CSC12107 – INFORMATION SYSTEMS FOR BUSINESS INTELLIGENCE

Building and Mining Data Warehouse

Project 2024 – 2025

21HTTT2 – Ho Chi Minh City University of Science - VNUHCM

Instructors: MSc. Nguyen Ngoc Minh Chau, MSc. Tiet Gia Hong, MSc. Ho Thi Hoang Vy

Group 11

✦

Ho Chi Minh City, November 2024

Table of Contents

Contents

Introduction

- GitHub repository: https://github.com/kru01/ISBI_AirQualityDW
- Demo playlist: https://youtube.com/playlist?list=PLnfnJ5OTnHtOGSF_7IGO4POBs6xLR9M6w&si=PDYsodbme1MaR0Ba

Table 0. Student information and Task assignment table.

ID	Name	Task	Status (%)
21127004	Tran Nguyen An Phong	*Stage to NDS:* - Data cleaning and transformation. - Create NDS schema and db. - ETL for STATE. *Cube and OLAP:* - Plot all graphs in Power BI. *Data Mining:* - Create Mining schema and table. - Mine data in SSAS. *Report:* - Sect. 2.2.1 – Data cleaning and transformation. - Sect. 2.3 – DDS schema design and ETL. - Sect. 3.2, 3.3 – Dimensions. - Sect. 5.9 to 5.12 – Analysis. - Sect. 6 – Data Mining.	100
21127135	Diep Huu Phuc	*Source to Stage:* - Create METADATA and STAGE dbs. - ETL for 10_STATE_AQI. *NDS to DDS:* - Create DDS schema and db. - ETL for DIM_GEO. *Cube and OLAP:* - Build cube. - MDX for Q1 to 4, and 10. *Report:* - Sect. 4 – Scheduling. - Sect. 5 to 5.2 – MDX preface and Analysis.	100
21127296	Dang Ha Huy	*Source to Stage:* - ETL for 2B_USCOUNTIES. *NDS to DDS:* - Generate DIM_DATE. - ETL for FACT_AIR_QUALITY. *Report:*	100

		- Sect. 2, 2.1 – Preface, Source (.csv) to Stage. - Sect. 3, 3.1 – Measures and Calc Members. - Sect. 5.3 to 5.5 – Analysis.	
21127385	Pham Uyen Nhi	*Stage to NDS:* - ETL for COUNTY, and AIR_QUALITY. *Cube and OLAP:* - MDX for Q5 to 9, 11, and 12. *Report:* - Sect. 1 – Data Description. - Sect. 2.2.2 – NDS schema design and ETL. - Sect. 5.6 to 5.8 – Analysis.	100

Installation

Regarding our setup,

- **SQL Server 2019** and **SSMS 19**.
- **Visual Studio 2022** for **SSIS**, and **VS2019** for **SSAS** and **Data Mining**.
- For all server connections, we just use a "**.**" (**localhost**) and Windows authentication.

1 – After successful extraction, our project should have the following core components,

- **Data** – Folder storing the .csv(s), preserved exactly as when they are provided.
- **Docs** – Folder storing the Project Assignment and Report.
- **SQL** – Folder storing 5 scripts for warehouse creation, and one for deletion.
- **SSAS_Group11_2019** – Folder storing VS2019 solution for SSAS.
- **SSIS_Group11** – Folder storing VS2022 solution for SSIS.
- **OLAP_Visual.pbix** – Power BI Desktop file for analysis graphs and charts.
- **OLAP.mdx** – MDX query file for analysis questions.

2 – Run scripts in the **SQL** folder in the following order to set up the Data Warehouse,

1_create_metadata → 2_create_stage → 4_create_nds → 5_create_dds

- **3_clean_data.sql** *MUST* only be run after *completing ETL from Source to Stage*.
- **6_create_mining.sql** is for SSAS's Data Mining, *MUST* only be run after *the whole data warehouse is completed*.

3 – For **SSIS**, if your server is not on **localhost**, all the packages' **Connection Managers** must be changed. Otherwise,

- In **CsvSrc_Stage**, edit the **10_STATE_AQI - LOOP CSVs** and change the **Folder** path in the **Collection** tab.
- Also in the same package, reselect the flat file connection **2b_uscounties_csv**'s **File name** field.

4 – For **SSAS**, double-click **21BI11 DDS.ds** in the **Solution Explorer**, navigate to the **Impersonation Information** tab then change the Windows **User name** and **Password**.

- Be aware that this is the password associated with your **Microsoft account** (or the personal **Outlook** mail in our case), *not whatever PIN or password for the Windows Lock Screen*.

5 – For **Power BI**, ensure that the cube has been properly processed and deployed to SSAS, and if the server is not on base **localhost**, may need to alter the **Data Source**. Enable the following **Options**,

- **Security** → **Use Map and Filled Map visuals**.
- **Preview features** → **Shape map visual**.

1 Data Description

All .csv files under **Air Quality Data** folder possess identical data structures. Each entry describes the AQI value, measured by some criteria, in a state's county on a specific date. It comprises of,

- **State Name**, and **county Name** – (string) Name of the state, and county where the monitor resides.
- **State Code**, and **County Code** – (int) ID of said state, and county.
- **Date** – (string) The date that the measure was taken. Although this technically should belong to type **date**, we shall see clearer in data cleaning that the 2023 entries have erroneous formats.
- **AQI** – (int) The result that was calculated by the monitor.
- **Category** – (string) The six-category classification based on the ranges of AQI values. From Good (0 – 50), Moderate (51 – 100), Unhealthy for Sensitive Groups (101 – 150), Unhealthy (151 – 200), Very Unhealthy (201 – 300), to Hazardous (≥ 301).
- **Defining Parameter** – (string) The specific pollutant (e.g., ozone, PM2.5, PM10) that most heavily influences the AQI for that location and date.
- **Defining Site** – (string) If multiple sites, the ID (or name) of the one with maximum AQI value in the state's county.
- **Number of Sites Reporting** – (int) The number of monitors used, and by extension the number of sites.
- **Created**, and **Last Updated** – (datetime) The moment that the record is created, and last updated.

(2B)uscounties.csv is an unrelated compilation of *almost* (also explained in data cleaning) all counties in the US.

- **county**, and **county_ascii** – (string) Name of the county, and its ascii representation.

- **county_full** – (string) The full formal specification of the name.
- **county_fips** – (string) Five-digit ID of the county, only certainly unique within a state.
- **state_id** – (string) Two-letter ID name for the state, from USPS postal abbreviation.
- **state_name** – (string) Name of the state containing the county.
- **lat**, and **lng** – (float) Latitude, and longitude of the county.
- **population** – (int) An estimate of the county's population.

Because of different code pages between the files and our machines, to conserve as much originality as possible when translating type **string** to SQL, we prefer using **NVARCHAR** in the database and **Unicode string** during ETL's Data Conversion.

2 Data Warehouse

To facilitate Incremental Load and Auditing (e.g., tracking data sources, update time, etc.), we will maintain a database named **21BI11_METADATA**.

- **Three tables** corresponding to the phases (STAGE, NDS, DDS) with each row storing the phase's table name and its **LSET** (Last Successful Extraction Time). CET (Current Extraction Time) is not included since it can be gotten through GETDATE() during ETL.
- **SRC_SYS** table to define the source systems (the .csv files), and **SCD_STATUS** for state (inactive, active) when handling Slowly Changing Dimension. As cross-database foreign key is not possible, these tables are purely for reference, although we can still cross-database join to filter out invalid sources, or states, when needed.

2.1 Source (.csv) to Stage

21BI11_STAGE has two tables which are,

- **10_STATE_AQI** has the same columns as the .csv(s) in Air_Quality_Data and mimics their data types. Except, **DATE** is set as **NVARCHAR** to preserve data integrity, the main reason for this will be discussed further when cleaning data.
- **2B_USCOUNTIES** also shares the same columns with (2B)uscounties.csv but with the addition of CREATED and LAST_UPDATED during ETL.
- For tracing purposes, a **SOURCE_ID** column will also be added to both tables.

The ETL process goes through the following steps,

1. **Get LSET, CET** for the corresponding table from the metadata.
2. **Truncate the table** to clear old data and reset identity column.
3. **Load only new data** (in the time range between LSET and CET) from Source to Stage.
4. **Update LSET** metadata to CET for the table.

Firstly, with **10_STATE_AQI**, we use a **Foreach Loop Container** to iterate through the .csv(s). And because **Flat File Source** must be used for .csv, we can't extract data through a SQL

command (like with Excel Source). That's why a **Conditional Split** is employed to check if the data is new, and only then is it loaded. A **Data Conversion** node is also required for the data to be compliant with the Stage's schema. In addition, a counter variable is increased with each loop to provide the current file index for **Derived Column** to make SOURCE_ID.

As for **2B_USCOUNTIES**, it is much simpler since we only need to ensure compatible data types and the inclusion of CREATED, LAST_UPDATED, and SOURCE_ID.

2.2 Stage to NDS

2.2.1 Data cleaning and transformation

*Refer to **SQL/3_clean_data.sql** for the detailed process and implementation.*

To start off, **STAGE.10_STATE_AQI**'s 2023 entries have a **DATE** format of **DD-MM-YYYY**, while the rest is **YYYY-MM-DD**. With an **UPDATE** and some string manipulations, all the affected rows can be corrected. Next, there exist trailing whitespaces in certain columns. Specifically, all **COUNTY_NAME** rows of the state **Alaska**. Nevertheless, for good measure, every **NVARCHAR** column will be trimmed during the ETL flow.

Initially, the air quality data in **STAGE.10_STATE_AQI** has 194971 rows, but if we try to query **distinct** entries while ignoring SOURCE_ID, the number is now 194964. Taking a sample for confirmation, we can conclude that the 7 affected rows have exactly matching data, albeit from different sources.

	STATE_NAME	COUNTY_NAME	STATE_CODE	COUNTY_CODE	DATE	AQI	CATEGORY	DEFINING_PARAMETER	DEFINING_SITE	NUMBER_OF_SITES_REPORTING	CREATED	LAST_UPDATED	SOURCE_ID
1	Virginia	Fauquier	51	61	2022-03-29	7	Good	Ozone	51-061-0002	1	2022-03-29 16:30:00.000	2022-12-31 23:54:00.000	2
2	Virginia	Fauquier	51	61	2022-03-29	7	Good	Ozone	51-061-0002	1	2022-03-29 16:30:00.000	2022-12-31 23:54:00.000	3

Figure 221a. Query for STATE_CODE=51, COUNTY_CODE=61, and DATE=2022-03-29.

Since identical data does not offer any value, we will only retain one record, this can be addressed during ETL to NDS. As for **STAGE.2B_USCOUNTIES**, no duplication is found.

Now on the topic of mismatched data, we have verified that both 10_STATE_AQI's **CATEGORY** and **DEFINING_PARAMETER** conform with the AQI basics table (given in the Assignment), and WHO's pollution measures. However, there are many misclassifications between **AQI** and **CATEGORY**, e.g., rows with AQI < 201 but labeled as "Very Unhealthy," or AQI < 301 but as "Hazardous." For a thorough solution, all entries will be recategorized based on AQI basics table.

Likewise, certain **(State, County)** pairs from 10_STATE_AQI don't align with (or exist in) 2B_USCOUNTIES.

	STATE_NAME	COUNTY_NAME	STATE_CODE	COUNTY_CODE
1	Connecticut	Tolland	9	13
2	Connecticut	Windham	9	15
3	Illinois	Saint Clair	17	163
4	Virginia	Bristol City	51	520
5	Virgin Islands	St Croix	78	10
6	Virgin Islands	St John	78	20
7	Country Of Mexico	BAJA CALIFORNIA NORTE	80	2
8	Country Of Mexico	SONORA	80	26

Figure 221b. List of 10_STATE_AQI's (State, County) pairs that can't be identified in 2B_USCOUNTIES.

First, after investigating, we learn that **Connecticut**'s **Tolland** is part of the **Southeastern Connecticut Planning Region**, and **Windham** the **Capitol Planning Region**. These regions do exist in 2B_USCOUNTIES, and despite consisting of many small towns and cities, our data does not reach that level of detail. So, we might as well map Tolland and Windham to their respective regions' names.

Moving on, **Illinois**' **Saint Clair** should be changed to **St. Clair**. And **Virginia**'s "**Bristol City**" is the county's full name, but for 10_STATE_AQI, all the counties use shortened name, thus, it will be contracted to **Bristol**.

Lastly, the **Country of Mexico** is not a state of the US, therefore, all its entries will be removed during ETL to NDS. Yet, **Virgin Islands** is certainly one, so to prevent the loss of its AQI data leading to our skewed estimate of US's air quality, we consider manually creating data for **the state and its counties** in **2B_USCOUNTIES**.

- Data for **St. Croix** and **St. John** will be compiled from multiple sources. Still, the USPS doesn't have an abbreviation for Virgin Islands, since it lies outside the US's Customs Territory. Fortunately, we can use **VI** for its **STATE_ID** as there isn't a state denoted with that ID in 2B_USCOUNTIES.
- Since this new data is handcrafted, we will update **METADATA.SRC_SYS** to include a new entry recording which script is the origin.

The final **row count** for **NDS.AIR_QUALITY** should be,

194964 (unique records) – 1472 (of Mexico) = **193492**.

2.2.2 NDS schema design and ETL

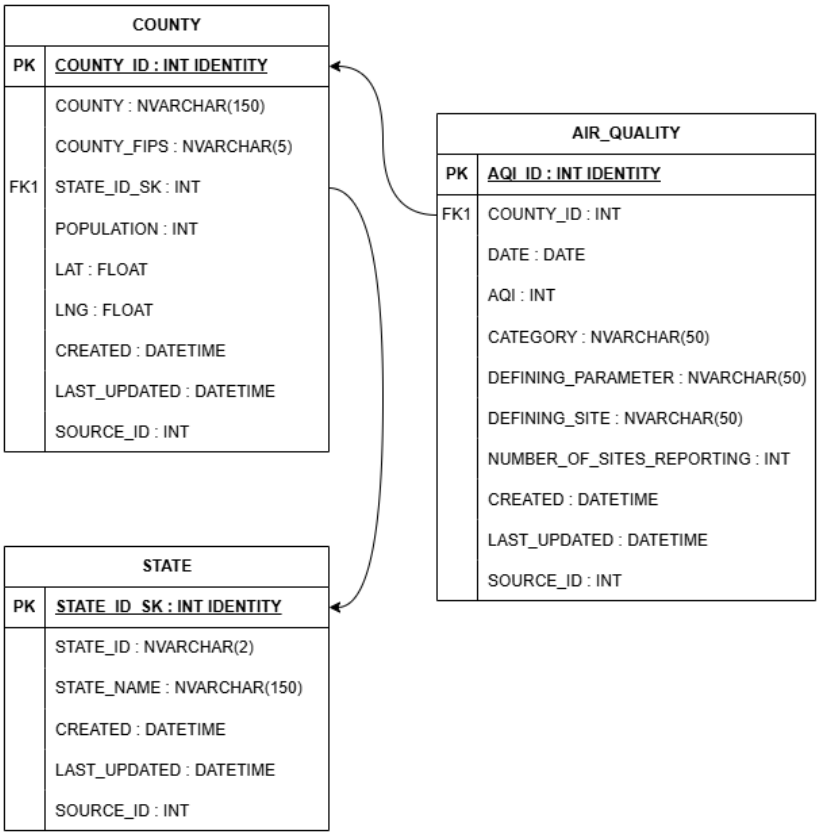


Figure 222a. Detailed NDS schema, with foreign keys and data types.

The schema can be derived from normalizing STAGE's to at least 3NF, removing certain superfluous columns, then attaching Surrogate Keys (SK).

