Week 1 Project Report: Project Planning and Data Collection

Project Title: Time Series Analysis for Weather Forecasting

During the first week of our capstone project, we focuses on time series analysis for forecasting weather conditions such as temperature, humidity, dew point, and wind speed. The project aims to use historical data to predict future weather patterns, helping to understand trends and seasonality in weather data.

This week, our goal was to collect the relevant time series data and perform initial data exploration to better understand its characteristics. This phase is crucial as it sets the foundation for later analysis and modelling.

For this project, we collected historical weather data from a trusted weather API and online datasets. The data consists of hourly weather records, including:

- Datetime: Timestamp of each observation.
- Temperature (Temp): Hourly temperature readings in Celsius.
- Dew Point (Dew): Temperature at which air becomes saturated with moisture.
- Humidity: Percentage of moisture in the air.
- Wind Speed: Hourly wind speed in km/h.

...

Source of Data:

The data was gathered from [OpenWeather API](https://openweathermap.org/) and a public weather dataset from [Kaggle](https://www.kaggle.com/).

Data Format:

The data is in EXCEL format and contains approximately 9,000 rows of hourly records over a period of 1 year.

Week 2 Project Report: Data Exploration

First, we loaded the data into Python using the ‘pandas’ library for analysis. This gave us a quick look at the dataset's structure, ensuring all columns are present and properly formatted.

```
import pandas as pd
 #Load the dataset
df = pd.read_csv('weather_data.csv')
#Display the first few rows
print(df.head())
```

Next, we checked if the dataset contains any missing or null values, which could cause issues in the analysis. Since the data is fetch from API we don’t have any missing values.

```
# Check for missing values
print(df.isnull().sum())
```

It’s important to ensure that the data types are correct, especially for the ‘datetime’ column. After ensuring the ‘datetime’ column was formatted correctly, we set it as the index for time series analysis.

```
# Convert 'datetime' column to datetime type
df['datetime'] = pd.to_datetime(df['datetime'])
# Set 'datetime' as the index
df.set_index('datetime', inplace=True)
```

We calculated basic statistics like mean, median, and standard deviation for each column to get a sense of the data distribution.

```
# Summary statistics
print(df.describe())
```

We created visualizations to understand trends, seasonality, and possible correlations between variables. We plotted each feature to observe its behaviour over time.

```
import matplotlib.pyplot as plt

#Plot temperature, dew point, humidity, and wind speed over time

df[['temp', 'dew', 'humidity', 'windspeed']].plot(subplots=True, figsize=(15, 10))

plt.show()
```

We also calculated the correlation between variables to explore any potential relationships.

```
# Correlation matrix

print(df.corr())
```

Week 3 Project Report: Data Pre-processing

The goal for this week was to prepare the collected weather data for model development by performing data pre-processing and feature engineering. These steps are essential for enhancing model accuracy and capturing relevant patterns within the dataset.

After no additional missing values were found, confirming that the data is complete and ready for further analysis. We checked for any outliers or extreme values in the data, as these could affect model performance. Outliers in time series data may occur due to measurement errors or rare events. We used box plots to visualize outliers.

```
import seaborn as sns

#Boxplots to visualize outliers for each feature

sns.boxplot(data=df[['temp', 'dew', 'humidity', 'windspeed']])

plt.show()
```

The ‘temp’, ‘dew’, and ‘humidity’ columns displayed no extreme outliers, but a few high wind speed values were observed. Since wind speed fluctuations can be natural, we did not remove these values as they may provide useful information for the forecasting model.