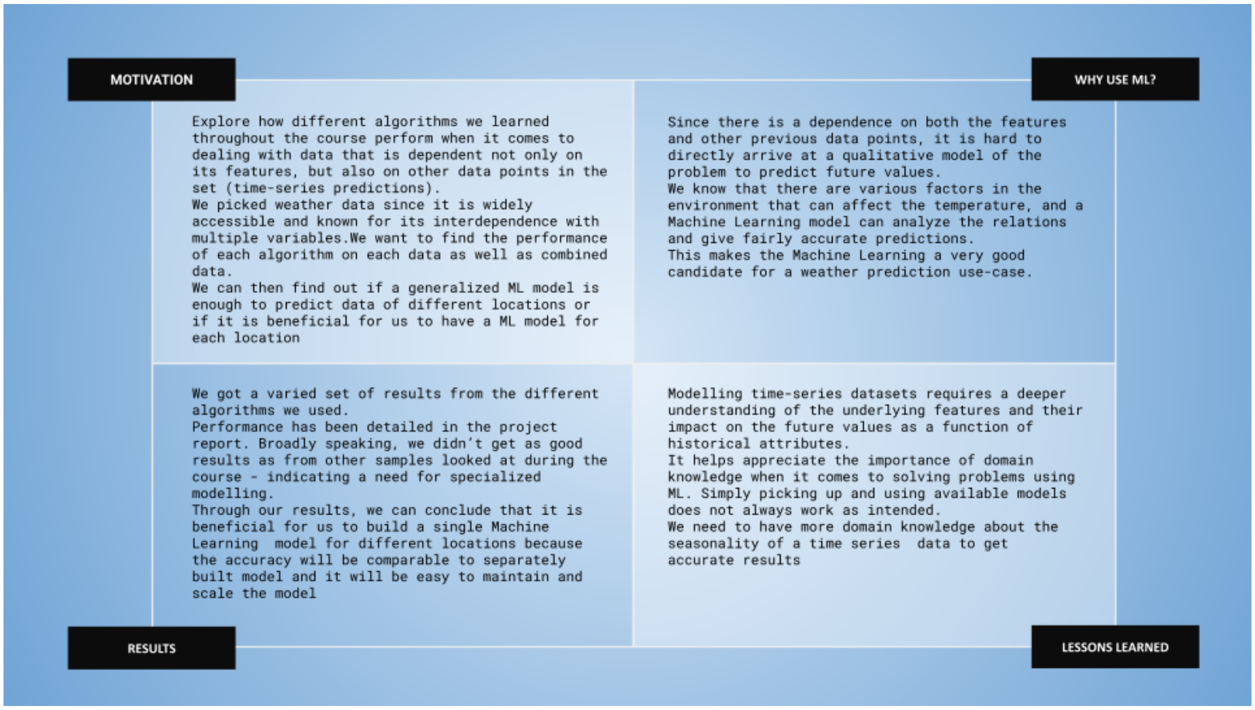


MOTIVATION

Explore how different algorithms we learned throughout the course perform when it comes to dealing with data that is dependent not only on its features, but also on other data points in the set (time-series predictions).
We picked weather data since it is widely accessible and known for its interdependence with multiple variables.We want to find the performance of each algorithm on each data as well as combined data.
We can then find out if a generalized ML model is enough to predict data of different locations or if it is beneficial for us to have a ML model for each location

WHY USE ML?

Since there is a dependence on both the features and other previous data points, it is hard to directly arrive at a qualitative model of the problem to predict future values.
We know that there are various factors in the environment that can affect the temperature, and a Machine Learning model can analyze the relations and give fairly accurate predictions.
This makes the Machine Learning a very good candidate for a weather prediction use-case.

