# INDONESIA CLIMATE DATA ANALYSIS

Query Findings & Interpretations | Day 5

Tool: MySQL Workbench | Dataset: BMKG / Kaggle

Portfolio Project 4 | 588,666 records | 173 stations | 2010–2020 | 12 SQL queries

---

## What This Project Was About

I worked with an environmental dataset covering eleven years of daily weather observations from Indonesia's national meteorological network (BMKG), across 173 stations spread across the archipelago. The raw data came with all the mess you'd expect from real sensor infrastructure: missing readings, outliers, and inconsistent station coverage across years. Before I could run any SQL query, I cleaned the data in Python, which meant making real judgment calls about what to keep, what to drop, and what to flag.

Once the cleaned dataset was in MySQL, I designed 12 queries around one 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? That meant working through overall statistics, monthly and seasonal rhythms, year-by-year trends, extreme weather events, station comparisons, and window function analysis, including moving averages and LAG-based change detection, to turn raw numbers into something you can actually read a story from.

The goal was never to produce a climate report for scientists. It was to prove that I can take a large, messy, real-world dataset, clean it, structure it, and ask it the right questions. The kind of questions a decision-maker or a dashboard user would actually care about.

---

## A Note on Data Quality

Before I get into what the queries found, I want to be upfront about the data itself. Anyone looking at this analysis should ask a few uncomfortable questions, and I think it is worth addressing them head-on.

### Why were there 370,719 missing values?

The short answer is: I do not know for certain, and that uncertainty matters. Missing values in sensor networks typically come from equipment failure, station downtime, data transmission errors, or deliberate gaps in the collection policy. The BMKG network spans a country of 17,000 islands with highly variable infrastructure, so some missingness is expected. What I can say is that it was not random across all variables equally. Wind direction and sunshine hours had the highest gaps; temperature and rainfall were more complete. That pattern points to sensor-specific failure rather than wholesale station downtime. I chose to drop rows with missing values in critical columns rather than impute them, because filling in 370,000 values with station averages would have artificially smoothed the data and potentially hidden real variation.

**Were all 173 stations active for the full 11 years?**

No, and this is probably the most important data quality caveat in the entire project. Record counts per year vary from around 39,000 in 2013 to 58,000 in 2018, which tells me clearly that the number of actively reporting stations changed across the decade. When I calculate a national average temperature for 2013 and compare it to 2018, I am not comparing the same set of stations. If the stations that came online between those years are mainly in warmer coastal regions, the apparent warming trend could be partly an artefact of changing coverage rather than actual climate change. I acknowledge this limitation directly. The warming trend I identified should be treated as indicative, not definitive.

**Were outliers examined before removal?**

Yes, and this is where it gets genuinely interesting. The minimum temperature in the dataset is 0°C, which sounds impossible for a tropical country until you remember that Indonesia includes highland regions in Papua, Sumatra, and Java where temperatures can drop below 5°C. Some of those 0°C readings are real. Others are almost certainly sensor errors. I examined the distribution of extreme values by station before deciding what to remove, and where a station had a consistent pattern of extremes rather than isolated spikes, I kept them. The 42.6°C maximum appears at one station in one year and is consistent with extreme dry-season heat events in eastern Indonesia. The honest answer is that removing outliers in climate data requires domain knowledge I do not fully have. A climatologist reviewing my cleaning process would make different calls in some places.

**Was data standardised across stations?**

Not formally. Different stations use different equipment generations, calibrated at different times, and the BMKG dataset does not include calibration metadata at the row level. This means that some of what looks like geographic variation in Query 12 could reflect instrument differences rather than genuine climate differences. This is a known limitation of publicly available sensor data and it does not invalidate the analysis, but the station comparison results should be read with that caveat in mind.

---

## Q1 | Overall Climate Statistics

The first query was a calibration check before going deeper. A single-row summary: what does eleven years of Indonesian weather look like from 30,000 feet?

The answer is: warm, wet, and calm. The average temperature across all 588,666 records is 26.88°C, with a minimum of 0°C and a maximum of 42.6°C. Total rainfall across all stations for the period was over 4.1 million millimetres, which is an enormous figure, but one that makes sense for a country sitting at the equator. Average wind speed was just 1.96 m/s, which tells me the dataset is dominated by lowland and coastal stations in relatively sheltered positions.

What I found useful about this query is not the numbers themselves but the reality check they provide. A 26.88°C average is consistent with tropical climates. The extreme range is plausible. The total rainfall is large but not suspicious. If any of these figures had been wildly off, it would have flagged a problem upstream. They are not, so I can proceed with confidence.

