Cleaning data in SQL

CBN exchange rate dataset

Richard Ogoma/ July 15, 2022

Problem:

Dirty exchange rate dataset **unusable** for descriptive and predictive analysis

Solution:

Help the analytics team **transform** and **clean** the exchange rate dataset to make it more usable for analysis

Data processing goals

WHY: How dirty is the data?

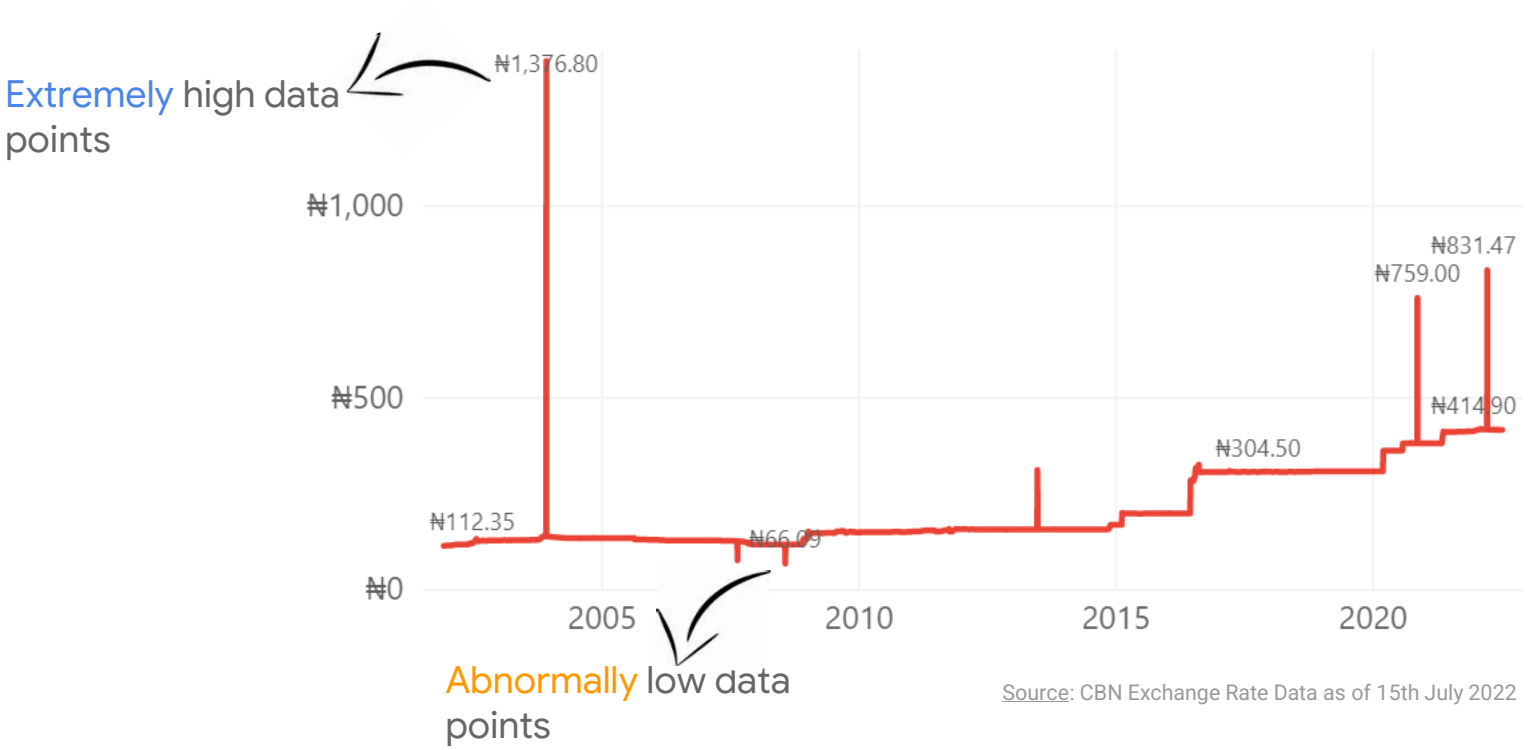
HOW: Demonstrate the data transformation from dirty to clean

Data processing goals

WHY: How dirty is the data?

HOW: Demonstrate the data transformation from dirty to clean

Unprocessed NGN/USD Central Exchange Rates Since 2001



Source: CBN Exchange Rate Data as of 15th July 2022

Business Day Data

Exchange rates were **typically** recorded on Monday through Friday; with **only 9 records on Sunday** and no record on Saturday.

