**AWS Project Report**

***Tal Benhamou, Sheng-Han Yueh, Bitterlein Konnoth Biju, Yudhishna Kuppala***

***Group 1 Section 1***

# **Table of contents**

[**Table of contents 2**](#_9bfp1w93l9k)

[**Introduction: 3**](#_76b3ukjbys50)

[**Datasets informations: 3**](#_tr5wigks7xfz)

[**Conceptual Model: 4**](#_dmsw2qv8pnrq)

[**AWS Solution: 5**](#_e373r2e3ms3g)

[**Data Retrieval and Pre-processing: 6**](#_9n5n1vxoqiry)

[**ETL Process: 7**](#_exon6ywgr0b4)

[**Analysis: 8**](#_u7vffycchf41)

[1. How does the presence of high-value Certificate of Fitness (COF) types, correlate with the average sale prices of properties? 8](#_6ndusx7s5xvu)

[2. How do the average sale prices vary across different boroughs for specific Fire Safety Inspection Certificate types, and what insights can be drawn regarding the property characteristics and market trends within each borough based on these variations? 9](#_sab70vm9s3sf)

[3. How do the count of commercial units and the sale prices of properties associated with specific Fire Safety Inspection Certificate (Cof) types in Manhattan correlate, and what implications does this relationship have for investment strategies and future commercial development in the area? 11](#_jfk96xg6dizi)

[4. What is the relationship between the average sale price of properties and the number of residential units associated with specific Fire Safety Inspection Certificate (Cof) types, and how can this correlation inform investment decisions and development strategies in the residential real estate market? 12](#_p6gz55mh1h3a)

[**Conclusion: 13**](#_5h4wx6lykf4w)

[**Sources: 14**](#_ryxrk151fgx8)

# **Introduction:**

For this project, we selected three datasets concerning fire prevention and apartment sales in New York to create an AWS environment for analyzing the safest boroughs in New York. The datasets included are as follows:

1. Fire Inspection (API)

2. Bureau of Fire Prevention - Certificates of Fitness (Historical), comprising 128,293 records and 18 variables.

3. NYC Citywide Annualized Calendar Sales Update, comprising 606,260 records and 28 variables.

The source of the API is [source], while the other two files are from NYC's open-source platform (NYC OpenData). For this project, we utilized the Amazon Web Services platform to upload these files.

This led us to the following question: How do fire incident history, property values, and community safety interact within New York City neighborhoods, and what insights can be gleaned to identify the safest boroughs?

# **Datasets informations:**

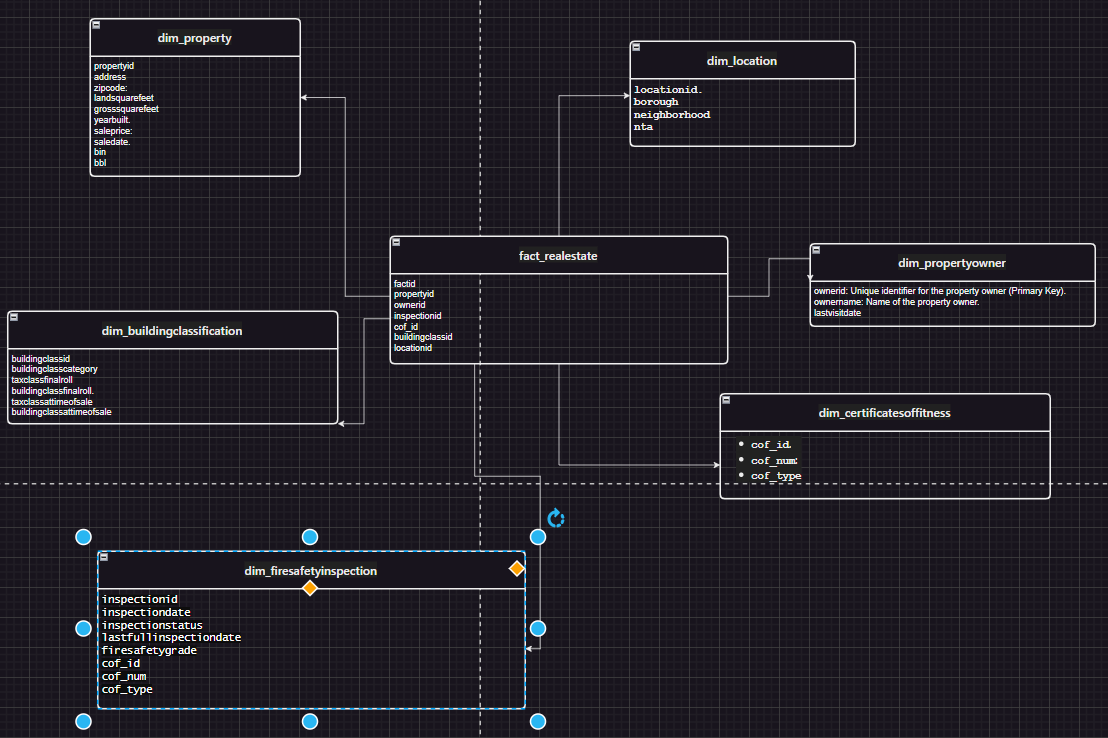
The dataset "NYC Citywide Annualized Calendar Sales Update" as of April 4, 2024, comprises 606,260 records across 31 fields, including address, building classification, sale prices, and sale date.

