

Universidade do Minho
Escola de Engenharia

Engenharia de Dados para Suporte à Tomada de Decisão

Entrega Final

Grupo 1

Elementos do Grupo



João Pereira A92937
a92937@alunos.uminho.pt



Bruno Barroso A95110
a95110@alunos.uminho.pt



Carlos Rodrigues A91663
a91663@alunos.uminho.pt



Rúben Martins A97724
a97724@alunos.uminho.pt



Leonor Machado A96521
a96521@alunos.uminho.pt

Índice

| | |
|--|-----|
| Introdução | 5 |
| Datasets | 6 |
| Questões Analíticas | 6 |
| KPIS | 7 |
| Estrutura do DataLakeHouse | 7 |
| Estrutura do HDFS | 8 |
| Analise dos Datasets | 10 |
| 2015 Street Tree Census | 10 |
| 2005 Street Tree Census | 22 |
| 1995 Street Tree Census | 35 |
| Infraestruturas Verdes | 40 |
| Localização dos quintais | 46 |
| Informação do quintal | 53 |
| GreenStreets | 72 |
| BlockLot | 88 |
| Potencial | 90 |
| Transformações em Silver | 95 |
| 2015 Street Tree Census | 95 |
| 2005 Street Tree Census | 95 |
| 1995 Street Tree Census | 95 |
| Infraestruturas Verdes | 96 |
| Localização dos quintais | 96 |
| BlockLot | 96 |
| Potencial | 97 |
| Informação do quintal | 97 |
| GreenStrets | 97 |
| Transformações em Gold | 98 |
| Tabelas das árvores | 98 |
| Tabelas da localização dos quintais | 99 |
| Tabelas dos Edifícios Verdes | 99 |
| Tabelas das informações dos quintais | 100 |
| GreenStreets | 100 |
| Dashboards | 101 |
| Árvores | 101 |
| Localização dos Jardins | 104 |

| | |
|------------------------------|-----|
| Edifícios verdes..... | 107 |
| Informação dos quintais..... | 109 |
| GreenStreets | 112 |
| Anexos..... | 115 |

Introdução

No âmbito da cadeira de Engenharia de Dados para Suporte à Tomada de Decisão foi proposta a elaboração de um projeto onde serão aplicados conhecimentos obtidos no decorrer da unidade curricular, cujo objetivo se traduz na construção de um conjunto de dashboards que auxiliam na tomada de decisões.

O tema escolhido pelo grupo foi “Áreas Verdes” e pretende analisar o número de árvores, quintais, ruas verdes e infraestrutura verdes presentes em Nova York.

Num primeiro momento começamos por analisar a qualidade dos dados presentes nos datasets escolhidos e apresentamos algumas soluções caso essa informação esteja comprometida. Para além disso, também escolhemos as KPIs e as questões analíticas que serão abordados pelo nosso trabalho.

Na segunda fase do projeto começamos por resolver os problemas assinalados na fase 1 e construímos a zona de silver.

Na fase final do projeto construímos as tabelas de gold que iriam responder às nossas questões analíticas e construímos as dashboards para a visualização dos resultados.

As dashboards construídas permitem responder às questões analíticas e tirar outras conclusões dos dados analisados.

Datasets

- 1995 Street Tree Census <https://data.cityofnewyork.us/Environment/1995-Street-Tree-Census/7gmq-dbas>
- 2005 Street Tree Census <https://data.cityofnewyork.us/Environment/2005-Street-Tree-Census/29bw-z7pj>
- 2015 Street Tree Census <https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35>
- Informação da localização do quintal
<https://data.cityofnewyork.us/dataset/GreenThumb-Garden-Info/p78i-pat6>
- Informação do quintal
<https://data.cityofnewyork.us/Environment/GreenThumb-Site-Visits/xqbk-beh5>
- Ruas Verdes <https://data.cityofnewyork.us/dataset/Greenstreets/vzj6-pcjy>
- Edifícios verdes <https://data.cityofnewyork.us/Environment/DEP-Green-Infrastructure/spjh-pz7h>
- BlockLot <https://data.cityofnewyork.us/dataset/GreenThumb-Block-Lot/fsjc-gfyh>
- Potencial <https://data.cityofnewyork.us/Environment/City-owned-sites-that-are-available-and-potential/qchy-end3>

Questões Analíticas

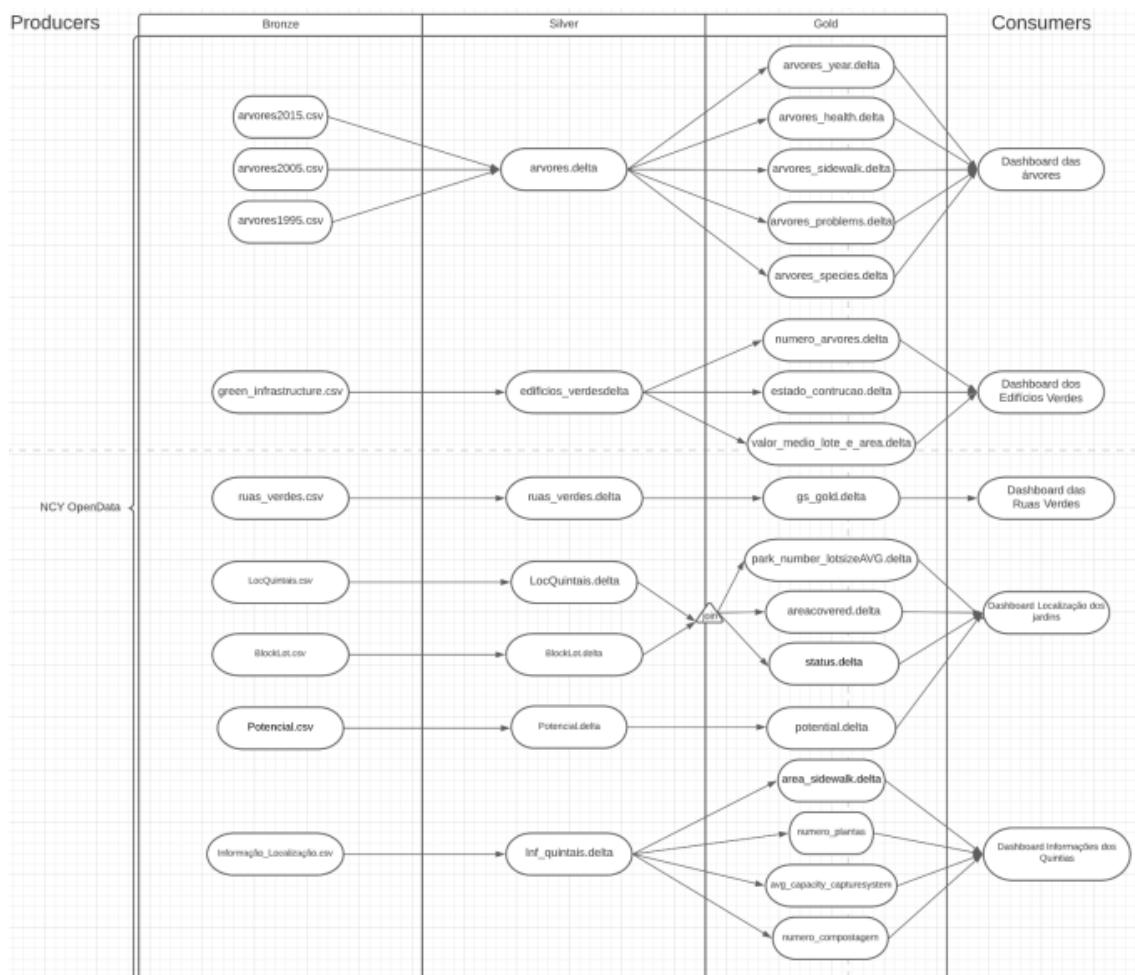
- Quais são as espécies de árvores mais comuns em Nova York?
- O número de árvores catalogadas aumentou ao longo do tempo?
- Em que estado de saúde as árvores estão?
- Quais são os problemas que as árvores têm?
- Quais são os bairros com mais árvores?
- Em que estado se encontra o passeio em que as árvores estão plantadas?
- Qual a área ocupada por edifícios verdes nos bairros de Nova York?
- Qual o estado de construção dos edifícios verdes nos bairros de Nova York?
- Quantidade de cada espécie de árvores utilizadas nos edifícios verdes em Nova York
- Qual o valor médio do lote em cada bairro de Nova York?
- Qual o bairro com mais jardins?
- Qual o tamanho médio dos jardins em cada bairro?
- Qual dos bairros têm uma maior área ocupada por jardins?
- Quais os bairros com o maior número de jardins abertos?
- Qual o bairro com maior potencial para a criação de novos jardins?
- Qual é a capacidade média dos sistemas de captura de água dos jardins em cada bairro?
- Qual é a quantidade dos tipos de plantas existentes em jardins nos bairros de New York?

- Qual é a percentagem dos tipos de compostagem existentes em jardins dos bairros de New York?
- Qual é o valor médio da área dos passeios em cada bairro?
- Qual é o bairro com maior número de ruas convertidas?
- Qual é o bairro que tem mais área total convertida?
- Qual é o bairro que apresenta maior densidade de área de rua convertida por área total?
- Qual é o grupo de comunidade responsável por mais áreas convertidas?
- Existe um desequilíbrio entre No de precincts e grupos comunitários a manter greenstreets?

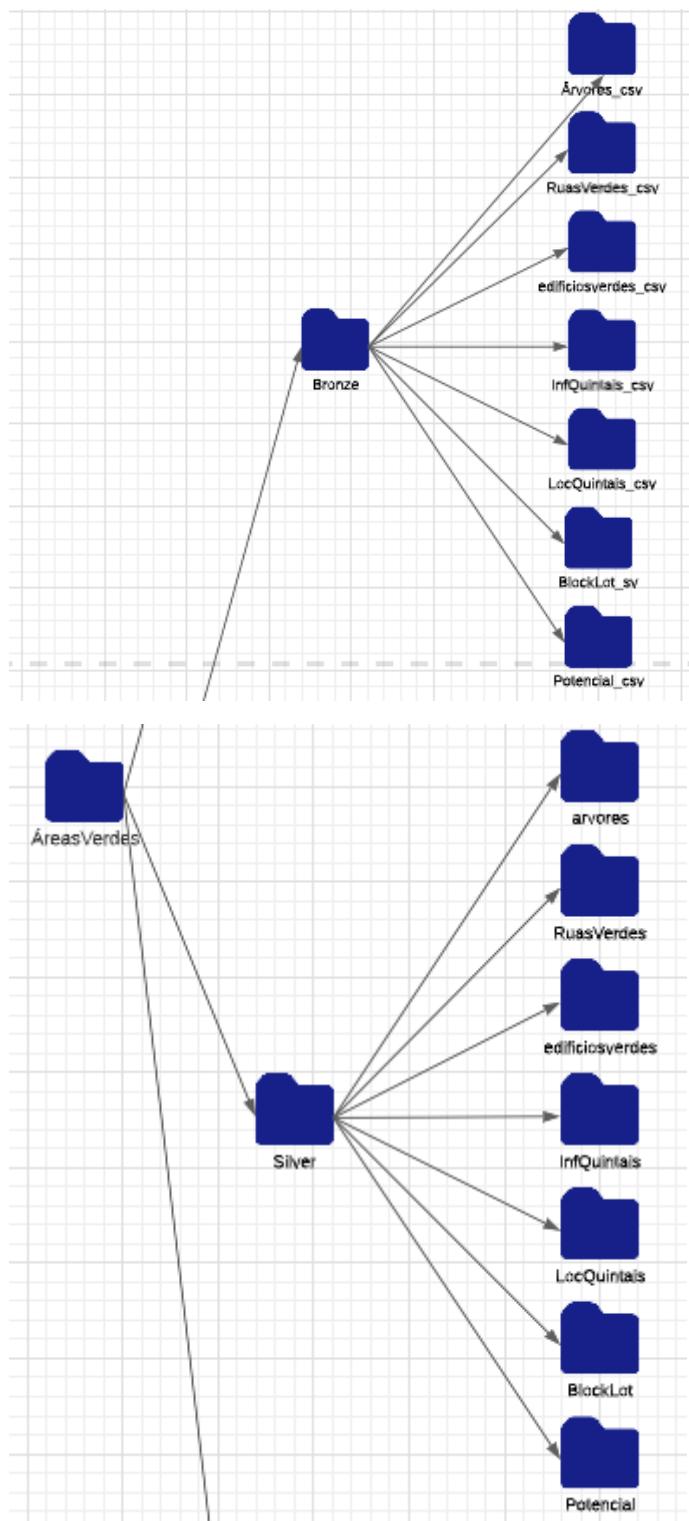
KPIs

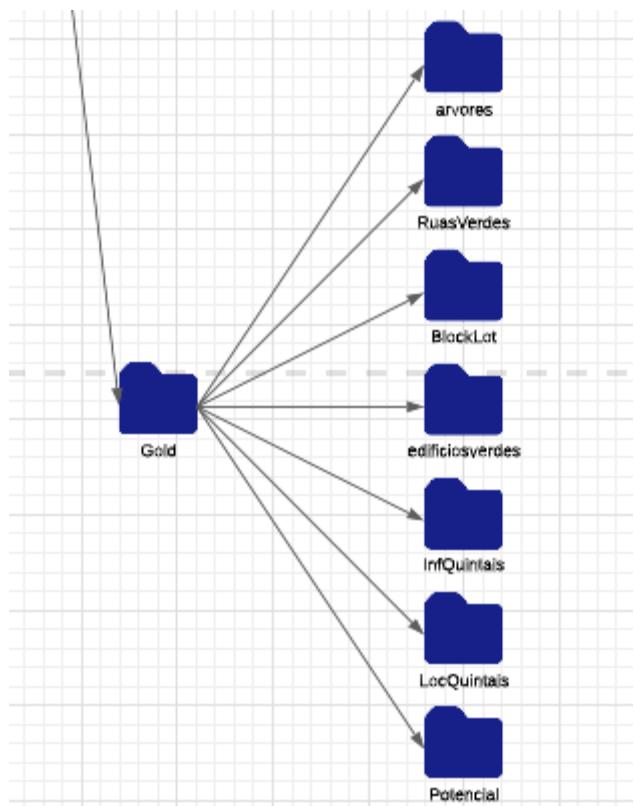
- 1 milhão de árvores catalogadas
- Taxa de 0.07 hectares de ruas convertidas por KM2

Estrutura do DataLakeHouse



Estrutura do HDFS





Analise dos Datasets

2015 Street Tree Census

Dicionário do Dataset

| Coluna | Tipo | Descrição |
|-------------------|---------|--|
| tree_id | Integer | Identificador único da árvore |
| block_id | Integer | Identificador que liga cada árvore ao bloco em que está mapeada |
| created_at | Date | Data em que a árvore foi registrada |
| tree_dbh | Integer | Diâmetro da árvore medido a aproximadamente 137cm do chão |
| stump_diam | Integer | Diâmetro do tronco |
| curb_loc | String | Localização da árvore em relação ao meio-fio |
| status | String | Indica se a árvore está viva ou morta |
| health | String | Identifica a saúde da árvore |
| spc_latin | String | Nome da árvore em latim |
| spc_common | String | Nome da árvore em inglês |
| steward | String | Identifica o número de sinais realizados pela administração para cada árvore |
| guards | String | Indica a prestação do guarda |
| sidewalk | String | Indica se o passeio perto da árvore tem danos |
| user_type | String | Categoria do usuário que catalogou esta árvore |
| root_stone | String | Indica se existe problemas na raiz causados por pedras |
| root_grate | String | Identifica se existe problemas na raiz causadas por grades metálicas |

| | | |
|-------------------|---------|--|
| root_other | String | Identifica se existem outros tipos de problemas na raiz |
| trnk_wire | String | Identifica se existem problemas no tronco causados por cordas enroladas no mesmo |
| trnk_light | String | Identifica problemas no tronco causados pela iluminação |
| trnk_other | String | Identifica outros problemas no tronco |
| brch_light | String | Identifica problemas na rama causados por luzes (luzes festivas) |
| brch_shoe | String | Indica problemas na rama causados por tênis em ramos |
| brch_other | String | Indica outros problemas na rama |
| address | String | Endereço da árvore |
| zipcode | Integer | CEP de 5 dígitos em que a árvore está localizada |
| zip_city | String | Cidade como derivado do CEP (freguesia) |
| cb_num | Integer | Conselho comunitário |
| borocode | Integer | Código do Conselho |
| Boroname | String | Nome do Conselho |
| cncldist | Integer | Distrito do conselho |
| st_assem | Integer | Distrito da Assembleia Estadual |
| st_senate | Integer | Distrito do Senado Estadual |
| nta | String | Código NTA correspondente ao bairro em que a árvore está localizada |
| nta_name | String | Nome da NTA correspondente ao bairro |
| boro_ct | String | Geocodificação para setores censitários |
| state | String | Estado onde as árvores foram catalogadas |
| latitude | Double | Longitude da árvore |
| longitude | Double | Latitude da árvore |
| x_sp | Double | Coordenada X |
| y_sp | Double | Coordenada Y |

Analise do Dataset

Coluna tree_id: Esta coluna não apresenta dados a null. Esta coluna não será descartada.

| Label | Count | % |
|------------|--------|---------|
| Row Count | 683788 | 100.00% |
| Null Count | 0 | 0.00% |

Coluna block_id: Esta coluna não apresenta dados a null e todos os dados na coluna são constituídos por 6 números. Esta coluna não será descartada.

| Value | Count | % |
|--------|--------|---------|
| 999999 | 683788 | 100.00% |

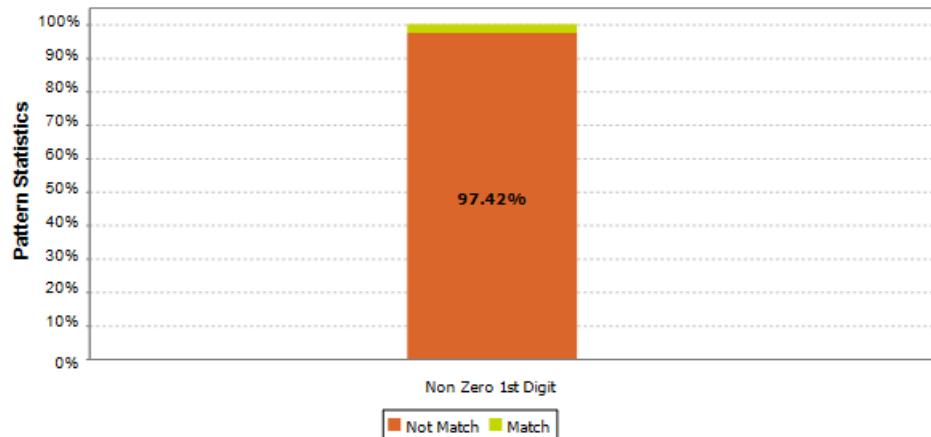
| Label | Count | % |
|------------|--------|---------|
| Row Count | 683788 | 100.00% |
| Null Count | 0 | 0.00% |

Coluna created_at: Esta coluna não apresenta dados a null e todos os dados são constituídos por NN/NN/NNNN (em que “N” representa um número). Esta coluna não será descartada.

| Value | Count | % |
|------------|--------|---------|
| 99/99/9999 | 683788 | 100.00% |

Coluna tree_dbh: Esta coluna não apresenta valores a null, contudo apresenta algumas medidas a 0 que não serão consideradas. Esta coluna não será descartada.

Coluna stum_diam: Esta coluna não apresenta valores a null, contudo apresenta uma grande quantidade de valores a 0 e por este motivo esta coluna não será considerada.



Coluna curb_loc: Esta coluna não apresenta dados a null ou em branco. Só existem 2 resultados possíveis para esta coluna (onCurb ou OffsetFromCurb). Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|----------------|----------------|--------|--------|
| OffsetFromCurb | 1 | 26892 | 3.93% |
| OnCurb | 1 | 656896 | 96.07% |

Coluna status: Esta coluna não apresenta dados a null ou em branco. Só existem 3 resultados possíveis para esta coluna (Dead, Alive ou Stump). Esta coluna será unida com a coluna health.

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| Dead | 1 | 13961 | 2.04% |
| Alive | 1 | 652173 | 95.38% |
| Stump | 1 | 17654 | 2.58% |

Coluna health: Esta coluna não apresenta dados a null mas contem 4.62% dos dados em branco. Sendo assim esta coluna só pode ter 4 resultados (Empty, Poor, Fair e Good). Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|--------|--------|
| Empty field | 1 | 31616 | 4.62% |
| Poor | 1 | 26818 | 3.92% |
| Fair | 1 | 96504 | 14.11% |
| Good | 1 | 528850 | 77.34% |

Coluna spc_latin: Esta coluna apresenta 4,62% dos dados em branco, contudo estas linhas não serão deletadas porque contem outras informações relevantes para o estudo das árvores. Esta coluna não será descartada.

Coluna spc_common: Esta coluna apresenta 4,62% dos dados em branco, contudo estas linhas não serão deletadas porque contem outras informações relevantes para o estudo das árvores. Esta coluna não será descartada.

Coluna steward: Esta coluna apresenta 4,62% dos dados em branco, sendo assim só é possível obter 4 resultados (Empty, 3or4, 4orMore e none). Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|--------|--------|
| 3or4 | 2 | 162740 | 23.80% |
| Empty field | 1 | 31615 | 4.62% |
| 4orMore | 1 | 1610 | 0.24% |
| None | 1 | 487823 | 71.34% |

Coluna guards: Esta coluna apresenta 4,62% dos dados em branco, sendo assim só é possível obter 5 resultados (Empty, Harmfull, Helpful, None e Unsure). Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|--------|--------|
| Empty field | 1 | 31616 | 4.62% |
| Harmful | 1 | 20252 | 2.96% |
| Helpful | 1 | 51866 | 7.59% |
| None | 1 | 572306 | 83.70% |
| Unsure | 1 | 7748 | 1.13% |

Coluna sidewalk: Esta coluna apresenta 4,62% dos dados em branco. Só é possível obter 3 resultados (Empty, NoDamage e Damage). Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|--------|--------|
| Empty field | 1 | 31616 | 4.62% |
| NoDamage | 1 | 464978 | 68.00% |
| Damage | 1 | 187194 | 27.38% |

Coluna user_type: Esta coluna não apresenta dados a null ou em branco. Só é possível obter 3 resultados (NYC Parks Staff, Volunteer, TreesCount Staff). Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|------------------|----------------|--------|--------|
| NYC Parks Staff | 1 | 169986 | 24.86% |
| Volunteer | 1 | 217518 | 31.81% |
| TreesCount Staff | 1 | 296284 | 43.33% |

Coluna problems: Esta coluna apresenta 4.62% dos dados em branco. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|--------|--------|
| TrunkOther | 4 | 15867 | 2.32% |
| BranchOther | 3 | 38595 | 5.64% |
| Sneakers | 2 | 116 | 0.02% |
| MetalGrates | 2 | 3110 | 0.45% |
| Stones | 2 | 139999 | 20.47% |
| WiresRope | 2 | 7531 | 1.10% |
| RootOther | 2 | 20626 | 3.02% |
| Empty field | 1 | 31664 | 4.63% |
| None | 1 | 426280 | 62.34% |

Coluna root_stone: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| No | 1 | 543789 | 79.53% |
| Yes | 1 | 139999 | 20.47% |

Coluna root_grate: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| No | 1 | 680252 | 99.48% |
| Yes | 1 | 3536 | 0.52% |

Coluna root_other: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| No | 1 | 653466 | 95.57% |
| Yes | 1 | 30322 | 4.43% |

Coluna trnk_wire: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| No | 1 | 670514 | 98.06% |
| Yes | 1 | 13274 | 1.94% |

Coluna trnk_light: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|--------|--------|
| No | 1 | 682757 | 99.85% |
| Yes | 1 | 1031 | 0.15% |

Coluna trnk_other: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

Coluna brnch_other: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

Coluna brnch_shoe: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

Coluna brnch_other: Esta coluna não apresenta dados em branco ou null e só apresenta “Yes” ou “No”. Esta coluna não será descartada.

