

Fundamentals of Data Science Final Project



Lecturer:

**D6211_NUNUNG NURUL QOMARIYAH, S.Kom.,
M.T.I., Ph.D.**

Arranged by :

2702345932_Daffa

2702357996_Edbert Tan

COMP6784001_Fundamentals of Data Science

Computer Science Faculty

Binus International University

Project name

Solar Power Plant optimization

Name and Group Members:

1. Daffa
2. Edbert Tan

1. Problem Analysis

In the current day and age, temperatures in various areas are always constantly changing. Some areas become colder while other areas become hotter over time. That statement holds true here in Indonesia too, as many areas in the span of the past few decades have had changes and fluctuations in the average temperature every year. Another thing that should be studied regarding the temperature changes in areas like Indonesia is the speed of which the temperature changes from year to year, as every area has different temperature changing speeds. A factor that can be analyzed for this is the location and coordinates of which the area is located to see if where the area is on the map affects the heating and cooling of that area and the speed of its changes. The result that is hoped to be gotten from this research is the estimated future temperature of each area in Indonesia and the data regarding which areas have the fastest heating and cooling speeds.

2. related work

<https://www.mckinsey.com/id/our-insights/how-to-power-indonesias-solar-pv-growth-opportunities>

<https://climateknowledgeportal.worldbank.org/country/indonesia/climate-data-historical>

<https://www.adb.org/sites/default/files/publication/700411/climate-risk-country-profile-indonesia.pdf>

By analyzing the website we researched we will know the estimations of the temperature in Indonesia from 2000 until 2023, and by knowing the temperature we will predict how it will change in the near future.

3. Dataset and Preprocessing

The following data is the data of average temperatures from the year 2000 to 2023 across 32 provinces of Indonesia taken from 32 BMKG stations from across Indonesia.

Data of each area from 2000 to 2010

Provinsi	latitude	longitude	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Aceh (Blang Bintang St)	5.5500	954.167	27.18	28.93	29.21	29.48	28.69	26.8	26.73	N/A	27	26.9	27.1
Bali (Ngurah Rai St)	-87.481	1.151.675	27.59	28.67	28.9	29.1	29.4	27	26.73	26.9	26.8	27.1	27.6
Banten (Serang St)	-61.200	1.061.503	27.35	27.88	28.48	29.85	29.32	27	26.98	26.7	26.6	27.2	27.1
Bengkulu (Pulau Baai St)	-309.042	1.023.069	24.23	26.8	28.51	29.51	29.71	25.7	26.1	26.4	26.3	26.5	26.8
DI Yogyakarta (Sleman St)	-76.817	1.103.233	28.2	29	28.48	29.87	29.42	N/A	N/A	25.8	26.1	26.1	N/A
DKI Jakarta (Tanjung Priok St)	-61.152	1.068.743	23.94	29.16	29	30.92	30.34	28.8	28.47	28.3	27.9	28.3	28
Gorontalo (Jalaludin St)	6.372	1.228.499	N/A	28.02	28.05	29.09	29.29	27	26.9	28.6	26.5	27.3	N/A
Jambi (Sungai Duren St)	-16.008	1.035.000	27.76	28.12	28.83	29.39	28.95	26.8	26.77	26.6	26.5	27.1	27.1
Jawa Barat (Husein St)	-69.006	1.075.764	N/A	27.26	28.33	28.73	28.47	N/A	N/A	N/A	N/A	N/A	N/A
Jawa Tengah (Semarang St)	-69.900	1.104.225	28.14	29.04	28.99	30.22	29.24	27.9	27.2	27.9	27.4	27.9	27.9
Jawa Timur (Juanda St)	-73.797	1.127.869	27.06	28.75	29.06	30.46	30.28	N/A	28.68	29.1	28	28.2	N/A
Kalimantan Barat (Supadio St)	-1.506	1.094.039	28.69	27.95	28.03	28.56	28.4	26.5	26.9	26.7	26.4	27.1	27.1
Kalimantan Selatan (Banjarbaru St)	-34.389	1.148.309	26.85	26.23	28.35	29.02	27.95	N/A	27.28	26.7	26.4	26.7	N/A
Kalimantan Tengah (Tjilik Riwat St)	-22.250	1.139.425	N/A	28.77	28.57	29	27.87	27	27.17	27.4	27.2	26.9	23.7
Kalimantan Timur (Temindung St)	-4.814	1.171.558	N/A	28.35	28.05	28.85	29.43	N/A	N/A	27.5	26.7	28.4	27.1
Kepulauan Bangka Belitung (Pangkal Pinang St)	-21.000	1.061.000	N/A	28.38	28.13	29.37	27.94	27.3	27.28	26.7	26.4	27.3	27
Kepulauan Riau (Batam St)	1.1300	1.040.531	N/A	N/A	N/A	28.88	28.88	N/A	N/A	N/A	N/A	N/A	N/A
Lampung (Radin Inten II/Branti St)	-52.425	1.051.789	28.2	26.72	27.5	29.67	29.66	N/A	26.18	27.6	26.4	26.7	26.7
Maluku (Pattimura St)	-37.103	1.280.891	25.72	28.48	28.46	28.14	29.13	N/A	26.18	27	26	26.6	26.9
Maluku Utara (Babullah St)	8.319	1.273.806	N/A	28.09	28.61	28.7	28.05	26.9	26.79	26.7	26.6	27.4	27.1

