Analysis and prediction of average temperature based on meteorological data

Mid Bootcamp Project

L. T. OZGUR YILDIRIM

Outline

- Introduction

- Materials and Methods

- Results and Discussion

- Conclusion

Introduction

This study analyses and predicts the average temperature of two European cities, Dusseldorf and Muenchen in Germany using machine learning models. Dataset consists of meteorological data ranging from 2000 to 2010.

The main objectives of this study include the following investigations:

- Analyzing relationships between meteorological parameters

- Hypothesis testing to decide whether the data sufficiently support the hypothesis

- Building linear and logistic regression models and time series forecasting to predict the average temperature

Dusseldorf data: train the linear regression model + develop logistic regression model + time series forecasting

Muenchen data: test the linear regression model

Materials and Methods

Meteorological parameters in the dataset are average temperature, precipitation, humidity, cloud cover, wind speed, sunshine, and barbeque weather. The studies are performed on .csv and excel files using Python, Mysql, and Tableau softwares.

The methods of this study include the followings:

- Get data

- Exploratory data analyses

- Data cleaning

- Data visualization

- Hypothesis testing $\longrightarrow$
  - **One-sided hypothesis testing:** 11.17℃ is the avg. temperature in Dusseldorf between 2000-2009.
- Data wrangling
  - Null hypothesis: avg. temperature in Dusseldorf between 2000-2010 =< 10.8°C
- Build model
  - Alternative hypothesis: avg. temperature in Dusseldorf between 2000-2010 > 10.8 ℃