Coluna address: Esta coluna não apresenta dados em branco ou null.

| Label | Count | % |
|-------------|--------|---------|
| Row Count | 683788 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

Coluna zip_city: Esta coluna não apresenta dados a null ou em branco. Também não apresenta dados fora do contexto da coluna. Esta coluna será descartada.

Coluna boroname: Esta coluna não apresenta dados a null ou em branco. Também não apresenta dados fora do contexto da coluna. Esta coluna não será descartada.

Coluna nta: Esta coluna não apresenta dados a null ou em branco. Também não apresenta dados fora do contexto da coluna. Esta coluna será descartada.

Coluna nta_name: Esta coluna não apresenta dados a null ou em branco. Também não apresenta dados fora do contexto da coluna. Esta coluna será descartada.

Coluna state: Esta coluna não apresenta dados a null ou em branco. Também não apresenta dados fora do contexto da coluna. Esta coluna não será descartada.

Coluna zipcode: Esta coluna apresenta alguns dados a null. Também apresenta dados fora do formato de 5 dígitos (estes dados não serão considerados). Esta coluna será descartada.

Coluna cb_num: Esta coluna apresenta alguns dados a null. Esta coluna será descartada.

| Value | Count | % |
|------------|--------|-----------|
| 999 | 620496 | 90.74% |
| Null field | 50576 | 7.40% |
| 99999 | 12711 | 1.86% |
| 99 | 5 | 7.312E-4% |

Coluna borocode: Esta coluna apresenta alguns dados a null (1,90%) que não serão considerados. Esta coluna não será descartada.

Coluna cncldist: Esta coluna apresenta alguns dados a null (7,02%) que não serão considerados. Esta coluna será descartada.

Coluna st_assem: Esta coluna apresenta alguns dados a null (1,87%) que não serão considerados. Esta coluna será descartada.

Coluna st_senate: Esta coluna apresenta alguns dados a null (0,37%) que não serão considerados. Esta coluna não será descartada.

Coluna boro_ct: Esta coluna apresenta alguns dados a null (8,84%) que não serão considerados. Esta coluna será descartada.

Coluna latitude: Esta coluna apresenta alguns dados a null (7,39%) que não serão considerados. Esta coluna será descartada.

| Value | Count | % |
|------------|--------|--------|
| 99.999999 | 383440 | 56.08% |
| 99.99999 | 213252 | 31.19% |
| Null field | 50524 | 7.39% |
| 99.9999 | 21364 | 3.12% |
| 9999999.9 | 12716 | 1.86% |
| 99.999 | 2203 | 0.32% |
| 99.99 | 216 | 0.03% |
| 99.9 | 73 | 0.01% |

Coluna longitude: Esta coluna apresenta alguns dados a null (1,90%). Esta coluna será descartada.

| Value | Count | % |
|--------------|--------|-----------|
| -99.99999999 | 558522 | 81.68% |
| -99.9999999 | 55787 | 8.16% |
| 99.99999999 | 42977 | 6.29% |
| Null field | 13014 | 1.90% |
| -99.999999 | 5592 | 0.82% |
| 99.9999999 | 4307 | 0.63% |
| 9999999.9 | 2524 | 0.37% |
| -99.99999 | 539 | 0.08% |
| 99.999999 | 418 | 0.06% |
| -99.9999 | 53 | 7.751E-3% |

Coluna x_sp: Esta coluna apresenta alguns dados a null (0,38%). Esta coluna será descartada.

| Value | Count | % |
|------------|--------|--------|
| 9999999.9 | 256991 | 37.58% |
| 999999.9 | 154280 | 22.56% |
| 9999999.99 | 116633 | 17.06% |
| 999999.99 | 92841 | 13.58% |
| -99.99999 | 32728 | 4.79% |
| -99.999999 | 11312 | 1.65% |
| 99.999999 | 7826 | 1.14% |
| 99.99999 | 4406 | 0.64% |
| -99.9999 | 3319 | 0.49% |
| Null field | 2575 | 0.38% |

Coluna x_sy: Esta coluna apresenta alguns dados a null (0,04%). Esta coluna será descartada.

| Value | Count | % |
|-------------|--------|--------|
| 999999.99 | 527055 | 77.08% |
| 999999.9 | 107506 | 15.72% |
| 9999999.9 | 21212 | 3.10% |
| 99999999.99 | 10589 | 1.55% |
| -99.99999 | 8662 | 1.27% |
| -99.999999 | 3077 | 0.45% |
| 999999.999 | 1933 | 0.28% |
| 99.999999 | 1574 | 0.23% |
| -99.9999 | 863 | 0.13% |
| 99.99999 | 857 | 0.13% |

2005 Street Tree Census

Dicionário do Dataset

| Coluna | Tipo | Descrição |
|-------------------|---------|---|
| Objectid | Integer | Id da árvore |
| cen_year | Integer | Data em que a árvore foi catalogada |
| Tree_dbh | Integer | Diâmetro da árvore medido a aproximadamente 137cm do chão |
| Tree_loc | String | Localização da árvore em relação ao endereço fornecido |
| Pip_type | String | Tipo do “poço” em que a árvore esta a crescer |
| Soil_lvl | String | Nível do solo |
| Status | String | Condições da árvore |
| Spc_latin | String | Nome da árvore em latim |
| Spc_coomon | String | Nome da árvore em inglês |
| Vert_other | String | Tratamento vertical presente |
| Vert_pgrd | String | Guarda presente |
| Vert_tgrd | String | Alta guarda presente |
| Vert_wall | String | Guarda presente (outro tipo de guarda) |
| Horz_blk | String | Pavimentos de presentes |
| Horz_grate | String | Grades da árvore presentes |
| Horz_plnt | String | Plantações presentes |
| Horz_other | String | Outro tratamento horizontal |
| Sidw_crack | String | Passeio rachado |
| Sidw_raise | String | Elevações no passeio |
| Wire_htap | String | Indica a presença de fios elétricos de casas |
| Wire_prime | String | Indica a presença de fios primários |
| Wire_2nd | String | Indica a presença de fios secundários |
| Wire_other | String | Indica a presença de outros fios |
| Inf_canopy | String | Objetos presentes na |

| | | |
|-------------------|---------|---|
| | | rama |
| Inf_guard | String | Grade presente no chão |
| Inf_wires | String | Fios presentes |
| Inf_paving | String | Pavimentação fechada |
| Inf_outlet | String | Tomada elétrica |
| Inf_shoes | String | Ténis presente |
| Inf_lights | String | Luzes presentes |
| Inf_other | String | Outros problemas com infraestruturas |
| Trnk_dmg | String | Dano no tronco |
| Zipcode | Integer | CEP de 5 dígitos em que a árvore esta localizada |
| Zip_city | Integer | Cidade como derivado do CEP (freguesia) |
| Cb_num | Integer | Conselho comunitário |
| borocode | Integer | Código do conselho |
| Boroname | String | Nome do Conselho |
| cncldist | Integer | Distrito do conselho |
| St_assem | Integer | Distrito da Assembleia Estadual |
| St_senate | Integer | Distrito da Assembleia Estadual |
| Nta | String | Código NTA correspondente ao bairro em que a árvore esta localizada |
| Nta_name | String | Nome da NTA correspondente ao bairro |
| Boro_ct | String | Geocodificação para setores censitários |
| State | String | Estado onde a árvore foi catalogada |
| X_sp | Double | Coordenada X |
| Y_sp | Double | Coordenada Y |
| Latitude | Double | Latitude da árvore |
| Longitude | Double | Longitude da árvore |

Analise do dataset

Coluna OBJECTID: Esta coluna apresenta uma linha a null. Esta coluna não será descartada.

| Label | Count | % |
|------------|-------|-----------|
| Row Count | 49438 | 100.00% |
| Null Count | 1 | 2.023E-3% |

Coluna cen_year: Esta coluna não apresenta dados a null. Todos os dados respeitam a estrutura NNNN (N = número). Esta coluna não será descartada.

Coluna tree_dbh: Esta coluna não apresenta problemas na estrutura dos seus dados. Esta coluna não será descartada.

Coluna address: Esta coluna não apresenta dados a null ou em branco. Esta coluna não será descartada.

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 49437 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

Coluna tree_loc: Esta coluna não apresenta dados a null ou em branco. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|-------|--------|
| Assigned | 1 | 276 | 0.56% |
| Adjacent | 1 | 1158 | 2.34% |
| Side | 1 | 6613 | 13.38% |
| Rear | 1 | 1181 | 2.39% |
| Front | 1 | 32698 | 66.14% |
| Side/Across | 1 | 587 | 1.19% |
| Median | 1 | 2391 | 4.84% |
| Across | 1 | 4191 | 8.48% |
| Side/Median | 1 | 341 | 0.69% |

Coluna pit_type: Esta coluna não apresenta dados a null ou em branco. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|----------------|----------------|-------|--------|
| Sidewalk Pit | 1 | 42873 | 86.72% |
| Lawn | 1 | 2202 | 4.45% |
| Continuous Pit | 1 | 4361 | 8.82% |

Coluna soil_lvl: Esta coluna não apresenta dados em branco ou null. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|-------|--------|
| Below Level | 1 | 6515 | 13.18% |
| Above Level | 1 | 31548 | 63.82% |
| Above | 1 | 11373 | 23.01% |

Coluna status: Esta coluna não apresenta dados a null ou em branco. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-----------|----------------|-------|--------|
| Excellent | 1 | 15770 | 31.90% |
| Dead | 1 | 659 | 1.33% |
| Poor | 1 | 4915 | 9.94% |
| Good | 1 | 28092 | 56.82% |

Coluna spc_latin: Esta coluna não apresenta irregularidades nos seus dados. Esta coluna não será descartada.

Coluna spc_common: Esta coluna não apresenta irregularidades nos seus dados. Esta coluna não será descartada.

Coluna vert_other: Esta coluna não apresenta irregularidades nos seus dados. Esta coluna será descartada.

Coluna Verth_pgrd: Esta coluna não apresenta dados a null ou em branco. Nesta coluna só é possível obter dois tipos de dados “yes” ou “no”. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|-------|--------|
| No | 1 | 39533 | 79.97% |
| Yes | 1 | 9903 | 20.03% |

Coluna Verth_tgrd: Esta coluna não apresenta dados a null ou em branco. Nesta coluna só é possível obter dois tipos de dados “yes” ou “no”. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|-------|--------|
| No | 1 | 40446 | 81.81% |
| Yes | 1 | 8990 | 18.19% |

Coluna vert_wall: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” ou “no”. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-------|----------------|-------|--------|
| No | 1 | 46278 | 93.61% |
| Yes | 1 | 3158 | 6.39% |

Coluna horz_b1ck: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (74,72%) ou “no”(25,28%). Esta coluna será descartada.

Coluna horz_grate: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (75,54%) ou “no”(24,46%). Esta coluna será descartada.

Coluna horz_plant: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (81,35%) ou “no”(18,65%). Esta coluna será descartada.

Coluna horz_other: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (84,37%) ou “no”(15,63%). Esta coluna será descartada.

Coluna sidw_crack: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (87,43%) ou “no”(12,57%). Esta coluna não será descartada.

Coluna sidw_raise: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (88,44%) ou “no”(11,56%). Esta coluna não será descartada.

Coluna wire_htap: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (97,5%) ou “no”(2,5%). Esta coluna será descartada.

Coluna wire_prime: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (99,83%) ou “no”(0,17%). Esta coluna será descartada.

Coluna wire_2nd: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (99,91%) ou “no”(0,09%). Esta coluna será descartada.

Coluna wire_other: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (99,54%) ou “no”(0,46%). Esta coluna será descartada.

Coluna inf_canopy: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (98,15%) ou “no”(1,85%). Esta coluna não será descartada.

Coluna inf_guard: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (97,24%) ou “no”(2,76%). Esta coluna não será descartada.

Coluna inf_wires: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (98,33%) ou “no”(1,67%). Esta coluna não será descartada.

Coluna inf_paving: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (97,60%) ou “no”(2,40%). Esta coluna não será descartada.

Coluna inf_outlet: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (98,12%) ou “no”(1,88%). Esta coluna não será descartada.

Coluna inf_shoes: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (98,77%) ou “no”(1,23%). Esta coluna não será descartada.

Coluna inf_other: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (96,28%) ou “no”(3,72%). Esta coluna não será descartada.

Coluna inf_lights: Não existem problemas de qualidade dos dados nesta coluna. Nesta coluna só é possível obter dois tipos de resultados “yes” (99,26%) ou “no”(0,74%). Esta coluna não será descartada.

Coluna trunk_dmg: Esta coluna apresenta algumas linhas em branco. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|-------------|----------------|-------|--------|
| Empty field | 1 | 426 | 0.86% |
| No | 1 | 22434 | 45.38% |
| Cavity | 1 | 1839 | 3.72% |
| Torn Bark | 1 | 2513 | 5.08% |
| Yes | 1 | 1258 | 2.54% |
| Trunk Wound | 1 | 5404 | 10.93% |
| None | 1 | 15562 | 31.48% |

Coluna zip_city: Esta coluna apresenta um grande número de dados que não fazem sentido no contexto da coluna. Estes dados serão considerados “New York”. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|----------|----------------|-------|--------|
| 11234 | 46 | 23898 | 48.34% |
| New York | 1 | 25529 | 51.64% |
| Brooklyn | 1 | 9 | 0.02% |

Coluna boroname: Esta coluna apresenta um elevado número de dados com o número “3” que serão considerados “Brooklyn”. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|-----------|----------------|-------|--------|
| 3 | 2 | 23692 | 47.92% |
| Brooklyn | 1 | 9 | 0.02% |
| Manhattan | 1 | 25735 | 52.06% |

Coluna nta: Esta coluna apresenta alguns dados em branco. Esta coluna será descartada.

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 49436 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 520 | 1.05% |

Coluna nta_name: Esta coluna apresenta alguns dados em branco. Esta coluna será descartada.

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 49436 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 971 | 1.96% |

Coluna state: Esta coluna só deveria apresentar o resultado “New York”, logo resultados diferentes serão convertidos. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|----------|----------------|-------|--------|
| 3070600 | 284 | 23692 | 47.92% |
| New York | 1 | 25744 | 52.08% |

Coluna Location_1: Esta coluna não consta no dicionário do dataset e por esse motivo não será considerada na análise dos dados.

Coluna zipcode: Esta coluna apresenta uma grande quantidade de dados a null. Esta coluna será descartada.

| Value | Count | % |
|------------|-------|--------|
| 99999 | 25538 | 51.66% |
| Null field | 23692 | 47.92% |
| 9 | 206 | 0.42% |

Coluna cb_num: Esta coluna apresenta uma grande quantidade de dados a null e 1,05% dos dados com um formato que não é aceitável pela coluna. Esta coluna será descartada.

| Value | Count | % |
|------------|-------|--------|
| 999 | 25224 | 51.02% |
| Null field | 23692 | 47.92% |
| 9 | 520 | 1.05% |

Coluna corocode: Esta coluna não apresenta irregularidades nos dados. Esta coluna não será descartada.

Coluna cncldist: Esta coluna apresenta uma grande quantidade de dados a null. Esta coluna será descartada.

| Value | Count | % |
|------------|-------|--------|
| 9 | 24175 | 48.90% |
| Null field | 23692 | 47.92% |
| 99 | 1569 | 3.17% |

Coluna st_assem: Esta coluna não apresenta irregularidades nos dados. Esta coluna será descartada.

Coluna st_senate: Esta coluna não apresenta irregularidades nos dados. Esta coluna será descartada.

Coluna boro_ct: Esta coluna apresenta uma grande quantidade de dados a null. Esta coluna será descartada.

Coluna latitude: Esta coluna apresenta uma grande quantidade de dados a null. Esta coluna será descartada.

| Label | Count | % |
|------------|-------|---------|
| Row Count | 49436 | 100.00% |
| Null Count | 23692 | 47.92% |

Coluna longitude: Esta coluna não apresenta irregularidades na qualidade de dados, contudo a informação desta coluna só tem valor em conjunto com a latitude (que apresenta muitos dados a null).

Coluna y_sp: Esta coluna não apresenta irregularidades na qualidade dos dados.

Coluna x_sp: Esta coluna apresenta uma grande quantidade de dados a null. Esta informação em falta vai condicionar a informação da coluna y_sp.

Coluna objectid_1: Esta coluna não consta no dicionário do dataset e por isso não será considerada.

Coluna bin: Esta coluna não consta no dicionário do dataset e por isso não será considerada.

Coluna censos_trat: Esta coluna não consta no dicionário do dataset e por isso não será considerada.

Coluna bbl: Esta coluna não consta no dicionário do dataset e por isso não será considerada.

1995 Street Tree Census

Dicionário do dataset

| Coluna | Tipo | Descrição |
|---------------------------|---------|--|
| _RecordID | Integer | Identificador único da árvore |
| House_Number | Integer | Identifica o número da porta da casa |
| Street | String | Rua em que a árvore está localizada |
| Postcode_original | Integer | Código postal |
| diameter | Integer | Diâmetro do tronco |
| Site | String | Localização da árvore em relação ao meio-fio |
| Sidewalk_condition | String | Indica as condições do passeio |
| Condition | String | Identifica a saúda da árvore |
| Species | String | Nome da árvore |
| Support_Structure | String | Estrutura de suporte |
| latitude | Double | Longitude da árvore |
| longitude | Double | Latitude da árvore |
| x_sp | Double | Coordenada X |
| y_sp | Double | Coordenada Y |
| Borough | String | Bairro |
| Zip_new | Integer | Cidade como derivado do CEP (freguesia) |
| Spc_latin | String | Nome da árvore em latim |
| Spc_common | String | Nome da árvore em inglês |

Nota: Algumas colunas foram descartadas por não constarem no dicionário do dataset.

Analise do dataset

Coluna _RecordId: Esta coluna não apresenta irregularidades nos dados. Esta coluna não será descartada.

Coluna House_number: Esta coluna apresenta uma grande quantidade de dados a null e por esse motivo deve ser descartada.

| Label | Count | % |
|------------|--------|---------|
| Row Count | 516989 | 100.00% |
| Null Count | 222944 | 43.12% |

Coluna Postcode_Original: Esta coluna apresenta alguns dados a null e outros fora da estrutura de 5 números. Esta coluna será descartada.

| Value | Count | % |
|------------|--------|-----------|
| 99999 | 515080 | 99.63% |
| 9 | 1904 | 0.37% |
| Null field | 5 | 9.671E-4% |

Coluna Diameter: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna não será descartada.

Coluna x_sp: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

Coluna y_sp: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

Coluna Longitude: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

Coluna latitude: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

Coluna Zip_new: Esta coluna apresenta alguns dados com uma estrutura incorreta (a correta são 5 números). Esta coluna será descartada.

Coluna Address: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna não será descartada.

Coluna Street: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

Coluna site: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|---------------|----------------|--------|-----------|
| Assigned/Side | 4 | 20278 | 3.92% |
| Across/Side | 3 | 47737 | 9.23% |
| Adjacent | 1 | 47154 | 9.12% |
| Side | 1 | 63837 | 12.35% |
| Side/Adjacent | 1 | 499 | 0.10% |
| Rear | 1 | 537 | 0.10% |
| 412 | 1 | 5 | 9.671E-4% |
| Median/Side | 1 | 357 | 0.07% |
| Front | 1 | 326958 | 63.24% |
| Median | 1 | 9627 | 1.86% |

Coluna species: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna não será descartada.

Coluna condition: Esta coluna apresenta 5 linhas com resultados que não são relevantes para a coluna. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|----------------|----------------|--------|-----------|
| 3 | 1 | 5 | 9.671E-4% |
| Excellent | 1 | 100286 | 19.40% |
| Dead | 1 | 12859 | 2.49% |
| Poor | 1 | 38567 | 7.46% |
| Fair | 1 | 327 | 0.06% |
| Shaft | 1 | 303 | 0.06% |
| Critical | 1 | 2 | 3.869E-4% |
| Planting Space | 1 | 15231 | 2.95% |
| Good | 1 | 332561 | 64.33% |
| Stump | 1 | 6087 | 1.18% |

Coluna sidewalk_condition: Esta coluna não apresenta irregularidades na qualidade dos dados. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|--------|----------------|--------|-----------|
| NA | 1 | 21273 | 4.11% |
| Good | 1 | 450444 | 87.13% |
| None | 1 | 5 | 9.671E-4% |
| Raised | 1 | 45267 | 8.76% |

Coluna Support_Strucure: Esta coluna não apresenta irregularidades nos dados. Esta coluna será descartada.

| Value | Distinct Count | Count | % |
|------------------|----------------|--------|-----------|
| Stakes / Wires | 1 | 23008 | 4.45% |
| Perimeter Guard | 1 | 1 | 1.934E-4% |
| Guard Strangling | 1 | 624 | 0.12% |
| Good | 1 | 5 | 9.671E-4% |
| None | 1 | 493351 | 95.43% |

Coluna Borough: Esta coluna não apresenta irregularidades nos dados. Esta coluna não será descartada.

| Value | Distinct Count | Count | % |
|----------------|----------------|--------|-----------|
| Stakes / Wires | 1 | 1 | 1.934E-4% |
| Queens | 1 | 227547 | 44.01% |
| Staten Island | 1 | 76634 | 14.82% |
| Brooklyn | 1 | 117101 | 22.65% |
| Manhattan | 1 | 47215 | 9.13% |
| None | 1 | 4 | 7.737E-4% |
| Bronx | 1 | 48487 | 9.38% |

Coluna SPC_Common: Esta coluna não apresenta irregularidades nos dados. Esta coluna não será descartada.

Coluna SPC_Latin: Esta coluna não apresenta irregularidades Esta coluna não será descartada.

Infraestruturas Verdes

Dicionário do Dataset

| Nome | Tipo | Descrição |
|-----------------------------|--------|--|
| The_geom | String | Localização geográfica. |
| GI_ID | String | ID exclusivo a um contrato ou fase. |
| DEP_Contra | String | Número de contrato específico do departamento para o projeto. |
| Project_Ty | String | A classe do projeto em relação à sua localização e a fonte de financiamento/desenho/propriedade da área afetada. |
| Row_Onsite | String | Projetos de direito de passagem (ROW) diferenciados de todos os outros projetos (no local). |
| Project_Na | String | O nome do projeto que contém o contrato do ativo. |
| Asset_Type | String | A classe de infraestrutura verde que define as especificidades do design e da finalidade de um ativo. |
| Status | String | Uma designação de nível que representa o estado de projeto e/ou conclusão da construção do ativo. |
| Borough | String | Bairro onde se localiza o ativo. |
| Sewer_Type | String | O tipo de sistema de esgoto na área que contém o ativo. |
| Outfall | String | O afluente em que o ativo está localizado. |
| Waterbody | String | Corpo de água envolvido. |
| Street_Add | String | O endereço físico mais próximo do ativo. |
| Nearest_In | String | A rua transversal mais próxima do endereço do ativo. |
| Secondary_Assembly_D | String | Não consta no dicionário de dados. |
| GI_Feature | String | O distrito de montagem em que o ativo está localizado. |
| Tree_Latin | String | A categoria de infraestrutura verde à qual o ativo pertence. |
| Tree_Como | String | Espécie de árvore incluída no ativo, nome latino. |
| Constructi | String | Espécie de árvore incluída no ativo, nome comum. |
| Construc_1 | String | Número do contrato de construção DEP. |

| | | |
|-------------------|---------|---|
| | | DEP. |
| Status_Gro | String | Significado não consta no dicionário de dados. |
| Asset_Id | Integer | Id do ativo. |
| DEP_Cont_1 | Integer | A designação de fase para o ativo, denotando seu progresso dentro do contrato. |
| Community | Integer | O conselho comunitário em que o ativo está localizado. |
| City_Counc | Integer | O distrito do conselho da cidade em que o ativo está localizado. |
| Asset_X_Co | Float | O X ou coordenada longitudinal do ativo, no sistema State Plane EPSG #3104. |
| Asset_Y_Co | Float | O Y ou coordenada latitudinal do ativo, no sistema State Plane EPSG #3104. |
| Asset_leng | Float | O comprimento da área de cobertura do ativo em pés. |
| Asset_Widt | Float | A largura da área de cobertura do ativo em pés. |
| Asset_Area | Float | A área de cobertura do ativo em pés quadrados. |
| BBL | Long | Valor do lote de quarteirão em que o ativo se encontra ou imediatamente à frente. |

Analise do dataset

The_geom- 0% null e 0% blanck. Soundex Frequency efetuado encontrando apenas um tipo de dado.

