# Air Quality Visuals With MongoDB

MongoDB tool exploration

Kevin Liao
Big Data Tools & Techniques
2024

# Goals

- Explore Visuals with MongoDB
- Get more comfortable with cloud based utilities instead of self hosted tools
  - Technology constantly changes, should learn on IaaS/SaaS more
- Not spend money

# Steps

1. Obtain Data: https://www.kaggle.com/datasets/threnjen/40-years-of-air-quality-index-from-the-epa-daily/data
2. Cut/Clean Data
    a. Too large for Free Atlas (reduce data)
    b. awk -F ',' ' ($2 >= "1980-01-01" && $2 <= "2011-12-31") {next} {print} ' aqi_daily_1980_to_2021.csv > aqi_daily_2011_to_2021.csv
3. Register a MongoAtlas account
    a. https://cloud.mongodb.com
4. Download MongoDB Compass (Client Interface)
    a. https://www.mongodb.com/try/download/compass
5. Connect Compass with Atlas
    a. Configure Compass to connect to your MongoDB Atlas cluster using your cluster credentials.
6. Familiarize with MongoDB Atlas Charts

# Data

**Content**

This set includes daily reports of air quality index from various US Metro areas, as well as geographic data for the collection locations. This set has over 10 million rows.

Want a smaller set? I also have ANNUAL metrics available here. The yearly dataset is ~34k rows.

**Acknowledgements**

Data sourced directly from the Environmental Protection Agency
Banner: Photo by Maxim Tolchinskiy on Unsplash


```
_id: ObjectId('662d5f6d15f4125655a7e794')
State Name : "Alabama"
Date : 2021-01-01T00:00:00.000+00:00
AQI : 30
Category : "Good"
Defining Parameter : "Ozone"
Latitude : 34.289001
Longitude : -85.970065
County Name : "DeKalb"
```

# Cut Data

- Limited to 500 Mb
- Had to cut daily data from *10 million records* from 1980-2021 to *1.37 million records* 2017-2021
- `awk -F ',' ' ($2 >= "1980-01-01" && $2 = "2016-12-31") {next} {print} '  $input_file  > $output_file`

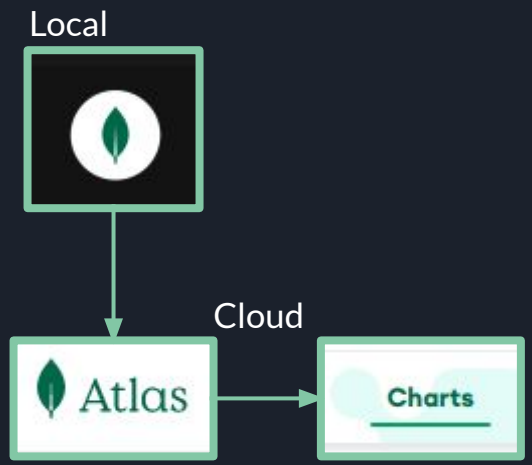
```
cheerio@cheerio-precision:~/Rowan_Local/BigData/BigDataFP/data/air_quality/air_quality (main)$ cat aqi_daily_2017_to_2021.csv | wc -l
1379828
cheerio@cheerio-precision:~/Rowan_Local/BigData/BigDataFP/data/air_quality/air_quality (main)$ cat aqi_daily_1980_to_2021.csv | wc -l
10158528
```

```
cheerio@cheerio-precision:~/Rowan_Local/BigData/BigDataFP/data/air_quality/air_quality (main)$ ll
total 926556
drwxrwxr-x 2 cheerio cheerio      4096 Apr 27 16:40 ./
drwxrwxr-x 3 cheerio cheerio      4096 Apr 27 15:45 ../
-rw-rw-r-- 1 cheerio cheerio 669029200 Aug 28  2021 aqi_daily_1980_to_2021.csv
-rw-rw-r-- 1 cheerio cheerio 190672791 Apr 27 15:40 aqi_daily_2011_to_2021.csv
-rw-rw-r-- 1 cheerio cheerio  89074250 Apr 27 16:14 aqi_daily_2017_to_2021.csv
```

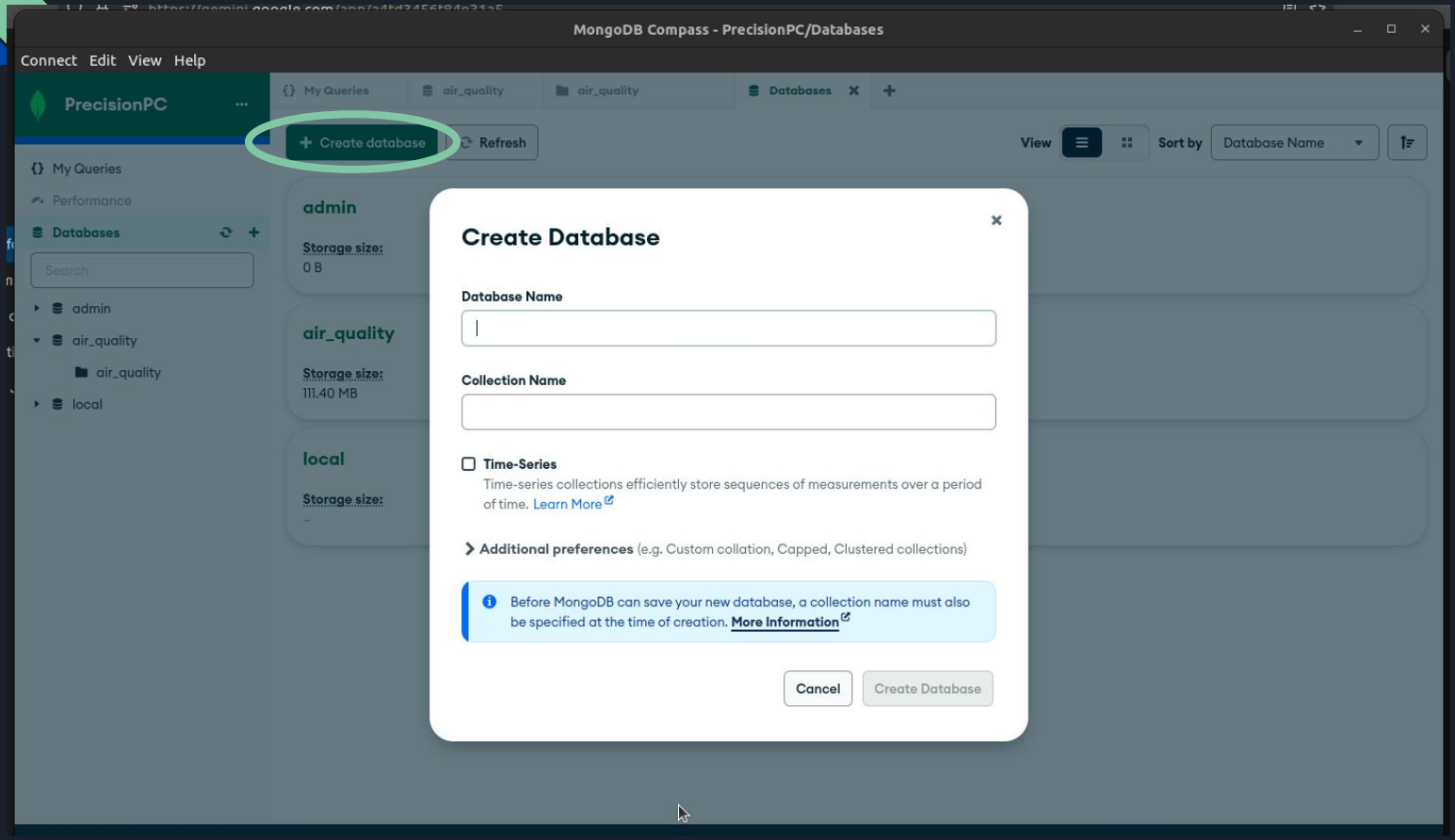
```
)$ ls
air_quality.zip       aqi_daily_1980_to_2021.csv
air_quality.zip.orig  aqi_daily_2017_to_2021.csv
cheerio@cheerio-precision:~/Rowan_Local/BigData/BigDataFP/data/air_quality
)$ cat aqi_daily_2017_to_2021.csv | egrep -v "2021" | wc -l
1339545
```

# MongoDB

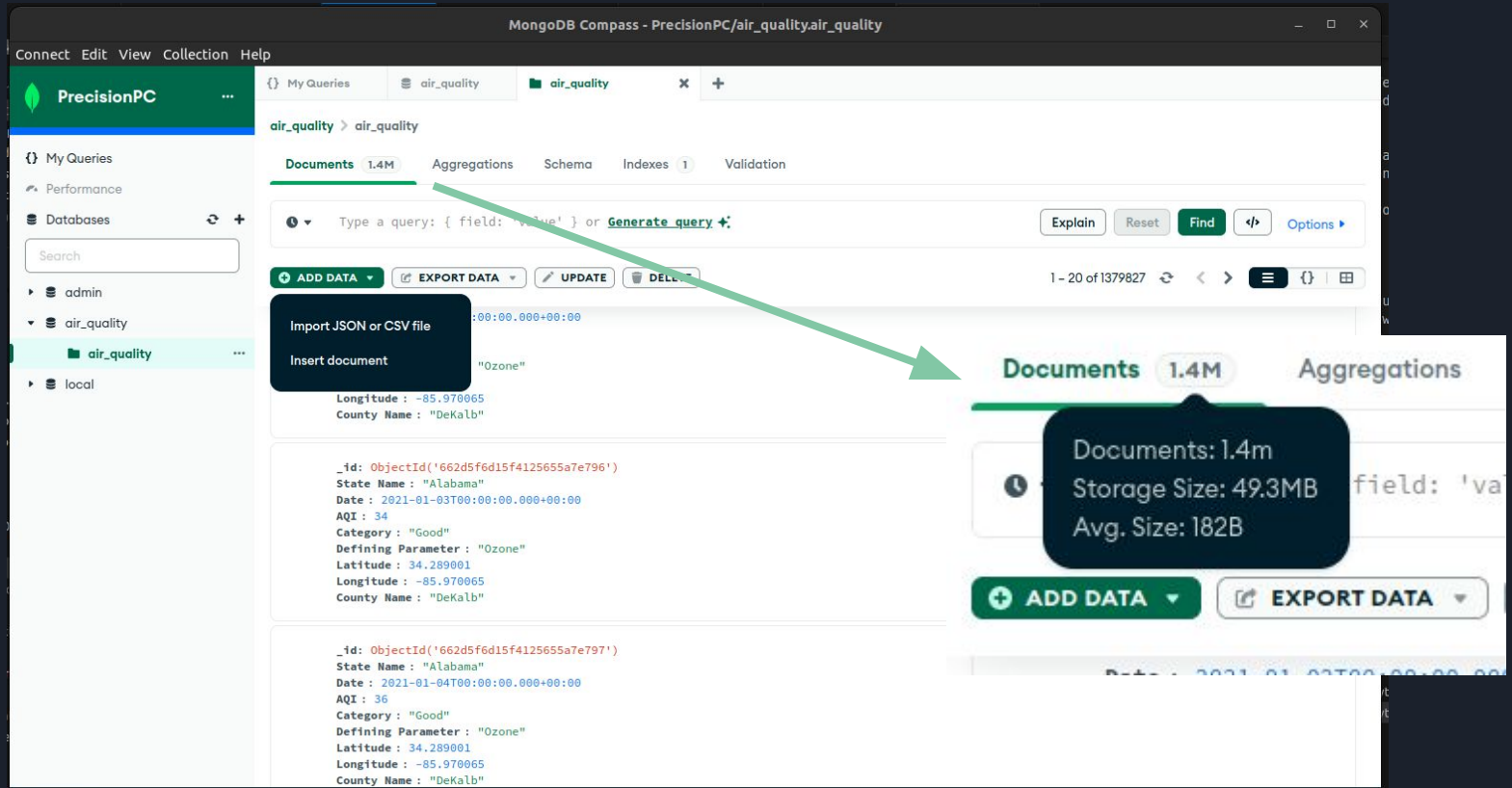
- MongoDB Compass
  - Mongos GUI Desktop Utility
- MongoDB Atlas
  - Mongos online SaaS
  - Billing structure like Azure
- MongoDB Charts
  - Online SaaS visualization tools like Tableau
  - Free tier limits