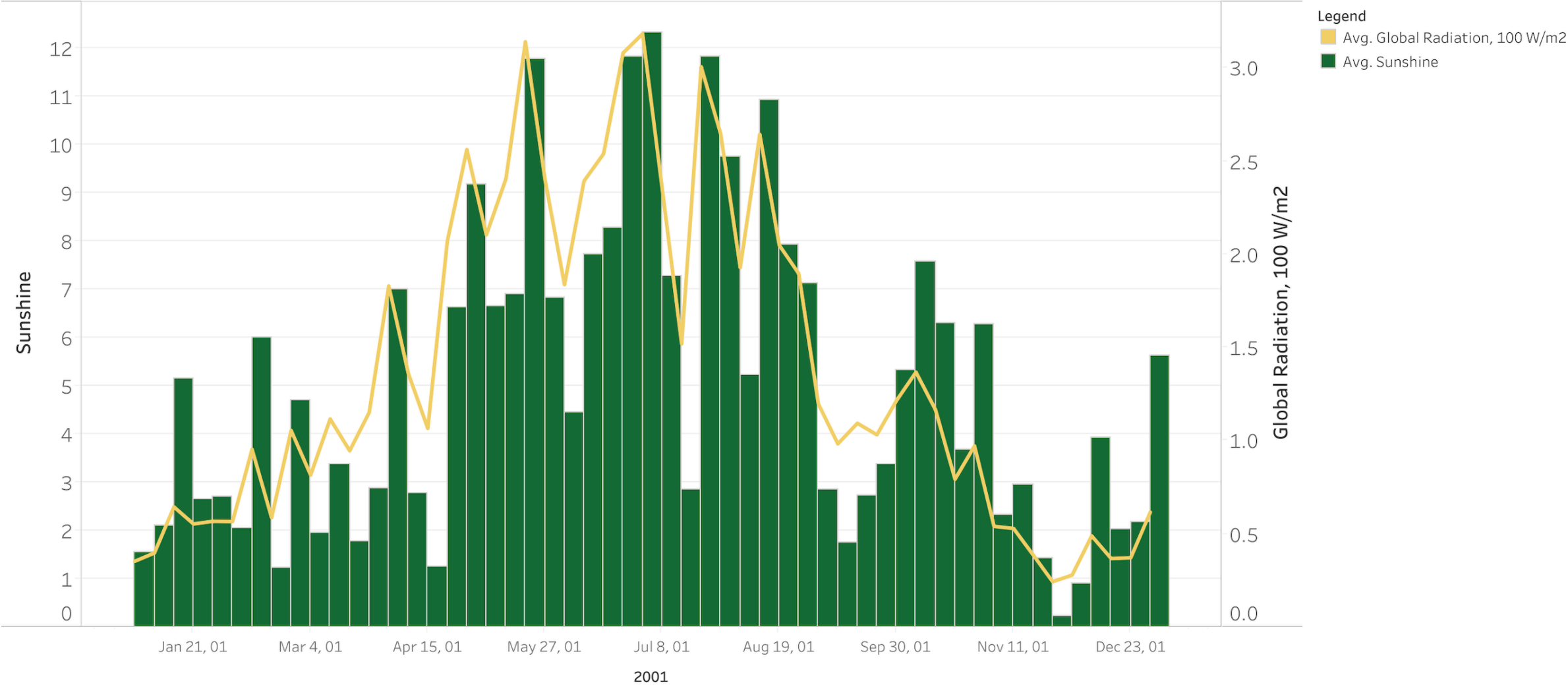P-value: 9.57×10^{-4}

Materials and Methods