Nusa Tenggara Barat (Selaparang St)	-85.708	1.161.125	26.81	28.73	29.02	29.66	30.09	27.8	26.75	26.6	27.3	27.4	27.9
Nusa Tenggara Timur (Lasiana St)	-101.417	1.236.689	27.56	27.37	28.28	29.13	28.93	27.2	N/A	27.3	27.2	27.3	N/A
Papua (Jayapura St)	-30.000	1.399.500	N/A	N/A	N/A	28.49	28.14	N/A	27.07	27.2	27	27	27
Papua Barat (Manokwari St)	-8.667	1.340.833	N/A	N/A	N/A	28.25	28.1	27.9	27.5	27.4	27.3	27.1	27.3
Riau (Sultan Syarif Qasim St)	4.608	1.014.444	N/A	29.05	28.21	29.64	29.85	N/A	27.7	27.1	27.4	27.7	27.7
Sulawesi Barat (Majene St)	-13.181	1.193.751	N/A	N/A	N/A	29.37	29.46	N/A	N/A	27.2	27.1	27.5	27.6
Sulawesi Tengah (Mutitara St)	-9.186	1.199.097	28.8	28.85	29.6	29.62	28.49	N/A	26.58	25.2	26.6	27.6	27.7
Sulawesi Tenggara (Wolter Monginsidi St)	-40.816	1.224.182	N/A	27.21	26.43	28.39	28.36	N/A	25.7	26.7	26.5	27.7	28
Sulawesi Utara (Kayuwatu St)	1.1161	1.249.588	26.97	28.8	28.76	28.67	26.9	26.7	26.24	26.3	26	26.6	26.3
Sumatera Barat (Sicincin St)	-5.750	1.002.870	25.87	28.18	28.2	29.17	29.02	25.9	26.05	25.7	25.4	25.5	25.8
Sumatera Selatan (Kenten St)	-29.275	1.047.719	N/A	28.83	28.24	29.64	28.73	27.1	27.1	27	26.9	27.4	N/A
Sumatera Utara (Polonia St)	3.5993	986.975	27.7	28.04	28.45	29.59	29.27	27	27.8	27	26.9	27.1	N/A

Data of each area from 2011-2023

Provinsi	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Aceh (Blang Bintang St)	27.1	26.9	27	27.1	27.1	28.23	28.23	27.66	27.97	N/A	27	27.2	27
Bali (Ngurah Rai St)	26.8	26.9	27.4	27.4	27.3	N/A	N/A	26.73	N/A	N/A	N/A	N/A	28.17
Banten (Serang St)	27	27.1	27	27.3	27.3	27.56	27.68	27.92	30.04	28	29.28	29.14	30.18
Bengkulu (Pulau Baai St)	N/A	26.9	26.74	26.9	27	N/A	N/A	28.8	27.35	28.45	27.5	26.7	27.2
DI Yogyakarta (Sleman St)	26	26.6	26.2	26.3	26.1	27.68	N/A	N/A	N/A	N/A	27.2	27.5	27.3
DKI Jakarta (Tanjung Priok St)	N/A	N/A	28.96	29.44	29.27	29.05	29.23	29.28	29.37	29.74	29.15	N/A	N/A
Gorontalo (Jalaludin St)	27	27	27	26.7	27.3	28.64	28.27	27.88	27.98	28.06	28.33	27.1	27.7
Jambi (Sungai Duren St)	26.9	26.7	26.75	27.2	27	28.15	27.83	27.95	N/A	N/A	26.9	27	27
Jawa Barat (Husein St)	N/A	N/A	N/A	N/A	N/A	N/A	25.19	24.93	25.18	N/A	28.1	27.9	28
Jawa Tengah (Semarang St)	27.7	28	28.02	28.02	28.5	N/A	30.23	28.57	29.17	28.35	28.5	27.9	28.5
Jawa Timur (Juanda St)	N/A	28	27.9	28	28	N/A	N/A	N/A	N/A	N/A	28.8	28.1	N/A
Kalimantan Barat (Supadio St)	26.6	27.1	26.9	26.8	26.9	29	29.3	28.75	29.2	29.1	27.6	26.7	27.4
Kalimantan Selatan (Banjarbaru St)	27.1	26.6	26.7	26.8	27	N/A	26.63	26.75	27.25	28.64	N/A	N/A	N/A
Kalimantan Tengah (Tjilik Riwt St)	N/A	27.3	27.4	27.4	27.7	N/A	N/A	N/A	N/A	N/A	27.6	27.2	27.6
Kalimantan Timur (Temindung St)	27.3	28	27.43	27.7	27.9	N/A	27.65	29.6	27.83	27.87	27.7	27.2	27.7
Kepulauan Bangka Belitung (Pangkal Pinang St)	27.4	27	27	27.2	27.3	N/A	N/A	N/A	27.81	27.92	27.86	27.77	27.5
Kepulauan Riau (Batam St)	N/A	N/A	N/A	N/A	N/A	27.4	27.95	30.46	28.12	28.27	28.07	27.81	27.75
Lampung (Radin Inten II/Branti St)	26.8	26.8	26.7	25.79	27.1	27.18	26.91	26.81	27.2	27.16	27.13	34.35	29.55
Maluku (Pattimura St)	26.6	26.4	26.5	26.6	26.5	26.7	N/A	N/A	N/A	27.08	27.9	28	28
Maluku Utara (Babullah St)	26.9	N/A	27	27	27.3	N/A	N/A	N/A	N/A	N/A	28.4	28	28.3
Nusa Tenggara Barat (Selaparang St)	26.6	25.8	28.25	26.9	26.1	27.2	26.35	27	27.55	N/A	27.4	27.4	27.4
Nusa Tenggara Timur (Lasiana St)	27	27.2	27.5	27.4	27.5	N/A	N/A	27.2	27.35	27.95	28.3	28	28
Papua (Jayapura St)	27.1	27	27.9	28.1	27.8	27.5	27.1	27	26.8	N/A	N/A	N/A	N/A
Papua Barat (Manokwari St)	27.1	27.2	27.3	27.5	27.4	27.89	27.98	N/A	N/A	N/A	26.1	25.9	26.1
Riau (Sultan Syarif Qasim St)	27	27.3	26	27.2	27.2	27.4	27.51	27.34	27.47	27.3	27.1	26.8	27

