Analiza przestępczości w Nowym Jorku

Adam Frej

Jan Gąska

Dane kontekstowe – Miasto Nowy York

- Założony w 1624 r.
- Powierzchnia 1213.3 km^2
- Populacja dla NYC 8 336 817 (2019)
- Dla stanu Nowy York 19 840 421 (2018)
- Gęstość zaludnienia – 10 636 os./km^2 (dla porównania Warszawa 3 372)
- Jeden z najbardziej zróżnicowanych stanów w USA
- Centrum handlu międzynarodowego
- Najbogatszy stan pod względem PKB per capita ($83,388)
- Trzeci najbogatszy stan pod względem całkowitego PKB ($1,705,127,000)

Szybki słownik

- Kondominium – pojedyncza jednostka nieruchomości w zabudowie wielomieszkaniowej, w której osoba ma zarówno odrębną własność jednostki, jak i niepodzielny udział we wspólnych elementach budynku.

NYC Open Data

- Ogólnodostępne dane dla Miasta Nowy York
- Dane aktualizowane wraz z postanowieniem administratora
- Dane udostępnione do publicznego użytku i komercjalizacji
- Zbierane przez administracyjny organ miasta Nowy York
- Obydwie nasze ramki danych pochodzą z NYC OpenData

Dane:
NYPD Complaint Data Historic

7.83M wierszy, 35 kolumn, odświeżane codziennie

CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	ADDR_PCT_CD	RPT_DT	KY_CD	OFNS_DESC	PD_CD
506547392	03/29/2018	20:30:00	NaN	NaN	32	03/30/2018	351	CRIMINAL MISCHIEF & RELATED OF	254
629632833	02/06/2018	23:15:00	NaN	NaN	52	02/07/2018	341	PETIT LARCENY	333
787203902	11/21/2018	00:15:00	11/21/2018	00:20:00	75	11/21/2018	341	PETIT LARCENY	321
280364018	06/09/2018	21:42:00	06/09/2018	21:43:00	10	06/10/2018	361	OFF. AGNST PUB ORD SENSBLTY &	639
985800320	11/10/2018	19:40:00	11/10/2018	19:45:00	19	11/10/2018	341	PETIT LARCENY	333

SUSP_AGE_GROUP	SUSP_RACE	SUSP_SEX	TRANSIT_DISTRICT	Latitude	Longitude	Lat_Lon	PATROL_BORO
NaN	NaN	NaN	NaN	40.81087724100007	-73.94106415099996	(40.810877241, -73.941064151)	PATROL BORO MAN NORTH
45-64	BLACK	F	NaN	40.87367103500002	-73.90801364899994	(40.873671035, -73.908013649)	PATROL BORO BRONX
25-44	WHITE HISPANIC	F	NaN	40.651782232000066	-73.88545676099994	(40.651782232, -73.885456761)	PATROL BORO BKLYN NORTH
25-44	WHITE HISPANIC	M	NaN	40.75931039900007	-73.99470607199999	(40.759310399, -73.994706072)	PATROL BORO MAN SOUTH
<18	BLACK HISPANIC	F	NaN	40.76453553900007	-73.97072838799994	(40.764535539, -73.970728388)	PATROL BORO MAN NORTH