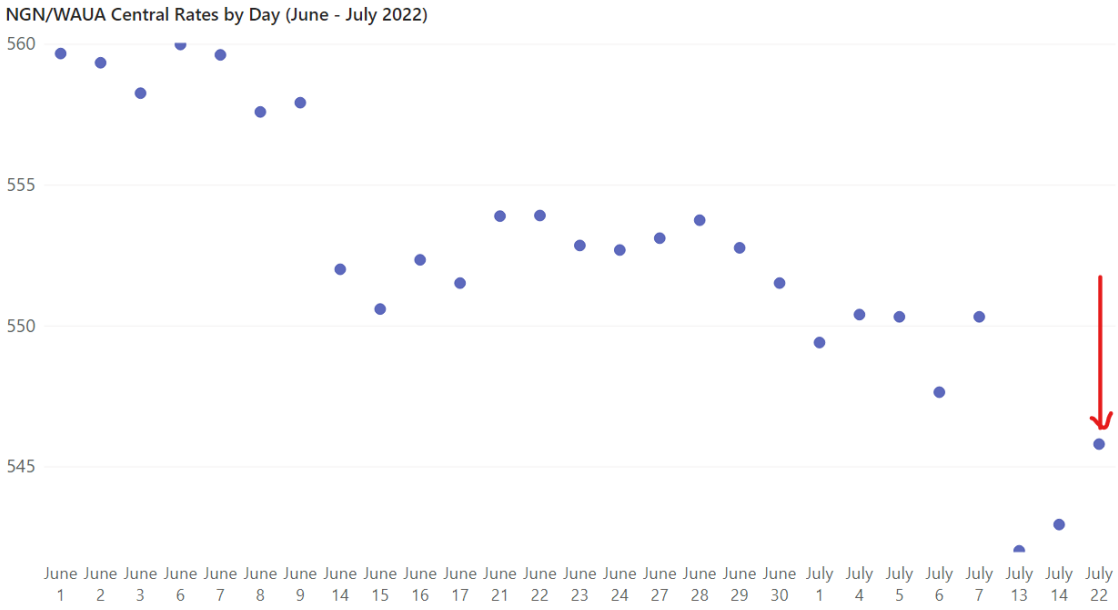
Observations by Weekday

Weekday	Observations	Contribution (%)
Wednesday	10,180	20.50%
Thursday	10,107	20.35%
Tuesday	9,903	19.94%
Friday	9,833	19.80%
Monday	9,629	19.39%
Sunday	9	0.02%
Total	**49,661**	**100.00%**

Source: CBN Exchange Rate Data as of 15th July 2022

Futuristic Dates

Exchange rate data was **erroneously** recorded beyond the current date.



Source: CBN Exchange Rate Data as of 15th July 2022

Trailing/Leading Spaces

There were **trailing or leading** spaces in the currency names, and some were **misspelt.**

	Currency	CharLength	Observations
1	CFA	3	5005
2	DANISH KRONA	12	2418
3	DANISH KRONER	13	1092
4	EURO	4	5026
5	EURO	5	1
6	JAPANESE YEN	12	1
7	NAIRA	5	8
8	POESO	5	3
9	POUND STERLING	14	5
10	POUNDS STERLING	15	5028
11	RIYAL	5	4496
12	SDR	3	3498
13	SDR	4	1
14	SOUTH AFRICAN RAND	18	1312
15	SWISS FRANC	11	4081
16	SWISS FRANC	12	1
17	US DOLLAR	9	5034
18	WAUA	4	5031
19	YEN	3	5027
20	YUAN/RENMINBI	13	2583

Source: CBN Exchange Rate Data as of 15th July 2022

Insufficient Data

The NAIRA*, the POESO, and the JAPANESE YEN contributed just **0.03%** to the total observations since 2001.