The fire prevention file consists of 128,963 observations and 17 variables, such as types of certification, date of certification, building block, street number, and more.

The dataset "Bureau of Fire Prevention - Certificates of Fitness Historical" as of April 4, 2024, contains 128,293 records. Each record spans 18 fields, including the certificate number, type, holder name, expiration date, as well as the address and geographical coordinates of the premises.

All the datasets were merged based on the Building Identification Number variable ‘BIN’.

# **Conceptual Model:**



Our conceptual model represents a structured database designed for managing real estate and fire safety information. The model comprises several dimension tables and a few fact tables that facilitate detailed reporting and analysis.

**Dimension Tables:**

* dim\_property: Contains detailed attributes of properties, such as property ID, address, zip code, gross square feet, land square feet, year built, sale price, sale date, bin, and bbl.
* dim\_buildingclassification: Includes building classification details like building class ID, category, classification role, tax class at time of sale, and building classification at the time of sale.
* dim\_location: Holds location-specific data, such as location ID, borough, neighborhood, and other location attributes.
* dim\_propertyowner: Stores information about property owners, including a unique identifier for the owner, the owner's name, and the last update date for ownership data.
* dim\_certificatesoffitness: Consists of data related to certificates of fitness, including certificate ID, number, and type.
* dim\_firesafetyinspection: Contains details from fire safety inspections, including inspection ID, inspection date, inspection status, last full inspection date, fire safety grade, and associated certificate IDs.

**Fact Tables:**

* fact\_realestate: Acts as a transactional ledger recording details from various dimension tables such as property ID, owner ID, building classification ID, certificate ID, and location ID, facilitating a comprehensive view of real estate transactions and attributes.

This model is thought to be well-organized to handle complex queries and reporting needs related to property management, fire safety compliance, and real estate transactions. Each dimension table serves a specific aspect of the real estate and fire safety domains, ensuring data integrity and supporting detailed analytics.

# 

# **Physical model:**

# 

# 

# 

# 

# **AWS Solution:**

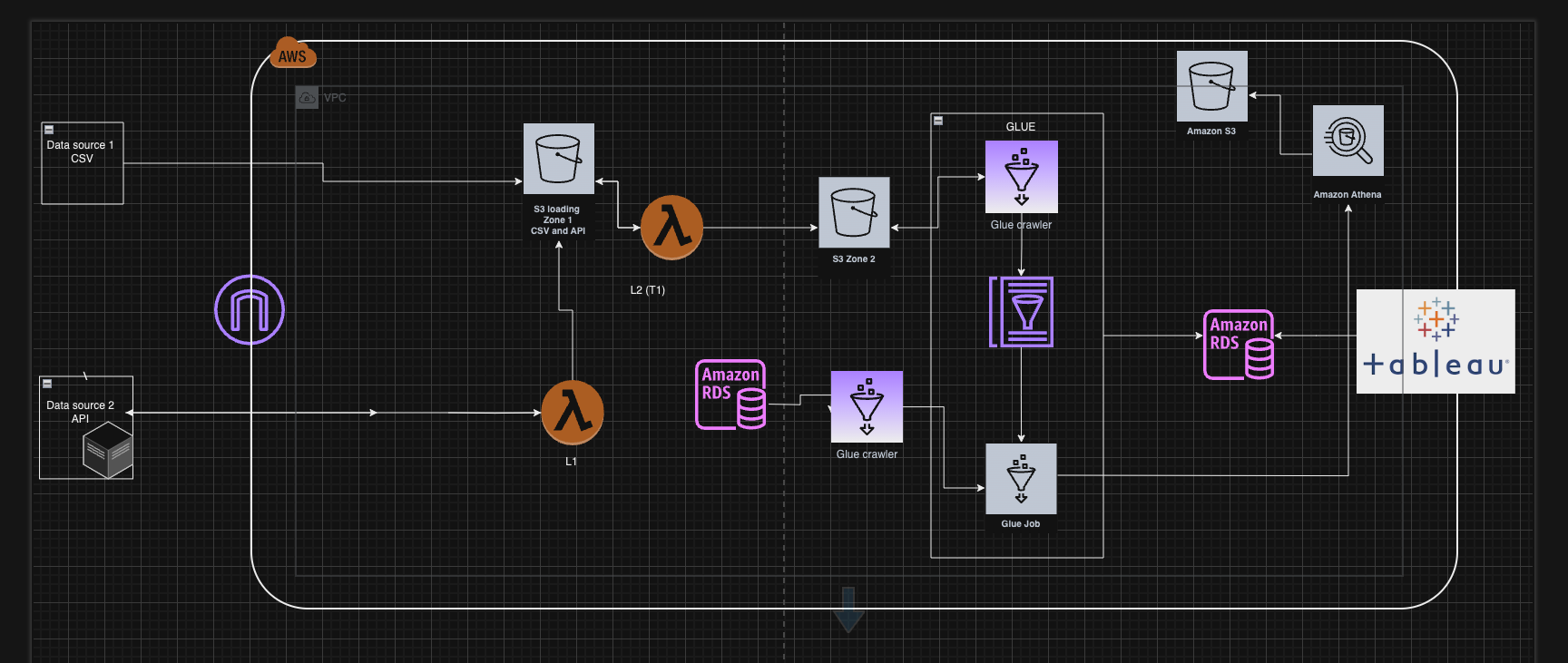
In the cloud architecture presented, data is ingested and processed through a series of coordinated services on AWS. Initially, two sources provide data: the first directly offers CSV files, while the second delivers data via an API. Both data streams are directed into an AWS S3 bucket in the us-east-1 region, serving as the primary collection and staging area.