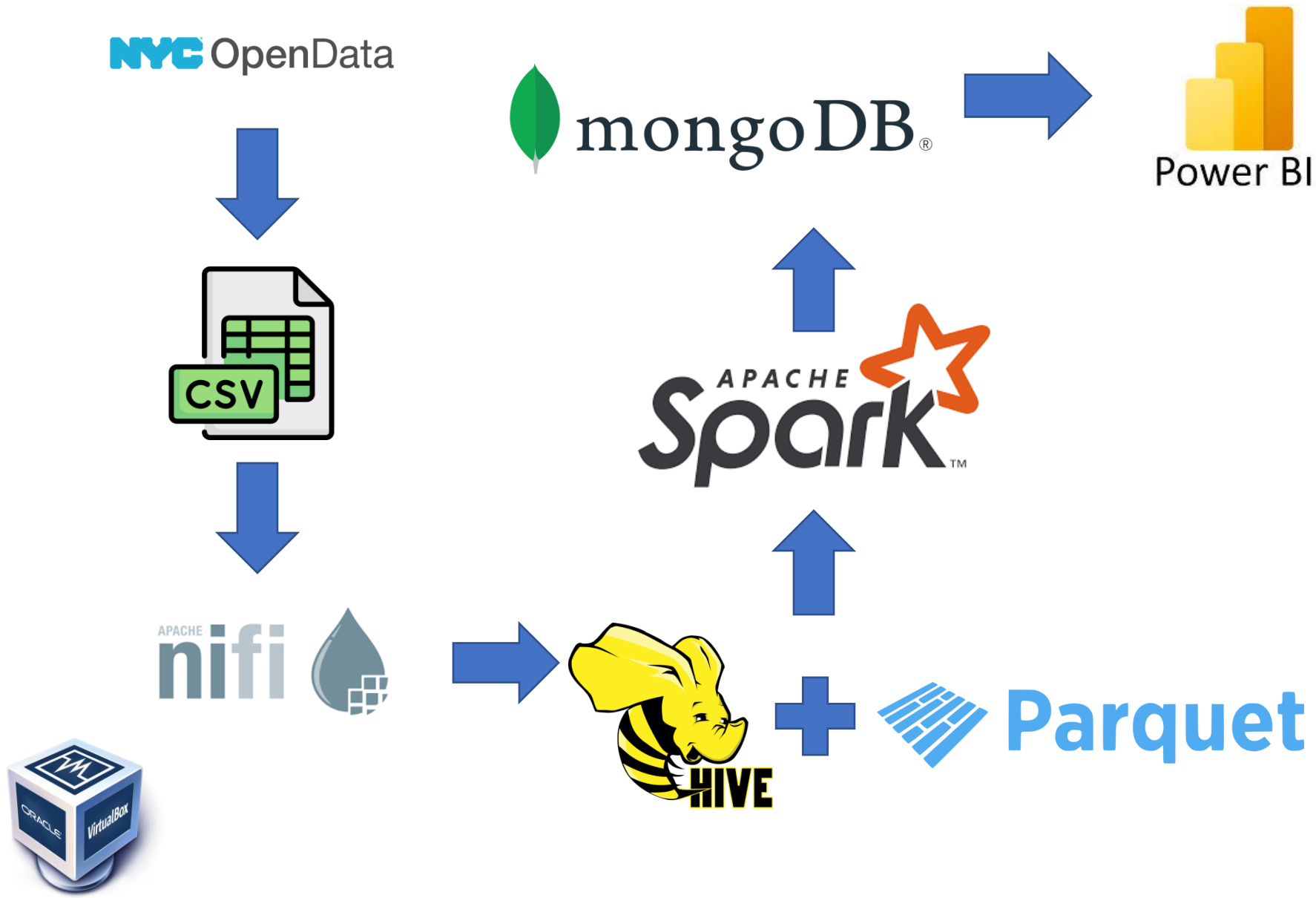
Dane:
Condominium Rental Income in NYC

28.5K wierszy, 61 kolumn, odświeżane rocznie

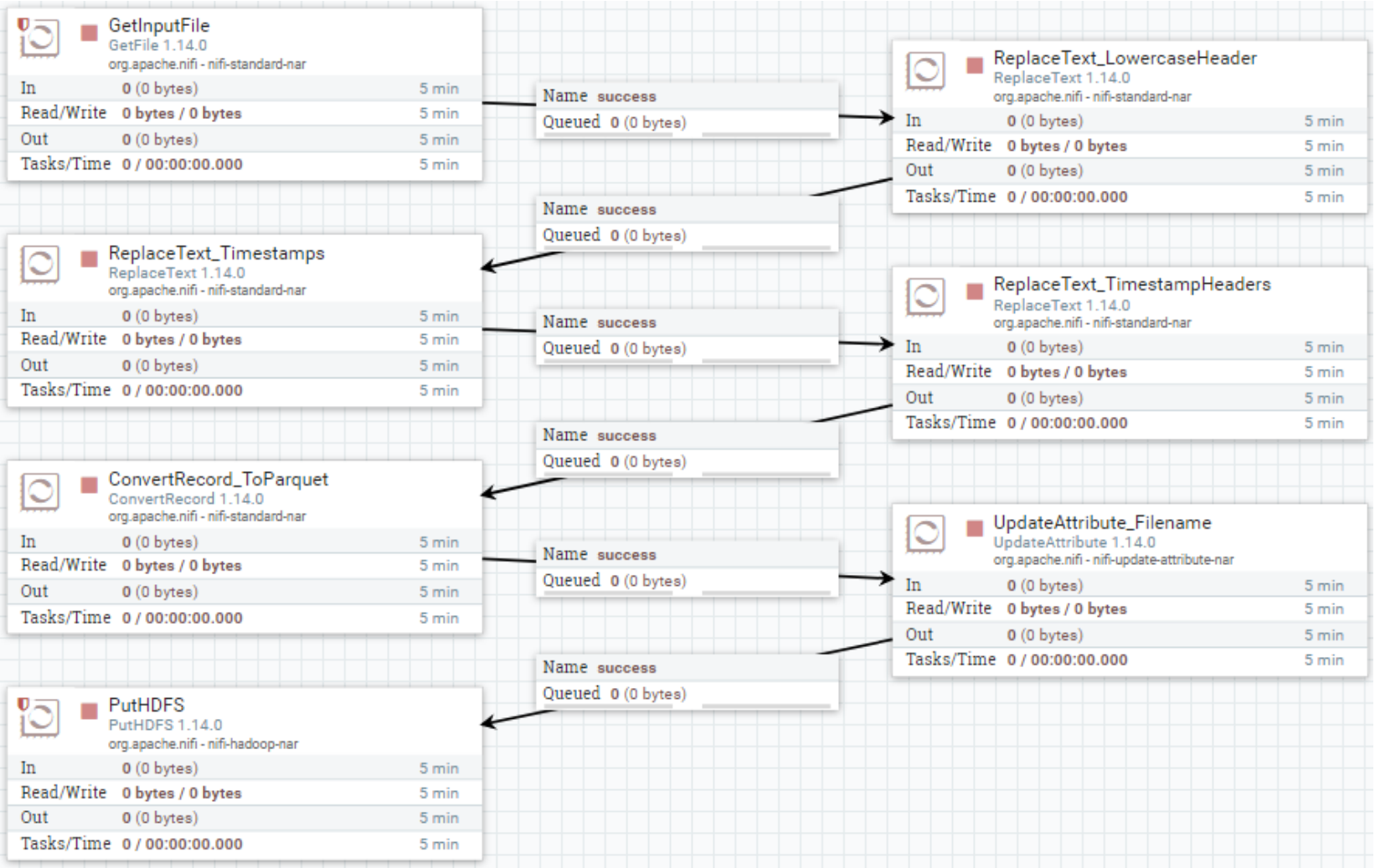
Boro-Block-Lot	Condo Section	Address	Neighborhood	Building Classification	Total Units	Year Built	Gross SqFt	Estimated Gross Income	Gross Income per SqFt	Estimated Expense	Expense per SqFt	Net Operating Income	Full Market Value	Market Value per SqFt
1-00576-7501	0003-R1	60 WEST 13 STREET	GREENWICH VILLAGE-CENTRAL	R4 - ELEVATOR	70	1966	82017	4452703	54.29	1729739	21.09	2722964	22115002	269.64
1-01271-7501	0007-R2	1360 6 AVENUE	MIDTOWN WEST	R4 - ELEVATOR	183	1963	141738	7113830	50.19	2361355	16.66	4752475	38596999	272.31
1-00894-7501	0009-R1	77 PARK AVENUE	MURRAY HILL	R4 - ELEVATOR	109	1924	158571	7329152	46.22	2854278	18	4474874	36343010	229.19
1-00631-7501	0018-R1	712 GREENWICH STREET	GREENWICH VILLAGE-WEST	R9 - CONDOPS	20	1910	53943	2132906	39.54	666196	12.35	1466710	11912000	220.83
1-00868-7501	0019-R1	35 EAST 38 STREET	MURRAY HILL	R4 - ELEVATOR	113	1961	88230	4288860	48.61	1055231	11.96	3233629	26261996	297.65

Boro-Block-Lot 1	Address 1	Neighborhood 1	Building Classification 1	Total Units 1	Year Built 1	Gross SqFt 1	Estimated Gross Income 1	Gross Income per SqFt 1	Estimated Expense 1	Expense per SqFt 1	Net Operating Income 1	Full Market Value 1	Market Value per SqFt 1	Distance from Condo in miles
1-00573-0011	60 WEST 10 STREET	GREENWICH VILLAGE-CENTRAL	D1 - ELEVATOR	27	1910	20797	1130317	54.35	453375	21.8	676942	5205000	250.28	0.14
1-01043-0005	369 WEST 52 STREET	MIDTOWN WEST	D5 - ELEVATOR	48	1940	37030	1858536	50.19	616920	16.66	1241616	11335000	306.1	0.54
1-00865-0032	20 PARK AVENUE	MURRAY HILL	D7 - ELEVATOR	102	1939	101306	4742134	46.81	2057525	20.31	2684609	26269000	259.3	0.18
1-00631-0030	697 GREENWICH STREET	GREENWICH VILLAGE-WEST	D5 - ELEVATOR	53	1979	51200	2086912	40.76	833024	16.27	1253888	14188000	277.11	0.06
1-00915-	210 EAST 35 STREET	MURRAY HILL	D1 - ELEVATOR	11	1920	9342	452807	48.47	100800	10.79	352007	2657000	284.41	0.37