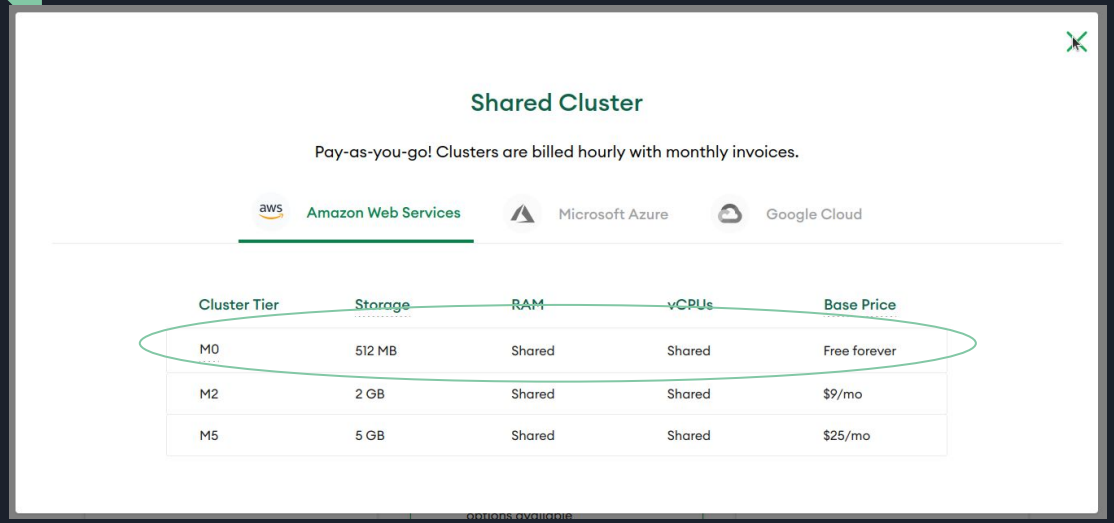
# Import Data w/ MongoDB Compass



# Import Data w/ MongoDB Compass



# MongoDB Atlas Free Tier



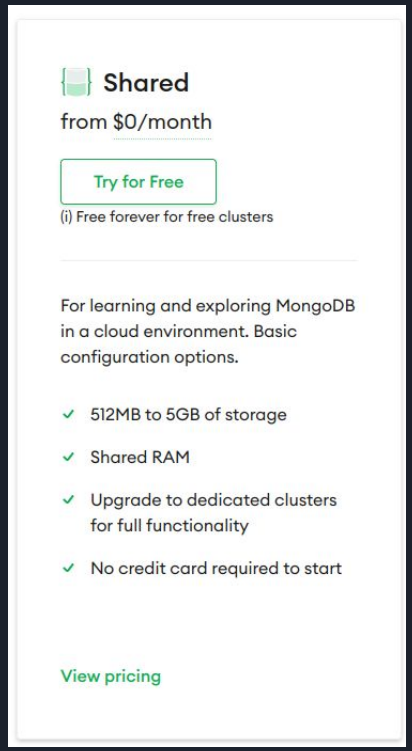
## Shared Cluster

Pay-as-you-go! Clusters are billed hourly with monthly invoices.

aws Amazon Web Services | Microsoft Azure | Google Cloud

| Cluster Tier | Storage | RAM | vCPUs | Base Price |
|---|---|---|---|---|
| M0 | 512 MB | Shared | Shared | Free forever |
| M2 | 2 GB | Shared | Shared | $9/mo |
| M5 | 5 GB | Shared | Shared | $25/mo |

options available



## Shared

from $0/month

Try for Free

(i) Free forever for free clusters

For learning and exploring MongoDB in a cloud environment. Basic configuration options.

- ✓ 512MB to 5GB of storage
- ✓ Shared RAM
- ✓ Upgrade to dedicated clusters for full functionality
- ✓ No credit card required to start

View pricing

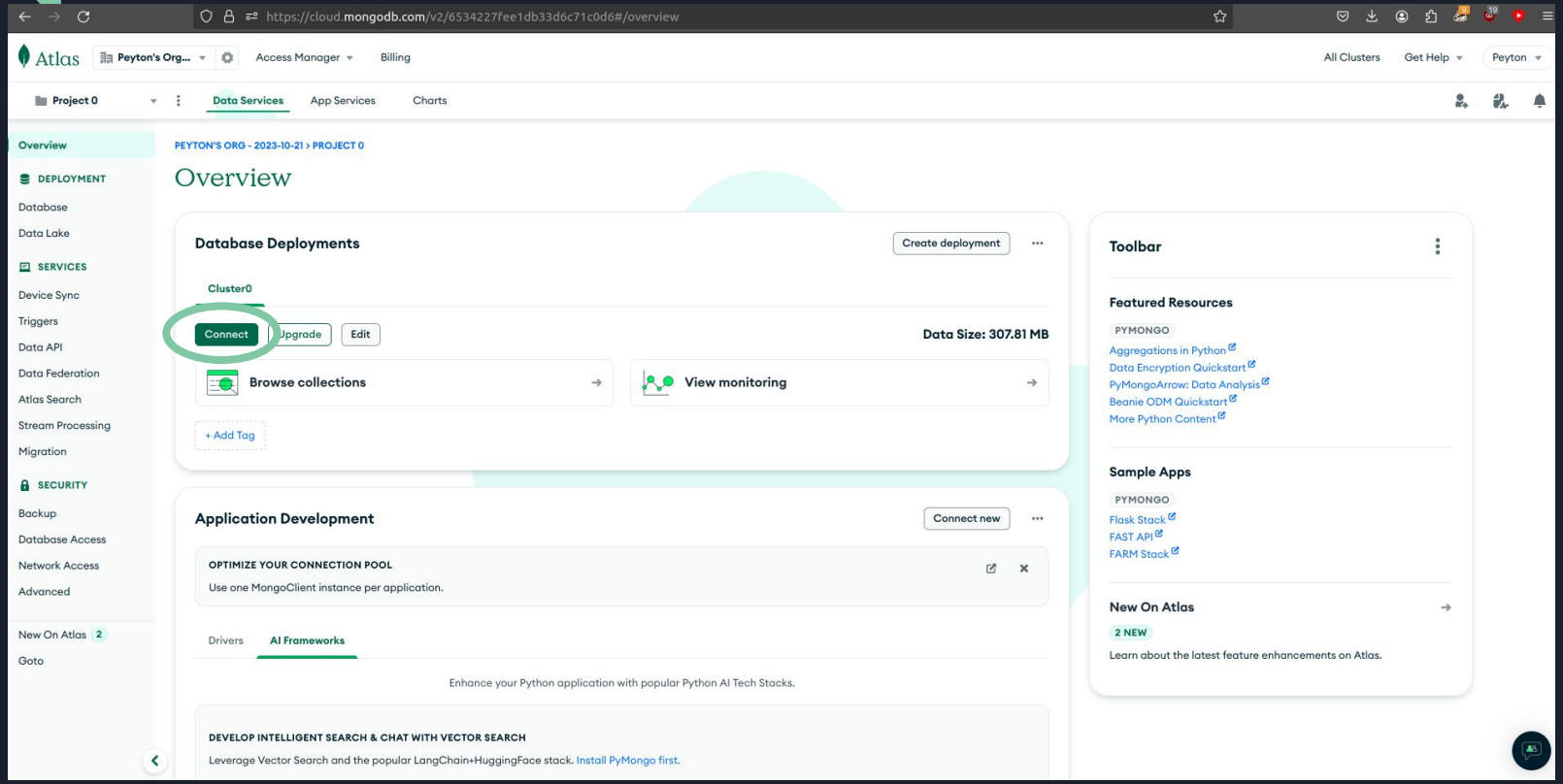
https://www.mongodb.com/pricing

# MongoDB Atlas Free Tier

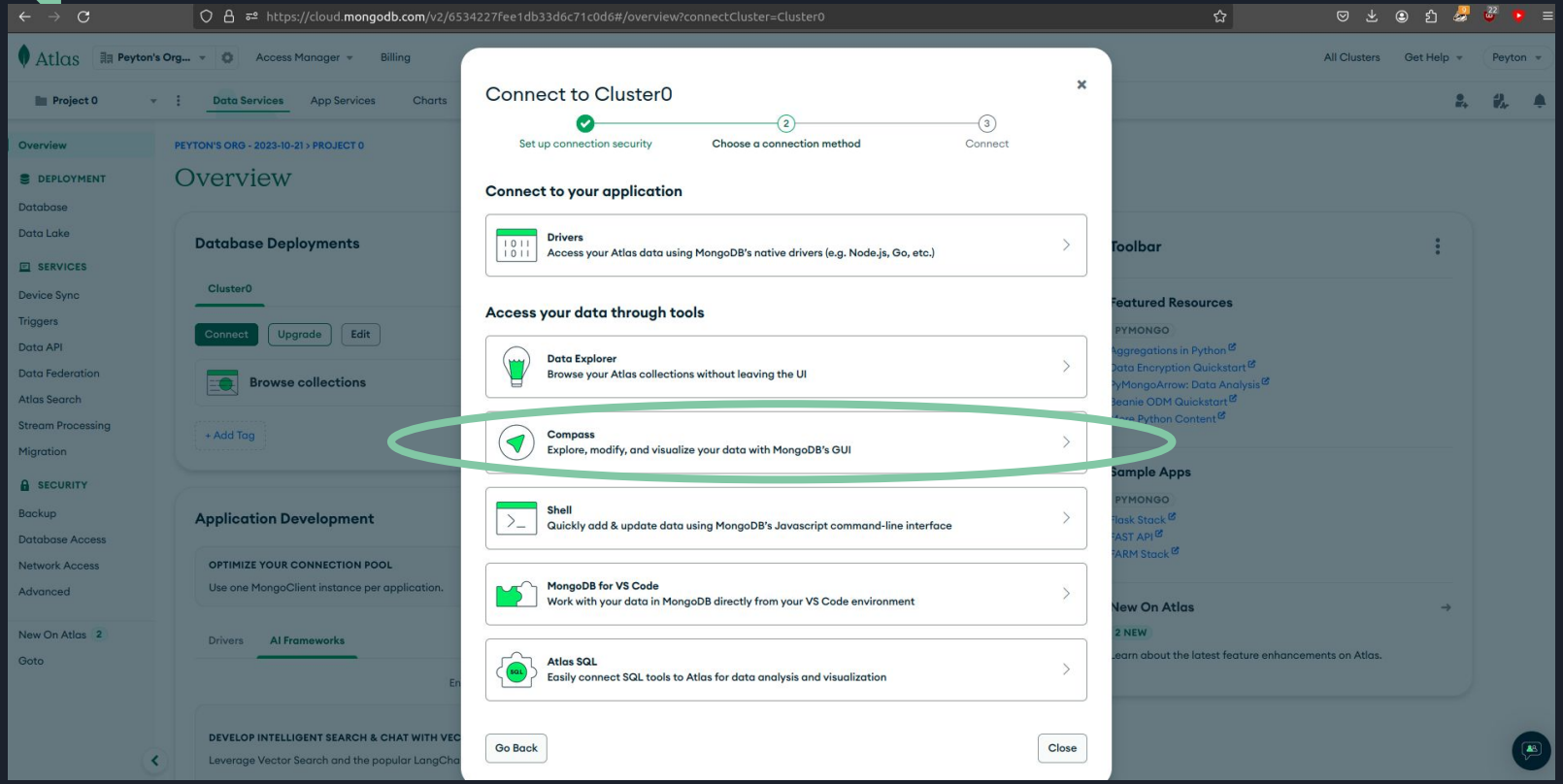
| | |
|---|---|
| **MongoDB Version and Storage Engine** | Atlas uses MongoDB 6.0 for `M0/M2/M5` clusters. |
| **MongoDB Version Upgrade** | You can't upgrade the MongoDB version that `M0/M2/M5` clusters run. Atlas upgrades `M0` free clusters or `M2/M5` shared clusters to the newest MongoDB version after several patch versions become available for that version. To learn more, see MongoDB Versioning. |
| **Cluster Tier** | You must select the `M0` cluster tier to deploy a free cluster.<br>**NOTE**<br>You can deploy only one `M0` cluster per project. |
| **Cluster Memory** | You can't configure memory for `M0` free clusters or `M2/M5` shared clusters. |
| **Cluster Storage** | You can't configure storage size for `M0` free clusters or `M2/M5` shared clusters. |
| **Replication Factor** | Replication Factor is set to `3 Nodes` and you can't modify it for `M0` free clusters or `M2/M5` shared clusters. |
| **Replica Set Tags** | `M0` free clusters and `M2/M5` shared clusters don't have pre-defined replica set tags. |
| **Do You Want A Sharded Cluster** | You can't deploy a `M0` free cluster or `M2/M5` shared cluster as a Sharded Cluster. |
| **Do You Want To Enable Backup** | You can't enable backups on `M0` free clusters. |