The process begins with the "L1" Lambda function, which retrieves data from the API and likely stages it into S3. From there, the "L2" Lambda function takes over, handling tasks such as cleaning, validating, transforming, and merging both the API and CSV data. This ensures the data is uniform and of high quality, suitable for advanced processing.

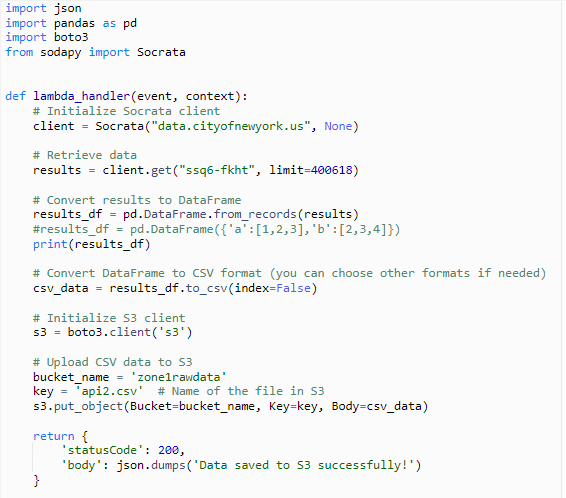
Once the data is processed by the L2 function, it moves to another S3 bucket, referred to as "S3 Zone 2," where it is stored for subsequent operations. AWS Glue then plays a crucial role, starting with a Glue Crawler that examines the data to deduce schema and other metadata, storing this information in the Glue Catalogue. This metadata integration supports the Glue Jobs that follow, performing scalable and managed data transformations.

Post transformation, the data is managed through two pathways: it is stored in Amazon RDS for structured query operations, and simultaneously, it is available in S3 for ad-hoc querying via Amazon Athena. This setup provides robust data querying capabilities.

Finally, the transformed and organized data is visualized through AWS QuickSight, offering rich, interactive visualizations and insights. This comprehensive setup not only facilitates extensive data analysis from varied inputs but also supports dynamic, data-driven decision-making within the organization. All these operations are securely managed within a VPC in the AWS us-east-1 region, ensuring data integrity and security.



# **Data Retrieval and Pre-processing:**



We used a two-step process using AWS Lambda functions to handle and preprocess data.

**Lambda 1:**

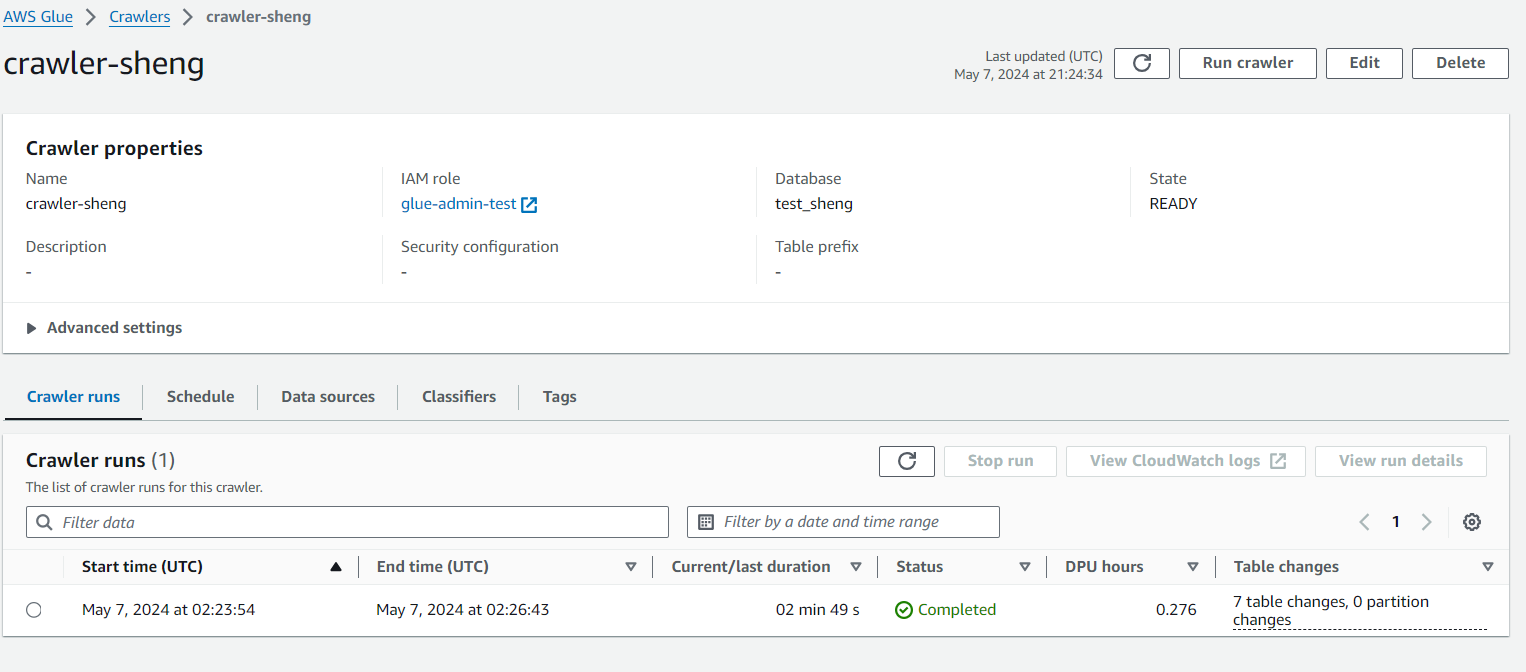
This function is responsible for retrieving data. It uses the Python libraries pandas and *sodapy* to interact with an API and process the data. After fetching the data from the API, Lambda 1 converts the data into a DataFrame format using pandas. It then converts the DataFrame to a CSV format, which is subsequently stored in an Amazon S3 bucket located in Zone 1.

