

BDAD Project

Data Ingestion, Cleaning and Profiling Report

Amitkumar Dineshbhai Patel (ap7986)

Property Valuation and Assessment Data

About the Dataset

Dataset Name: Property Valuation and Assessment Data Tax Classes 1, 2, 3, 4

Agency Name: Department of Finance

Update Frequency: Annually

Dataset Description: Real Estate Assessment Property data

Dataset Keywords: Property, Assessment, Evaluation

Dataset Category: City Government

This dataset contains detailed information on property valuation and assessment data for tax classes 1, 2, 3, and 4 in New York City. It includes market-assessed values, tax values, exemption values, and other related metrics necessary for calculating property tax, determining eligibility for exemptions, and assessing property valuations. The dataset is maintained by the Department of Finance and is updated annually.

The dataset is crucial for understanding the financial aspects of property ownership and taxation in New York City. It serves as a valuable resource for property owners, real estate professionals, policymakers, and researchers interested in analyzing property market trends, evaluating tax implications, and assessing the economic value of real estate assets.

For our Big Data Analysis project, we selected a specific range of data from the tax years 2021 to 2023 (years 2020 to 2022) to study the patterns across two other datasets: Google restaurant reviews and subway ridership. This selection allows us to analyze the impact of crime complaints on property valuations in NYC neighborhoods

during the same time frame. By integrating property valuation data with crime complaint data, restaurant reviews, and subway ridership data, we aim to uncover correlations and insights that can inform urban planning, public safety measures, and real estate investment strategies. This multi-dimensional analysis harnesses the power of big data to provide a comprehensive understanding of the complex interplay between crime, property values, consumer behavior, and public transportation in urban environments like New York City.

Relevance to Project Theme

Our project theme revolves around determining the impact of crime complaints on various aspects of New York City neighborhoods, including property valuation, restaurant reviews, and subway ridership. By leveraging the Property Valuation and Assessment Data from tax years 2021 to 2023, we can assess how crime incidents influence property values across different neighborhoods.

This dataset serves as a fundamental component in understanding the financial dynamics of urban areas, providing insights into how crime rates correlate with property assessments. By analyzing this data alongside crime complaint records, restaurant reviews, and subway ridership data, we aim to uncover patterns and relationships that illuminate the broader socio-economic landscape of NYC neighborhoods.

Through our Big Data Analysis approach, we seek to offer valuable insights that can inform policy decisions, urban development strategies, and real estate investment choices, ultimately contributing to the creation of safer, more prosperous communities.

Data Procurement

Initial Filtering at Source

To streamline the export process from the NYC Open Data querying tool, I selected and exported the following columns from the dataset:

- PARID
- BORO

- BLOCK
- LOT
- YEAR
- PYMKTLAND
- PYMKTTOT
- PYACTLAND
- PYACTTOT
- PYACTEXTOT
- PYTRNLAND
- PYTRNTOT
- PYTRNEXTOT
- PYTXBTOT
- PYTXBEXTOT
- FINMKTLAND
- FINMKTTOT
- FINACTLAND
- FINACTTOT
- FINACTEXTOT
- FINTRNLAND
- FINTRNTOT
- FINTRNEXTOT
- FINTXBTOT
- FINTXBEXTOT
- CURMKTLAND
- CURMKTTOT
- CURACTLAND
- CURACTTOT
- CURACTEXTOT
- CURTRNLAND
- CURTRNTOT
- CURTRNEXTOT
- CURTXBTOT
- CURTXBEXTOT
- HOUSENUM_LO
- HOUSENUM_HI
- STREET_NAME
- ZIP_CODE
- LAND_AREA
- NUM_BLDGS

- YRBUILT
- EXTRACRDT
- BLD_STORY
- UNITS
- GROSS_SQFT

These columns were chosen to provide comprehensive information on property valuation and assessment, including market-assessed values, tax values, exemption values, and physical characteristics of the properties. This subset of columns allows for meaningful analysis while managing the size of the dataset for easier handling and processing.

Data Ingestion

After exporting the CSV files containing property valuation and assessment data for tax years 2021-2023 from the querying tool of NYC Open Data, I initiated the data ingestion process. This involved uploading the exported CSV files to the data ingest server. Once uploaded, I utilized relevant commands to transfer the datasets to my Hadoop Distributed File System (HDFS) workspace. The datasets were stored under a single directory within the HDFS, facilitating easy access and management for subsequent data processing tasks. This ingestion process ensured that the datasets were readily available for further analysis and exploration using big data technologies and tools.