Temperature (°C) vs. Precipitation (mm) in Dusseldorf (2000-2010)
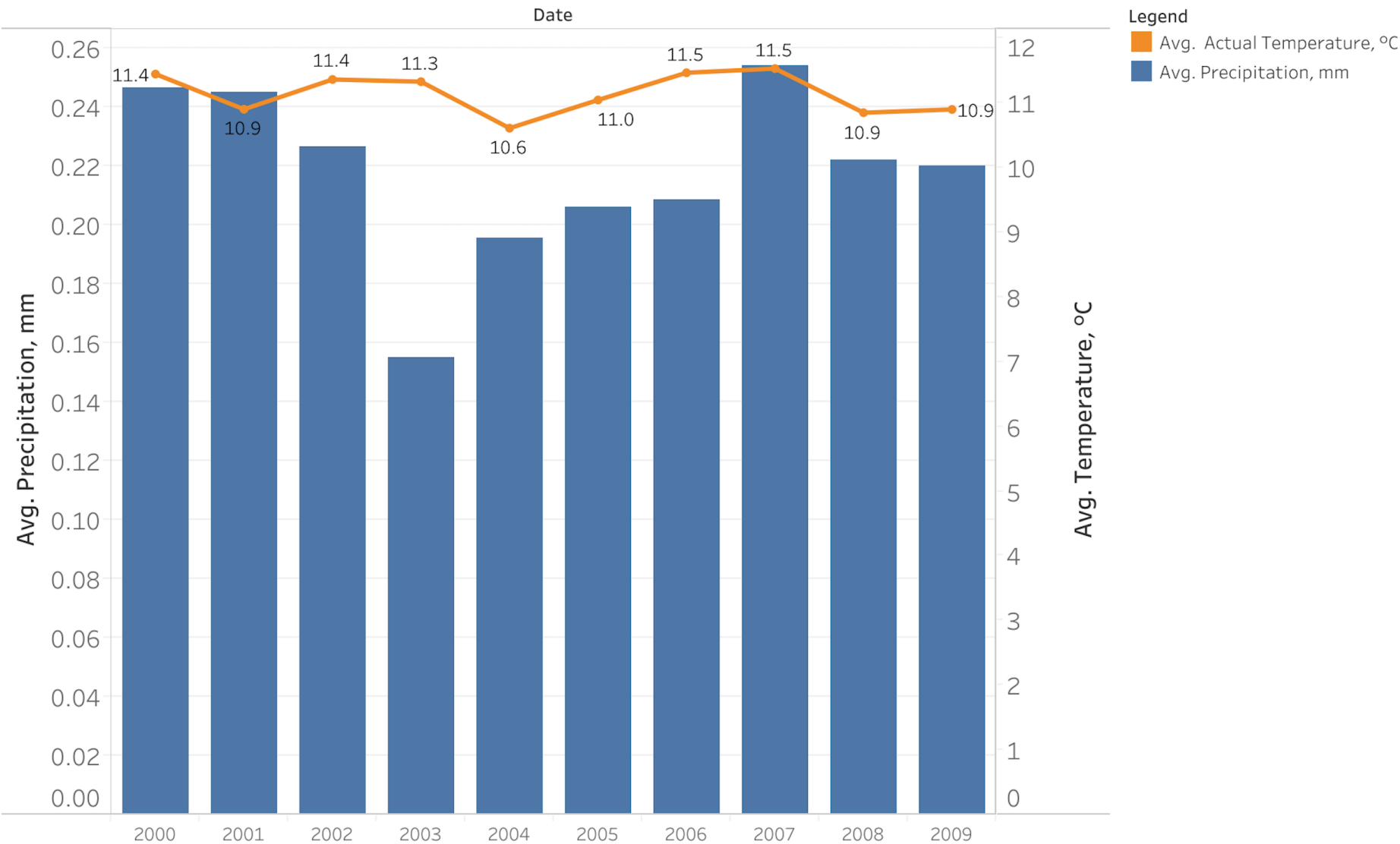