JOWIM
**International Journal of Wireless and Multimedia Communications**
Vol. 1 No. 2 (2024), July 2024, pp. 1-6
ISSN: 3047-2172

# Simple Regression Model Analysis of the Effect of Temperature on Rainfall in Padang City Using Scikit-Learn

Nelvidawati[a]

[a] *Environmental Engineering, Institut Teknologi Padang, Padang, 25143, Indonesia*

[a*] *nelvidaus11@gmail.com*

***Abstract*— Temperature is one of the factors that influences the intensity of rainfall in an area. The higher the temperature in an area, the higher the evaporation so the greater the chance of rain, especially if this occurs at sea. The location of Padang City close to the sea influences the intensity of rainfall that occurs. Research was conducted to model the effect of temperature on rainfall that occurs in Padang City using simple linear regression using the Python programming language. The methodology used to model the relationship between these two variables is the Cross-Industry Standard Process for Data Mining. The simple regression model equation obtained is y = -26.73 x + 1119.98. The evaluation carried out on the model resulted in a relationship that was not close and had a negative correlation. The error rate of the simple linear regression model used in this study is quite large with a mean model error value of 40%, an MAE value of 124.2311, MSE 23489.97 and RMSE 153.2644. A good model has MAE, MSE and RMSE close to zero. Further research is needed to find out a suitable model to describe the relationship between temperature and rainfall that occurs in Padang City.**

***Keywords*— Temperature; Rainfall; Model; Simple Regression; Prediction.**

*Manuscript received 15 Mei. 2024; revised 29 Mei. 2024; accepted 18 June. 2024. Date of publication 1 July. 2024.*


## I. INTRODUCTION

The increase in the amount of greenhouse gases in the atmosphere causes the temperature on the earth's surface to increase [1]. Temperature affects rainfall directly and indirectly [2]. In arid areas, high temperatures cannot result in evaporation so that rainfall in these areas is low [3]. High temperatures in areas close to seas and lakes cause more evaporation to occur, resulting in higher rainfall [4].

West Sumatra has an average temperature increase of 0.007 $^{o}$C – 0.01 $^{o}$C/year with daily maximum temperatures increasing around 0.058 $^{o}$C - 0.066 $^{o}$C/year and daily minimum air temperatures increasing around 0.028 $^{o}$C - 0.045 $^{o}$C/year. Temperature changes in West Sumatra are influenced by Monsoon and Elnino [5].

Temperature changes that occur in an area result in climate change [6]. Changes in rain patterns make seasons unpredictable which can cause disasters (floods or droughts) [6], [7]. Padang City is one of the cities in West Sumatra that has high vulnerability due to the impacts of climate variability and change. Padang City's low adaptive capacity and sensitivity to climate change can result in flooding [8], [9] and sea level rise [10]. Climate change disaster mitigation can be done through good environmental management strategies based on climate change data and analysis [11]. Historical climate studies and projections are needed, both short and long term. Climate change studies can analyze climate factors that are indicators of climate change such as temperature, rainfall and sea level rise [12].

Indication of climate change through time series analysis involves various statistical methods and data analysis techniques to understand trends and patterns of climate change [13]. Approaches that can be used are trend analysis, seasonal analysis, cycle analysis, anomaly analysis, modeling and forecasting and climate modeling. The aim of this research is to model the relationship between temperature and rainfall in Padang City using linear regression. Apart from that, model evaluation and predictions of rainfall were also carried out.

## II. MATERIAL AND METHODS

### A. *Studi Area*

Research location in Padang City which has an area of 1,414.96 km$^2$ with a population of 919,145 people (2022). Padang City is a medium-sized city and functions as the capital

of West Sumatra Province which has more facilities than other cities or districts in West Sumatra Province, especially

education and health facilities. Based on previous research using Spearman correlation, it was stated that there was no significant relationship between temperature and rainfall in Padang City [14]. The map of Padang city can be seen in Figure 1.

