Introduction to data visualization

DATA VISUALIZATION IN DATABRICKS

Gang Wang
Senior Data Scientist

datacamp

Your data visualization partner

Gang Wang

Senior Data Scientist

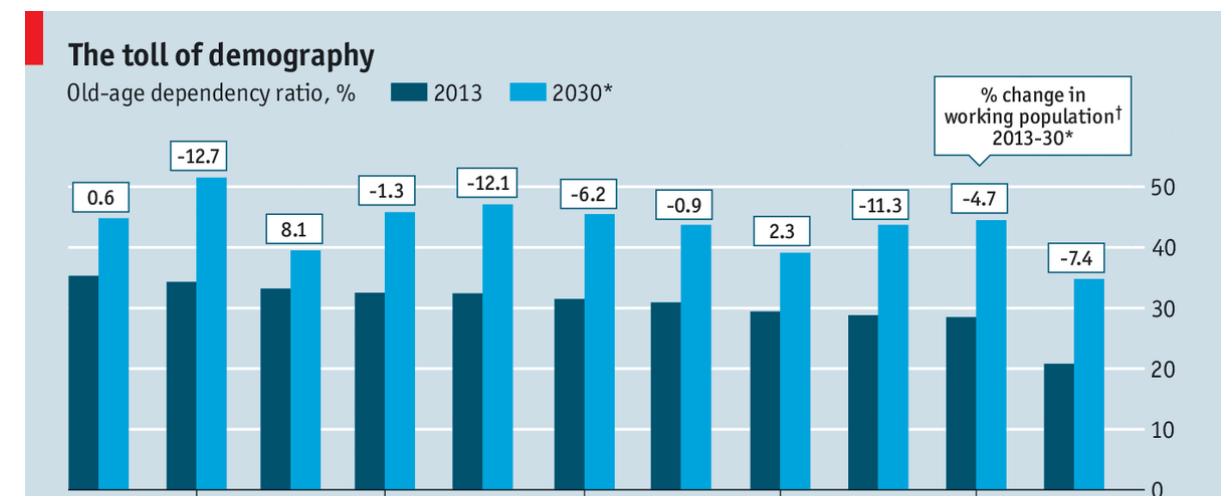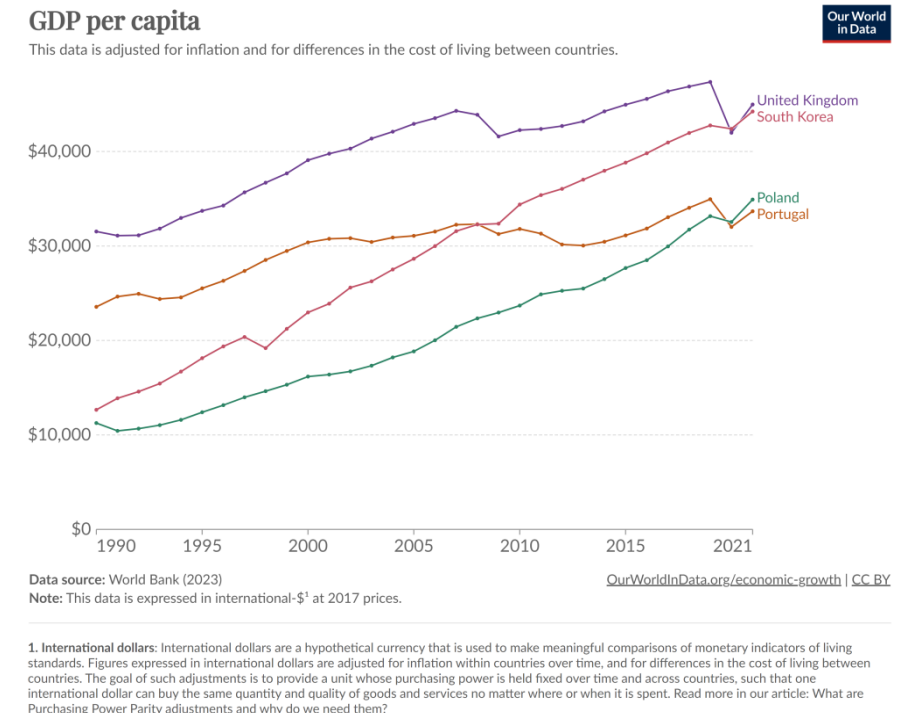Origin Energy, Australia (2021-Present)