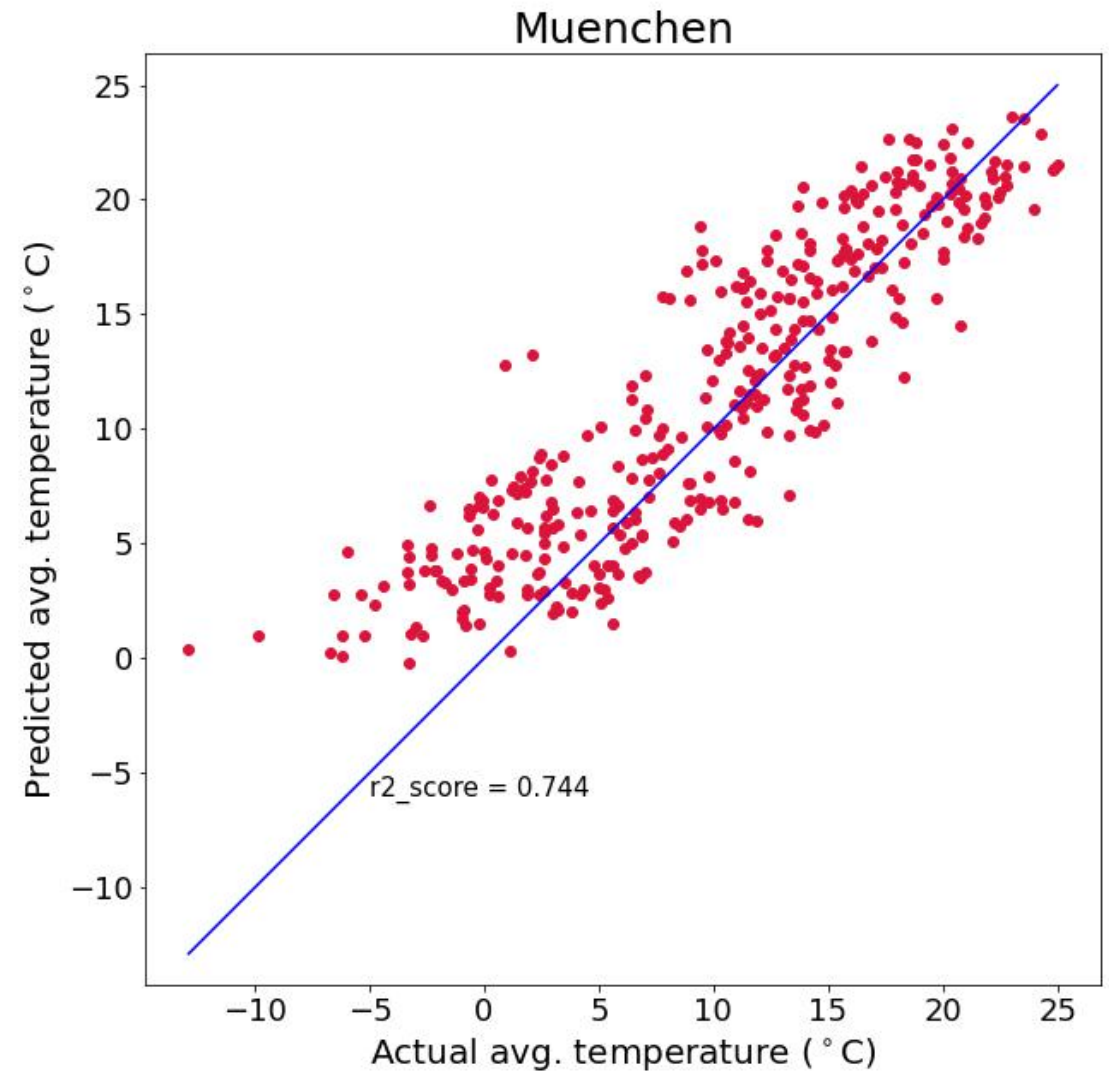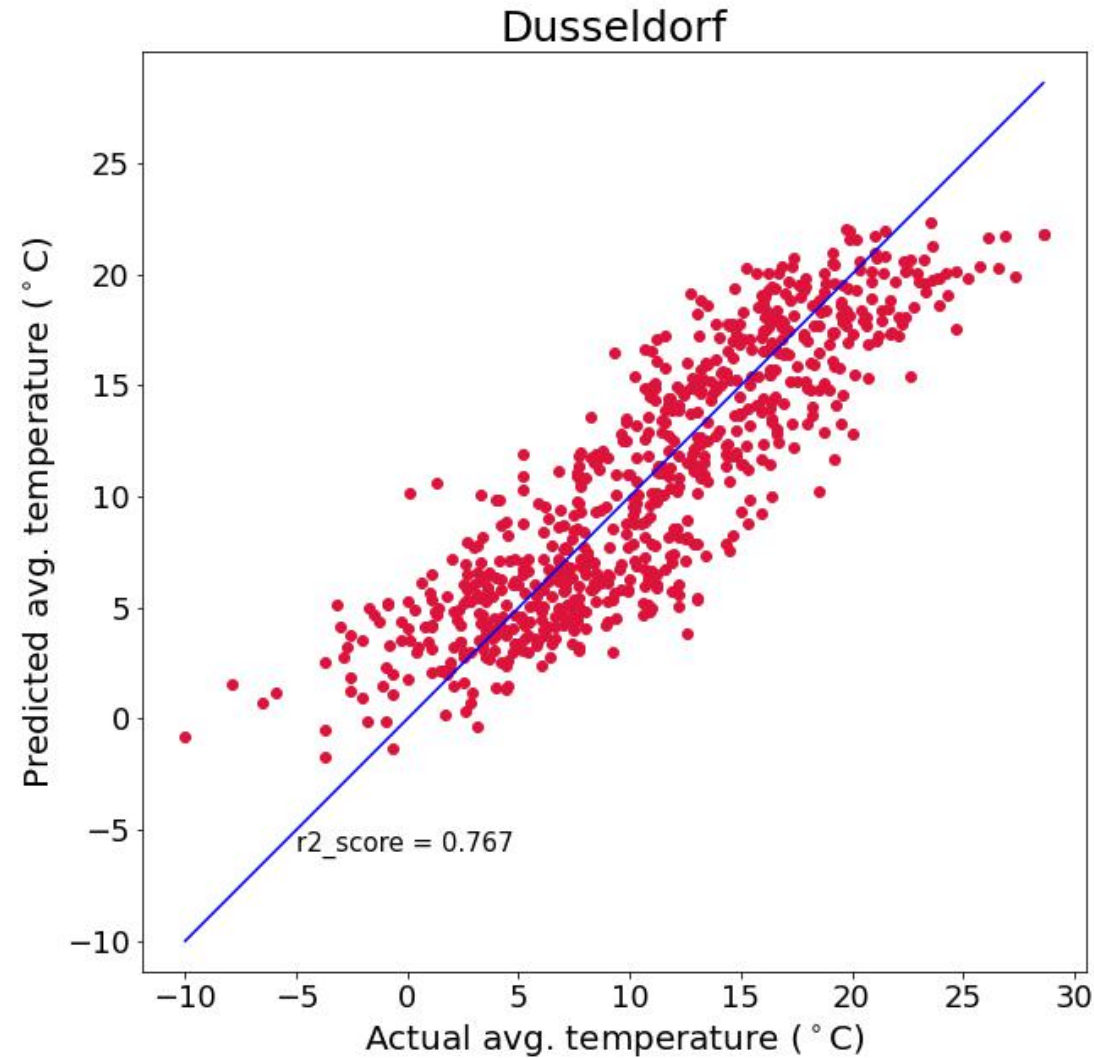
Materials and Methods