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 14270 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

GI_ID- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|--------|----------------|-------|--------|
| IB993a | 3401 | 3614 | 25.33% |
| A3 | 2471 | 3053 | 21.39% |
| BP-B1 | 1625 | 1849 | 12.96% |
| Ch-c | 899 | 977 | 6.85% |
| IB993C | 839 | 862 | 6.04% |
| DA9-E | 627 | 673 | 4.72% |
| IB975D | 422 | 434 | 3.04% |
| GS986C | 412 | 416 | 2.92% |
| E-95a | 373 | 385 | 2.70% |
| 992 | 224 | 237 | 1.66% |

DEP_Contra- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|------------------|----------------|-------|--------|
| GXWI71-01 | 58 | 6186 | 43.35% |
| GXHP33-01 | 38 | 2213 | 15.51% |
| BEPA_CPI_9 | 34 | 34 | 0.24% |
| B541-115M | 26 | 39 | 0.27% |
| QG-518M | 26 | 44 | 0.31% |
| GK26W5-06-OS8 | 25 | 76 | 0.53% |
| GQNC83-04 | 21 | 1610 | 11.28% |
| XG-418M | 20 | 37 | 0.26% |
| OGI-DESIGN-3-OS5 | 19 | 749 | 5.25% |
| M402-116M | 16 | 24 | 0.17% |

Project_Ty- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|-----------|----------------|-------|--------|
| Grant | 1 | 117 | 0.82% |
| Onsite | 1 | 591 | 4.14% |
| Area-Wide | 1 | 12877 | 90.24% |
| External | 1 | 685 | 4.80% |

Row_Onsite- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|--------|----------------|-------|--------|
| Onsite | 1 | 889 | 6.23% |
| ROW | 1 | 13381 | 93.77% |

Project_Na- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|----------------------------------|----------------|-------|--------|
| P.S. 91 Q | 36 | 90 | 0.63% |
| EDC HP-033 | 19 | 5054 | 35.42% |
| Brownsville Playground | 9 | 15 | 0.11% |
| DPR HP-009 | 9 | 421 | 2.95% |
| P.S. 361 M | 8 | 28 | 0.20% |
| I.S. 73 Q | 8 | 13 | 0.09% |
| Queens In-House Parkland Retr... | 6 | 7 | 0.05% |
| Flushing-Gowanus EBP 2010 | 6 | 24 | 0.17% |
| OGI Design Area 3 (BB-008) | 5 | 713 | 5.00% |
| DDC TI-03-23 Phase 1 | 5 | 2112 | 14.80% |

Asset_Type- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|------------------------------------|----------------|-------|-----------|
| Subsurface Storage | 4 | 280 | 1.96% |
| ROW Infiltration Basin with Gra... | 3 | 4984 | 34.93% |
| Detention System (Connected t... | 2 | 12 | 0.08% |
| ROWSGS | 2 | 531 | 3.72% |
| ROW Porous Concrete | 2 | 92 | 0.64% |
| ROWEB | 2 | 7501 | 52.56% |
| Green Roof | 1 | 63 | 0.44% |
| Ext. Masonry & Roofs)" | 1 | 1 | 7.008E-3% |
| Engineered Soil Tree Pit | 1 | 4 | 0.03% |
| Lighting and Seating Area Cons... | 1 | 2 | 0.01% |

Status- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|----------------------------------|----------------|-------|-----------|
| 90% Design Submitted | 3 | 1952 | 13.68% |
| Constructed (In Guarantee) | 3 | 10792 | 75.63% |
| and Belt Parkway" | 2 | 6 | 0.04% |
| or Combined Sewers in Various... | 1 | 4 | 0.03% |
| Sport Courts and AFA Construc... | 1 | 1 | 7.008E-3% |
| & Playgrounds" | 1 | 1 | 7.008E-3% |
| Rain Garden | 1 | 1 | 7.008E-3% |
| Green Roof | 1 | 1 | 7.008E-3% |
| In Construction | 1 | 1510 | 10.58% |
| Permeable Pavers | 1 | 1 | 7.008E-3% |

Borough- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|---------------|----------------|-------|--------|
| 992279.42492 | 9 | 12 | 0.08% |
| Queens | 1 | 7348 | 51.49% |
| Staten Island | 1 | 37 | 0.26% |
| Brooklyn | 1 | 5810 | 40.71% |
| Constructed | 1 | 4 | 0.03% |
| Manhattan | 1 | 118 | 0.83% |
| Bronx | 1 | 941 | 6.59% |

Sewer_Type- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|--------------------|----------------|-------|-----------|
| 231139.32332 | 11 | 13 | 0.09% |
| Staten Island | 1 | 2 | 0.01% |
| Non-combined | 1 | 118 | 0.83% |
| Combined | 1 | 14039 | 98.38% |
| Bronx | 1 | 1 | 7.008E-3% |
| On-site management | 1 | 3 | 0.02% |
| MS4 | 1 | 94 | 0.66% |

Outfall- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|---------|----------------|-------|--------|
| HP-637 | 26 | 906 | 6.35% |
| NCM-087 | 13 | 55 | 0.39% |
| TI-660 | 13 | 2490 | 17.45% |
| BB-031 | 12 | 1972 | 13.82% |
| CI-637 | 12 | 1957 | 13.71% |
| 8 | 11 | 40 | 0.28% |
| OH-023 | 10 | 103 | 0.72% |
| ROC-680 | 9 | 30 | 0.21% |
| NR-043 | 8 | 33 | 0.23% |
| NCB-083 | 8 | 1520 | 10.65% |

Waterbody- 0% null e 0% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|--------------------------|----------------|-------|-----------|
| Flushing Creek | 2 | 3128 | 21.92% |
| 55 | 1 | 2 | 0.01% |
| Hutchinson River | 1 | 204 | 1.43% |
| Gowanus Canal | 1 | 123 | 0.86% |
| HP-033 | 1 | 1 | 7.008E-3% |
| Coney Island Creek | 1 | 45 | 0.32% |
| Alley Creek | 1 | 10 | 0.07% |
| Bronx River | 1 | 430 | 3.01% |
| Combined | 1 | 7 | 0.05% |
| East River / Open Waters | 1 | 1870 | 13.10% |

Street_Add- 0% null e 73.22% blanck, o que me fez optar por descartar esta coluna. Soundex Frequency efetuado:

| Label | Count | % | Value | Distinct Count | Count | % |
|-------------|-------|---------|-----------------|----------------|-------|-------|
| Row Count | 14270 | 100.00% | 99-09 98th St | 513 | 594 | 4.16% |
| Null Count | 0 | 0.00% | 99-89 60th Ave | 327 | 381 | 2.67% |
| Blank Count | 10449 | 73.22% | E 229th St | 80 | 101 | 0.71% |
| | | | 99-90 66th Rd | 65 | 79 | 0.55% |
| | | | 95-19 101st St | 56 | 61 | 0.43% |
| | | | 98-03 103rd Ave | 55 | 64 | 0.45% |
| | | | 97-40 102nd St | 53 | 59 | 0.41% |
| | | | 97-34 103rd St | 52 | 56 | 0.39% |
| | | | 98-77 41st Ave | 51 | 60 | 0.42% |
| | | | 88-29 70th Dr | 44 | 47 | 0.33% |

Nearest_In- 0% null e 72.68% blanck , o que me fez optar por descartar esta coluna.Soundex Frequency efetuado:

| | Value | Distinct Count | Count | % |
|-------------|---------------------------|----------------|-------|-------|
| Label | 97th Ave & 98th St | 144 | 286 | 2.00% |
| Row Count | 98th St & 99th Ave | 130 | 247 | 1.73% |
| Null Count | E 98th St & Winthrop St | 75 | 200 | 1.40% |
| Blank Count | 99th St & Corona Ave | 55 | 87 | 0.61% |
| | 80th Rd & Grenfell St | 54 | 118 | 0.83% |
| | 90th St & Roosevelt Ave | 50 | 97 | 0.68% |
| | 92nd St & Corona Ave | 48 | 82 | 0.57% |
| | 91st St & Roosevelt Ave | 40 | 75 | 0.53% |
| | 99th St & Northern Blvd | 39 | 61 | 0.43% |
| | 73rd Ave & Woodhaven Blvd | 38 | 70 | 0.49% |

Secondary- 0% null e 86.90% blanck o que me fez optar por descartar esta coluna. Soundex Frequency efetuado, encontrando apenas um tipo de dado.

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 14270 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 12400 | 86.90% |

Assembly_D- 0% null e 1.54% blanck. Soundex Frequency efetuado tendo apenas 1.93% de Not found o que não é

grave na minha opinião.

| Label | Count | % | Value | Distinct Count | Count | % |
|-------------|-------|---------|-----------|----------------|-------|--------|
| Row Count | 14270 | 100.00% | 87 | 67 | 13995 | 98.07% |
| Null Count | 0 | 0.00% | Not Found | 1 | 275 | 1.93% |
| Blank Count | 220 | 1.54% | | | | |

GI_Feature- 0% null e 12.79% blanck. Soundex Frequency efetuado:

| Label | Count | % | Value | Distinct Count | Count | % |
|-------------|-------|---------|----------------------------------|----------------|-------|--------|
| Row Count | 14270 | 100.00% | 99 | 16 | 1841 | 12.90% |
| Null Count | 0 | 0.00% | Type B/C - Stormwater Inlet/S... | 3 | 22 | 0.15% |
| Blank Count | 1825 | 12.79% | Type DA | 2 | 251 | 1.76% |
| | | | Type C - SW Chamber | 1 | 2664 | 18.67% |
| | | | Type A - Stone Columns | 1 | 2029 | 14.22% |
| | | | Standard | 1 | 7463 | 52.30% |

Tree_Latin- 0% null e 10.07% blanck. Soundex Frequency efetuado:

| Label | Count | % | Value | Distinct Count | Count | % |
|-------------|-------|---------|------------------------------------|----------------|-------|--------|
| Row Count | 14270 | 100.00% | Quercus velutina | 13 | 828 | 5.80% |
| Null Count | 0 | 0.00% | Prunus x yedoensis 'Cascade S... | 12 | 171 | 1.20% |
| Blank Count | 1437 | 10.07% | 920 | 11 | 1450 | 10.16% |
| | | | Acer rubrum 'Redpointe' | 7 | 132 | 0.93% |
| | | | Zelkova serrata 'Wireless' | 7 | 225 | 1.58% |
| | | | Liquidambar styraciflua 'Worpil... | 6 | 165 | 1.16% |
| | | | Amelanchier x grandiflora | 6 | 231 | 1.62% |
| | | | Ginkgo biloba 'Shangri-la' | 5 | 122 | 0.85% |
| | | | Cornus mas | 5 | 27 | 0.19% |
| | | | Malus 'Royal Raindrops' | 5 | 18 | 0.13% |

Tree_Commo- 0% null e 10.10% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|----------------------------------|----------------|-------|--------|
| 800 | 5 | 1445 | 10.13% |
| American Sycamore | 5 | 98 | 0.69% |
| Eastern Redcedar | 4 | 116 | 0.81% |
| Japanese Zelkova | 3 | 249 | 1.74% |
| Amur Maple | 2 | 34 | 0.24% |
| Cherokee Sweetgum | 2 | 11 | 0.08% |
| Fastigiata Goldenraintree | 2 | 11 | 0.08% |
| Cherokee Princess Flowering D... | 2 | 6 | 0.04% |
| Chinese Sweet Gum | 2 | 4 | 0.03% |
| European Hornbeam | 2 | 27 | 0.19% |

Constructi- 0% null e 8.80% blanck. Soundex Frequency efetuado:

| Label | Count | % | Value | Distinct Count | Count | % |
|-------------|-------|---------|---------------------|----------------|-------|--------|
| Row Count | 14270 | 100.00% | GXHP33-01 | 18 | 2216 | 15.53% |
| Null Count | 0 | 0.00% | GKC105-3A | 7 | 3161 | 22.15% |
| Blank Count | 1256 | 8.80% | GQT103-23 | 7 | 1718 | 12.04% |
| | | | 53320022-05 | 6 | 3202 | 22.44% |
| | | | GI-NYR-5 | 5 | 123 | 0.86% |
| | | | BB05-02-A | 4 | 833 | 5.84% |
| | | | GKNC15-05 | 4 | 1105 | 7.74% |
| | | | GNCB14-2A | 3 | 495 | 3.47% |
| | | | GCT110-4B | 2 | 678 | 4.75% |
| | | | GXHP12-01-GXHP16-01 | 2 | 82 | 0.57% |

Construc_1- 0% null e 8.80% blanck. Soundex Frequency efetuado:

| Value | Distinct Count | Count | % |
|-------------------|----------------|-------|-----------|
| Package_3 | 6 | 2906 | 20.36% |
| 5 | 6 | 8941 | 62.66% |
| Phase-4-Stage-1 | 5 | 798 | 5.59% |
| Stage-4 | 4 | 978 | 6.85% |
| Phase-3 | 3 | 300 | 2.10% |
| Work-Order-4 | 2 | 233 | 1.63% |
| Arnold Tulip Tree | 1 | 1 | 7.008E-3% |
| Adams Crabapple | 1 | 2 | 0.01% |
| Provost | 1 | 2 | 0.01% |
| Dawn Redwood | 1 | 1 | 7.008E-3% |

Status_Gro- 0% null e 0.11% blanck. Soundex Frequency efetuado, sendo a percentagem de empty field muito reduzida para descartar a coluna.

| Value | Distinct Count | Count | % |
|-----------------|----------------|-------|--------|
| Empty field | 1 | 16 | 0.11% |
| Constructed | 1 | 10792 | 75.63% |
| Final Design | 1 | 1952 | 13.68% |
| In Construction | 1 | 1510 | 10.58% |

Asset_Id- 0% null.

DEP_Cont_1- 4.04% null.

Community_- 1.60% null.

City_Counc- 1.61% null.

Asset_X_Co- 0.11% null.

Asset_Y_Co- 0.09% null.

Asset_leng- 0.01% null.

Asset_Widt- 0% null.

Asset_Area- 0% null.

BBL- 0% null.

Resultados desta analise

Em geral, todas as colunas com uma percentagem de null e blanck reduzida e no caso das strings, de um soundex frequency que mostre que os dados que podemos obter são em grande parte úteis eu escolhi manter as colunas. No decorrer do projeto, provavelmente irei eliminar mais algumas colunas por falta de utilidade nos dados nela encontrados, para já baseei-me apenas na qualidade dos dados por coluna.

Localização dos quintais

Dicionário dataset

| Coluna | Tipo | Descrição |
|--------------------------|--------------|---|
| assemblydist | Integer | Distrito onde está localizado o jardim |
| borough | String | Bairro onde está localizado o jardim |
| communityboard | Integer | Community Board onde está localizado o jardim |
| congressionaldist | Integer | Congressional District onde está localizado o jardim |
| coundist | Integer | Council district onde está localizado o jardim |
| gardenname | String | Nome do jardim |
| juris | String | Entidade com poder legal de aplicação de leis sobre o jardim |
| multipolygon | Multipolygon | Forma do jardim |
| openhrsf | String | Horário de funcionamento às sextas-feiras |
| openhrsm | String | Horário de funcionamento às segundas-feiras |
| openhrssa | String | Horário de funcionamento aos sábados |
| openhrssu | String | Horário de funcionamento aos domingos |
| openhrsth | String | Horário de funcionamento às quintas-feiras |
| openhrstu | String | Horário de funcionamento às terças-feiras |
| openhrsw | String | Horário de funcionamento às quartas-feiras |
| parksid | String | Número de identificação de cada parque |
| policeprecinct | String | Força policial que patrulha a área onde está localizado cada jardim |
| statesenatedist | Integer | Distrito do senado do estado onde está localizado o jardim |
| status | String | Estado de atividade de cada jardim |
| zipcode | Integer | Código-postal do local do jardim |

Análise Dataset

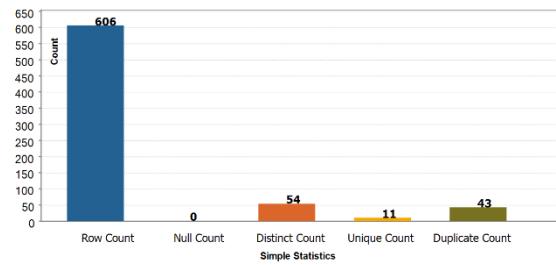
As colunas **openhrsf**, **openhrsm**, **openhrssa**, **openhrssu**, **openhrsth**, **openhrstu**, **openhrsw** não serão consideradas uma vez que a maioria dos seus dados são nulos.

As colunas **policeprecint** e **juris** também não serão consideradas nesta análise uma vez que não são relevantes ao tema.

Coluna assemblydist: esta coluna não apresenta dados a *null*, e todos eles são constituídos por dois dígitos, representados por “99”.

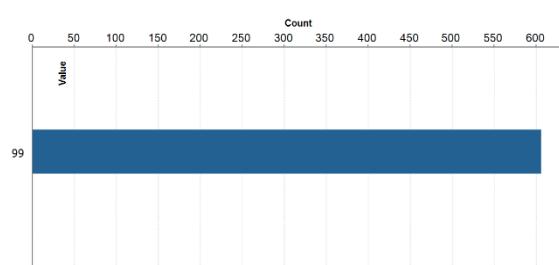
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 54 | 8.91% |
| Unique Count | 11 | 1.82% |
| Duplicate Count | 43 | 7.10% |



Pattern Frequency

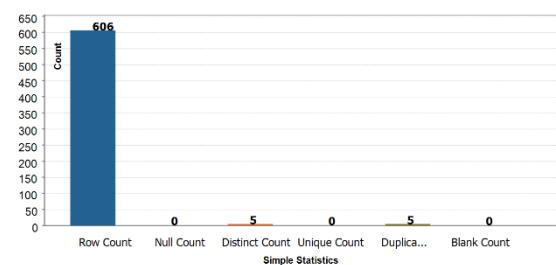
| Value | Count | % |
|-------|-------|---------|
| 99 | 606 | 100.00% |



Coluna borough: esta coluna não apresenta dados a *null*, e todos eles são constituídos por uma letra, representada por “A”.

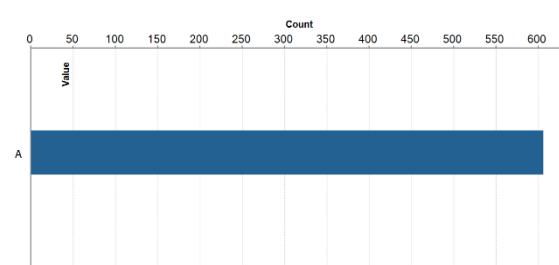
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 5 | 0.83% |
| Unique Count | 0 | 0.00% |
| Duplicate Count | 5 | 0.83% |
| Blank Count | 0 | 0.00% |



Pattern Frequency

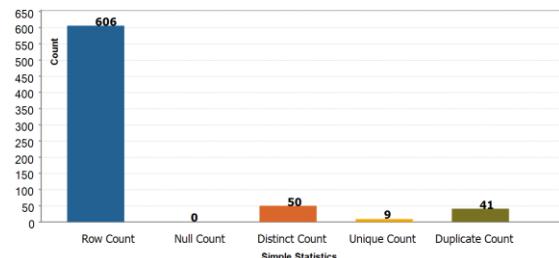
| Value | Count | % |
|-------|-------|---------|
| A | 606 | 100.00% |



Coluna communityboard: esta coluna não apresenta dados a *null*, e todos eles são constituídos por dois dígitos, representados por “999”.

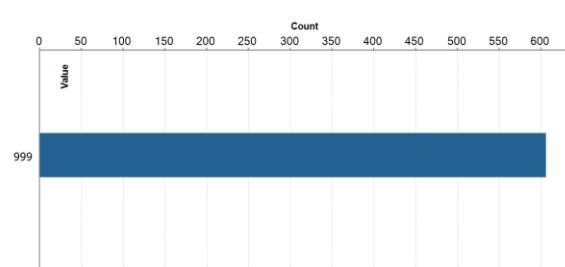
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 50 | 8.25% |
| Unique Count | 9 | 1.49% |
| Duplicate Count | 41 | 6.77% |



▼ Pattern Frequency

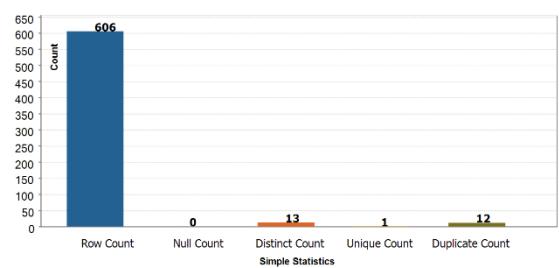
| Value | Count | % |
|-------|-------|---------|
| 999 | 606 | 100.00% |



Coluna congressionaldist: esta coluna não apresenta dados a *null*, sendo os dados constituídos por dois e um digito, representados por “99” e “9”, respectivamente.

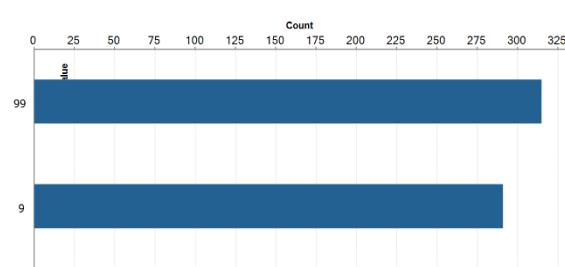
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 13 | 2.15% |
| Unique Count | 1 | 0.17% |
| Duplicate Count | 12 | 1.98% |



▼ Pattern Frequency

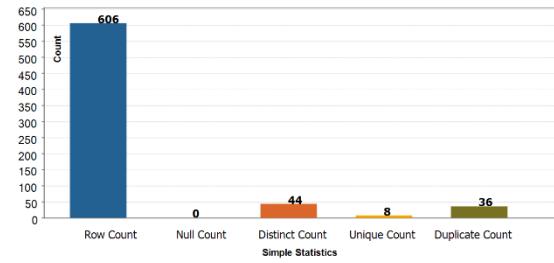
| Value | Count | % |
|-------|-------|--------|
| 99 | 315 | 51.98% |
| 9 | 291 | 48.02% |



Coluna coundist: esta coluna não apresenta dados a *null*, sendo os dados constituídos por dois e um dígito, representados por “99” e “9”, respetivamente.