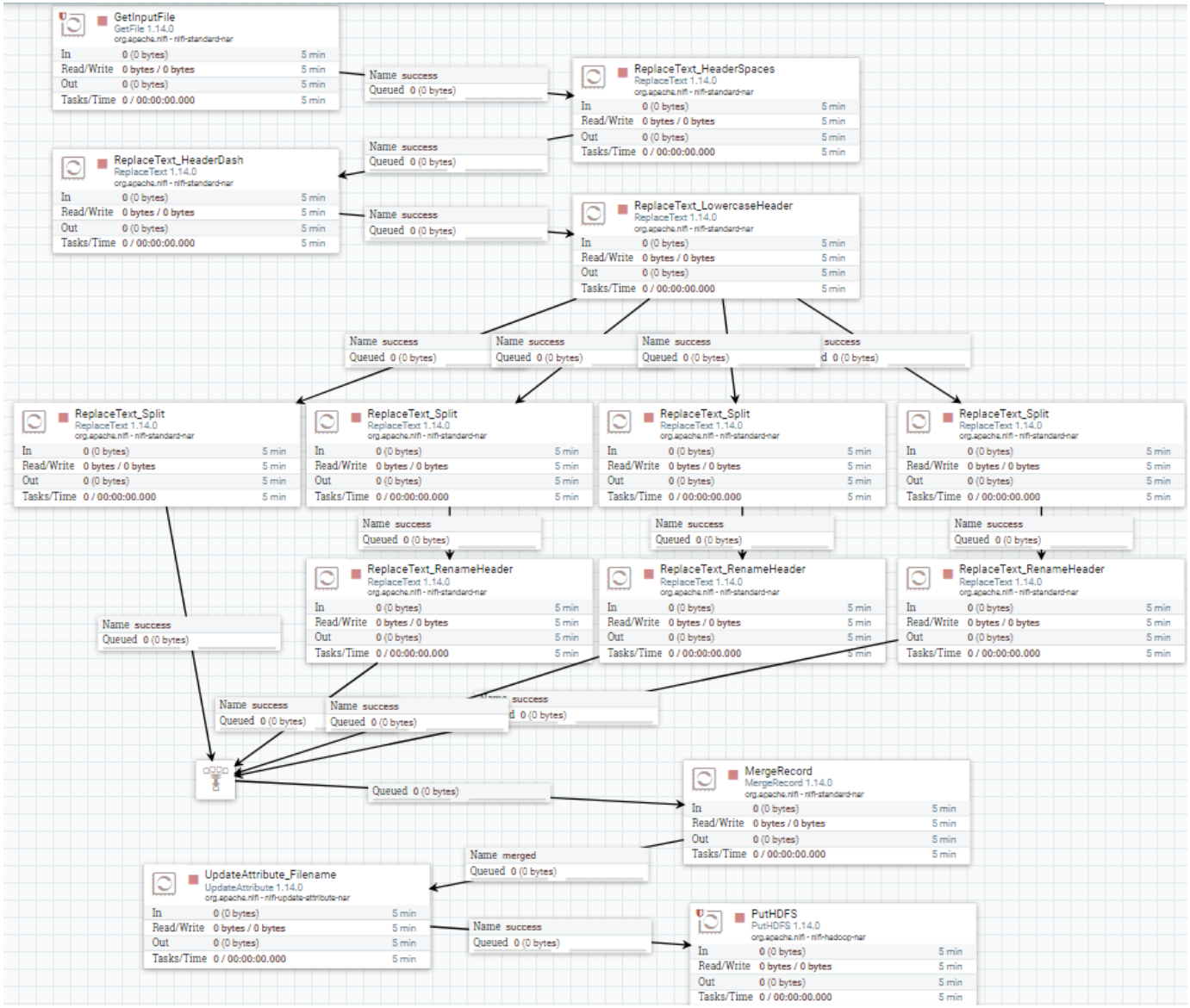
Przyjęte rozwiązanie architektoniczne



Nifi - NYPD

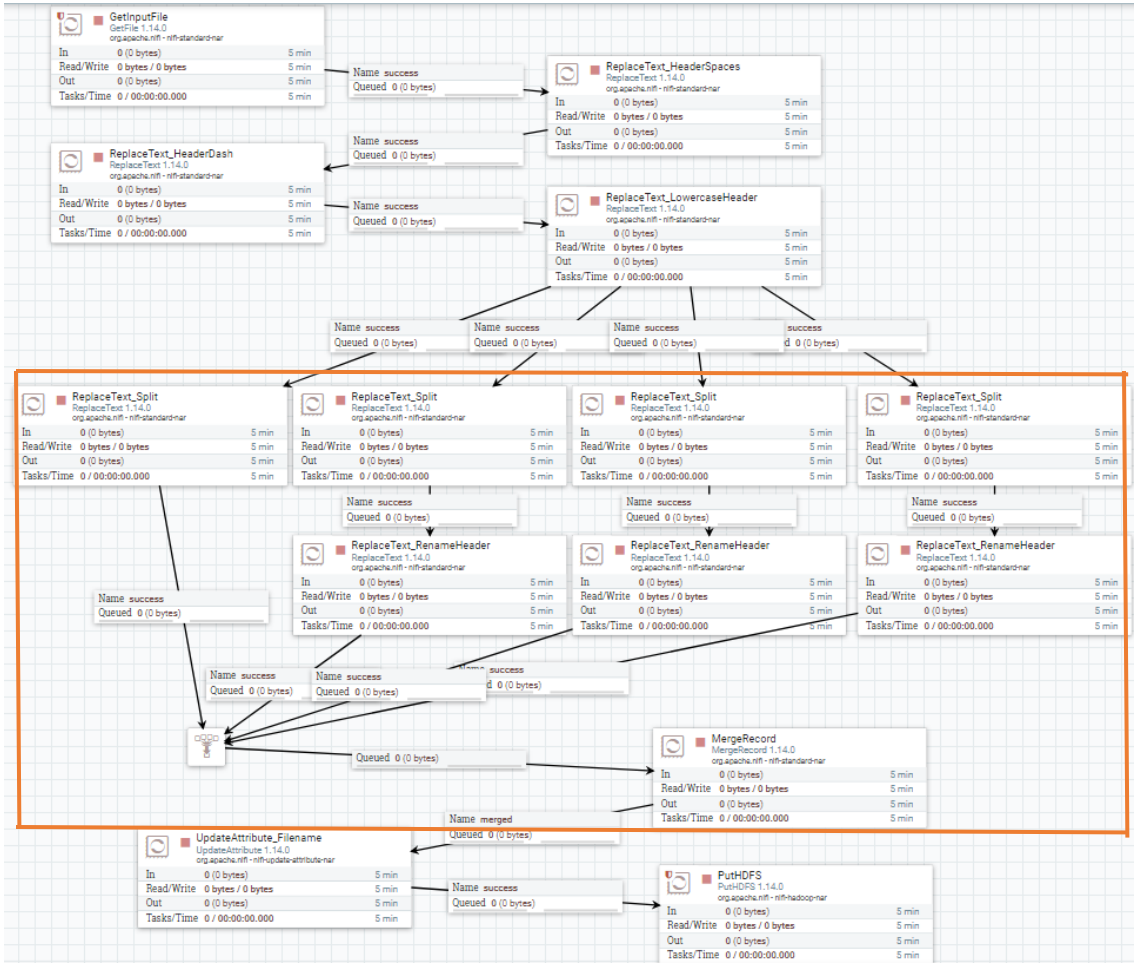

GetInputFile
GetFile 1.14.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
ReplaceText_LowercaseHeader
ReplaceText 1.14.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
ReplaceText_Timestamps
ReplaceText 1.14.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
ReplaceText_TimestampHeaders
ReplaceText 1.14.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
ConvertRecord_ToParquet
ConvertRecord 1.14.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
UpdateAttribute_Filename
UpdateAttribute 1.14.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
PutHDFS
PutHDFS 1.14.0
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Nifi - Condominiums



Problem with condominium data

- Konieczność poważniejszej modyfikacji danych
- Dane nie były zatomizowane
- W preprocessing'u w nifi należało podzielić ramkę kolumnowo a następnie ją złączyć



Hive

- External tables
- Parquet files

```
CREATE EXTERNAL TABLE external_table_condo
(condo_section string,
boro_block_lot string,
address string,
neighborhood string,
building_classification string,
total_units int,
year_built int,
gross_sqft int,
estimated_gross_income int,
gross_income_per_sqft double,
estimated_expense int,
expense_per_sqft double,
net_operating_income int,
full_market_value int,
market_value_per_sqft double,
report_year int)
STORED AS PARQUET
LOCATION '/projekt/external_table_condo';
```

```
CREATE EXTERNAL TABLE external_table_nypd
(cmplnt_num int,
cmplnt_fr_ts timestamp,
cmplnt_to_ts timestamp,
addr_pct_cd int,
rpt_dt date,
ky_cd int,
ofns_desc string,
pd_cd int,
pd_desc string,
crm_atpt_cptd_cd string,
law_cat_cd string,
boro_nm string,
loc_of_occur_desc string,
prem_typ_desc string,
juris_desc string,
jurisdiction_code int,
parks_nm string,
hadevelopt string,
housing_psa string,
x_coord_cd int,
y_coord_cd int,
susp_age_group string,
susp_race string,
susp_sex string,
transit_district double,
latitude double,
longitude double,
patrol_boro string,
station_name string,
vic_age_group string,
vic_race string,
vic_sex string)
STORED AS PARQUET
LOCATION '/projekt/external_table_nypd';
```

