Python Meteostat Data Quality Report

Overview

This report will outline the initial data quality findings on hourly weather data obtained by the meteodata python package for locations in New York City from 2021-present. Meteodata reportedly uses data from NOAA for US weather. This report will include an overview of the data, a review of the continuous and categorical features, including histograms and bar charts, and a comparison to manually pulled data. On initial review, this package provides clean and well documented data. There is no missing data, and the data appears to be reasonable and logical.

Summary

This dataset consists of hourly weather data, including temperature, dewpoint, precipitation, snow, windspeed, and pressure. There are no missing values in any of the columns. There are no duplicate rows. Distribution of the data is consistent with expectations, and there appear to be few outliers in this dataset. When comparing with data from meteorological stations around New York, it appears that this data is being pulled from Newark Airport, instead of the closest station. This is fine for the purposes of this app, however if extra time allows it is recommended to pull the data directly from the meteorological stations. As wind speed and wind direction appear to vary around the city, these features will need to be dropped. Data was taken from 5 different latitude/longitude coordinates in NYC, including one for Central Park.

Review Logical Integrity

Test 1: No timestamps before Jan 2021 or after 28th May 2024.

- 0 instances.

Test 2: No missing hour/location combinations.

- 0 instances.

Test 3: No temperature in central park in 2023 below NYC central park min (-16.1 C) or above NYC central park max (+33.9 C).

- 25 instances. These are from the recorded hottest times of the year (27th Aug and 7th Sept). It appears this is due to the fact that meteostat pulls data from Newark Airport and not the Central Park meteorological station.

Test 4: Wind direction not below 0 or above 360

- 0 Insances.

Review Continuous Features

Descriptive Statistics

There are 12 continuous features in this dataset:

- Time

    - This has an even distribution across all months. There are 19,891 unique timestamps in this dataset, which is slightly less than the 19,895 hours that are between the 1st Feb 2022 and 9th May 2024. This means there are 4 hours that has 0 subway ridership data.
- Temperature
    - Temperature has a mean of 13.8 C with a min of -15 and a max of 38.9, which lie within the expected temperatures of NYC. There are no missing values.
- Dew Point
    - Dew point has a mean of 5 C with a min of -25.6 and a max of 24.5. There are no missing values.
- Relative Humidity
    - Relative humidity has a mean of 60%, with a min of 6% and a max of 76%. There are no missing values.
- Precipitation
    - Precipitation has a mean of 0.16 mm, with a min of 0 mm and a max of 36.8 mm. 0.007% of values are missing.
- Snow
    - 100% of values are missing.
- Wind Direction
    - Wind direction has a mean of 180 deg, with a min of 0 deg and a max of 360 deg. No values are missing.
- Wind Speed
    - Wind speed has a mean of 15 km/h, with a min of 0 km/h and a max of 38.4 km/h. No values are missing.
- Wind Peak Gust
    - 100% of values are missing.
- Pressure
    - Pressure has a mean of 1016 hPa, with a min of 982 hPa and a max of 1040 hPa.
- Total Sunshine Duration
    - 100% of values are missing.

Histograms

All Histograms can be found in the appendix as a summary sheet. All features show a plausible distribution.

Review Categorical Features

There is 1 categorical feature in the dataset:

- Weather Condition Code
    - This has 19 unique values. Most common code is 2 (fair), followed by 3 (cloudy)

Comparison with Data Obtained from NOAA

Summary

Data was collected from multiple NOAA meterological stations around New York between Jan and Jun 2023 and compared to meteostat data collected at the same time point at the coordinates of Central Park (40.7789, -73.9692). Temperature, relative humidity, wind speed,

precipitation and wind direction were compared between the meteostat data and multiple New York meteorological stations. The meteostat data aligns with data collected from Newark airport, and this appears to be the station that meteostat collects NYC weather data from. Box plots of temperature, relative humidity, wind speed and wind direction deltas vs Newark and Central park are shown in the appendix. Temperature, precipitation, and relative humidity are tightly correlated, however wind speed and wind direction are not. For this reason, wind speed and wind direction are recommended to be dropped from the data.