Global radiation (100 W/m2) vs. Sunshine in Muenchen (2001)

Results and Discussion

- **Linear regression model**
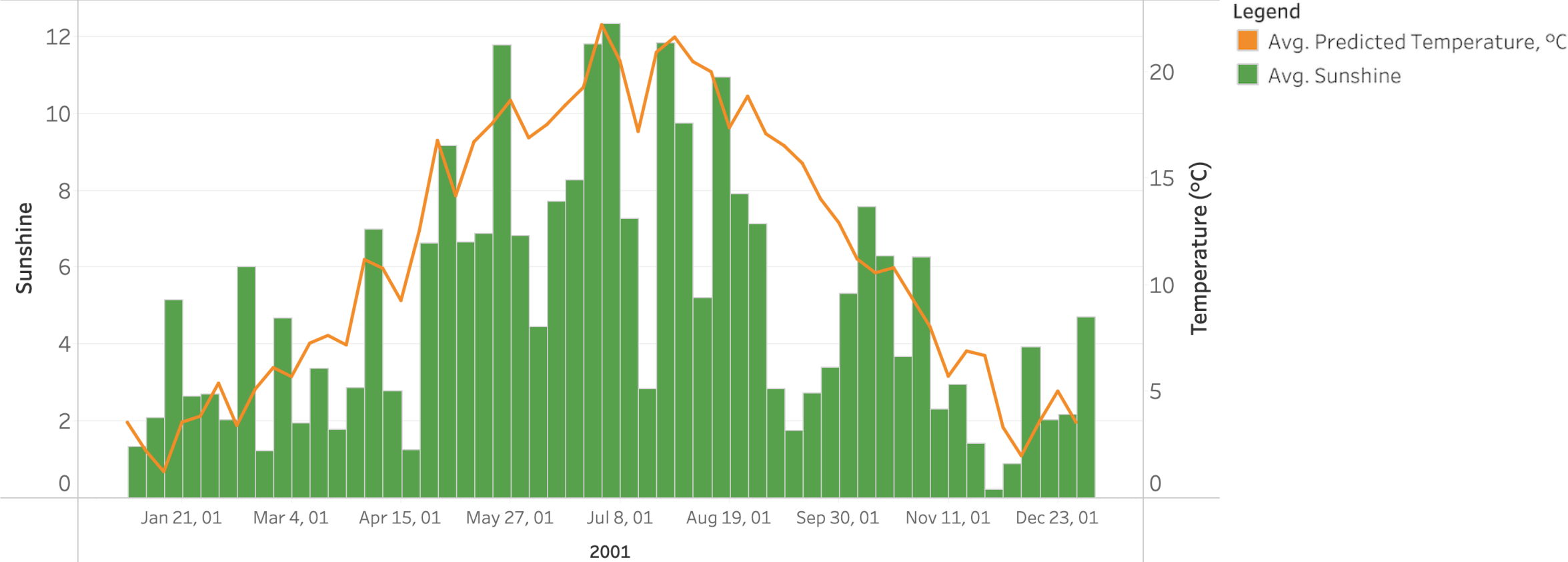
Results and Discussion

- **Linear regression model**

Actual Temperature (°C) vs. Predicted Temperature (°C) in Muenchen (2001)
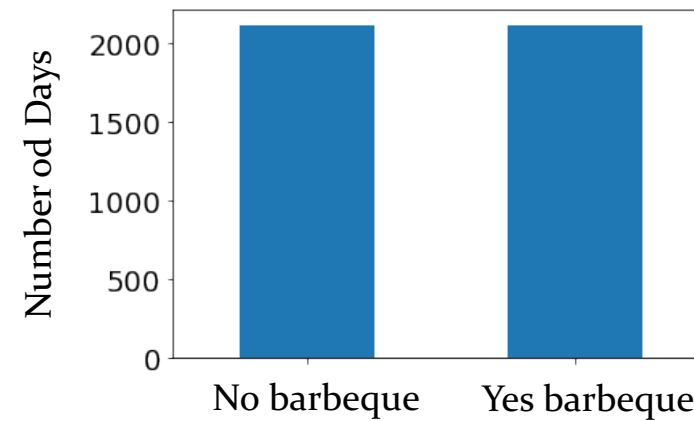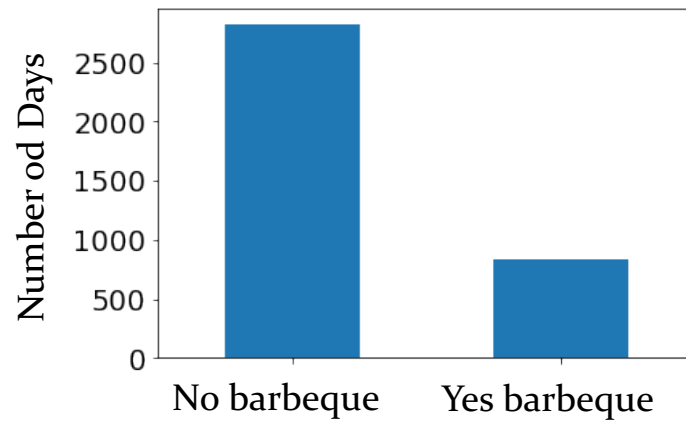
Results and Discussion

Avg. Predicted Temperature (°C) vs. Sunshine in Muenchen (2001)

Results and Discussion

- ## Logistic regression model

  - Unevenly distributed data : no barbeque in Dusseldorf in 80% of the days between 2000-2010.