9+ Years post-PhD experience

What is data visualization?

Data visualization is the practice of representing data in a visual format.

Formats: Includes charts, graphs, maps, and infographics.

Main goal:

- Make complex data more accessible.
- Enhance understanding and usability.

[1] Pictures: Economist Writing Every Day, The Economist

Why we need data visualization?

Benefits:

- Simplifies complex data

- Highlights key patterns and trends

- Enhances visual processing

- Improves understanding and retention

- Supports decision-making and planning

- Boosts data accessibility and collaboration
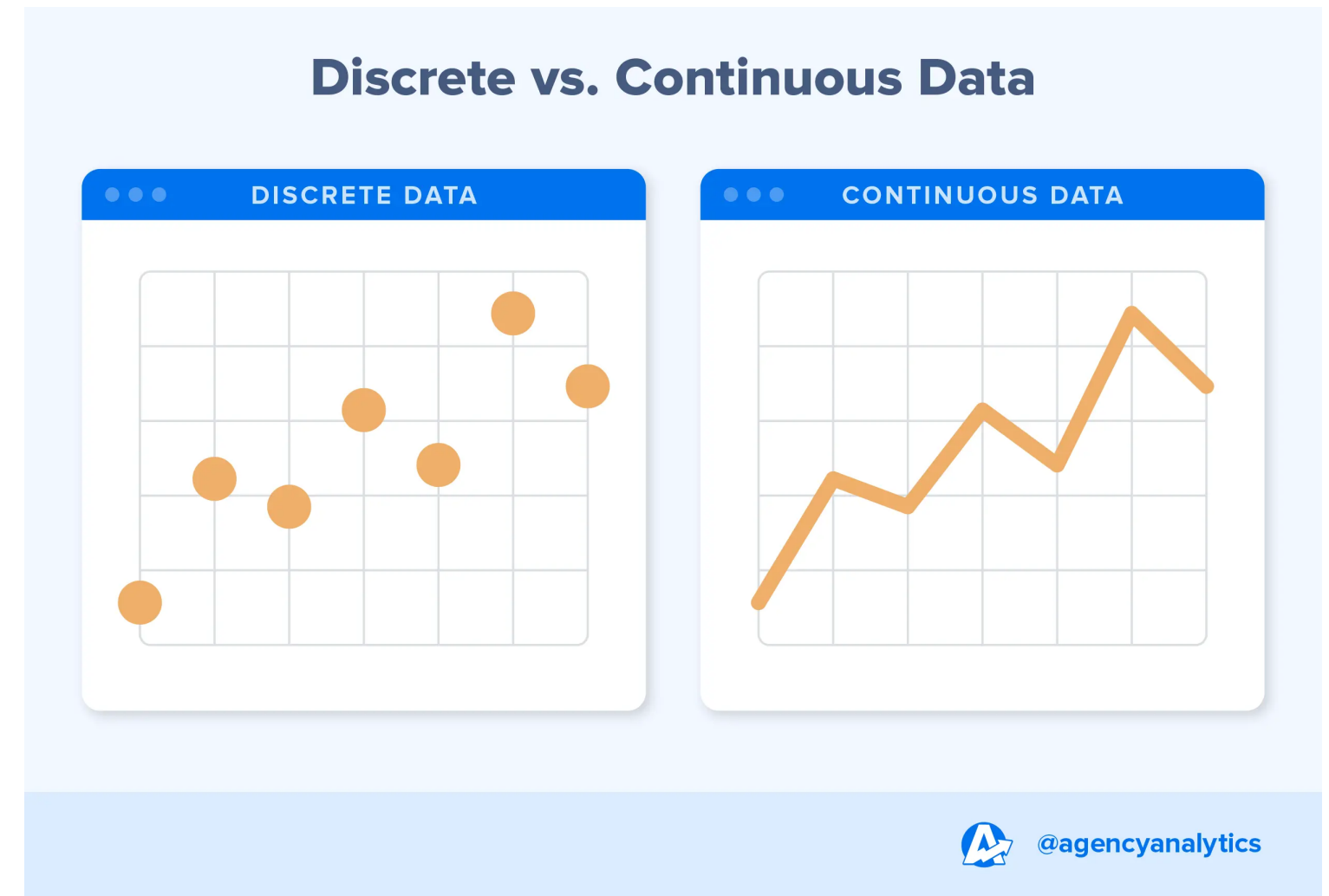
[1] Pictures: Kovair