- **STAGE.10_STATE_AQI**'s **STATE_CODE** and **COUNTY_CODE** are discarded since they don't correlate with anything in **STAGE.2B_USCOUNTIES**, and the latter's **STATE_ID** and **COUNTY_FIPS** together can uniquely identify a county.
- **STAGE.2B_USCOUNTIES**'s **COUNTY_ASCII** and **COUNTY_FULL** are also redundant as the first is identical to **COUNTY**, while the second adds very little information. Moreover, **COUNTY** alone is sufficient for it matches the shortened **COUNTY_NAME** that **STAGE.10_STATE_AQI** uses to designate a county.

Worth noting as well is that, in Fig. 222a, the absence of STATUS columns is because SCD isn't used during this phase. If a record exists in the NDS, it'll be directly updated. The ETL process must go from **STATE**, **COUNTY**, then **AIR_QUALITY** due to foreign key constraints. And we technically don't require **LSET** and **CET** (i.e., no incremental load) here, since **STAGE** doesn't store old entries and is truncated every time new data from **Sources** is "ETL in".

2.2.2a STATE

STATE will be populated by **STAGE.2B_USCOUNTIES**, we will first remove any repeating state (that might have come from different sources) by **sorting STATE_ID**, and trim trailing spaces through **Derived Column**. Then with **Lookup**, check (by its id) whether the state is already in **NDS.STATE**. If it does, we grab the **STATE_ID_SK** to update its entry, otherwise, with the new **CREATED** and **LAST_UPDATED** appended, the state is inserted.

2.2.2b COUNTY

Data for COUNTY also comes from **STAGE.2B_USCOUNTIES**, yet, a county is only unique within its state, i.e., by **(STATE_ID, COUNTY_FIPS)** alone can the county be identified. In addition, COUNTY has a foreign key to **NDS.STATE** but notice that, for the latter did come from the same STAGE's table, their STATE_IDs will be identical. This enables the use of **Merge Join**, for slightly better performance than **Lookup**, as extra rows won't be produced.

From **STAGE.2B_USCOUNTIES**, trimmed rows are sorted by (STATE_ID, COUNTY_FIPS) to filter duplicate counties, and prepare for merge join with sorted states from NDS.STATE. Joining will net us the foreign key **STATE_ID_SK**, which is next combined with **COUNTY_FIPS** for the **Lookup** to decide whether to **update-or-insert** process.

2.2.2c AIR_QUALITY

Unlike the previous two, AIR_QUALITY's entries are sourced from STAGE.10_STATE_AQI, if we attempt to merge join for the foreign key to **NDS.COUNTY**, surplus rows will arise due to both sides' inconsistent data.

From STAGE.10_STATE_AQI, trimmed rows are sorted by everything **except SOURCE_ID** to eliminate redundancy. Since here we don't have state's id and county's FIPS, a **Lookup** is performed with **(STATE_NAME, COUNTY_NAME)** on a joined table of NDS's STATE and COUNTY to get the corresponding **COUNTY_ID**.

Now, owing to the STAGE's table not having an obvious natural key, it proves difficult to judge what qualifier to use for the **update-or-insert Lookup**. However, after some inspections, we've determined that (STATE_NAME, COUNTY_NAME, **DATE, DEFINING_PARAMETER**) can uniquely identify any entry. And in AIR_QUALITY, (STATE_NAME, COUNTY_NAME) translates to solely **COUNTY_ID**.

2.3 NDS to DDS

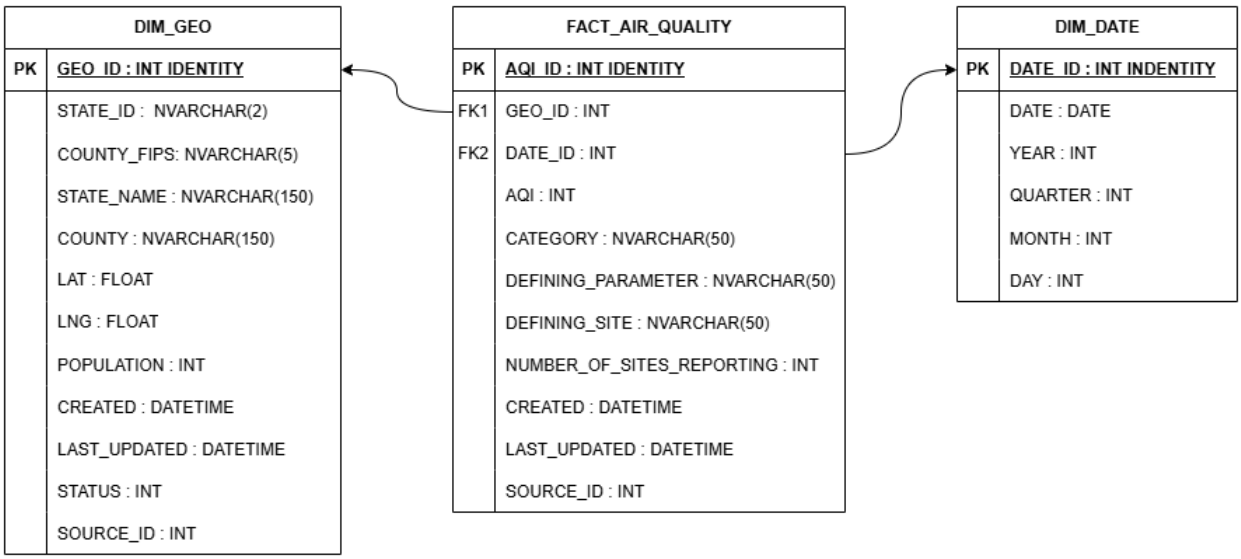


Figure 23a. Base DDS Star schema, with foreign keys and data types.

By comprehending requirements and de-normalizing NDS's, we arrive at this schema. It is designated as "**Base**" since here we don't include any potential **measure**, or **calculated member**. They will be documented later during Cube building, where we also delve into their related questions.

Currently, our DDS can satisfy both **Geography** and **Date** dimensions' hierarchies. Apart from the obvious merging of **STATE** into **COUNTY** to form DIM_GEO, we would like to note some degenerate dimensions (or **fact dimensions**) in FACT_AIR_QUALITY.

- **CATEGORY** and **DEFINING_PARAMETER**, per a lack of additional information, are put straight into the fact table to reduce joining works. They will be used to form the appropriate dimensions when configuring our cube.
- For **DEFINING_SITE**, despite technically also a fact dimension, is not demanded in any section of the project.

Additionally, the missing **STATUS** column in FACT_AIR_QUALITY is because we consider these facts to represent a moment in time, i.e., once an entry is recorded then it is final. To alter a past data would be highly uncommon.

2.3.1 DIM_GEO

The source for DIM_GEO is a joint of **NDS**'s **COUNTY** and **STATE**, where all rows with timestamp in the range of $(LSET, CET]$ and their relevant columns are extracted. Because changes are tracked separately for COUNTY and STATE, we must check if any of their **CREATED**s and **LAST_UPDATED**s satisfy.

Moving on to **Derived Column**, we append **STATUS** with value **1** (for SCD) and reassign LAST_UPDATED to be the latest date between COUNTY's and STATE's. To clarify, for this dimension, every row essentially represents a specific county (of a state), so if this county is updated in the NDS, only one DDS record is modified accordingly. But if it's an NDS state, then we want to alter all related DDS rows (i.e., all counties of that state).

Finally, we configure the **SCD** node with business keys **(STATE_ID, COUNTY_FIPS)**, and the following attributes,

- **Historical**: STATE_NAME, COUNTY.
- **Changing**: LAT, LNG, POPULATION.

For changing attributes, outdated records won't be updated, and for historical, a STATUS of **1** signifies **current** and **0 expired**. Inferred member support is left enabled (as default) for potential Early Arriving Facts.

2.3.2 DIM_DATE and FACT_AIR_QUALITY

Considering the time gaps that will appear if **DIM_DATE** is only made up of dates from Sources, leading to inflexibility, we choose to populate it through a stored procedure **USP_DIM_DATE_GEN**. This procedure, declared and ran along with the DDS's creation, generates all dates within 2 given years, breaks them down into date parts (year, quarter, month, day), and inserts the resulting rows into DIM_DATE.

When querying new rows (i.e., those having CREATED or LAST_UPDATED in $(LSET, CET]$) from **NDS.AIR_QUALITY** for the fact table, we must join it with **COUNTY** and **STATE** to also attain the **COUNTY_FIPS** and **STATE_ID**, both of which will form a pair to be used for looking up **GEO_ID** from DDS.DIM_GEO. Then similarly, we use **DATE** to look up **DATE_ID** from DDS.DIM_DATE and insert the completed entry.

3 SSAS Cube

When first initialized from the DDS, our Cube only has 2 valid measures which are,

- **AQI** – From the original AQI column of FACT_AIR_QUALITY, this serves a base to make other measure and is almost never used directly in query.
- **FACT AIR QUALITY Count** – Automatically generated by SSAS, it is simply a row count of the fact table. Since every entry in fact belongs to a date, the count can also double as a **Count of Days**, which appears in analysis questions (Q) 3, 4, 11, and 12.

Working through the requirements, we start to formulate the following, through the **New Measure** wizard,

- **Minimum** and **Maximum AQI** (Q1, 9) – Derived from AQI by taking min and max.

- **SUM AQI** and **SUM AQI SQUARED** – Derived from AQI and AQI_SQUARED by summing. Both are purely supports for **STDEV AQI**, which will be explained shortly.

3.1 Calculated Members

Q2, 3, 5, 6, 8, 9, and 10 all demand the AQI mean, but after ascertaining that using the **Average over time** does not produce desired results, we opt to manually compute **AVG AQI** by simply dividing **SUM AQI** by **FACT AIR QUALITY Count**, making sure to set the output as NULL if the count is 0 as well.

Now, **STDEV AQI** for Q2, and 9 is a tricky one since although the Stdev function does exist, not only is it operationally expensive but also requires a fixed Set_Expression, meaning no context-based filtered querying. To avoid these complexities, we choose to just apply the standard deviation formula,

$$\sqrt{\frac{\sum(X^2) - \frac{(\sum X)^2}{N}}{N}} = \left(\frac{SUM\ AQI\ SQUARED - \frac{SUM\ AQI \times SUM\ AQI}{Count}}{Count} \right)^{0.5}$$

Notice that we have **SUM AQI** and **FACT AIR QUALITY Count**, and power of 0.5 as there is no square root function, but we are still lacking **SUM AQI SQUARED**. For it is impossible to make a new measure from a calculated member, the sole route left for us is adding an **AQI_SQUARED** column to **DDS.FACT_AIR_QUALITY** and updating the **NDS_DDS** ETL package duly. Afterwards, the cube is reprocessed, and the New Measure wizard is once again employed to make SUM AQI SQUARED. Of course, we also treat the result when the count becomes 0 as NULL.

3.2 DIM GEO and DIM DATE

In **DIM GEO**, we define **Hierarchy GEO**'s level going from **STATE NAME** to **COUNTY** and link the **Attribute Relationships** suitably, their **RelationshipType**(s) are put as **Rigid**. For each attribute's properties, the **KeyColumns** is arranged to conform with the hierarchy, and **NameColumn** is set to said attribute's name.

Likewise, all steps are repeated for **DIM DATE** and **Hierarchy DATE** (with the level of YEAR, QUARTER, MONTH, DAY, and DATE). Furthermore, we expect SSAS to recognize this as a Time Dimension, so the property **Type** for DIM DATE is switched to **Time** and the attributes to the appropriate type that each represents.

Lastly, to tackle Q10, **DAYLIGHT_SAVING** is covered in this dimension, we also add the corresponding column to **DDS.DIM_DATE** with the **BIT** type and amend the date generating procedure to provide it data, being 1 (True) between Mar. 12th, 2023, and Nov. 5th, 2023.

3.3 Fact (Degenerate) dimensions

As mentioned in Sect. 2.3, out of the 3 dimensions, CATEGORY, DEFINING_PARAMETER, and DEFINING_SITE, only the former two are asked during analysis. Thus, they ascend to standalone dimensions **DIM CATEGORY** and **DIM DEFINING PARAMETER** with the key columns and attributes being merely themselves, and no related table.

4 Scheduling

Scheduling can be carried out with the help of **SSIS Catalog** and **SQL Server Agent**. Back in SSIS, we've created a new **Controller** package to guide the flow, starting from executing the packages **Stage_NDS** to **NDS_DDS** and ending with **processing** the whole **SSAS** database, through the Execute Package Task and Analysis Services Processing Task nodes.

- The exclusion of **CsvSrc_Stage** can be attributed to us defining Stage as a storage for messy and inconsistent data, compiled from many sources, before getting manually transformed, and prepared for loading to inner data stores. In that subsequent sources may vary in data quality, we should not automate this process.

After, the entire SSIS project is deployed to a premade catalog **SSISDB** in Integration Services Catalogs. We then make a new Job in SQL Server Agent with a singular step pointing to the Controller package, which now lies in the SSIS Catalog's SSISDB. Our job is scheduled to run Weekly at 12AM on Sunday.

Now, prior to starting the job, one important point is granting SSAS's **Process database** and **Read definition** privileges to the **NT SERVICE\SQLSERVERAGENT** as it will be the account attempting to process the Cube.

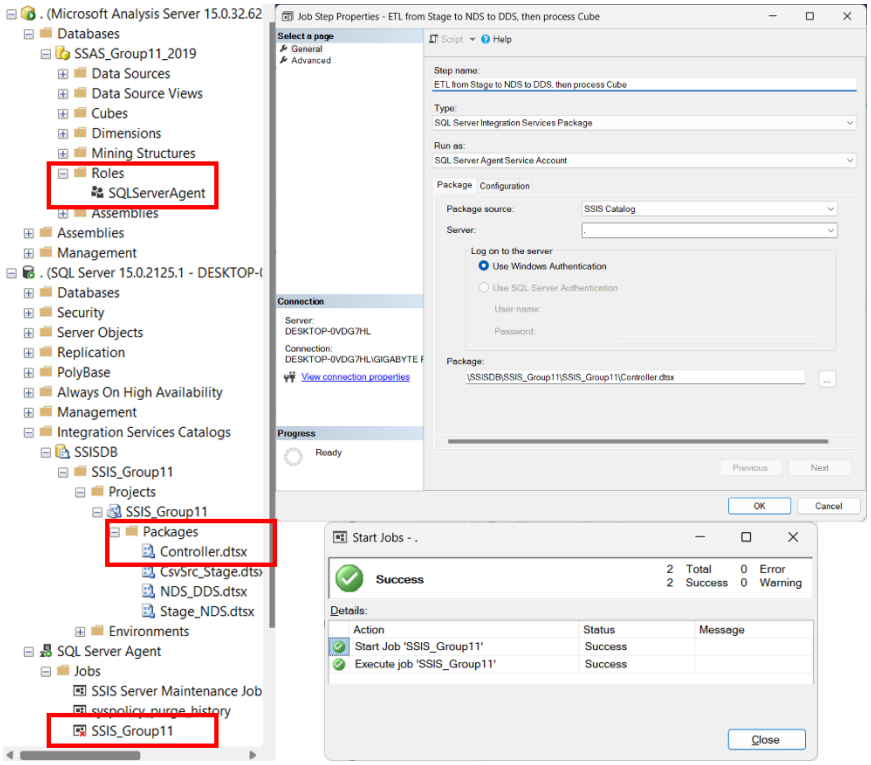


Figure 4. Left – All correctly setup components (in red) in SSMS's Object Explorer. Top Right – Configurations of the Job's ETL and Cube process step. Bottom Right – A successful run of the Job.