https://www.mongodb.com/docs/atlas/reference/free-shared-limitations/

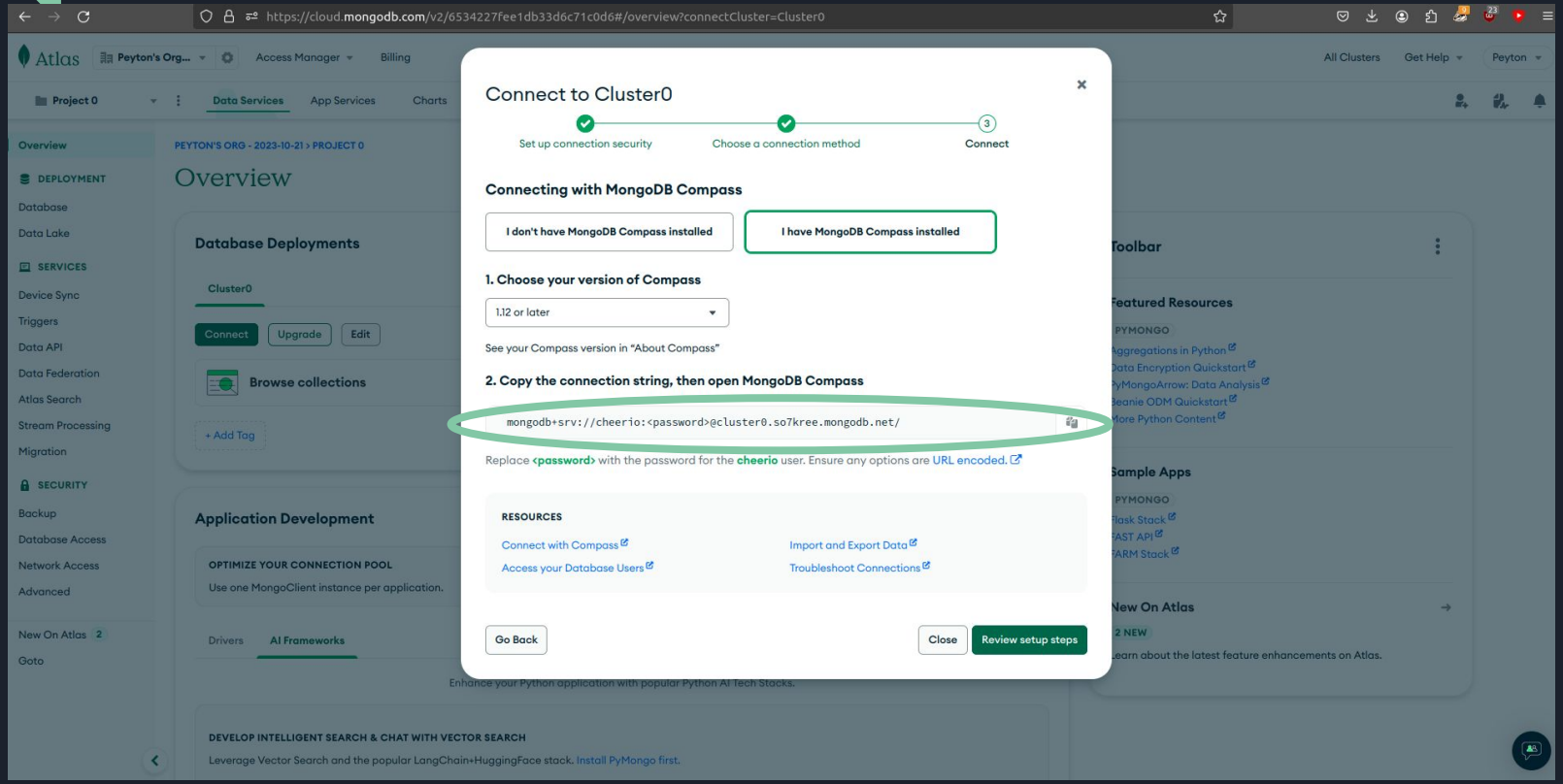
# MongoDB Atlas



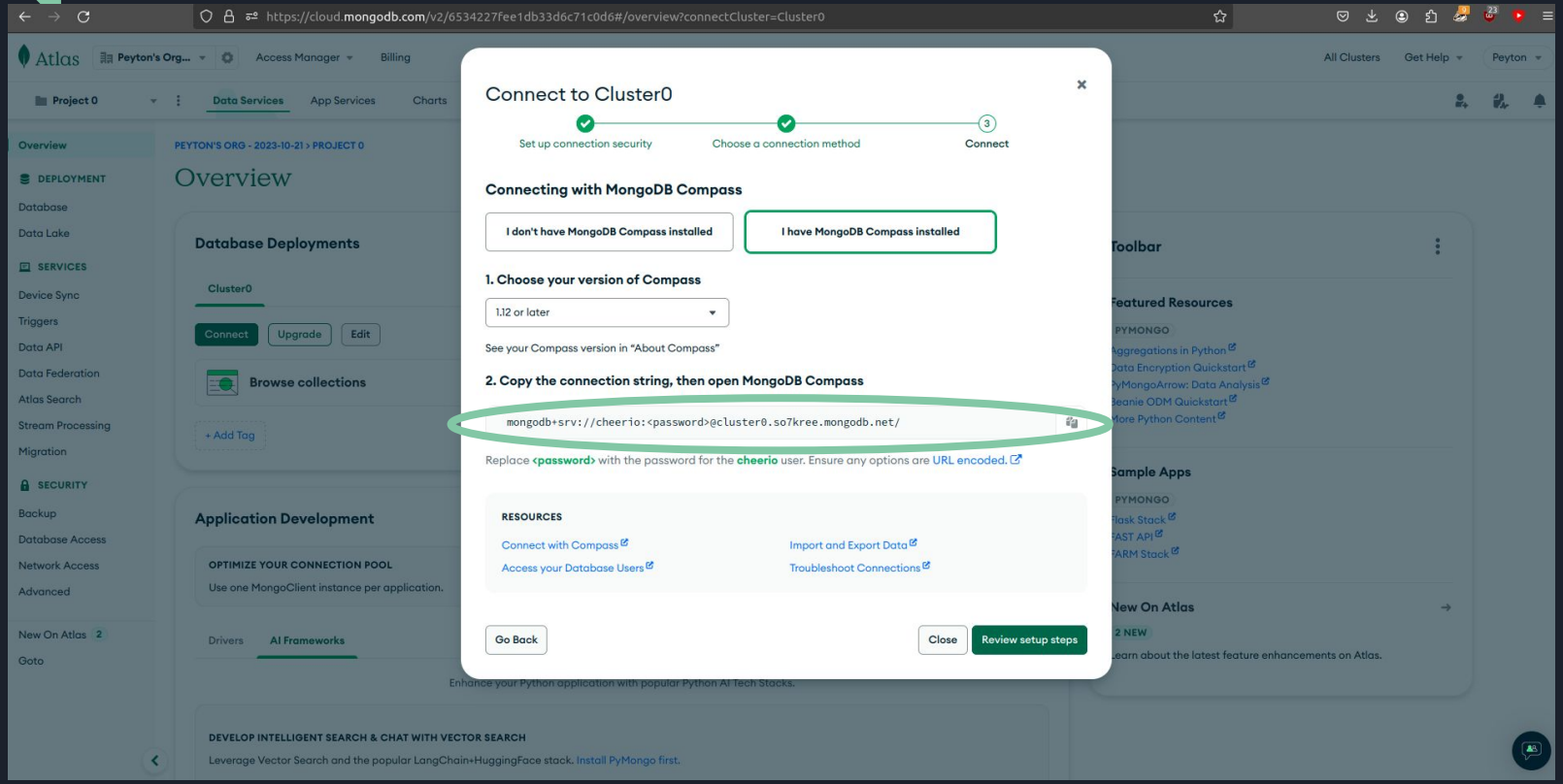
# Connecting Compass to Atlas



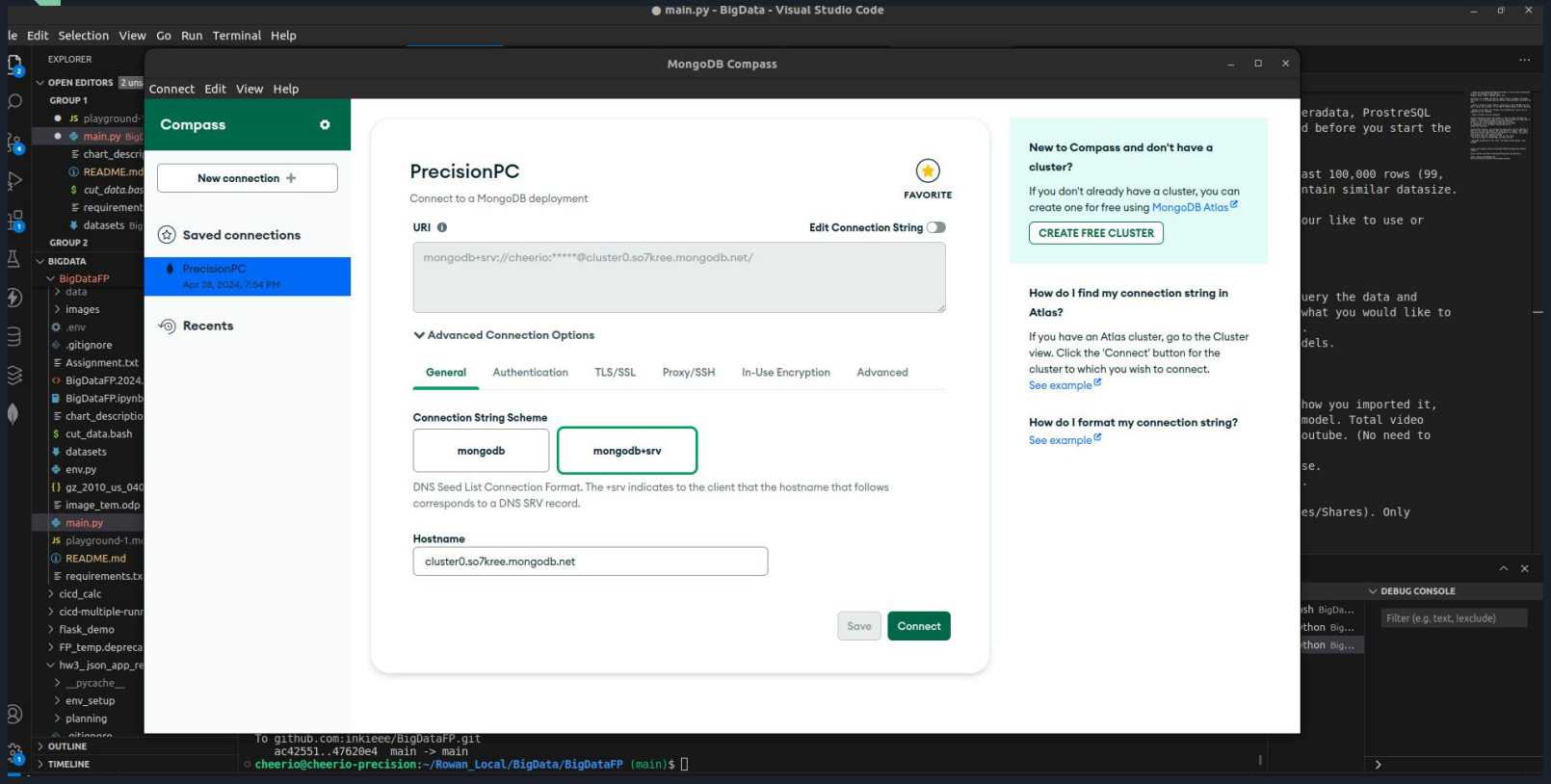
# Connecting Compass to Atlas



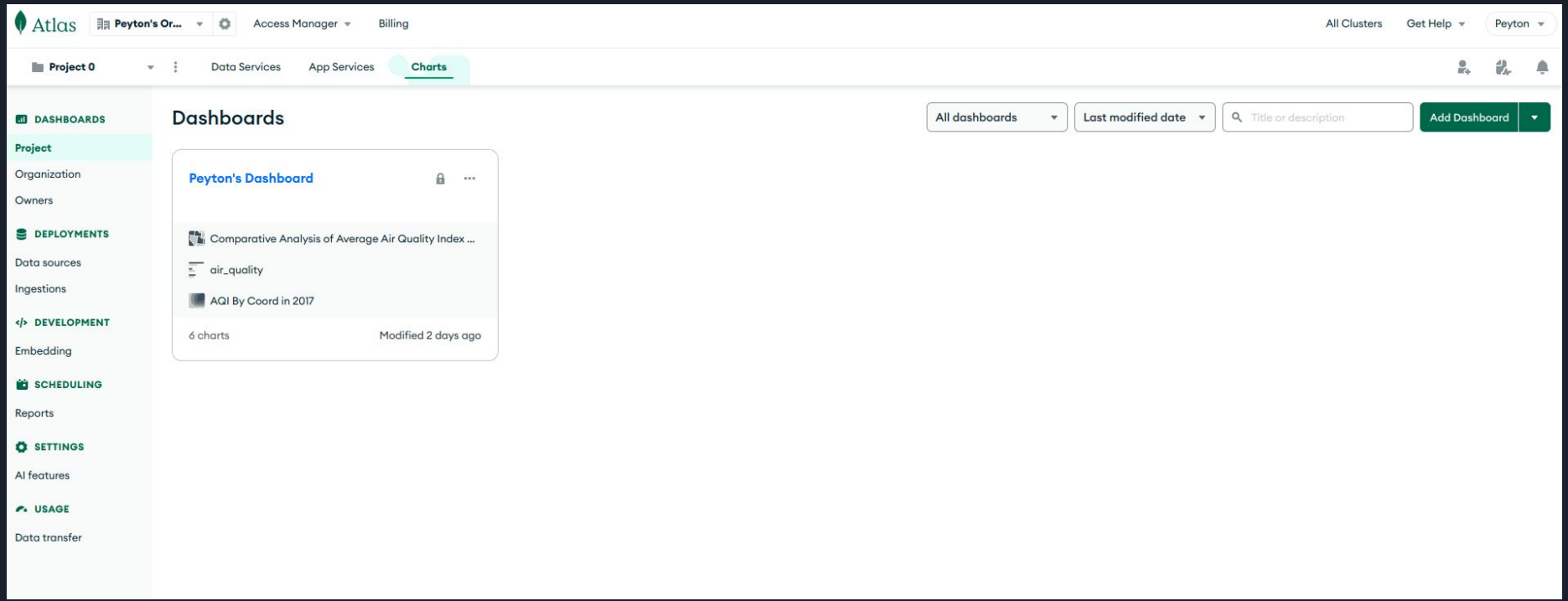
# Connecting Compass to Atlas



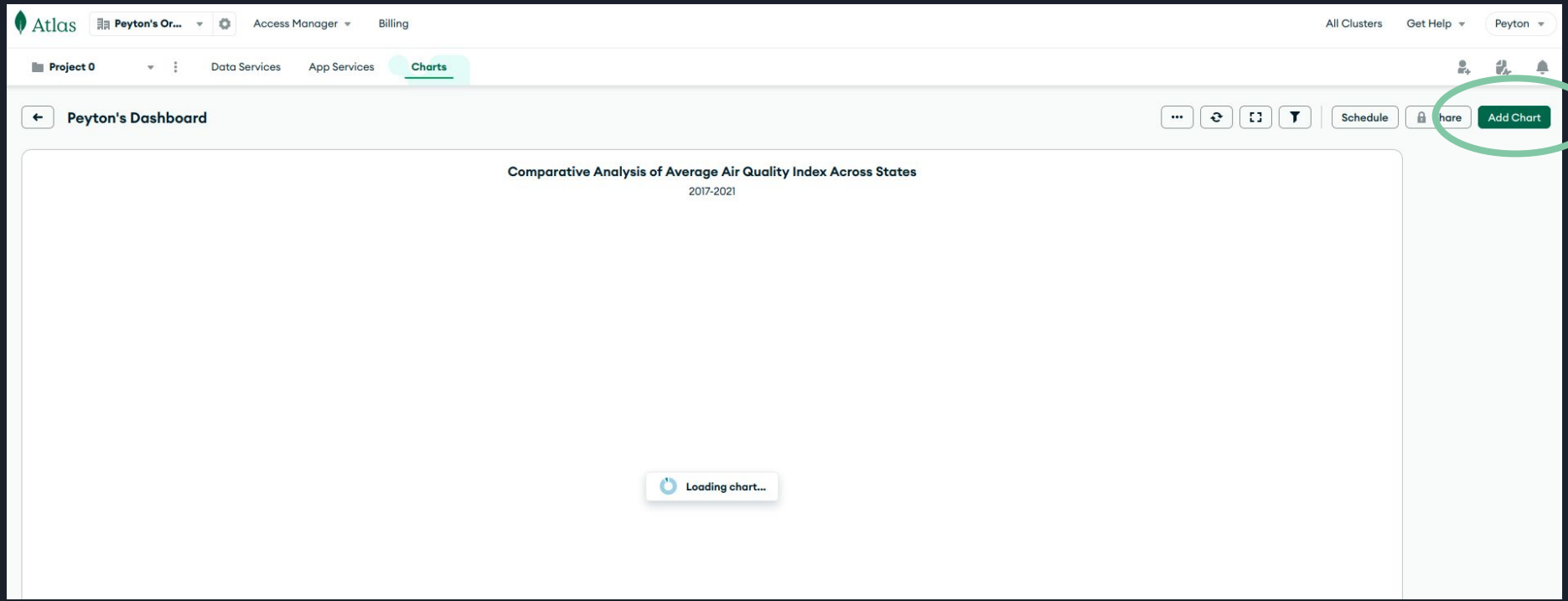
# MongoDB Compass Connection



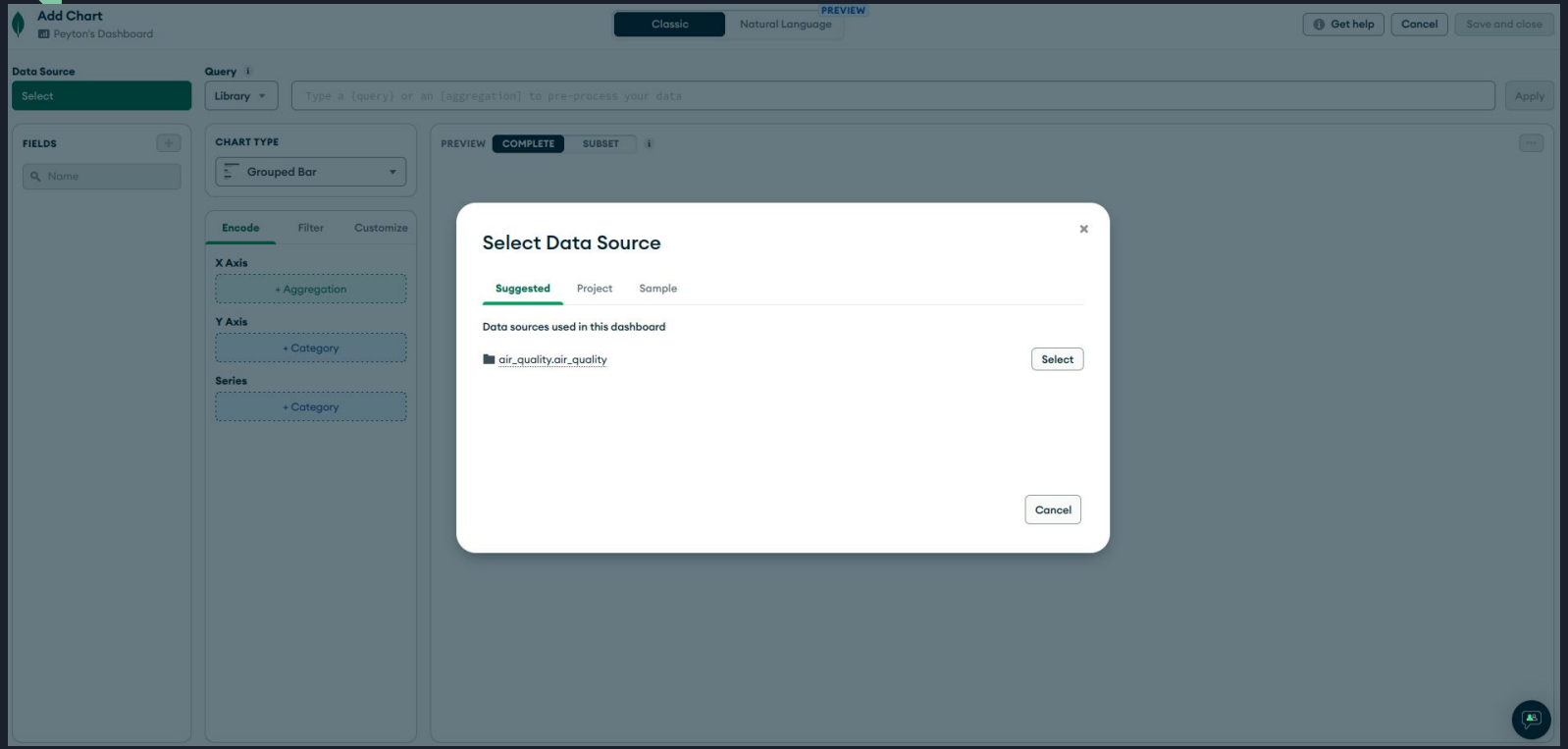
# MongoDB Charts



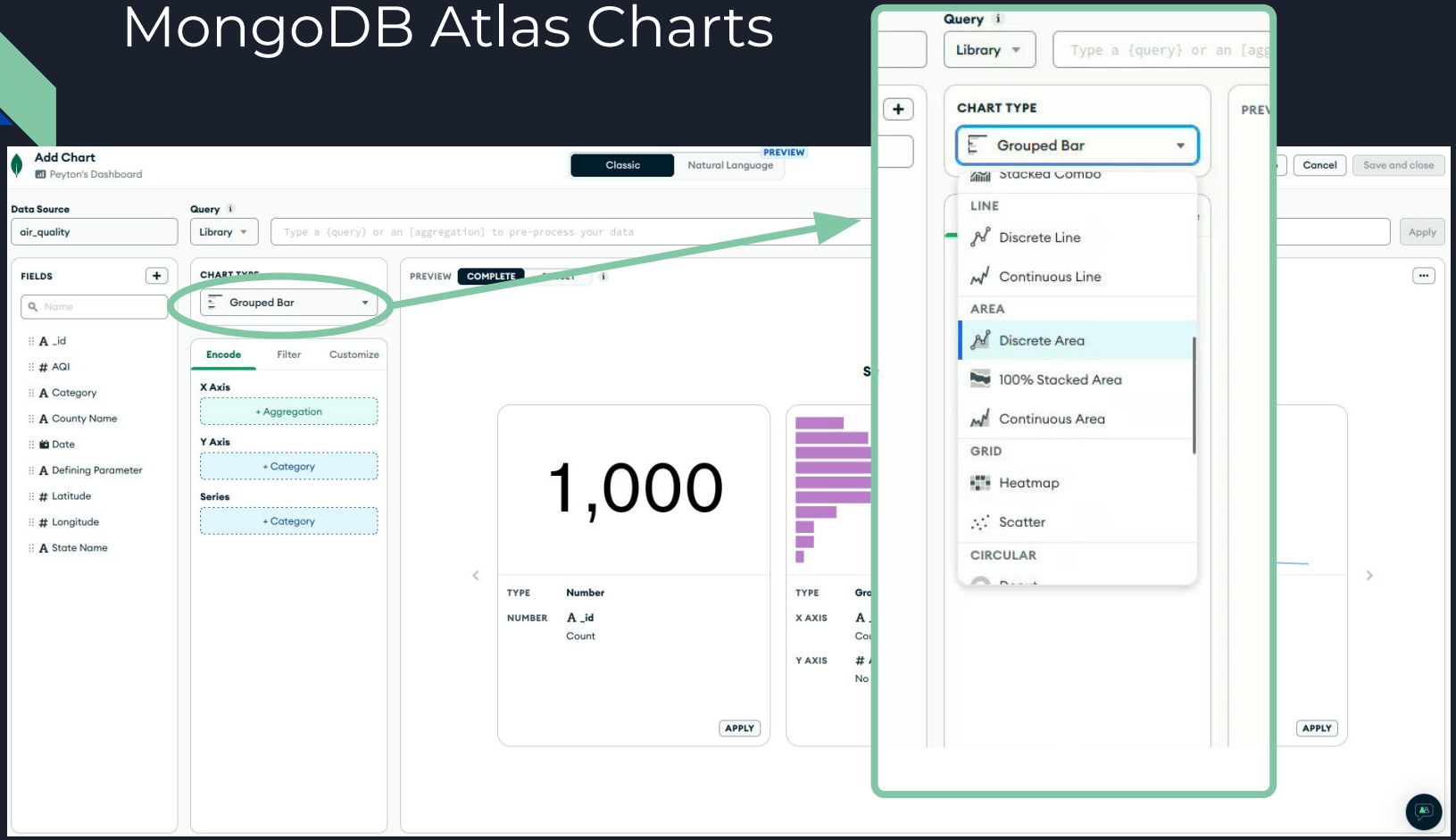
# MongoDB Charts



# MongoDB Atlas Charts

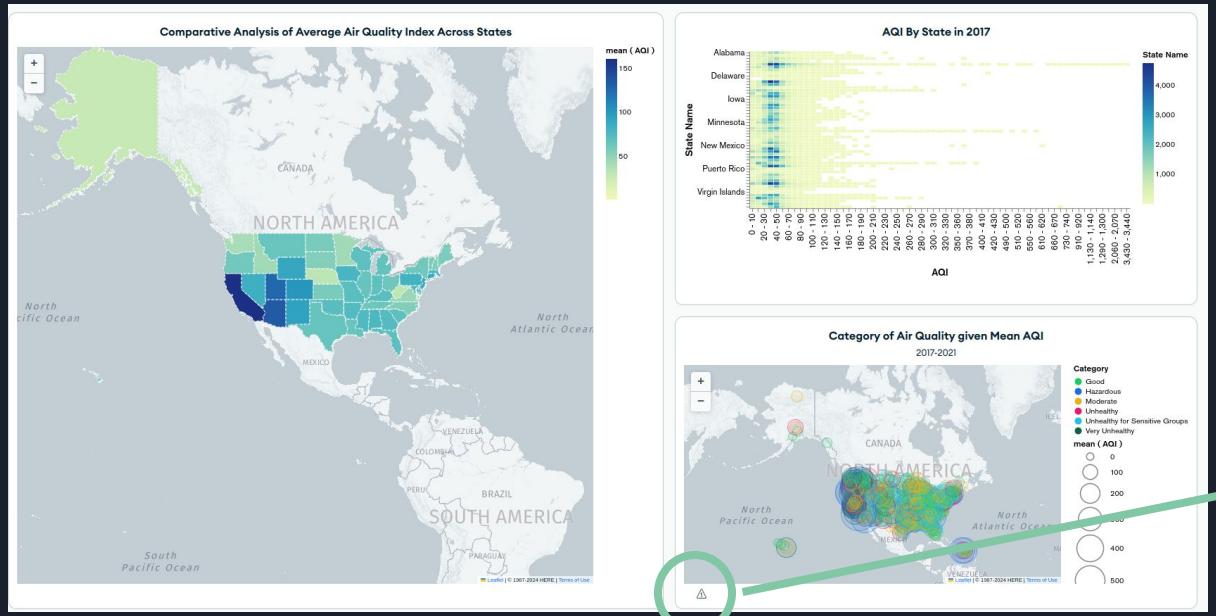


# MongoDB Atlas Charts

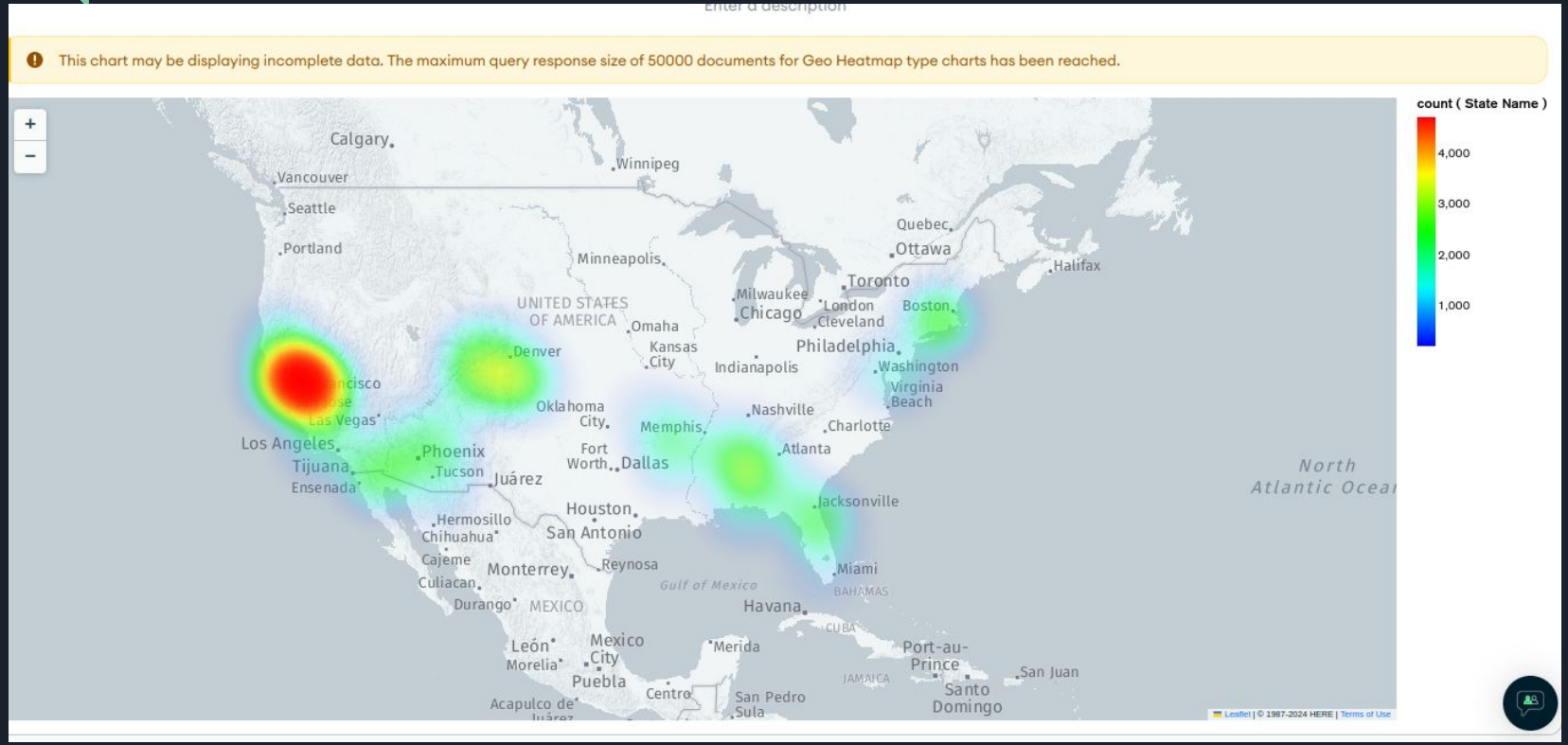


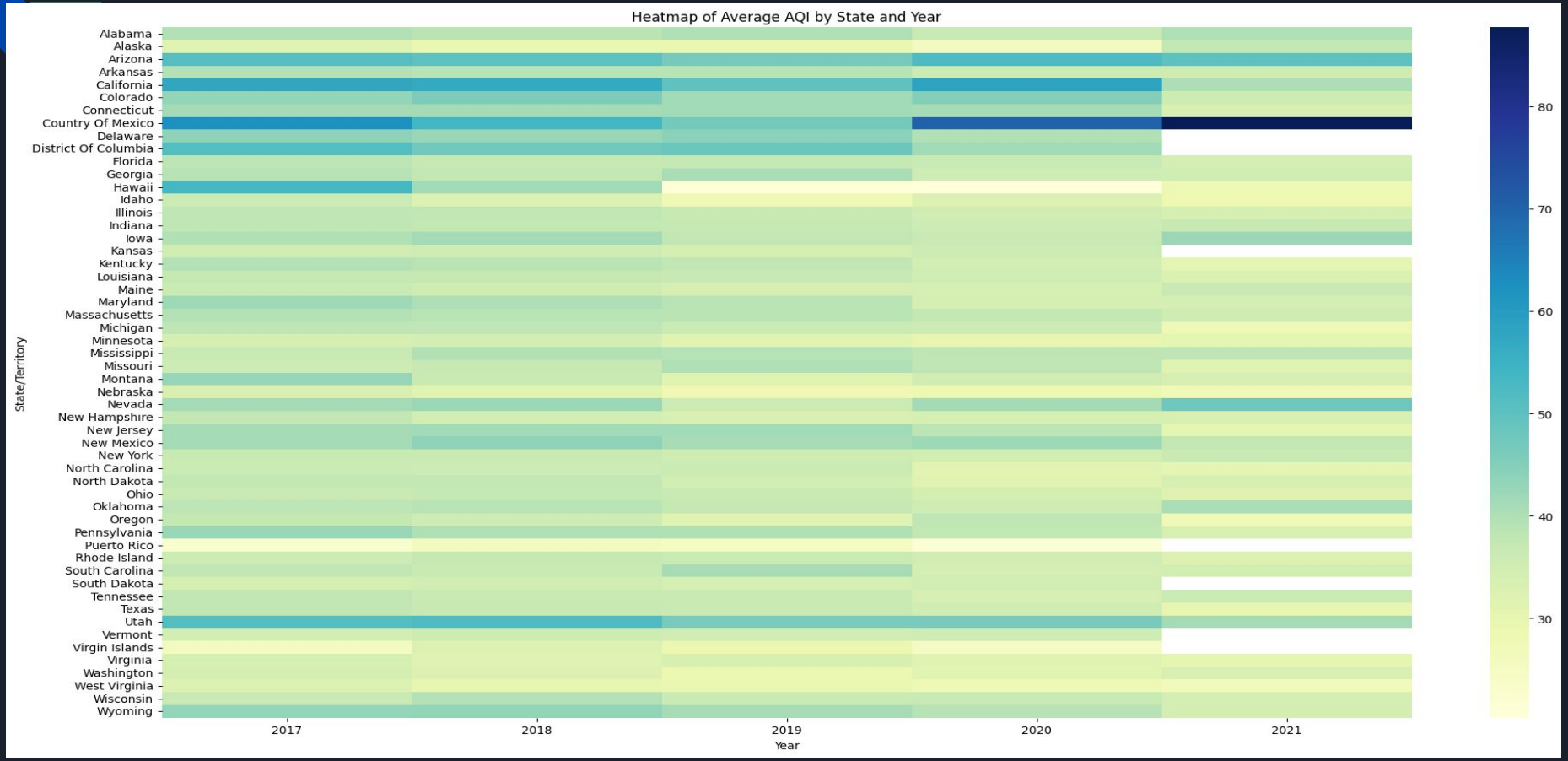
