Aviation Weather Forecasting Using METAR Data

This report details the analysis of the METAR dataset of KMIA (Miami international airport)

by Gagan Aryan

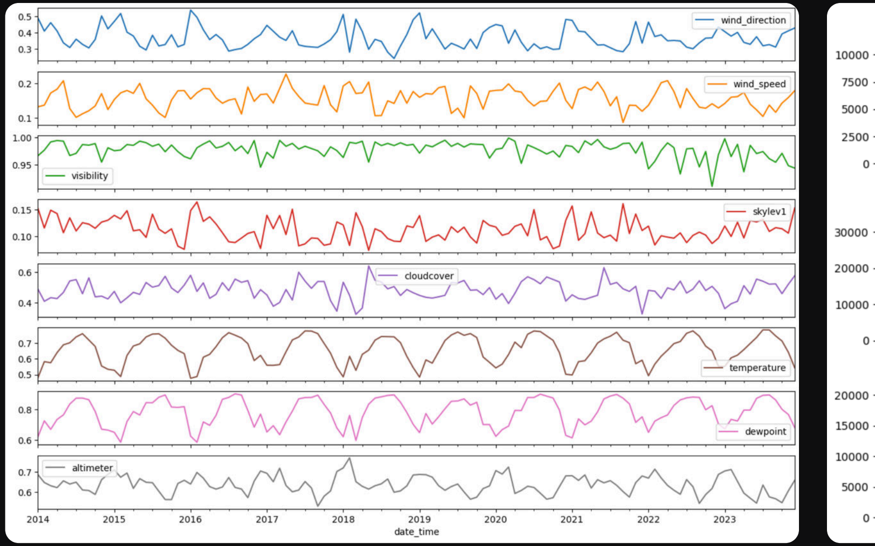
Data preprocessing

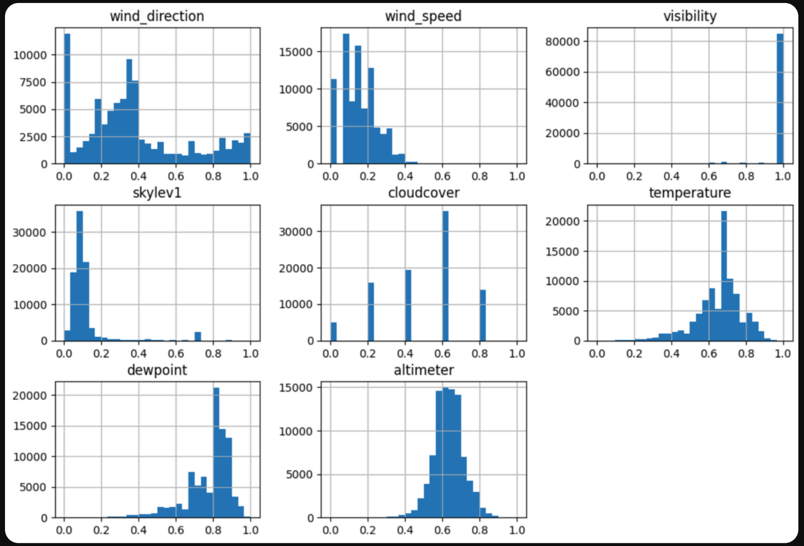
1. Data present in the **given dataset** is hard to perform machine learning analysis upon. To tackle this we are using the metpy library to process the data into a format that separates **each of the metric** in the metar format. **Link to script**.
2. Handling missing values
   1. Continuous variables: Here we imputed the values with mean for all of the variables except wind speed, skylevel 2, 3 and 4 since these columns had too many missing values and imputing would have created significant bias. Hence these columns were dropped.
   2. Categorical variables: Here we assigned the missing values as the mode of the dataset
3. Each of the continuous variables were normalised using Min-Max scaling and categorical variables were converted into one hot encoding.
4. For time based metric, we derived the month, hour and day from the METAR data. For rolling averages, 24-hour rolling average was imputed for each of the continuous metrics.

Seasonality and general distribution

Here are the distributions of some of the variables. We can notice a roughly normal distribution for temperature and altimeter datapoints. Whereas for the dewpoint we have a right skewed distribution and vice-versa for wind speed.