RESULTS

We got a varied set of results from the different algorithms we used.
Performance has been detailed in the project report. Broadly speaking, we didn't get as good results as from other samples looked at during the course - indicating a need for specialized modelling.
Through our results, we can conclude that it is beneficial for us to build a single Machine Learning model for different locations because the accuracy will be comparable to separately built model and it will be easy to maintain and scale the model

LESSONS LEARNED

Modelling time-series datasets requires a deeper understanding of the underlying features and their impact on the future values as a function of historical attributes.
It helps appreciate the importance of domain knowledge when it comes to solving problems using ML. Simply picking up and using available models does not always work as intended.
We need to have more domain knowledge about the seasonality of a time series data to get accurate results

Project Report

Analysis of Time-Series Data Predictions using Weather Data

ECE-GY 6143 Machine Learning

Palak Keni - pk2539
Chinmay Nivsarkar - cmn8525

Introduction

Time series data is a type of data that is collected over a period of time, typically at regular intervals. This data consists of a sequence of measurements, observations, or events that are recorded at specific times, and can be used to understand trends and patterns in the data over time. Analytics for time series involves a range of techniques and methods that are designed to extract useful insights and information from the data, such as statistical analysis, machine learning algorithms, and data visualization techniques.

In order to predict future values for any dataset using existing datasets, we need to use some form of regression. Regression models are statistical models that are used to predict a continuous outcome variable based on one or more predictor variables. These models are commonly used in time series data analytics to forecast future values of a given variable based on its past values.

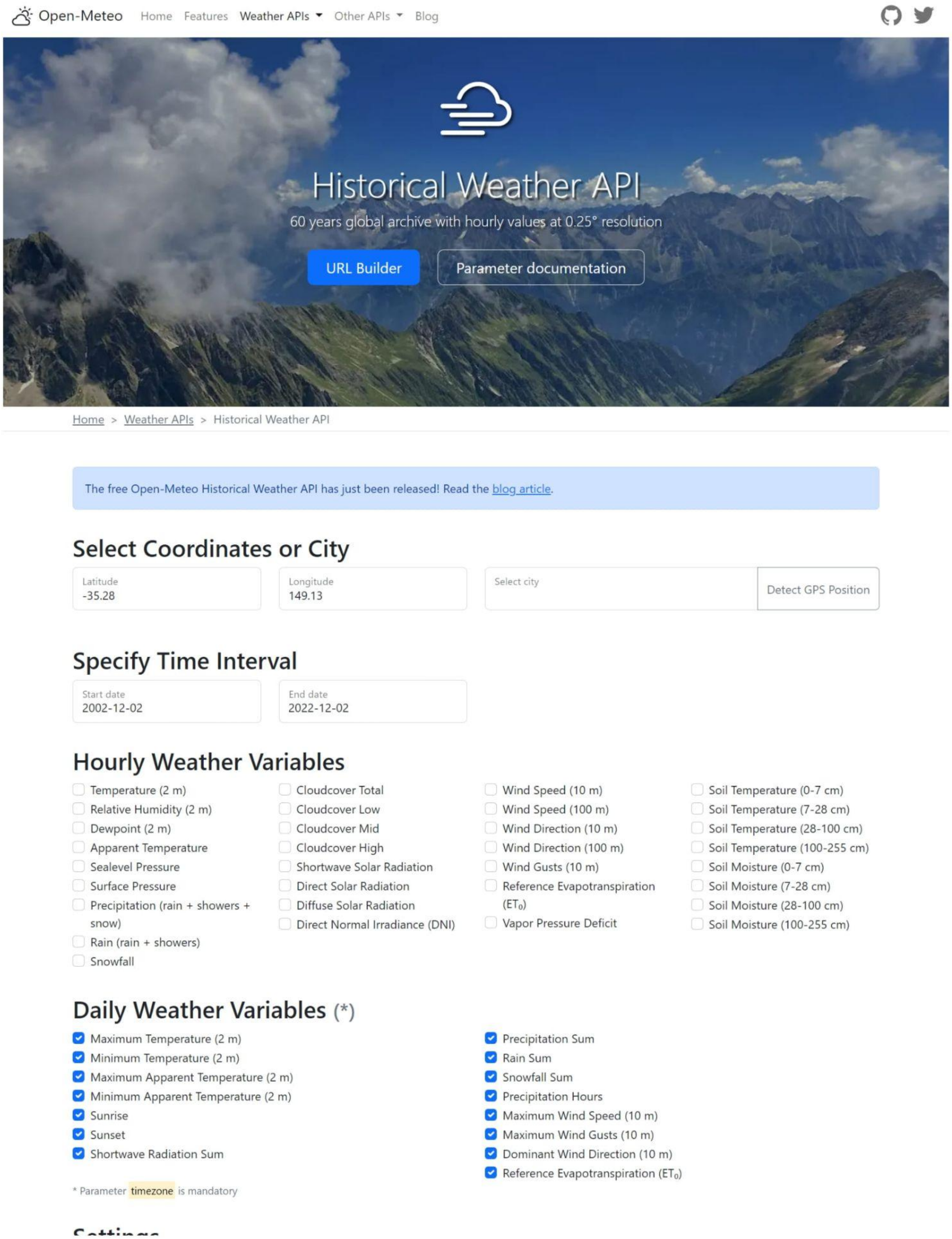
Linear regression is one of the most commonly used regression models for time series data. This model assumes that there is a linear relationship between the predictor variable (the time series data) and the outcome variable (the value being predicted). Linear regression models can be used to make predictions about future values of a time series based on a set of past values.