Actions to take

1 action will be taken:

- Missing snow data.
    - Drop snow feature.
- Missing wind peak gust data.
    - Drop wind peak gust feature.
- Missing total sun duration data.
    - Drop total sun duration feature.
- Missing precipitation values.
    - Impute value as mode value (0).
- Weather condition code missing values
    - Impute code as mode value (2).
- Inconsistent wind speed and wind direction between different NOAA meterological sites and meteostat data.
    - Drop wind speed and wind direction features.

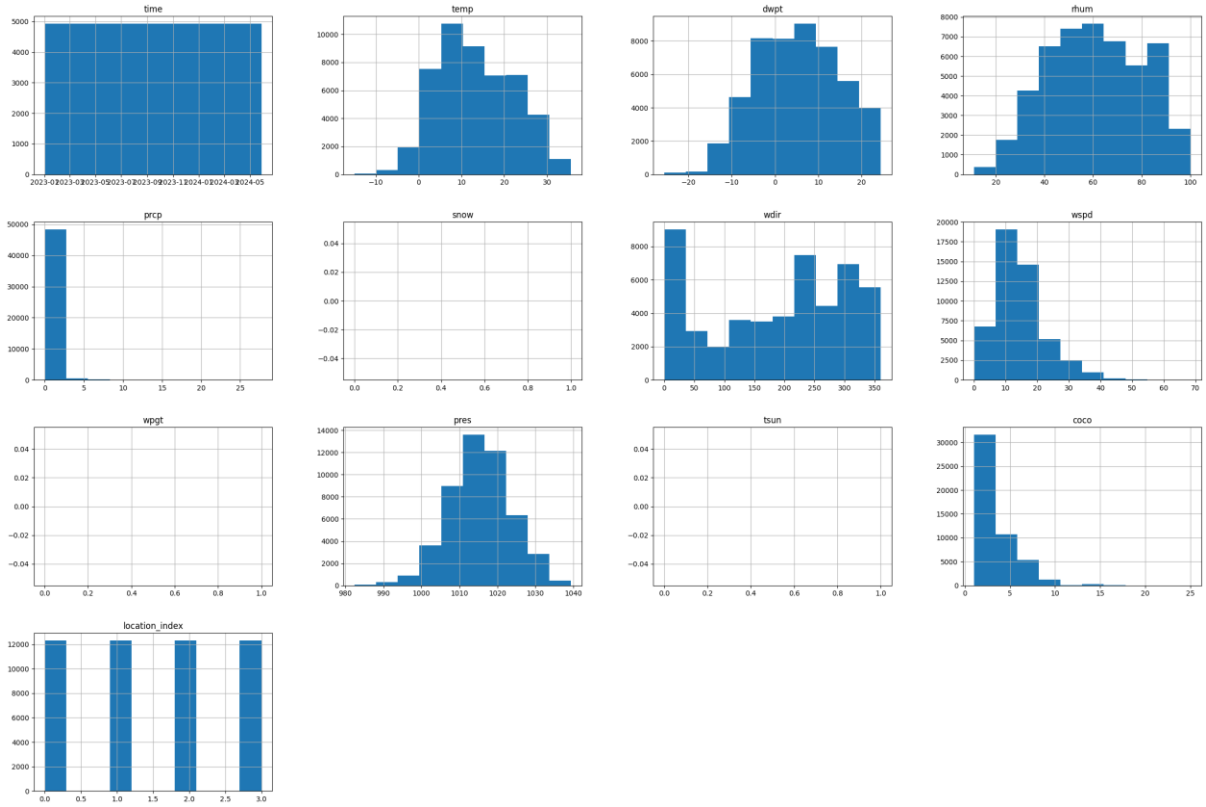
Appendix

Continuous Features

Descriptive Statistics

Feature	count	mean	min	25%	50%	75%	max	std	%missing	card
time	149165	14/09/2022 12:00	01/01/2021 00:00	07/11/2021 18:00	14/09/2022 12:00	22/07/2023 06:00	28/05/2024 00:00		0	29833
temp	149165	13.826272919250497	-15	6.1	13.3	22.2	38.9	9.82183982	0	116
dwpt	149165	5.391405490564139	-25.6	-2.7	6.1	14	24.5	10.2720429	0	445
rhum	149165	60.14336473033218	6	45	60	76	100	19.7860951	0	93
prcp	149155	0.15517414769870272	0	0	0	0	36.8	0.88203845	0.006704	101
snow	0								100	0
wdir	149165	189.00915094023398	0	90	220	280	360	113.518617	0	181
wspd	149165	14.945432909864914	0	9.4	13	20.5	68.4	8.49673194	0	45
wpgt	0								100	0
pres	149165	1016.3569470049945	982.4	1011.4	1016.2	1021.6	1040.9	7.79763232	0	528