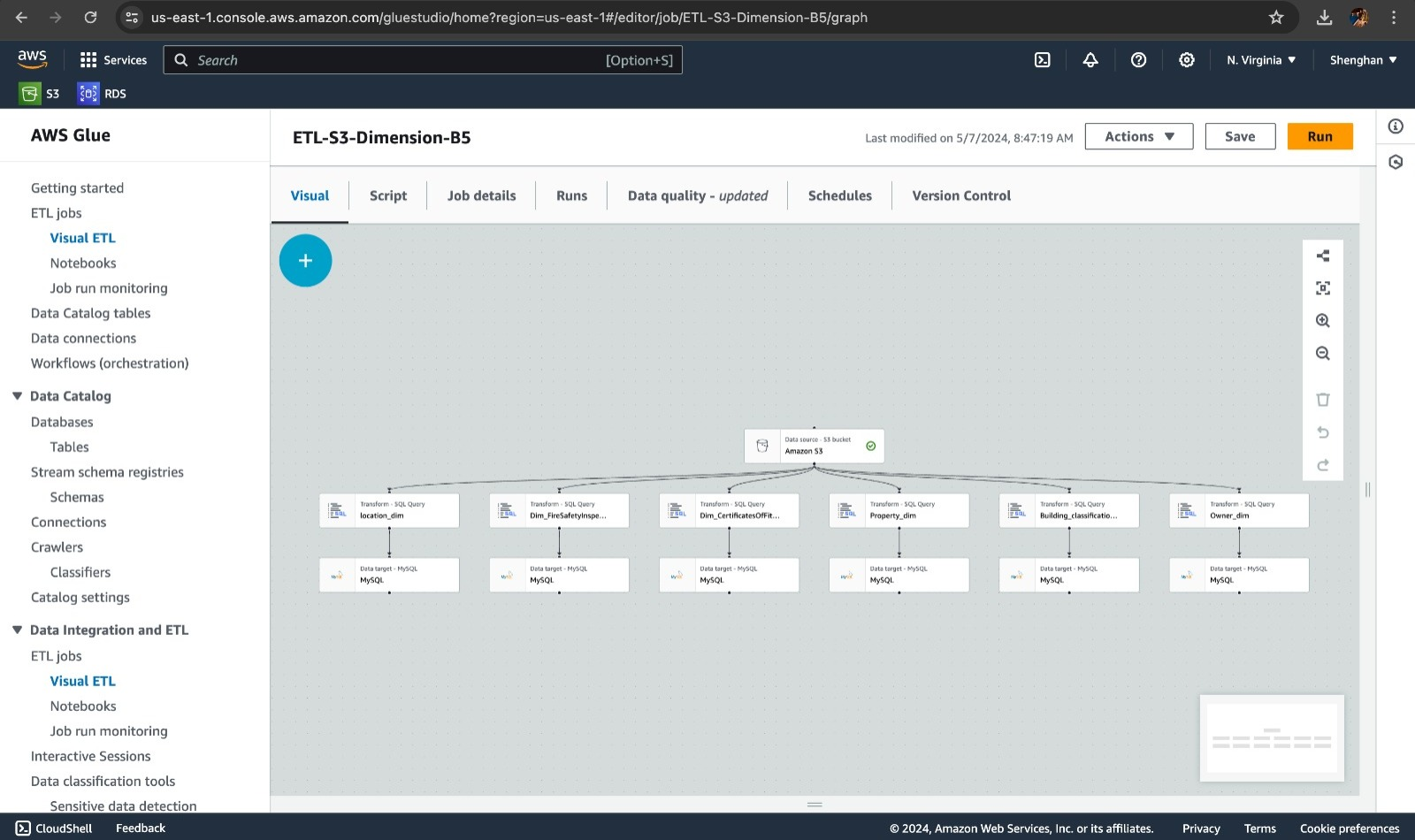
**Lambda 2:**

The second Lambda function takes over from where the first one left off. It reads the CSV file from the S3 bucket in Zone 1. The main tasks of Lambda 2 are to clean the data, handle missing values (imputation), and perform data mapping such as mapping borough names to their respective codes. After these preprocessing steps, Lambda 2 merges the cleaned data with other relevant datasets as needed for further analysis or processing. Finally, the processed data is stored in another S3 bucket, designated as Zone 2.