As for visibility we have majority of the normalised datapoints around 1. (figure 2)

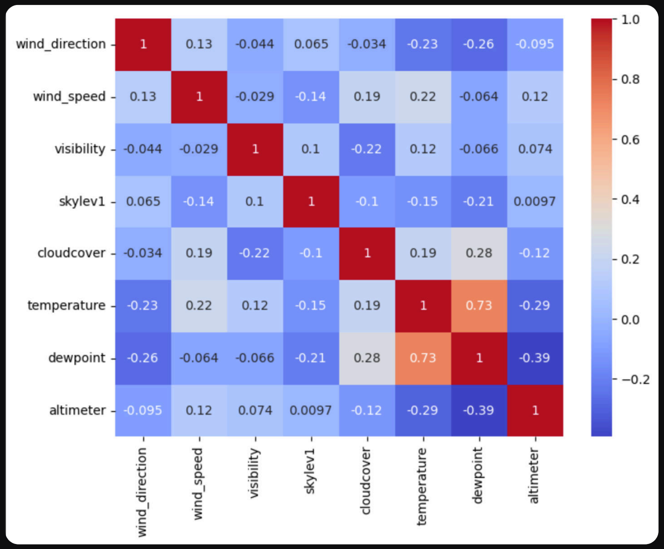


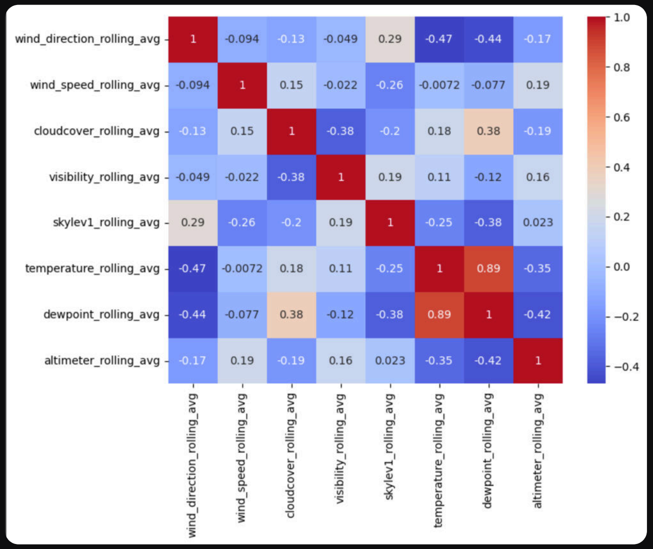


We can notice that when sampled acrosss months, the seasonality in temperature and dewpoint becomes very apparent. We can also notice some seasonality in the altimeter readings indicating that the atmospheric pressure is typically higher in winters compare to the summers. (figure 1)