Observations by Currency

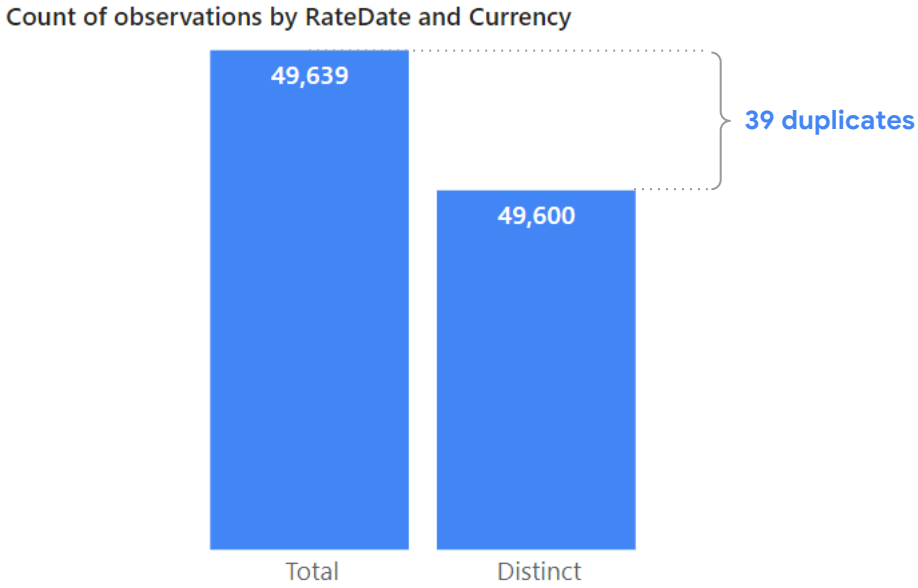
Currency	Observations	Contribution(%)
US DOLLAR	5034	10.14%
POUND STERLING	5033	10.14%
WAUA	5031	10.13%
EURO	5027	10.12%
YEN	5027	10.12%
CFA	5005	10.08%
RIYAL	4496	9.06%
SWISS FRANC	4082	8.22%
DANISH KRONE	3510	7.07%
SDR	3499	7.05%
YUAN/RENMINBI	2583	5.20%
SOUTH AFRICAN RAND	1312	2.64%
NAIRA	8	0.02%
POESO	3	0.01%
JAPANESE YEN	1	0.00%

Source: CBN Exchange Rate Data as of 15th July 2022

* The NAIRA shouldn't be in the dataset, because the rates represent the value of foreign currencies in terms of the NAIRA.

Duplicates

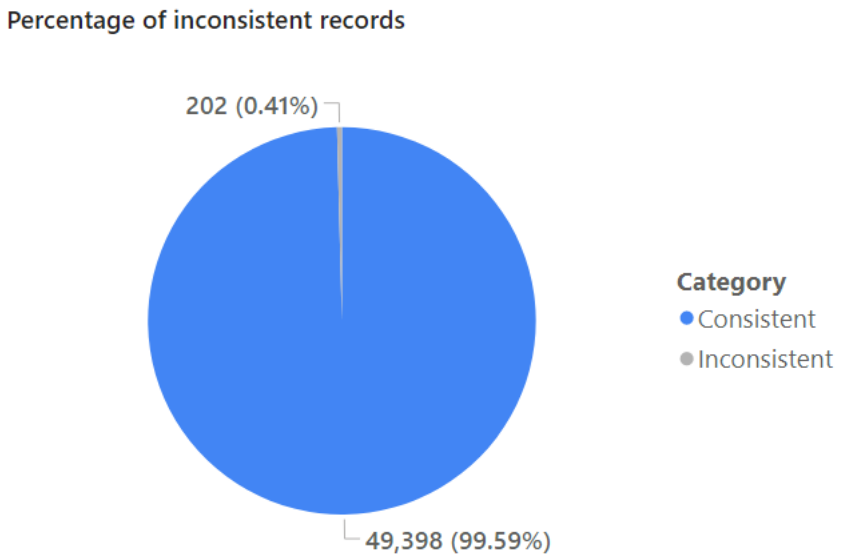
There ought to be a **singular** currency observation for each day, however, there are **39 duplicate** observations.