Sulawesi Barat (Majene St)	27.7	27.6	27.58	27.9	27.9	27.9	27.8	27.6	27.6	27.55	27.7	27.4	28
Sulawesi Tengah (Mutia St)	27.6	27.7	26.7	26.7	28.4	28.23	27.5	27.98	28.2	27.66	28	27	28
Sulawesi Tenggara (Wolter Monginsidi St)	27.5	27.3	27	26.8	26.9	26.9	26.9	26.9	26.9	27	27	27	27.55
Sulawesi Utara (Kayuwatu St)	26.1	26.1	26.37	26.6	27	26.66	N/A	N/A	N/A	N/A	28.8	27.9	28.3
Sumatera Barat (Sicincin St)	27	25.2	25.13	25.67	26.5	N/A	N/A	27	27	27	26.9	26.5	26.7
Sumatera Selatan (Kerten St)	27.3	27.4	27.3	24.2	27.7	27	27	27	27	27	27	27	27
Sumatera Utara (Polonia St)	27.2	27.3	28.77	27.9	27.4	27	27	27	27	27	28.4	27.9	28.2

Preprocessing the data requires splitting the dataset into 3 columns of location, year and average temperature. Before processing the data into a linear regression algorithm, outliers are first calculated then replaced with NaN values. After the outliers are replaced with NaN values, the program then erases all rows which contain a NaN value

4. Model and Techniques

Model Choice: Ridge Regression

The code uses Ridge Regression as the machine learning model. Ridge Regression is a type of linear regression that adds a penalty term to the loss function, which helps to prevent overfitting and improve the model's generalization to new data. It's a good choice for this problem because:

- **Simplicity and Interpretability:** Ridge Regression is relatively simple to understand and interpret. It provides coefficients for each predictor variable, indicating their influence on the target variable
- **Regularization:** The penalty term in Ridge Regression helps to prevent overfitting, which can be a concern when working with time series data or when there are many predictor variables.
- **Efficiency:** Ridge Regression is computationally efficient, making it suitable for backtesting over multiple iterations.

Why Ridge Regression over Other Techniques:

- **Linear Relationship:** The code assumes a linear relationship between the predictor variables and the target variable. Ridge Regression is well-suited for such relationships.
- **Time Series Data:** While more complex time series models (e.g., ARIMA) exist, Ridge Regression provides a reasonable starting point and is less prone to overfitting with limited data.

How Ridge Regression Works:

Ridge Regression finds the best-fitting line (or hyperplane in higher dimensions) by minimizing the sum of squared errors between the predicted and actual values. However, it adds a penalty term (controlled by the alpha parameter) to the loss function that discourages large coefficients. This penalty helps to prevent the model from becoming overly complex and sensitive to noise in the training data.

Tools Used and Justification:

1. Pandas (pd): Used for data manipulation, cleaning, and loading (via `read_csv`). Pandas is essential for working with tabular data in Python and provides a wide range of functions for data wrangling and analysis.
2. Scikit-learn (sklearn): Used for implementing the Ridge Regression model (`Ridge`) and for calculating performance metrics (`mean_absolute_error`, `mean_squared_error`). Scikit-learn is a comprehensive library for machine learning in Python.
3. IPython and Jupyter: Used for interactive development and code execution. Jupyter notebooks provide a convenient environment for data exploration, visualization, and sharing code.

Employment of Tools and Features:

- Pandas: Employed extensively for loading the weather data, handling missing values, creating new features (rolling averages, expanding means), and selecting predictors. Features like `groupby`, `rolling`, and `expanding` were particularly useful for time series analysis.
- Scikit-learn: Used to train the Ridge Regression model (`fit`) and make predictions (`predict`). The alpha parameter was set to 0.1, indicating a moderate level of regularization.
- Jupyter: The code is written and executed within a Jupyter notebook, enabling interactive exploration and visualization of the results.

Potential Improvements:

Feature Engineering: Explore additional features that might improve the model's accuracy, such as lagged temperatures, weather and season patterns, or external factors.

Model Tuning: Experiment with different values of the alpha parameter in Ridge Regression to find the optimal level of regularization.

Cross-Validation: Implement a more robust cross-validation strategy within the backtesting process to better estimate the model's generalization performance.

Other Models: Consider exploring other time series models like ARIMA or Prophet for potentially better predictions.

Error Analysis: Conduct a more in-depth analysis of the prediction errors to identify patterns and areas where the model struggles. This might suggest areas for improvement in feature engineering or model selection.

Data Quality: Investigate the quality of the weather data and address any potential issues like outliers or inconsistencies.

5. Evaluation Method

Evaluation Technique: Backtesting

The code employs a technique called backtesting to evaluate the performance of the machine learning model (Ridge Regression in this case).