Dataset

For the weather data, we referred to the Open-Meteo Weather API (https://open-meteo.com/en/docs/historical-weather-api) . The idea was to gain access to a reliable datastream which could be used as a long term feed for the model once we were able to train a reliable data model for our predictions.

In the short term, we have prepared a data dump for the last ~20 years of data to use in the code in a static manner. We’ve prepared extracts for the following 5 cities: New York, Mumbai, London, Canberra, and Los-Angeles.

Below are the parameters for the extract for New York city location:



Home > Weather APIs > Historical Weather API

The free Open-Meteo Historical Weather API has just been released! Read the blog article.

Select Coordinates or City

Latitude: -35.28

Longitude: 149.13

Select city:

Specify Time Interval

Start date: 2002-12-02

End date: 2022-12-02

Hourly Weather Variables

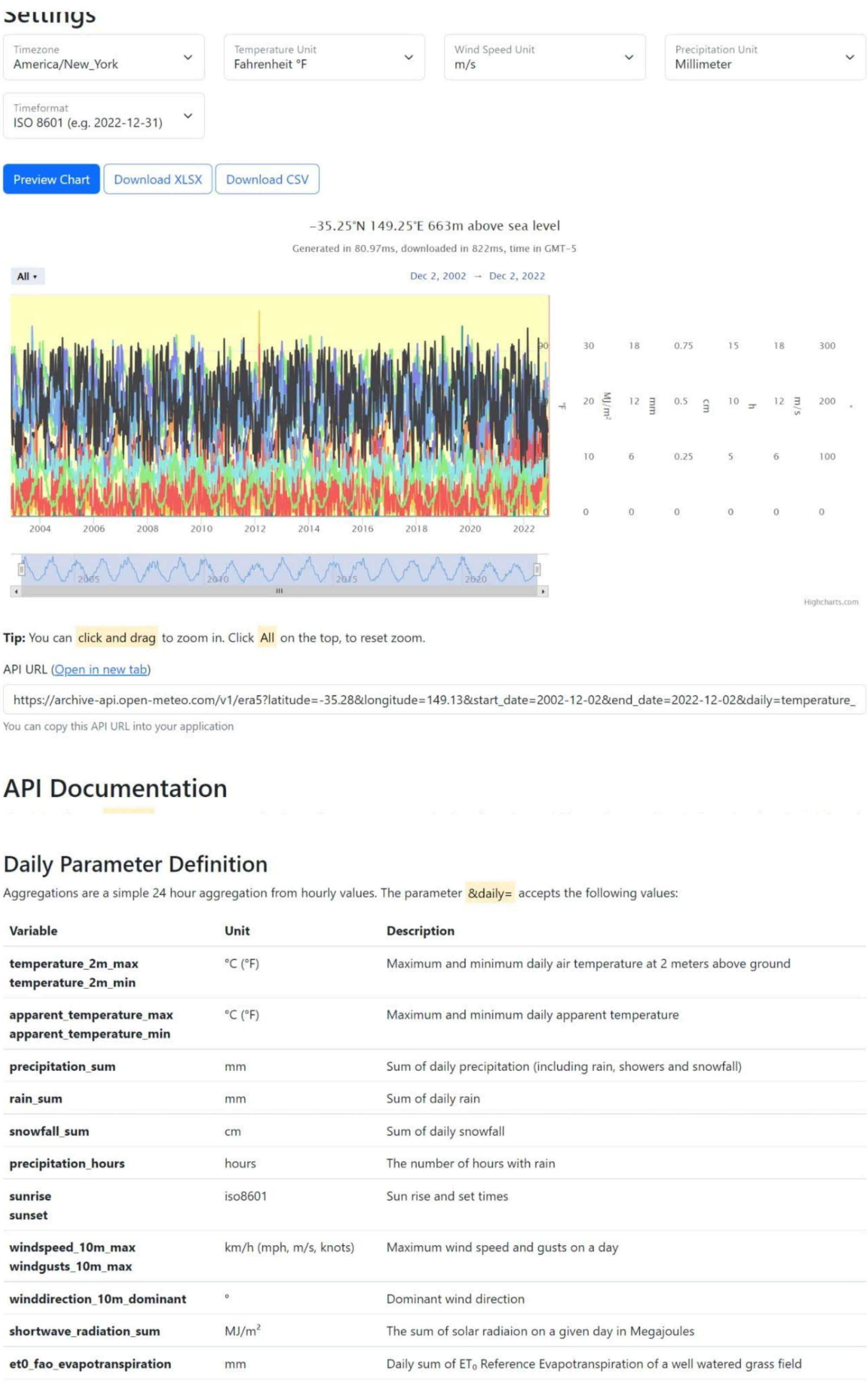
- □ Temperature (2 m)
- □ Relative Humidity (2 m)
- □ Dewpoint (2 m)
- □ Apparent Temperature
- □ Sealevel Pressure
- □ Surface Pressure
- □ Precipitation (rain + showers + snow)
- □ Rain (rain + showers)
- □ Snowfall
- □ Cloudcover Total
- □ Cloudcover Low
- □ Cloudcover Mid
- □ Cloudcover High