  - Data is manipulated.

  - Logistic regression score: 0.932.

|	Precision	Recall	F1
Before data manipulation	0.87	0.83	0.85
After data manipulation	0.80	0.97	0.87

**Results and Discussion**

- **Time series forecasting**

- This plot shows that the temperature is unstable, along with its obvious seasonality.

- Trend shows the overall direction of data.

- Seasonality is a periodic component. It shows fluctuation in a the temperature sequence.

- The difference between trend and seasonal is residual.

Additive Decomposition
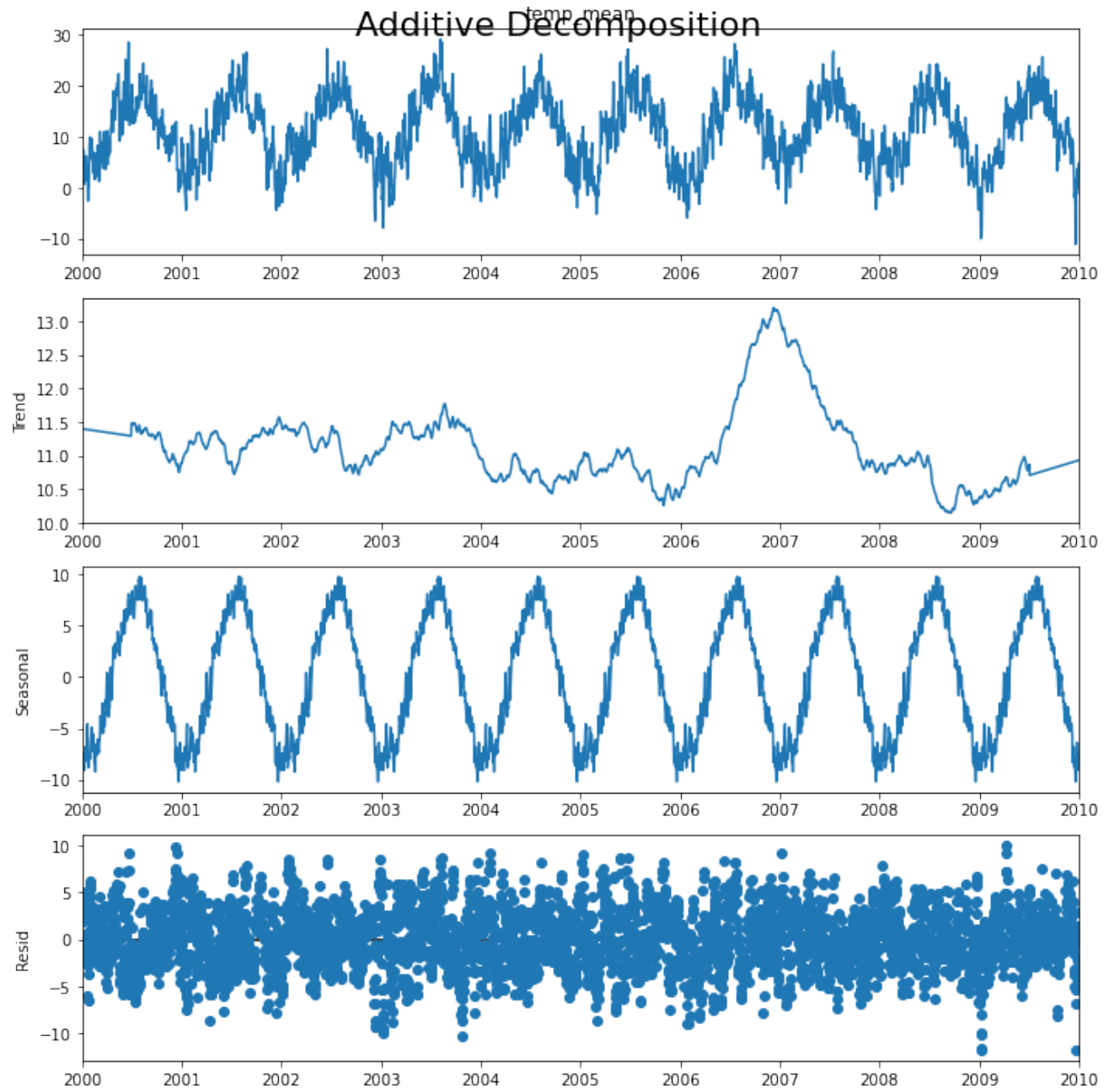
Results and Discussion