How it works in this project:

1. **Iterative Training and Testing:** The backtest function simulates how the model would perform over time. It iteratively splits the data into training and testing sets, moving the split point forward with each iteration (controlled by the start and step parameters).
2. **Model Training and Prediction:** In each iteration, the model is trained on the training data and then used to make predictions on the testing data.
3. **Performance Evaluation:** The predictions are compared to the actual values in the testing data using metrics like mean absolute error (MAE).
4. **Aggregation:** The results from each iteration are combined to provide an overall performance assessment.

Data Splitting:

The dataset is not split into traditional train/validation/test sets like in many machine learning scenarios. Instead, backtesting dynamically creates these splits within the backtest function.

1. **Training Data:** In each iteration, the train variable contains all data points before the current split point (i).
2. **Testing Data:** The test variable contains a subset of data points after the current split point (i), determined by the step parameter.

Performance Measures:

The code uses the following performance measures:

1. **Mean Absolute Error (MAE):** This metric calculates the average absolute difference between the predicted values and the actual values. It provides a measure of the model's overall prediction accuracy.
2. **Mean Squared Error (MSE):** This metric calculates the average squared difference between the predicted values and the actual values. It gives more weight to larger errors.
3. **Difference Distribution:** The code also analyzes the distribution of prediction errors ("diff") to

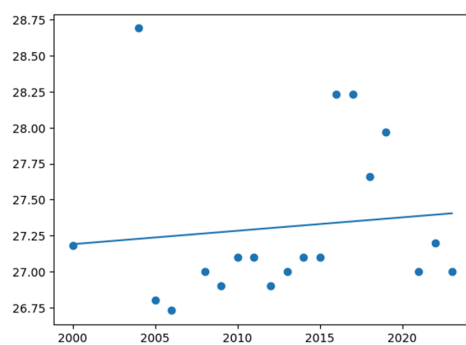
understand how frequently the model makes predictions within a certain range of accuracy.

In summary, the code uses backtesting with a rolling window approach to evaluate a Ridge Regression model for predicting the average temperature. MAE and MSE are the primary performance measures used, along with an analysis of the distribution of prediction errors. This allows for a comprehensive evaluation of the model's performance over time.

6. Results and Discussion

Below are the results of the preprocessed data being entered through the program which will show the graph made from linear regression, the slope of the line which represents how fast the area's heating or cooling changes are, and the predicted temperature of that area in 2030.