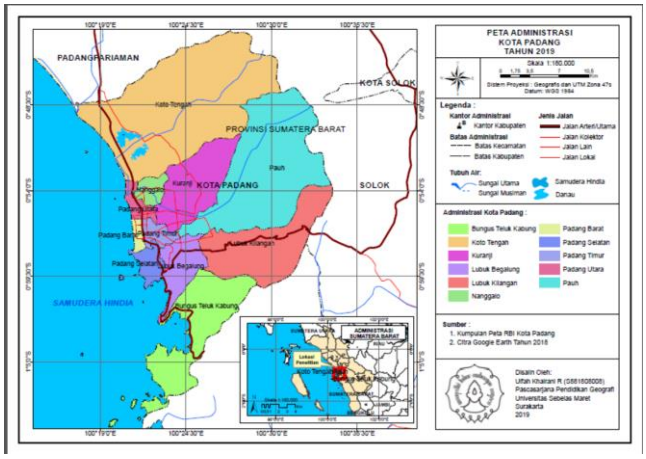


Fig 1. Map of Padang city

B. *Data And Method*

The data used in this research are monthly rainfall and average monthly temperature data obtained from climatecharts.net. The data mining methodology used is CRISP-DM (Cross-Industry Standard Process for Data Mining). This methodology is designed to provide a systematic and organized structure for dealing with complex data mining projects [15][16]. The stages of the CRISP-DM methodology can be seen in Figure 2.

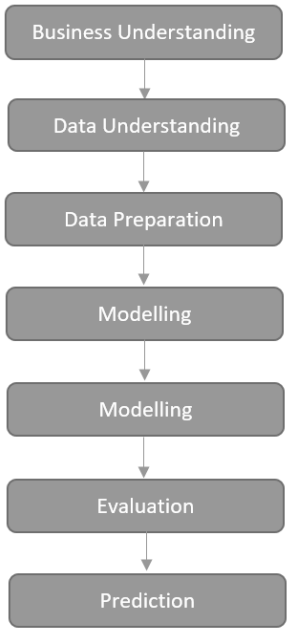


Fig 2. The stages of the CRISP-DM

1) Business understanding, namely identifying research objectives.
2) Data understanding, namely collecting data, identifying relevant data and understanding the characteristics and quality of the data.
3) Data Preparation, namely cleaning data, integrating data and transforming data into a format that is more suitable for analysis.
4) Modeling, namely carrying out modeling using linear regression to determine a model of the relationship between temperature and rainfall that occurs in Padang City based on data from 1991-2020.
   Simple linear regression is used to forecast and predict quality and quantity [17][18].
   The equation used is:

$$Y = a + bX \quad (1)$$

   Constants (a) and regression coefficients (b) can be calculated with the following equation:

$$a = \frac{(\sum Y)(\sum X^2)-(\sum X)(\sum XY)}{n\sum X^2-(\sum X)^2} \quad (2)$$

$$b = \frac{n(\sum XY)-(\sum X)(\sum Y)}{n(\sum X^2)-(\sum X)^2} \quad (3)$$

5) Evaluation, namely assessing the quality and performance of the model that has been developed by measuring model performance through measuring accuracy, precision, recall and relevant performance metrics. For regression, model evaluation is carried out by calculating MAE, MSE and RMSE [19][20].
   Mean Absolute Error (MAE), namely the average absolute value of the difference between the prediction and the actual value. The equation used is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Yi - \hat{Y}\imath\right| \quad (4)$$

   Mean Squared Error (MSE), which is the average of the squares of the difference between the prediction and the actual value. The equation used is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Yi - \hat{Y}\imath\right)^2 \quad (5)$$

   Root Mean Squared Error (RMSE), namely the square root of MSE which provides a more intuitive interpretation. The equation used is:

$$RMSE = \sqrt{MSE} \quad (6)$$

6) Application of results to predict rainfall.

## III. RESULT

Linear Regression Analysis of the Relationship Between Temperature and Rainfall in Padang City using Scikit-Learn.

a. Data information

The data used is monthly rainfall data and average temperature in one month with a total of 360 data and is a float data type (data with commas) which can be seen in Figures 3 and 4. The data is the result of satellite recordings which can be accessed on the climate chart.

| | Temperature | Rainfall |
|---|---|---|
| 0 | 26.8 | 401.9 |
| 1 | 26.9 | 208.9 |
| 2 | 27.2 | 557.7 |
| 3 | 27.2 | 407.0 |
| 4 | 27.1 | 311.1 |
| ... | ... | ... |
| 355 | 27.6 | 261.0 |
| 356 | 27.2 | 479.3 |
| 357 | 27.7 | 516.8 |
| 358 | 27.2 | 626.0 |
| 359 | 27.1 | 359.7 |

360 rows × 2 columns

Fig 3. Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 360 entries, 0 to 359
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Temperature  360 non-null    float64
 1   Rainfall     360 non-null    float64
dtypes: float64(2)
memory usage: 5.8 KB
```

Fig 4. Type of Data

b. Descriptive Statistics

Descriptive statistical data can be seen in Figure 5. Temperature data quartiles (Q1= 26.9 ºC, Q2=27.2 ºC and Q3=27.5 ºC) and Rainfall data quartiles (Q1=271.3 mm, Q2=371.3 mm and Q3=492.72 mm).

| | Temperature | Rainfall |
|---|---|---|
| count | 360.000000 | 360.000000 |
| mean | 27.193889 | 389.324167 |
| std | 0.425073 | 154.104684 |
| min | 26.300000 | 65.300000 |
| 25% | 26.900000 | 271.300000 |
| 50% | 27.200000 | 371.300000 |
| 75% | 27.500000 | 492.725000 |
| max | 28.700000 | 873.900000 |

Fig 5. Descriptive Statistic Data

c. Missing Value

There were no missing values found in the data as seen in Figure 6.