Source: CBN Exchange Rate Data as of 15th July 2022

Inconsistent Records

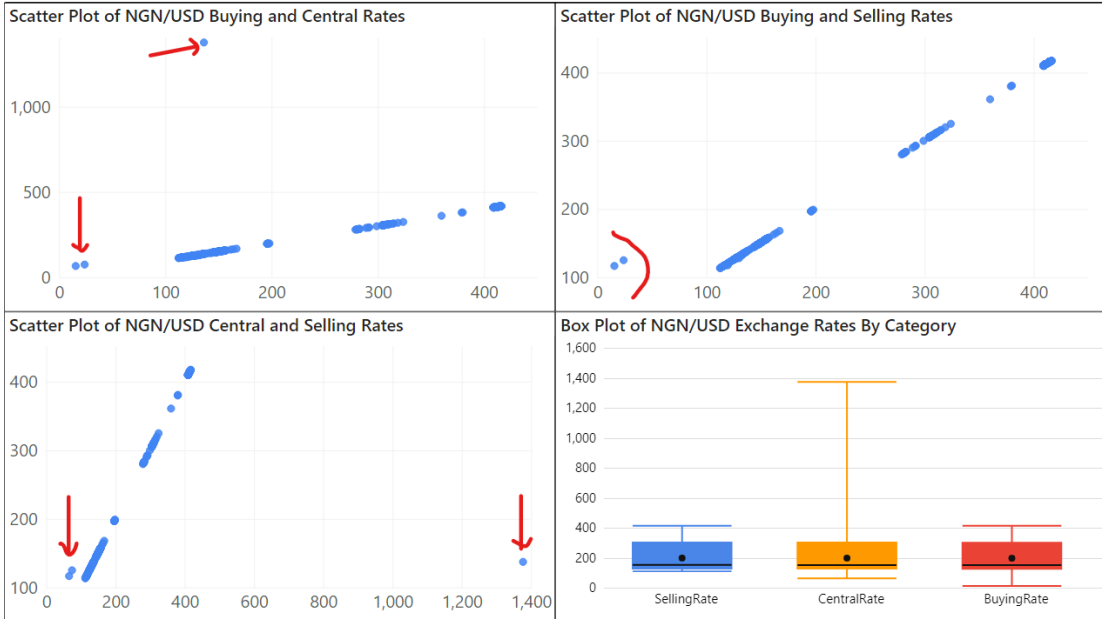
The RateYear and RateMonth fields **ought to be consistent** with the derived Year and Month features from the RateDate, but there are **202** inaccurate records.



Source: CBN Exchange Rate Data as of 15th July 2022

Outliers

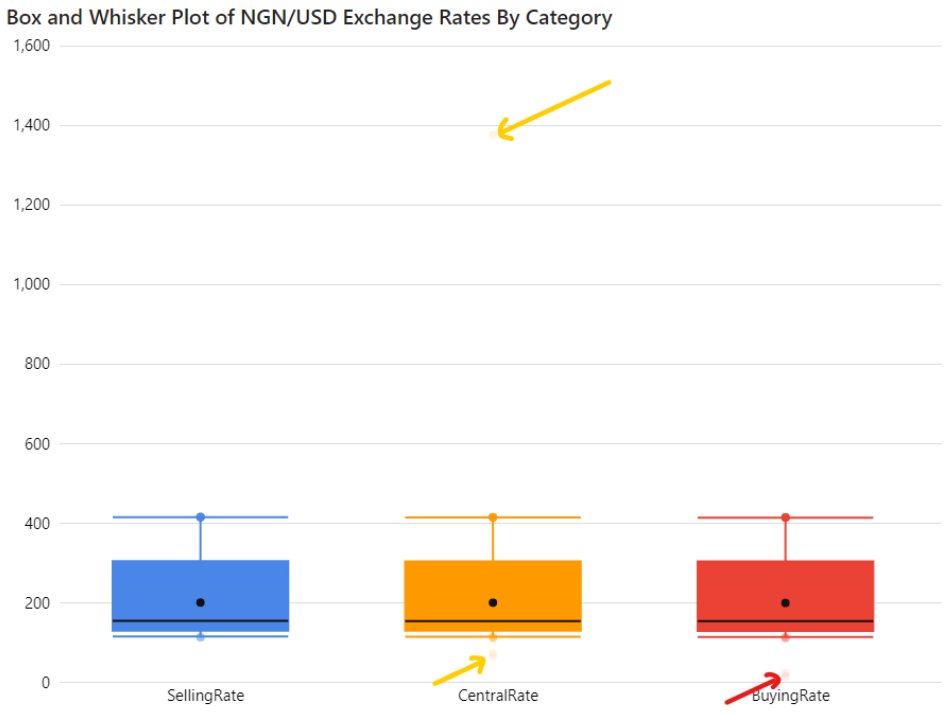
There are datapoints that differ substantially from the rest of the data. The maximum rate is **1376.80NGN/USD**, and the least is **15.59NGN/USD** in contrast to a typical average rate of **201.35NGN/USD.**



Source: CBN Exchange Rate Data as of 15th July 2022

Outliers

Our data is distributed between **115.10NGN/USD and 416.09NGN/USD** for each rate category. However, there are **abnormal datapoints** that fall outside this range.