Key statistical concepts for visualization

Discrete versus continuous data

- Discrete data: countable, distinct values

- Continuous data: Measurable quantities with a range of values

Descriptive statistics

- Summarize data to reveal trends, patterns, and outliers

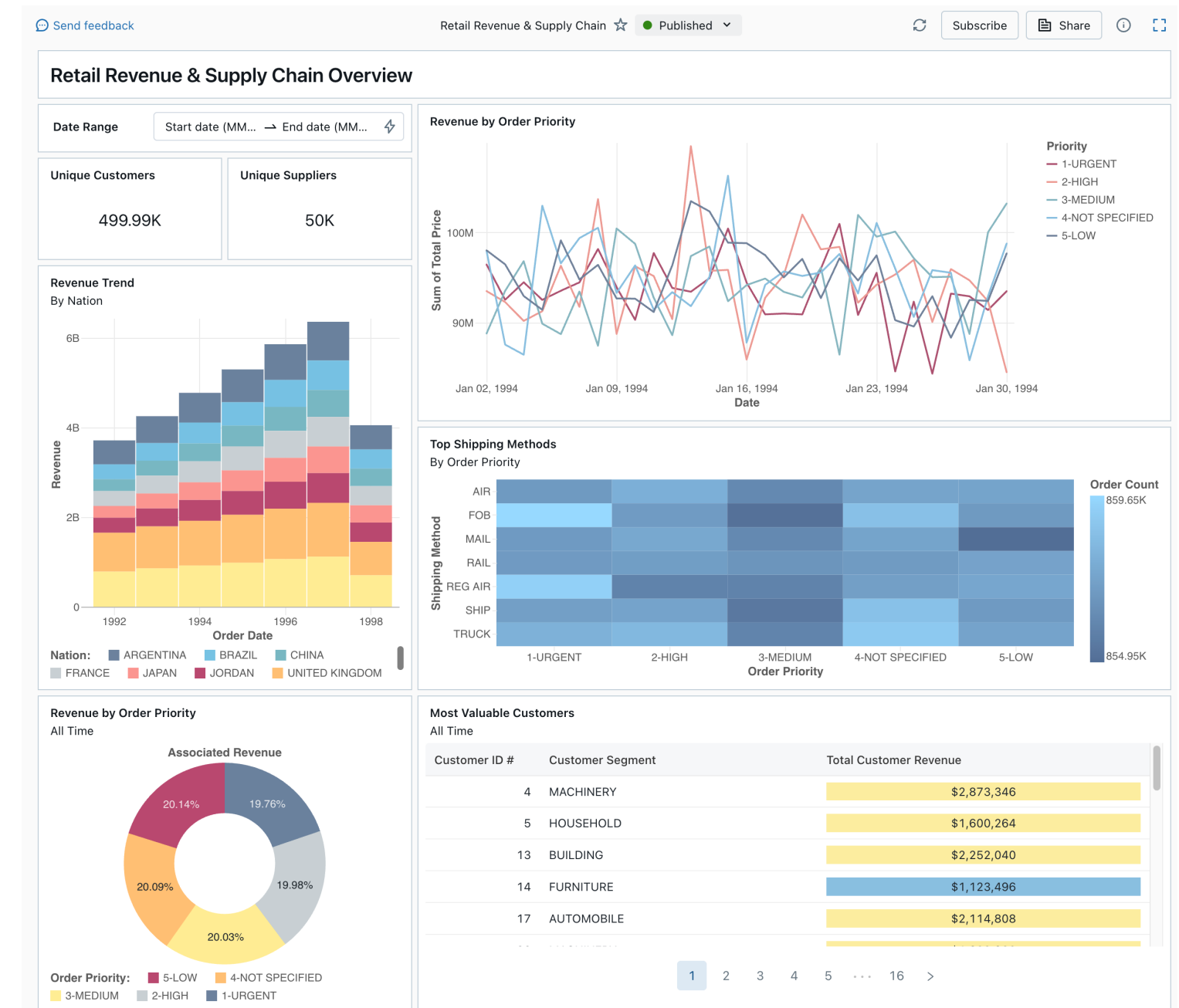- Examples: mean, median, frequency distributions