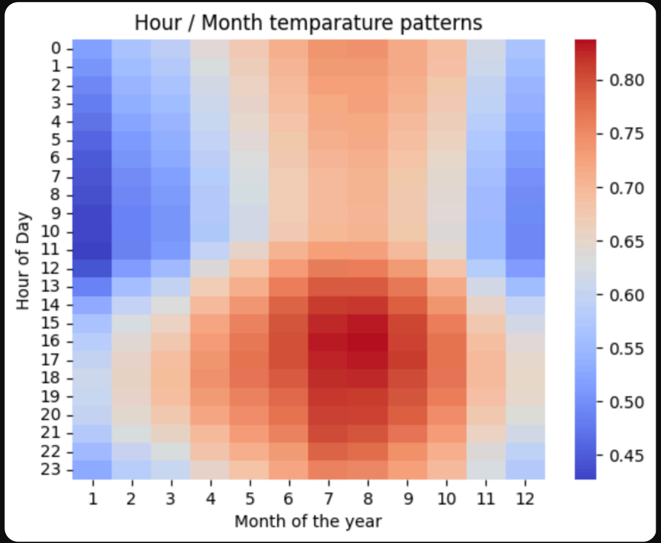
Correlation analysis

1. Temperature and Dewpoint - These two datapoints show tha strongest correlation (0.73) because as the air temperature increases, it can hold more moisture, leading to a higher dew point, which signifies more moisture in the air. Conversely, a decrease in temperature reduces the air's capacity to hold moisture, often bringing the temperature closer to the dew point and increasing the likelihood of dew, fog, or cloud formation. This correlation underscores the fundamental meteorological principle that warmer air supports more moisture, reflecting the direct but complex relationship between temperature and atmospheric moisture content. Interestingly when we see the correlation of rolling averages of these two datapoints it gets even stronger to about 0.89.
2. Cloudcover and Dewpoint - These two datapoints show a correlation of 0.28 and about 0.38 with rolling averages. This moderate relationship suggests that while cloud cover and atmospheric moisture content (as indicated by the dew point) are related, other factors also significantly influence cloud formation. A higher dew point indicates more moisture in the air, which can contribute to cloud development, but the relatively low correlation implies that cloud cover is also dependent on additional atmospheric conditions such as air temperature, pressure, and upper-level atmospheric dynamics. This correlation shows that while increased moisture can support cloud formation, the presence or extent of cloud cover is not solely determined by the dew point. Therefore, the 0.22 correlation reflects the multifaceted nature of cloud formation, where moisture is a contributing factor but not the only determinant.





Weather pattern throughout the day and year



We can see from this heatmap that the highest temperature across the hours of the day and month of the year seems to be around 1600 hours and July-August respectively with temperature going north of 30C

Outlier analysis

To pinpoint the outliers we scanned the datapoints that lie on either side of 10th and 90th percentile and found the following two notable events

1. For wind speed, September 10th 2017 is an outlier. This is consistent with the weather report here that reports a Hurricane in Miami, Florida - https://www.weather.gov/mfl/hurricaneirma#:~:text=On%20Sunday%20morning%2C%20September%2010,3%20with%20115%20mph%20winds.
2. For temperature, the anamoly is 22nd Jan 2020 which is consistent with the weather report here - https://www.cbsnews.com/miami/news/cold-snap-coming-to-south-florida-feeling-the-30s-this-weekend/.

Weather prediction - baseline model

We used a baseline model that would predict values for a particular hour of the day by scanning the values at the same hour in the past years. Of the 10 years of data, 9 years of data was chosen as the training set and the data of 2023 was chosen as the test set.

For continuous variables, the value of the datapoint would be the median in the past 9 years and for categorical values it would be the mode.

Weather prediction - XGBoost

Reasons for using XGBoost:

1. **Handling Non-linear Relationships**: Weather data often involve complex, non-linear relationships that XGBoost can capture.
2. **Dealing with Variability and Noise**: Weather datasets can be noisy and variable; XGBoost's robustness to overfitting makes it suitable for these conditions.
3. **Flexibility:** XGBoost can handle a mix of categorical and continuous data and is robust to missing data. While preprocessing (like one-hot encoding for categorical variables) is often recommended, XGBoost can naturally handle various data types, making it versatile for different modeling tasks.

Comparison of performance

Continuous datapoints

Metric	Baseline Modal RMSE	XGBRegressor RMSE
Wind Direction	0.2939	0.1820
Wind Speed	0.0895	0.0455
Cloudcover	0.2277	0.0113
Skylev1	0.1303	0.0854
Visibility	0.1461	0.0533
Temperature	0.0786	0.0339
Dewpoint	0.0863	0.0308
Altimeter	0.0765	0.0209

Categorical datapoints

Metric	Baseline Accuracy	XGBClassifier Accuracy
current_wx1	0.8886	0.9152
urrent_wx2	0.9728	0.9936
current_wx3*	1.0000	1.0000
skyc1	0.5129	0.7641
skyc2	0.3585	0.7716
skyc3	0.5946	0.8798
skyc4	0.8862	0.9439

Feature importance for each of the metrics are provided in this colab notebook:
https://colab.research.google.com/drive/1FGaHAevB4YEW6qhYYqGOXzVP3V9dLuBT#scrollTo=yXfhbTToD5mu

* accuracy of 1.0 is because of majority of the datapoints being absent

Links

1. Code: https://github.com/encrypted-soul/tinkering-with-data/tree/main
2. Colab Notebook: https://colab.research.google.com/drive/1FGaHAevB4YEW6qhYYqGOXzVP3V9dLuBT?usp=sharing