Spark


```
-----------+--------------------+------------+-------------+--------------------+-------+
|cmplnt_num|       cmplnt_fr_ts|       cmplnt_to_ts|addr_pct_cd|    rpt_dt|ky_cd|           ofns_desc|pd_cd|             pd_des
c|crm_atpt_cptd_cd| law_cat_cd|  boro_nm|loc_of_occur_desc|       prem_typ_desc|         juris_desc|jurisdiction_code|parks_nm|
hadevelopt|housing_psa|x_coord_cd|y_coord_cd|susp_age_group|     susp_race|susp_sex|transit_district|          latitude|
longitude|         patrol_boro|station_name|vic_age_group|            vic_race|vic_sex|
+----------+------------------+------------------+-----------+----------+-----+--------------------+-----+------------------
-+---------------+-----------+---------+-----------------+-------------------+------------------+----------------+--------+
----------+-----------+----------+----------+--------------+--------------+--------+----------------+------------------+-------
-----------+--------------------+------------+-------------+--------------------+-------+
| 506547392|2018-03-29 20:30:00|               null|         32|2018-03-30|  351|CRIMINAL MISCHIEF...|  254|MISCHIEF, CRIMIN
A...|       COMPLETED|MISDEMEANOR|MANHATTAN|         FRONT OF|PARKING LOT/GARAG...|   N.Y. POLICE DEPT|                0|    nu
ll|      null|       null|   1000565|    234704|          null|          null|    null|            null| 40.81087724100007|-73.
94106415099996|PATROL BORO MAN N...|        null|        25-44|               WHITE|      F|
| 629632833|2018-02-06 23:15:00|               null|         52|2018-02-07|  341|       PETIT LARCENY|  333|LARCENY,PETIT FR
O...|       COMPLETED|MISDEMEANOR|    BRONX|           INSIDE|    DEPARTMENT STORE|   N.Y. POLICE DEPT|                0|    nu
ll|      null|       null|   1009690|    257590|         45-64|         BLACK|       F|            null| 40.87367103500002|-73.
90801364899994|   PATROL BORO BRONX|        null|      UNKNOWN|             UNKNOWN|      D|


--+--------------------+--------------------+----------------+---------------+--------------------+----------------+-----
---------------+-----------+
|condo_section|boro_block_lot|             address|        neighborhood|building_classification|total_units|year_built|gross_sq
ft|estimated_gross_income|gross_income_per_sqft|estimated_expense|expense_per_sqft|net_operating_income|full_market_value|marke
t_value_per_sqft|report_year|
+-------------+--------------+--------------------+--------------------+-----------------------+-----------+----------+--------
--+----------------------+---------------------+-----------------+----------------+--------------------+-----------------+-----
---------------+-----------+
|      0003-R1|  1-00576-7501|   60 WEST 13 STREET|GREENWICH VILLAGE...|          R4  -ELEVATOR|         70|      1966|     820
17|               4452703|                54.29|          1729739|           21.09|             2722964|         22115002|
269.64|       2019|
|      0007-R2|  1-01271-7501|       1360 6 AVENUE|        MIDTOWN WEST|          R4  -ELEVATOR|        183|      1963|    1417
38|               7113830|                50.19|          2361355|           16.66|             4752475|         38596999|
272.31|       2019|
|      0009-R1|  1-00894-7501|      77 PARK AVENUE|         MURRAY HILL|          R4  -ELEVATOR|        109|      1924|    1585
71|               7329152|                46.22|          2854278|            18.0|             4474874|         36343010|
229.19|       2019|
|      0018-R1|  1-00631-7501|712 GREENWICH STREET|GREENWICH VILLAGE...|          R9  -CONDOPS|         20|      1910|     539
43|               2132906|                39.54|           666196|           12.35|             1466710|         11912000|
220.83|       2019|
```

Spark – utworzone agregacje

- W Jupyter Notebook
- Cztery główne kolekcje a gregacyjne dla danych (+ 4 pomocnicze)
- Jedna kolekcja dla danych condominium
- Jedna kolekcja dla danych przestępczości
- Dwie dla zespolonych danych

In [14]:
```
#agregacje dla rasy, przedziału wiekowego, płci, dystryktu gdzie zaszedł incydent oraz jego typ (skala powagi) dla ofiar
df_sub_vic = df_nypd.select(col('vic_race'),col('vic_age_group'),col('vic_sex'),col('boro_nm'),col("law_cat_cd"))
df_vic = df_sub_vic.groupBy('vic_race', 'vic_age_group', 'vic_sex', 'boro_nm',"law_cat_cd").count()
df_vic = df_vic.sort(col('count').desc())
df_vic.sort(col('count').desc()).show()
df_vic.write.format("mongodb").mode("append").save()
```

```
+-------------+-------------+-------+---------+-----------+-----+
|      vic_race|vic_age_group|vic_sex|  boro_nm| law_cat_cd|count|
+-------------+-------------+-------+---------+-----------+-----+
|       UNKNOWN|      UNKNOWN|      D|MANHATTAN|MISDEMEANOR| 3960|
|       UNKNOWN|      UNKNOWN|      D| BROOKLYN|MISDEMEANOR| 2447|
|       UNKNOWN|      UNKNOWN|      E|    BRONX|MISDEMEANOR| 2156|
|       UNKNOWN|      UNKNOWN|      E|MANHATTAN|MISDEMEANOR| 2072|
|       UNKNOWN|      UNKNOWN|      E| BROOKLYN|MISDEMEANOR| 2067|
|       UNKNOWN|      UNKNOWN|      D|   QUEENS|MISDEMEANOR| 1717|
|       UNKNOWN|      UNKNOWN|      D|    BRONX|MISDEMEANOR| 1503|
|       UNKNOWN|      UNKNOWN|      D|MANHATTAN|     FELONY| 1466|
|       UNKNOWN|      UNKNOWN|      E| BROOKLYN|     FELONY| 1430|
|         BLACK|        25-44|      F| BROOKLYN|MISDEMEANOR| 1341|
|       UNKNOWN|      UNKNOWN|      E|   QUEENS|MISDEMEANOR| 1032|
|       UNKNOWN|      UNKNOWN|      D| BROOKLYN|     FELONY|  988|
|         BLACK|        25-44|      F|    BRONX|MISDEMEANOR|  964|
|WHITE HISPANIC|        25-44|      F|    BRONX|MISDEMEANOR|  952|
|         BLACK|        25-44|      M| BROOKLYN|MISDEMEANOR|  890|
|       UNKNOWN|      UNKNOWN|      E|MANHATTAN|     FELONY|  880|
|       UNKNOWN|      UNKNOWN|      E|    BRONX|     FELONY|  840|
|         BLACK|        25-44|      F| BROOKLYN|  VIOLATION|  774|
|         BLACK|        25-44|      F| BROOKLYN|     FELONY|  760|
|         WHITE|        25-44|      M| BROOKLYN|MISDEMEANOR|  724|
+-------------+-------------+-------+---------+-----------+-----+
only showing top 20 rows
```

MongoDB

- Dane zagregowane przetrzymywane są w bazie danych MongoDB.
- Zapisywane są z pysparka poprzez inicjalizację sesji do MongoDB

In [3]:
```
try:
    spark.sparkContext.stop()
except NameError:
    print("")
spark = (SparkSession.builder.appName("projekt")
         .config("spark.jars.packages", "org.mongodb.spark:mongo-spark-connector:10.0.5")
         .config("spark.mongodb.uri", "mongodb://127.0.0.1:27017/")
         .config("spark.mongodb.database", "projekt")
         .config("spark.mongodb.collection", "df")
         .enableHiveSupport()
         .getOrCreate())
```