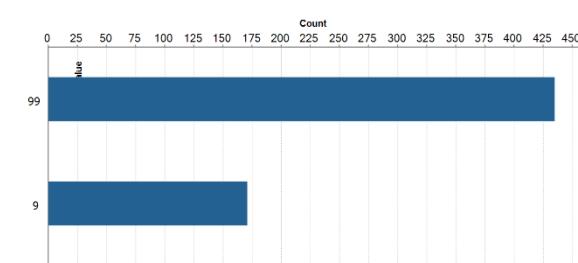
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 44 | 7.26% |
| Unique Count | 8 | 1.32% |
| Duplicate Count | 36 | 5.94% |



▼ Pattern Frequency

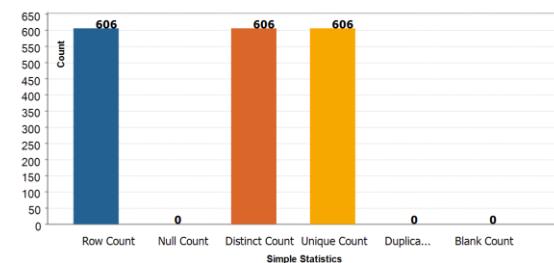
| Value | Count | % |
|-------|-------|--------|
| 99 | 435 | 71.78% |
| 9 | 171 | 28.22% |



Coluna gardenname: esta coluna não apresenta dados a *null*, sendo os dados constituídos por palavras cujas letras são representadas por “A”, tal como se verifica na tabela/gráfico abaixo.

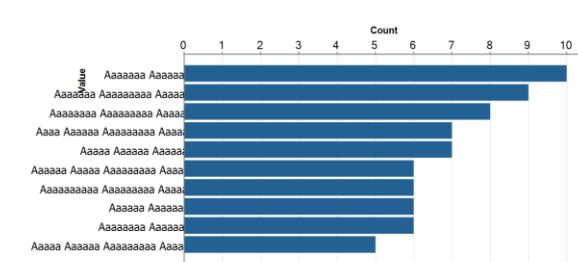
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 606 | 100.00% |
| Unique Count | 606 | 100.00% |
| Duplicate Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |



▼ Pattern Frequency

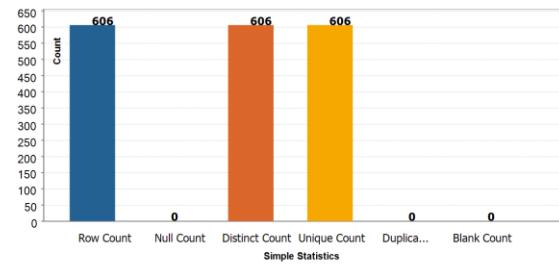
| Value | Count | % |
|------------------------------|-------|-------|
| Aaaaaaa Aaaaaaa | 10 | 1.65% |
| Aaaaaaaaa Aaaaaaa | 9 | 1.49% |
| Aaaaaaaaaa Aaaaaaa | 8 | 1.32% |
| Aaaa Aaaaaa Aaaaaaaaaa Aa... | 7 | 1.16% |
| Aaaaa Aaaaaa Aaaaaaa | 7 | 1.16% |
| Aaaaaa Aaaaa Aaaaaaaaa A... | 6 | 0.99% |
| Aaaaaaaaaa Aaaaaaaaaa Aa... | 6 | 0.99% |
| Aaaaaaa Aaaaaa | 6 | 0.99% |
| Aaaaaaaaa Aaaaaaa | 6 | 0.99% |
| Aaaaaa Aaaaaa Aaaaaaa A... | 5 | 0.83% |



Coluna multipolygon: esta coluna não apresenta dados a *null*, sendo os dados constituídos por palavras cujas letras são representadas por “A” seguindo-se por um conjunto de dígitos representados por “9”, tal como se verifica na tabela/gráfico abaixo

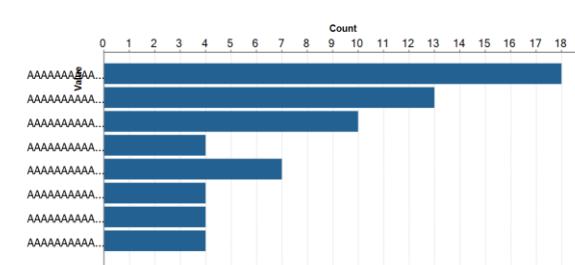
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 606 | 100.00% |
| Unique Count | 606 | 100.00% |
| Duplicate Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |



▼ Pattern Frequency

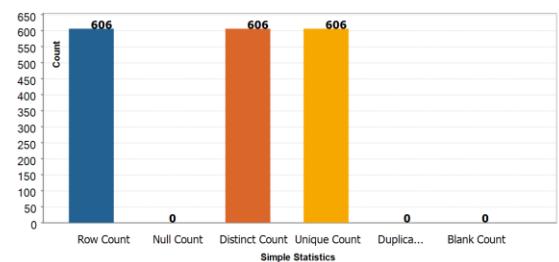
| Value | Count | % |
|--------------------------------|-------|---|
| AAAAAAAAAAAA ((-99.999... 18 | 2.97% | |
| AAAAAAAAAAAA ((-99.999... 13 | 2.15% | |
| AAAAAAAAAAAAAA ((-99.999... 10 | 1.65% | |
| AAAAAAAAAAAAAA ((-99.999... 10 | 1.65% | |
| AAAAAAAAAAAAAA ((-99.999... 9 | 1.49% | |
| AAAAAAAAAAAAAA ((-99.999... 7 | 1.16% | |
| AAAAAAAAAAAAAA ((-99.999... 4 | 0.66% | |



Coluna parksid: esta coluna não apresenta dados a null e todos eles são constituídos por um conjunto de letras e dígitos, representados por “A” e “9”, respetivamente.

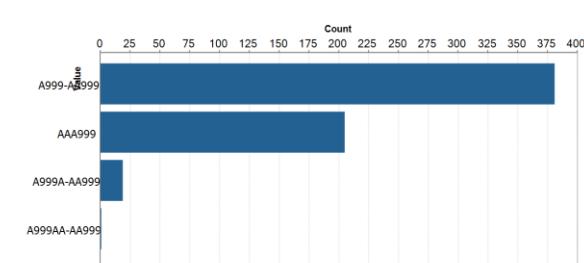
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 606 | 100.00% |
| Unique Count | 606 | 100.00% |
| Duplicate Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |



▼ Pattern Frequency

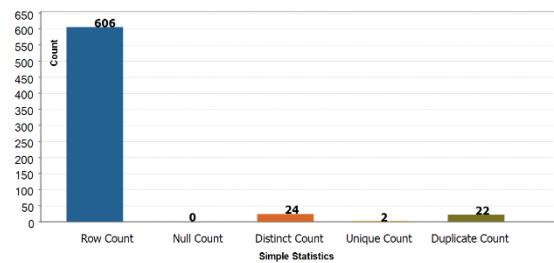
| Value | Count | % |
|--------------|-------|--------|
| A999-AA999 | 381 | 62.87% |
| AAA999 | 205 | 33.83% |
| A999A-AA999 | 19 | 3.14% |
| A999AA-AA999 | 1 | 0.17% |



Coluna statesenatedist:

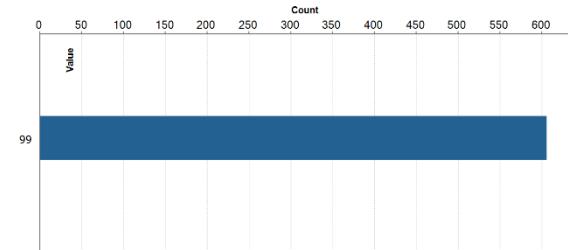
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 24 | 3.96% |
| Unique Count | 2 | 0.33% |
| Duplicate Count | 22 | 3.63% |



Pattern Frequency

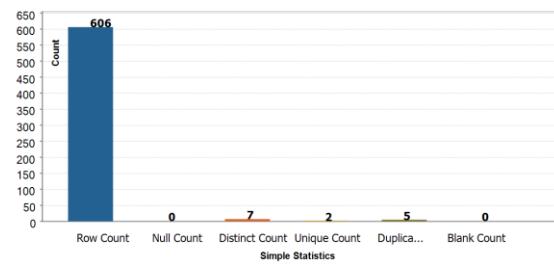
| Value | Count | % |
|-------|-------|---------|
| 99 | 606 | 100.00% |



Coluna status: esta coluna não contém linhas em branco, sendo os seus dados constituídos, maioritariamente, por um conjunto de letras, representados por “Aaaaaa”.

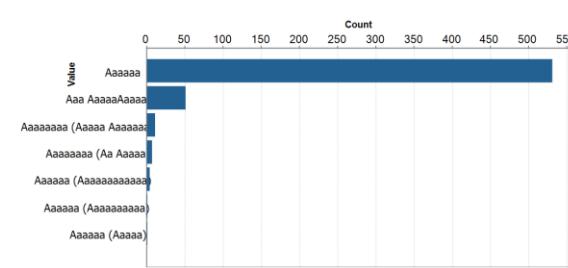
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 7 | 1.16% |
| Unique Count | 2 | 0.33% |
| Duplicate Count | 5 | 0.83% |
| Blank Count | 0 | 0.00% |



Pattern Frequency

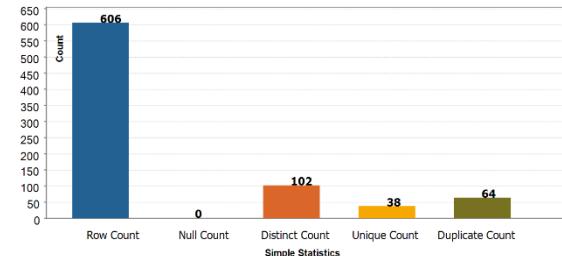
| Value | Count | % |
|---------------------------|-------|--------|
| Aaaaaa | 531 | 87.62% |
| Aaa AaaaaAaaaa | 51 | 8.42% |
| Aaaaaaaaa (Aaaaa Aaaaaaa) | 11 | 1.82% |
| Aaaaaaaaaa (Aa Aaaaa) | 7 | 1.16% |
| Aaaaaaa (Aaaaaaaaaaaa) | 4 | 0.66% |
| Aaaaaaa (Aaaaaaaaaaa) | 1 | 0.17% |
| Aaaaaaa (Aaaaa) | 1 | 0.17% |



Coluna zipcode: esta coluna não apresenta dados a null e todos os dados são constituídos por 5 dígitos representados por "99999".

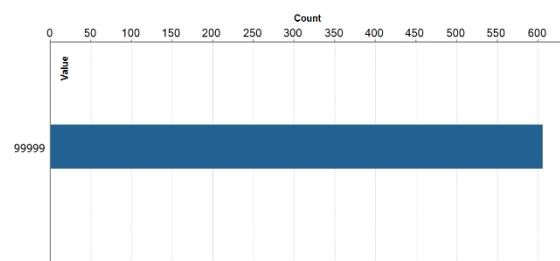
▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 606 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 102 | 16.83% |
| Unique Count | 38 | 6.27% |
| Duplicate Count | 64 | 10.56% |



▼ Pattern Frequency

| Value | Count | % |
|-------|-------|---------|
| 99999 | 606 | 100.00% |



Informação do quintal

Dicionário do dataset

| Coluna | Tipo | Descrição |
|---------------------|---------|--|
| ParksID | String | Número de identificação único para cada propriedade do parque. Identificação por bairro (B = Brooklyn; M = Manhattan; Q = Queens; R = Richmond (Staten Island); X = Bronx) e seguido de um número. GT = GreenThumb |
| InspectionID | Integer | Identificador único de cada visita ao parque. |
| NextToAnotherGarden | Boolean | O jardim está localizado perto de outro jardim GreenThumb existente? |
| TotalFenceLength | Float | Comprimento total das cercas que estão no jardim. |
| TotalSidewalkArea | Float | Área total dos passeios que rodeiam o jardim. |
| TotalSidewalkLength | Float | Comprimento total dos passeios que rodeiam o jardim. |
| OnSiteService | Boolean | Há fonte de água no jardim? |
| HydrantW_in15ft | Boolean | Há uma boca de incendio localizada em 15 metros do jardim? |
| HydrantOnGardenSide | Boolean | A Boca de incendio e o jardim estão localizados na mesma rua? |
| RainHarvesting | Boolean | Há sistema de captura de água? |
| RainGallons | Float | Capacidade do sistema de captura de água nos galões de coleta de chuva. |
| SolarPanels | Boolean | Jardim tem painéis solares? |
| CompostSystem | Boolean | Há compostagem em caixotes? |
| CompostTumblers | Boolean | Há um barril de |

| | | |
|-------------------------------|---------|--|
| | | compostagem que gira? |
| NonFoodPlants | Boolean | O jardim tem plantas não comestíveis? |
| Food | Boolean | O jardim produz comida? |
| OpenLawnOrCommunalArea | Boolean | Há uma secção aberta no jardim não dedicada a plantas? |
| PavedArea | Boolean | Área do jardim coberta por pedras, tijolos, asfalto,etc.. |
| TreesInGarden | Boolean | Há árvores sem fruto dentro do jardim? |
| FruitTrees | Boolean | Há árvores de fruto dentro do jardim? |
| StreetTrees | Boolean | Há árvores nos passeios á beira do jardim? |
| EmptyTreePits | Boolean | Há poços de árvore sem uma árvore? |
| Murals | Boolean | Há quadros ou outros trabalhos de arte executados na parede de frente ao jardim? |
| BlankShed | Boolean | Há uma parede em branco que pode ser usada para trabalhos de arte? |
| ParksSign | Boolean | Para cada jardim de comunidade sobre a lei dos parques de NYC, GreenThumb providencia um aviso que explica que aquele jardim está registado na GreenThumb, tal como as regras e/ou informações em múltiplas línguas. |
| Chickens | Boolean | Há galinhas no jardim? |
| Pond | Boolean | Há um poço no jardim? |
| FishInPond | Boolean | Há peixes no poço? |
| Turtles | Boolean | Há tartarugas no jardim? |
| Aquaponics | Boolean | Há um sistema de hidracultura em qual o lixo produzido pelos peixes ou outros animais aquáticos, dá nutrientes às plantas nascidas em água, essas mesmas que |

| | | |
|------------------------------------|---------|--|
| | | purificam a água? |
| FarmersMarket | Boolean | O Jardim tem um mercado de jardineiros? |
| CSApickup | Boolean | O Jardim é uma localização para CSA (Community Supported Agriculture) pick-up? |
| Composting | Boolean | O Terreno combina dados do copo de compostagem e dos sistemas de compostagem? |
| Greenhouse | Boolean | O Jardim tem uma estufa? |
| StructureForSeasonExtension | Boolean | O Jardim tem uma estrutura que permite cultivar comida durante o inverno? |

Análise do dataset

ParksID

0 % dos valores estão em branco/ ou são nulos.

0.56 % dos valores são duplicados (ParksID é um identificador único)

A999-AA999 aparece 95.25%.

A999A-AA999 aparece 4.75%.

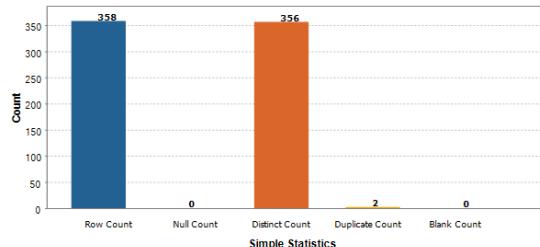
Esta coluna será mantida.

Analysis Results

Column: metadata._ParksID

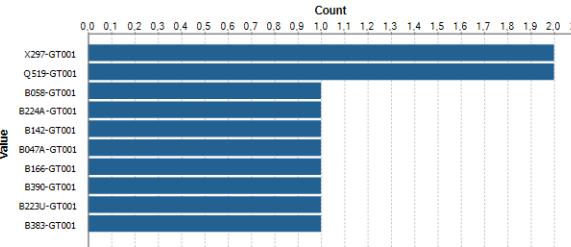
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 356 | 99.44% |
| Duplicate Count | 2 | 0.56% |
| Blank Count | 0 | 0.00% |



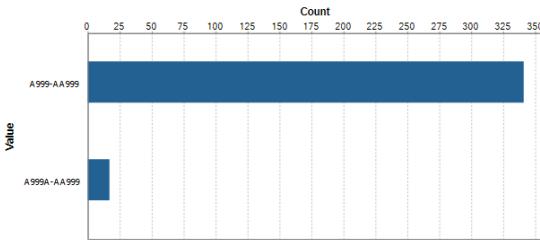
Value Frequency

| Value | Count | % |
|--------------|-------|-------|
| X297-GT001 | 2 | 0.56% |
| Q519-GT001 | 2 | 0.56% |
| B058-GT001 | 1 | 0.28% |
| B224A-GT001 | 1 | 0.28% |
| B142-GT001 | 1 | 0.28% |
| B047A-GT001 | 1 | 0.28% |
| B166-GT001 | 1 | 0.28% |
| B390-GT001 | 1 | 0.28% |
| P2221L-GT001 | 1 | 0.28% |



Pattern Frequency

| Value | Count | % |
|-------------|-------|--------|
| A999-AA999 | 341 | 95.25% |
| A999A-AA999 | 17 | 4.75% |



Go to page 1/1

InspectionID

0.56 % dos valores são nulos.

0.28 % dos valores são duplicados (InspectionID é um identificador único).

999 aparece 86.59%.

99 aparece 11.45%.

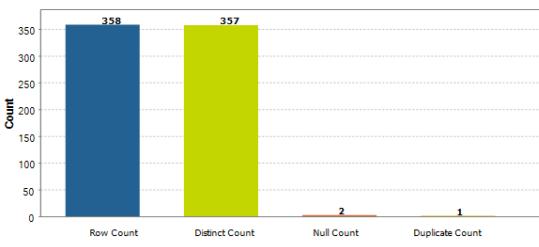
9 aparece 1.40%.

Esta coluna será mantida.

Column: metadata.InspectionID

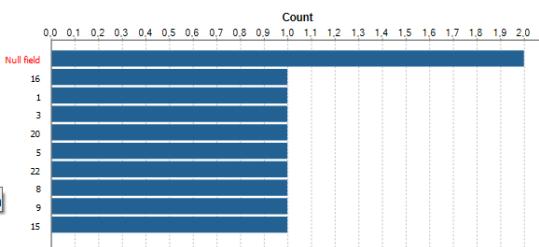
Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Distinct Count | 357 | 99.72% |
| Null Count | 2 | 0.56% |
| Duplicate Count | 1 | 0.28% |



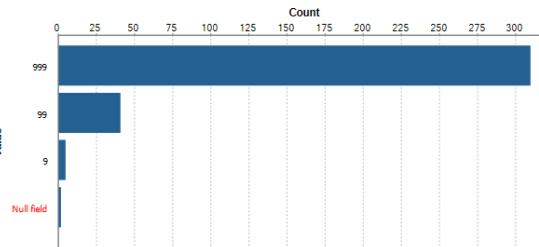
Value Frequency

| Value | Count | % |
|------------|-------|-------|
| Null field | 2 | 0.56% |
| 16 | 1 | 0.28% |
| 1 | 1 | 0.28% |
| 3 | 1 | 0.28% |
| 20 | 1 | 0.28% |
| 5 | 1 | 0.28% |
| 22 | 1 | 0.28% |
| 8 | 1 | 0.28% |
| 0 | 1 | 0.28% |



Pattern Frequency

| Value | Count | % |
|------------|-------|--------|
| 999 | 310 | 86.59% |
| 99 | 41 | 11.45% |
| 9 | 5 | 1.40% |
| Null field | 2 | 0.56% |



NextToAnotherGarden

0.84 % dos valores estão em branco.

Esta coluna será removida devido a não trazer informação relevante e até o próprio data dictionary indicar que esta coluna não é necessária.

Analysis Results

Column: metadata.NextToAnotherGarden

Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |
| Distinct Count | 3 | 0.84% |

Count

Row Count

Null Count

Blank Count

Distinct Count

Simple Statistics

Go to page

1/1

TotalFenceLength

10.34 % dos valores estão em branco.

As patterns mais frequentes são a 999,9,99.

Esta coluna será mantida.

Value Frequency

| Value | Count | % |
|-------------|-------|--------|
| 0 | 63 | 17.60% |
| Empty field | 37 | 10.34% |
| 100 | 10 | 2.79% |
| 200 | 7 | 1.96% |
| 250 | 6 | 1.68% |
| 25 | 5 | 1.40% |
| 300 | 5 | 1.40% |
| 50 | 4 | 1.12% |
| 150 | 4 | 1.12% |

Value

Count

Empty field

0

100

200

250

25

300

50

150

1.12%

1.40%

1.68%

1.96%

2.79%

10.34%

17.60%

Value Frequency

| Value | Count | % |
|-------------|-------|--------|
| 999 | 173 | 48.32% |
| 9 | 63 | 17.60% |
| 99 | 47 | 13.13% |
| Empty field | 37 | 10.34% |
| 999.9 | 18 | 5.03% |
| 999.99 | 12 | 3.35% |
| 99.9 | 4 | 1.12% |
| 9,999 | 3 | 0.84% |
| 0.000.0 | 1 | 0.29% |

Value

Count

999

9

99

Empty field

999.9

999.99

99.9

9,999

0.000.0

48.32%

17.60%

13.13%

10.34%

5.03%

3.35%

1.12%

0.84%

0.29%

Analysis Results

Column: metadata.TotalFenceLength

Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 183 | 51.12% |
| Blank Count | 37 | 10.34% |

Count

Row Count

Null Count

Distinct Count

Blank Count

Simple Statistics

TotalSidewalkArea

4.19 % dos valores estão em branco.

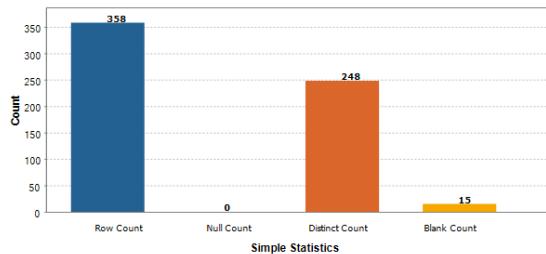
As patterns mais frequentes são 999 e 9999.

Esta coluna será mantida.

Column: metadata.TotalSidewalkArea

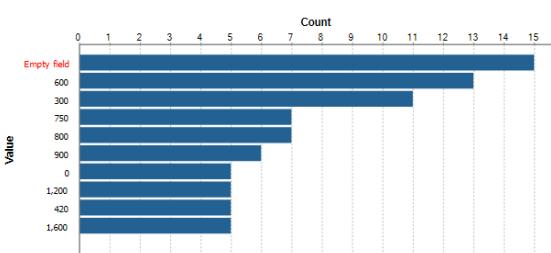
Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 248 | 69.27% |
| Blank Count | 15 | 4.19% |



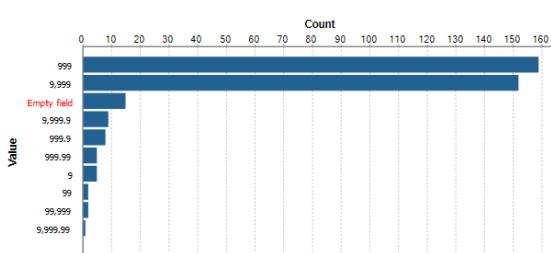
Value Frequency

| Value | Count | % |
|-------|-------|-------|
| 300 | 11 | 3.07% |
| 750 | 7 | 1.96% |
| 800 | 7 | 1.96% |
| 900 | 6 | 1.68% |
| 0 | 5 | 1.40% |
| 1,200 | 5 | 1.40% |
| 420 | 5 | 1.40% |
| 1,600 | 5 | 1.40% |



Pattern Frequency

| Value | Count | % |
|-------------|-------|--------|
| 999 | 159 | 44.41% |
| 9,999 | 152 | 42.46% |
| Empty field | 15 | 4.19% |
| 9,999.9 | 9 | 2.51% |
| 999.9 | 8 | 2.23% |
| 999.99 | 5 | 1.40% |
| 9 | 5 | 1.40% |
| 99 | 2 | 0.56% |
| 00.000 | 2 | 0.56% |



TotalSidewalkLength

4.47 % dos valores são nulos.