This process streamlines the workflow from data retrieval to preprocessing, ensuring that the data is ready for analysis in a clean and structured format.

# **ETL Process:**







We have implemented an AWS Glue crawler named `crawler-sheng`, which is configured to work with our database named `test\_sheng` under the IAM role `glue-admin-test`. This crawler is crucial for identifying and processing changes in our data landscape, effectively updating or ingesting new data into our systems. It recently completed a run that resulted in 7 table changes, indicating significant updates and data ingestion activities.

In the transformation and loading phases, we utilize AWS Glue's visual interface to configure and manage multiple ETL jobs. These jobs are specifically designed to handle data from Amazon S3 sources, transforming it through a series of SQL query manipulations to fit our operational and analytical needs. The transformed data includes various dimensions such as `Dim\_FireSafetyInspection`, `Dim\_CertificatesOfFitness`, `Property\_dim`, and others, which are then structured appropriately for storage in MySQL databases.

Our ETL jobs are part of a broader workflow within AWS Glue that ensures not only the transformation but also the orchestration of data processing tasks. This setup allows us to maintain a flexible, scalable, and efficient data management environment. It supports our analytics and business intelligence capabilities by ensuring that data is consistently processed, transformed, and made ready for analysis and decision-making processes.

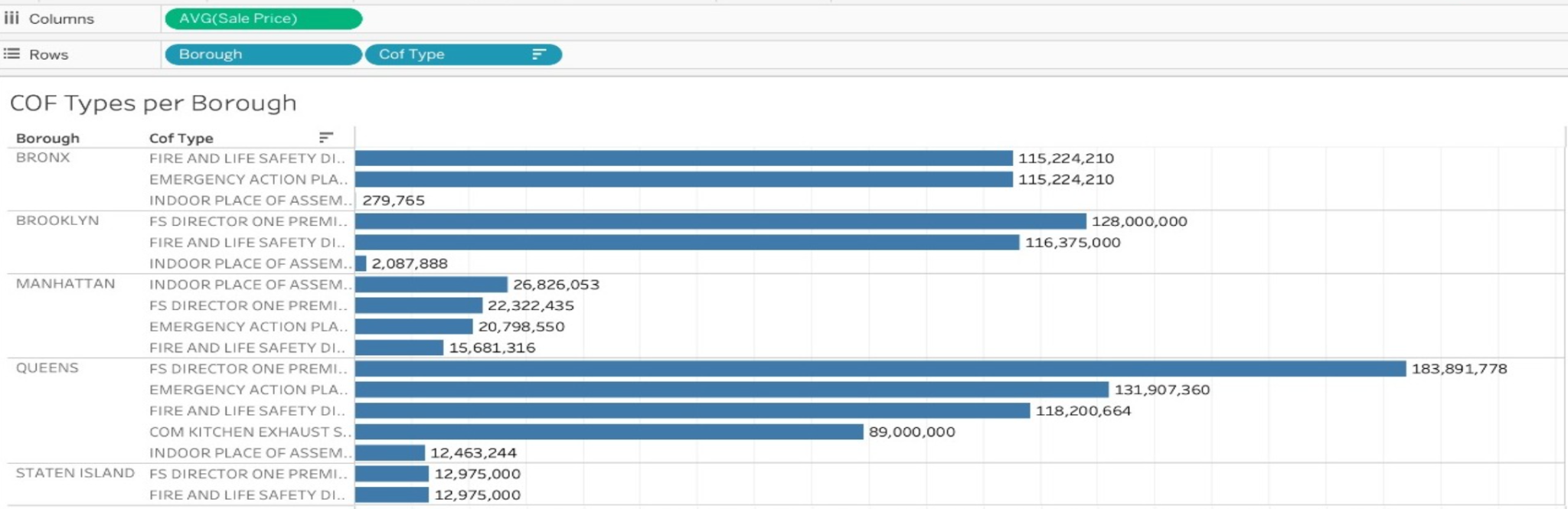
Overall, this approach exemplifies our commitment to leveraging advanced cloud-based technologies to handle complex data transformations, ensuring that our data is accurate, timely, and aligned with our strategic objectives.

# **Analysis:**

In our analysis, we answered different questions:

## ***How does the presence of high-value Certificate of Fitness (COF) types, correlate with the average sale prices of properties?***

The data from New York City reveals a correlation between the presence of high-value Certificate of Fitness (COF) types and the average sale prices of properties across different boroughs. Manhattan stands out with the highest property sale prices; for instance, properties associated with "Indoor Place of Assembly FS Director One Premise" COF type have an average sale price of approximately $183.9 million. Other COFs in Manhattan, like "Emergency Action Plan Director," also show high average sale prices exceeding $115 million. Brooklyn and Queens display moderate correlations, with significant property values linked to similar COF types, around $128 million in Brooklyn and $131.9 million in Queens. In contrast, the Bronx and Staten Island show lower average sale prices for similar COFs. This trend suggests that properties requiring extensive safety certifications tend to command higher market values, especially in areas like Manhattan, where stringent safety measures are paramount due to high-density environments.