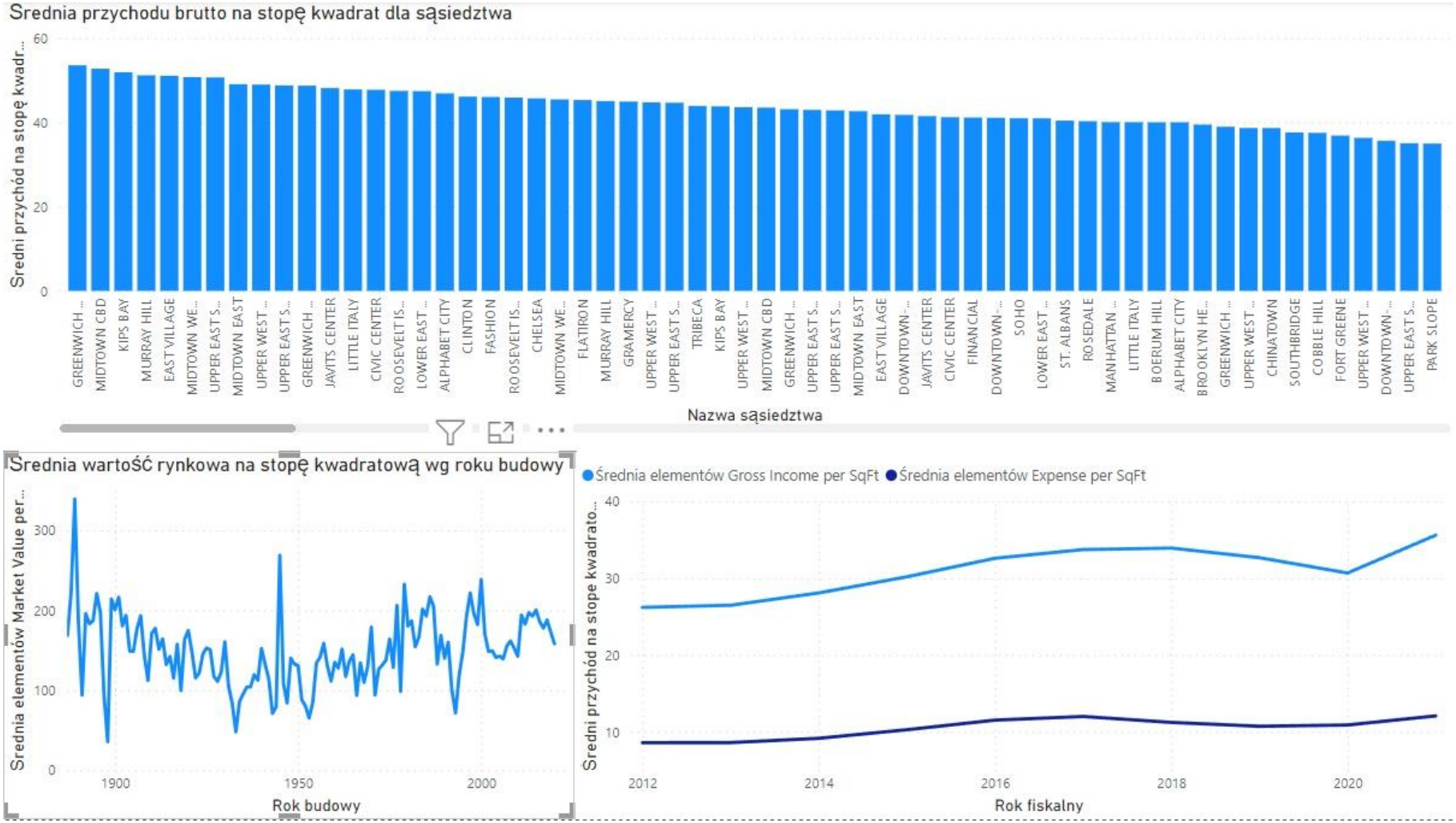
In [16]:
```
spark.read.format("mongodb").load().show(5)
```

```
+--------------------+--------------------+------------------------+------------------------+------------------------+---
--------------------+----------+----------+
|                 _id|avg(expense_per_sqft)|avg(gross_income_per_sqft)|avg(market_value_per_sqft)|avg(net_operating_income)|bui
lding_classification|     neighborhood|year_built|
+--------------------+--------------------+------------------------+------------------------+------------------------+---
--------------------+----------+----------+
|63bb1471b8309200c...|                8.79|                   35.15|                  214.77|                604356.0|
R4  -ELEVATOR|   FLUSHING-NORTH|      2020|
|63bb1471b8309200c...|               14.02|                   36.96|                  187.01|                525028.0|
R4  -ELEVATOR|    CROWN HEIGHTS|      2020|
|63bb1471b8309200c...|               12.58|                   27.95|                   99.29|               1932993.0|
RR  -CONRENT|         JAMAICA|      2020|
|63bb1471b8309200c...|               15.71|                   31.95|                  131.96|                223982.0|
R4  -ELEVATOR|        GRAVESEND|      2020|
|63bb1471b8309200c...|                5.91|                   32.82|                  219.05|                554050.0|
R4  -ELEVATOR|WILLIAMSBURG-EAST|      2019|
+--------------------+--------------------+------------------------+------------------------+------------------------+---
--------------------+----------+----------+
only showing top 5 rows
```

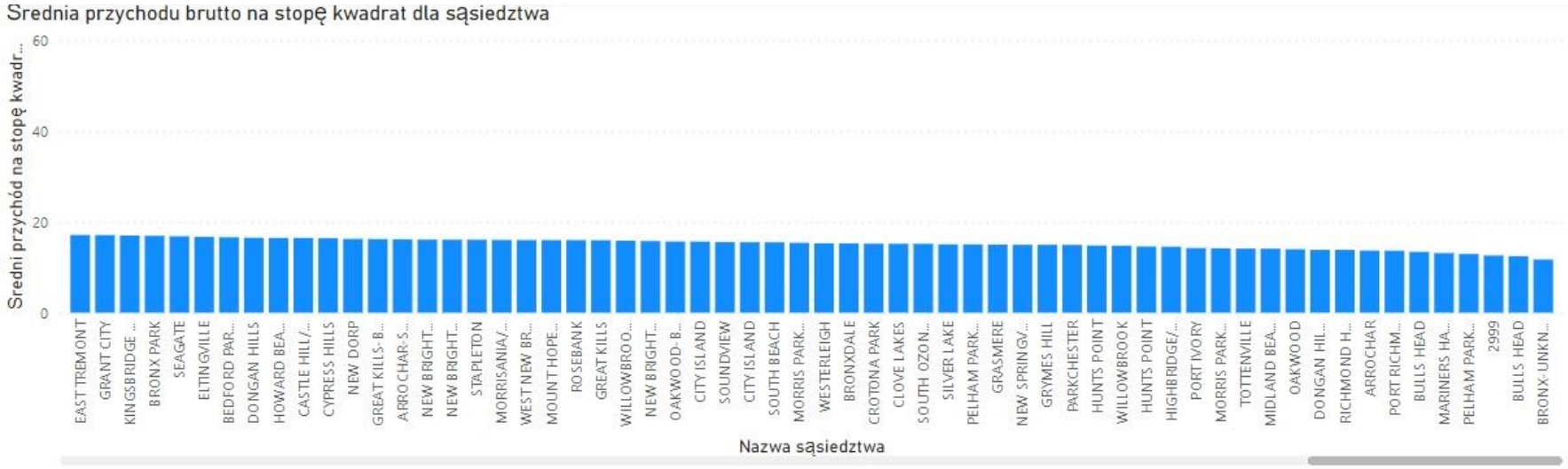
PowerBI

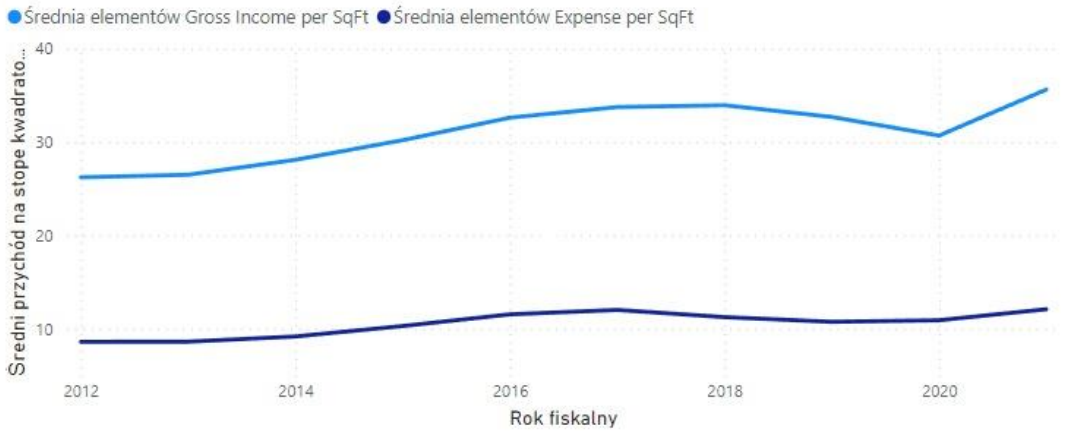
- Połączenie bazy MondoDB do PowerBI w celach wizualizacyjnych
- W PowerBI dokonujemy wyłącznie wizualizacji, nie przekształcamy już danych
- Przystępne narzędzie do wizualizacji oraz komponuje się z czasowym odświeżaniem danych pod wpływem ich napływu