- **Time series forecasting**

  - Time and average temperature of Dusseldorf between 2000-2010 are used for time series forecasting.

  - ARIMA Model => The RMSE of static forecasting: 2.09 and the RMSE of dynamic forecasting: 7.00.
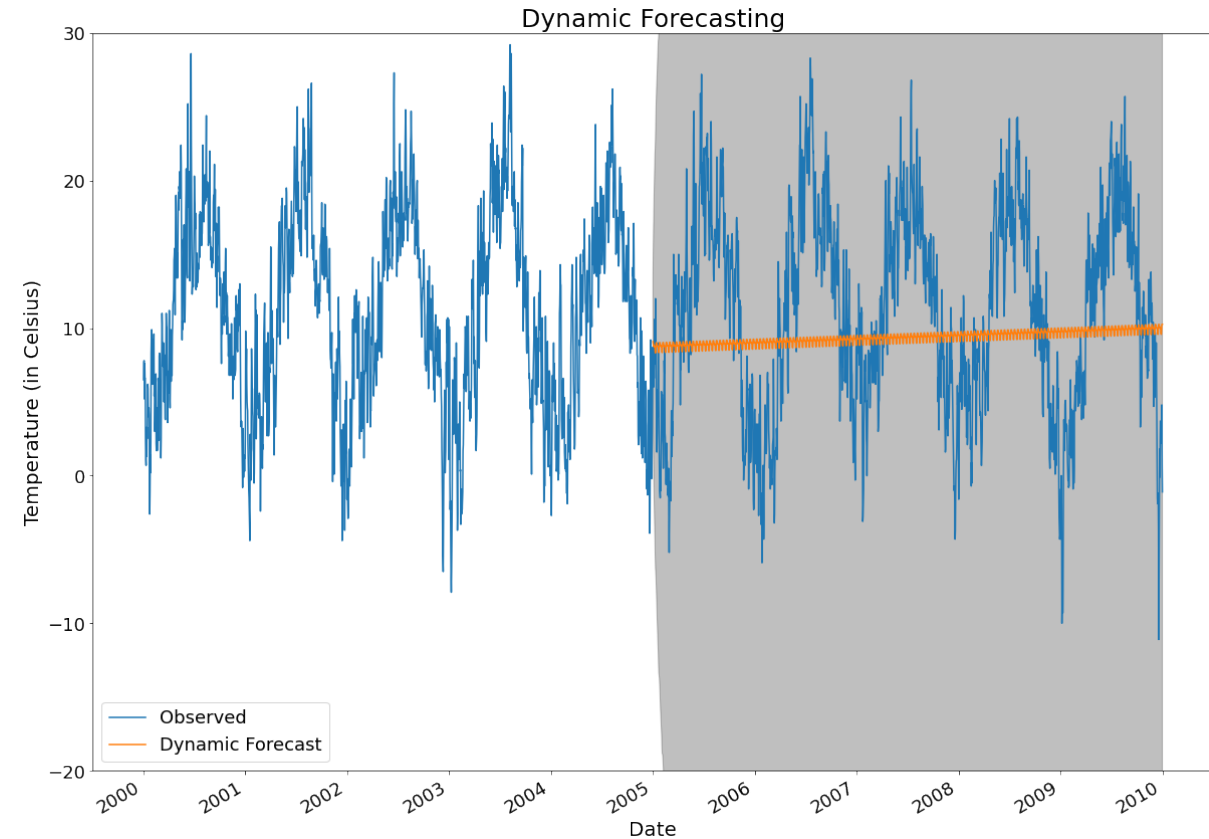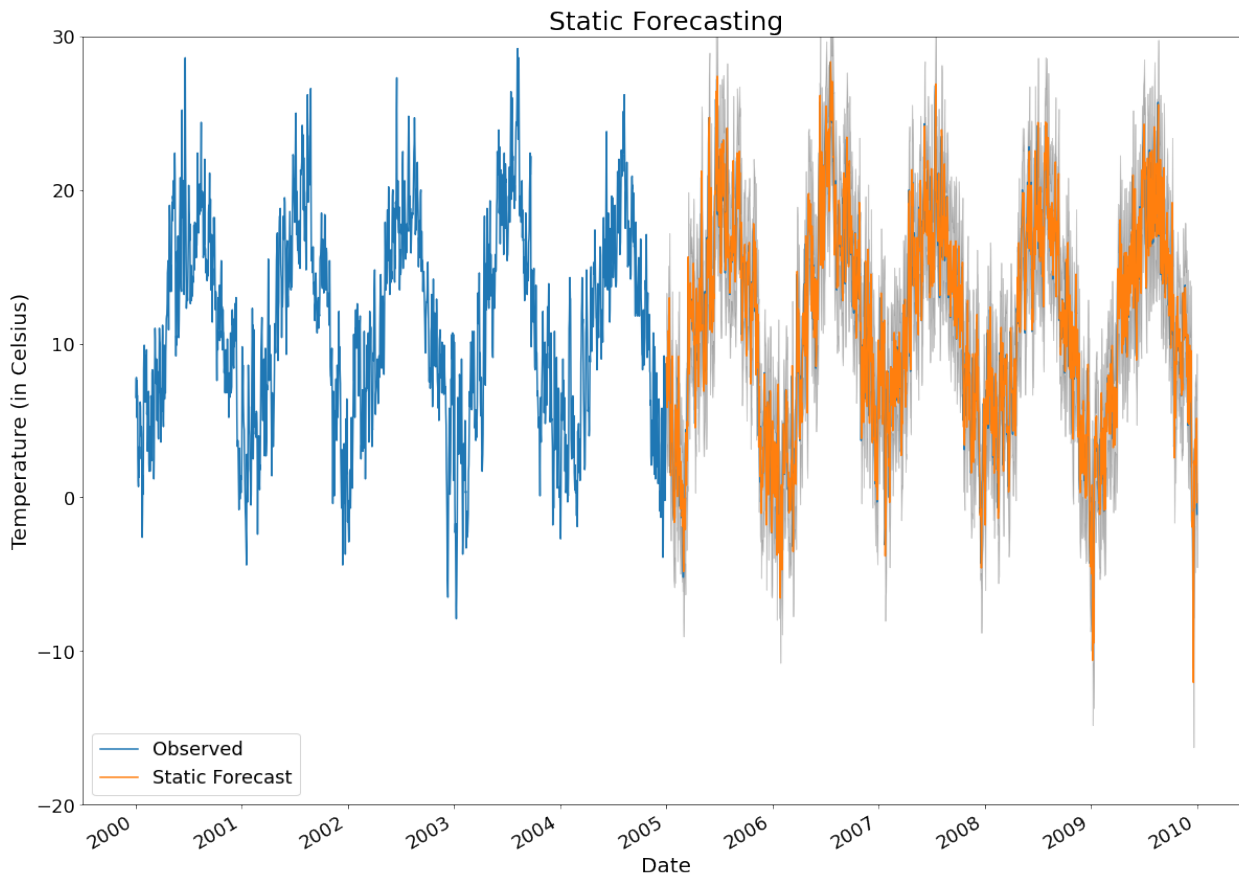
Conclusion

This study investigates the average temperatures for Dusseldorf and Muenchen cities based on meteorological parameters using linear and logistic regression models.

The linear regression model (0.75 r2 score) predicts that Muenchen has annual temperature of 11.50 °C in 2001.

Logistic regression model is useful before making plans outside as it classifies if the weather is suitable for outdoor activities or not.

Based on RMSE values, static forecasting of ARIMA model (RMSE=2.09) predicts better than the linear regression model (RMSE=3.20) in this project.

References

- Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment.

- Lai, 2020. Time Series Analysis and Weather Forecast in Python. Retrieved from: https://medium.com/@llmkhoa511/time-series-analysis-and-weather-forecast-in-python-e80b664c7f71.

Thank you!

Questions and Comments

Ironhack
Data Analytics Bootcamp