5 OLAP and Reporting

Our **MDX** queries are **elementary** and **self-documented** since they **only call** upon the **components** discussed during Cube building and **don't carry** any **complex operation**. They are simply **to verify the cube** and serve as **templates for graphs**. Besides, the **raw results** are too **verbose** and **obscure** for analysis. Hence, we will only provide an **MDX figure** in each section **for reference** but not delve into it; the **actual studies** are **conducted on the graphs**.

5.1 AQI's Min and Max of states during each year's quarters

> "Report the **min and max** of AQI value for each **State** during **each quarter of years**. *Analysis hints:* How do the AQI values fluctuate during the year? Pay attention to the values (max, min). Are any unusually large or small?"

In **Fig. 51b**, **most states** have their **maximum AQI** values **below 250** throughout the year, **except** for **Arizona** and **California**. California often surpasses this threshold by a large margin. Only a **few states** have **minimum AQIs greater than zero**, but at the same time, their maximum AQIs are among the smallest.

The **AQI** values **vary** for **each state**, yet the **dataset does not cover all quarters** of a **year** for **some states**. Therefore, it is very difficult to draw precise conclusions or identify any

significant trends. Fortunately, **certain insights** can still be gathered by **reviewing a few** with **sufficient data points**. Specifically, our analysis will be conducted on **Delaware**, **Illinois**, and **Texas** going forward.

```
6  SELECT
7      NON EMPTY
8          [DIM GEO].[STATE NAME].[STATE NAME] *
9          {[Measures].[Minimum AQI], [Measures].[Maximum AQI]} ON ROWS,
10     NON EMPTY
11         [DIM DATE].[YEAR].[YEAR] *
12         [DIM DATE].[QUARTER].[QUARTER] ON COLUMNS
13 FROM [21BI11 DDS];
```

		2021	2021	2021	2021	2022	2022	2022	2022	2023	2023	2023	2023
		1	2	3	4	1	2	3	4	1	2	3	4
Alabama	Minimum AQI	(null)	(null)	(null)	(null)	(null)	(null)	4	3	2	4	10	4
Alabama	Maximum AQI	(null)	(null)	(null)	(null)	(null)	(null)	51	46	46	132	86	107
Alaska	Minimum AQI	0	0	1	0	0	1	0	1	0	0	0	1
Alaska	Maximum AQI	189	86	183	157	132	183	189	174	131	183	93	189
Arizona	Minimum AQI	3	5	3	1	4	3	3	1	4	9	3	1
Arizona	Maximum AQI	297	313	239	217	236	313	239	297	236	313	239	297
Arkansas	Minimum AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)	33	34	33	23	25
Arkansas	Maximum AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)	81	81	84	151	147
California	Minimum AQI	1	0	2	0	0	2	0	1	0	0	1	0
California	Maximum AQI	500	500	500	500	500	375	500	500	500	500	323	500
Colorado	Minimum AQI	(null)	(null)	(null)	(null)	24	15	(null)	(null)	(null)	(null)	(null)	(null)
Colorado	Maximum AQI	(null)	(null)	(null)	(null)	159	75	(null)	(null)	(null)	(null)	(null)	(null)
Connecticut	Minimum AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)	30	37	29	30	29
Connecticut	Maximum AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)	171	171	90	169	187
Delaware	Minimum AQI	22	21	21	24	21	25	23	21	24	23	21	25
Delaware	Maximum AQI	89	112	144	72	100	89	80	144	68	89	112	144
Georgia	Minimum AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	10	14	11	12

Figure 51a. MDX query and result.

In **Fig. 51c**, the **minimum AQI**s for **each state** remain **unchanged** throughout the year, apart from **Illinois** which has an **abnormal spike** during **summer**. Broadly, **seasonal** factors **do not** seem to **impact** the **minimum** score.

As for **maximum AQI**, it is a **different** story. **Fall** and **winter** often **hold** the **highest** scores; although much **less frequent**, **summer** also **shares** this **trait**. However, **spring** has **never had** the highest maximum score, thus, spring may be the **quarter** of the year with the **lowest overall AQI** score.

Despite **being unable to predict** which **season** will have the **maximum** or **minimum** AQI, **many reports suggest** that **hot and cold weather** can **affect air quality**, resulting in **summer** and **winter** boasting the **highest AQI** values. An interesting fact is that **in recent years**, **fall temperatures** in the **U.S.** have been **on the rise**, which could also justify the peak in fall's AQI values.

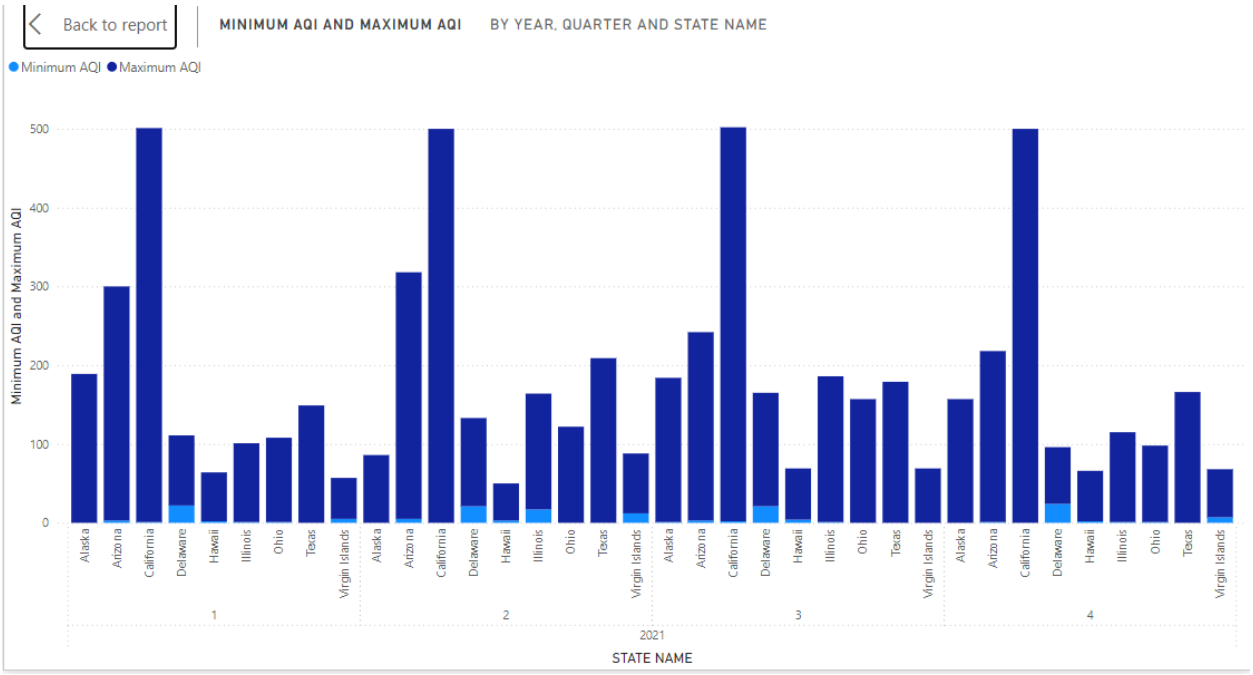


Figure 51b. Minimum and Maximum AQI report in 2021.

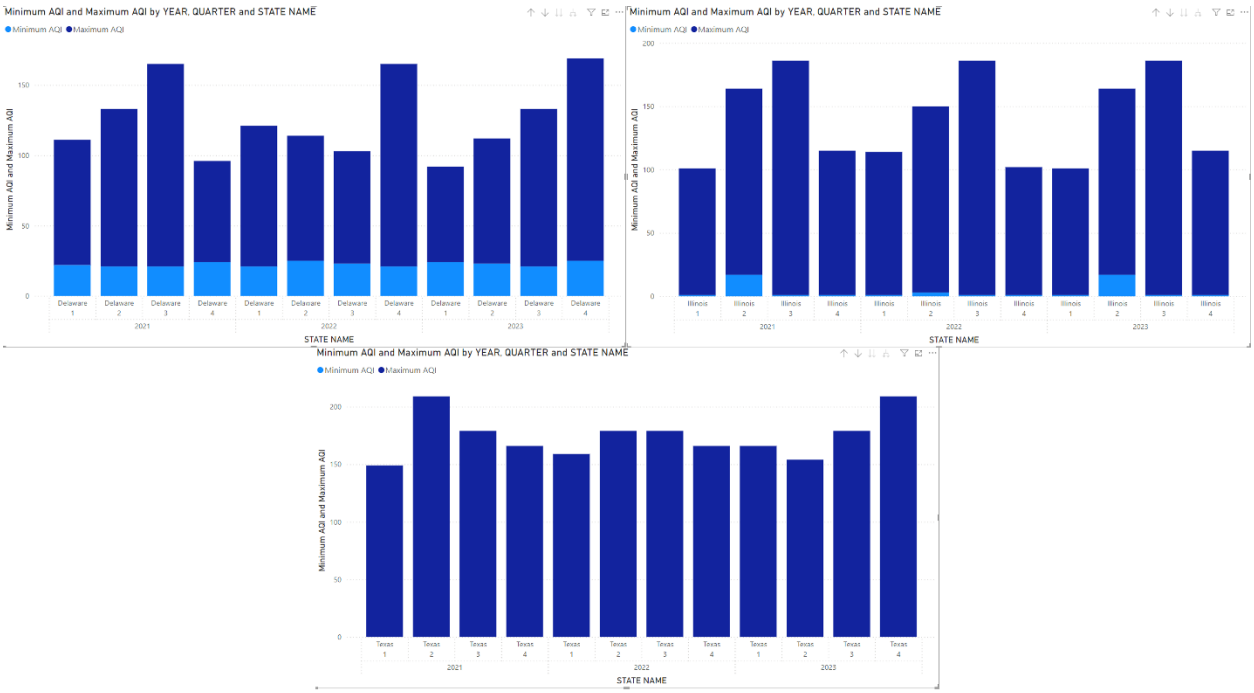


Figure 51c. Top Left – Delaware's Min and Max AQIs. Top Right – Illinois' Min and Max AQIs. Bottom – Texas' Min and Max AQIs.

5.2 AQI's Mean and StDev of states during each year's quarters

"Report the **mean and the standard deviation** of AQI value for each **State** during **each quarter of years**. *Analysis hints:* How do the AQI values fluctuate during the year? Pay attention to the values (mean, std, max, min). Are any unusually large or small?"

```
20 SELECT
21     NON EMPTY
22         [DIM GEO].[STATE NAME].[STATE NAME] *
23         {[Measures].[AVG AQI], [Measures].[STDEV AQI]} ON ROWS,
24     NON EMPTY
25         [DIM DATE].[YEAR].[YEAR] *
26         [DIM DATE].[QUARTER].[QUARTER] ON COLUMNS
27 FROM [21BI11 DDS];
```

		2021	2021	2021	2021	2022	2022	2022
		1	2	3	4	1	2	3
Alabama	AVG AQI	(null)	(null)	(null)	(null)	(null)	(null)	15.9272727272
Alabama	STDEV AQI	(null)	(null)	(null)	(null)	(null)	(null)	9.40377979219
Alaska	AVG AQI	37.2346368715084	24.5223880597015	22.4338235294118	33.4166666666667	25.1877394636015	25.1082089552239	34.2481203007
Alaska	STDEV AQI	29.787029151755	15.713724592401	16.6869763437919	25.6661345543688	16.5389127550713	19.2445826783487	27.0407395278
Arizona	AVG AQI	49.0061674008811	60.4520780322307	54.9722921914358	48.6193277310924	55.6649572649573	57.9516129032258	53.2787573467
Arizona	STDEV AQI	34.5716768904899	34.1983691114102	36.4877178838954	32.9521496150697	38.9806950165403	28.8317560069515	40.3558421891
Arkansas	AVG AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Arkansas	STDEV AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)
California	AVG AQI	48.7270815811606	53.1810163796724	77.344726357727	53.1915118243243	52.460025651988	57.8566277836691	58.5009499683
California	STDEV AQI	20.3786989083012	27.9480602284584	51.3257807186637	30.1037735677513	24.1946380380003	30.3103742459813	35.8138416405
Colorado	AVG AQI	(null)	(null)	(null)	(null)	60.3777777777778	41.8484848484849	(null)
Colorado	STDEV AQI	(null)	(null)	(null)	(null)	27.1749794143959	16.2455051352137	(null)
Connecticut	AVG AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Connecticut	STDEV AQI	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Delaware	AVG AQI	43.5924528301887	45.7069597069597	49.2644927536232	43.6884057971015	41.4481481481481	46.9650655021834	43.3423913043
Delaware	STDEV AQI	12.5674009119522	13.6816977258446	15.8564315151573	11.3949523052195	11.3494614122619	12.6594150068037	9.27345127929

Figure 52a. MDX query and result.

Fig. 52b illustrates the **average** and **standard deviation** of **AQI** values in **2021** and **2022** (2023 is omitted as it shares similar features). Our dataset has **very high standard deviation** values, indicating that the **data is widely spread** or **contains many outliers**. That the **highest values** occur **in fall** or **winter supports** our **previous observation** of these seasons **often experiencing** harsh days with **unusually high** or **low AQI** values.

Overall, the **average AQI** remains **below 100** for **most states** and **quarters**, while those with **elevated AQI**s **still present**, likely reflecting **temporary air quality issues**. The **AQI** values usually **increase** in the **final quarters of the year**.

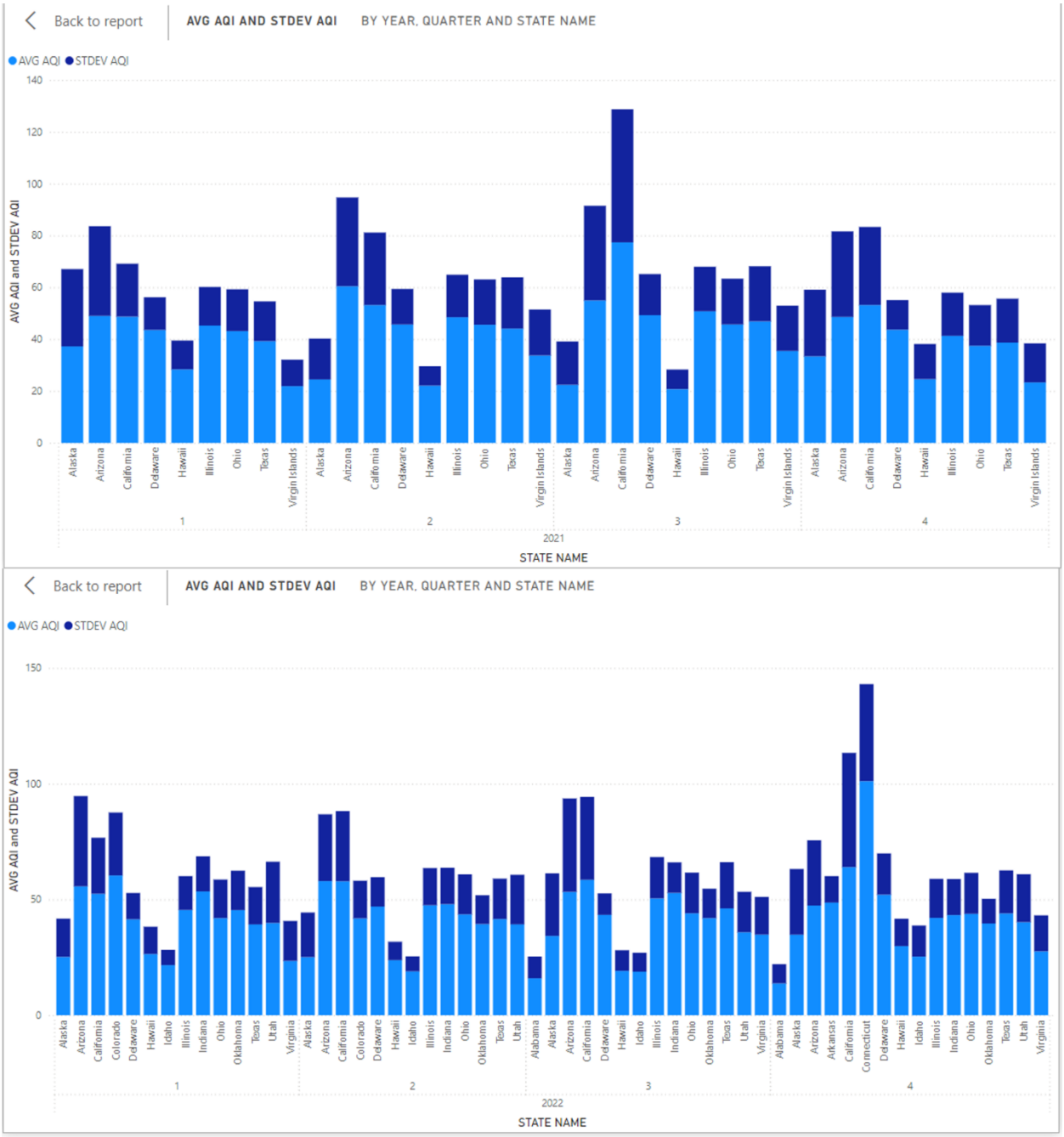


Figure 52b. Top – 2021's Mean and StDev AQIs. Bottom – 2022's Mean and StDev AQIs.