Data Cleaning

The data cleaning process involved several steps to ensure the quality and consistency of the property valuation and assessment data:

1. **Deduplication:** The raw data was checked for duplicate records using Spark DataFrame's `dropDuplicates` function.
2. **Column Dropping:** Certain columns deemed irrelevant or redundant for the analysis were dropped from the dataset. These included details such as specific house numbers, street names, and various property assessment values.
3. **Zero Value Filtering:** Rows containing zero values in critical columns related to property attributes, such as land area, building area, and total assessed values, were filtered out to remove invalid entries.

4. Borough and Block Filtering: Data integrity was maintained by filtering out records where the block numbers were not within the valid ranges corresponding to their respective boroughs.
5. ZIP Code Correction: ZIP codes were standardized and corrected where possible. Rows with missing or invalid ZIP codes were addressed manually. In cases where the ZIP code was null but the Borough, Block, and Lot (BBL) data were available, external queries were made to the NYC Department of Finance's Property Information Portal (<https://propertyinformationportal.nyc.gov/>) to obtain the missing ZIP code. Geolocation services were utilized when direct ZIP code information was not available online.
6. ZIP Code Trimming: ZIP codes with lengths greater than 5 characters were trimmed to ensure uniformity.
7. Invalid ZIP Code Removal: Rows with invalid or zero ZIP codes, for which automated extraction was not feasible, were removed from the dataset to maintain accuracy.

Data Profiling

The data profiling phase involved analyzing the properties of individual columns within the cleaned dataset:

- Overall Statistics: Descriptive statistics were computed for columns expected to contain numerical data, including mean, minimum, maximum, and standard deviation.
- Borough-wise Analysis: Statistical insights were derived for numerical columns based on boroughs, providing a deeper understanding of property valuation trends across different areas.
- Distinct Value Counts: For categorical or non-numerical columns, the count of distinct values and their occurrences were examined to identify any anomalies or patterns.

The cleaning and profiling processes ensured that the property valuation dataset was prepared for subsequent analysis and integration with other datasets for the overarching project on evaluating the impact of crime complaints on various neighborhood factors.

Insights from Cleaning and Profiling

After processing a raw dataset comprising 6,883,138 rows, several cleaning and profiling steps were undertaken. Initially, the dataset was checked for duplicates, but no duplicate rows were found. Subsequently, specific columns deemed irrelevant for analysis were dropped, including HOUSENUM_LO, HOUSENUM_HI, STREET_NAME, and various others, resulting in a dataset of 6,883,138 rows.

Filtering was then applied to remove rows with invalid values in critical columns, resulting in a dataset of 5,419,134 rows. Further refinement involved filtering out rows with invalid block numbers, maintaining the same number of rows.

Another cleaning step involved deleting rows with invalid zip codes, which reduced the dataset slightly to 5,416,467 rows. This process ensured data integrity by removing entries with incomplete or inaccurate location information.

Profiling the dataset provided valuable insights into its composition. For instance, the distribution of properties across boroughs revealed significant disparities, with Queens and Brooklyn having the highest number of properties, followed by Staten Island, the Bronx, and Manhattan. Similarly, analysis of block and lot numbers demonstrated a wide range of values, highlighting the diversity in property sizes and configurations.

Moreover, borough-wise statistics for property valuation metrics such as CURMKTLAND and CURMKTTOT showcased considerable variation, with Manhattan consistently exhibiting higher average values compared to other boroughs. This suggests differing property market dynamics and investment potential across New York City.

Additionally, examining attributes like land area and the number of buildings per property provided insights into spatial patterns and urban development trends. Staten Island, for example, exhibited a higher average land area compared to other boroughs, potentially indicating more spacious properties or suburban-like neighborhoods. Furthermore, analysis of extracurricular activity dates (EXTRACRDT) revealed patterns in property transactions over time, with certain dates showing higher transaction volumes compared to others. This temporal analysis can be instrumental in understanding seasonal variations and market trends within the real estate sector.

Overall, the cleaning and profiling process not only ensured data quality and integrity but also facilitated deeper insights into the underlying characteristics and dynamics of the dataset, laying the groundwork for subsequent analyses and decision-making processes.