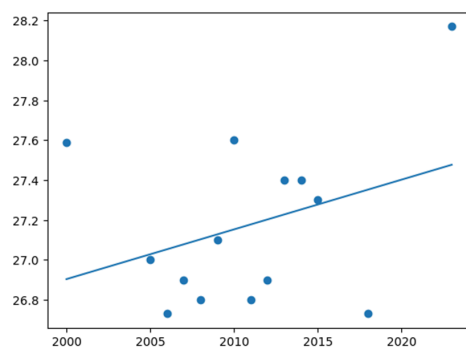
•Aceh



Slope: 0.009307477288609336

predicted heat in 2030: 27.47018658280922

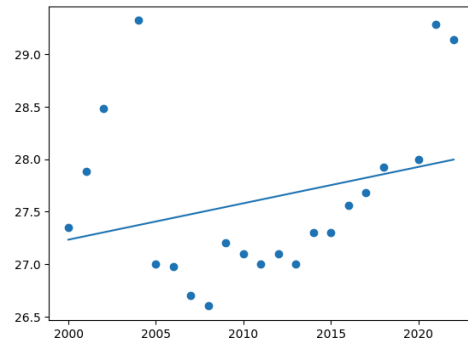
•Bali



slope: 0.024930110179246876

predicted heat in 2030: 27.65187140272982

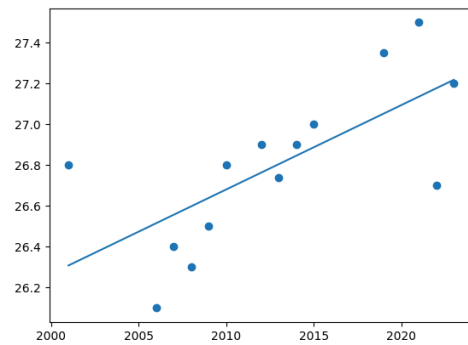
•Banten



slope: 0.03467194570135748

predicted heat in 2030: 28.272576492135315

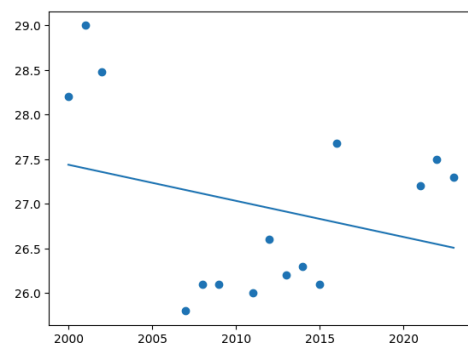
•Bengkulu



slope: 0.04136111111111107

predicted heat in 2030: 27.508333333333326

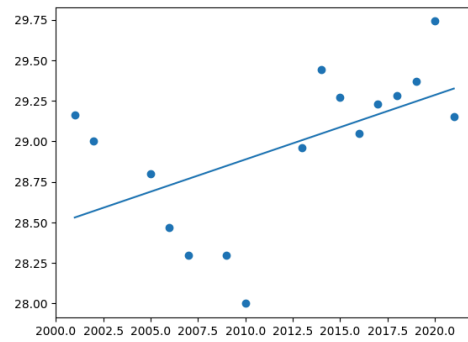
•DI Yogyakarta



slope: -0.040445278969957074

predicted heat in 2030: 26.226473533619455

•DKI Jakarta



slope: 0.03973367782967305

predicted heat in 2030: 29.68272284606975

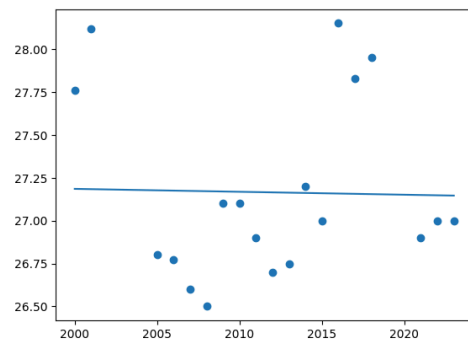
•Gorontalo



slope: 0.01860050890585239

predicted heat in 2030: 27.883638676844786

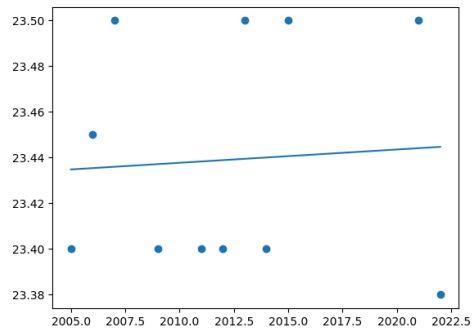
•Jambi



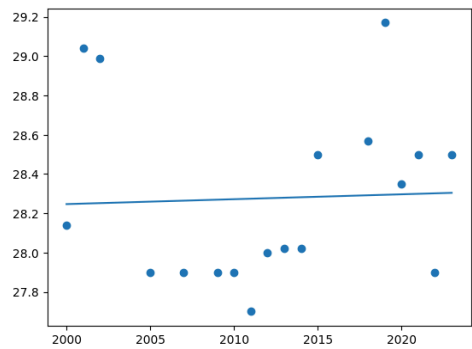
slope: -0.0017042606516291697

predicted heat in 2030: 27.134060150375937

•Jawa Barat (West Java)

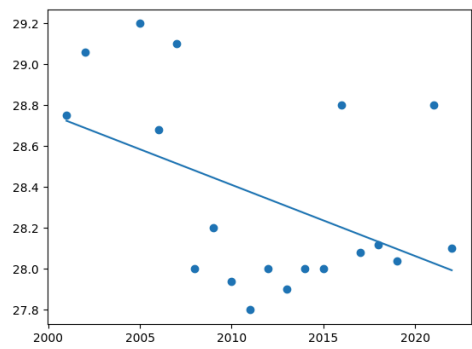


slope: 0.0005815972222222525
 predicted heat in 2030: 23.449401041666665
 •Jawa Tengah (Central Java)



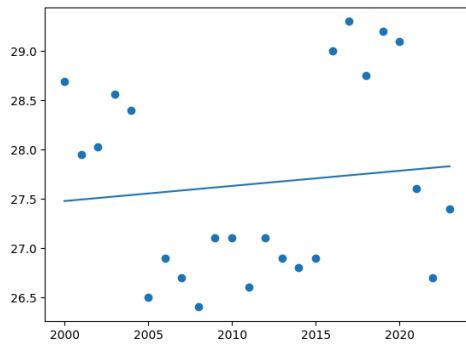
slope: 0.002492721979621605
 predicted heat in 2030: 28.321815866084425

•Jawa Timur (East Java)



slope: -0.03478848232206267
 predicted heat in 2030: 27.715934846379497

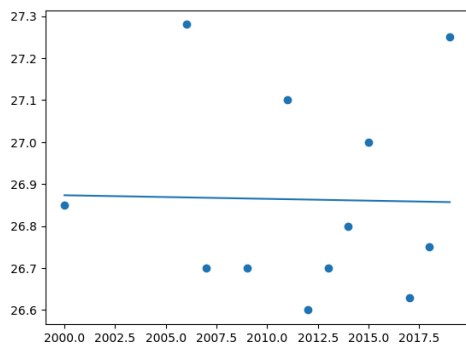
•Kalimantan Barat (West Kalimantan)



slope: 0.01536521739130436

predicted heat in 2030: 27.937589855072467

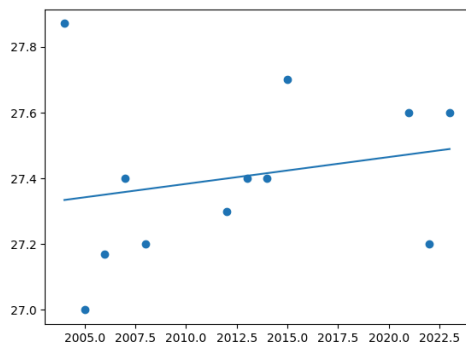
•Kalimantan Selatan (South Kalimantan)



slope: -0.0008573540280858037

predicted heat in 2030: 26.847686622320765

•Kalimantan Tengah (Central Kalimantan)



slope: 0.008151093439363832

predicted heat in 2030: 27.545977468522203

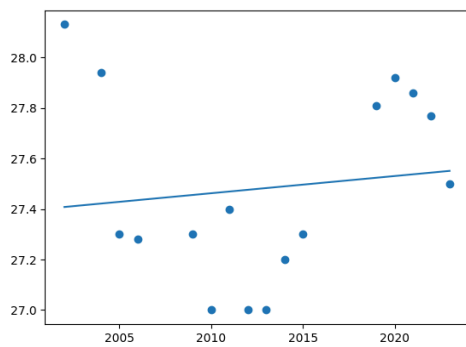
•Kalimantan Timur (East Kalimantan)