5.3 No. of days, AQI Mean when "very unhealthy" or worse by county

"Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County. *Analysis hint:* What is the AQI limit above which air quality is "very unhealthy" or worse?"

```
34 SELECT
35     NON EMPTY
36         [DIM GEO].[STATE NAME].[STATE NAME] *
37         [DIM GEO].[COUNTY].[COUNTY] ON ROWS,
38     NON EMPTY
39         {[Measures].[FACT AIR QUALITY Count],
40          [Measures].[AVG AQI]} ON COLUMNS
41 FROM [21BI11 DDS]
42 WHERE
43     {[DIM CATEGORY].[CATEGORY].&[Very Unhealthy],
44      [DIM CATEGORY].[CATEGORY].&[Hazardous]};
```

		FACT AIR QUALITY Count	AVG AQI
Arizona	Gila	1	297
Arizona	La Paz	73	215.575342465753
Arizona	Maricopa	145	215.013793103448
Arizona	Navajo	1	313
Arizona	Pima	2	313
California	Amador	3	244.333333333333
California	Butte	6	244.333333333333
California	Calaveras	1	225
California	Colusa	2	225
California	Del Norte	2	324.5
California	El Dorado	7	296.714285714286
California	Fresno	3	241.333333333333
California	Humboldt	3	342
California	Imperial	13	330.538461538462
California	Inyo	13	314
California	Kern	5	326.8
California	Kings	2	500
California	Lake	1	281

Figure 53a. MDX query and result.

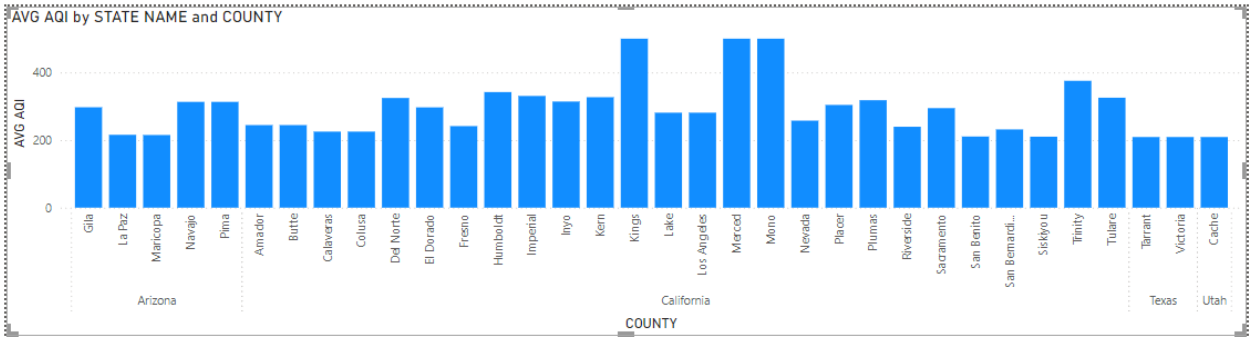


Figure 53b. Average (Mean) AQI on bad days.

The **sole category worse than "very unhealthy"** in AQI index is **"hazardous"**, so only **days belonging to one of these two** are handled. **Fig. 53b** shows only **4 states** have recorded **daily AQI** values **falling into these categories**. All of them have **mean AQI** values **above 200** – the **threshold** of **"very unhealthy"** or **worse**. Most counties' values **fluctuate around 250** (very unhealthy), while **some counties in California exceed 400** (hazardous).

Although **California suffers extremely high mean AQIs on bad days** (AQI above 200), in **Fig. 53c**, **many of the days remain unhealthy**, and **the number of days** this occurs is **very low**, with only a **few counties** having **notable values**. On the other hand, while **Arizona's counties haven't reached** a **hazardous** level yet, the **number of days** with **very unhealthy** AQI is **significantly higher**. But for all counties, AQI values above 200 rarely ever occur.

This information together expresses that the **air quality in California** is **unstable**, and it may be wise to **keep two of Arizona's counties on the checklist**.

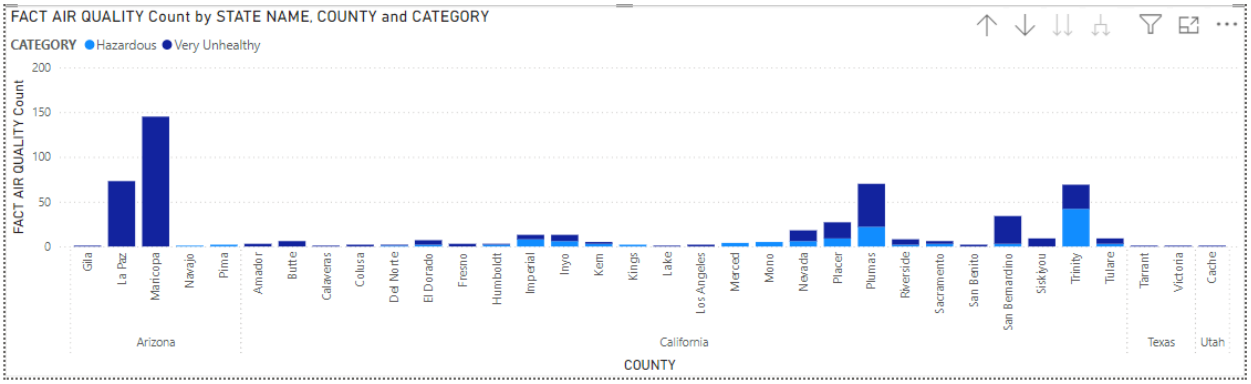


Figure 53c. Number of days having bad AQI record.

5.4 No. of days in each Category by County for 4 states

"For the four following states: Hawaii, Alaska, Illinois and Delaware, **count the number of days** in each air quality **Category** (Good, Moderate, etc.) by **County**. *Analysis hints:* Comparing the data of the states and counties, focus on the distribution of the harmful air condition. What could you conclude about the differences?)"

```
52 SELECT
53     NON EMPTY
54         {[DIM GEO].[STATE NAME].&[Hawaii],
55          [DIM GEO].[STATE NAME].&[Alaska],
56          [DIM GEO].[STATE NAME].&[Illinois],
57          [DIM GEO].[STATE NAME].&[Delaware]} *
58         [DIM GEO].[COUNTY].[COUNTY] ON ROWS,
59     NON EMPTY
60         [DIM CATEGORY].[CATEGORY].[CATEGORY] ON COLUMNS
61 FROM [21BI11 DDS]
62 WHERE [Measures].[FACT AIR QUALITY Count];
```

		Good	Moderate	Unhealthy	Unhealthy for Sensitive Groups
Hawaii	Hawaii	1029	59	(null)	(null)
Hawaii	Honolulu	1092	3	(null)	(null)
Hawaii	Kauai	454	(null)	(null)	(null)
Hawaii	Maui	1023	4	(null)	(null)
Alaska	Aleutians East	299	60	(null)	1
Alaska	Anchorage	939	138	7	11
Alaska	Denali	854	158	13	26
Alaska	Fairbanks North Star	800	247	19	29
Alaska	Juneau	966	94	1	2
Alaska	Kenai Peninsula	319	38	1	2
Alaska	Matanuska-Susitna	973	77	1	2
Alaska	North Slope	324	10	(null)	1
Illinois	Adams	440	47	(null)	(null)
Illinois	Clark	897	136	(null)	(null)
Illinois	Cook	499	554	2	40
Illinois	Champaign	738	311	(null)	(null)
Illinois	DuPage	451	619	1	24
Illinois	Effingham	570	146	(null)	2

Figure 54a. MDX query and result.

In **Fig. 54b**, **Good** and **Moderate** are the **dominant** categories in the distribution; while instances of **Unhealthy** and **Unhealthy for Sensitive Groups** do **exist**, they **make up only a fraction** in **most counties**, primarily of the states **Alaska** and **Illinois**.

Albeit having the **highest number** of **unhealthy cases** overall, **Alaska** shows a **higher share** of **Good** days in its distribution **compared to Delaware** and **Illinois**. This indicates that **Alaska's air quality** is **generally better**, though it **includes** a certain **number of very poor** days. In contrast, **Hawaii** exhibits **the best air quality**, with only **a few Moderate** days in **Hawaii County**, while the rest is mostly recorded as **Good**.

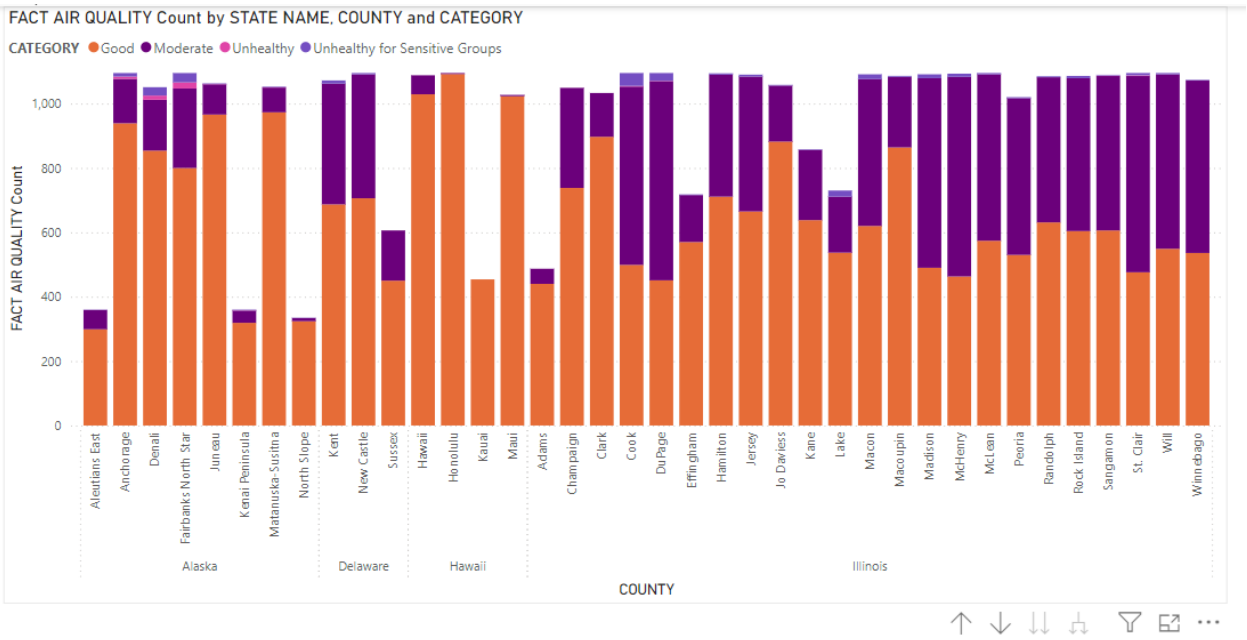


Figure 54b. The number of days in each category for the requested states' counties.

5.5 Mean AQI by quarters for 4 states

"For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the **mean AQI** value by **quarters**. *Analysis hints:* Comparing the data of the states over the year. What could you conclude about the fluctuations?"

In **Fig. 55b**, **Hawaii** generally has **the lowest average AQI over the year** and **Alaska** is its **runner-up**. Although **Alaska** keeps **a smaller AQI mean** than **Delaware** and **Illinois**, it faces **a huge value gap between quarters**, while **the latter two's AQI mean** are **more consistent** throughout the year. This implies **seasonal elements greatly impact Alaska**'s air quality.

Delaware's AQI does not seem to **fluctuate in any pattern**, yet there are clear trends in **Hawaii's** and **Illinois'**. **Hawaii's mean AQI** is **lowest in summer** and **fall**, then **highest during the remaining months**. For **Illinois**, there is a bit different as its **AQI mean values increase** in a **precise order** from **winter to spring, summer, then fall**. It appears that **Alaska's mean AQI imitates Hawaii's pattern**, but the **third quarter of 2022 breaks this mold**. There may have been **unusual temperatures** at that time **causing people to rely more on heaters** and **impair the air quality**.

```
69 SELECT
70     NON EMPTY
71         [DIM DATE].[YEAR].[YEAR] *
72         [DIM DATE].[QUARTER].[QUARTER] ON ROWS,
73     NON EMPTY
74         {[DIM GEO].[STATE NAME].&[Hawaii],
75          [DIM GEO].[STATE NAME].&[Alaska],
76          [DIM GEO].[STATE NAME].&[Illinois],
77          [DIM GEO].[STATE NAME].&[Delaware]} ON COLUMNS
78 FROM [21BI11 DDS]
79 WHERE [Measures].[AVG AQI];
```

		Hawaii	Alaska	Illinois	Delaware
2021	1	28.3944444444444	37.2346368715084	45.2623843222646	43.5924528301887
2021	2	22.1208791208791	24.5223880597015	48.5098795180723	45.7069597069597
2021	3	20.7635869565217	22.4338235294118	50.7884250474383	49.2644927536232
2021	4	24.6164383561644	33.4166666666667	41.3112653497064	43.6884057971015
2022	1	26.4334277620397	25.1877394636015	45.4502014968336	41.4481481481481
2022	2	23.7838827838828	25.1082089552239	47.5063419583968	46.9650655021834
2022	3	19.16	34.2481203007519	50.4547263681592	43.3423913043478
2022	4	29.8644688644689	34.804780876494	42.0454797559623	52.0710382513661
2023	1	25.1385767790262	26.0362595419847	45.4394939493949	42.2888888888889
2023	2	20	28.0166358595194	48.4846743295019	45.9642857142857
2023	3	18.3150183150183	24.2802893309222	50.3023696682464	45.5434782608696
2023	4	23.2478260869565	34.4495412844037	41.6941928609483	51.437125748503

Figure 55a. MDX query and result.