- □ Shortwave Solar Radiation
- □ Direct Solar Radiation
- □ Diffuse Solar Radiation
- □ Direct Normal Irradiance (DNI)
- □ Wind Speed (10 m)
- □ Wind Speed (100 m)
- □ Wind Direction (10 m)
- □ Wind Direction (100 m)
- □ Wind Gusts (10 m)
- □ Reference Evapotranspiration (ET_0)
- □ Vapor Pressure Deficit
- □ Soil Temperature (0-7 cm)
- □ Soil Temperature (7-28 cm)
- □ Soil Temperature (28-100 cm)
- □ Soil Temperature (100-255 cm)
- □ Soil Moisture (0-7 cm)
- □ Soil Moisture (7-28 cm)
- □ Soil Moisture (28-100 cm)
- □ Soil Moisture (100-255 cm)

Daily Weather Variables (*)

- ☑ Maximum Temperature (2 m)
- ☑ Minimum Temperature (2 m)
- ☑ Maximum Apparent Temperature (2 m)
- ☑ Minimum Apparent Temperature (2 m)
- ☑ Sunrise
- ☑ Sunset
- ☑ Shortwave Radiation Sum
- ☑ Precipitation Sum
- ☑ Rain Sum
- ☑ Snowfall Sum
- ☑ Precipitation Hours
- ☑ Maximum Wind Speed (10 m)
- ☑ Maximum Wind Gusts (10 m)
- ☑ Dominant Wind Direction (10 m)
- ☑ Reference Evapotranspiration (ET_0)

* Parameter timezone is mandatory

Settings



Settings

Timezone: America/New_York

Temperature Unit: Fahrenheit °F

Wind Speed Unit: m/s

Precipitation Unit: Millimeter

Timeformat: ISO 8601 (e.g. 2022-12-31)

Tip: You can click and drag to zoom in. Click All on the top, to reset zoom.