Data Cleaning and Profiling Output Screenshots

1. Snippet of Source Input Data

[illegible]


2. Cleaned Data Snippet

```

cp /$HOME/nyu_edu@nyu-dataproo-m:~/bdad_projects/hadoop fs -head bdad_project/cleaned/project_valuation_data.csv/part-00000-d66b95f0-ec47-4768-b651-269a11751cd8-e000.csv
BOROUGH, BLOCK, BORO, LOT, YEAR, CURMKTLT, CURMKMTO, CURACTLAND, CURACTTOT, ZIP_CODE, LAND AREA, NUM_BLDGS, EXTRACTED, BLD_STORY, UNITS, GROSS_SQFT
MANHATTAN, 851, 1, 63, 2021, 2500000, 7889000, 1125000, 3550050, 10010, 4605, 1, 01/10/2020, 6, 17, 25170
MANHATTAN, 855, 1, 1053, 2021, 29302, 262018, 13186, 11709, 10010, 9877, 1, 01/10/2020, 0, 1, 705
MANHATTAN, 859, 1, 1311, 2021, 12202, 639433, 5491, 287745, 10016, 8451, 1, 01/10/2020, 34, 1, 1303
MANHATTAN, 861, 1, 69, 2021, 1000000, 3649000, 450000, 1642050, 10016, 2158, 1, 01/10/2020, 8, 10, 18418
MANHATTAN, 863, 1, 1154, 2021, 75735, 475309, 34081, 213889, 10016, 10752, 1, 01/10/2020, 34, 1, 891
MANHATTAN, 865, 1, 69, 2021, 1140000, 4951000, 513000, 2227950, 10016, 2469, 1, 01/10/2020, 5, 8, 9267
MANHATTAN, 868, 1, 1030, 2021, 1107, 121276, 689, 36074, 10016, 9890, 1, 01/10/2020, 0, 1, 522
MANHATTAN, 869, 1, 1323, 2021, 22957, 262225, 10331, 118002, 10016, 17280, 1, 01/10/2020, 0, 1, 777
MANHATTAN, 870, 1, 1013, 2021, 53853, 355002, 24234, 160156, 10003, 76503, 1, 01/10/2020, 0, 1, 714
MANHATTAN, 870, 1, 1238, 2021, 42442, 280522, 19099, 126235, 10003, 76503, 1, 01/10/2020, 0, 1, 670
MANHATTAN, 870, 1, 1328, 2021, 3ap7986 nyu edu@nyu-dataproo-m:~/bdad_projects


```


3. Data Cleaning Summary


 SSH-in-browser

UPLOAD FILE

DOWNLOAD FILE







```
Number of rows in raw data: 6883138
24/04/20 02:46:21 WARN org.apache.spark.sql.catalyst.util.package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
Number of rows after deduplication: 6883138
Dropped columns List(HOUSENUM_LO, HOUSENUM_HI, STREET_NAME, PYMKTLAND, PYMKTLOT, PYACLAND, PYACTTOT, PYACTEXTOT, PYTRNLAND, PYTRNTOT, PYTRNXTOT, PYTXBTOT, PYTXBEXTOT, FIBATLAND, FIBACTTOT, FIBATNLAND, FIBACTTOT, FIBACTEXTOT, FIBTRNLAND, FIBTRNTOT, FIBTRNXTOT, FIBTXBTOT, FIBTXBEXTOT, CURACTEXTOT, CURTRNLAND, CURTRNTOT, CURTRNXTOT, CURTXBTOT, CURTXBEXTOT, FARID, YRBUHID)
Number of rows after filtering rows having invalid value in critical columns: 5419134
Number of rows after filtering block numbers: 5419134
Number of rows after deleting rows with invalid zip code: 5416467
```