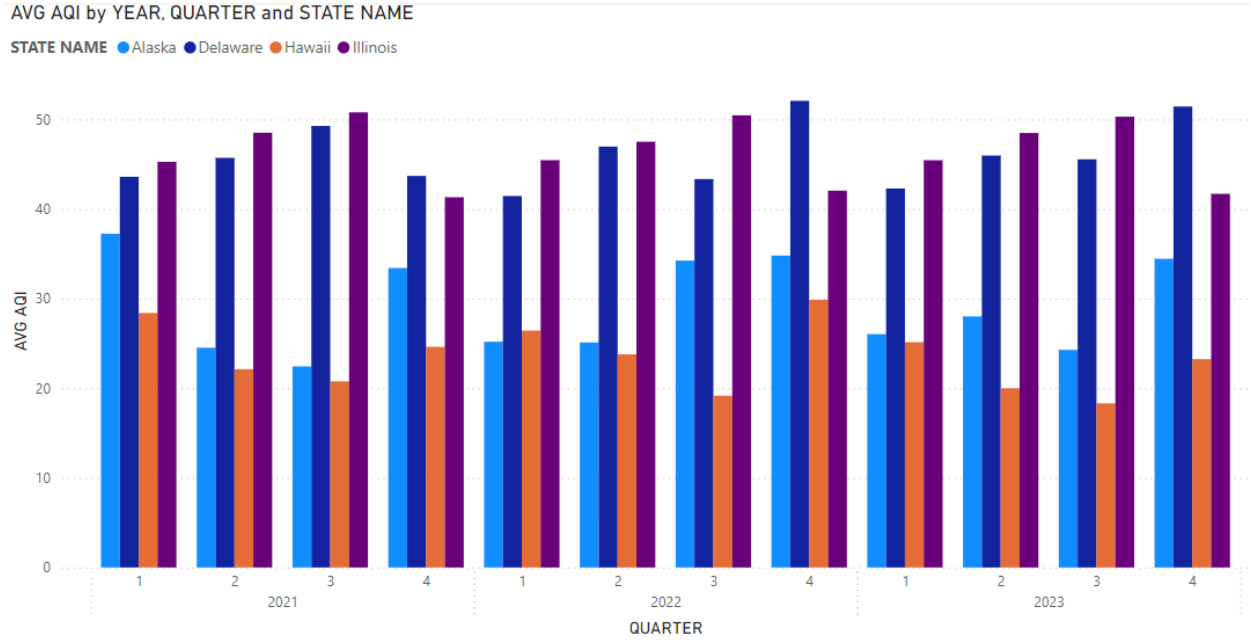


Figure 55b. Mean AQI by quarter for the 4 states.

5.6 AQI fluctuation trends over the year for 4 states

> "Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California. *Analysis hint:* Give your opinion about the fluctuations of AQI value."

Since this requirement has been largely addressed in Sect. 5.5, we will only be focusing on the 4 states' AQI trend over the whole years, instead of specific quarters.

From **Fig. 56b**, it is evident that the **AQI values** for **each state** show **minimal fluctuation**. If **any change** is observed, it tends to **be limited to a slight drop** in the **most recent year**. This trend reveals that **measures have been implemented** to **control and improve air quality** in these states.

```
86  SELECT
87      NON EMPTY
88          {[DIM GEO].[STATE NAME].&[Hawaii],
89           [DIM GEO].[STATE NAME].&[Alaska],
90           [DIM GEO].[STATE NAME].&[Illinois],
91           [DIM GEO].[STATE NAME].&[California]} ON ROWS,
92      NON EMPTY
93          [DIM DATE].[YEAR].[YEAR] ON COLUMNS
94  FROM [21BI11 DDS]
95  WHERE [Measures].[AVG AQI];
```

	2021	2022	2023
Hawaii	23.9533287577213	24.9114139693356	21.6060019361084
Alaska	29.3871117292536	29.7791586998088	28.2024965325936
Illinois	46.6543772963385	46.5103044807871	46.6543772963385
California	58.1278452399727	58.2029239766082	58.2009722741599

Figure 56a. MDX query and result.

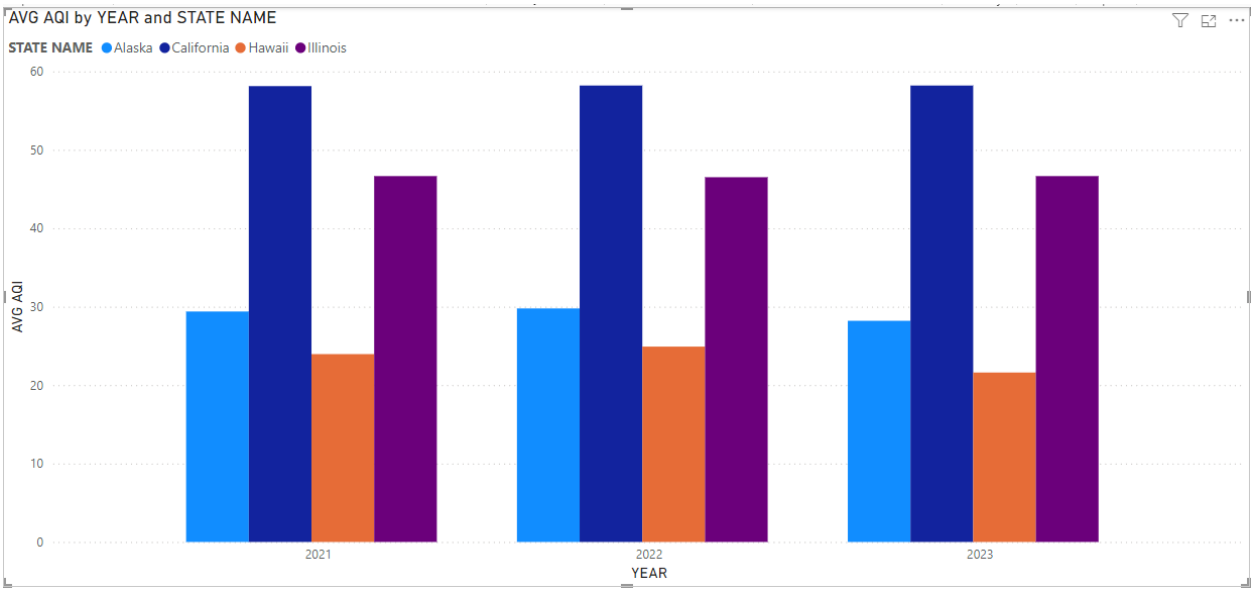


Figure 56b. AVG AQI by year for the 4 states.

5.7 Build graphs/charts for the above reports

Here just for completion's sake, the graphs are incorporated into their respective questions.

5.8 Regional map for AQI Mean in regions during a year

"Use a regional map to visually represent (by color) the mean AQI value in regions during a year. Example:

US mean AQI of four states: Alaska, Delaware, Hawaii, Illinois over the year 2023

Month	Alaska	Delaware	Hawaii	Illinois
2023-01	40.339	43.151	38.581	44.517
2023-02	28.032	47.893	34.405	40.057
2023-03	29.077	46.570	29.600	45.179
2023-04	24.994	53.278	25.500	48.929
2023-05	25.632	50.699	28.364	65.065
2023-06	20.050	82.956	24.435	90.900
2023-07	22.762	56.355	24.462	55.857
2023-08	30.117	53.000	26.544	50.501
2023-09	17.956	47.822	25.256	48.688
2023-10	29.303	41.032	18.419	40.935
2023-11	30.620	46.012	24.156	41.243
2023-12	36.558	43.284	23.593	35.737"

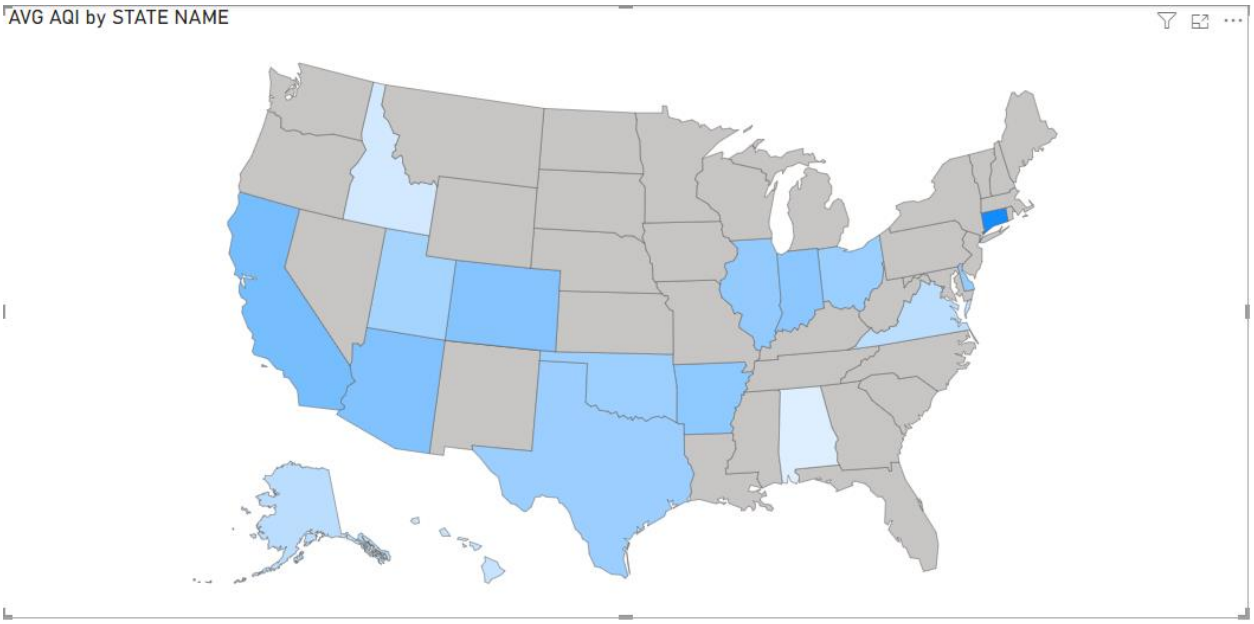


Figure 58. AVG AQI of states in 2022.

Due to **a lot of information missing** from both **the states** and **certain years' AQI values**, we **cannot derive any trend** in the **AQI values by geographic features**. The following remarks can be said about the map,

- **Connecticut's mean AQI** value is **outstandingly high**.
- **Western** and **Southwestern** parts of the **USA** seem to have **slightly higher mean AQI** values than **other parts** of the country (based on some candidates like California, Arizonia, and Colorado).

5.9 AQI's Mean, StDev, Min, Max by county during each year's quarters

> "Report the **mean, the standard deviation, min and max** of AQI value group by **State** and **County** during **each quarter of the year**. *Analysis hints:* Pay attention to the values (mean, std, max, min). Are any unusually large or small? Compare the standard deviation values between question 1 and 2, explain."

Most noticeable in **Fig. 59b** is the **huge gap** between **maximum AQI** and **other measures**, especially with **mean AQI**s being **many times smaller**. Again, **reinforcing** our conclusion about the **existence of spike days** (when the AQI value is unusually high). If **maximum AQI** occurred **due to seasonal effects alone**, we **would expect** the **mean AQI** to be **more proportionate**. Whereas the **minimum AQI** value for **each state** always **appears very small compared to the other values**. Because of this, the **standard deviation** remains **high**, making it easier to spot in every county.

Put together with the analysis in Sect. 5.1, it can be observed that the **maximum and minimum AQI** values **for the same county across different quarters** tend to **be similar**. Additionally, these values are **comparable across counties within the same state**. While

the **exact values for each county** may **vary**, it is reasonable to **estimate a county's maximum** and **minimum AQI** values just **by looking at its belonging state**.

```
123 SELECT
124     NON EMPTY
125         [DIM GEO].[STATE NAME].[STATE NAME] *
126         [DIM GEO].[COUNTY].[COUNTY] *
127         {[Measures].[AVG AQI], [Measures].[STDEV AQI],
128          [Measures].[Minimum AQI], [Measures].[Maximum AQI]}
129         ON ROWS,
130     NON EMPTY
131         [DIM DATE].[YEAR].[YEAR] *
132         [DIM DATE].[QUARTER].[QUARTER] ON COLUMNS
133 FROM [21BI11 DDS];
```

			2021	2021	2021	2021	2022	
			1	2	3	4	1	
Alabama	Sumter	AVG AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Sumter	STDEV AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Sumter	Minimum AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Sumter	Maximum AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Tuscaloosa	AVG AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Tuscaloosa	STDEV AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Tuscaloosa	Minimum AQI	(null)	(null)	(null)	(null)	(null)	
Alabama	Tuscaloosa	Maximum AQI	(null)	(null)	(null)	(null)	(null)	
Alaska	Aleutians East	AVG AQI	17.5	13.7096774193548	13.6	15.3333333333333	35.1	41.933
Alaska	Aleutians East	STDEV AQI	11.1497384124771	5.8264887250971	8.92038863129479	9.85073789302289	26.0145472124092	16.029
Alaska	Aleutians East	Minimum AQI	4	4	3	2	2	
Alaska	Aleutians East	Maximum AQI	51	30	46	39	132	
Alaska	Anchorage	AVG AQI	38.7777777777778	29.8021978021978	23.304347826087	37.5978260869565	28.2666666666667	24.659
Alaska	Anchorage	STDEV AQI	18.5512969164948	16.666884835347	14.7226084131757	18.6087051757391	15.9114214743163	15.135
Alaska	Anchorage	Minimum AQI	4	10	7	5	10	
Alaska	Anchorage	Maximum AQI	132	86	107	81	86	

Figure 59a. MDX query and result.

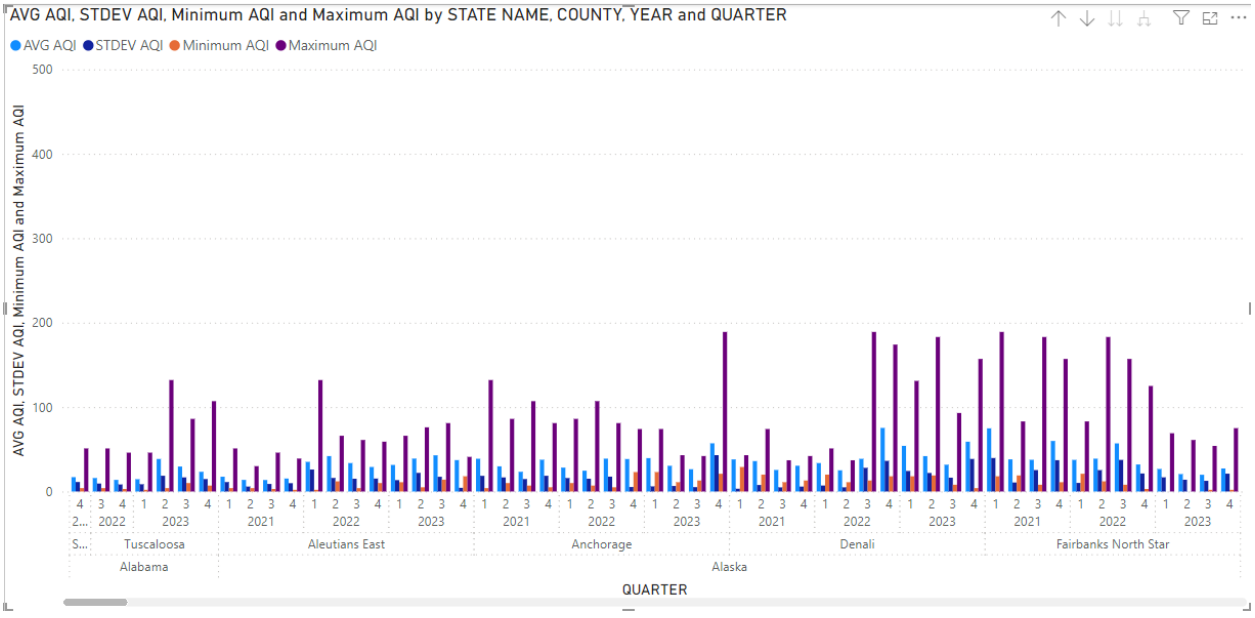


Figure 59b. A slice of the graph on AVG, STDEV, Min, and Max AQI for each state's county by quarters.