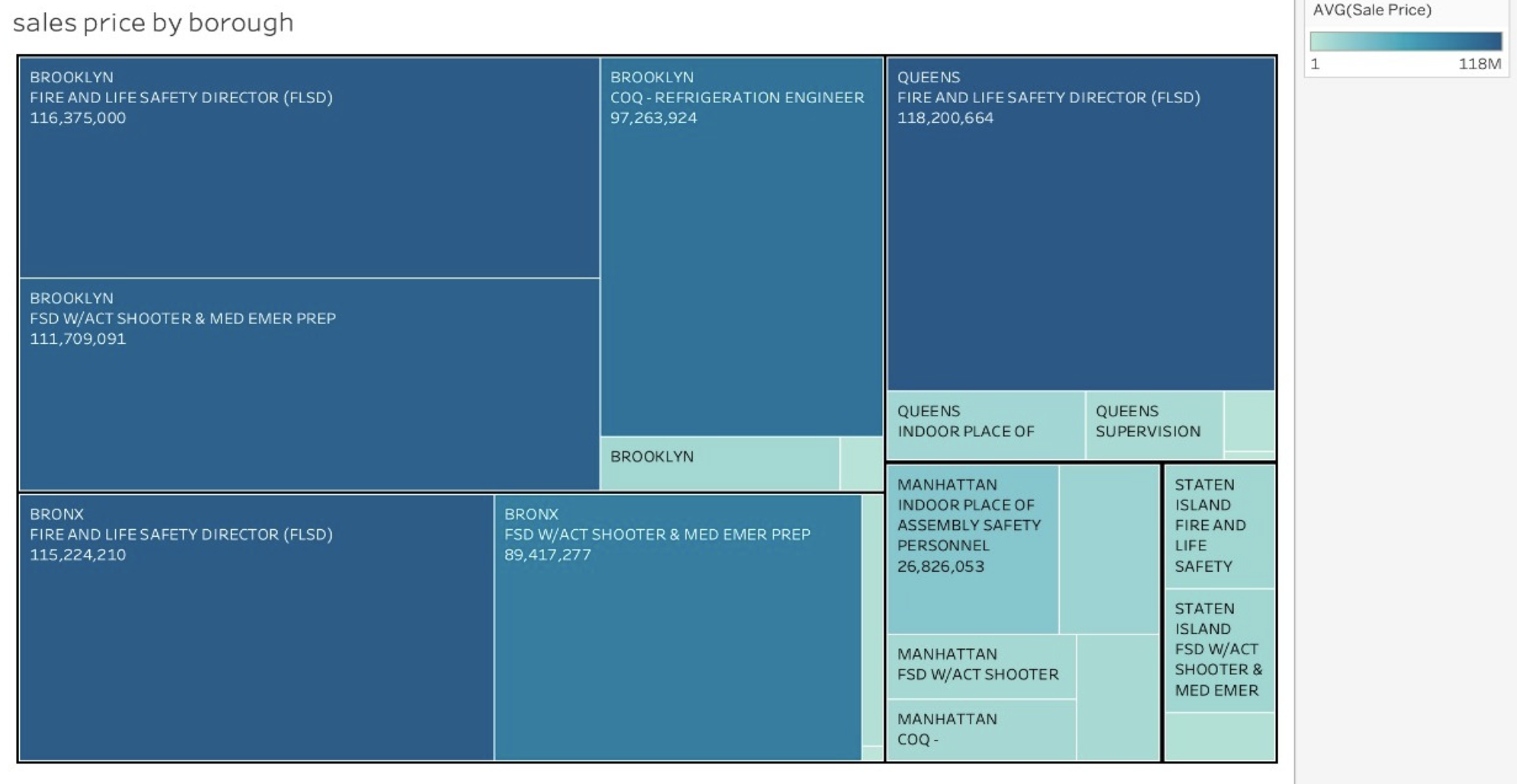


## ***How do the average sale prices vary across different boroughs for specific Fire Safety Inspection Certificate types, and what insights can be drawn regarding the property characteristics and market trends within each borough based on these variations?***

The treemap visualization of average sale prices for properties with specific Fire Safety Inspection Certificates (COF) across New York City boroughs reveals significant insights into market trends and property characteristics:

* Manhattan shows the highest property sale prices, especially for properties requiring extensive safety measures, like "Indoor Place of Assembly Safety Personnel" at $26,826,053. This underscores Manhattan's high-density, high-value real estate market where large, strategically located properties necessitate advanced safety features.
* Brooklyn presents a diverse range of high-value COF types with substantial sale prices, such as "Fire and Life Safety Director (FLSD)" at $116,375,000. This diversity highlights Brooklyn's evolution into a mixed-use hub that prioritizes safety in both residential and commercial developments.
* Queens also exhibits high sale prices for safety-focused properties, indicating robust demand for modern amenities and safety in its mixed-use developments, reflecting its growing commercial and residential sectors.
* Bronx and Staten Island display lower sale prices for similar COF types compared to Manhattan and Brooklyn, suggesting markets with potentially lower property values but a similar emphasis on safety, influenced by economic factors and local commercial activities.

Overall, high-value COF types correlate with higher property sale prices, particularly in densely populated boroughs with active commercial sectors like Manhattan and Brooklyn. The data implies that safety certifications are highly valued across NYC, driven by regulatory demands and a premium on tenant safety in high-risk environments. These insights are crucial for investors and developers planning to capitalize on market demands and safety trends in various boroughs.