slope: -0.023870062161620277

predicted heat in 2030: 27.38371365550431

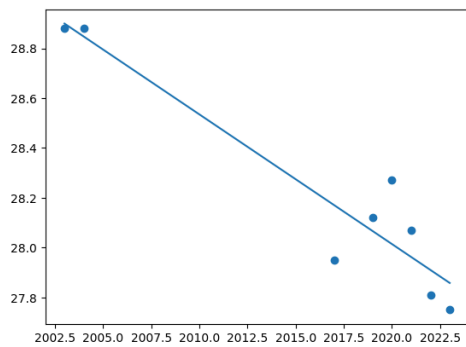
•Kepulauan Bangka Belitung



slope: 0.006802133137182761

predicted heat in 2030: 27.598361529974255

•Kepulauan Riau



slope: -0.052055137844611486

predicted heat in 2030: 27.493984962406017

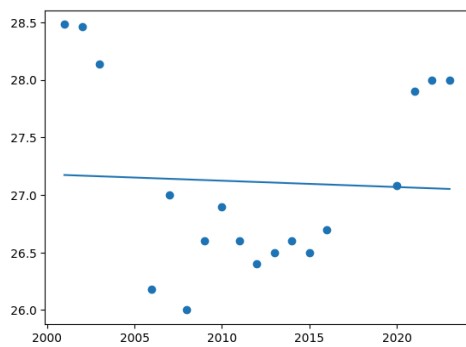
•Lampung



slope: 0.004365339049983073

predicted heat in 2030: 27.03979126706533

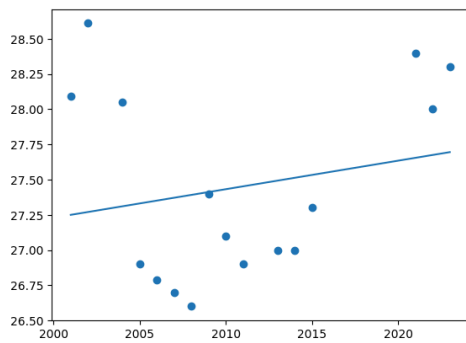
•Maluku



slope: -0.005504121750158594

predicted heat in 2030: 27.01334178820545

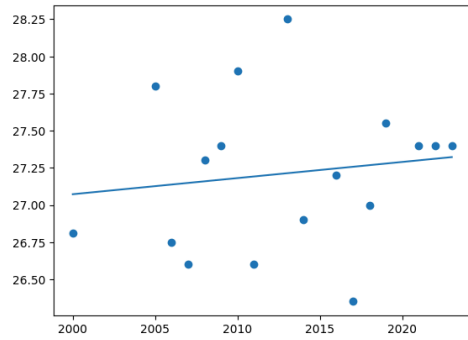
•Maluku Utara



slope: 0.020241787122207644

predicted heat in 2030: 27.837169513797633

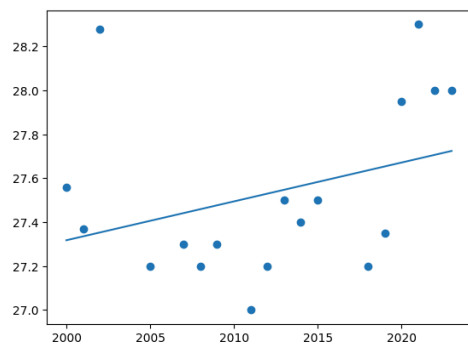
•Nusa Tenggara Barat (West Nusa Tenggara)



slope: 0.01086258776328983

predicted heat in 2030: 27.398294884653964

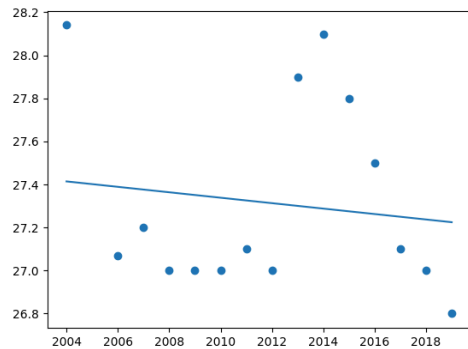
•Nusa Tenggara Timur (East Nusa Tenggara)



slope: 0.0176775890935183

predicted heat in 2030: 27.84815713944033

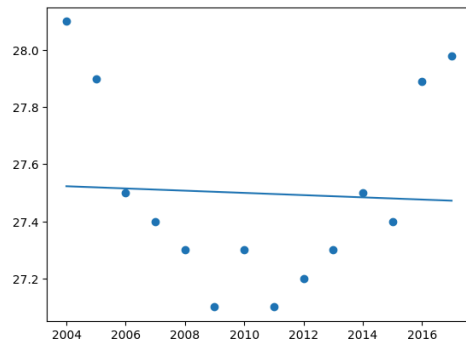
•Papua



slope: -0.01263336347197103

predicted heat in 2030: 27.08575723327306

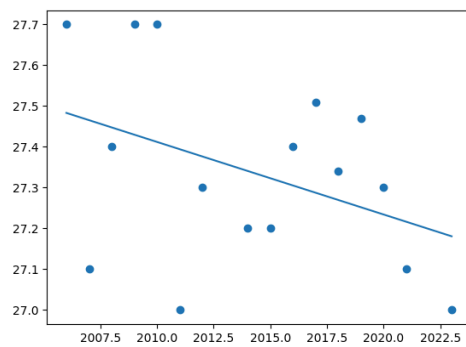
•Papua Barat (West Papua)