4. Portal to Manually Querying Zip Code from BBL information

Find a Property

Select

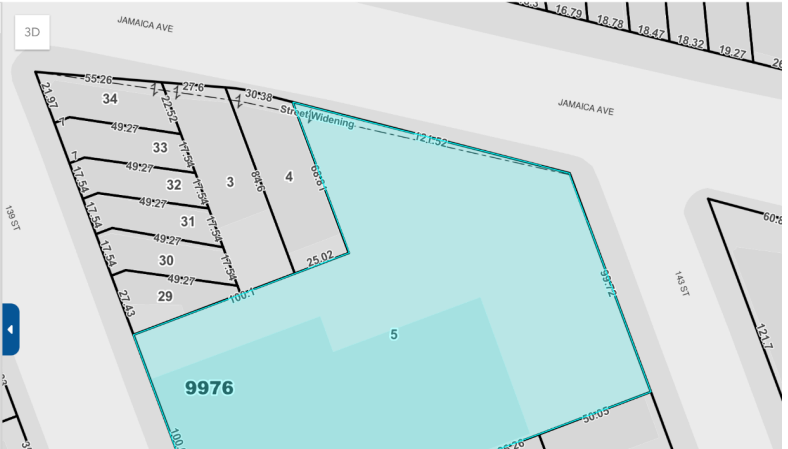
Borough / Block / Lot

Borough
Manhattan

Block

Lot

Search



5. Overall and Borough-wise Profiling of Specific Dataset Columns


```
Distinct values for column: BOROUGH
+-----+
| BOROUGH| count|
+-----+
| QUEENS|2012737|
| BROOKLYN|1710026|
| BRONX| 507347|
| MANHATTAN| 448397|
| STATEN ISLAND| 737960|
+-----+

Total distinct values for BOROUGH: 5

Distinct values for column: BLOCK
+-----+
| BLOCK|count|
+-----+
| 2136| 408|
| 3210| 678|
| 1090| 888|
| 1512| 276|
| 1436| 1782|
| 5645| 1308|
| 1572| 210|
| 691| 1108|
| 829| 1146|
| 8304| 580|
| 1159| 971|
| 2294| 282|
| 2088| 246|
| 2162| 444|
| 8433| 204|
| 467| 570|
| 675| 398|
| 6194| 450|
| 9583| 450|
| 9993| 62|
+-----+
only showing top 20 rows

Total distinct values for BLOCK: 13697
```

```
Distinct values for column: BORO
+-----+
| BORO| count|
+-----+
| 3|1710026|
| 5| 737960|
| 1| 448397|
| 4|2012737|
| 2| 507347|
+-----+

Total distinct values for BORO: 5

Distinct values for column: LOT
+-----+
| LOT|count|
+-----+
| 1159| 740|
| 1090| 897|
| 467| 144|
| 296| 426|
| 1436| 221|
| 1572| 108|
| 1512| 297|
| 691| 18|
| 2162| 30|
| 4032| 18|
| 2069| 42|
| 2136| 42|
| 2088| 54|
| 3210| 29|
| 829| 30|
| 2294| 30|
| 125| 6226|
| 1372| 215|
| 1394| 183|
| 451| 168|
+-----+
only showing top 20 rows

Total distinct values for LOT: 4474
```

```
Distinct values for column: YEAR
+-----+
| YEAR| count|
+-----+
| 2022|1804973|
| 2023|1820274|
| 2021|1791220|
+-----+

Total distinct values for YEAR: 3

Overall statistics for column: CURMKTLAND
+-----+
| summary| CURMKTLAND|
+-----+
| count| 5416467|
| mean| 356474.1071312721|
| stddev|2633926.2027289453|
| min| 5|
| max| 738553000|
+-----+

Borough-wise statistics for column: CURMKTLAND
+-----+
| BOROUGH| Average|Minimum|Maximum|Standard Deviation|
+-----+
| QUEENS| 266628.2496426508| 10000| 99991589987.3089832277| |
| BROOKLYN| 281220.5719538767| 10000| 9997| 914027.7190985366|
| BRONX|230252.1922238147| 1000| 999630|1588113.2878142374|
| MANHATTAN|1427012.8508999837| 10000| 999997| 7825712.762543917|