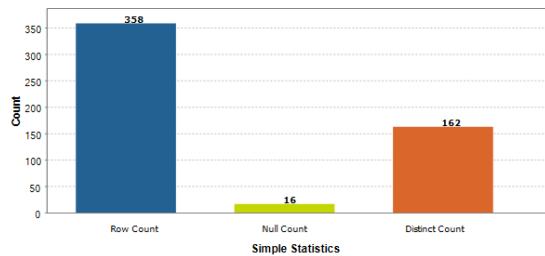
As patterns mais frequentes são 999,9 e 99,9.

Esta coluna será mantida.

Column: metadata.TotalSidewalkLength

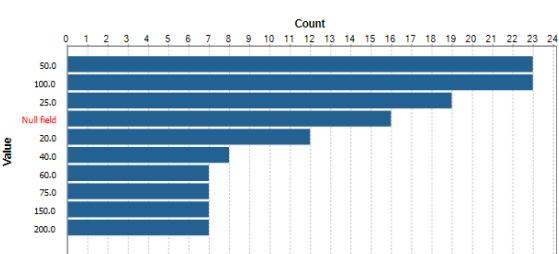
Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 16 | 4.47% |
| Distinct Count | 162 | 45.25% |



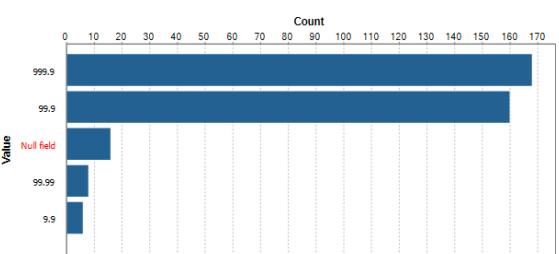
Value Frequency

| Value | Count | % |
|------------|-------|-------|
| 50.0 | 23 | 6.42% |
| 100.0 | 23 | 6.42% |
| 25.0 | 19 | 5.31% |
| Null field | 16 | 4.47% |
| 20.0 | 12 | 3.35% |
| 40.0 | 8 | 2.23% |
| 60.0 | 7 | 1.96% |
| 75.0 | 7 | 1.96% |
| 150.0 | 7 | 1.96% |



Pattern Frequency

| Value | Count | % |
|------------|-------|--------|
| 999.9 | 168 | 46.93% |
| 99.9 | 160 | 44.69% |
| Null field | 16 | 4.47% |
| 99.99 | 8 | 2.23% |
| 9.9 | 6 | 1.68% |



OnSiteService

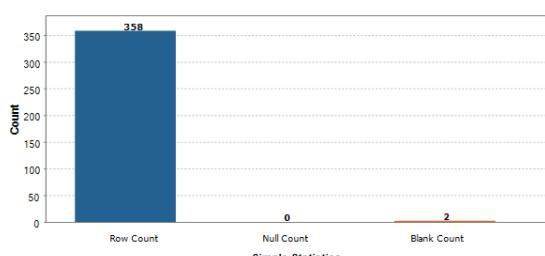
0.56 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.OnSiteService

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 2 | 0.56% |



HydrantW_in15ft

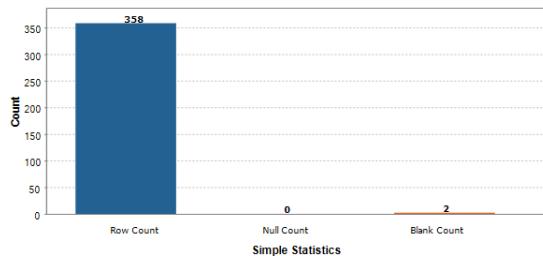
0.56 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.HydrantW_in15ft

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 2 | 0.56% |



HydrantOnGardenSide

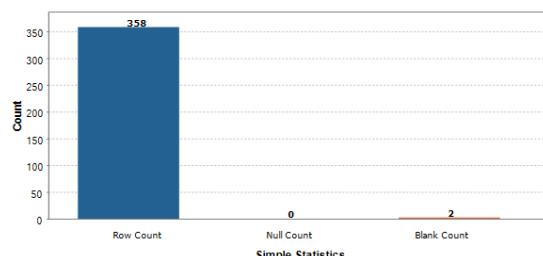
0.56 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.HydrantOnGardenSide

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 2 | 0.56% |



RainHarvesting

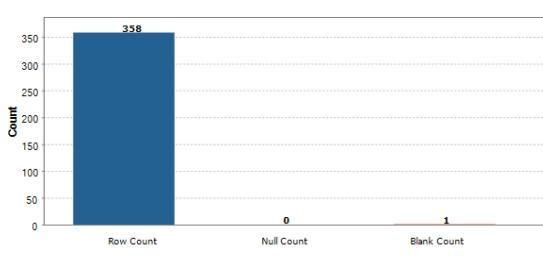
0.28 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.RainHarvesting

Simple Statistics

| Label | Count | % |
|-------------|----------------------------------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | right click and select an action | 0.28% |



RainGallons

48.32 % dos valores estão em branco.

As patterns mais frequentes são “empty field” e 999.

Esta coluna será removida, pois praticamente metade dos valores estão em branco.

Analysis Results

Column: metadata.RainGallons

Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 51 | 14.25% |
| Blank Count | 173 | 48.32% |

A bar chart titled "Simple Statistics" showing the count for each category. The categories are Row Count, Null Count, Distinct Count, and Blank Count. The values are 358, 0, 51, and 173 respectively. The Y-axis is labeled "Count" and ranges from 0 to 350. The X-axis is labeled "Simple Statistics".

Value Frequency

| Value | Count | % |
|-------------|-------|--------|
| Empty field | 173 | 48.32% |
| 1,000 | 20 | 5.59% |
| 120 | 17 | 4.75% |
| 500 | 12 | 3.35% |
| 300 | 11 | 3.07% |
| 100 | 9 | 2.51% |
| 60 | 9 | 2.51% |
| 150 | 8 | 2.23% |
| 250 | 9 | 2.23% |

A horizontal bar chart titled "Value Frequency" showing the count for each value. The values are Empty field, 1,000, 120, 500, 300, 100, 60, 150, and 250. The Y-axis is labeled "Value" and the X-axis is labeled "Count". The counts range from 8 to 173.

Pattern Low Frequency

| Value | Count | % |
|-------------|-------|--------|
| 9 | 8 | 2.23% |
| 99 | 18 | 5.03% |
| 9,999 | 35 | 9.78% |
| 999 | 124 | 34.64% |
| Empty field | 173 | 48.32% |

A horizontal bar chart titled "Pattern Low Frequency" showing the count for each value. The values are 9, 99, 9,999, 999, and Empty field. The Y-axis is labeled "Value" and the X-axis is labeled "Count". The counts range from 8 to 173.

SolarPanels

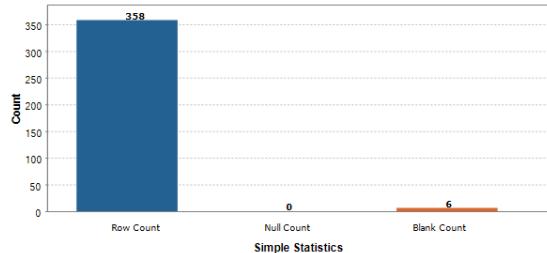
1.68 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.SolarPanels

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 6 | 1.68% |



CompostSystem

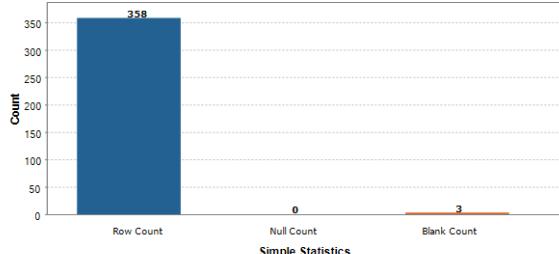
0.84 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.CompostSystem

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



CompostTumblers

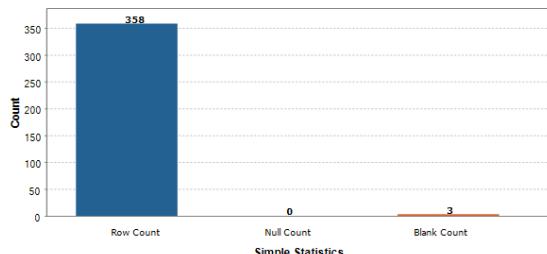
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.CompostTumblers

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



NonFoodPlants

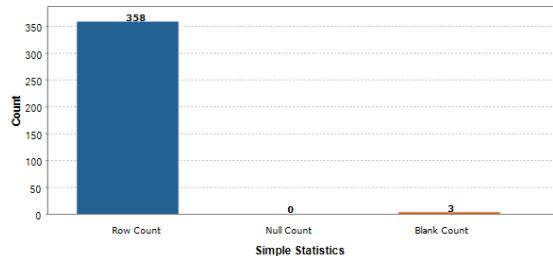
0.84 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.NonFoodPlants

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |
| | | |
| | | |
| | | |
| | | |
| | | |



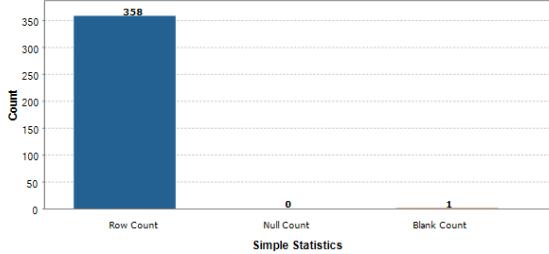
Food

0.28 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.Food

▼ Simple Statistics



OpenLawnOrCommunalArea

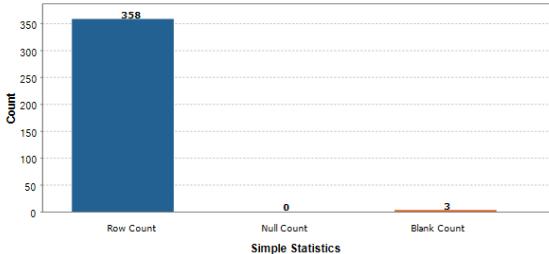
0.84 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.OpenLawnOrCommunalArea

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |
| | | |
| | | |
| | | |
| | | |
| | | |



PavedArea

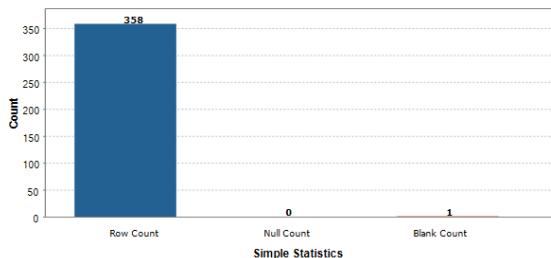
0.28 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.PavedArea  

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



TreesInGarden

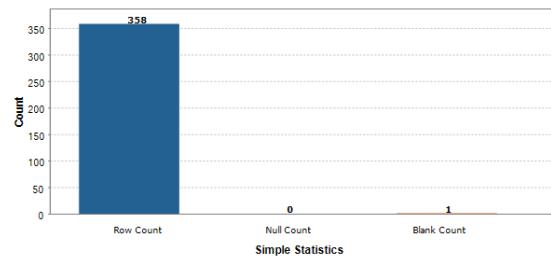
0.28 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.TreesInGarden  

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



FruitTrees

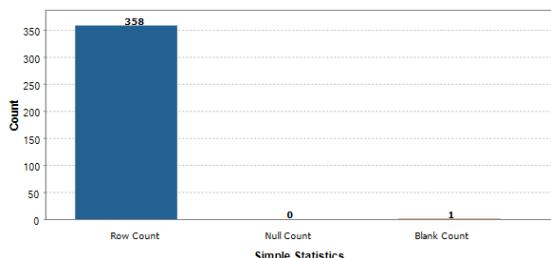
0.28 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.FruitTrees  

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



StreetTrees

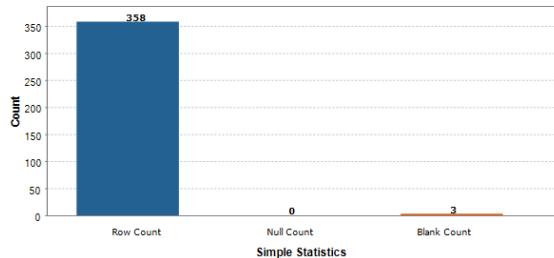
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.StreetTrees

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



EmptyTreePits

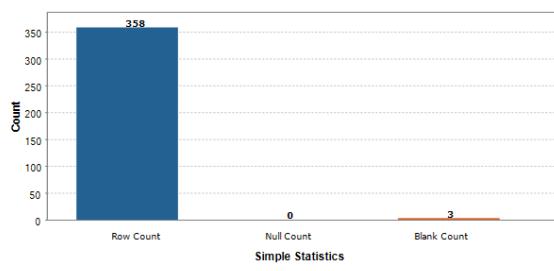
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.EmptyTreePits

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



Murals

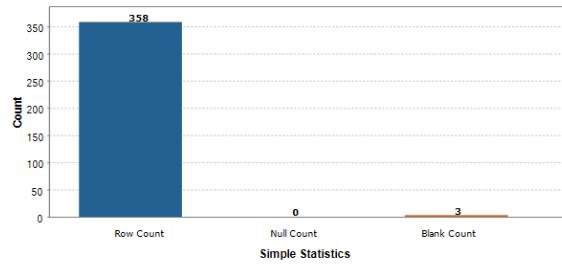
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Murals

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



BlankShed

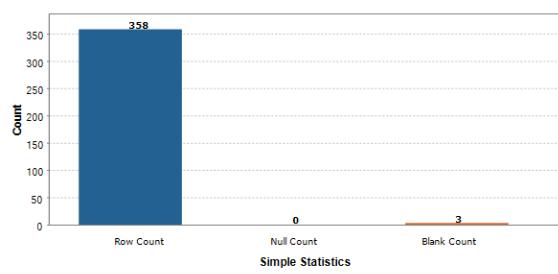
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.BlankShed

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



ParksSign

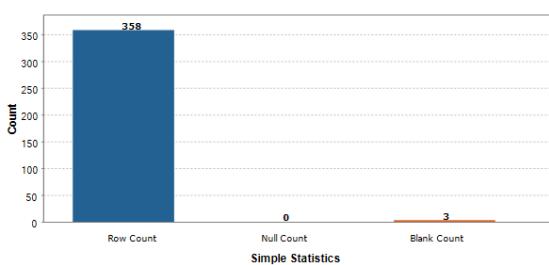
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.ParksSign

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |
| | | |
| | | |



Chickens

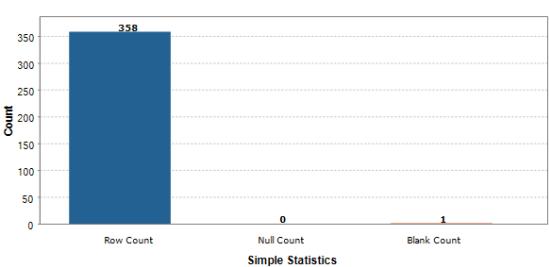
0.28 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Chickens

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |
| | | |
| | | |



Pond

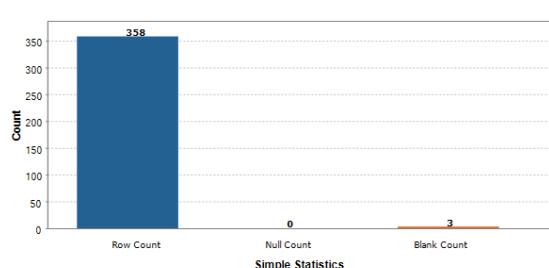
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Pond

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |
| | | |
| | | |



FishInPond

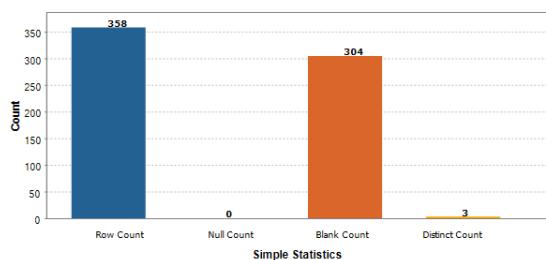
84.92 % dos valores estão em branco.

Esta coluna será removida, pois maioria dos dados estão em branco.

Column: metadata.FishInPond

Simple Statistics

| Label | Count | % |
|----------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 304 | 84.92% |
| Distinct Count | 3 | 0.84% |



Turtles

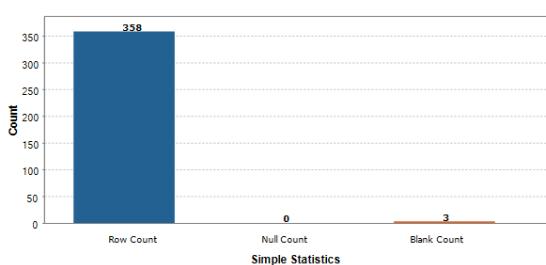
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Turtles

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



Aquaponics

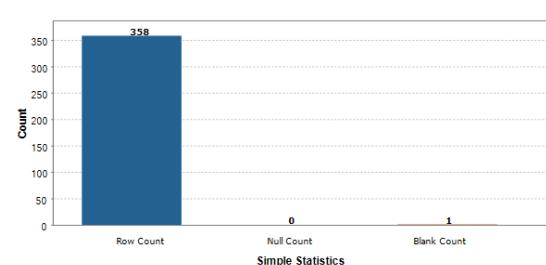
0.28 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Aquaponics

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



FarmersMarket

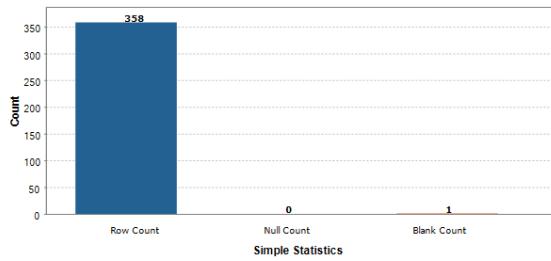
0.28 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.FarmersMarket

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



CSApickup

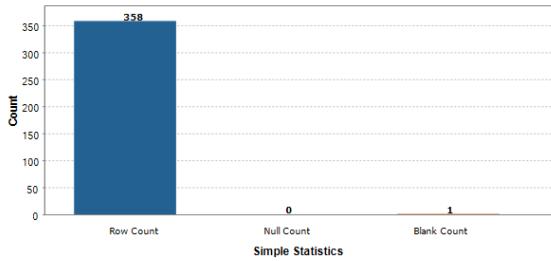
0.28 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.CSApickup

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 1 | 0.28% |



Composting

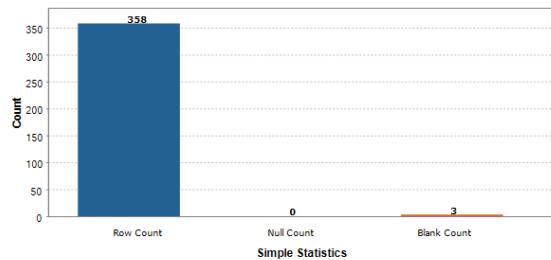
0.84 % dos valores estão em branco.

Esta coluna será mantida.

Column: metadata.Composting

Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



Greenhouse

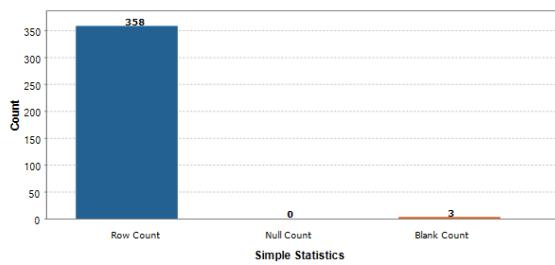
0.84 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.Greenhouse

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.84% |



StructureForSeasonExtension

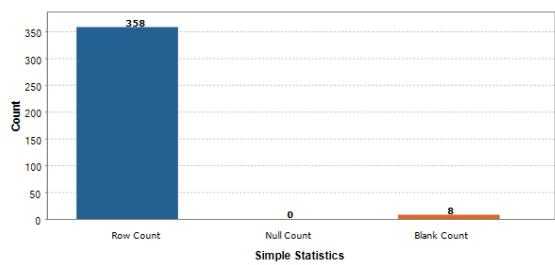
2.23 % dos valores estão em branco.

Esta coluna será mantida.