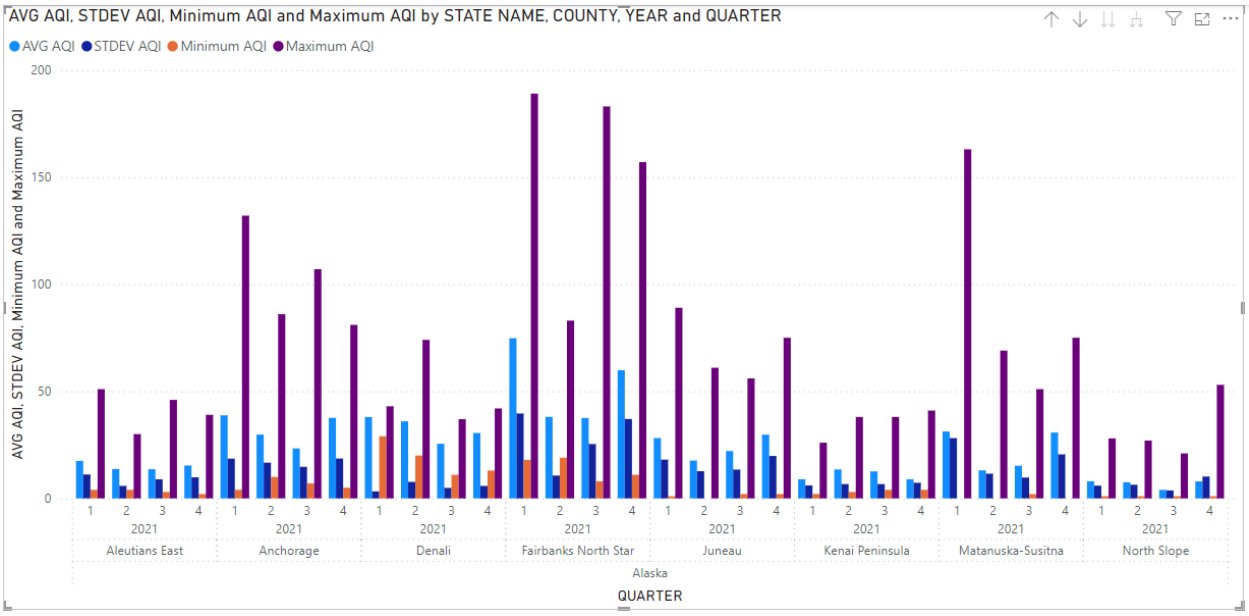


Figure 59c. AVG, STDEV, Min, and Max AQI of Alaska's counties in 2021 by quarters.

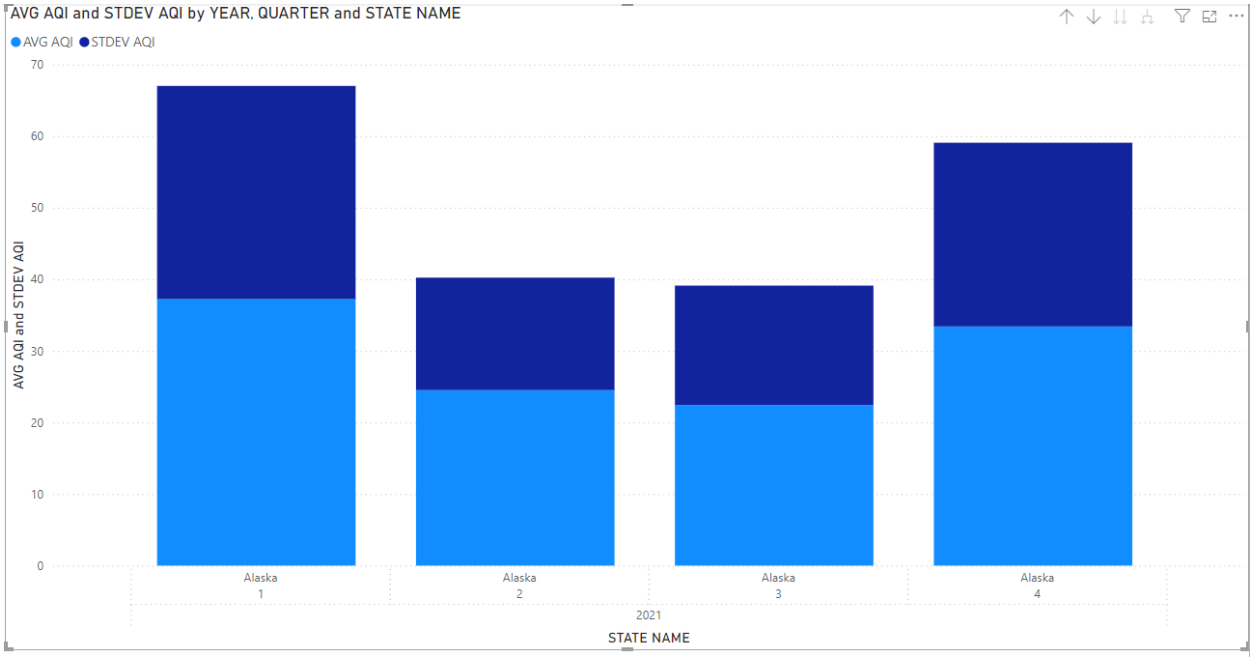


Figure 59d. AVG, STDEV, Min, and Max AQI of Alaska in 2021 by quarters.

During Sect. 5.2, we observed a **relatively high standard deviation compared to** the **average**. This **high variability** is **evident in each state and county** (e.g., **Alaska's** in **Fig.59c**). Still, the **significant variation** in **counties' data** is **challenging** when **assessing county-level averages** or **standard deviations** based **solely on state-level values**.

For instance, the **standard deviation of the values** in **Fig. 59e** is nearly **half the mean AQI** value in **Fig. 59d** (compared in the same quarter, the former is **18.77** while the latter **37.23**).

These statistics highlight the fact that, while it's **feasible** to **evaluate county-level AQIs based on state-level values** to a certain extent, **such estimations** are **not accurate** and may **not** be the **best practice**.

```
SELECT
    NON EMPTY
        [DIM GEO].[STATE NAME].&[Alaska] *
        [DIM GEO].[COUNTY].[COUNTY] ON ROWS,
    NON EMPTY
        [DIM DATE].[YEAR].[YEAR] *
        [DIM DATE].[QUARTER].&[1]&[2021] ON COLUMNS
FROM [21BI11 DDS]
WHERE [Measures].[AVG AQI];
```

		2021
		1
Alaska	Aleutians East	17.5
Alaska	Anchorage	38.7777777777778
Alaska	Denali	38
Alaska	Fairbanks North Star	74.7888888888889
Alaska	Juneau	28.1555555555556
Alaska	Kenai Peninsula	8.93103448275862
Alaska	Matanuska-Susitna	31.2444444444444
Alaska	North Slope	8

Figure 59e. AVG AQI for Alaska's counties in 2021's first quarter.

5.10 AQI by state, Category, DayLightSaving over years

"Create a new attribute, **DayLightSaving**, in a suitable table. **DayLightSaving** may have two values:

True: Between March 12, 2023, and November 5, 2023

False: Otherwise

Report the mean AQI value by State, Category, DayLightSaving over years. Analysis hint: Is there any notable difference on the air quality during the DaylightSaving period compared to the other?"

```
143 SELECT
144     NON EMPTY
145         [DIM GEO].[STATE NAME].[STATE NAME] *
146         [DIM DATE].[YEAR].[YEAR] *
147         [DIM DATE].[DAYLIGHT SAVING].[DAYLIGHT SAVING] ON ROWS,
148     NON EMPTY
149         [DIM CATEGORY].[CATEGORY].[CATEGORY] ON COLUMNS
150 FROM [21BI11 DDS]
151 WHERE [Measures].[AVG AQI];
```

Messages Results

			Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Alabama	2022	False	14.5436893203883	(null)	51	(null)	(null)	(null)
Alabama	2023	False	16.248322147651	(null)	56	(null)	107	(null)
Alabama	2023	True	24.9435897435897	(null)	57.7209302325581	(null)	132	(null)
Alaska	2021	False	22.013093289689	(null)	62.1684210526316	164.357142857143	120.76	(null)
Alaska	2022	False	22.2221594120972	(null)	62.2077464788732	164.357142857143	120.76	(null)
Alaska	2023	False	20.5737976782753	(null)	65.5445544554455	164.6	122.888888888889	(null)
Alaska	2023	True	21.6635145784082	(null)	61.1842105263158	163.75	112	(null)
Arizona	2021	False	35.312324929972	313	66.0524934383202	176.112676056338	118.552941176471	215.575342465753
Arizona	2022	False	35.6382676606319	313	66.0524934383202	176.112676056338	118.552941176471	215.575342465753
Arizona	2023	False	38.0275715800636	(null)	62.9106699751861	173.484848484848	126.8	216.774193548387
Arizona	2023	True	38.0130718954248	313	67.2518050541516	178.394736842105	116.507462686567	214.690476190476
Arkansas	2022	False	38.7931034482759	(null)	58.8571428571429	(null)	(null)	(null)
Arkansas	2023	False	38.78125	(null)	61.448275862069	(null)	106.5	(null)
Arkansas	2023	True	39.0212765957447	(null)	60.8992805755396	151	113.8	(null)
California	2021	False	36.4540637914591	439.634146341463	66.2028549962434	167.10401891253	120.239476145931	235.925373134328
California	2022	False	36.4708972022725	439.634146341463	66.2184607582097	167.1567695962	120.272300469484	235.925373134328
California	2023	False	35.1064250411862	447.307692307692	66.9685844057532	167.951690821256	120.789592760181	239.022727272727
California	2023	True	37.0554493307839	426.333333333333	65.9810664605873	166.362790697674	119.987055016181	230

Figure 510a. MDX query and result.

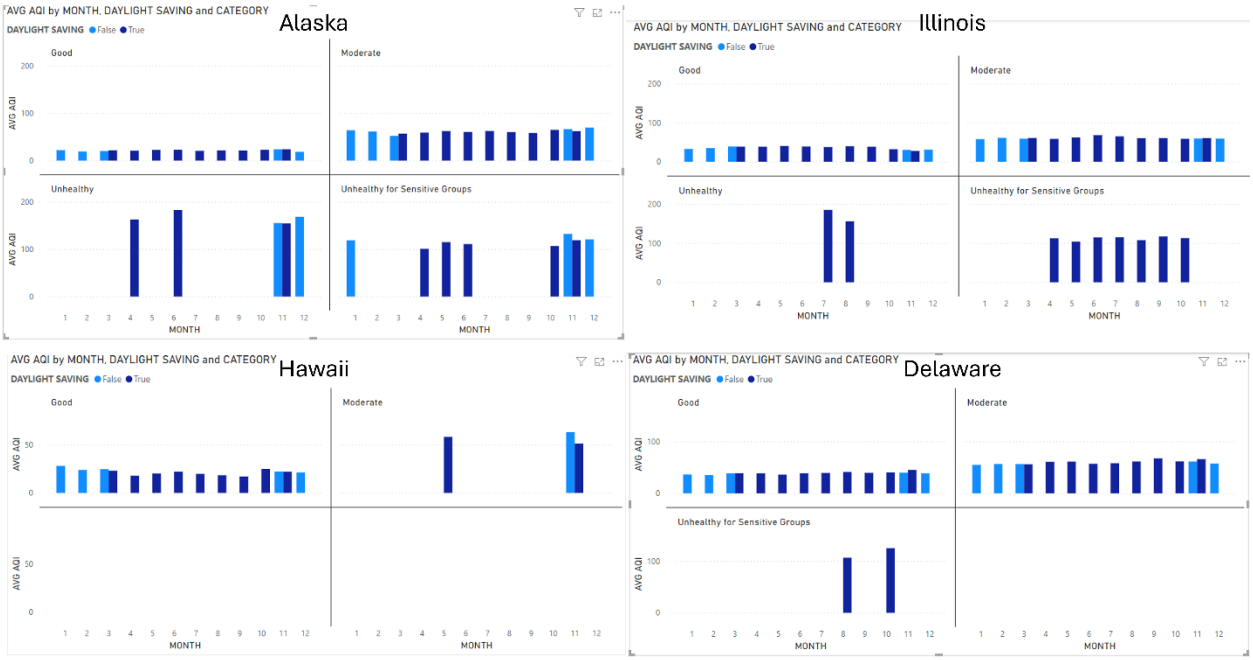


Figure 510b. AVG AQI by State, Category, DayLightSaving over the years for 4 states.

For **DayLightSaving** (hereafter, **DST** or Daylight-Saving Time) is only **applied** to **a range of days in 2023**, it is logical to just focus on said relevant year. An **exhaustive graph** has been plotted but due to **being clustered** by **various attributes**, it proves **tedious for inspection**. Ergo, we think it's **best** to **examine one state at a time**. In this case, **Alaska**, **Illinois**, **Hawaii**, and **Delaware** are **picked** as representatives, presented in **Fig. 510b**.

Since **Category** has already **reflected** the **mean AQI level**, and the **differences** within the **same category** are marginally **negligible**, it is safe to **analyze** based on **Category**, rather than senselessly scrutinizing the average AQI values.

The **data from these 4 states draws** a **trend** that **is present in** many **other states** as well. Specifically, categories like "**Unhealthy for Sensitive Groups**" or **worse** tend to **occur more frequently** during the **DST period**. In opposition, "**Good**" and "**Moderate**" **react similarly** across **all months**, but many **instances** where the **mean AQI** values **during DST** are **higher than** in **normal months** also exist.

At first, this may **seem odd** seeing that **DST** was **originally introduced** with **energy saving** as **one of the goals**. Yet, **one factor** can once again be **the effects season plays on** those **DST months**, or we simply have **more DST months** than regular ones; **recent reports** from the **U.S.** government also **show** that the **energy-saving effects** of **DST** are **insignificant**.

5.11 No. of days by state, Category in each month

> "Count the number of days by State, Category in each month. *Be caution: The Category in the data set is calculated for each County, not State.*"

```
157 SELECT
158     NON EMPTY
159         [DIM GEO].[STATE NAME].[STATE NAME] *
160         [DIM DATE].[YEAR].[YEAR] *
161         [DIM DATE].[MONTH].[MONTH] ON ROWS,
162     NON EMPTY
163         [DIM CATEGORY].[CATEGORY].[CATEGORY] ON COLUMNS
164 FROM [21BI11 DDS]
165 WHERE [Measures].[FACT AIR QUALITY Count];
```

Messages Results

			Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Alabama	2022	8	24	(null)	1	(null)	(null)	(null)
Alabama	2022	9	30	(null)	(null)	(null)	(null)	(null)
Alabama	2022	10	30	(null)	(null)	(null)	(null)	(null)
Alabama	2022	11	10	(null)	(null)	(null)	(null)	(null)
Alabama	2022	12	9	(null)	(null)	(null)	(null)	(null)
Alabama	2023	1	31	(null)	(null)	(null)	(null)	(null)
Alabama	2023	2	28	(null)	(null)	(null)	(null)	(null)
Alabama	2023	3	31	(null)	(null)	(null)	(null)	(null)
Alabama	2023	4	21	(null)	8	(null)	1	(null)
Alabama	2023	5	21	(null)	10	(null)	(null)	(null)
Alabama	2023	6	24	(null)	6	(null)	(null)	(null)
Alabama	2023	7	16	(null)	15	(null)	(null)	(null)
Alabama	2023	8	31	(null)	(null)	(null)	(null)	(null)
Alabama	2023	9	30	(null)	(null)	(null)	(null)	(null)
Alabama	2023	10	27	(null)	4	(null)	(null)	(null)
Alabama	2023	11	29	(null)	1	(null)	(null)	(null)
Alabama	2023	12	55	(null)	1	(null)	1	(null)
Alaska	2021	1	132	(null)	36	8	11	(null)

Figure 511a. MDX query and result.