```
# Information Missing Value
df.isnull().sum()
```

```
Temperature    0
Rainfall       0
dtype: int64
```

Fig 6. Missing Value Information

d. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach to data analysis that aims to analyze datasets and summarize their characteristics using visual and descriptive methods. The main goal is to understand the structure of the data, identify interesting patterns, find anomalies, and test initial hypotheses that can be obtained from the data. EDA is important in the data analysis process because it can provide deep initial insight into the data before applying more complex statistical models or machine learning techniques. By performing EDA carefully, data analysts can make better decisions about next analysis steps and understand the limitations and potential of the dataset being studied.

The type of EDA used is graphical univariate. Graphical univariate refers to a type of data visualization that focuses on a single variable in its analysis. The purpose of graphical univariates is to understand the distribution, patterns and characteristics of these variables separately. The type of univariate graphic used is the Box Plot which can present statistical summaries of numerical variables such as quartiles, medians and outliers. Box plots also help in detecting the presence of outliers and seeing the distribution of data visually. From Figure 7 it can be seen that the mean and median temperature is at 27 ºC which is normally distributed. The high standard deviation is due to the long whiskers boxplot as seen in Figure 8. The temperature data outlier amounts to one data.

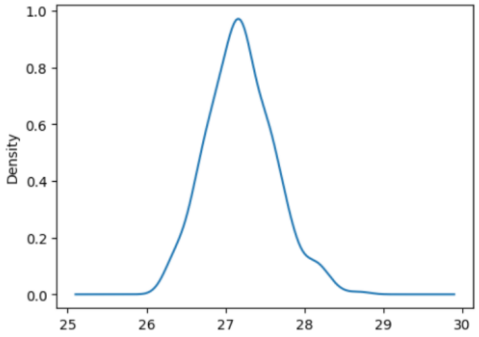


Fig 7. Temperature Plot Distribution

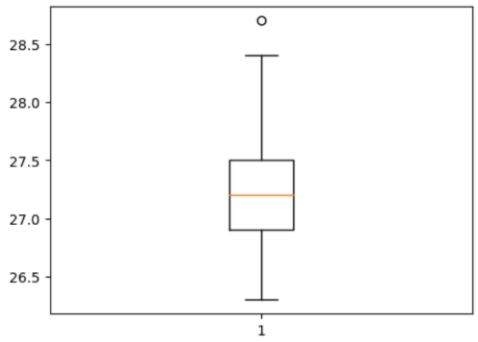


Figure 8. Temperature Boxplot

The mean and median of the rainfall data are around 400mm with the distribution being more to the right skew and the data spread is less even and has a high standard deviation as seen in Figure 9. There are 2 outliers in the rainfall data as seen in Figure 10.

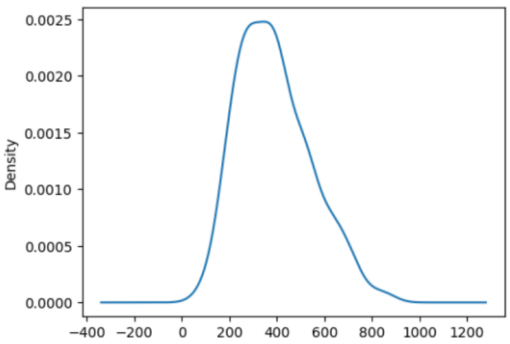


Fig 9. Rainfall Plot Distribution

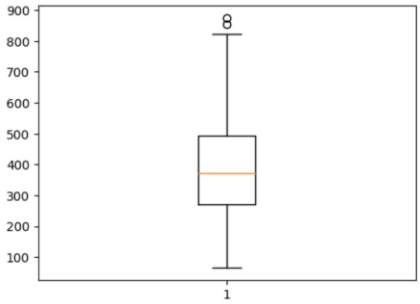


Figure 10. Rainfall Boxplot

Bivariate analysis is an approach to data analysis that examines the relationship between two variables simultaneously. The main goal of bivariate analysis is to explore how two variables are related to each other and how changes in one variable can affect the other variable. The bivariate analysis used is the Bivariate Line Plot or Area Plot. This technique is used to show trends or patterns of change in two numerical variables over time or other variables. It helps to see how changes in one variable affect other variables. Based on Figure 11, it can be seen that the data is negatively correlated. Rainfall increases when the average monthly temperature in Padang City decreases.

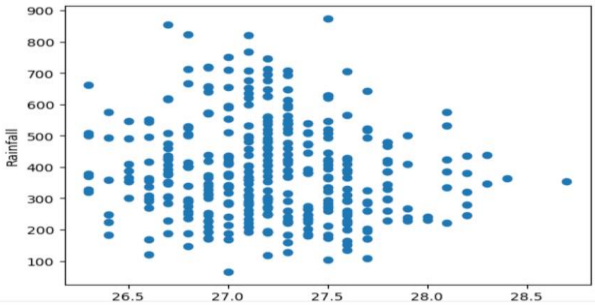


Fig 11. Bivariate Line Plot Temperature And Rainfall

e. Modeling

In Figure 12 it can be seen that the correlation between temperature and rainfall in Padang City is 0.096 and this shows that the relationship between temperature and rainfall is not strong.