# MongoDB Chart Limitations

This chart may be displaying incomplete data. The maximum query response size of 1000 documents for Geo Scatter type charts has been reached.



# MongoDB Chart Limitations



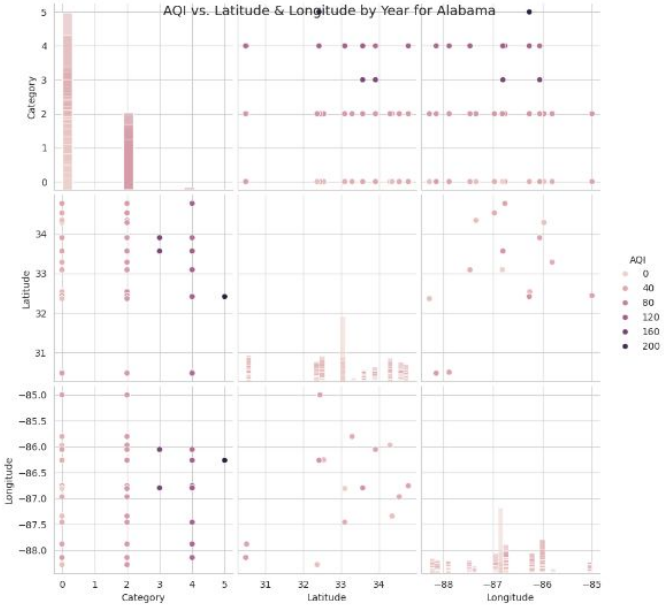

Heatmap of Average AQI by State and Year
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Country Of Mexico
Delaware
District Of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virgin Islands
Virginia
Washington
West Virginia
Wisconsin
Wyoming
State/Territory
2017
2018
2019
2020
2021
Year
80
70
60
50
40
30

```
new_data_predictions = model.predict(new_data)
predicted_categories = le.inverse_transform(new_data_predictions)
print(f"Predicted categories for new data: {predicted_categories}")
```

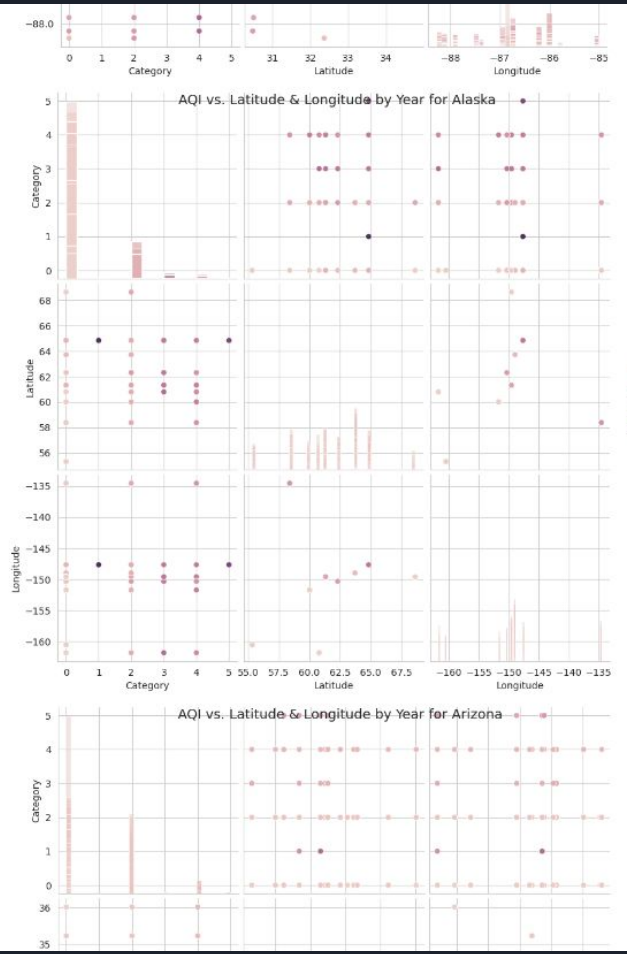
```
Model Accuracy: 1.0
Predicted categories for new data: ['Good' 'Unhealthy for Sensitive Groups' 'Very Unhealthy']
```

In [25]:

```
# Function to create a pairplot for a specific state
def state_pairplot(state_data):
  # Filter data for the specific state
  data_filtered = state_data[state_data['State Name'] == state]

  # Create a pairplot with AQI as hue variable (colors by AQI value)
  sns.pairplot(data_filtered, hue='AQI', height=3, diag_kind='hist')  # Updated call

  # Add a title with state name
  plt.suptitle(f"AQI vs. Latitude & Longitude by Year for {state}", fontsize=14)
  plt.show()

# Group data by state (assuming 'State Name' is the grouping variable)
state_groups = data.groupby('State Name')

# Iterate through each state group
for state, state_data in state_groups:
  state_pairplot(state_data.copy())  # Pass a copy to avoid modifying original data
```