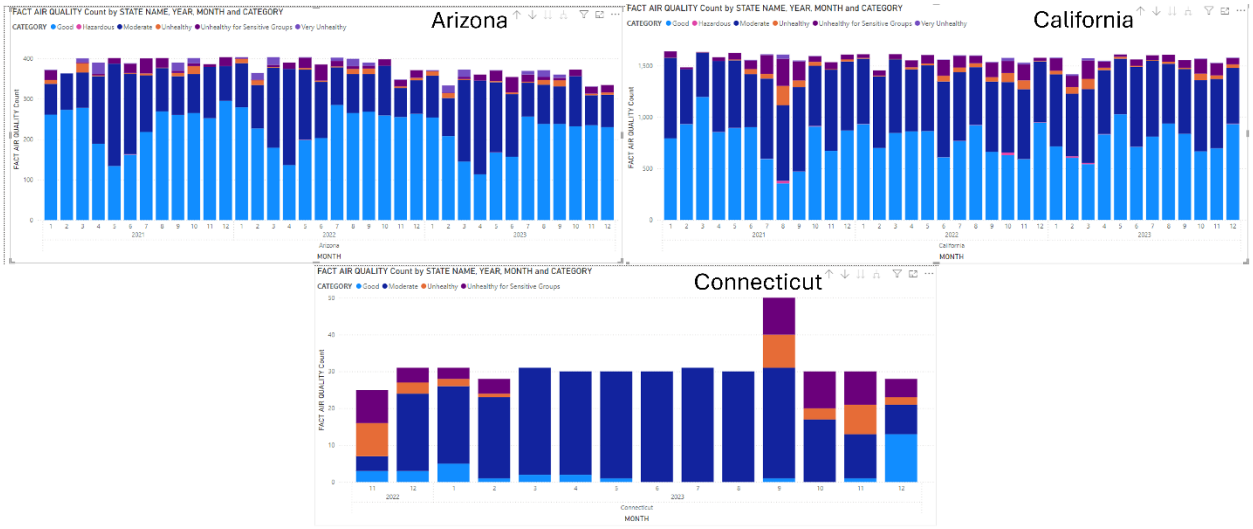


Figure 511b. No. of days by state and category in each month for 3 states.

In that the **data**, again, **varies heavily by geographic elements**, we elect some **candidate states** for analysis. To **avoid overusing** exact **states** repeatedly, new faces of **Arizona**, **California**, and **Connecticut** will be put on the table (although their data is not as sufficient as the previous sections' states'), presented in **Fig. 511b**.

Arizona's air **contains many spike days** under "**Unhealthy for Sensitive Groups**" or **worse**, they appear to be **unpredictable** and **lack any trend**. It's a **different** case **for** "**Good**" days as **compared to other categories**, **its ratio** exhibits a **very visible flow**. This

ratio **usually falls the lowest** around **March** to **July**, **reaching** its **bottom** in **April** or **May**, which signals that **season** plays an **extreme role** in **Arizonia's AQI**.

Things are also **not favorable** for **California**, yet it **does not trace Arizona's pattern**. The **worst spikes don't occur** during the **same time each year**, hinting at many **significant factors contributing to the air quality crisis**, and **California's** is deemed **polluted**.

There **isn't enough data** on **Connecticut** from **January** to **October 2021**, but overall, it **stands out** as having **the best air quality among the three states** so far. Interestingly, Connecticut **didn't record** any **day worse than "Unhealthy."** Even more impressive is **the stretch from March to August**, where every day is either "**Moderate**" or "**Good**". This suggests that **Connecticut** has **relatively cleaner air during that period**, which could be **thanks to** a combination of **effective management** or supportive **environmental factors**.

In conclusion, here again **seasonal aspects** prove to be **vital in AQI matters**. Moreover, in **states with oddly fluctuating air quality** (e.g., California), it must have been **the involvement of pollutants** such as factories, emissions, etc.

5.12 No. of days by Category and Defining Parameter

"Report the number of days by Category and Defining Parameter. *Analysis hints:* What is your opinion on the pollution situation in the United States as a whole? Additionally, please identify the primary factors that the country should consider in order to enhance air quality."

```
172 SELECT
173     NON EMPTY [DIM CATEGORY].[CATEGORY].[CATEGORY] ON ROWS,
174     NON EMPTY
175         [DIM DEFINING PARAMETER].[DEFINING PARAMETER].[DEFINING PARAMETER]
176         ON COLUMNS
177 FROM [21BI11 DDS]
178 WHERE [Measures].[FACT AIR QUALITY Count];
```

	CO	NO2	Ozone	PM10	PM2.5
Good	223	2899	63200	4378	50398
Hazardous	(null)	(null)	26	18	82
Moderate	11	748	28016	2190	34667
Unhealthy	(null)	7	875	84	652
Unhealthy for Sensitive Groups	(null)	15	2742	251	1587
Very Unhealthy	(null)	(null)	245	34	144

Figure 512a. MDX query and result.

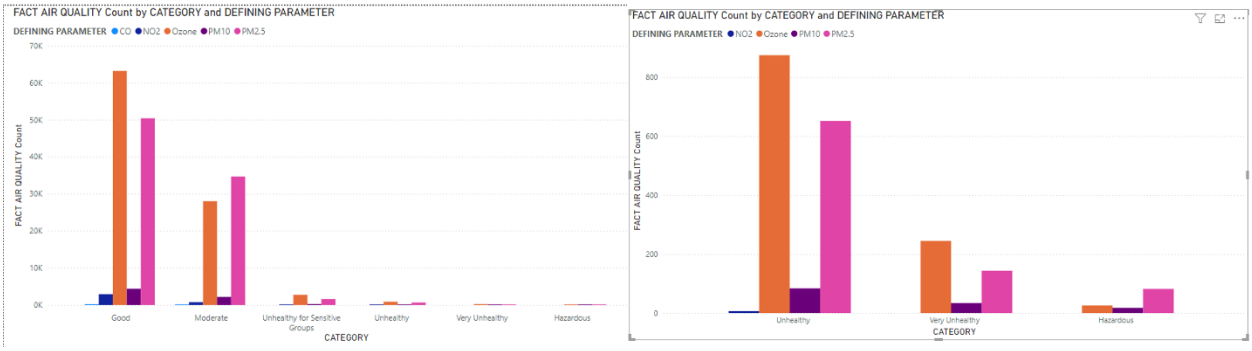


Figure 512b. Left – No. of days by Category and Defining Parameter. Right – Only for "Unhealthy" and worse.

Usually, a **country's general air quality** is **judged by** taking the cases' dominant trait. In **Fig. 512b**, the "**Good**" and "**Moderate**" categories absolutely **triumph over** the rest (96.5% of

the dataset precisely); "**Good**" is **recorded** nearly **twice as often** as "**Moderate**." So, it is acceptable to say that, altogether, **U.S. air quality** is **quite good**, and the **pollution** state is **far from extreme**. In fact, it stands in the **102nd** spot on one of **the most polluted countries in the world**'s rankings.

Now, for def. parameters, **Ozone** and **PM2.5** are always the **primary contributors** not only in **the unhealthy group** but across **all categories**. While **not** necessarily **the only ones with poor records**, **being a def. parameter ensures** they have **the worst influence** on the **site's AQI**. Therefore, **Ozone** and **PM2.5** would be **the firsts to consider** when pinpointing **main problems**. To attain **better insights**, we will **explore the sources creating these parameters** and **causing their levels** to be **so high**.

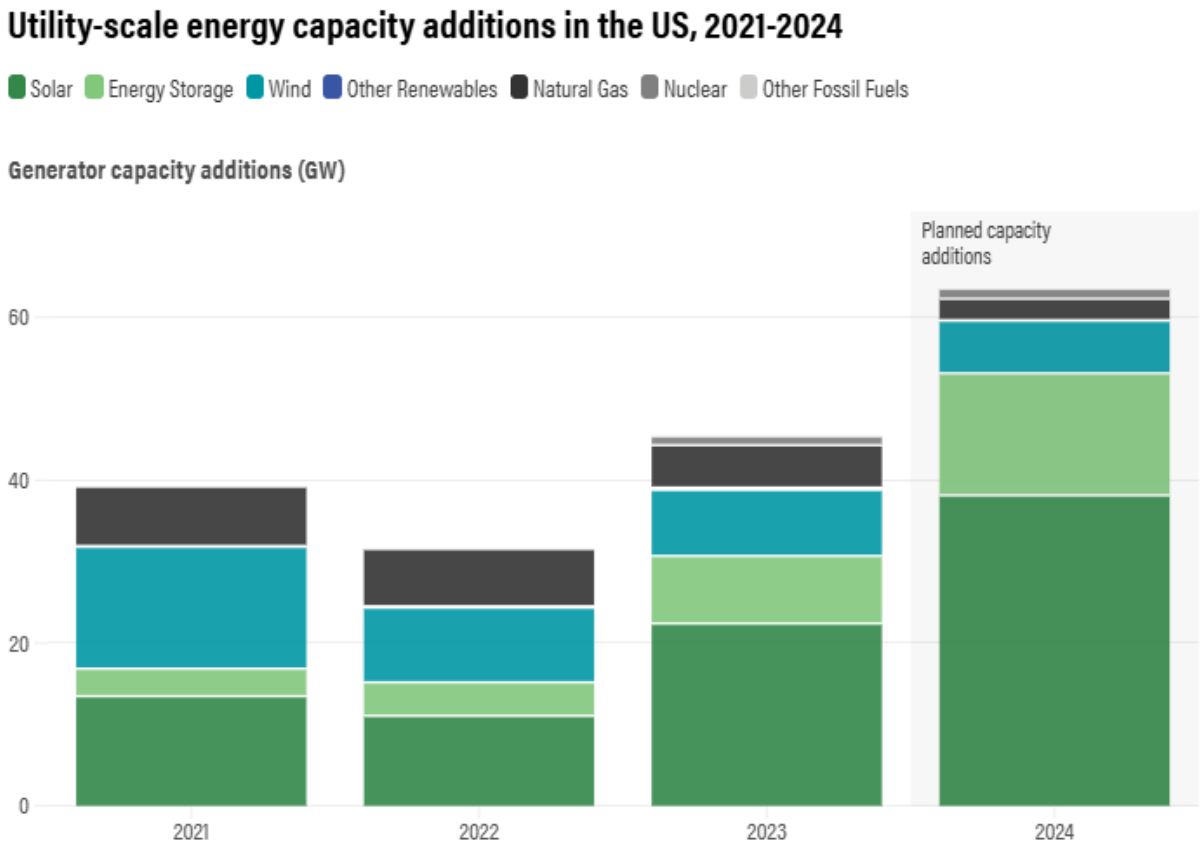


Figure 512c. U.S.'s energy capacity by types from 2021 to 2024.

- "**Ground-level ozone forms** when **nitrogen oxides** and **volatile organic** compounds **react** with each other **in sunlight** and **hot temperatures**. This pollution comes **from vehicles**, **industry**, and other sources and **contributes to smog formation**."
- "**Emissions from combustion of gasoline**, **oil**, **diesel fuel** or **wood produce** much of the **PM2.5** pollution **found in outdoor air**."

In short, they are **largely caused by demands** for **transportation**, **industrial activities**, and primary **energy sources**, while **most pollution from vehicles** and **industry comes from** the **combustion of fossil fuels** like gas, oil, and coal. In the end, the **key issue** lies in **energy consumption**. Addressing this, the **U.S. should focus** on **two chief measures**. First, **aiming to reduce unnecessary power usage**, by **promoting public transportation** and **improving the efficiency of industrial production**. Second, ultimately **transitioning to green energy sources**. In truth, the U.S. has **already taken significant stride in both areas**, rapidly **shifting toward cleaner energy sources** in recent years (**Fig. 512c**).

6 Data Mining

Data Mining to implement a model for predicting the **average AQI** of **a state's county** in **a specific month and year**. As previously observed, AQI values when broken down into daily information are inconsistent. Hence, the lowest level on which predictions should be performed is month. In this case, the applicable features are **County**, **State**, **Month**, **Year**, and the target will be **AQI**.

6.1 Preparation

To facilitate mining with only the relevant data while not affecting the warehouse's, we isolate all the particulars into **DDS.MINING_AIR_QUALITY**. This process is done simply through plain SQL queries. In SSAS, the table then acts as a new **Data Source View** from which our **Mining Structure** will extract data.

Results Messages

	AQI_ID	GEO_ID	DATE_ID	AQI	CATEGORY	DEFINING_PARAMETER	DEFINING_SITE	NUMBER_OF_SITES_REPORTING	STATE_NAME	COUNTY	YEAR	QUARTER	MONTH	DAY
1	1	90	1070	7	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	6
2	2	90	1071	8	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	7
3	3	90	1072	11	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	8
4	4	90	1073	4	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	9
5	5	90	1074	4	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	10
6	6	90	1075	7	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	11
7	7	90	1076	13	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	12
8	8	90	1077	51	Moderate	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	13
9	9	90	1078	14	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	14
10	10	90	1079	26	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	15
11	11	90	1080	12	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	16
12	12	90	1081	15	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	17
13	13	90	1082	33	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	18
14	14	90	1083	17	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	19
15	15	90	1084	22	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	20
16	16	90	1085	36	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	21
17	17	90	1086	4	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	22
18	18	90	1087	9	Good	PM2.5	01-119-0003	1	Alabama	Sumter	2023	4	12	23

Query executed successfully. (local) (15.0 RTM) DESKTOP-0VDG7HL\GIGABY... master 00:00:00 193,492 rows

Figure 61. DDS.MINING_AIR_QUALITY after data population, should have the same row count with DDS.FACT_AIR_QUALITY.

6.2 Algorithm

Decision tree is a machine learning algorithm that uses a tree-like structure to model decisions and their possible consequences. Its design consists of,

- **Nodes**: Represent conditions on the features.
- **Branches**: Represent the outcomes of those conditions.
- **Leaves**: Represent the predicted value (AQI in this case).

Decision tree splits the dataset into subsets based on the values of input features. Then, it recursively divides the data until either each subset contains a consistent set of outcomes or reaches a predefined stopping criterion.

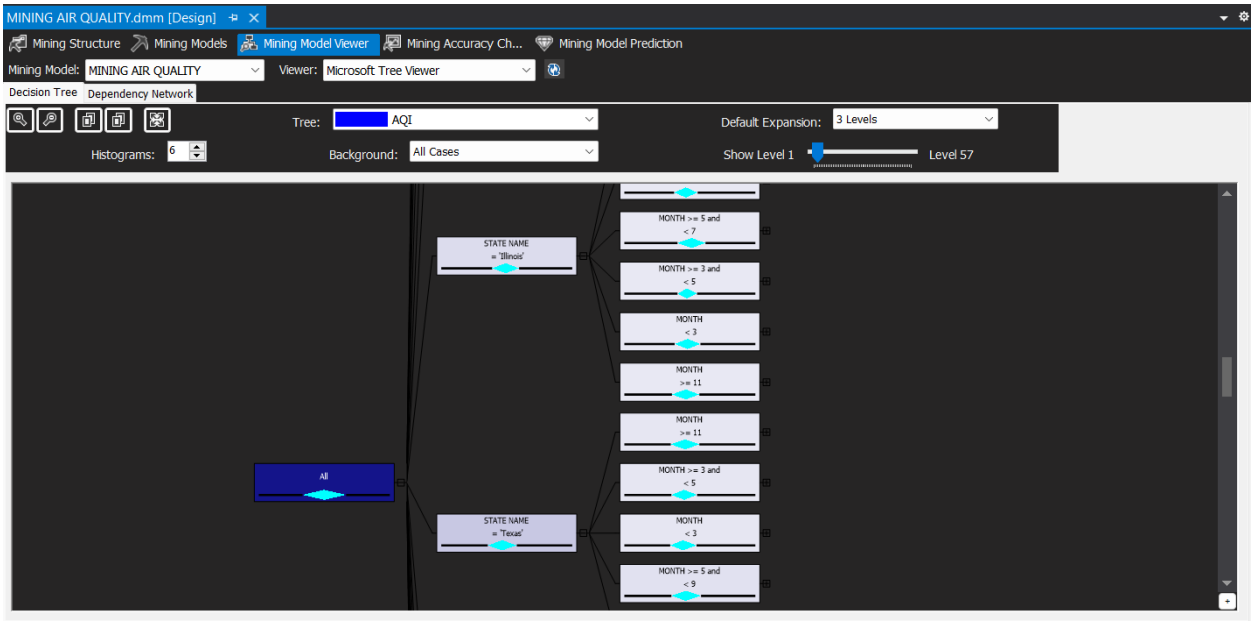


Figure 62. Decision Tree Model visualized in SSAS.