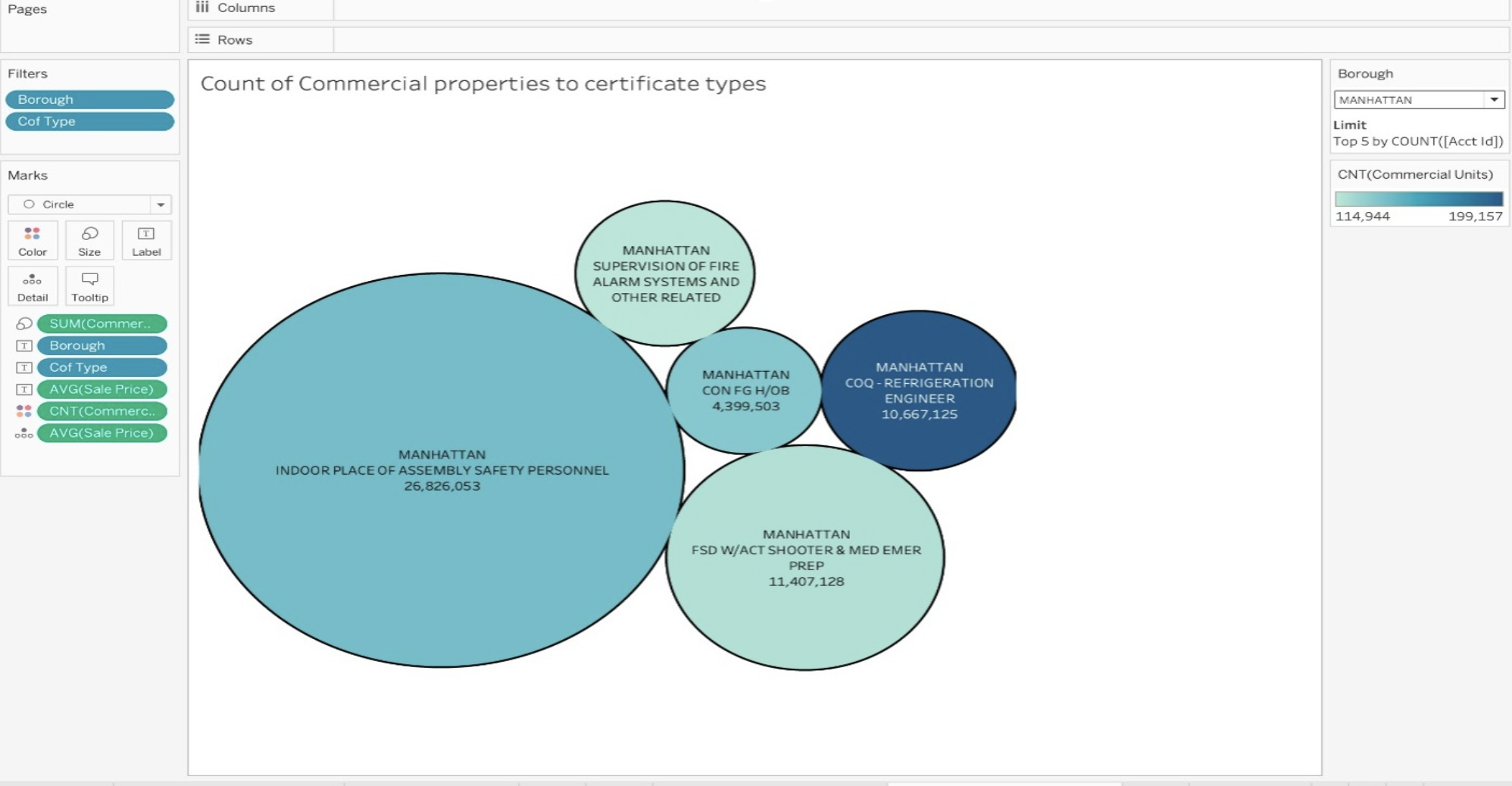


## ***How do the count of commercial units and the sale prices of properties associated with specific Fire Safety Inspection Certificate (Cof) types in Manhattan correlate, and what implications does this relationship have for investment strategies and future commercial development in the area?***

The data from Manhattan shows a clear correlation between the count of commercial units and the sale prices of properties with specific Fire Safety Inspection Certificates (COF) types. Notably, properties associated with the "Indoor Place of Assembly Safety Personnel" COF type indicate a high value, with an average sale price of $26,826,053, suggesting these units are highly valued for their comprehensive safety measures.

Other COF types, like "FSD w/Active Shooter & Medical Emergency Preparedness" and "Supervision of Fire Alarm Systems," also command high sale prices, pointing to the market's premium on advanced safety features. This trend implies that properties equipped with top-tier safety certifications are seen as lower-risk and potentially higher-return investments, attracting premium tenants and justifying higher lease rates.

