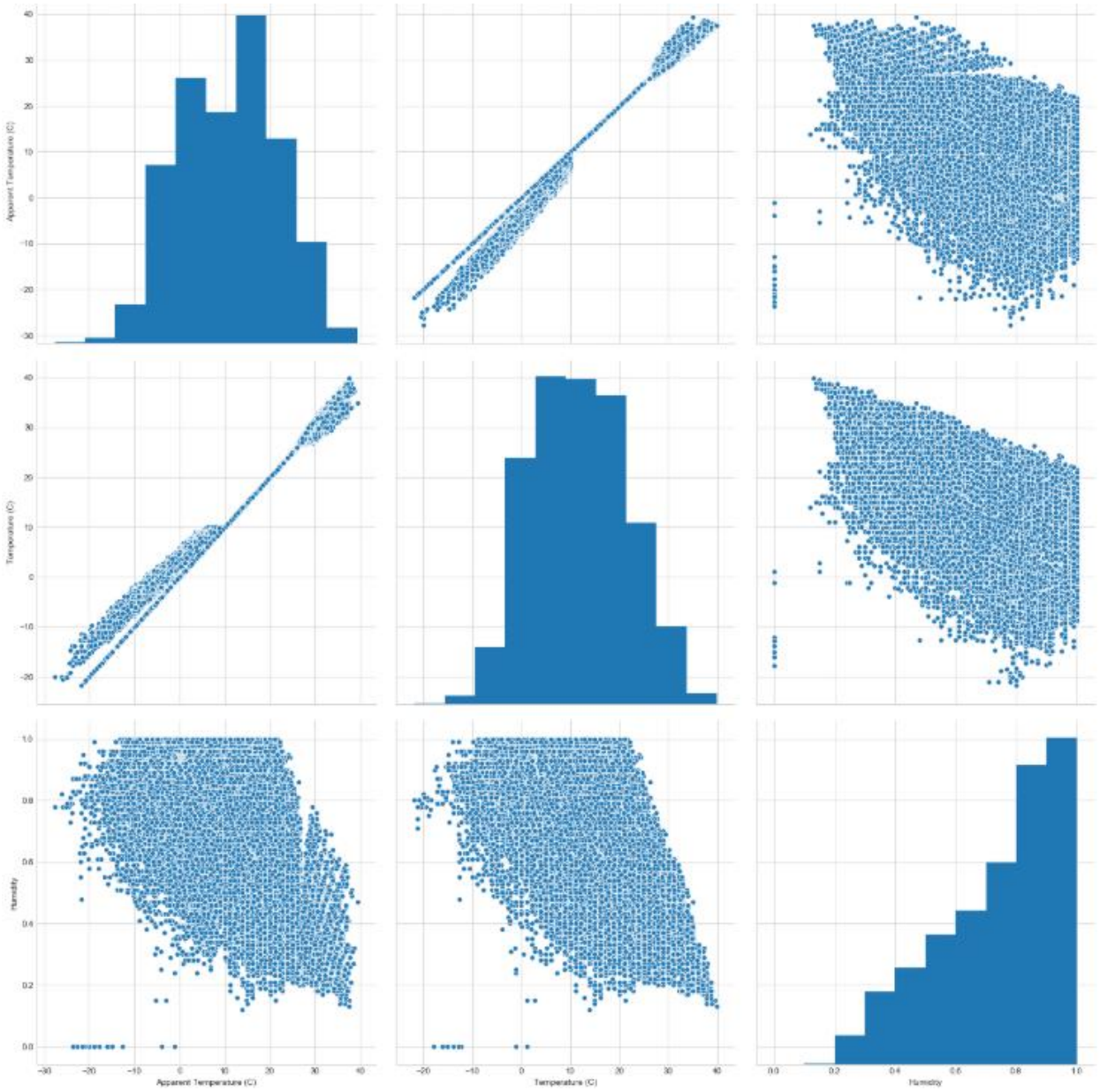
**Proportion of most common categories per daily summary**.