API URL (Open in new tab)

https://archive-api.open-meteo.com/v1/era5?latitude=-35.28&longitude=149.13&start_date=2002-12-02&end_date=2022-12-02&daily=temperature_

You can copy this API URL into your application

API Documentation

Daily Parameter Definition

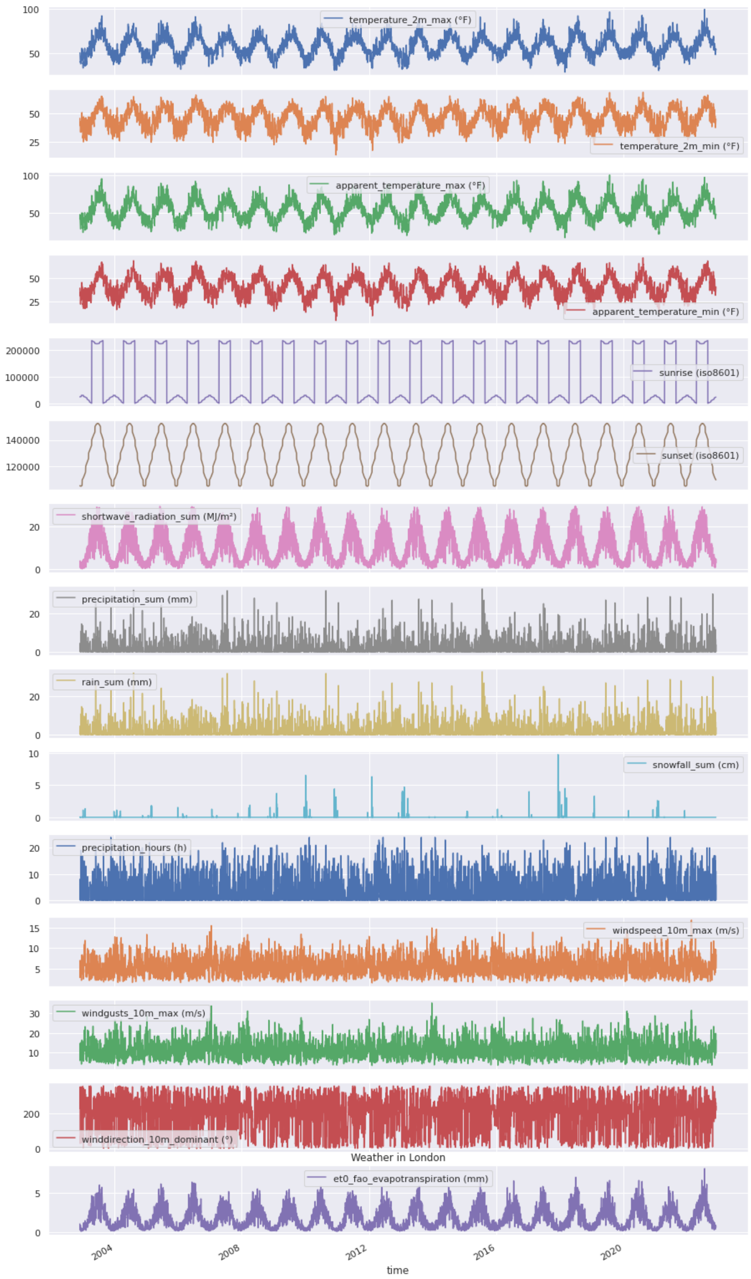
Aggregations are a simple 24 hour aggregation from hourly values. The parameter &daily= accepts the following values:

Variable	Unit	Description
temperature_2m_max **temperature_2m_min**	°C (°F)	Maximum and minimum daily air temperature at 2 meters above ground
apparent_temperature_max **apparent_temperature_min**	°C (°F)	Maximum and minimum daily apparent temperature
precipitation_sum	mm	Sum of daily precipitation (including rain, showers and snowfall)
rain_sum	mm	Sum of daily rain
snowfall_sum	cm	Sum of daily snowfall
precipitation_hours	hours	The number of hours with rain
sunrise **sunset**	iso8601	Sun rise and set times
windspeed_10m_max **windgusts_10m_max**	km/h (mph, m/s, knots)	Maximum wind speed and gusts on a day
winddirection_10m_dominant	°	Dominant wind direction
shortwave_radiation_sum	MJ/m^2	The sum of solar radiaion on a given day in Megajoules
et0_fao_evapotranspiration	mm	Daily sum of ET$_0$ Reference Evapotranspiration of a well watered grass field

Identical parameters are used for other cities extract

Solution Approach

Since we have already gained access to a dataset, the first step is to visualize and see if we can gauge any trends in the data since this will be able to help us in model selection.