slope: -0.003890109890109889

predicted heat in 2030: 27.421999999999997

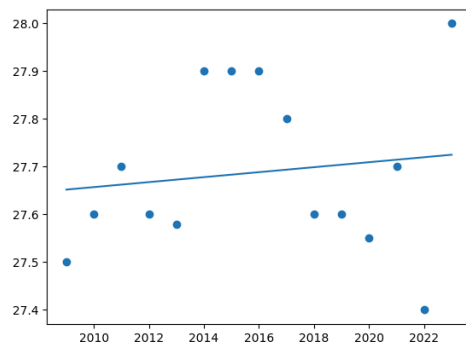
•Riau



slope: -0.017787610619468985

predicted heat in 2030: 27.056371681415925

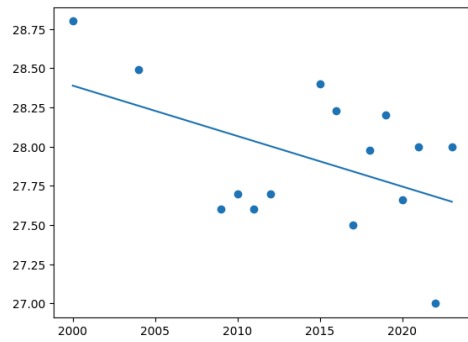
•Sulawesi Barat (West Sulawesi)



slope: 0.005214285714285704

predicted heat in 2030: 27.761666666666667

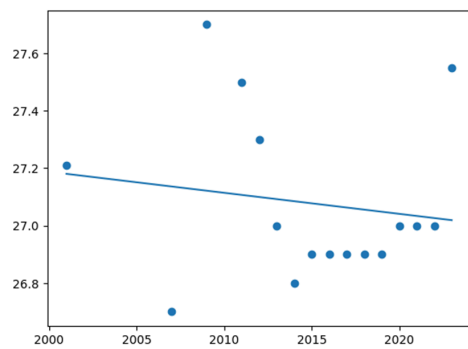
•Sulawesi Tengah (Central Sulawesi)



slope: -0.03217813423385395

predicted heat in 2030: 27.424166314900802

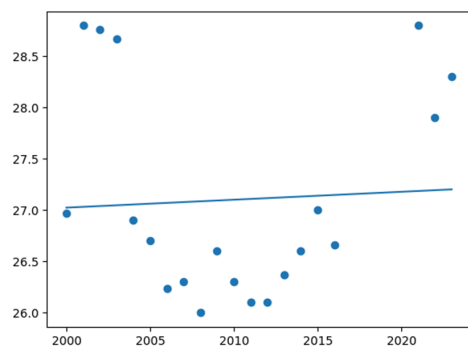
•Sulawesi Tenggara (South East Sulawesi)



slope: -0.0073289287399717126

predicted heat in 2030: 26.967899952807926

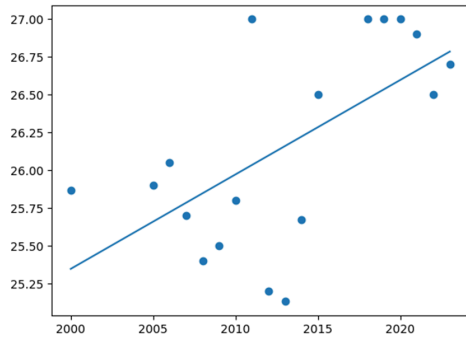
•Sulawesi Utara (North Sulawesi)



slope: 0.007730270389096528

predicted heat in 2030: 27.257332380743016

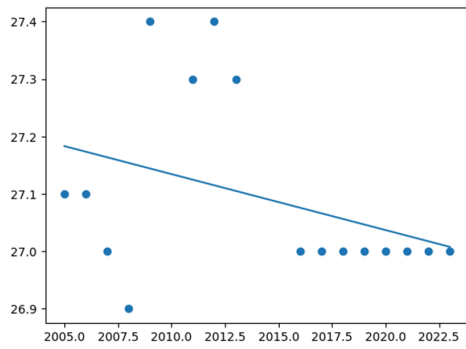
•Sumatera Barat (West Sumatera)



slope: 0.06257864018797844

predicted heat in 2030: 27.223980140983855

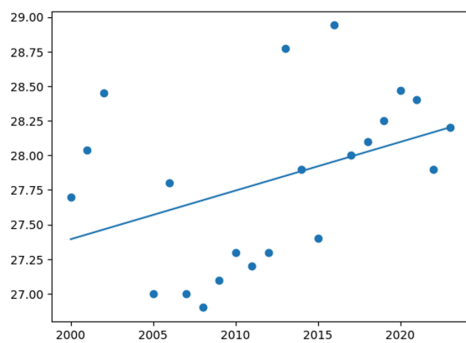
•Sumatera Selatan (South Sumatera)