Wizualizacja I

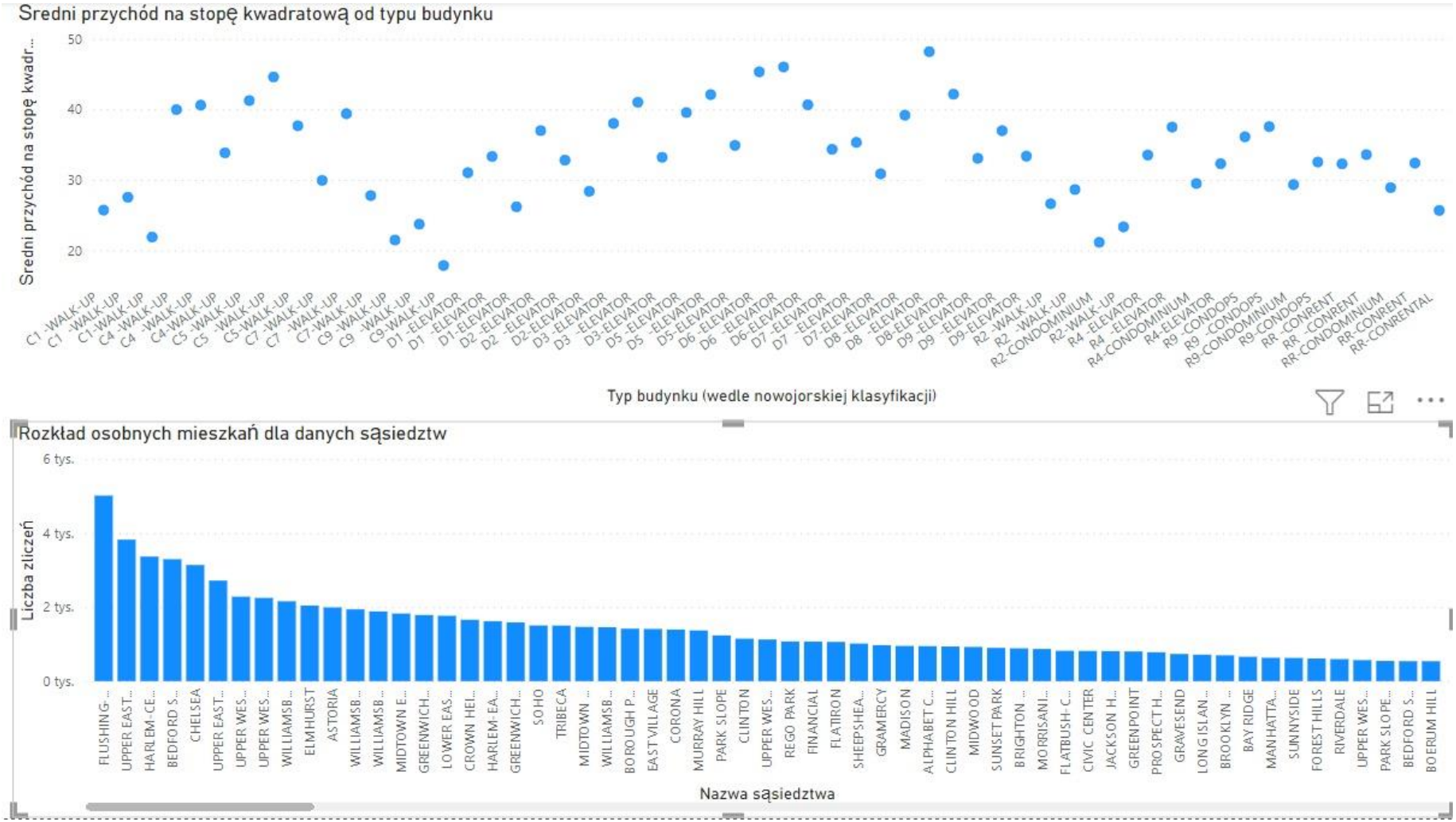


Wizualizacja I a)