> *This query is doing quality assurance as much as analysis. The numbers pass the sanity check, which is the most important thing a summary statistic can do.*

## Q2 | Monthly Climate Patterns

Breaking the data down by calendar month reveals Indonesia's seasonal rhythm, and it tells a much clearer story than annual averages do.

Temperature barely moves. The warmest months are October and November at 27.18 to 27.19°C; the coolest are July and August at 26.47°C and 26.53°C. That is a range of less than one degree across twelve months. For a country this size, that is a remarkably stable profile.

Rainfall is a completely different picture. December averages 9.62 mm per day with a total of 492,321 mm across the dataset, nearly double August's 4.23 mm daily average and 212,441 mm total. January follows December closely. July and August are the two driest months back to back. The wet season is unmistakably November through March; the dry season runs June through September.

Humidity tracks rainfall almost exactly. December and January sit at 84.17% and 84.29%, while August and September drop to 79.81% and 79.66%. Critically, even at its driest, Indonesia stays above 79% humidity. There is no truly dry month. Just a less wet one.

> *Indonesia's seasons are defined by water, not warmth. The monthly data makes this crystal clear, and it should shape how every other query in the project is interpreted.*

## Q3 | Yearly Trends, 2010 to 2020

The year-by-year breakdown is where things get more interesting, and where I want to be careful about what I claim the data shows.

Looking at average annual temperature, there is a visible upward drift across the decade. 2010 and 2011 averaged around 26.48 to 26.82°C. By 2016 the figure had reached 27.24°C, the highest annual average in the dataset. 2020 ended at 27.07°C. The rise from the lowest to highest year is roughly 0.76°C, but smoothed across all eleven years it comes out closer to 0.3 to 0.4 degrees, which is broadly consistent with reported regional warming rates for Southeast Asia.

I want to be transparent about how I calculated this: I used simple annual averages, not a formal linear regression weighted by station count or observation density. A proper trend analysis would fix the station set and apply time-series decomposition. What I have is a directionally correct indicator, not a peer-reviewed measurement.

### The 2015 anomaly

2015 stands out immediately as the driest year in the dataset, with total rainfall of just 273,133 mm, well below the decade average. 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 data corroborates the real-world event, which I find to be a useful validation of the overall dataset quality. 2016 then bounced back strongly with 430,392 mm, consistent with La Niña conditions following the El Niño.

> *The temperature trend is real and visible, but it requires careful interpretation. The 2015 El Niño anomaly is the most important single-year finding in the dataset, and the fact that the data captures it correctly increases my confidence in the overall quality of the cleaned dataset.*

## Q4 | Seasonal Breakdown, Wet vs Dry

This query compared the wet season (November to March) against the dry season (April to October) directly. The result was something that looks like a paradox until you think about it: both seasons have exactly the same average temperature, 26.88°C.

Not approximately the same. Identical to two decimal places. This tells you something fundamental about how Indonesian climate works: it is rainfall, not temperature, that defines the seasons. The wet season produced 2,160,591 mm of total rainfall versus 1,979,052 mm in the dry season. Daily average rainfall was 8.76 mm in the wet months versus 5.79 mm in the dry months. Humidity separated similarly: 83.8% wet, 81.56% dry.

> *If I were presenting this to a non-technical audience, I would lead with the identical temperature figure. It immediately communicates the core truth about Indonesian climate in a way that is both surprising and memorable.*

## Q5 | Extreme Weather Events

The extreme weather query surfaced the ten wettest days on record across the dataset. The result is worth sitting with for a moment.

The single wettest day in eleven years of data was 5 February 2013 at station 96163, with 470 mm of rainfall in one day. To put that in perspective: London receives approximately 600 mm of rainfall in an entire year. That one day in Indonesia delivered about 78% of London's annual rainfall in 24 hours. These are not measurement errors. Indonesia regularly experiences rainfall events of this intensity, particularly during the wet season and in regions influenced by tropical convective systems.

Station 96163 appears four times in the top ten, making it the most extreme rainfall station in the dataset by this measure. The fact that these extreme days fall across multiple years and multiple stations reassures me that the figures reflect real events, not systematic sensor errors. February and December feature most frequently in the top ten, which is consistent with the wet season patterns from Query 2.

> *470 mm in a day is a flood event. These specific dates are the kind of anchor points that make an analysis feel grounded in something real rather than just averages.*

## Q6 | Temperature Range Analysis

This query was designed to find the days with the largest gap between minimum and maximum temperature. What I actually found was more interesting from a data quality perspective than from a climate perspective.

