Forecasting Temperature in New York City

DS 4002, Project 2: Time Series Data, 03/24/2025

Group 2 - Maggie Crowner (Group Leader), Ella Thomasson, Emily McMahon

CONTEXT & HYPOTHESIS

MOTIVATIONS

We know the climate is changing, and temperatures are trending upwards. We also believe that precipitation can have an effect on daily temperature.

We want to be able to forecast temperature a year in advance. NYC exhibits all four seasons, so it is an appropriate choice of city for this analysis.

GOALS

Forecast the monthly average maximum daily temperature for 2023 in NYC, based on daily temperature measures from 1970-2022 and precipitation measures from 1970-2023.

RESEARCH QUESTIONS

Can we forecast daily temperatures for 2023 with a SARIMAX Time Series Model, using precipitation as an exogenous variable? Is precipitation a significant predictor for temperature?

HYPOTHESIS

We can forecast NYC temperature for 2023 within a visualized 95% confidence interval based on precipitation and previous temperature measurements.

MODELING APPROACH

SARIMAX Time Series Model

Time series data

Likely seasonality component

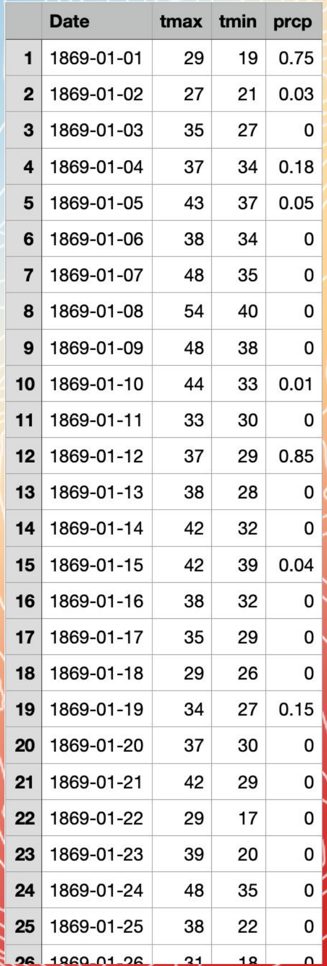
Exogenous variable: precipitation

We want to predict temperature for 2023, so we will visualize our predictions to ensure the actual 2023 temperature falls within our calculated confidence interval.

DATA ACQUISITION & PREPROCESSING

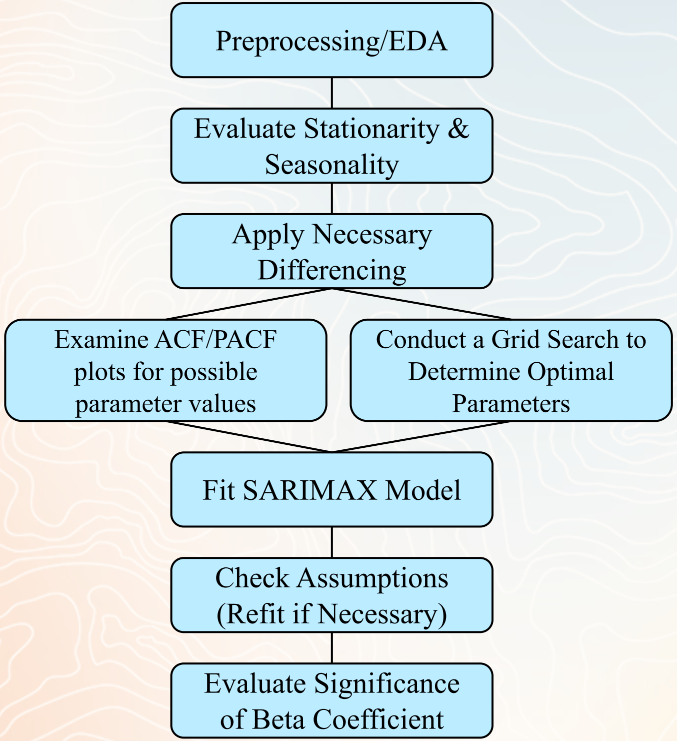
This data is from Carnegie Mellon University, and includes all temperature maximum, minimum, and precipitation measurements for each day from 1869-2024.