```
[ ] from sklearn.model_selection import train_test_split
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

[ ] from sklearn.linear_model import LinearRegression
    regression = LinearRegression ()
    regression.fit (x_train, y_train)

    ▾ LinearRegression
    LinearRegression()

# Correlation
df.corr()
```

| | Temperature | Rainfall |
|---|---|---|
| **Temperature** | 1.000000 | -0.095991 |
| **Rainfall** | -0.095991 | 1.000000 |

Fig 12. Correlation Temperature and Rainfal

The regression model equation obtained is:

$$y = -26.73\ x + 1119.98$$

with slope (m) is -26.73 and intercept (b) is 1119.98. This model gets an accuracy score of 0.57%.

Regression line is a straight line used in regression analysis to show the statistical relationship between two variables. This line is obtained by techniques such as simple linear regression or multiple linear regression, which try to find patterns or relationships between the dependent variable (y) and the independent variable (x).

g. Data visualization using data testing

Regression Line visualization uses testing data as seen in Figure 13.

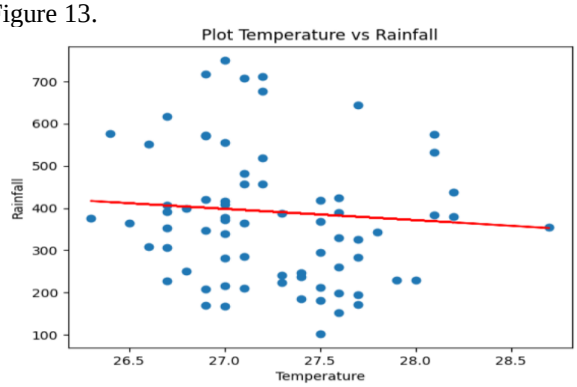


Fig 13. Regression Line Temperature And Rainfall

h. Evalution Model

The error rate of the simple linear regression model used in this study is quite large with a mean model error value of 40%, an MAE value of 124.2311, MSE 23489.97 and RMSE 153.2644. A good model has MAE, MSE and RMSE close to zero.

i. Data Predictions

The data prediction uses a temperature value of 27.9 ºC and the rainfall obtained is 374 mm as seen in Figure 14.