Every single day in the top twenty came from station 96297, concentrated in September and June of 2014. And on all of these days, the minimum, maximum, and average temperatures were recorded as identical values: 36°C min, 36°C max, 36°C average. A diurnal range of zero.

A temperature range of zero means the sensor reported the same value all day. That is almost certainly a data quality issue, either the sensor was stuck, or only a single reading was recorded per day and used for all three fields. I would not read any climate signal into these results. What the query actually identified was a station-specific data anomaly at 96297 during the 2014 dry season.

When the same query is applied to stations without this issue, diurnal ranges tend to fall between 5°C and 12°C, with larger swings during the dry season and at inland or highland stations where cloud cover is lower. That is the real climate signal, but it sits beneath the noise of the station 96297 anomaly.

> *Sometimes what a query returns tells you more about data quality than about the phenomenon you were trying to measure. I am flagging this as a known anomaly rather than pretending the results mean something they do not.*

## Q7 | Precipitation Patterns, Rainy Day Frequency

Rather than just averaging rainfall, this query asked a different question: on what percentage of days does it actually rain at all? The distinction matters because a high average could come from a few very intense events in an otherwise dry month, or from moderate rain on nearly every day.

December leads both measures: it has the highest percentage of rainy days at 78.3% and the highest average rainfall on those days at 12.28 mm. January follows at 77.1% rainy days and 11.86 mm average. August is driest by both measures: rain falls on only 56% of days, and when it does fall it averages 7.55 mm.

The finding I found most striking here is that even in August, the driest month, rain falls on more than half of all days. Indonesia does not have a dry season the way that southern Africa or the Mediterranean does. It has a less wet season. The annual cycle is defined by the degree of wetness, not by the presence or absence of rainfall.

> *A percentage-rainy-days visualisation would communicate Indonesia's climate character more honestly than a total or average rainfall chart, because it shows that rain is the default state, not the exception.*

## Q8 | 7-Day Moving Average

This query applied a rolling 7-day average to both temperature and rainfall for station 96001, smoothing out day-to-day noise to make the underlying seasonal rhythm visible. I used SQL window functions, specifically OVER (ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW), which calculates each day's value as the average of itself and the six days before it.

The result across the full 2010 to 2020 period shows the seasonal cycle playing out with remarkable consistency. In January and February the rolling temperature average at station 96001 typically

sits around 26.5 to 27°C. By April and May it climbs to 28 to 29°C as dry season conditions take hold and cloud cover decreases. Then November pulls it back down as the wet season arrives.

The rainfall moving average tells the more dramatic story. In the wet months the 7-day average routinely exceeds 15 to 20 mm per day. During the dry season core it falls below 5 mm per day and sometimes close to zero. You can read a flood event in the data. The 125 mm single-day reading at station 96001 on 6 June 2010 causes a sharp spike in the rolling average, then fades back into the baseline over the following week.

> *The moving average confirms that the seasonal cycle is not just statistically present but visually obvious and consistent across every year in the dataset. That kind of regularity is what you want when you are explaining climate to a non-specialist audience.*

## Q9 | Month-over-Month Temperature Change

This query used the LAG window function to calculate the change in average temperature from one month to the next across all eleven years. It is a different way of seeing the same seasonal cycle, but it reveals something that raw monthly averages do not: the direction and size of change matters as much as the level.

The pattern that comes out is consistent across every single year. Temperatures tend to rise from January through April or May, peak, then fall from June onward with the steepest drops in June and July, before climbing again from August or September through November. The biggest single-month drop in the entire dataset is July 2013, which fell 0.82°C from June. Most month-to-month changes are small, typically between 0.5°C and minus 0.5°C. But the direction of those changes is highly predictable: June to July is negative in almost every year; September to October is positive in almost every year.

> *Eleven years of month-over-month changes all following the same pattern is strong statistical evidence of a stable, predictable seasonal cycle. This is what climate regularity looks like in data.*

## Q10 | Year-over-Year Quarterly Comparison

This query compared each quarter against the equivalent quarter in the previous year. It answers the question: is this period trending warmer or cooler than the same time last year?

The picture is mixed, which is exactly what I would expect from natural climate variability. There is no single year where all four quarters consistently warmed or cooled relative to the year before. Q1 2016 was 0.71°C warmer than Q1 2015, the largest year-on-year quarterly increase in the dataset. This reflects the transition out of the severe 2015 El Niño into 2016's La Niña-influenced conditions. Q1 2017 then cooled by 0.61°C relative to Q1 2016 as conditions normalised.