Source: CBN Exchange Rate Data as of 15th July 2022

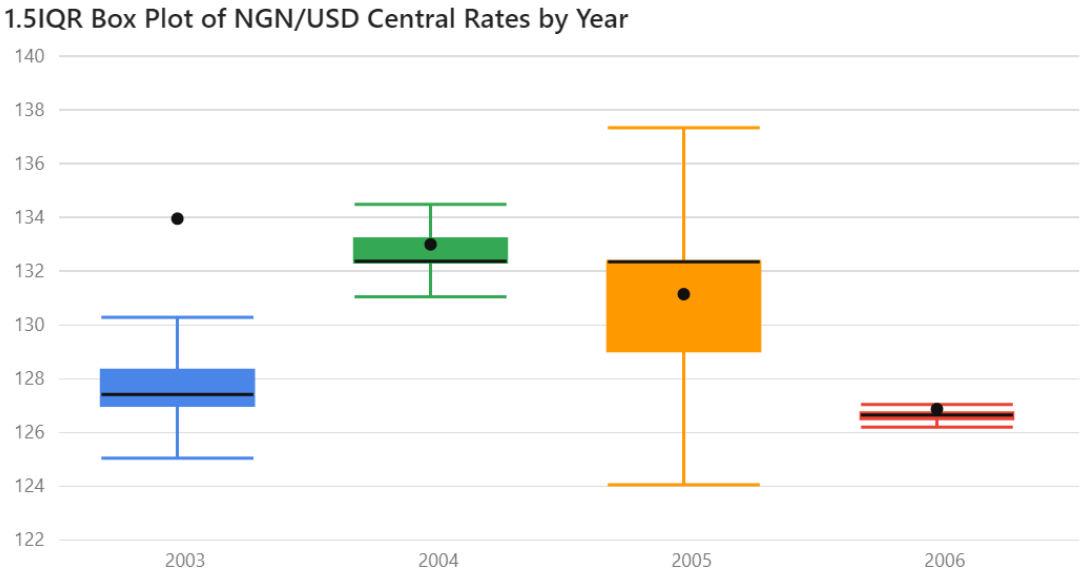
Data processing goals

WHY: How dirty is the data?

HOW: Demonstrate the data transformation from dirty to clean

The Interquartile Range

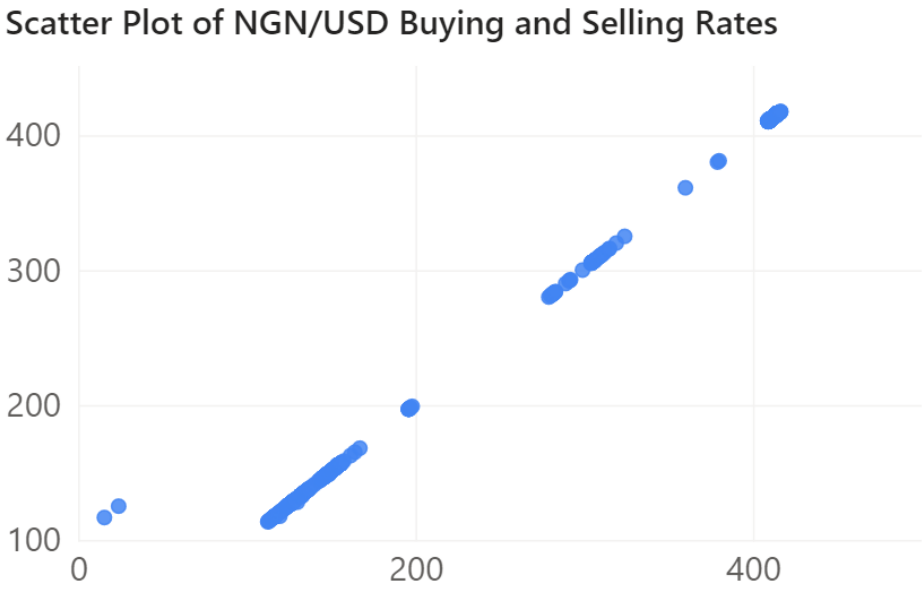
One statistical method of **identifying outliers** is by the interquartile range, or IQR. When we find rates that fall outside of **1.5 times** the range between our first and third quartiles, we typically consider these to be outliers.