| STATEN ISLAND| 212202.7810152312| 1000| 999000|1542897.6361203867|
+-----+
```

```
Overall statistics for column: CURMKTTOT
+-----+-----+
|summary|CURMKTTOT|
+-----+-----+
|count|5416467|
|mean|1543146.6305973986|
|stddev|1.0270987359004166E7|
|min|95|
|max|1858141000|
+-----+-----+

Borough-wise statistics for column: CURMKTTOT
+-----+-----+-----+-----+-----+-----+
|BOROUGH|Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+-----+-----+
|QUEENS|1007632.4468388071|1000|99994|3809765.566747111|
|BROOKLYN|1354898.367392694|10000|99993|3430942.2212283765|
|BRONX|1180861.076482171|100000|9993000|6818994.929138571|
|MANHATTAN|6437344.128774278|100000|999978|3.2749500455077928E7|
|STATEN ISLAND|715215.1026613909|1000|9996000|2754147.361327146|
+-----+-----+-----+-----+-----+-----+

Overall statistics for column: CURACTLAND
+-----+-----+
|summary|CURACTLAND|
+-----+-----+
|count|5416467|
|mean|85014.41880583782|
|stddev|1179954.3837994759|
|min|2|
|max|332348850|
+-----+-----+

Borough-wise statistics for column: CURACTLAND
+-----+-----+-----+-----+-----+-----+
|BOROUGH|Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+-----+-----+
|QUEENS|41542.214773713604|100|999900|709616.0827851704|
|BROOKLYN|43796.01888860169|10|9999|408170.50227823487|
|BRONX|53381.37018450883|1001|999900|713765.1144476952|
|MANHATTAN|564271.027058611|100|99994|3507307.7295759004|
|STATEN ISLAND|29637.91603067917|10020|9995|694685.1008848724|
+-----+-----+-----+-----+-----+-----+
```

```
Overall statistics for column: CURACTTOT
+-----+-----+
|summary|CURACTTOT|
+-----+-----+
|count|5416467|
|mean|405523.5419396075|
|stddev|4622001.5495873485|
|min|42|
|max|836163450|
+-----+-----+

Borough-wise statistics for column: CURACTTOT
+-----+-----+-----+-----+-----+-----+
|BOROUGH|Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+-----+-----+
|QUEENS|163293.50423229663|100007|999967|1705014.7304895355|
|BROOKLYN|257167.38596430697|100002|999994|1539264.6745206774|
|BRONX|326775.56693347945|10001700|999900|3079905.791492078|
|MANHATTAN|2670337.0918561006|100000000|99996|1.4732462636065971E7|
|STATEN ISLAND|87965.9384329774|10001250|99960|1243123.5451824465|
+-----+-----+-----+-----+-----+-----+
```