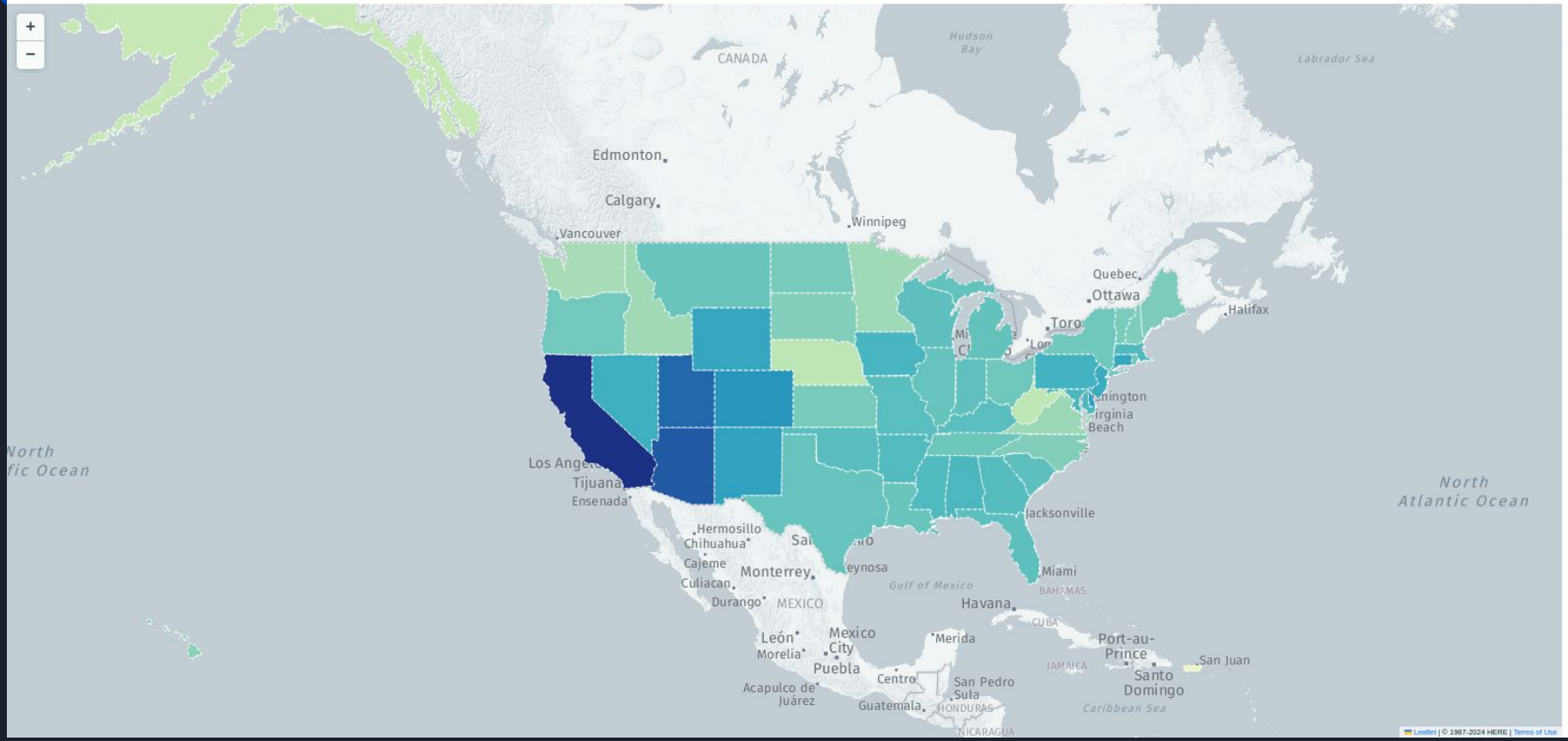






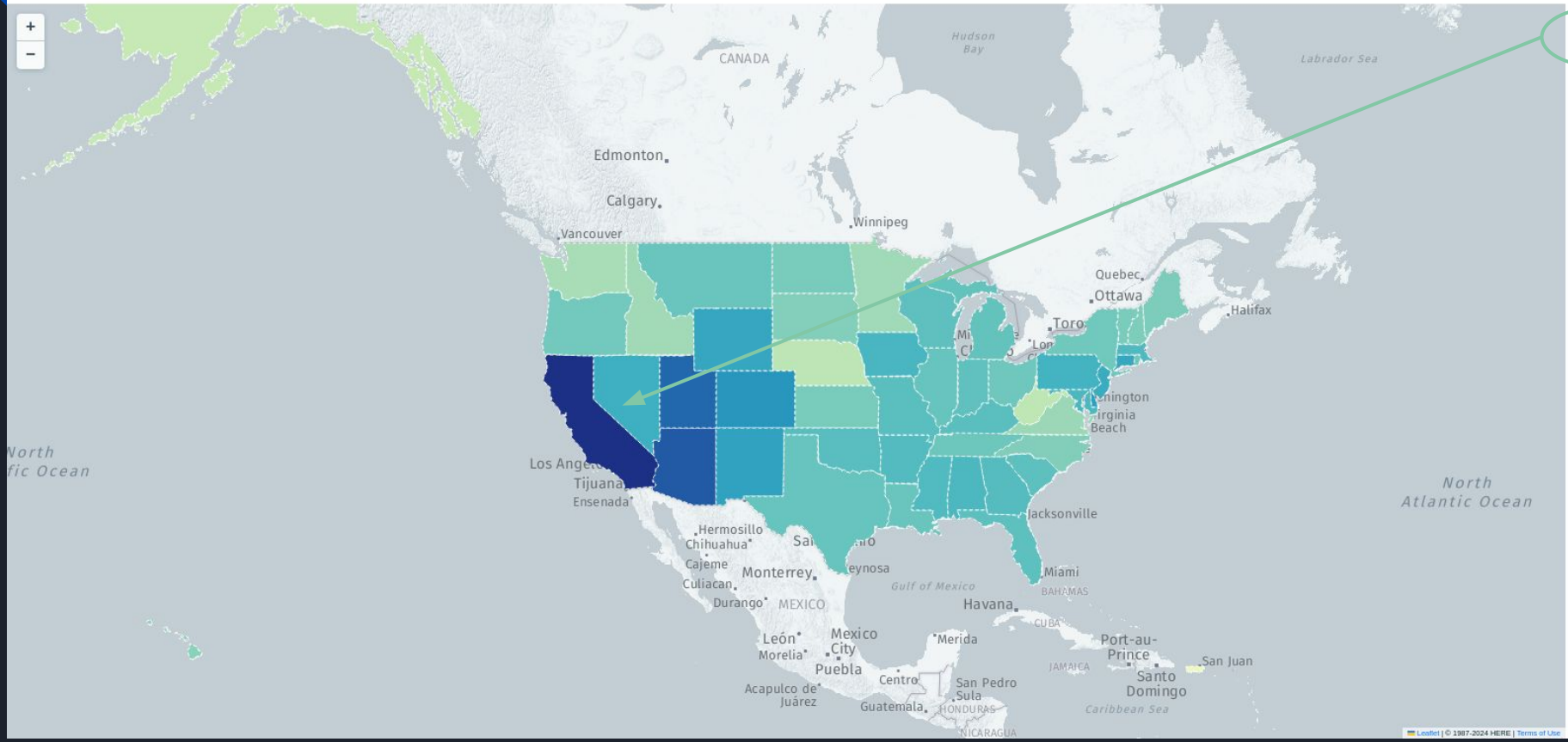
## Comparative Analysis of Average Air Quality Index Across States