▼ Column: metadata.StructureForSeasonExtension

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 358 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 8 | 2.23% |



Dicionário

| Coluna | Tipo | Descrição |
|-----------------------|-------------|--|
| SYSTEM | String | Identificador Único de cada Greenstreet |
| GISPROPNUM | String | Identificador único para a propriedade do Parque na qual a GreenStreet esta localizada |
| OMPPROPID | String | Identificador único para o programa de inspeção do parque |
| DEPARTMENT | String | Distrito de NYC na qual a Greenstreet esta localizada |
| PARENTID | String | ID do parente a que pertence (Caso aplicável) |
| LOCATION | String | Descrição da localização |
| SITENAME | String | Nome de cada Greenstreet |
| DESCRIPTION | String | Descrição de cada greenstreet |
| BOROUGH | Char | Bairro em que a Greenstreet se situa |
| ACRES | Float | Área em Acres (Unidade Imperial) |
| COMMISSIONDATE | Date | Data em que o Departamento do Parque tomou responsabilidade pela funcionalidade |
| GSGROUP | Integer | ID da Greenstreet group a que pertence |
| GSTYPE | String | Tipo de Greenstreet |
| MOU | String | Tipo de Memorandum of Understanding |
| SUBCATEGORY | String | Categoria atribuída pelo Programa de Inspeção do parque |
| US_CONGRESS | Integer | Distrito Congressional em que a Greenstreet se situa |
| NYS_ASSEMBLY | String | Distrito de Assembleia de NY pela qual a Greenstreet e abrangida |
| NYS_SENATE | Integer | Distrito do Senado de NY pela qual a Greenstreet e abrangida |

| | | |
|------------------------|---------|--|
| COMMUNITYBOARD | Integer | Grupo de comunidade em que a Greenstreet se situa |
| COUNCILDISTRICT | Integer | Distrito de Concelho de NY pela qual a Greenstreet é abrangida |
| PRECINCT | Integer | Esquadra da NYPD que abrange a Greenstreet |
| ZIPCODE | Integer | Código postal da Greenstreet |
| FEATURESTATUS | String | Indica o estado desta funcionalidade em relação a Greenstreet |
| STArea | Float | Área do polígono em questão |
| STLength | Float | Perímetro do polígono em questão |
| multipolygon | String | Tipo de polígono em questão |

Analise do dataset

SYSTEM

Esta coluna não apresenta dados em branco, nulos ou duplicados

GISPROPNUM

Esta coluna não apresenta dados em branco ou nulos, porém apresenta duplicados que não serão considerados

▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |
| Duplicate Count | 46 | 1.67% |

OMPPROPID

Esta coluna não apresenta dados em branco, nulos ou duplicados

▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Duplicate Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

DEPARTMENT

Esta coluna não apresenta linhas em branco ou nulas

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Blank Count | 0 | 0.00% |
| Null Count | 0 | 0.00% |

PARENTID

Esta coluna não apresenta valores em branco ou nulos

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

LOCATION

Esta coluna não apresenta valores em branco ou nulos

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

SITENAME

Devido ao número elevado de entradas “Greenstreets”, esta tabela é considerada redundante e será descartada

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 151 | 5.48% |

▼ Value Frequency

| Value | Count | % |
|-----------------------------|-------|--------|
| GREENSTREET | 1286 | 46.71% |
| Greenstreet | 1108 | 40.25% |
| Empty field | 151 | 5.48% |
| Broadway Malls | 17 | 0.62% |
| Jackson Avenue Median | 8 | 0.29% |
| Flatbush Malls | 8 | 0.29% |
| Schiff Malls | 7 | 0.25% |
| Sherman Creek | 5 | 0.18% |
| John P. Salogub Greenstreet | 5 | 0.18% |
| Triangle | 4 | 0.15% |

DESCRIPTION

Esta tabela contém um número elevado da mesma entrada “Greenstreets”, e por isso será também descartada

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

▼ Value Frequency

| Value | Count | % |
|-------------------------------|-------|--------|
| GREENSTREET | 1059 | 38.47% |
| Greenstreet | 871 | 31.64% |
| Greenstreet Inspection Group | 17 | 0.62% |
| Broadway Malls | 17 | 0.62% |
| 9th Avenue - 23rd st to 31 st | 15 | 0.54% |
| Park Circle Greenstreets | 12 | 0.44% |
| Greenwich Ave & 7th Ave | 12 | 0.44% |
| Mall Forty-two-Greenstreet | 11 | 0.40% |
| Jackson Avenue Medians | 8 | 0.29% |
| GREENSTREET-Flatbush Malls | 8 | 0.29% |

Borough

Valores que não sejam do tipo B, Q, R, X, M serão ignorados. Existe também uma variedade saudável de entradas de cada tipo

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 3 | 0.11% |

Value Frequency

| Value | Count | % |
|-------------|-------|--------|
| Q | 974 | 35.38% |
| B | 522 | 18.96% |
| X | 469 | 17.04% |
| M | 437 | 15.87% |
| R | 348 | 12.64% |
| Empty field | 3 | 0.11% |

ACRES

A tabela não apresenta valores nulos

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |

COMMISSIONDATE

Esta coluna apresenta alguns valores nulos, que serão descartados. Não foi possível fazer uma verificação do formato da data devido ao tipo de dados utilizado e aos métodos de averiguação pré-carregados no Talend

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 24 | 0.87% |

GSGROUP

Devido ao elevado número de valores nulos e relativa pouca importância, esta tabela será descartada

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 1138 | 41.34% |

GSTYPE

Apesar de haver muitos valores em branco e muitos “greenstreet”, ainda existe a oportunidade de distinguir os vários tipos de greenstreets

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 987 | 35.85% |

Value Frequency

| Value | Count | % |
|----------------|-------|--------|
| Greenstreet | 1485 | 53.94% |
| Empty field | 987 | 35.85% |
| Ped_Ref | 94 | 3.41% |
| SA_Tri | 65 | 2.36% |
| SW_Greenstreet | 62 | 2.25% |
| TreeOnlyMed | 60 | 2.18% |

MOU

Devido a sua pouca importância no contexto temático deste trabalho, e a maioria das entradas serem DOT, esta coluna será descartada

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Blank Count | 0 | 0.00% |

▼ Value Frequency

| Value | Count | % |
|-------|-------|--------|
| DOT | 1825 | 66.29% |
| None | 641 | 23.28% |
| Other | 286 | 10.39% |
| DEP | 1 | 0.04% |

SUBCATEGORY

Tal como em GSTYPE, embora haja numericamente pouca variedade de dados, estes são pertinentes para o dataset

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |

▼ Value Frequency

| Value | Count | % |
|-------------------------------|-------|--------|
| Type 1 | 1305 | 47.40% |
| Empty field | 835 | 30.33% |
| Type 2 | 563 | 20.45% |
| Type 1 Greenstreet | 18 | 0.65% |
| Sitting Area/Triangle/Mall | 17 | 0.62% |
| Type 1 Stormwater Greenstreet | 6 | 0.22% |
| Type 2 Greenstreet | 4 | 0.15% |
| Neighborhood Plgd | 3 | 0.11% |
| Pier | 1 | 0.04% |
| Greenstreet | 1 | 0.04% |

US CONGRESS

Estas entradas são consideradas de pouco relevantes no contexto da temática do projeto, e por isso serão descartadas

▼ Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 29 | 1.05% |
| Duplicate Count | 14 | 0.51% |

NYS ASSEMBLY

Esta coluna segue o mesmo princípio de US CONGRESS e por isso também não será tida em conta

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 7 | 0.25% |

NYS SENATE

A coluna também não tem importância

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 115 | 4.18% |

COMMUNITY BOARD

Esta coluna contém alguns valores nulos, que serão descartados.

▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 3 | 0.11% |

COUNCILSDISTRICT

Como colunas anteriores, sendo os dados de natureza burocrática não tem relevância para o tema, por isso serão ignorados

▼ Simple Statistics

| Label | Count | % |
|--------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Unique Count | 3 | 0.11% |
| Null Count | 46 | 1.67% |

PRECINCT

▼ Simple Statistics

| Label | Count | % |
|--------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Unique Count | 0 | 0.00% |

▼ Value Frequency

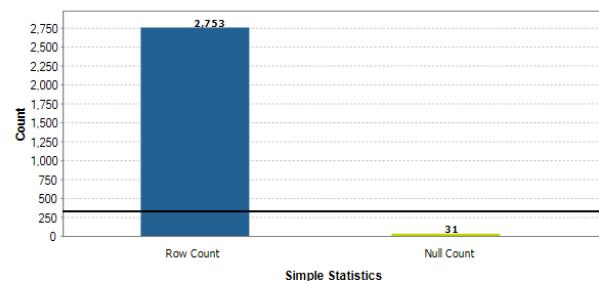
| Value | Count | % |
|-------|-------|-------|
| 122 | 214 | 7.77% |
| 105 | 157 | 5.70% |
| 107 | 139 | 5.05% |
| 111 | 94 | 3.41% |
| 120 | 93 | 3.38% |
| 45 | 90 | 3.27% |
| 109 | 76 | 2.76% |
| 108 | 74 | 2.69% |
| 100 | 71 | 2.58% |
| 115 | 69 | 2.51% |

ZIPCODE

Esta coluna contém valores nulos e valores que não considerados validos pelo Pattern Matching, serão então removidos.

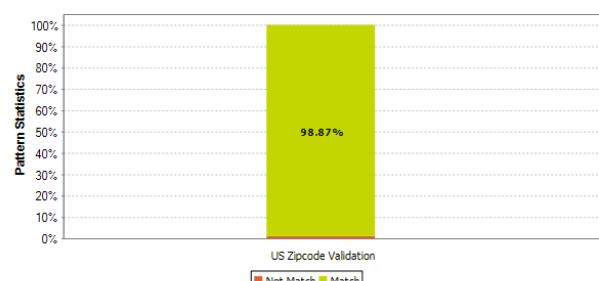
▼ Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 31 | 1.13% |



▼ Pattern Matching

| Label | Match% | Not Mat... | Match | Not Match |
|-----------------------|--------|------------|-------|-----------|
| US Zipcode Validation | 98.87% | 1.13% | 2722 | 31 |



FEATURESTATUS

Esta tabela não apresenta quaisquer valores inválidos

▼ Simple Statistics

| Label | Count | % |
|-------------|-------|---------|
| Row Count | 2753 | 100.00% |
| Null Count | 0 | 0.00% |
| Blank Count | 0 | 0.00% |

▼ Value Frequency

| Value | Count | % |
|----------|-------|--------|
| Active | 2208 | 80.20% |
| Inactive | 545 | 19.80% |

STArea, STLength e multypolygon

Estas colunas servem como referências para o programa de visualização no mapa, e como tal serão mantidas na integra

BlockLot

Dicionário dataset

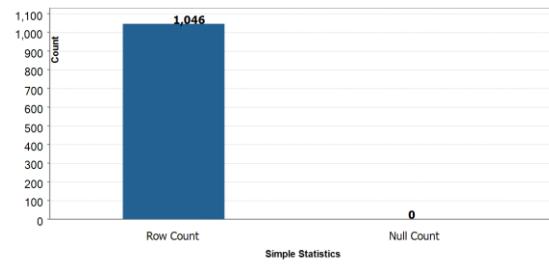
| Coluna | Tipo | Descrição |
|-------------|---------|--|
| Parksid | String | Número de identificação de cada parque |
| Block | Integer | Bairro onde se situa o local |
| Lotnum | Integer | Número do lote do local |
| Lotsize | String | Tamanho do lote |
| Areacovered | String | Área preenchida do lote |

Análise dataset

Coluna block: esta coluna não apresenta dados a null, e são constituídos por conjuntos de dígitos representados por “9”, tal como se pode observar na tabela “Pattern Frequency”.

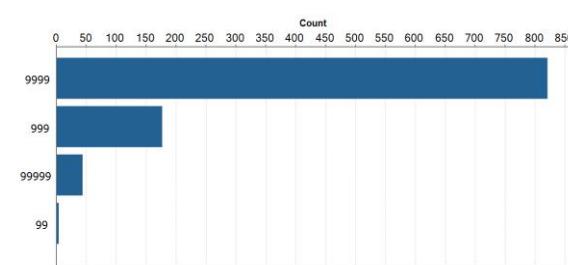
Simple Statistics

| Label | Count | % |
|------------|-------|---------|
| Row Count | 1046 | 100.00% |
| Null Count | 0 | 0.00% |

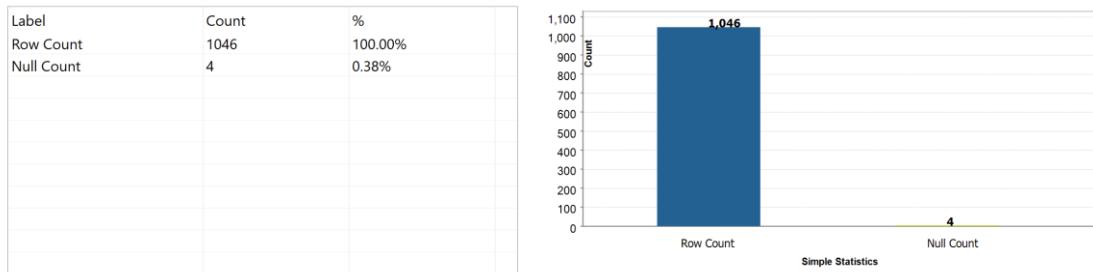


Pattern Frequency

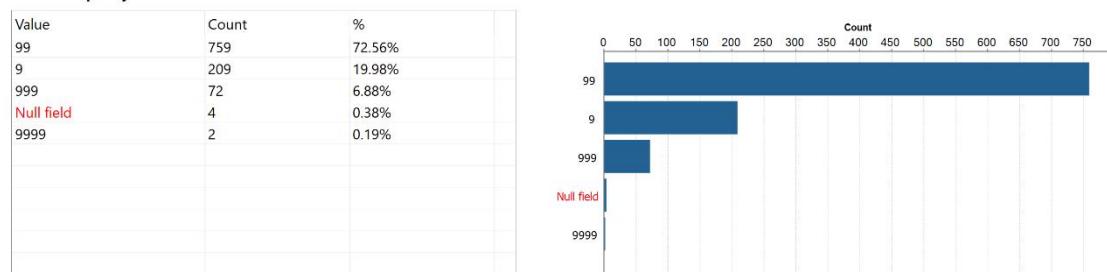
| Value | Count | % |
|-------|-------|--------|
| 9999 | 821 | 78.49% |
| 999 | 177 | 16.92% |
| 99999 | 44 | 4.21% |
| 99 | 4 | 0.38% |



Coluna lotnum: nesta coluna apenas 0.38% dos dados estão a null e os restantes são constituídos por um conjunto de 1, 2, 3 ou 4 dígitos representados por “9”.

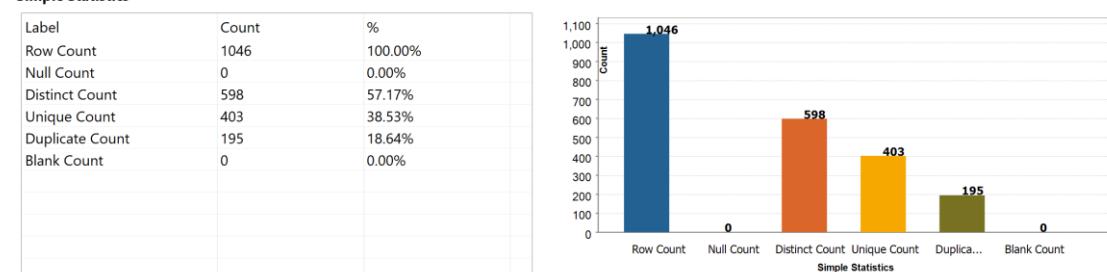


▼ Pattern Frequency

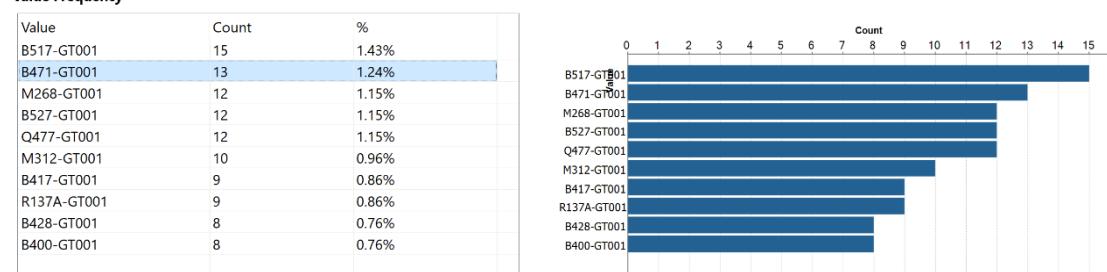


Coluna parksid: esta coluna não apresenta dados a null/blank e os seus dados são constituídos por um conjunto de letras e números que servem de identificador de cada um dos espaços.

▼ Simple Statistics



▼ Value Frequency



Potencial

Dicionário Dataset

| Coluna | Tipo | Descrição |
|---------------------------|---------|--|
| Data Created | String | Data em que o dataset foi atualizado |
| Borough | Integer | Bairro onde se situa o local |
| Block | Integer | Número do bloco do local |
| Lot | Integer | Número do lote do local |
| Address | String | Morada do local |
| Parcel Name | String | Nome do Local |
| Agency | String | Agencia à qual está associado |
| Total Area | Integer | Área do local em metros quadrados |
| Community Board | Integer | NYC Community Board a que pertence |
| Council District | Integer | NYC Council District a que pertence |
| Coordinates | String | Coordenadas da localização aproximada |
| Potential Urban Ag | String | O potencial para agricultura urbana |
| Latitude | Integer | Latitude |
| Longitude | Integer | Longitude |
| BIN | Integer | ID do edifício localizado no local, caso haja um |
| NTA | String | Neighborhood Tabulation Area do local onde está localizado |

As colunas *Data Created*, *Latitude*, *Longitude*, *Address*, *Parcel Name*, *Agency*, *Coordinates*, *Postcode*, *BIN*, *NTA* serão descartadas, uma vez que não possuem informação pertinente ou dados suficientes.

Analise do Dataset

Borough

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 4 | 13.79% |
| Unique Count | 0 | 0.00% |
| Duplicate Count | 4 | 13.79% |

▼ Value Frequency

| Value | Count | % |
|-------|-------|--------|
| 4 | 19 | 65.52% |
| 5 | 4 | 13.79% |
| 2 | 3 | 10.34% |
| 3 | 3 | 10.34% |

Block

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 21 | 72.41% |
| Unique Count | 16 | 55.17% |
| Duplicate Count | 5 | 17.24% |

Lot

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 27 | 93.10% |
| Unique Count | 25 | 86.21% |
| Duplicate Count | 2 | 6.90% |

Total Area

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 28 | 96.55% |
| Unique Count | 27 | 93.10% |
| Duplicate Count | 1 | 3.45% |

Community Board

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 10 | 34.48% |
| Unique Count | 4 | 13.79% |
| Duplicate Count | 6 | 20.69% |

Value Frequency

| Value | Count | % |
|-------|-------|--------|
| 414 | 14 | 48.28% |
| 410 | 3 | 10.34% |
| 407 | 2 | 6.90% |
| 503 | 2 | 6.90% |
| 201 | 2 | 6.90% |
| 316 | 2 | 6.90% |
| 501 | 1 | 3.45% |
| 502 | 1 | 3.45% |
| 313 | 1 | 3.45% |
| 205 | 1 | 3.45% |

Council District

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 1 | 3.45% |
| Distinct Count | 12 | 41.38% |
| Unique Count | 7 | 24.14% |
| Duplicate Count | 4 | 13.79% |

| Value | Count | % |
|-------|-------|--------|
| 31 | 14 | 48.28% |
| 32 | 3 | 10.34% |
| 17 | 2 | 6.90% |
| 51 | 2 | 6.90% |
| 19 | 1 | 3.45% |
| 24 | 1 | 3.45% |
| 41 | 1 | 3.45% |
| 42 | 1 | 3.45% |
| 14 | 1 | 3.45% |
| 47 | 1 | 3.45% |

Coordinates

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 7 | 24.14% |
| Unique Count | 6 | 20.69% |
| Duplicate Count | 1 | 3.45% |
| Blank Count | 0 | 0.00% |

Potencial Urban Ag

Simple Statistics

| Label | Count | % |
|-----------------|-------|---------|
| Row Count | 29 | 100.00% |
| Null Count | 0 | 0.00% |
| Distinct Count | 2 | 6.90% |
| Unique Count | 0 | 0.00% |
| Duplicate Count | 2 | 6.90% |
| Blank Count | 0 | 0.00% |

Transformações em Silver

2015 Street Tree Census

Neste dataset foram corrigidos alguns problemas de qualidade de dados como nulls, strings vazias e Nones (As linhas que continham estes erros passaram a ter “Unknown”). Depois a coluna “Year” que apresentava a data em que a árvore foi catalogada foi alterada para apresentar apenas o ano 2015. Para finalizar as transformações deste dataset foram criadas as colunas “problems_canopy_general”, “problems_guard_general”, “problems_wires_general”, “problems_paving_general”, “problems_outlet_general”, “problems_shoes_general”, “problems_lights_general”, “problems_trunk_general”. Como o dataset não apresentava informação sobre algumas colunas que foram criadas estas ficaram com todas as linhas “Unknown”.

2005 Street Tree Census

Neste dataset foram corrigidos alguns problemas de qualidade de dados como nulls, strings vazias, strings que não faziam sentido e nones (As linhas que continham estes erros passaram a ter “Unknown”). Depois a coluna “Year” que apresentava dois anos diferentes foi alterada para apresentar apenas o ano 2005. Na coluna boroname todas as linhas que apresentavam o número 5 foram alteradas para “Staten Island”. Para finalizar as transformações desta dataset foram criadas as restantes colunas que estão presentes no dataset árvores 20015 (com as linhas todas a “Unknown”) e as colunas “problems_trunk_general”, “health” e “sidewalk” tiveram algumas transformações para ficarem com a mesma estrutura do dataset arvores 20015.

1995 Street Tree Census

Neste dataset foram corrigidos alguns problemas de qualidade de dados como nulls, strings vazias, strings que não faziam sentido e nones (As linhas que continham estes erros passaram a ter “Unknown”). Depois as colunas “health” e “sidewalk” tiveram algumas transformações para ficarem com a mesma estrutura dos outros datasets. Para finalizar foram criadas as colunas presentes nos outros dataset com as linhas todas a “Unkown”.

Conclusão das transformações dos datasets das arvores:

Nestes datasets foram criadas e transformadas colunas com o objetivo de criar uma estrutura igual para cada um deles. Depois destas transformações fizemos um unio entre os 3 datasets e guardamos em silver na tabela arvores.delta.

[Infraestruturas Verdes](#)

Nesta fase de silver o primeiro passo foi selecionar as colunas do dataset que eu pretendia “aproveitar”, descartando assim, as que através da análise de qualidade efetuada no talend conclui serem pouco úteis para o objetivo pretendido.

Feita essa seleção segui para a fase de transformações das colunas tendo efetuado a substituição de nulls e blancks por “Não temos essa informação”. As outras alterações passaram basicamente por conversões de valores monetários que estavam em dólar para euro através de uma API ou medidas de comprimento, largura e área que estavam em pés e foram convertidas para metros através de operações de multiplicação.

No final das alterações a tabela foi guardada na bases de dados “Areas Verdes”, sendo o nome desta tabela “edificiosverdes”.

[Localização dos quintais](#)

Neste dataset comecei por criar a tabela e selecionar as colunas com que iria trabalhar. De seguida, substitui as linhas que não tinham informação com a palavra “Unknown”, nomeadamente nas colunas de “júris”, as colunas respondentes a cada dia da semana “openhrs” e a coluna “policeprecint”. Como os nomes dos bairros estavam identificados por apenas uma letra, alterei-os para os respetivos nomes. Por fim, guardei a tabela no HDFS.

[BlockLot](#)

Aqui criei novamente a tabela e selecionei as colunas que pretendia trabalhar. Na coluna “lotsize” preenchi as linhas que não tinha informação com o valor 0 (zero) e com “Unknown” as linhas da coluna “areacovered” que não tinha informação. Depois verifiquei se havia “parksi” duplicados e caso houvesse, removê-los. Finalmente guardei as colunas numa tabela no HDFS.

Potencial

Neste dataset voltei a criar a tabela e selecionar as colunas pretendidas. À semelhança do dataset “LocQuintais” os bairros estavam identificados por números e alterei para que se identificassem pelos respetivos nomes. As colunas “Latitude”, “Longitude”, “BIN”, “NTA” não tinham informação, ou eram muito poucas, então preenchi as linhas dessas colunas com “Unknown”. A última transformação deste dataset foi na coluna “Potencial_Urban_Ag” em que alterei a informação sobre o potencial conforme as informações que encontrei no dicionário deste dataset, ou seja, as linhas em que tínhamos a informação “Potential Suitable 1(...)” foram substituídas com “Most Potential” e as linhas com “Potential Suitable 3(...)” foram substituídas com “Least Potential”. Após estas transformações guardei a tabela no HDFS.

Informação do quintal

Na fase de silver, o primeiro passo feito, foi selecionar as colunas do dataset que eu pretendia utilizar, removendo assim, as colunas que não traziam relevância ou que tinham muito nulos/brancos.

Após a seleção das colunas, seguiu-se as transformações das colunas, tendo substituído os nulos e os brancos por 0 (em caso de Float e Double), e False (em caso de Boolean). Outras alterações realizadas foram a renomeação de colunas para exemplificar melhor a mudança da unidade de medida, e depois na coluna RainGallons (agora RainLitres), fez-se a conversão de galões para litros.

No final, estas alterações foram guardadas na base de dados “Areas Verdes”, sendo o nome desta tabela “InfQuintais” .

GreenStrets

Na etapa de Silver, pretende-se “limpar” as tabelas que se vão usar para construir o dataset. Para esse fim primeiro foram selecionadas as colunas consideradas pertinentes para o tema escolhido, com base no documento de análise de qualidade de dados.

Depois de esta seleção, na fase de transformações, as colunas em que apresentem dados em branco ou nulos foram substituídos por “Unknowns”, para os tipos String, e por “0” em tipos Int, tendo o cuidado de verificar se estes valores não interferem com outros valores presentes nas respetivas colunas. Colunas com valores em medida imperial foram convertidos para medida métrica, colunas que apresentem valores fora do esperável (por exemplo a coluna “Borough”) foram filtradas desses valores, colunas de valores numéricos com vírgulas foram transformados em valores

numéricos sem essas vírgulas e por fim colunas que representam IDs únicos foram filtradas de duplicados.