- Subsetted to only include data from 1970-2023
  - 2023 observations will be the validation set
- Chose ‘tmax’ as primary variable to model, and ‘prcp’ as possible exogenous variable
  - Literature suggests that precipitation may have an impact on temperature
- Created monthly aggregates to eliminate a possible second seasonality component
- Converted ‘Date’ to a datetime object to be used for time series analysis



	Date	tmax	tmin	prcp
1	1869-01-01	29	19	0.75
2	1869-01-02	27	21	0.03
3	1869-01-03	35	27	0
4	1869-01-04	37	34	0.18
5	1869-01-05	43	37	0.05
6	1869-01-06	38	34	0
7	1869-01-07	48	35	0
8	1869-01-08	54	40	0
9	1869-01-09	48	38	0
10	1869-01-10	44	33	0.01
11	1869-01-11	33	30	0
12	1869-01-12	37	29	0.85
13	1869-01-13	38	28	0
14	1869-01-14	42	32	0
15	1869-01-15	42	39	0.04
16	1869-01-16	38	32	0
17	1869-01-17	35	29	0
18	1869-01-18	29	26	0
19	1869-01-19	34	27	0.15
20	1869-01-20	37	30	0
21	1869-01-21	42	29	0
22	1869-01-22	29	17	0
23	1869-01-23	39	20	0
24	1869-01-24	48	35	0
25	1869-01-25	38	22	0
26	1869-01-26	31	18	0

ANALYSIS PLAN



Preprocessing/EDA

Evaluate Stationarity & Seasonality

Apply Necessary Differencing

Examine ACF/PACF plots for possible parameter values

Conduct a Grid Search to Determine Optimal Parameters

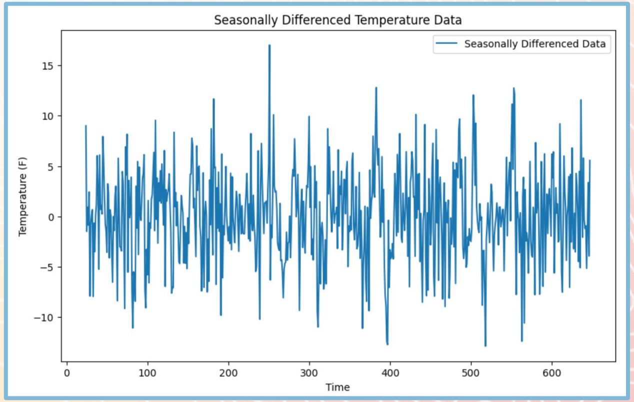
Fit SARIMAX Model

Check Assumptions (Refit if Necessary)

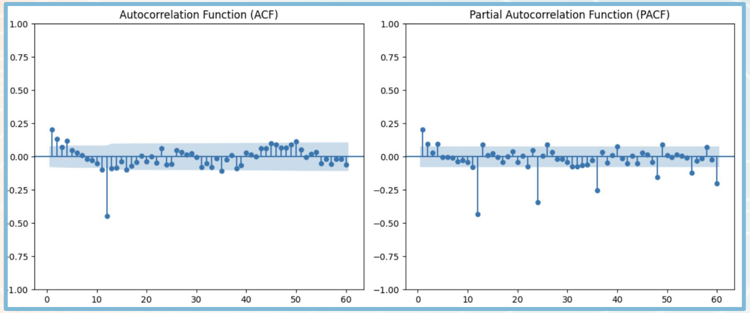
Evaluate Significance of Beta Coefficient

JUSTIFICATION

- SARIMAX model includes seasonality component, possible AR component, possible differencing, possible MA component, and exogenous variable
- We performed seasonal differencing to make our data stationary before proceeding with model building



CHOOSING PARAMETERS



Tricky analysis decision! 6 candidate models based on interpreting these graphs visually.

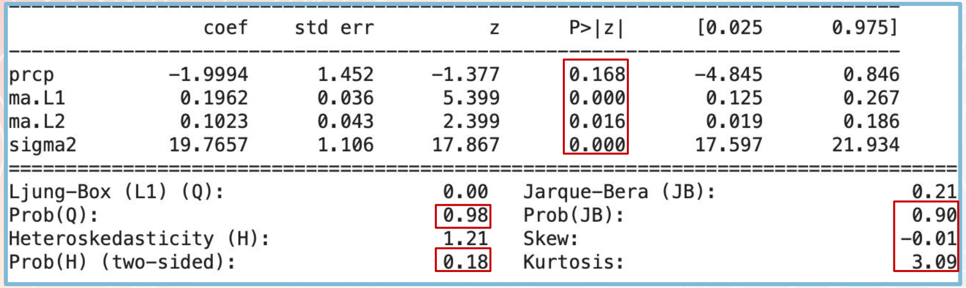
- SARIMAX(2,0,0) x (0,1,1,12)
- SARIMAX(2,0,0) x (0,1,0,12)
- SARIMAX(1,0,0) x (0,1,1,12)
- SARIMAX(1,0,0) x (0,1,0,12)
- SARIMAX(0,0,2) x (0,1,1,12)
- SARIMAX(0,0,2) x (0,1,0,12)