tsun	0								100	0
location_i	149165	2	0	1	2	3	4	1.41421830	0	5

Histograms

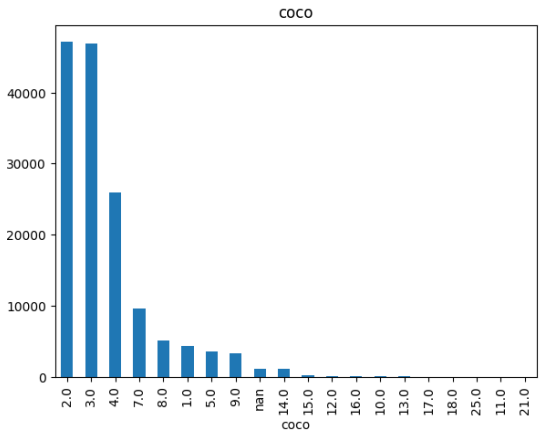


Categorical Features

Descriptive Statistics

Feature	count	unique	top	freq	second	second_freq	%missing	card
coco	148015	19	2	47135	3	46935	0.770958335	19

Box Plots



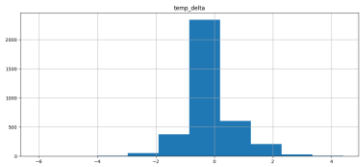
Comparison with NOAA Data

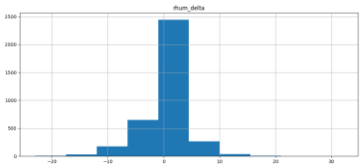
Newark vs Meteostat

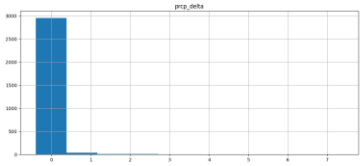
Descriptive Statistics

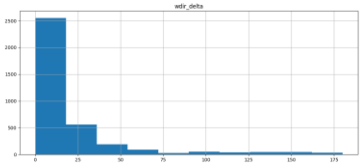
Feature	count	mean	std	min	25%	50%	75%	max
temp_delta	3625	0.00	0.865232	-6.10	-0.50	0.00	0.00	4.40
rhum_delta	3625	-0.19	4.012972	-23.00	-1.00	0.00	1.00	32.00
prcp_delta	3012	0.03	0.262327	-0.40	0.00	0.00	0.00	7.4
wdir_delta	3625	17.584828	33.40687	0.00	0.00	0.00	20.00	180.00
wspd_delta	3625	11.058317	6.508931	-4.10	6.8	10.40	14.80	38.70

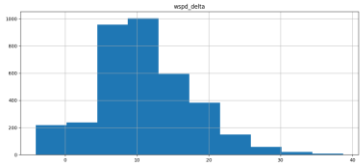
Box Plots











Central Park vs Meteostat

Descriptive Statistics

Feature	count	mean	std	min	25%	50%	75%	max
temp_delta	3625	0.506648	1.532693	-7.3	-0.5	0.6	1.2	5.6
rhum_delta	3625	0.259862	7.109542	-39	-4	0	4	59
prcp_delta	3500	0.057943	0.536325	-6.4	0	0	0	10.2
wdir_delta	3400	56.57824	52.9146	0	10	40	100	180
wspd_delta	3400	12.89944	7.930793	-4.1	7.6	11.7	17.4	47.9

Box Plots