Finalmente, a tabela foi guardada na base de dados “Areas Verdes”, com o nome “GreenStreets” em formato delta, particionado por Borough, devido a contagem de linhas para cada valor possível estar relativamente bem distribuído, como comprovado na análise de qualidade de dados anteriormente feita.

Transformações em Gold

Tabelas das árvores

As transformações em Gold têm com objetivo responder às questões analíticas. Para isso, neste dataset foi criada uma coluna “problems_general” onde foram agrupados os problemas mais comuns das árvores. Depois foi realizado um COUNT com um groupBy (“year”, “boroname”, “problems_general”) que deu origem a tabela arvores_problems que permite responder à questão “Quais são os problemas mais comuns das árvores em Nova York”.

Para responder às restantes questões foram criadas mais 4 tabelas. Sendo elas arvores_year, arvores_health, arvores_sidewalk e arvores_species. Estas tabelas seguem a mesma estrutura de código, contudo não era possível realizar um join entre as mesmas uma vez que as informações eram diferentes.

Para finalizar as transformações de Gold para este dataset foram criadas tabelas em preto com as mesmas colunas, uma vez que o Tableau não consegue ler diretamente tabelas no formato delta.

Tabelas da localização dos quintais

Nesta fase, fiz o join entre os datasets LocQuintais e BlockLot e selecionei as tabelas pretendidas.

Para a primeira tabela comecei por calcular o número de parques e a média do lote por cada bairro e guardei a tabela em HDFS. Na segunda tabela, foi feito um count para analisar quantos jardins se enquadravam nos possíveis estados de atividade, estes também foram agrupados por bairro e estado de atividade. Na tabela “areacovered”, agrupou-se a informação por bairros e por área ocupada para averiguar quantos jardins é que tinham uma ocupação total, parcial ou se era desconhecida esta informação.

Posteriormente, para o dataset “Potential” foi criada uma tabela para análise da área total de cada tipo de potencial, ou seja, a área dos que tinham mais potencial e a área dos que tinham menos potencial.

Tabelas dos Edifícios Verdes

Nesta parte de gold o objetivo é fazer as transformações necessárias para responder às questões analíticas, que no meu caso, são relativas aos edifícios verdes.

Para isso, criei as tabelas numero_arvores_gold que faz basicamente um count de árvores por bairro, estado_construcao_gold que faz um count do estado de construção dos edifícios por bairro e por fim media_valor_gold onde faço um sum das áreas ocupadas por edifícios verdes por bairro e também através do avg o valor médio do lote por bairro.

No final das transformações e tabelas criadas, guardei as tabelas e criei também as tabelas presto para utilizar na fase final do projeto para construção das dashboards no Tableau.

Tabelas das informações dos quintais

Na fase de gold, é necessário fazer as transformações que sejam necessárias de modo a responder às questões analíticas, que neste caso, são a informação dos quintais.

Para podermos responder às questões analíticas, foram criadas as tabelas area_sidewalk, que faz um avg (média) da área dos passeios em bairros de New York, numero_plantas, que faz um count (soma) dos tipos de plantas (Both, Food, NonFood, No Plants) existentes em jardins de bairros de New York, avg_capacity_capturesystem, que faz um avg (média) da capacidade do sistema de captura de água em bairros e por fim numero_compostagem, que faz um count (soma) dos tipos de compostagem existentes (Bins, Barrels, Both, No Composting System), que depois permite saber a percentagem dos tipos de compostagem existentes em jardins de um bairro de New York.

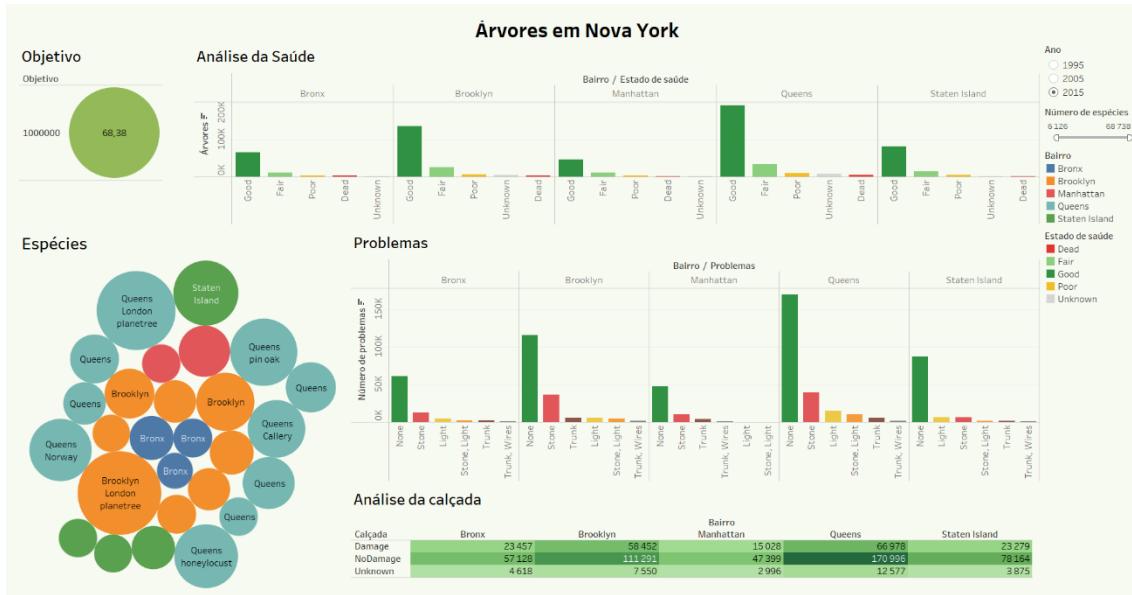
GreenStreets

Estas transformações são realizadas por forma a responder as questões analíticas e definir um dos KPIs. Para o efeito foram filtradas as linhas que tem o estado da funcionalidade como “ativo”, para apenas utilizar a informação que esta atualizada aquando da extração dos dados na camada bronze, foi criada uma coluna “Hectares_por_km2”, que associa a cada linha (Rua verde) um valor constituído pela sua área dividida pela área do bairro em que se localiza, proporcionando assim um valor de densidade que representaria a nossa Key Performance Indicator de hectares de rua verde por KM2 de Bairro e foram convertidos algumas colunas do tipo INT para String para facilitar a utilização do Tableau, como estes valores são apenas de identificação e não servem para qualquer efeito estatístico, podem ser convertidos sem afetar a veracidade dos dados. Finalmente selecionou-se apenas as colunas que permitirão responder às questões analíticas, por forma a reduzir tempo de processamento desnecessário. Infelizmente, foi necessário descartar uma a coluna COMMISSIONDATE, que incluiria informação temporal das ruas, devido a uma análise dos dados revelar que a esmagadora maioria dos campos tinham o mesmo valor, não permitindo tirar quaisquer conclusões relevantes.

Dashboards

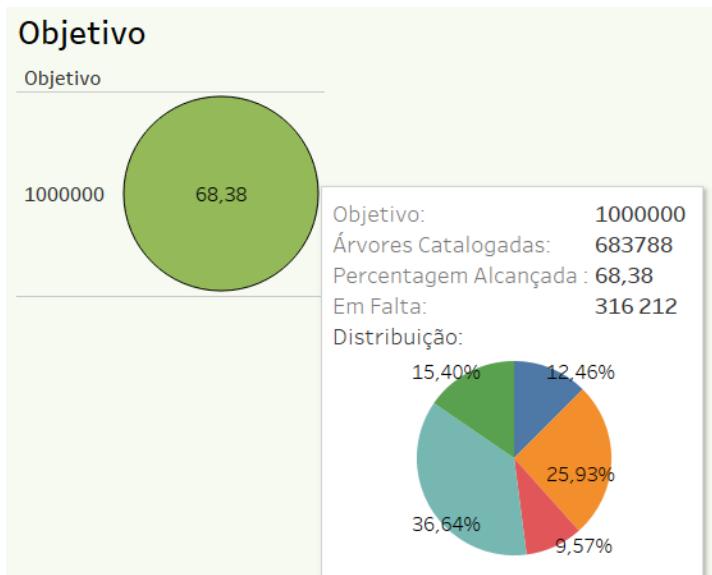
Árvores

Para responder às questões analíticas do dataset das árvores foi construída a seguinte dashboard:



Objetivo

O primeiro gráfico é alimentado pela tabela arvores_year.delta e mostra a percentagem alcançada em relação à nossa KPI de catalogar 1 milhão de árvores até 2015. Para além disso, ao passar o mouse por cima da percentagem alcançada aparece um segundo gráfico com a distribuição de árvores por bairro permitindo assim responder a questão analítica “Quais são os bairros que tem mais árvores?”.



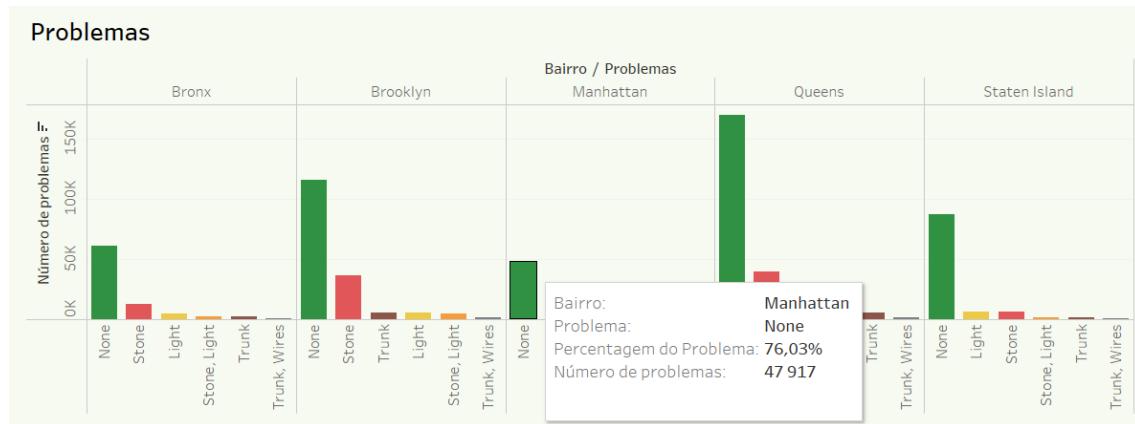
Análise da Saúde

O gráfico “Análise da Saúde” é alimentado pela tabela arvores_health.delta e permite responder à questão analítica “Em que estado de saúde se encontram as árvores?”. Ao passar o mouse por cima de alguma coluna deste gráfico aparece a percentagem de árvores nesse estado de saúde em relação ao total de árvores nesse bairro.



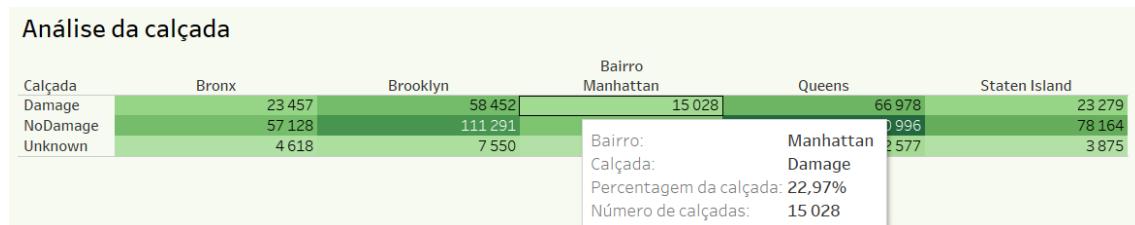
Problemas

O gráfico “Problemas” é alimentado pela tabela arvores_problems.delta e permite responder à questão analítica “Quais são os problemas que as árvores têm?”. Ao passar o mouse por cima de alguma coluna aparece a percentagem de árvores com problema em relação ao total de árvores nesse bairro.



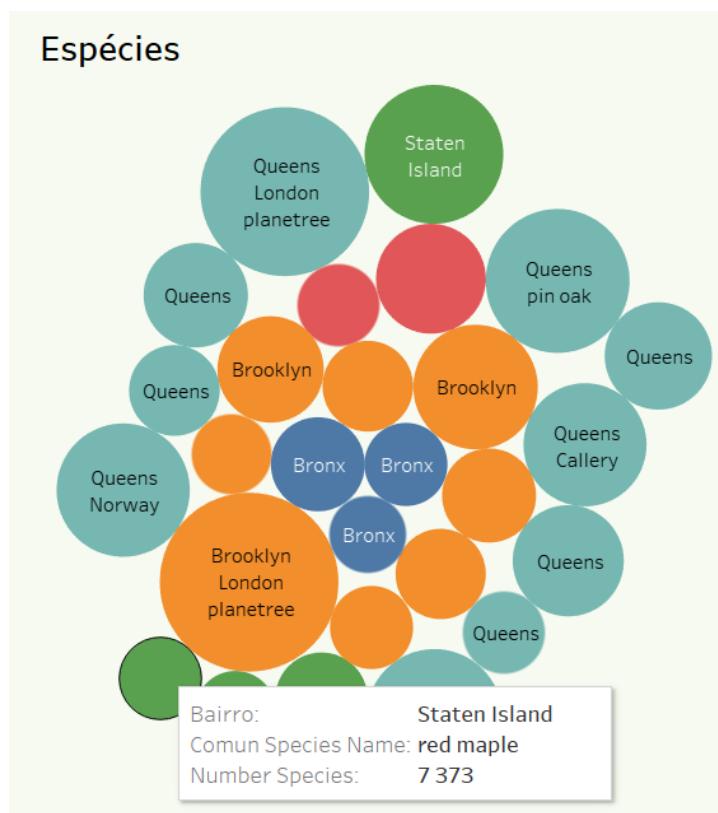
Análise da calçada

O gráfico “Analise da calçada” é alimentado pela tabela arvores_sidewalk.delta e permite responder à questão analítica “Em que estado se encontra o passeio em que as árvores estão plantadas?”. Ao passar o mouse por cima de alguma coluna aparece a percentagem de árvores com problema em relação ao total de árvores nesse bairro.



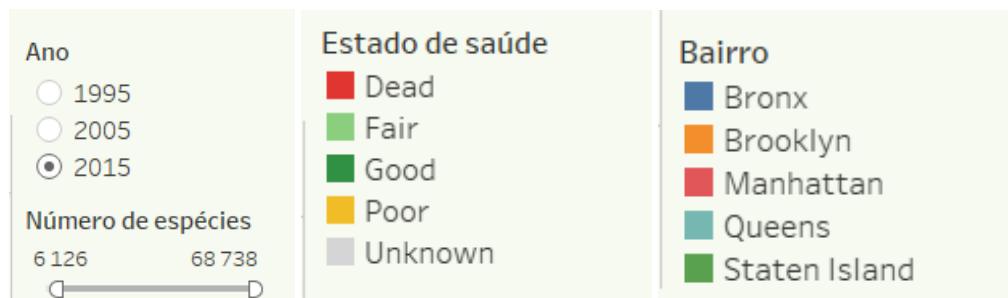
Espécies

O gráfico “Espécies” é alimentado pela tabela arvores_species.delta e permite responder à questão analítica “Quais são as espécies de árvores mais comuns em Nova York?”. Ao passar o mouse por cima de alguma coluna aparece o número de árvores dessa espécie nesse bairro.



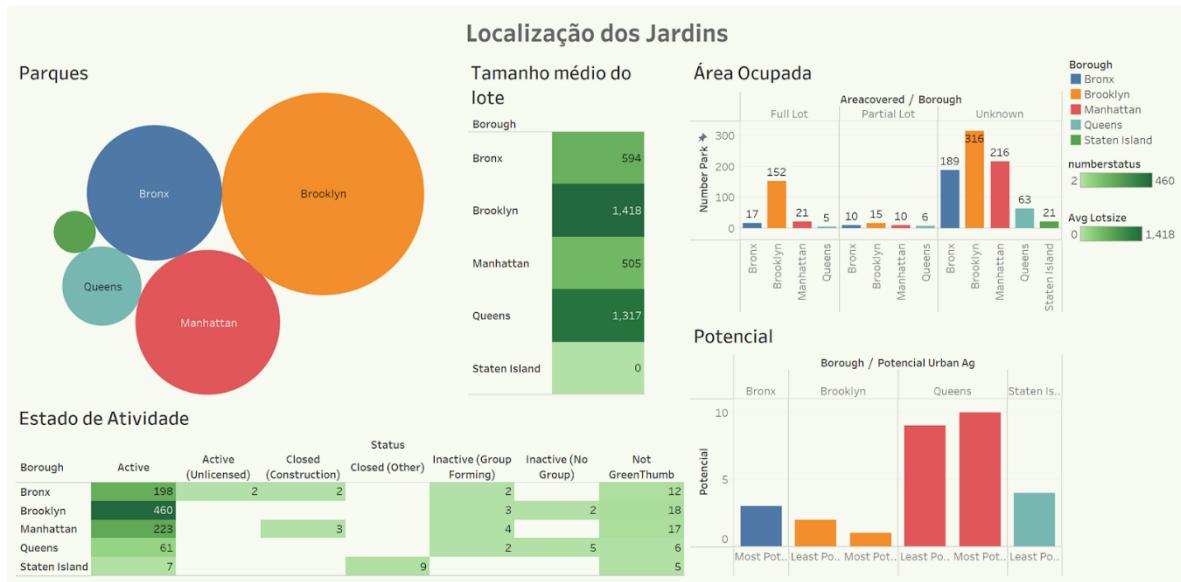
Filtros

Por fim, ainda estão disponíveis na dashboard um conjunto de filtros que permitem mudar a informação dos gráficos consoante o ano que se pretende avaliar, o número de árvores de cada espécie e o bairro que se pretende destacar.

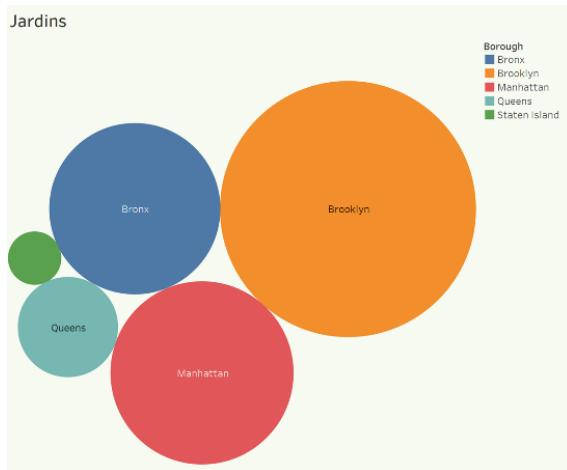


Localização dos Jardins

Dashboard completa



Jardins



Temos a distribuição dos jardins por cada bairro, conseguimos ver o número de jardins em cada e também a percentagem de cada um.

Brooklyn é o que possui mais jardins, com 483 jardins, 46.40%, seguido de Manhattan (247, 23.73%), Bronx (216, 20.75%), Queens (74, 7.11%) e por fim Staten Island com 21, 2.02%.

Tamanho médio do lote



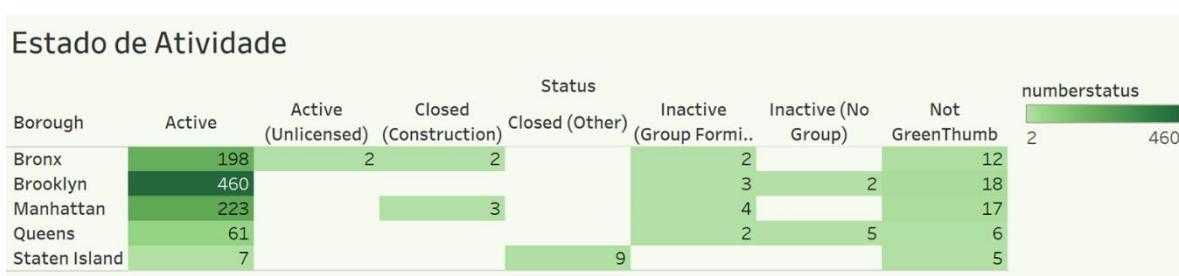
Em média, o bairro com a média mais alta é Brooklyn, com 1418, e de seguida, apesar de não ser o bairro com mais jardins, Queens com uma média de 1317, depois temos Bronx com 594 e Manhattan com 505. Para o bairro de Staten Island não possuímos dados que permitissem a determinação do tamanho médio dos jardins.

Área Ocupada



Neste caso faz se a referência a cada lote, se está totalmente ocupado ou apenas parcialmente. A maioria dos dados eram desconhecidos, mas pelo que conseguimos perceber, entre todos os bairros, exceto Staten Island, uma vez que não há informação disponível para os seus 21 jardins, 195 deles estão completamente ocupados e 41 parcialmente.

Estado de Atividade



Os vários jardins foram categorizados consoante o seu estado de atividade sendo os possíveis estados os seguintes:

Ativo, com (949) ou sem licença (2) – o jardim está aberto e há um grupo responsável pela sua manutenção

Fechado para construção (5) – o jardim encontra-se fechado temporariamente para algum tipo de construção

Fechado por outros motivos (9) – o jardim está fechado temporariamente, sem qualquer atividade ou construção planeada, mas devido a outros problemas que impede o jardim de estar operacional.

Inativo (Group Forming) (11) – o jardim está temporariamente fechado, porém um grupo com o qual os coordenadores GreenThumb estão a trabalhar para o desenvolvimento do jardim, definição de estatutos e registo de materiais.

Inativo (No Group) (7) – o jardim está temporariamente fechado e os coordenadores GreenThumb estão a trabalhar no sentido de formar um grupo para reativar o jardim.

Not GreenThumb (58) – estes jardins não se encontram associados com a GreenThumb

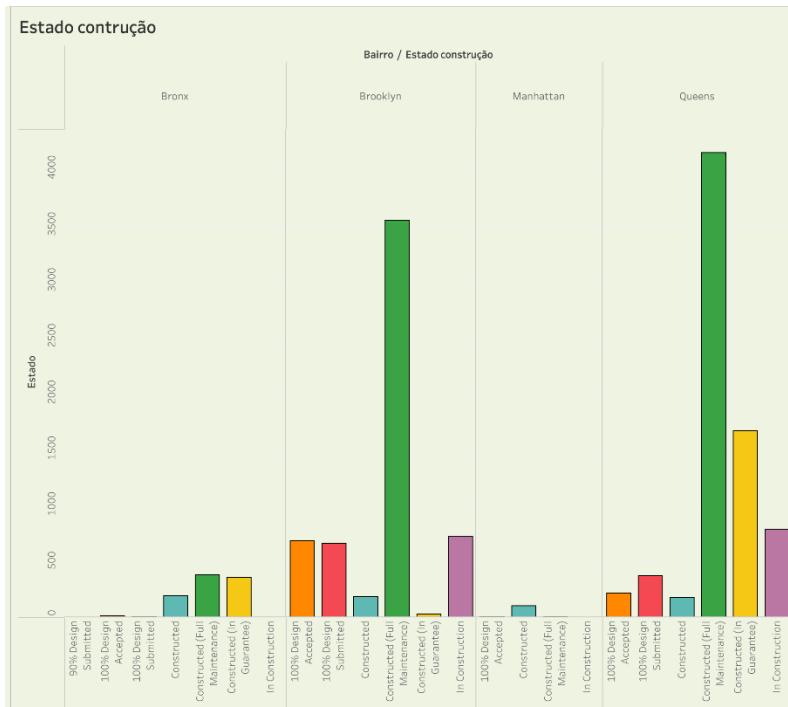
Edifícios verdes

Esta é a junção de todas as dashboards que eu fiz e das quais vou passar a explicar individualmente.