Apart from being one of the simplest algorithms available in SSAS, other reasons for Decision Tree to be our selection are,

- Easy to interpret and visualize – The influence each feature casts on AQI predictions can be clearly presented.
- Able to handle a mix of categorical data (county, state) and numerical data (month, year) – Eliminate the need to manually encode features.
- Computationally efficient – Especially for our dataset of moderate size.
- Capable of identifying the most important features for prediction – Illuminating how geographical and temporal factors affect AQI values.

Due to its nature, using Decision Tree to predict AQI categories may yield better results than the actual AQI values. However, abiding by the project's proposed approach, we still move forward with the latter.

6.3 Result

Fig. 63a compares **actual vs. predicted AQI values**, showing the model's performance. The **red diagonal line** indicates the **ideal prediction** while the **blue dots** are **predictions** made. The **overall trend** shows that **predictions align reasonably well with actual values**, especially in the lower AQI range. This suggests the **model captures general patterns** in the data **effectively**.

There is **a noticeable break away from the ideal line** for higher actual AQIs. These exemplify the **difficulties** the model is facing **when attempting to accurately predict higher AQI values**, possibly due to limited data in this range or model limitations.

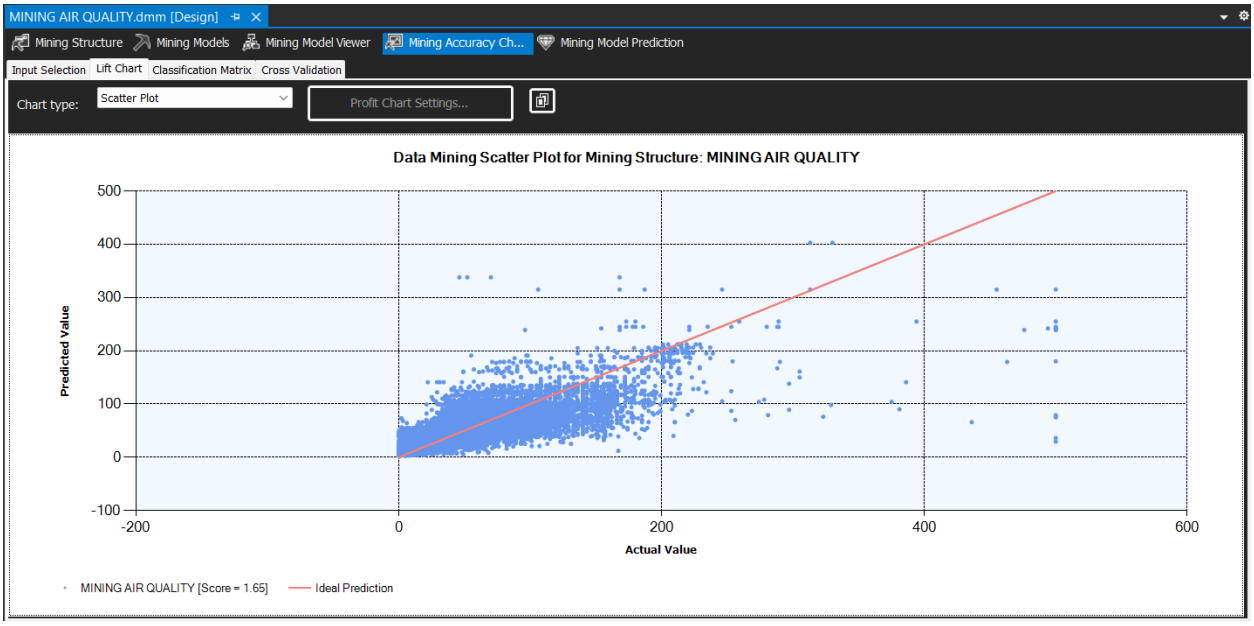
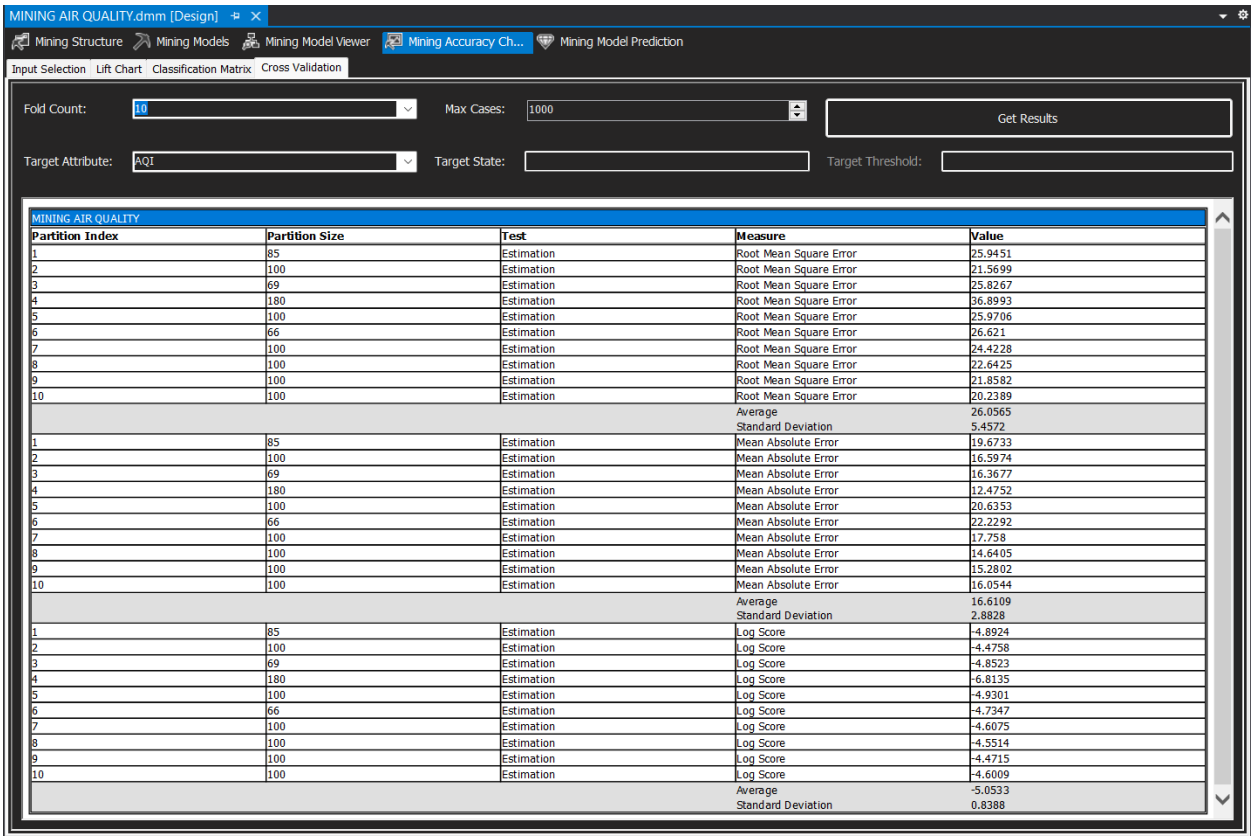


Figure 63a. Scatter plot built by SSAS.



MINING AIR QUALITY

Partition Index	Partition Size	Test	Measure	Value
1	85	Estimation	Root Mean Square Error	25.9451
2	100	Estimation	Root Mean Square Error	21.5699
3	69	Estimation	Root Mean Square Error	25.8267
4	180	Estimation	Root Mean Square Error	36.8993
5	100	Estimation	Root Mean Square Error	25.9706
6	66	Estimation	Root Mean Square Error	26.621
7	100	Estimation	Root Mean Square Error	24.4228
8	100	Estimation	Root Mean Square Error	22.6425
9	100	Estimation	Root Mean Square Error	21.8582
10	100	Estimation	Root Mean Square Error	20.2389
			Average	26.0565
			Standard Deviation	5.4572
1	85	Estimation	Mean Absolute Error	19.6733
2	100	Estimation	Mean Absolute Error	16.5974
3	69	Estimation	Mean Absolute Error	16.3677
4	180	Estimation	Mean Absolute Error	12.4752
5	100	Estimation	Mean Absolute Error	20.6353
6	66	Estimation	Mean Absolute Error	22.2292
7	100	Estimation	Mean Absolute Error	17.758
8	100	Estimation	Mean Absolute Error	14.6405
9	100	Estimation	Mean Absolute Error	15.2802
10	100	Estimation	Mean Absolute Error	16.0544
			Average	16.6109
			Standard Deviation	2.8828
1	85	Estimation	Log Score	-4.8924
2	100	Estimation	Log Score	-4.4758
3	69	Estimation	Log Score	-4.8523
4	180	Estimation	Log Score	-6.8135
5	100	Estimation	Log Score	-4.9301
6	66	Estimation	Log Score	-4.7347
7	100	Estimation	Log Score	-4.6075
8	100	Estimation	Log Score	-4.5514
9	100	Estimation	Log Score	-4.4715
10	100	Estimation	Log Score	-4.6009
			Average	-5.0533
			Standard Deviation	0.8388

Figure 63b. Cross Validation by SSAS with Max Cases of 1000 and Fold Count 10.

Fig. 63b displays another form of metricizing model performance using **Cross Validation**. With **Root Mean Square Error (RMSE)** average of **26.0565** and **Mean Absolute Error (MAE)** average of **16.6109**, the model demonstrates consistent performance across subsets, validating its robustness. Generally, the model is stable and capable of making fairly accurate predictions for the AQI value.

Conclusion

In closing, it can be inferred from previous sections' outcomes that **in 2023**, **US air quality** is **notably good** with an average AQI around **47.848** (calculated via the dataset). While **geographical** and **seasonal** factors do have many **unpredicted effects**, the **atmosphere** is **generally healthy** when judging on a country scale throughout the year.

Albeit still facing some **issues** in parameters such as **Ozone** and **PM2.5**, **America's air quality** is usually **on the cleaner end** when set side by side with other countries (Sect. 5.12). Compared to previous years, **its AQI standard** has been **successfully raised**, or at least **maintained**. Furthermore, the state is making **visible strides toward green energy** in 2023 and is expected to do even better onward. These advances together have convinced us to give **US air quality** a **strong good overall rating**.

Acknowledgement, Achievements, and potential Improvements

The authors would like to thank the lecturers of the *Information Systems for Business Intelligence* course at the Ho Chi Minh University of Science (MSc. Nguyen Ngoc Minh Chau, MSc. Tiet Gia Hong, and MSc. Ho Thi Hoang Vy) for their helpful instructions during theory and lab sessions as well as timely support during the project's duration.

Whilst there are endeavors to competently fulfill all the project's requirements (details in Introduction's Table 0), from Data comprehension, Warehouse design and Cube building, Scheduling, OLAP and Reporting, to Data mining, we have also identified a handful of areas where improvements could be made,

- Extend the phases from Source to NDS by implementing better metadata with data quality rules, firewall, and logging failures.
- Restructure metadata for a more fleshed out control of the ETL process.
- Remodel NDS's schema to further reduce repetitions.
- Explore other algorithms when data mining.

References

- **Slides, Resources, and Guides** provided during both Lecture and Lab sessions.

- Past lecture sessions and project of this course in 2023: https://www.youtube.com/@BinhPham-lh4rv

2 Data Warehouse

2.1 Source (.csv) to Stage

- Load files from folder to table: https://youtu.be/OkA85uaUiM0?si=spEKdtvuqHnB1Z42

2.2 Stage to NDS

- Refer to **SQL/3_clean_data.sql** file for certain links and their usages.
- On comparing Merge join and Lookup: https://stackoverflow.com/a/6739121
- On inner join returning more rows than expected: https://blog.sqlauthority.com/2012/02/09/sql-server-inner-join-returning-more-records-than-exists-in-table/
- On uniquely identifying records in air quality data: https://aqs.epa.gov/aqsweb/airdata/FileFormats.html#_monitors

2.3 NDS to DDS

- On updating fact tables: https://stackoverflow.com/a/67745448
- On getting all dates between two dates in MSSQL: https://stackoverflow.com/a/23291758
- On Unknown and Inferred Members, Early Arriving Fact or Late Arriving Dimension: https://youtu.be/weNidVsI6WQ?si=_iH7XTDCMT6IjJEs

3 SSAS Cube

- On SSAS's common warnings: https://www.sqlshack.com/warnings-in-ssas-cubes/
- On creating Average/Mean measure: https://insightextractor.com/2014/08/27/how-to-create-an-average-aggregation-in-sql-server-analysis-services/
- On creating Standard Deviation measure: https://dba.stackexchange.com/a/154385

4 Scheduling

- On deploying SSIS packages and scheduling: https://youtu.be/eNxbMwUGl1g?si=hdRvZUvIer9LAJkw&t=9423
- On scheduling the processing of SSAS Cube: https://youtu.be/UebBb64MAT4?si=Qp_F7DbfArwShZIe

5 OLAP and Reporting

5.1 AQI's Min and Max of states during each year's quarters

- AQI basis: https://www.airnow.gov/sites/default/files/2018-04/aqi_brochure_02_14_0.pdf
- How weather affects air quality: https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality
- On why fall's temperature rises:
  - https://edition.cnn.com/2024/08/28/weather/fall-forecast-heat-hurricane-climate/index.html
  - https://www.climatecentral.org/climate-matters/2023-fall-package

5.5 Mean AQI by quarters for 4 states

- On the worsening of Alaska's air quality in fall 2022: https://dec.alaska.gov/commish/newsroom/22-02-dec-fnsb-release-annual-report-on-air-quality-improvement-efforts

5.10 AQI by state, Category, DayLightSaving over years

- U.S report about impact of DST on energy consumption: https://www.energy.gov/sites/default/files/2022-04/Impact%20of%20Extended%20Daylight%20Saving%20Time%20on%20National%20Energy%20Consumption%2C%20Report%20to%20Congress.pdf

5.11 No. of days by state, Category in each month

- On causes of air pollution in California: https://ww2.arb.ca.gov/resources/sources-air-pollution

5.12 No. of days by Category and Defining Parameter

- Country air quality ranking (updated Nov. 21th 2024): https://www.iqair.com/world-most-polluted-countries?srsltid=AfmBOor5s420GPqcoNquHpLntequCI0cKyO7cG_sj6iCTZFMHWYioAtp
- On causes of Ozone pollution: https://ecology.wa.gov/air-climate/air-quality/air-quality-targets/air-quality-standards/ozone-pollution
- On causes of PM2.5 pollution: https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health
- On U.S.'s transition to clean energy: https://www.wri.org/insights/clean-energy-progress-united-states

6 Data Mining

- SSAS Data Mining Overview: https://youtu.be/C8NCvuufhOg?si=xN6BLxgAbYVqHiCN
- On Decision Tree: https://www.ultralytics.com/glossary/decision-tree