2017-2021



**Comparative Analysis of Average Air Quality Index Across States**

2017-2021



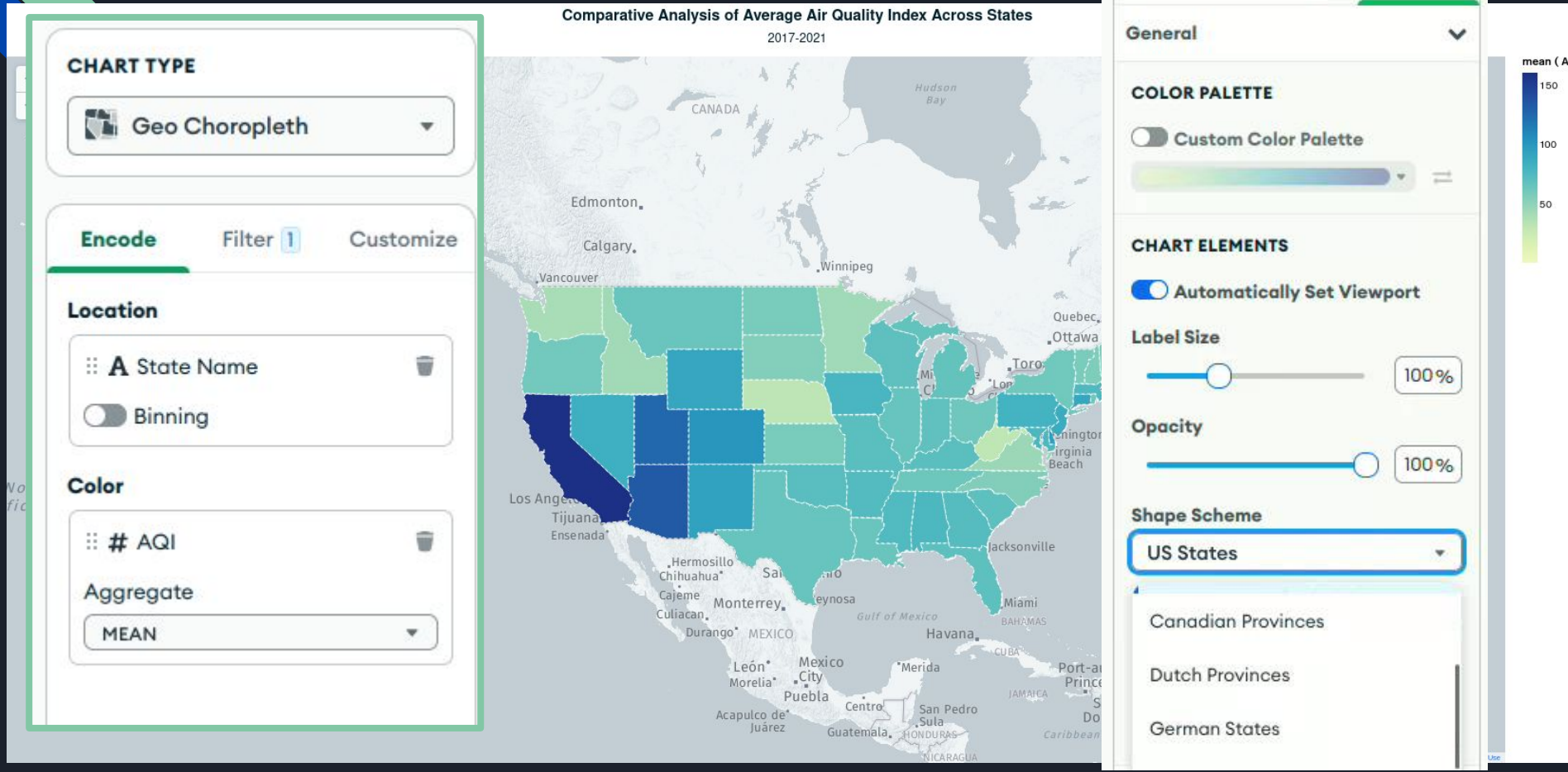
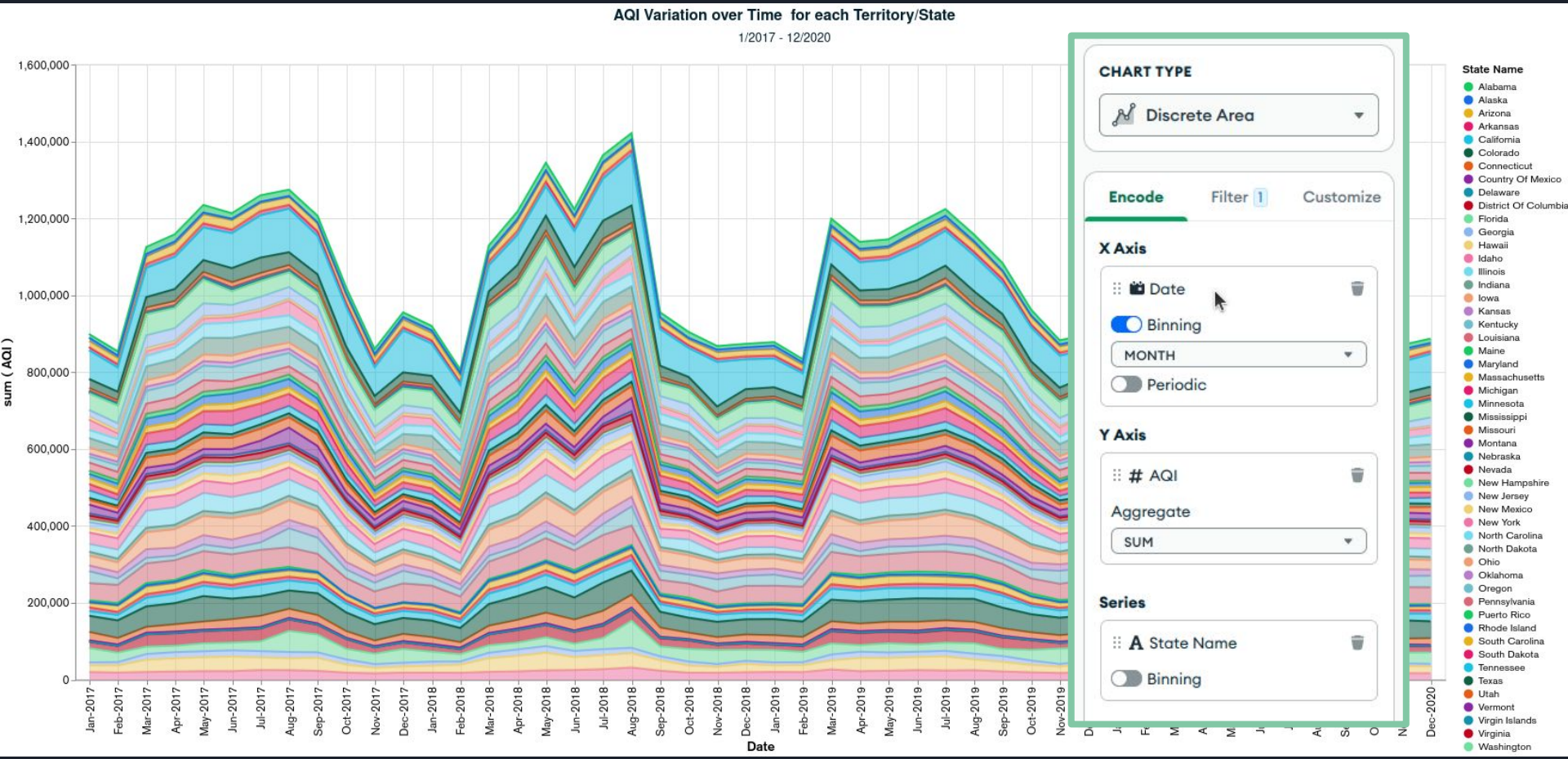
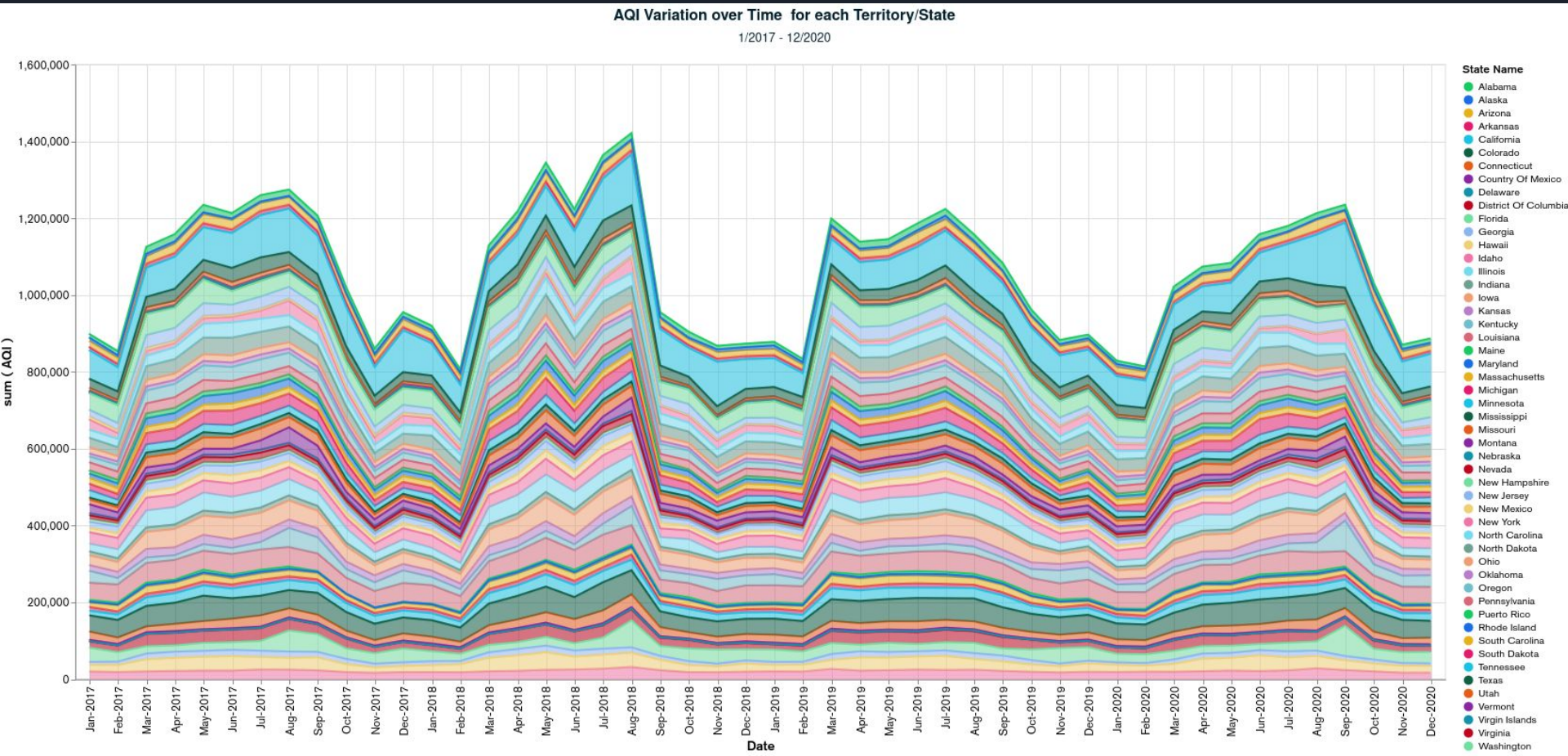
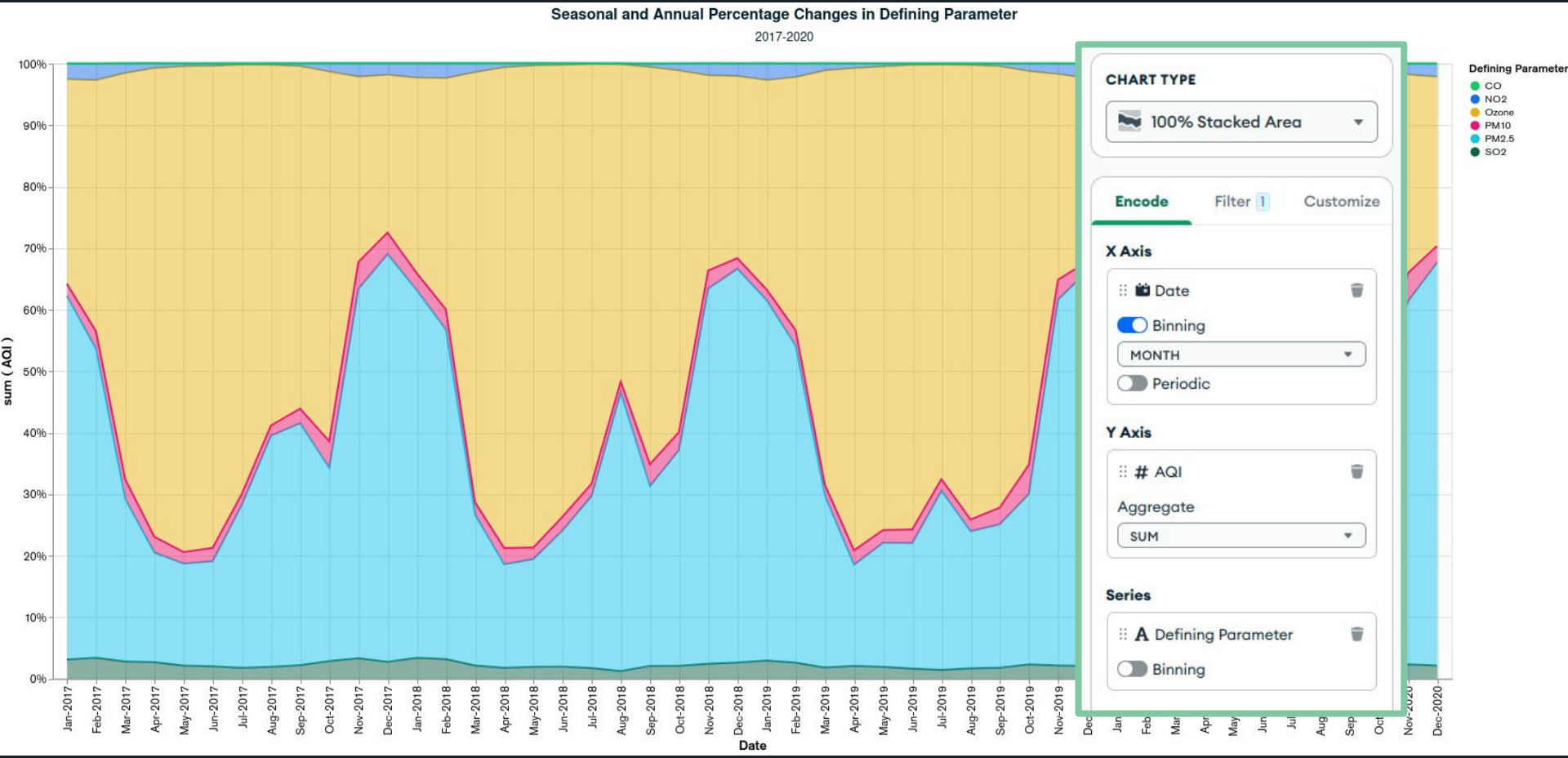
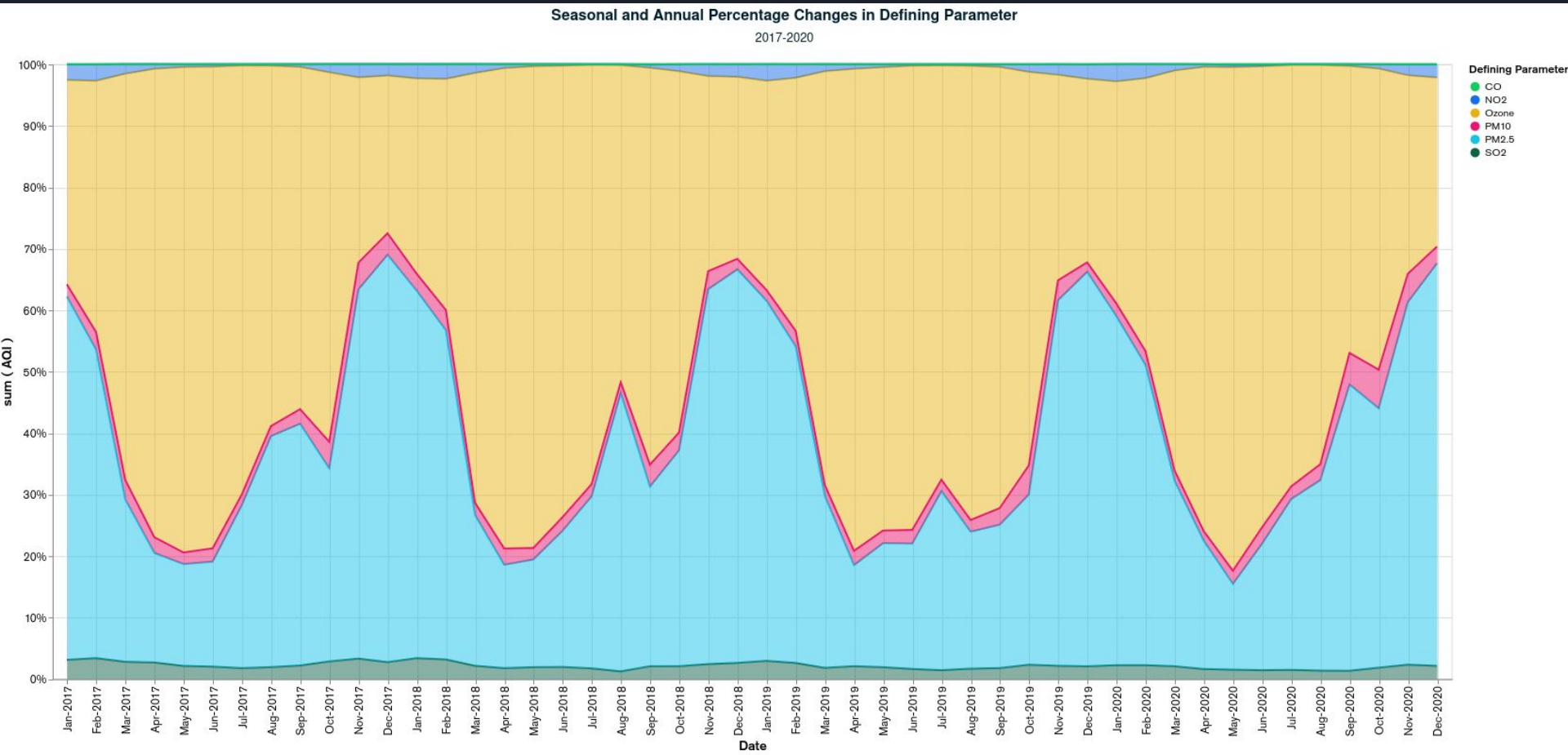


CHART TYPE
Geo Choropleth
Encode
Filter 1
Customize
Location
State Name
Binning
Color
AQI
Aggregate
MEAN
Comparative Analysis of Average Air Quality Index Across States
2017-2021
CANADA
Hudson Bay
Edmonton
Calgary
Vancouver
Winnipeg
Quebec
Ottawa
Toro
Los Angeles
Tijuana
Ensenada
Hermosillo
Chihuahua
Cajeme
Culiacan
Monterrey
Durango
MEXICO
Gulf of Mexico
Miami
BAHAMAS
Havana
CUBA
Jacksonville
León
Morelia
Mexico City
Puebla
Merida
Centro
Acapulco de Juárez
Guatemala
San Pedro Sula
HONDURAS
NICARAGUA
JAMAICA
Caribbean
Encode
Filter 1
Customize
General
COLOR PALETTE
Custom Color Palette
CHART ELEMENTS
Automatically Set Viewport
Label Size
100%
Opacity
100%
Shape Scheme
US States
Canadian Provinces
Dutch Provinces
German States
mean ( A
150
100
50


AQI Variation over Time for each Territory/State
1/2017 - 12/2020
sum ( AQI )
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Jan-2017
Feb-2017
Mar-2017
Apr-2017
May-2017
Jun-2017
Jul-2017
Aug-2017
Sep-2017
Oct-2017
Nov-2017
Dec-2017
Jan-2018
Feb-2018
Mar-2018
Apr-2018
May-2018
Jun-2018
Jul-2018
Aug-2018
Sep-2018
Oct-2018
Nov-2018
Dec-2018
Jan-2019
Feb-2019
Mar-2019
Apr-2019
May-2019
Jun-2019
Jul-2019
Aug-2019
Sep-2019
Oct-2019
Nov-2019
Dec-2020
Date
CHART TYPE
Discrete Area
Encode
Filter 1
Customize
X Axis
Date
Binning
MONTH
Periodic
Y Axis
# AQI
Aggregate
SUM
Series
A State Name
Binning
State Name
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Country Of Mexico
Delaware
District Of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virgin Islands
Virginia
Washington