Source: CBN Exchange Rate Data as of 15th July 2022

The Absolute Deviation

There is a **strong positive correlation** and "**typically" no difference** between the rates. So, rates having at least one-point absolute distance from the lateral mean and exceed 1.5 times the IQR, are considered anomalous datapoints.



Correlation coefficient, r = 0.999772

Source: CBN Exchange Rate Data as of 15th July 2022

How was the data cleaned?

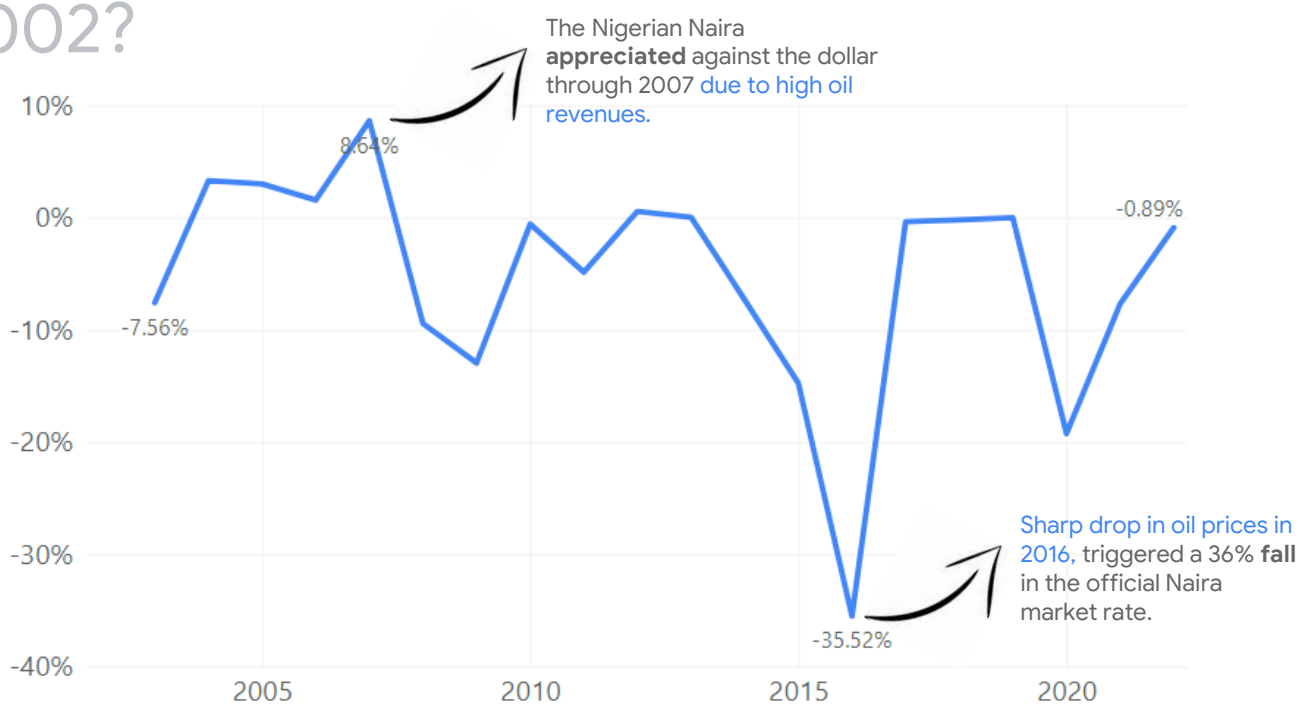
This describes the methodology used in processing the exchange rate data

- Rates that fall outside 1.5 times the IQR, having at least 1-point absolute deviation from the lateral mean were **identified** as anomalous data points
- Anomalous exchange rates were **replaced** with the longitudinal average of the preceding and following rates of the observed datapoints
- Due to the business day constraint, the rates were **transformed** to monthly averages to avoid the interpretation of a non-existent daily trend
- Other issues identified with the data, for example, duplicates and futuristic dates, were either **updated** with the correct values or **deleted** to avoid skewing the data

Processed Average Monthly NGN/USD Central Exchange Rates Since 2002

Source: CBN Exchange Rate Data as of 15th July 2022

How has the Nigerian NGN depreciated since 2002?



Source: CBN Exchange Rate Data as of 15th July 2022

Thank You