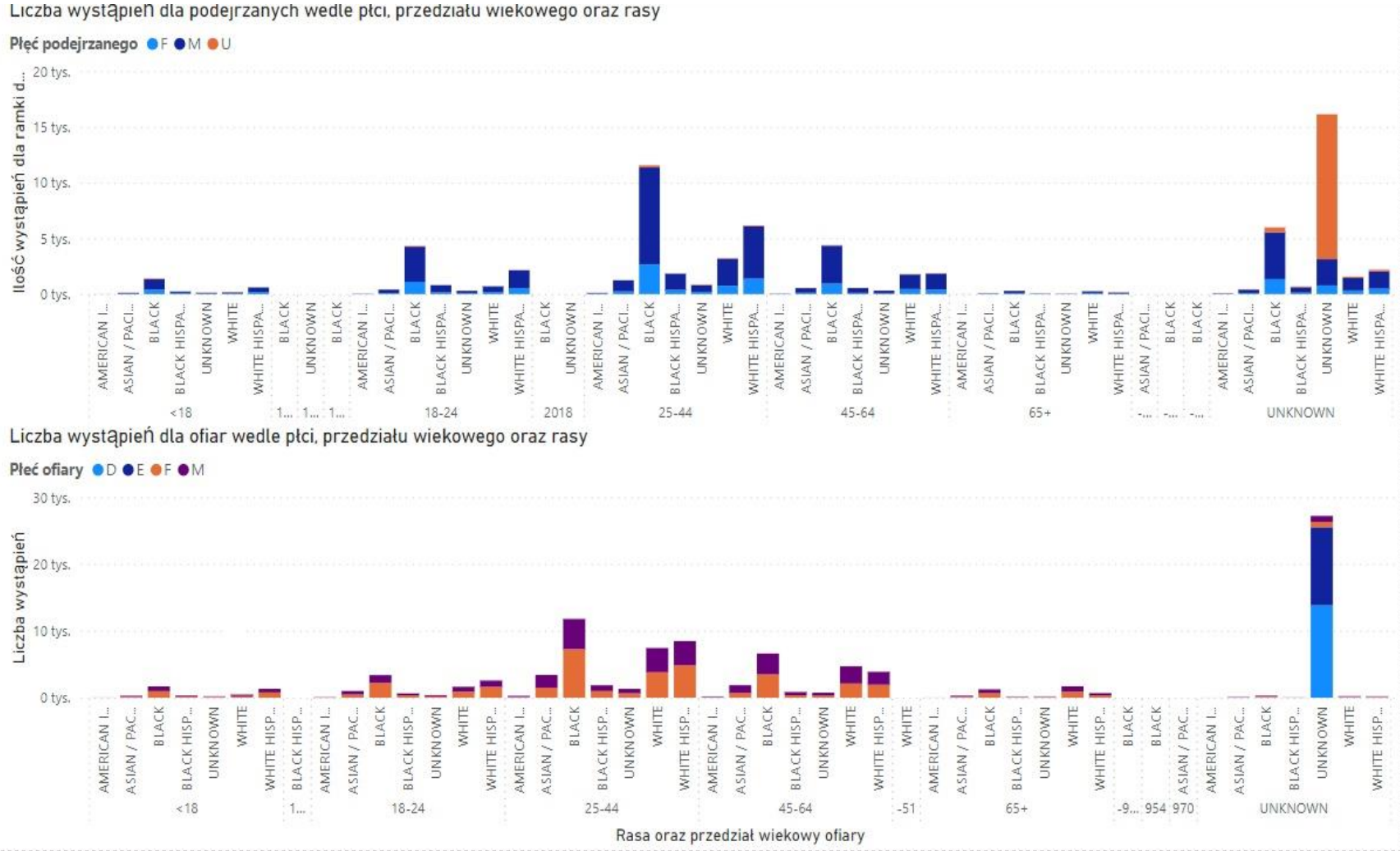




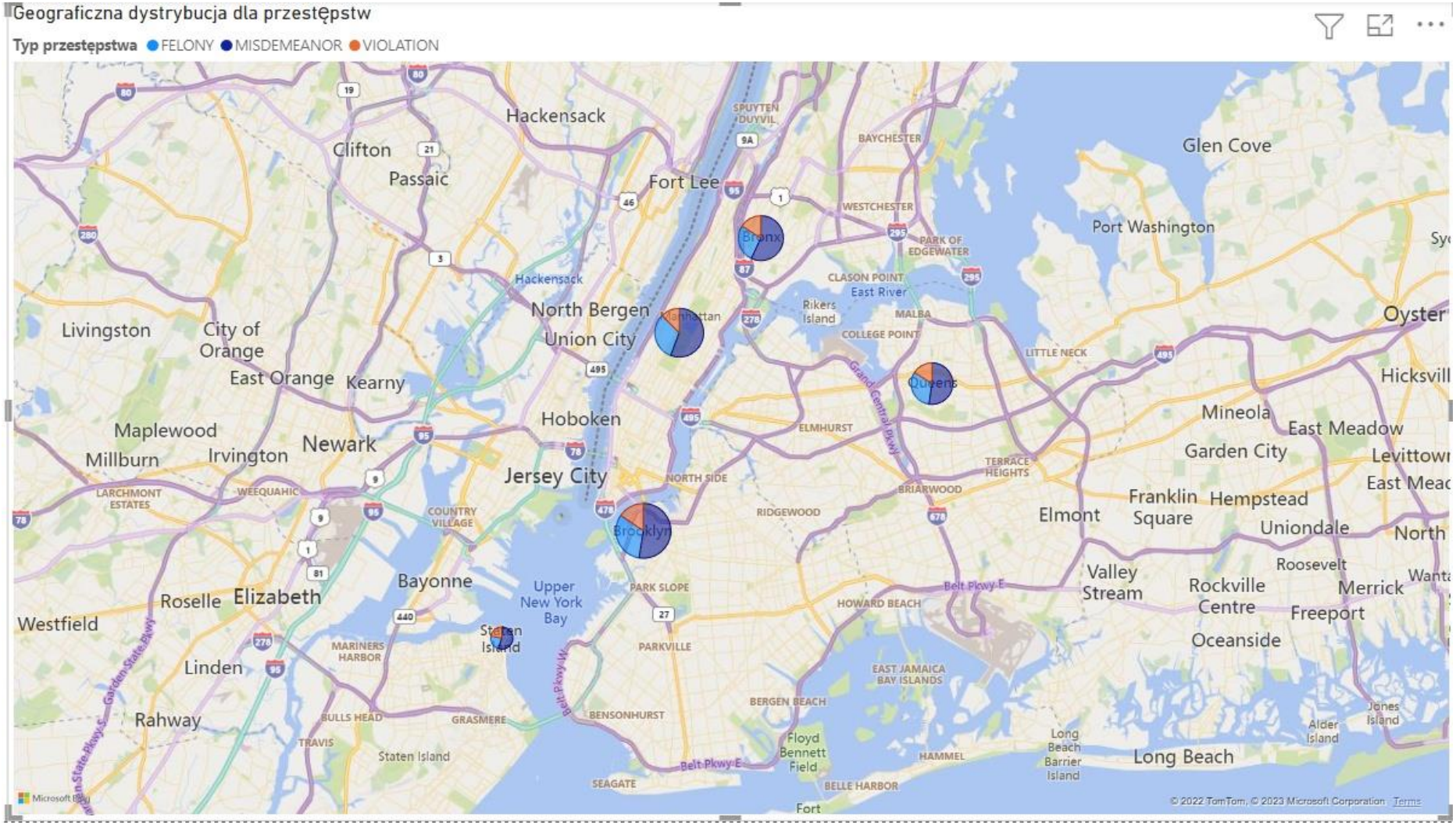
Wizualizacja II



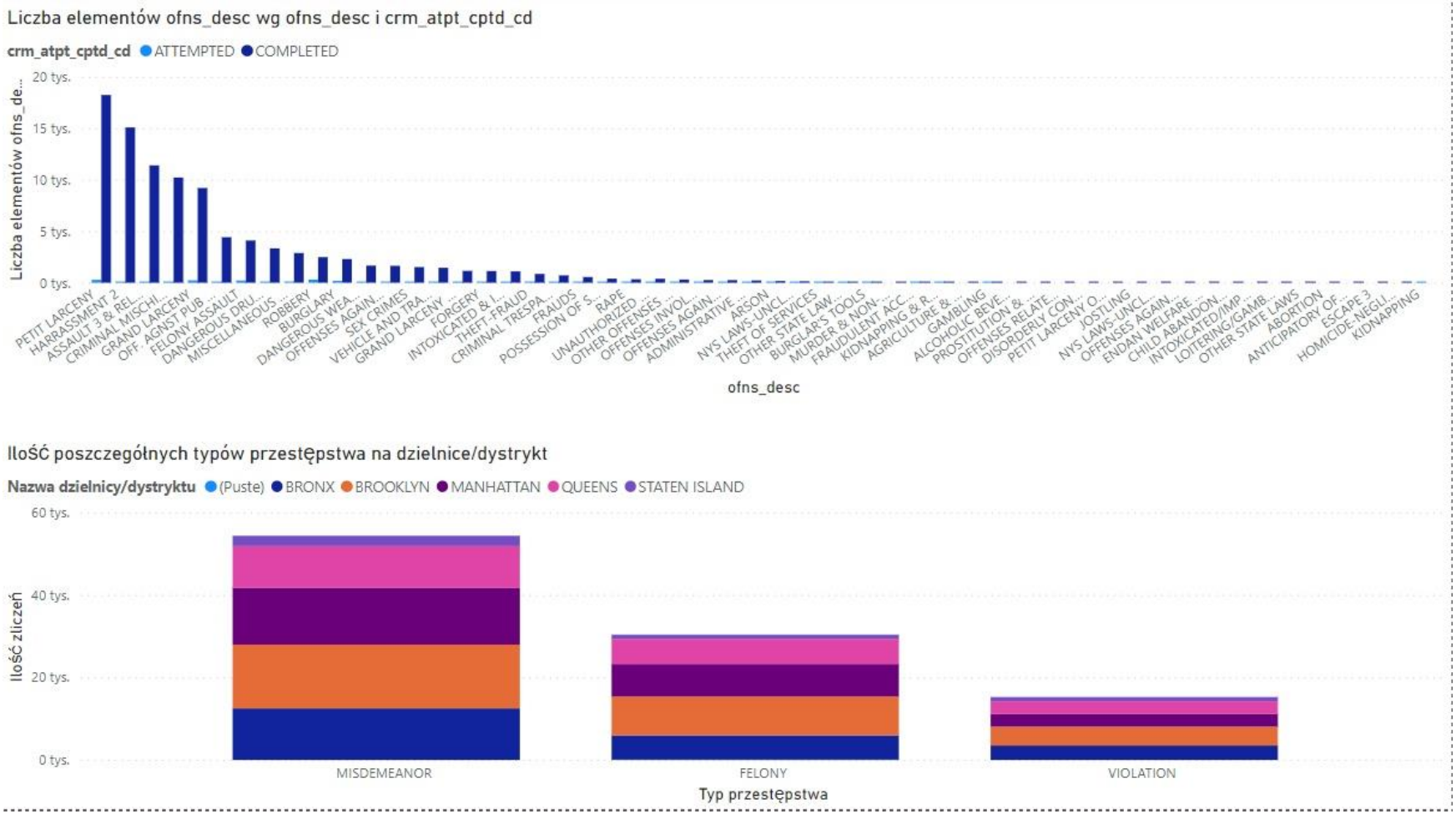
Wizualizacja III



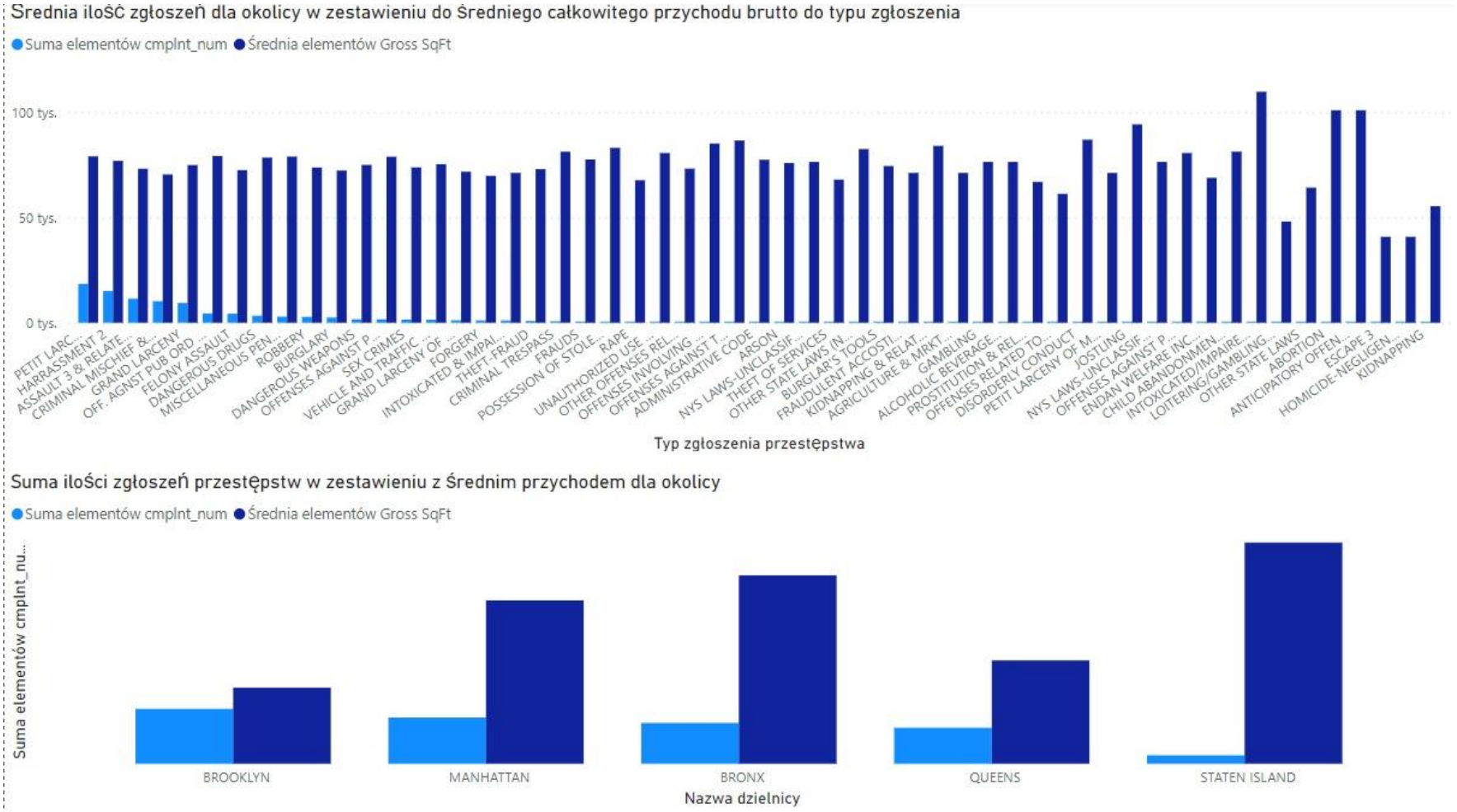
Wizualizacja IV



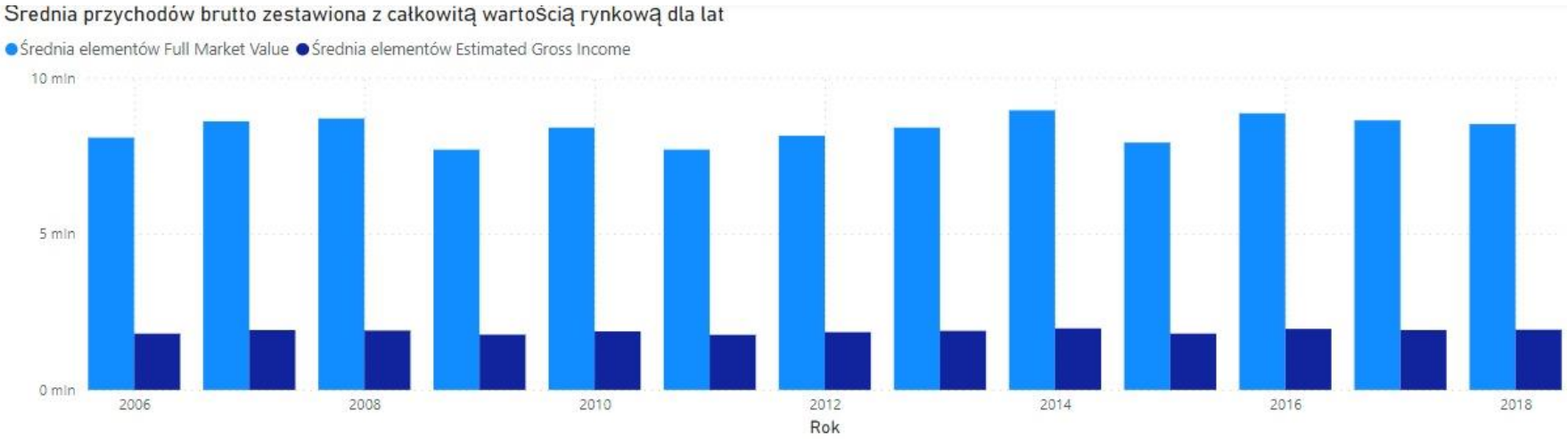
Wizualizacja V

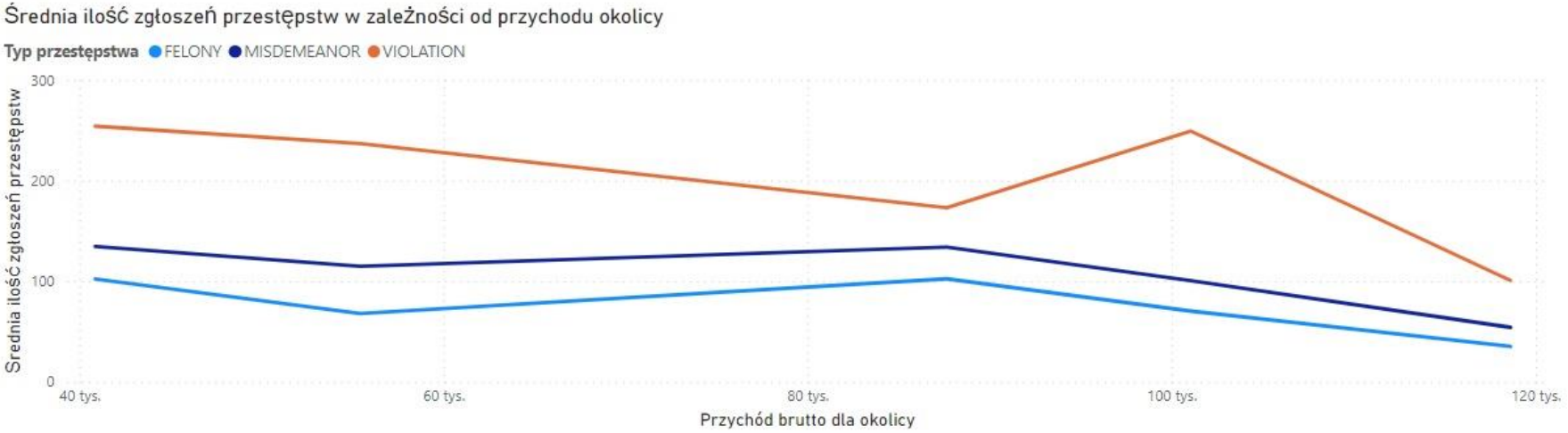


Wizualizacja VI



Wizualizacja VII





Wnioski i podsumowanie

- Dane wymagały dogłębniejszego preprocessingu by dokonać ich atomizacji
- Dane posiadały dużo braków miejscowych, które zostały usunięte w czasie processingu
- Dystrybucja lokalna zamożności w mieście Nowy York jest dość równomierna (wyjątkami jest Staten Island i Manhattan)
- W ogólności zwiększenie zamożności lokalnego sąsiedztwa zmniejsza przestępczość w danym rejonie (wyjątkiem jest Manhattan)
- Miasto jest dość dobrze zunifikowane i nie ma silnych trendów, choć można je lokalnie dostrzec

KONIEC

Dziękujemy za Uwagę I zapraszamy do pytań!