AQI Variation over Time for each Territory/State
1/2017 - 12/2020
sum ( AQI )
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Jan-2017
Feb-2017
Mar-2017
Apr-2017
May-2017
Jun-2017
Jul-2017
Aug-2017
Sep-2017
Oct-2017
Nov-2017
Dec-2017
Jan-2018
Feb-2018
Mar-2018
Apr-2018
May-2018
Jun-2018
Jul-2018
Aug-2018
Sep-2018
Oct-2018
Nov-2018
Dec-2018
Jan-2019
Feb-2019
Mar-2019
Apr-2019
May-2019
Jun-2019
Jul-2019
Aug-2019
Sep-2019
Oct-2019
Nov-2019
Dec-2019
Jan-2020
Feb-2020
Mar-2020
Apr-2020
May-2020
Jun-2020
Jul-2020
Aug-2020
Sep-2020
Oct-2020
Nov-2020
Dec-2020
Date
State Name
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Country Of Mexico
Delaware
District Of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virgin Islands
Virginia
Washington


Seasonal and Annual Percentage Changes in Defining Parameter
2017-2020
sum ( AQI )
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Jan-2017
Feb-2017
Mar-2017
Apr-2017
May-2017
Jun-2017
Jul-2017
Aug-2017
Sep-2017
Oct-2017
Nov-2017
Dec-2017
Jan-2018
Feb-2018
Mar-2018
Apr-2018
May-2018
Jun-2018
Jul-2018
Aug-2018
Sep-2018
Oct-2018
Nov-2018
Dec-2018
Jan-2019
Feb-2019
Mar-2019
Apr-2019
May-2019
Jun-2019
Jul-2019
Aug-2019
Sep-2019
Oct-2019
Nov-2019
Dec-2020
Date
Defining Parameter
CO
NO2
Ozone
PM10
PM2.5
SO2
CHART TYPE
100% Stacked Area
Encode
Filter 1
Customize
X Axis
Date
Binning
MONTH
Periodic
Y Axis
# AQI
Aggregate
SUM
Series
A Defining Parameter
Binning


Seasonal and Annual Percentage Changes in Defining Parameter
2017-2020
sum ( AQI )
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Jan-2017
Feb-2017
Mar-2017
Apr-2017
May-2017
Jun-2017
Jul-2017
Aug-2017
Sep-2017
Oct-2017
Nov-2017
Dec-2017
Jan-2018
Feb-2018
Mar-2018
Apr-2018
May-2018
Jun-2018
Jul-2018
Aug-2018
Sep-2018
Oct-2018
Nov-2018
Dec-2018
Jan-2019
Feb-2019
Mar-2019
Apr-2019
May-2019
Jun-2019
Jul-2019
Aug-2019
Sep-2019
Oct-2019
Nov-2019
Dec-2019
Jan-2020
Feb-2020
Mar-2020
Apr-2020
May-2020
Jun-2020
Jul-2020
Aug-2020
Sep-2020
Oct-2020
Nov-2020
Dec-2020
Date
Defining Parameter
CO
NO2
Ozone
PM10
PM2.5
SO2

# Github Repository

https://github.com/inkieee/BigDataFP

# Like/Dislike

- Liked
  - Exposure to a lot of different tools
  - CICD
  - Distributed HA concepts
  - Various DB types
  - Example git repos

- Disliked
  - Required using up free trials to complete assignments
  - Couldn’t download gitbook easily

# Thanks!

# Following slides here for posterity

# GitHub Limitations

```
size of 50.00 MB
remote: error: Trace: 45002baf118dcfb1debbb79f1ed4104a1cf17e4fadb257724ef3eb4298a3e8ce
remote: error: See https://gh.io/lfs for more information.
remote: error: File data/air_quality/air_quality/aqi_daily_2011_to_2021.csv is 181.84 MB; this exceeds GitHub's file size limit of 100.0
) MB
remote: error: File data/air_quality/air_quality/aqi_daily_1980_to_2021.csv is 638.04 MB; this exceeds GitHub's file size limit of 100.0
) MB
```

# GitHub Limitations

“Additionally, had some issues with uploading document data into github for free for too large file sizes.”

# Grafana Attempt

```
(Reading database ... 215041 files and directories currently installed.)
Preparing to unpack grafana_10.4.2_amd64.deb ...
Unpacking grafana (10.4.2) ...
Setting up grafana (10.4.2) ...
Adding system user `grafana' (UID 135) ...
Adding new user `grafana' (UID 135) with group `grafana' ...
Not creating home directory `/usr/share/grafana'.
### NOT starting on installation, please execute the following statements to con
figure grafana to start automatically using systemd
 sudo /bin/systemctl daemon-reload
 sudo /bin/systemctl enable grafana-server
### You can start grafana-server by executing
 sudo /bin/systemctl start grafana-server
```

## Installing on a local Grafana:

For local instances, plugins are installed and updated via a simple CLI command. Plugins are not updated automatically, however you will be notified when updates are available right within your Grafana.

### 1. Install the Data Source

Use the grafana-cli tool to install MongoDB from the commandline:

```
grafana-cli plugins install grafana-mongodb-datasource
```

Enterprise only

# Mongo w/ Grafana

You have uncapped usage until May 12, 2024. Upgrade plans to continue using Grafana Cloud with unlimited, pay-as-you-go usage. Upgrade now

Looks exactly like free self hosted :v

## Plugins

Extend the Grafana experience with panel plugins and apps. To find more data sources go to Connections.

Search

mongo × Clear

MongoDB

By Grafana Labs

Enterprise

grafana-mongodb-datasource-2

Type: MongoDB

Settings | Permissions | Insights | Cache

Name: grafana-mongodb-datasource-2 Default

Before you can use the MongoDB data source, you must configure it below or in the config file. For detailed instructions, view the documentation.

*Fields marked with * are required*

### Connection

Connection string * mongodb+srv://cluster.host.net/dbname?retryWrites=1

Please enter a valid URL

### Authentication

Authentication methods

Choose an authentication method to access the data source

Credentials

User * User

Password * Password

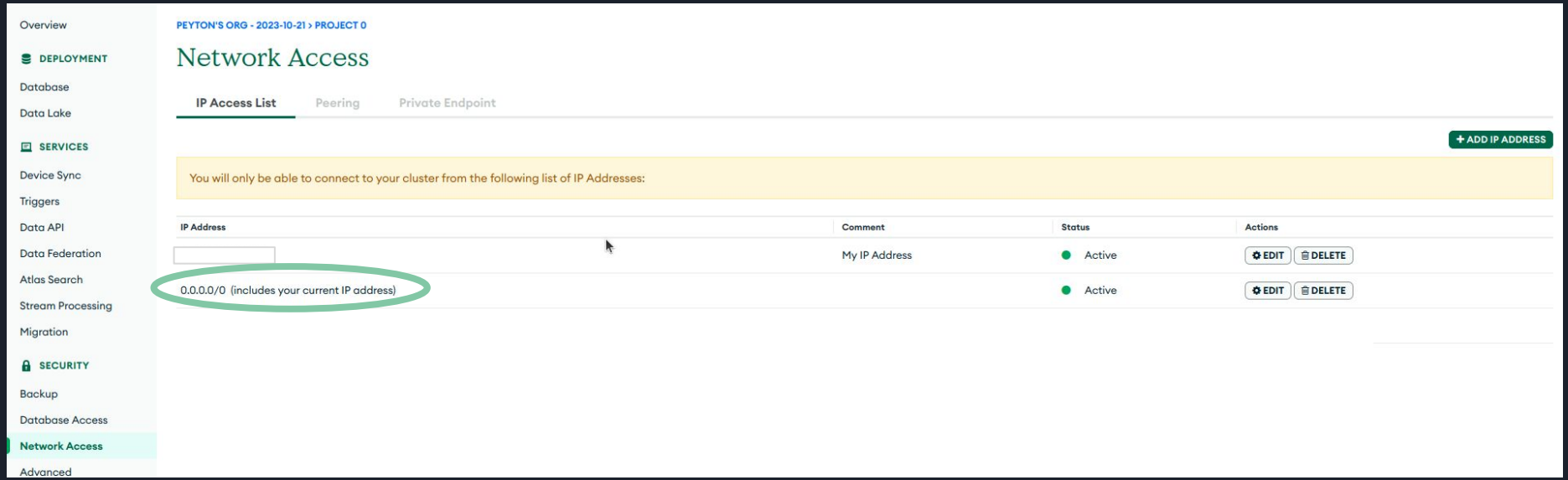
# Grafana Attempt

“

Initial project attempts included Grafana and Tableau, all of which were new tools to me, which I encountered other issues. For example Grafana requires a mongodb-plugin that is only available to Enterprise accounts of grafana, even when hosted locally from your machine, the plugin verifies against some enterprise key and if incorrect, will not allow mongodb connections. You can use the online Grafana SaaS service which also has it's own payment structure like Azure or MongoAtlas, but also requires an enterprise key. You can use a free trial for 2 weeks though.

”

# Mongo w/ Grafana

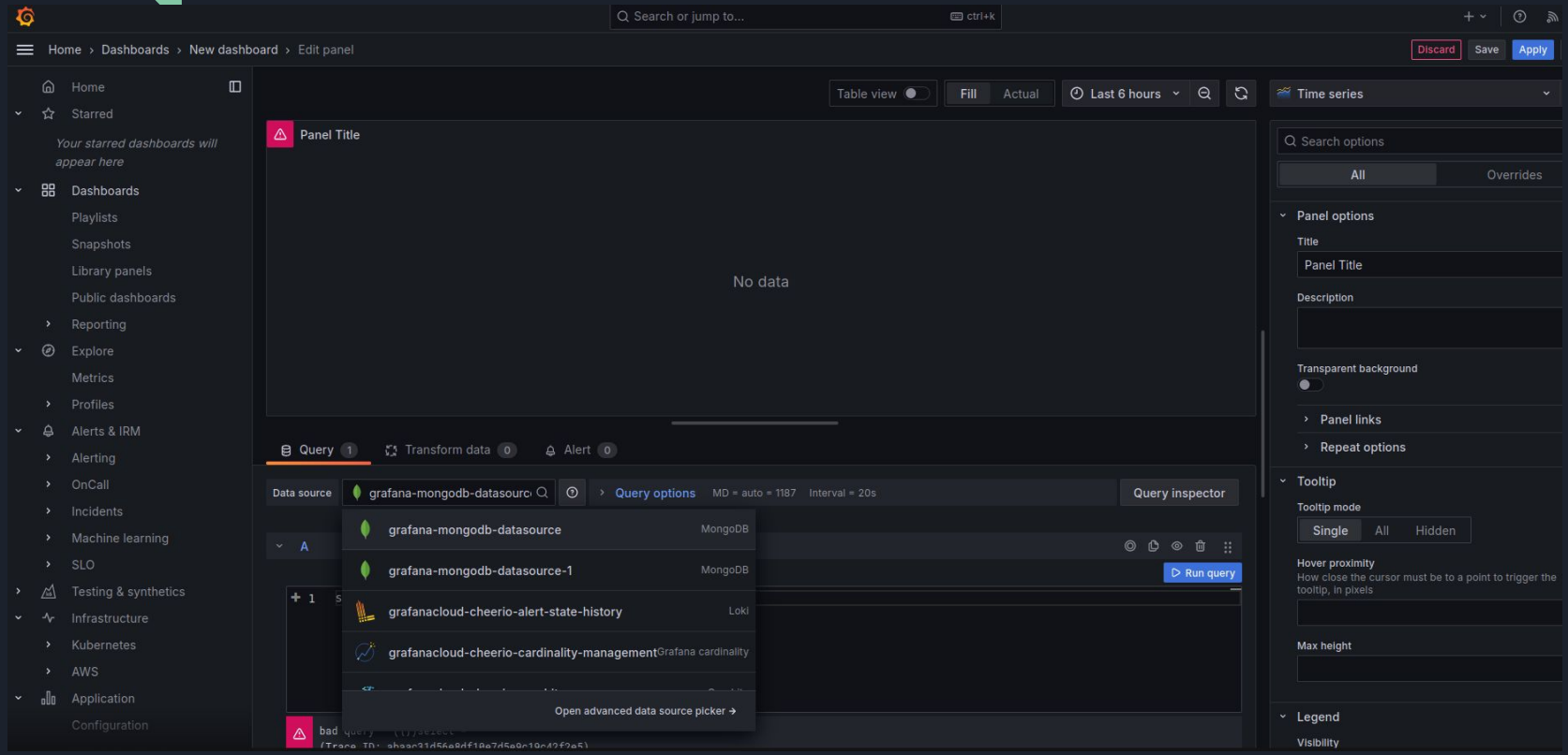


# Mongo w/ Grafana

“

Being hosted on Grafana servers though, we do have to modify our MongoDB network connection profile to be open to the internet (0.0.0.0 netmask 0) which allows all connections (including from grafana IPs).”

# Grafana issues w/ Cloud



# Grafana issues w/ Cloud

“

Annoyingly I had issues with generating queries of any type and decided not to invest too much time on something that would expire in 2 weeks anyway.

“