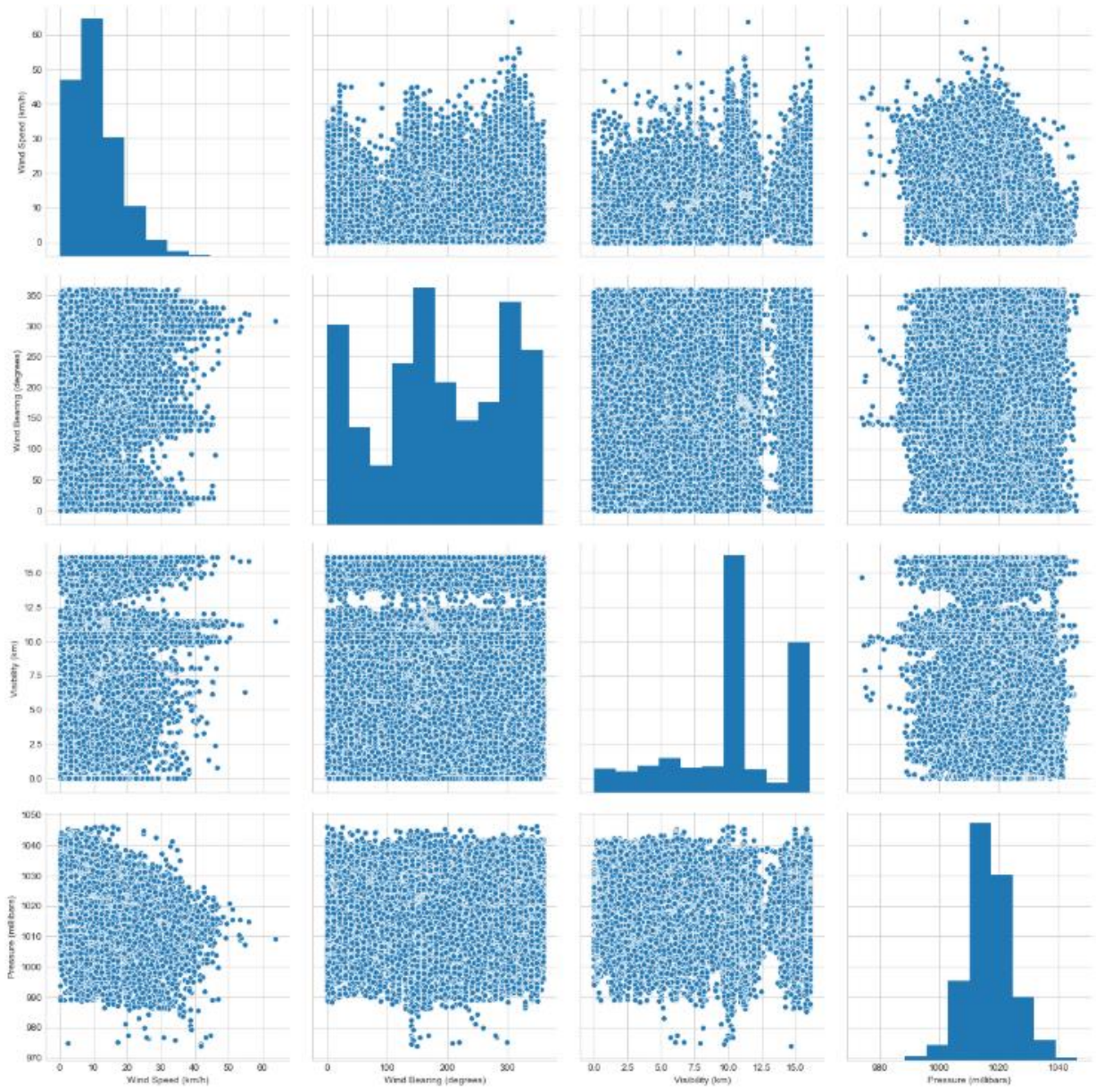
- Mostly cloudy throughout the day 21%
- Partly cloudy throughout the day 10%
- There were a total of 214 different Daily summaries.

| col_0<br>Daily Summary | count |
|---|---|
| Mostly cloudy throughout the day. | 0.208236 |
| Partly cloudy throughout the day. | 0.103480 |
| Partly cloudy until night. | 0.063959 |
| Partly cloudy starting in the morning. | 0.053746 |
| Foggy in the morning. | 0.043555 |
| Foggy starting overnight continuing until morning. | 0.037075 |
| Partly cloudy until evening. | 0.034089 |
| Mostly cloudy until night. | 0.032088 |
| Overcast throughout the day. | 0.030606 |
| Partly cloudy starting in the morning continuing until evening. | 0.029092 |

## b. Data Visualization




Apparent Temperature (C)
Temperature (C)
Humidity
Apparent Temperature (C)
Temperature (C)
Humidity


Wind Speed (km/h)
Wind Bearing (degrees)
Visibility (km)
Pressure (millibars)
Wind Speed (km/h)
Wind Bearing (degrees)
Visibility (km)
Pressure (millibars)