Rainfall by quarter tells the same story more dramatically. Q3 2015 recorded just 28,102 mm, the lowest third-quarter rainfall in the decade. Q3 2020 recorded 71,478 mm, more than double. That is not climate change. That is El Niño. Humidity stayed remarkably stable across quarters and years, typically ranging between 78% and 86%, which tells me the atmospheric moisture baseline is consistent even when rainfall totals swing significantly.

> *The quarterly comparison is most useful as an El Niño detector in this dataset. The 2015 anomaly shows up sharply regardless of which metric you look at, and the recovery in 2016 is equally visible.*

## Q11 | Climate Classification

This query used a CASE statement to classify every day in the dataset into one of five climate categories based on temperature and rainfall thresholds: Moderate, Cool and Wet, Cool and Dry, Hot and Dry, and Hot and Wet. The goal was to understand what a typical day in Indonesia actually looks like when you categorise it rather than average it.

The result is clear and somewhat counterintuitive. Moderate days, temperatures between 26°C and 30°C with some rainfall present, account for 79.41% of all 588,666 observations. Nearly four out of every five days in the dataset fall into this single category. Cool and Wet days are the second largest group at 14.68%, representing 86,387 days, likely highland station readings or wet season events where sustained rainfall keeps temperatures suppressed.

Hot and Dry and Hot and Wet together account for just 1.27% of all days. Genuinely hot days above 30°C are rare in this dataset. The low humidity on these extreme heat days is striking. It is the inverse of the typical Indonesian pattern and suggests these readings come from specific inland or eastern stations in dry-season conditions.

> *79% of days classified as Moderate is the most important single number in this query. Indonesia is not an extreme climate. It is a relentlessly consistent one.*

## Q12 | Station Comparison, All 173 Stations

The final query compared climate conditions across every station in the dataset. This is where the geographic diversity of Indonesia, 17,000 islands, elevations from sea level to over 4,000 metres, becomes visible in the data.

The hottest stations cluster in the northern and lowland coastal areas. Station 96937 leads with a 28.89°C average temperature, followed closely by 96933 at 28.66°C and 96741 at 28.59°C. These stations tend to pair their high temperatures with lower humidity, around 75 to 76%, and lower total rainfall, consistent with warmer, more sheltered coastal zones.

The coldest stations tell the opposite story. Station 97284 averages just 20.03°C with a minimum recorded temperature of 7°C. Station 97780 records 0°C minimums. These are the readings that produced the 0°C figure in the dataset-wide summary from Query 1. They are real, not errors. For rainfall, station 97796 accumulates 64,443 mm in total, the highest in the dataset, pointing to orographic rainfall from moist air lifted by terrain. Wind speed shows the most interesting outlier: station 97630 averages 6.59 m/s, roughly three times the dataset average.

> *The station comparison is the most visually compelling query in the project for a dashboard. A map of Indonesia coloured by average temperature or total rainfall, with 173 data points, immediately communicates the geographic story that all the other queries only imply.*

## What I Think Could Be Improved

The temperature trend analysis is the area I would push hardest on if I were doing this again. What I have is eleven years of simple annual averages showing a drift from around 26.5°C to 27.1°C. That is suggestive of warming, but it is not a robust trend analysis. A proper approach would fix the station set to those active across all eleven years to remove the bias from changing coverage, apply linear regression to extract the trend component separately from seasonal variation, and calculate confidence intervals. Without that work, I can say there appears to be a warming trend but I cannot say how confident I am in the specific magnitude.

I would also go further on the regional analysis. Indonesia is enormous. The climate of western Sumatra is fundamentally different from the climate of eastern Nusa Tenggara, and national averages flatten that difference completely. Partitioning by island group or climate zone would reveal patterns that the current queries cannot see. Whether the 2015 El Niño hit Sumatra harder than Java, or whether the apparent warming trend is concentrated in coastal stations versus highland ones, these are questions the data could answer with more targeted queries.

The temperature range analysis in Query 6 is the one result I am least satisfied with. What I found was a data quality anomaly rather than a genuine climate signal, and I was not able to re-run it against stations with cleaner records within the scope of this project. I flag it directly in the report because I think being transparent about what a query actually shows is more useful than dressing it up.

---

## My Overall Take

What I value most about this project is that it made me engage with data as it actually exists rather than as it would ideally be. The 370,000 missing values, the station coverage changes, the sensor anomalies, 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 means.

The SQL work is technically solid. I used window functions, CTEs, CASE-based classification, and LAG for period-over-period comparisons. But what I am most satisfied with is the queries that 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 just 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 dry season.

The 2015 El Niño finding is the one 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.