[1] Pictures: AgencyAnalytics

Databricks for data visualization

Benefits:

- Efficiently handling large datasets

- Built-in visualization options

- Interactive dashboards

- Collaborative environment

Understanding our dataset

Dataset: NYC Taxi dataset from Databricks

Includes: pick-up and drop-off locations, times, distance, and fares

Column Name	Details
tpep_pickup_datetime	Date and time when the ride began
tpep_dropoff_datetime	Date and time when the ride ended
trip_distance	Distance of the ride in miles
fare_amount	Fare charged for the ride in dollars
pickup_zip	ZIP code where the passenger was picked up
dropoff_zip	ZIP code where the passenger was dropped off

Let's practice!

DATA VISUALIZATION IN DATABRICKS

Create a Bar Chart

DATA VISUALIZATION IN DATABRICKS

Gang Wang
Senior Data Scientist

datacamp

Let's practice!

DATA VISUALIZATION IN DATABRICKS

Map visualizations

DATA VISUALIZATION IN DATABRICKS

Gang Wang
Senior Data Scientist

What are map visualizations?

- Visual representation of data on a map

- Key tool for analyzing geographical data

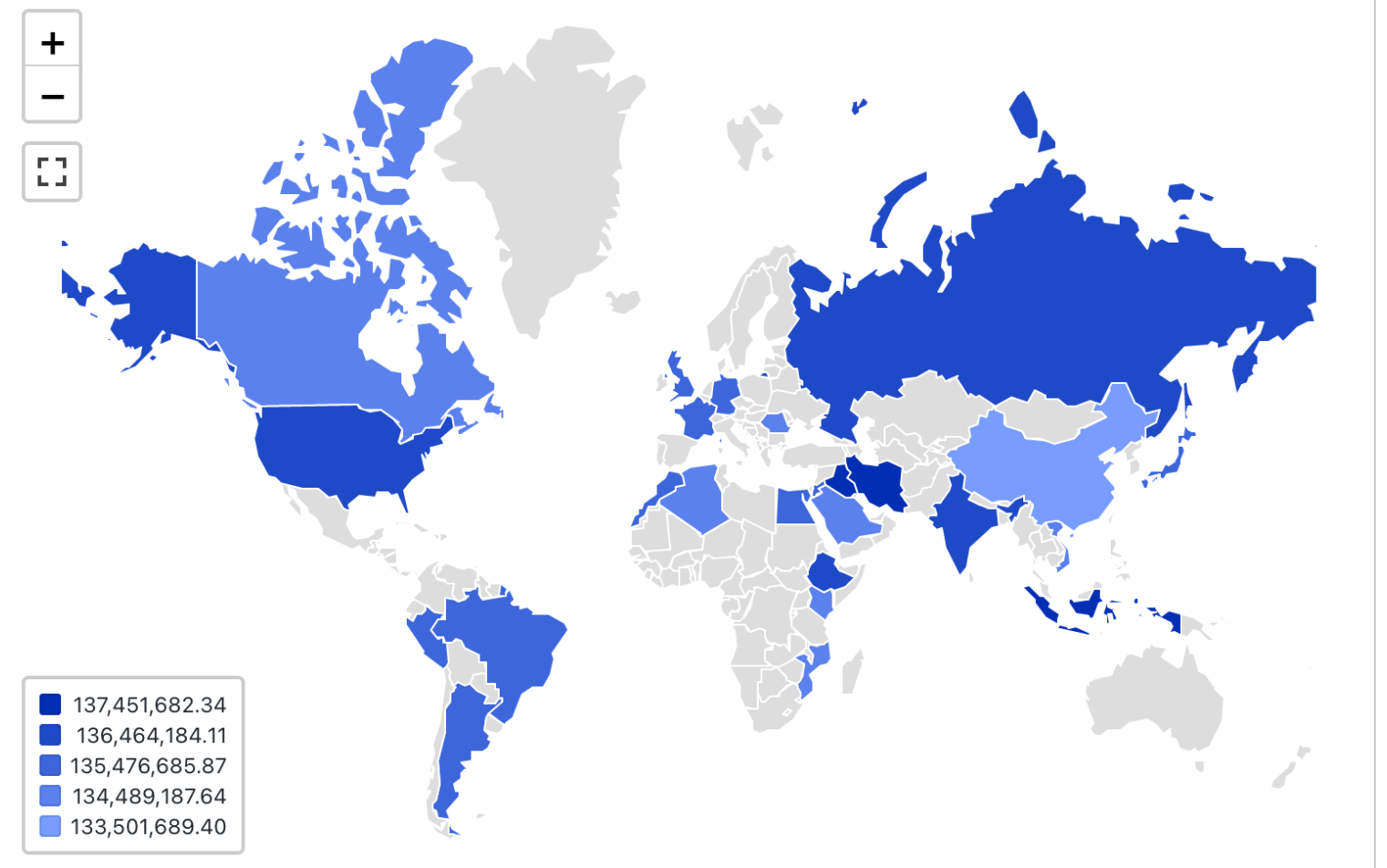- Useful for identifying spatial patterns and trends

A map visualization depicting the population density of countries around the world.