BIAS/UNCERTAINTY

- The ‘tmax’ variable (after preprocessing) is a monthly average of the daily maximum temperatures observed, which is difficult to interpret as it is not actually the monthly maximum or monthly average
- Parameters were chosen visually, meaning there is room for human error and misinterpretation
  - Performing grid search for optimization did not yield a successful model

RESULTS & CONCLUSIONS

Final model: **SARIMA(0,0,2) x (0,1,0,12)**

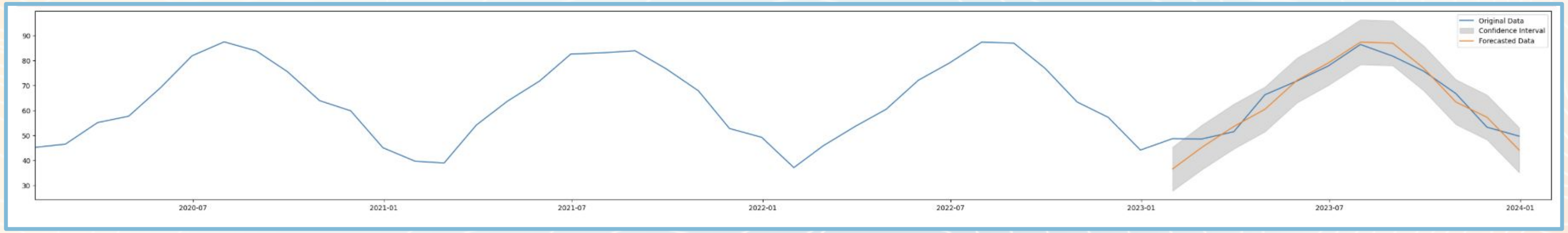


```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
prcp          -1.9994      1.452     -1.377      0.168      -4.845       0.846
ma.L1          0.1962      0.036      5.399      0.000       0.125       0.267
ma.L2          0.1023      0.043      2.399      0.016       0.019       0.186
sigma2        19.7657      1.106     17.867      0.000      17.597      21.934
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 0.21
Prob(Q):                              0.98   Prob(JB):                         0.90
Heteroskedasticity (H):               1.21   Skew:                            -0.01
Prob(H) (two-sided):                  0.18   Kurtosis:                         3.09
```

Assumptions are met!

'prcp' is not significant in the model → Precipitation does not have a significant effect on temperature, we can remove the exogenous variable and conclude with a SARIMA model

2023 forecasted temperature measurements versus actual temperature measurements:



NEXT STEPS

- Explore other exogenous variable options to find one that is significant and assists with accurate forecasting
- Predict further into the future to see if the model is still producing accurate forecasts
- Consider performing this analysis with minimum temperature in addition to maximum

REFERENCES

GitHub: https://github.com/maggiecrowner/DS4002-Project-2/tree/main

References:

[1] Melanie, “SARIMAX model: What is it? How can it be applied to time series?,” *Data Science Courses | DataScientest*, Mar. 18, 2024.
[2] J. A. Acero, P. Kestel, H. T. Dang, and L. K. Norford, “Impact of rainfall on air temperature, humidity and thermal comfort in tropical urban parks,” *Urban Climate*, vol. 56, p. 102051, Jul. 2024, doi: https://doi.org/10.1016/j.uclim.2024.102051.
[3] Y. Lai and D. Dzombak, “Compiled historical daily temperature and precipitation data for selected 210 U.S. cities,” *figshare*, Aug. 2024, doi: https://doi.org/10.1184/u002FR1/u002F7890488.v6.
[4] “Complete Guide To SARIMAX in Python,” *GeeksforGeeks*, Dec. 26, 2023. https://www.geeksforgeeks.org/complete-guide-to-sarimax-in-python/

THANKS!

QUESTIONS?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**