INDONESIA CLIMATE DATA ANALYSIS

Portfolio Project 4 | 173 Stations | 11 Years | 588,666 Records

Tool: MySQL Workbench | Dataset: BMKG / Kaggle

What This Project Was About

I worked with eleven years of daily weather observations from Indonesia's national meteorological network (BMKG), pulling data from 173 monitoring stations spread across an archipelago of over 17,000 islands. The raw data was exactly the kind of thing you'd expect from real sensor infrastructure: 370,719 missing values, questionable outliers, and stations that weren't all active at the same time. Before I could run a single SQL query, I had to clean the data in Python, and that meant making real judgment calls about what to keep, what to drop, and what to flag honestly rather than paper over.

Once I had the cleaned 588,666-record dataset loaded into MySQL, I built 12 queries around a central question: what does Indonesia's climate actually look like across a decade, and what patterns show up when you look at it from every angle? I worked through overall statistics, monthly and seasonal rhythms, year-on-year trends, extreme weather events, window functions for moving averages and period-over-period comparisons, climate classification, and a full station comparison across all 173 locations.

The goal was never to produce a report for climate scientists. It was to show that I can take a large, messy, real-world dataset, clean it, structure it, and ask the right questions. The kind of questions that actually matter to someone making a decision or reading a dashboard.

What I Found

The number I kept coming back to was 26.88°C, the average temperature across all records. That is the warm, stable baseline you'd expect from an equatorial country, but what genuinely surprised me was how little it moves across the calendar year. October and November come in as the warmest months at 27.19°C; July and August are the coolest at 26.47°C. That is less than one degree of range across twelve months. Indonesia does not have a warm season and a cold season. It has a warm season and a slightly less warm season, and I think that finding alone tells you something important about the climate.

Rainfall is a completely different story, and I find it more interesting. December averages 9.62 mm per day, nearly double August's 4.23 mm. When I ran the seasonal comparison query, both the wet and dry seasons came back with identical average temperatures: 26.88°C each. That one result communicates more about how Indonesian climate actually works than any summary statistic I could have led with. The seasons are defined by water, not warmth.

The 2015 anomaly was the most important single-year finding in the whole dataset. Total rainfall in 2015 was just 273,133 mm, well below every other year in the decade. This is not a data error. 2015 was one of the strongest El Niño years on record, and its effect on Indonesian rainfall was

severe and well-documented. The fact that the data captures this event accurately gave me confidence in the overall quality of the cleaned dataset. 2016 then bounced back to 430,392 mm, which lines up with La Niña conditions following the El Niño.

Looking across the eleven years, I can see a visible upward drift in average annual temperature, from around 26.48°C in 2010 to 27.24°C in 2016, settling at 27.07°C in 2020. That is roughly 0.3 to 0.4 degrees across the decade, which fits broadly with reported regional warming rates for Southeast Asia. I want to be careful about what I claim for this finding, and I address the limitations directly in the full analysis.

The station comparison in Query 12 was the most visually striking result in the project. Station 96937 in the north averages 28.89°C with 75% humidity and relatively low rainfall. Station 97284 in the highlands averages just 20.03°C and has recorded minimums of 7°C. Both are in Indonesia. On a map, these 173 data points would tell the country's geographic story more powerfully than anything I could write.

> *79% of all 588,666 observations were classified as Moderate, meaning temperatures between 26°C and 30°C with some rainfall present. Nearly four in five days across eleven years fall into a single category. Indonesia is not an extreme climate. It is a relentlessly consistent one.*

What I Think Could Be Improved

The temperature trend analysis is the part I would push hardest on if I were doing this again. What I have is eleven years of simple annual averages showing a directional drift upward. That is suggestive of warming, but it is not a robust trend analysis. The record counts per year vary from around 39,000 to 58,000, which tells me that the number of actively reporting stations changed across the decade. If the stations that came online during the period are mainly in warmer coastal regions, part of what looks like warming could be an artefact of changing coverage rather than actual climate change. A proper approach would fix the station set to those active for all eleven years before running any trend calculation. Without that, I can say there appears to be a warming trend, but I cannot confidently put a number on the magnitude.

The temperature range query (Query 6) is the result I am least satisfied with. It was designed to find days with the biggest gap between minimum and maximum temperature. What it actually found was a data quality anomaly: every single day in the top twenty came from station 96297, with minimum, maximum, and average temperatures recorded as the same value, 36°C across the board. A sensor reporting one value all day is almost certainly stuck or recording only a single measurement. I flag this directly in the full report rather than presenting it as a genuine finding, because I think being honest about what a query actually shows is more useful than dressing it up.

I would also go further on regional analysis. Indonesia covers more than 5,000 kilometres east to west and spans multiple climate zones. National averages flatten that variation almost entirely. Whether the 2015 El Niño hit Sumatra harder than Java, or whether the warming trend is concentrated in coastal stations versus highland ones, these are questions the data could answer with more targeted queries, and they are the logical next step for any spatial dashboard.

My Overall Take

What I value most about this project is that it made me work with data as it actually exists, not as I would have liked it to be. The 370,000 missing values, the changing station coverage, the sensor anomalies at station 96297: these are not problems I could clean my way out of entirely. They required judgment calls, and making those calls openly is what working with real data actually means.

The SQL work is technically solid. I used window functions, LAG for period-over-period comparisons, CASE-based climate classification, and rolling averages using ROWS BETWEEN. But what I am most satisfied with is when the queries ask the right question rather than just a technically impressive one. The rainfall frequency query is a good example. It would have been easier to average rainfall by month, but asking what percentage of days actually see rain produces a more honest and more interesting answer. Finding that even Indonesia's driest month still sees rain on more than 50% of days is not what most people would guess, and it completely reframes how you describe the so-called dry season.

The 2015 El Niño finding is what I would lead with in any presentation of this project. It is the moment where the data connects to something real and externally verifiable, where you can say the data shows this, and the historical record confirms it. That kind of validation is what separates analysis from description. The SQL is the tool. The story it tells is the work.