It is quite apparent from this visualization of London data that we have a seasonality to our attributes, along with a predictable range of variation for most of these attributes.

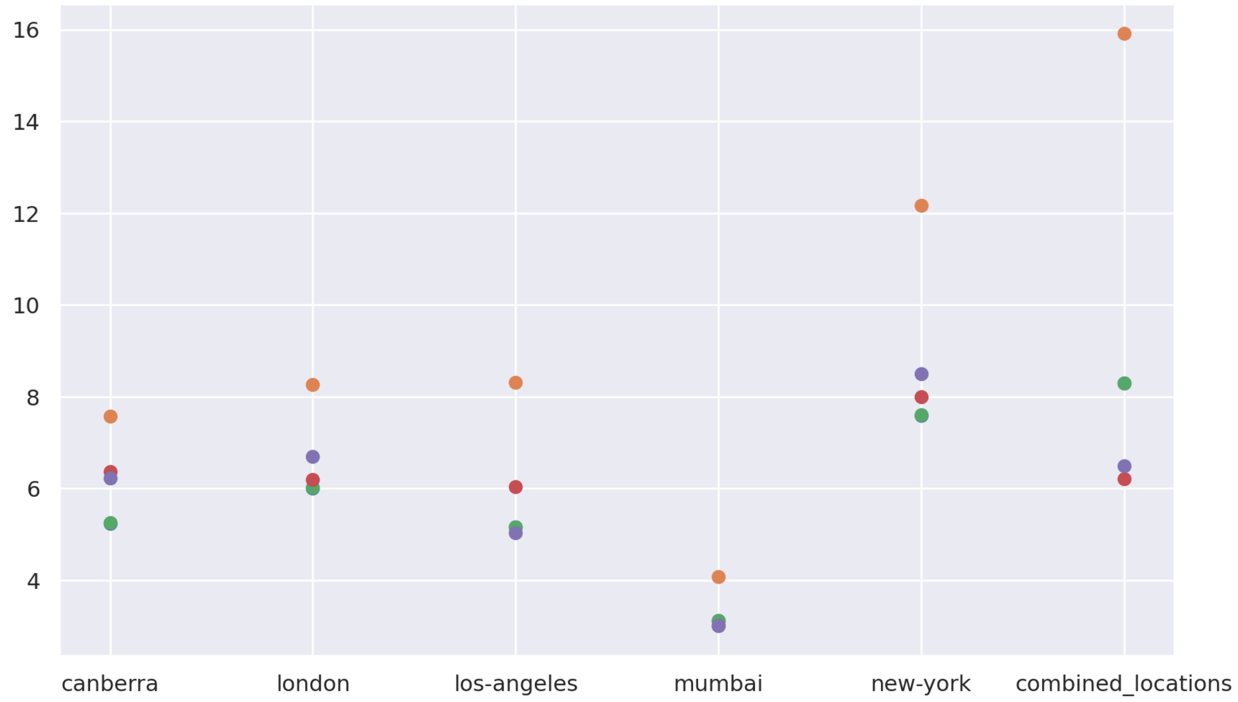
We also performed some causal analysis since we have time series data and found that all of our attributes seemed to be related to our target variables.

Results

We tried using the following models:

Linear Regression
Lasso Regression
Ridge Regression
K-Nearest Neighbors
Decision Tree
Vector AutoRegression (VAR)

We achieved best performance using linear and ridge regression. Despite being suited for time series data, VAR gave us the worst RMSE.



As expected, the model scores varied as the location changed. The best model score was consistently achieved on the Mumbai dataset, which leads us to the conclusion that this city may have the least complicated weather dynamic for predicting temperature.

References

- https://scikit-learn.org/
- https://www.statsmodels.org/dev/vector_ar.html
- https://www.statology.org/granger-causality-test-in-python
- https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/
- https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/
- https://keremkargin.medium.com/ridge-regression-fundamentals-and-modeling-in-python-bb56f4301f62
- https://machinelearningmastery.com/ridge-regression-with-python/
- https://medium.com/analytics-vidhya/lasso-regression-fundamentals-and-modeling-in-python-ad8251a636cd
- https://towardsdatascience.com/vector-autoregressive-for-forecasting-time-series-a60e6f168c70