[1] Our World in Data

Examples - Choropleth Maps

Choropleth Maps
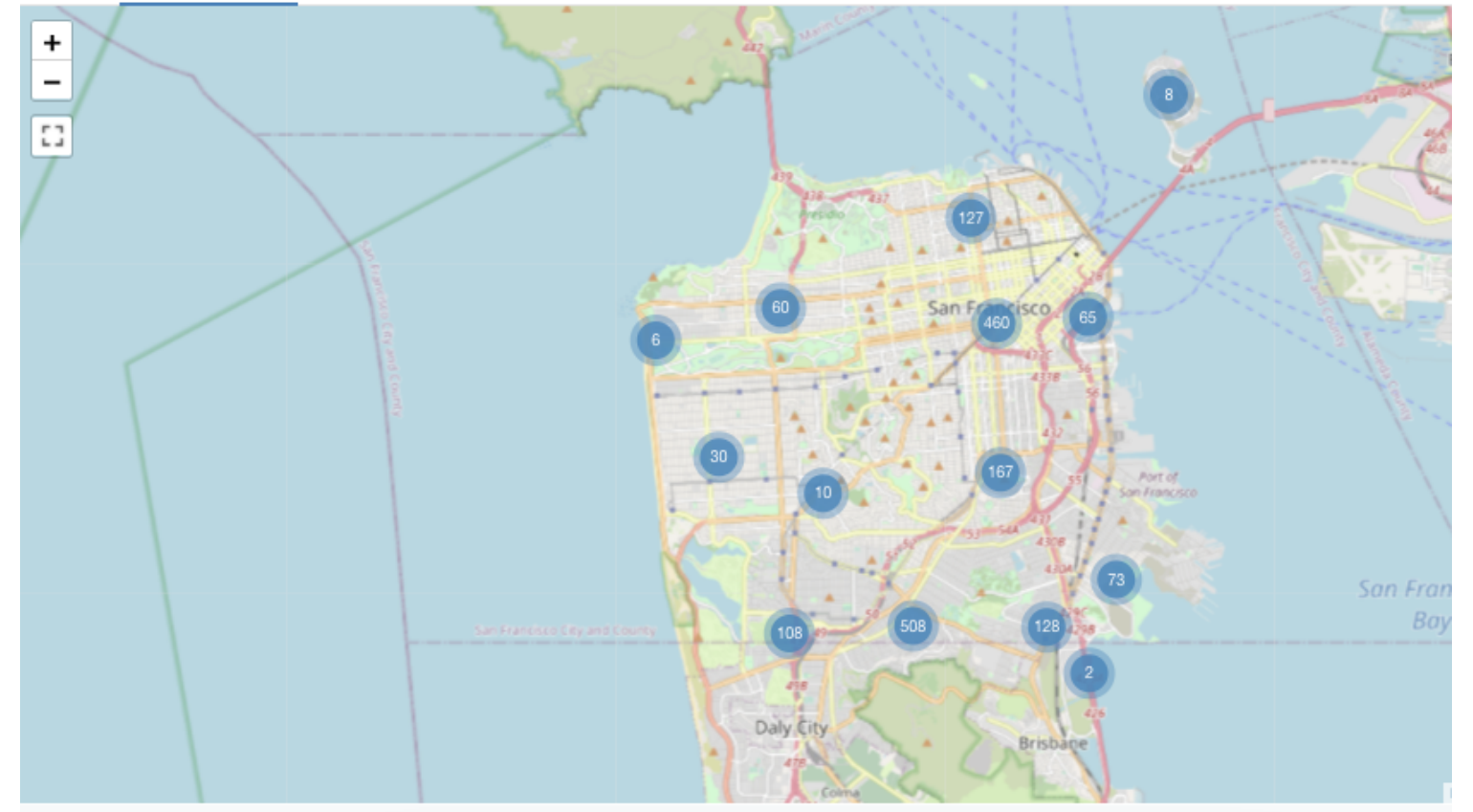
- Uses color gradients to represent data density or values

- Ideal for comparing data across regions

- Highlights regional differences clearly

137,451,682.34
136,464,184.11
135,476,685.87
134,489,187.64
133,501,689.40

Examples - Marker Maps

Marker Maps

- Represents individual data points with markers

- Ideal for visualizing specific locations

- Provides clear, location-based insights

Understanding our dataset

Dataset: Online retail dataset from Databricks

Column Name	Details
InvoiceNo	Unique identifier for each invoice
StockCode	Identifier for the stock item purchased
Quantity	Number of units of the stock item purchased in the transaction
InvoiceDate	Date and time when the transaction occurred
UnitPrice	Price per unit of the stock item
CustomerID	Unique identifier for the customer making the purchase
Country	Geographical location where the purchase was made
Latitude/Longitude	Latitude/Longitude coordinate of the transaction location

Let's practice!

DATA VISUALIZATION IN DATABRICKS

Building map visualizations in Databricks

DATA VISUALIZATION IN DATABRICKS

Gang Wang
Senior Data Scientist

datacamp

Let's practice!

DATA VISUALIZATION IN DATABRICKS