slope: -0.009740920918655973

predicted heat in 2030: 26.93972168797375

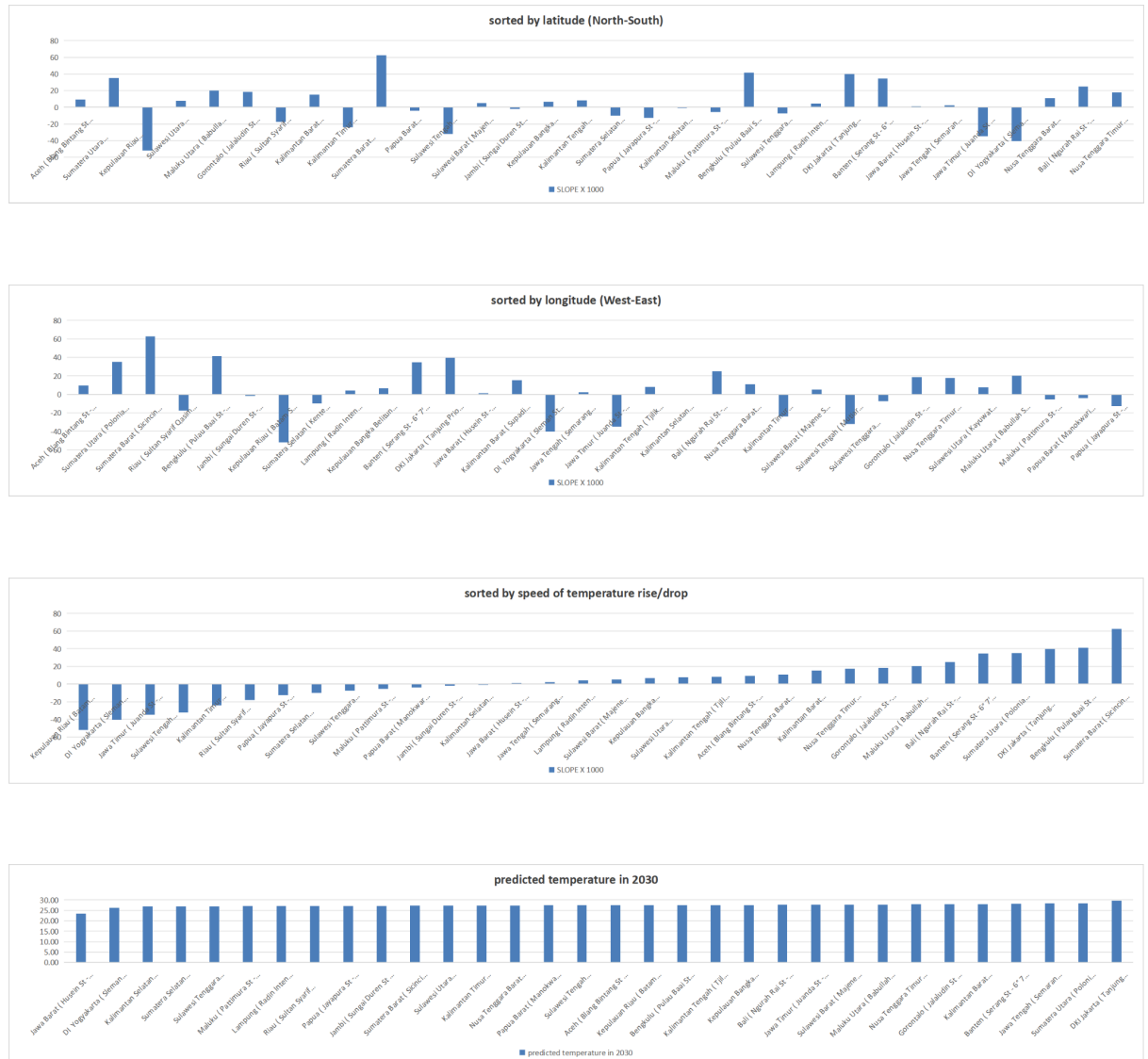
•Sumatera Utara (North Sumatera)



slope: 0.035135256785344564

predicted heat in 2030: 28.44808570013953

Below are the graphs made from the data of the slope from every area sorted by latitude from North to South, then sorted by longitude from West to East, then sorted by speed of temperature changes, and finally a separate graph of predicted temperatures in each area by 2030 sorted from lowest to highest temperature.



From the above three graphs we can conclude a few things. First is that based on the longitude sorted graph, areas on the west side of Indonesia have more extreme temperature changes than areas on the east side of Indonesia. The second is that location of the area doesn't individually affect the temperature change speed of every area. Third is that West Java is predicted to have the lowest temperature at 23.45 Celsius while Jakarta is predicted to have the highest temperature at 29.69 Celsius in 2030.

7. Conclusion and Recommendation:

The dataset results tell us that the speed of which temperature changes in each area is unique and the predicted temperatures in 2030 will still mostly be in the 20-30 degree Celsius range. The results also tell us that the speed of the temperature changes in each area is not affected by latitude at all and only slightly affected by longitude as areas in the Western part of Indonesia have more extreme and fast temperature changes than the Eastern part. Another result regarding the predicted temperatures in 2030 say that Jakarta will be the highest temperature area in Indonesia at that time and West Java will be the coldest, though the acceleration of temperature is fastest in West Sumatera meaning that given enough time, West Sumatera will be the highest temperature area in Indonesia.

Sources:

Badan Pusat Statistik. (2015). *Suhu Minimum, Rata-Rata, dan Maksimum di Stasiun Pengamatan BMKG (oC), 2000-2010*.

<https://www.bps.go.id/id/statistics-table/1/MTM0NyMx/suhu-minimum--rata-rata--dan-maksimum-di-stasiun-pengamatan-bmkg--oc---2000-2010.html>

Badan Pusat Statistik. (2017). *Suhu Minimum, Rata-Rata, dan Maksimum di Stasiun Pengamatan BMKG (oC), 2011-2015*.

<https://www.bps.go.id/id/statistics-table/1/MTk2MSMx/suhu-minimum--rata-rata--dan-maksimum-di-stasiun-pengamatan-bmkg--oc---2011-2015.html>

Link to source code : <https://github.com/EdbertTan/fundamental-of-data-science-final-project/>

Group member contact:

-Edbert: edbert.tan001@binus.ac.id

-Daffa: daffa.wiryawan@binus.ac.id