```
# Prediction Model With Temperature 27.9
lin_reg.predict([[27.9]])
```

```
array([[374.22324379]])
```

Fig 14. Prediction

## IV. CONCLUSION

The linear regression model between temperature and temperature in Padang City does not show a strong correlation and is negatively correlated. The amount of rainfall in Padang City is also influenced by other factors and further research is needed to find out a suitable model to determine the relationship between temperature and the amount of rainfall in Padang City.

## NOMENCLATURE

| | |
|---|---|
| Temperature | $^{o}$C |
| Rainfall | mm |

Subscripts

| | |
|---|---|
| a | Constant |
| b | Regression coefficient |
| Y | Dependent variable |
| X | Independent variable |

## REFERENCES


[1] T. K. Manik and P. B. Timotiwu, "Tracing Global Warming Path in Local Scale: Greenhouse Gas Emission or Land Use Changes?," *Eur. J. Environ. Earth Sci.*, vol. 3, no. 6, pp. 69–74, 2022, doi: 10.24018/ejgeo.2022.3.6.357.

[2] M. Ma, S. L. Collins, and G. Du, "Direct and indirect effects of temperature and precipitation on alpine seed banks in the Tibetan Plateau," *Ecol. Appl.*, vol. 30, no. 5, 2020, doi: 10.1002/eap.2096.

[3] A. C. Review, "Flood Risk Management in Arid and Semi-Arid Areas :," 2023.

[4] K. L. Findell, P. W. Keys, R. J. van der Ent, B. R. Lintner, A. Berg, and J. P. Krasting, "Rising temperatures increase importance of oceanic evaporation as a source for continental precipitation," *J. Clim.*, vol. 32, no. 22, pp. 7713–7726, 2019, doi: 10.1175/JCLI-D-19-0145.1.

[5] H. A. Musyayyadah and M. Vonnisa, "Analisa Pola Temperatur Udara Permukaan di Sumatera Barat Tahun 1980 - 2017," *J. Fis. Unand*, vol. 8, no. 1, pp. 91–97, 2019, doi: 10.25077/jfu.8.1.91-97.2019.

[6] R. Alokozay, "Impact of Climate Change on Water Resources in Kabul City," *Int. Res. J. Eng. Technol.*, no. May, pp. 1679–1688, 2020, [Online]. Available: www.irjet.net

[7] N. Thanh Son, H. Le Huong, T. Thi Phuong, and T. Trong Phuong, "Assessing Change Impacts on Flood and Drought hazard in Thua Thien Hue province using Standardized Precipitation Index (SPI)," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1345, no. 1, p. 012022, 2024, doi: 10.1088/1755-1315/1345/1/012022.

[8] S. Eka Putri, A. F. Corp, Rembrandt, Dasman Lanin, Genius Umar, and Mulya Gusman, "Kota Padang : Identifikasi Potensi Bencana Banjir Dan Upaya Mitigasi," *J. Ilm. Multidisiplin Nusant.*, vol. 1, no. 3, pp. 116–122, 2023, doi: 10.59435/jimnu.v1i3.56.

[9] W. Prarikeslan *et al.*, "The Impact of Climate Change on the City of Padang, Indonesia," *Nat. Environ. Pollut. Technol.*, vol. 22, no. 4, pp. 2223–2229, 2023, doi: 10.46488/NEPT.2023.v22i04.050.

[10] A. Y. Nofrizal, H. Rahman, and M. Hanif, "Prediction of Seawater Flooding Hazard on Settlement Areas in Padang City as a Climate Change Impact using GIS and Remote Sensing Technology," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 303, no. 1, 2019, doi: 10.1088/1755-1315/303/1/012025.

[11] S. Fawzy, A. I. Osman, J. Doran, and D. W. Rooney, "Strategies for mitigation of climate change: a review," *Environ. Chem. Lett.*, vol. 18, no. 6, pp. 2069–2094, 2020, doi: 10.1007/s10311-020-01059-w.

[12] K. Abbass, M. Z. Qasim, H. Song, M. Murshed, H. Mahmood, and I. Younis, "A review of the global climate change impacts, adaptation, and sustainable mitigation measures," *Environ. Sci. Pollut. Res.*, vol. 29, no. 28, pp. 42539–42559, 2022, doi: 10.1007/s11356-022-19718-6.

[13] M. Mudelsee, "Trend analysis of climate time series: A review of methods," *Earth-Science Rev.*, vol. 190, no. June 2018, pp. 310–322, 2019, doi: 10.1016/j.earscirev.2018.12.005.

[14] N. Nelvidawati and M. Kasman, "Penggunaan Korelasi Spearman Untuk Menguji Hubungan Suhu Dan Besarnya Curah Hujan Bulanan di Kota Padang".

[15] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[16] N. Lebkiri *et al.*, "Using Machine Learning for Prediction Students Failure in Morocco: an Application of the CRISP-DM Methodology," *Int. J. Educ. Inf. Technol.*, vol. 15, pp. 344–352, 2021, doi: 10.46300/9109.2021.15.36.

[17] Aviral Gupta, Akshay Sharma, and Dr. Amita Goel, "Review of Regression Analysis Models," *Int. J. Eng. Res.*, vol. V6, no. 08, pp. 58–61, 2017, doi: 10.17577/ijertv6is080060.

[18] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 140–147, 2020, doi: 10.38094/jastt1457.

[19] T. Esaki, "Appropriate evaluation measurements for regression models," *Chem-Bio Informatics J.*, vol. 21, pp. 59–69, 2021, doi: 10.1273/cbij.21.59.

[20] F. Emmert-Streib and M. Dehmer, "Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 521–551, 2019, doi: 10.3390/make1010032.