Distinct values for column: ZIP_CODE

```
+-----+-----+
|ZIP_CODE|count|
+-----+-----+
|11205|21958|
|11236|89004|
|10309|57823|
|11106|21831|
|11351|194|
|10110|6|
|11218|42157|
|10452|7828|
|11428|25608|
|11237|25863|
|11379|53738|
|10169|6|
|10177|6|
|11021|3|
|10803|6|
|11364|38908|
|11109|72|
|11249|17600|
|11001|7182|
|10012|10893|
+-----+-----+
```

only showing top 20 rows

Total distinct values for ZIP_CODE: 232

Overall statistics for column: LAND_AREA

```
+-----+-----+
|summary|LAND_AREA|
+-----+-----+
|count|5416467|
|mean|10634.102961949182|
|stddev|590028.0293649806|
|min|1|
|max|503315650|
+-----+-----+
```

```
Borough-wise statistics for column: LAND_AREA
+-----+-----+-----+-----+-----+
| BOROUGH| Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+
| QUEENS| 9124.640538729103| 1| 9999| 67397.23470527849|
| BROOKLYN| 5363.306877205376| 1| 9999| 85060.61892020836|
| BRONX| 9866.445716639697| 1000| 9996| 311634.07833677187|
| MANHATTAN| 25592.459590496816| 1| 99960| 1998833.4771042822|
| STATEN ISLAND| 18403.542468426473| 100| 9999| 177211.67407722157|
+-----+-----+-----+-----+

Overall statistics for column: NUM_BLDGS
+-----+-----+
| summary| NUM_BLDGS|
+-----+-----+
| count| 5416467|
| mean| 1.0270777981292971|
| stddev| 2.7565545515088017|
| min| 0|
| max| 2244|
+-----+-----+

Borough-wise statistics for column: NUM_BLDGS
+-----+-----+-----+-----+-----+
| BOROUGH| Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+
| QUEENS| 1.042993694655586| 0| 99| 4.448942556832905|
| BROOKLYN| 1.018621939081628| 0| 910.4174460405246258|
| BRONX| 1.0400160048251| 0| 911.1590504678038438|
| MANHATTAN| 1.0244203239539962| 0| 910.614455851893838|
| STATEN ISLAND| 0.995982167055125| 0| 910.4785061369095627|
+-----+-----+-----+-----+-----+
```

```
Borough-wise statistics for column: LAND_AREA
+-----+-----+-----+-----+-----+
| BOROUGH| Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+
| QUEENS| 9124.640538729103| 1| 9999| 67397.23470527849|
| BROOKLYN| 5363.306877205376| 1| 9999| 85060.61892020836|
| BRONX| 9866.445716639697| 1000| 9996| 311634.07833677187|
| MANHATTAN| 25592.459590496816| 1| 99960| 1998833.4771042822|
| STATEN ISLAND| 18403.542468426473| 100| 9999| 177211.67407722157|
+-----+-----+-----+-----+

Overall statistics for column: NUM_BLDGS
+-----+-----+
| summary| NUM_BLDGS|
+-----+-----+
| count| 5416467|
| mean| 1.0270777981292971|
| stddev| 2.7565545515088017|
| min| 0|
| max| 2244|
+-----+-----+

Borough-wise statistics for column: NUM_BLDGS
+-----+-----+-----+-----+-----+
| BOROUGH| Average|Minimum|Maximum|Standard Deviation|
+-----+-----+-----+-----+
| QUEENS| 1.042993694655586| 0| 99| 4.448942556832905|
| BROOKLYN| 1.018621939081628| 0| 910.4174460405246258|
| BRONX| 1.0400160048251| 0| 911.1590504678038438|
| MANHATTAN| 1.0244203239539962| 0| 910.614455851893838|
| STATEN ISLAND| 0.995982167055125| 0| 910.4785061369095627|
+-----+-----+-----+-----+-----+
```

```
Distinct values for column: EXTRACRDT
+-----+
| EXTRACRDT| count|
+-----+
|05/17/2020|896942|
|01/10/2020|894278|
|05/17/2022|911177|
|01/10/2022|909097|
|01/11/2021|900961|
|05/24/2021|904012|
+-----+

Total distinct values for EXTRACRDT: 6

Distinct values for column: BLD_STORY
+-----+
|BLD_STORY| count|
+-----+
|      8.5|    12|
|       7| 48507|
|      51|   410|
|    4.33|    59|
|      15|11817|
|      54|    70|
|      11|13535|
|    20.5|    18|
|       2.6|    10|
|     209|     6|
|     470|     6|
|      29|   627|
|      69|     8|
|     42| 2590|
|     0.3|    19|
|      73|   513|
|     1.85|    61|
|      31770178|
|      30|  6175|
|    22.5|     6|
+-----+

only showing top 20 rows

Total distinct values for BLD_STORY: 202
```

```
Overall statistics for column: UNITS
+-----+
|summary|          UNITS|
+-----+
| count|          5416467|
|  mean|4.163267679836322|
| stddev|34.03375323765625|
|   min|              1.0|
|   max|          10948.0|
+-----+

Borough-wise statistics for column: UNITS
+-----+
|BOROUGH|          Average|Minimum|Maximum|Standard Deviation|
+-----+
|  QUEENS|2.755946256266964|      1|    99|20.002959792837924|
|BROOKLYN|3.873618295862168|      1|   997|26.560940849895893|
|  BRONX|6.730157071984263|      1|    99| 62.39211104694712|
|MANHATTAN|12.967118424075094|      1|   995| 67.48719548952036|
|STATEN ISLAND|1.558725405171012|      1|    98| 16.36765296482242|
+-----+

Overall statistics for column: GROSS_SQFT
+-----+
|summary|          GROSS_SQFT|
+-----+
| count|          5416467|
|  mean|5764.13323961911|
| stddev|47277.36246342804|
|   min|              1|
|   max|          22051813|
+-----+

Borough-wise statistics for column: GROSS_SQFT
+-----+
|BOROUGH|          Average|Minimum|Maximum|Standard Deviation|
+-----+
|  QUEENS|3496.943922131903|      1|  9998|25552.964063692114|
|BROOKLYN|4866.398454175551|      1|  9998| 27197.06482274286|
|  BRONX|8513.996349638413|    100| 99985|64619.386266856374|
|MANHATTAN|21886.401900547953|      1| 9999|126394.29031050214|
|STATEN ISLAND|2341.310795978102|      1| 9995|11922.117828699937|
+-----+
```