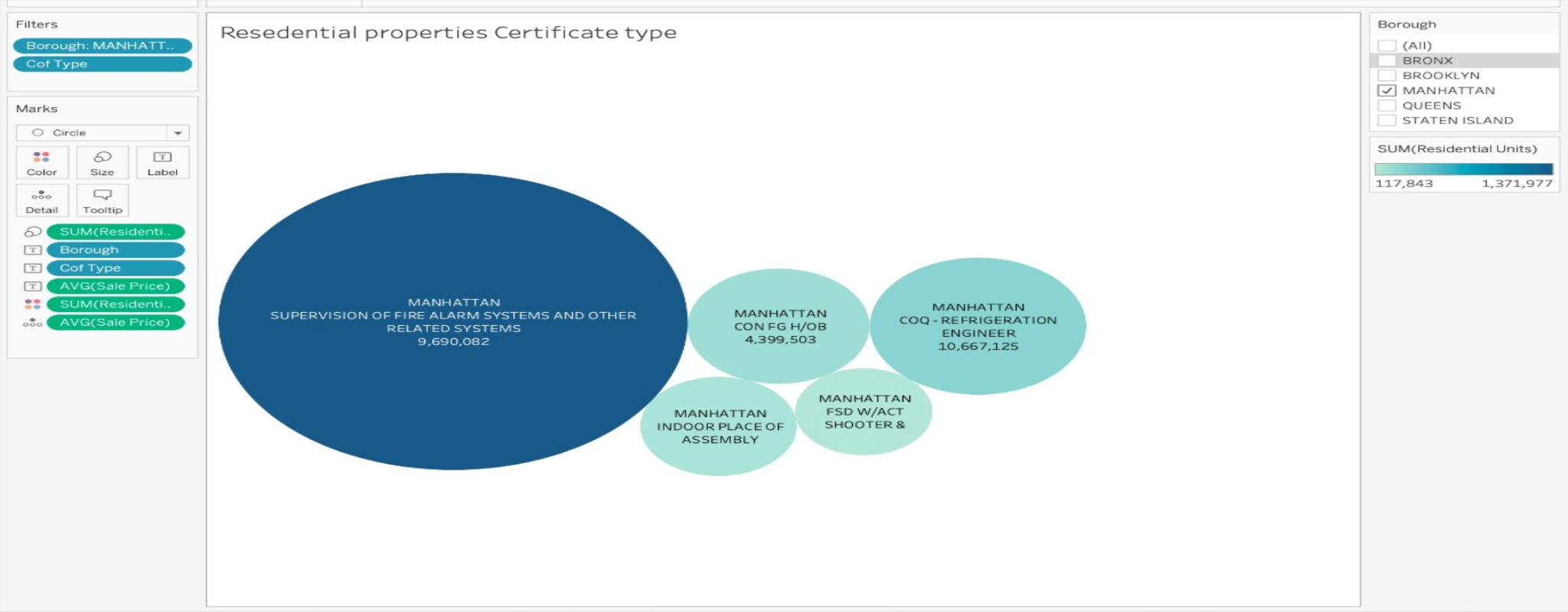
For investors, this correlation suggests that investing in properties with stringent safety features in Manhattan could be highly lucrative, offering stability and appeal due to mandatory safety standards. The strategic focus on enhancing safety features in commercial properties could influence future development trends in Manhattan, emphasizing the integration of advanced safety technologies to meet regulatory demands and increase property values. This approach not only meets a critical market demand but also positions these properties as attractive investment opportunities in a competitive real estate market.



## ***What is the relationship between the average sale price of properties and the number of residential units associated with specific Fire Safety Inspection Certificate (Cof) types, and how can this correlation inform investment decisions and development strategies in the residential real estate market?***

The data from Manhattan highlights the relationship between property sale prices and residential units with specific Fire Safety Inspection Certificates (COF). Properties equipped with advanced fire safety systems, such as "Supervision of Fire Alarm Systems and Other Related Systems," command a significant market share and a high average sale price of $9,690,082. Other specialized COF types like "CON FG H/OB" and "COQ - Refrigeration Engineer" also fetch high prices, reflecting the market's valuation of specialized safety features.

This trend suggests that safety enhancements in residential properties can substantially increase property values. Investors and developers can leverage this by integrating advanced safety systems to attract a higher-paying tenant base and increase investment returns. The premium on safety in Manhattan’s residential market is evident, indicating that properties with comprehensive safety and emergency preparedness features are highly sought after and command higher prices. This insight is crucial for shaping investment strategies and development plans in urban real estate markets.



# **Conclusion:**

In conclusion, this project has successfully demonstrated the power of integrating fire prevention and real estate datasets in a sophisticated AWS environment to discern the safest boroughs in New York City. By merging comprehensive data from fire inspections, certificates of fitness, and real estate transactions, we have developed a robust analytical framework that highlights the interplay between fire safety measures and real estate dynamics. Our findings indicate significant correlations between fire safety certifications and property values, underscoring the importance of rigorous safety standards in enhancing property desirability and market value.

Furthermore, the deployment of AWS technologies like Lambda functions, S3 buckets, AWS Glue, and QuickSight for data handling and visualization has streamlined our data processing and improved our analytical capabilities. This has allowed us to not only answer complex questions about safety and property values but also to provide actionable insights that can guide policy-making and real estate investments.

Ultimately, this project reflects a comprehensive approach to urban safety and real estate management, offering a blueprint for similar analyses in other metropolitan areas. The methodologies and technologies applied here promise to enhance urban planning and safety strategies, making a significant impact on the way cities manage and utilize their real estate and safety data for the betterment of their communities.

# **Sources:**

<https://data.cityofnewyork.us/Public-Safety/Bureau-of-Fire-Prevention-Inspections-Historical-/ssq6-fkht/about_data>

<https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu/about_data>

<https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu/about_data>