Estado construção

Esta dashboard mostra-nos quantos edifícios existem para cada estado de construção por cada bairro.



Número árvores

Esta dashboard mostra-nos o número de árvores por espécie que se encontram nos edifícios verdes de cada bairro.

Número de árvores

| Bairro | Tree Comm | | | | | | | | | | | |
|-----------|----------------|-------------|------------------|-------|----------------|---------|-----------|--------------|---------------------|-----------------|-----------|--------------------|
| | Eastern Redbud | Ginkgo Tree | Japanese Zelkova | N/A | Não temos in.. | Pin Oak | Red Maple | Sawtooth Oak | Shadblow Serviceb.. | Swamp White O.. | Sweetgu.. | Thornless Common.. |
| Bronx | | | | 502 | 279 | | | | | | | |
| Brooklyn | | | | 3 795 | 652 | 63 | | | 59 | 192 | 78 | 82 |
| Manhattan | | | | | 106 | | | | | | | |
| Queens | 83 | 56 | 101 | 5 003 | 378 | 136 | 58 | 55 | | 133 | 75 | 60 |

Valor médio por lote

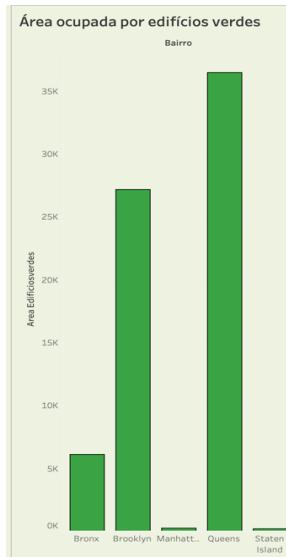
Esta dashboard mostra-nos o valor médio do lote onde se encontram os edifícios verdes ou do lote imediatamente à frente do mesmo, sendo esta média feito por bairro.

Valor médio do lote

| Bairro | Valor médio do lote |
|---------------|---------------------|
| Bronx | 1 962 343 279 |
| Brooklyn | 2 905 187 602 |
| Manhattan | 971 176 293 |
| Queens | 3 900 389 026 |
| Staten Island | 4 786 007 918 |

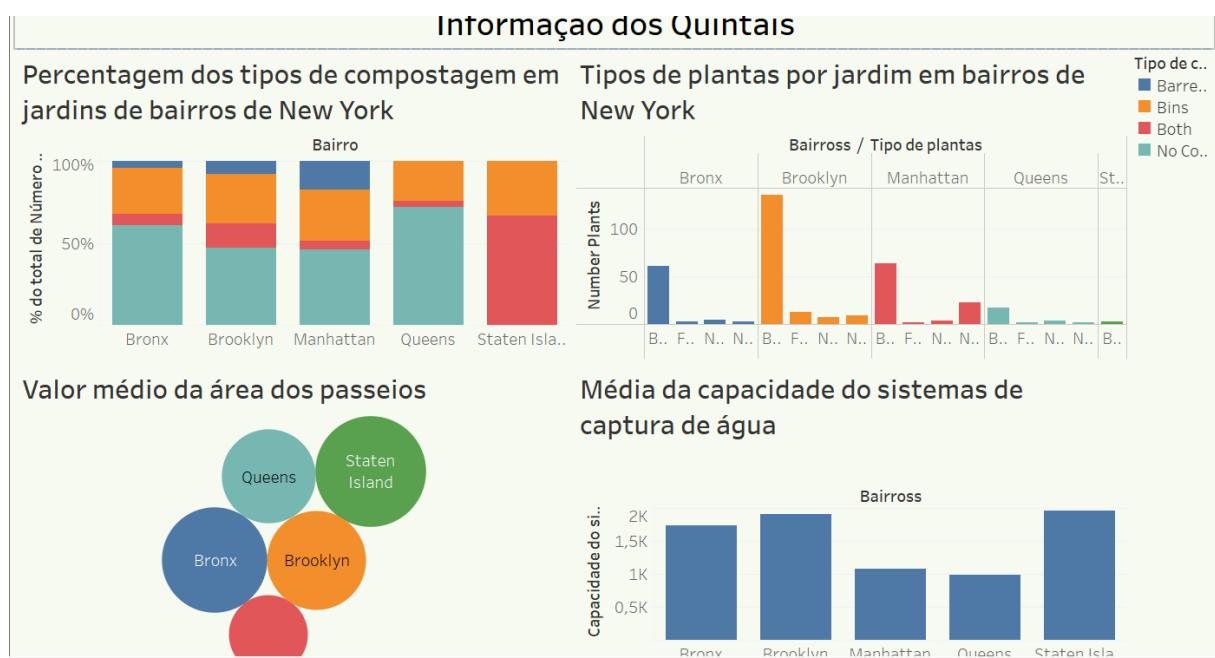
Área ocupada por edifícios verdes

Esta dashboard mostra-nos a área ocupada por edifícios verdes por cada bairro, o que nos pode levar à conclusão de que bairros estão mais ou menos explorados nesse sentido.



Informação dos quintais

O resultado da dashboard foi este, em que vai ser explicado cada uma individualmente.



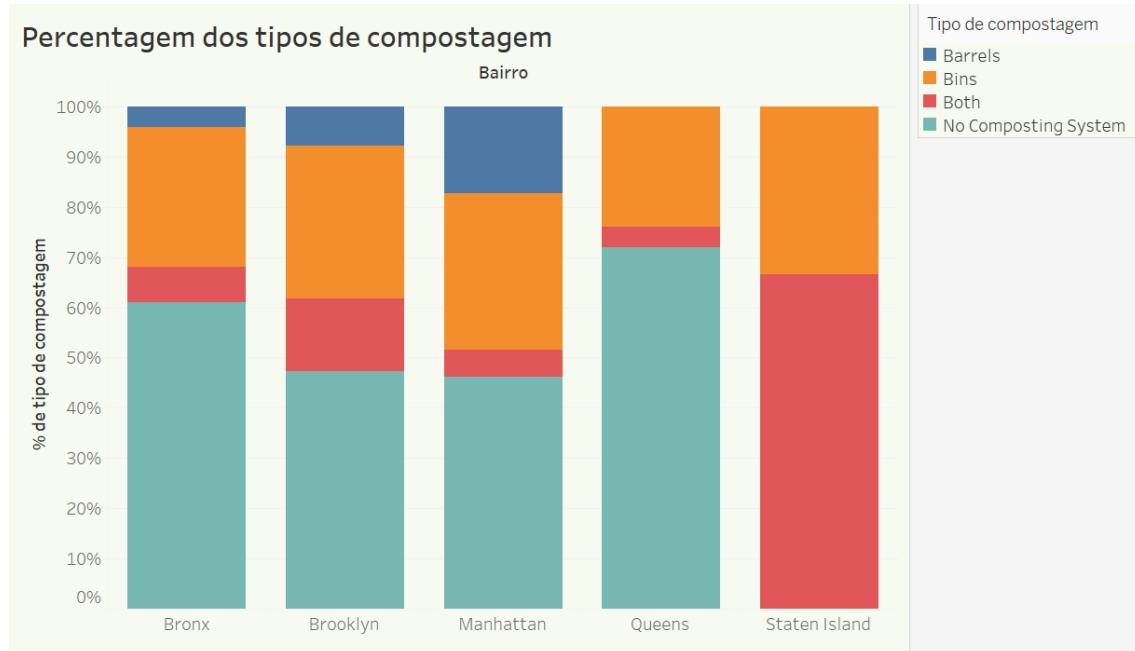
Tipos de plantas

Esta dashboard mostra a quantidade de tipos de plantas existentes em jardins por cada bairro.



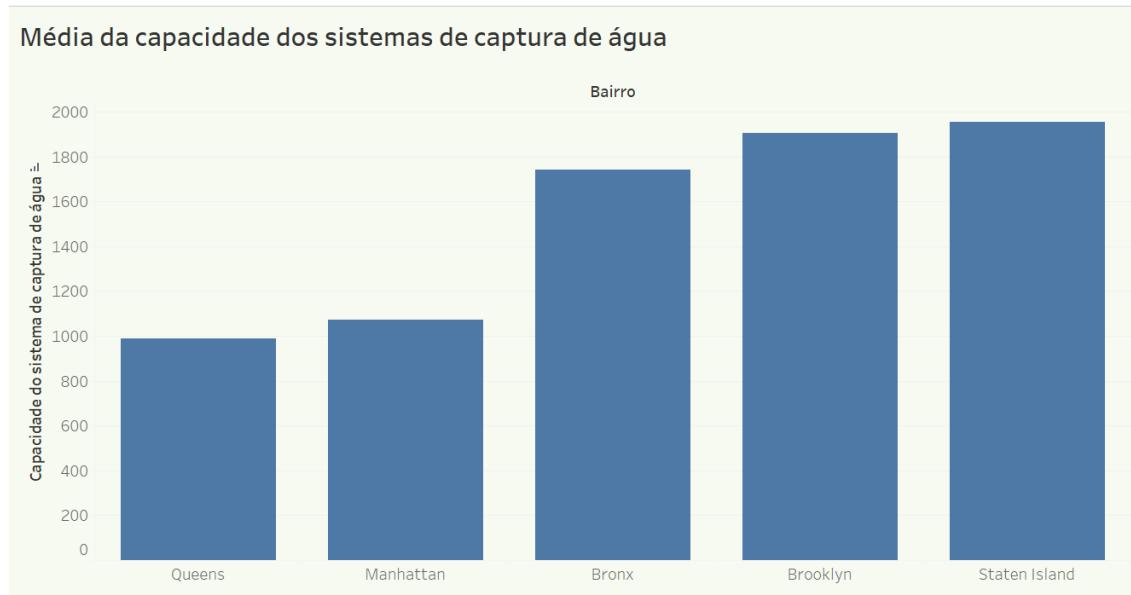
Tipos de compostagem

Esta dashboard mostra a percentagem dos tipos de compostagem existentes por cada bairro. (Aproximadamente 60% dos jardins de Bronx não tem nenhum tipo de compostagem).



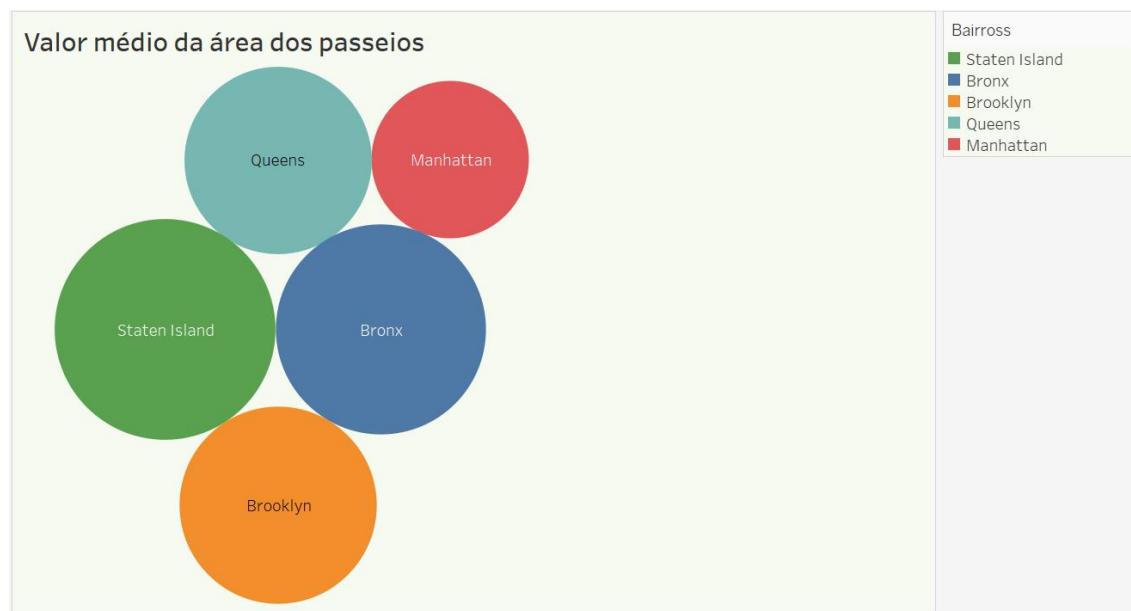
Capacidade dos sistemas de captura de água

Esta dashboard mostra a média da capacidade dos sistemas de captura de água por cada bairro, o que permite observar a performance dos sistemas de captura de água por cada bairro.

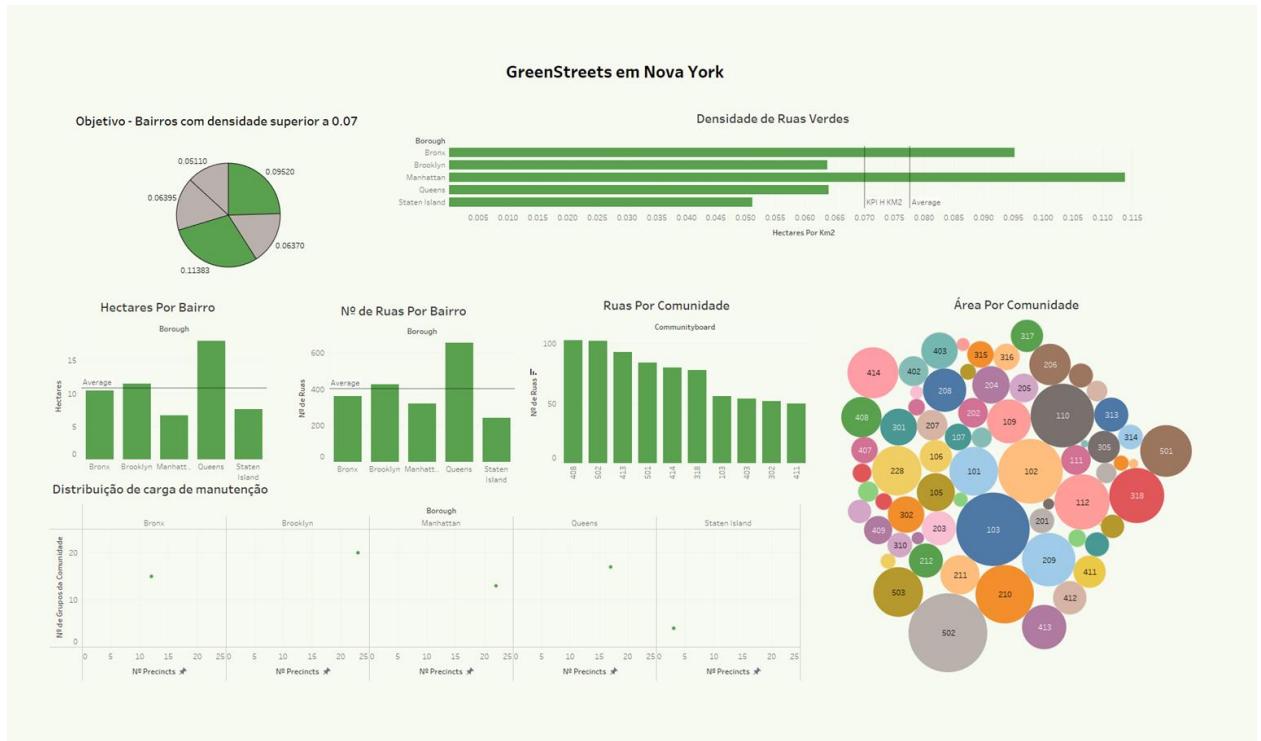


Área dos passeios

Esta dashboard mostra a média da área dos passeios por cada bairro, o que pode ajudar a construir novos passeios para bairros com poucos passeios (pouca área média).

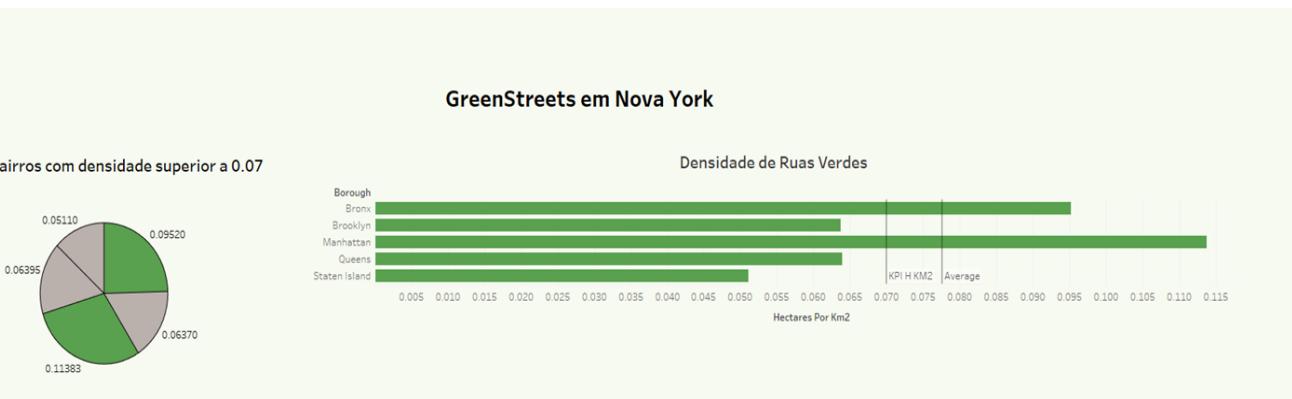


GreenStreets



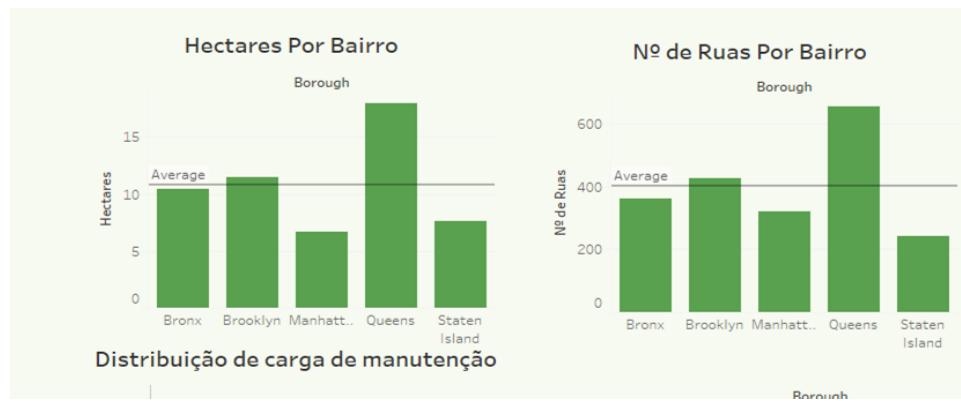
Com os objetivos e informação definida anteriormente foi possível construir a dashboard acima dividida em quatro secções principais

Secção da KPI



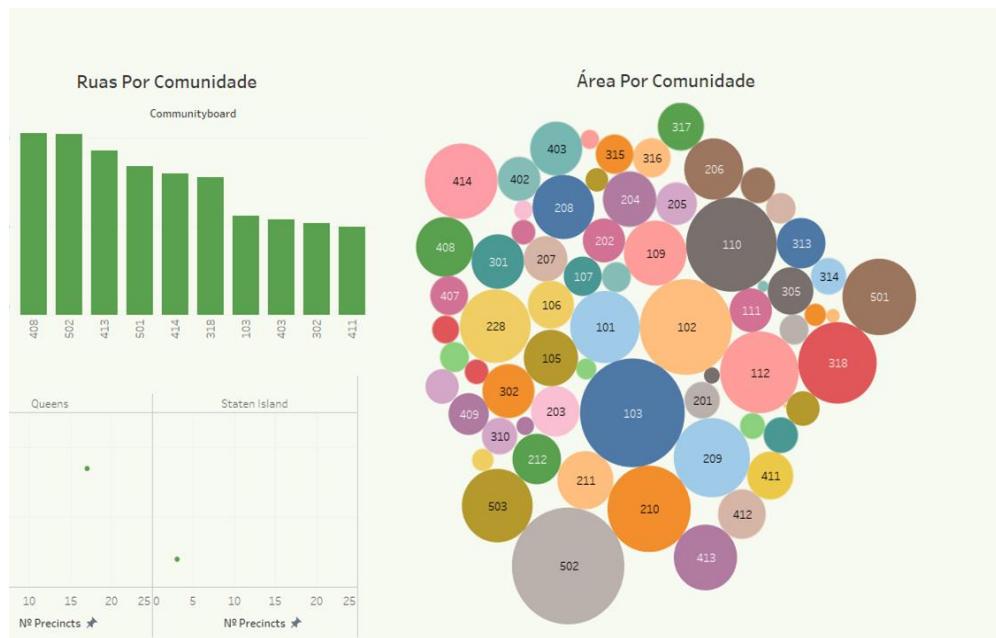
Esta secção, representando a nossa KPI mostra rapidamente que bairros cumprem o objetivo de 0.07 Hectares por KM2 no lado esquerdo, e contam através da nova tabela Hectares_por_km2 os hectares de rua por km2 de bairro no lado direito, respondendo à pergunta “Qual é o bairro que apresenta maior densidade de área de rua convertida por área total?”

Secção de Rua

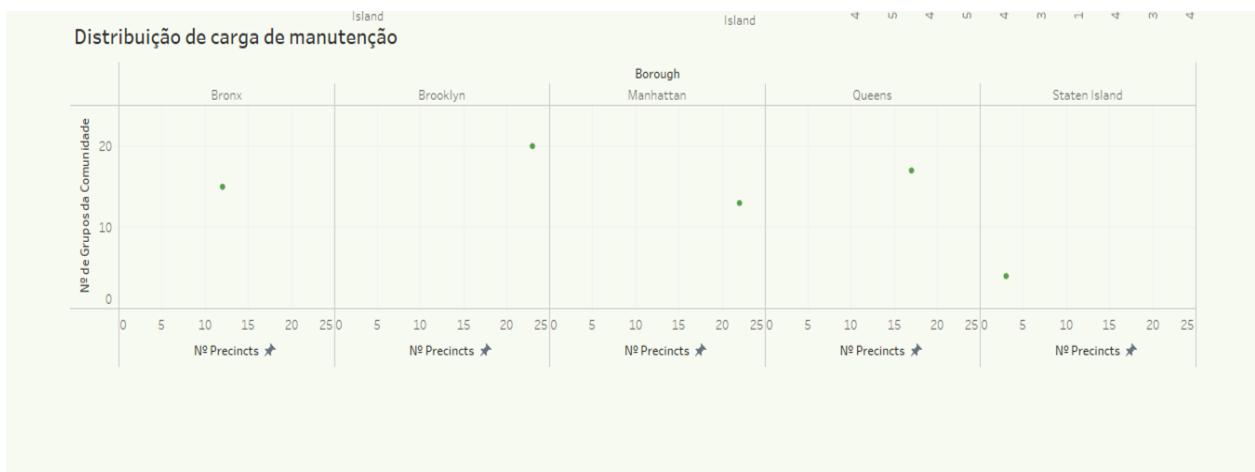


Estes gráficos mostram informação diversa sobre as Ruas, respondendo às questões “Qual é o bairro com maior número de ruas convertidas?” e “Qual é o bairro que tem mais área total convertida?”

Secção Comunidade



Secção Manutenção



Esta secção permite responder a última questão analítica pertinente a este dataset, “Existe um desequilíbrio entre Número de precincts e grupos comunitários a manter greenstreets?”. Este gráfico demonstra o equilíbrio entre número de Precincts e número de grupos comunitários a manter as greenstreets, sendo que quanto mais perto o ponto se encontra da linha diagonal de cada gráfico de cada bairro, mais equilibrado esta a distribuição de carga de manutenção.

Anexos

Vídeo da apresentação:

https://www.youtube.com/watch?v=O_jevp1dF_E&ab_channel=LeonorMachado