DOB Permit Issuance

```
%sh
python -m pip install socrata-py
python -m pip install openclean
python -m pip install openclean-geo
python -m pip install geopy
```

FINISHED

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: socrata-py in ./.local/lib/python3.7/site-packages (1.0.10)
Requirement already satisfied: requests in /share/apps/peel/python/3.7.9/gcc/lib/python3.7/site-packages (fro
m socrata-py) (2.25.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /share/apps/peel/python/3.7.9/gcc/lib/python3.7/site-
packages (from requests->socrata-py) (1.26.4)
Requirement already satisfied: idna<3,>=2.5 in /share/apps/peel/python/3.7.9/gcc/lib/python3.7/site-packages
(from requests->socrata-py) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /share/apps/peel/python/3.7.9/gcc/lib/python3.7/site-pac
kages (from requests->socrata-py) (2020.12.5)
Requirement already satisfied: chardet<5,>=3.0.2 in /share/apps/peel/python/3.7.9/gcc/lib/python3.7/site-pack
ages (from requests->socrata-py) (4.0.0)
WARNING: You are using pip version 20.1.1; however, version 21.3.1 is available.
You should consider upgrading via the '/share/apps/peel/python/3.7.9/gcc/bin/python -m pip install --upgrade
pip' command.
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: openclean in /share/apps/peel/openclean/0.2.1/lib/python3.7/site-packages (0.
2.1)
```

Took 6 sec. Last updated by nbuser at December 12 2021, 5:11:10 AM.

```
%pyspark
historical_dob_df = spark.read.options(header='true').csv('/user/CS-GY-6513/project_data/data-cityofnewyork
dob_cert_occupancy_df = spark.read.options(header='true').csv('/user/CS-GY-6513/project_data/data-cityofnew
housing_litigations_df = spark.read.options(header='true').csv('/user/CS-GY-6513/project_data/data-cityofne
housing_maintenance_code_complaints_df = spark.read.options(header='true').csv('/user/CS-GY-6513/project_da
housing_maintenance_code_violations_df = spark.read.options(header='true').csv('/user/CS-GY-6513/project_da
historical_dob_df.createOrReplaceTempView('historical_dob')
dob_cert_occupancy_df.createOrReplaceTempView('dob_cert_occupancy')
housing_litigations_df.createOrReplaceTempView('housing_litigations')
housing_maintenance_code_complaints_df.createOrReplaceTempView('housing_maintenance_code_complaints')
housing_maintenance_code_violations_df.createOrReplaceTempView('housing_maintenance_code_violations')
```

SPARK JOB FINISHED

Took 1 sec. Last updated by nbuser at December 12 2021, 5:41:33 AM.

```
%pyspark
historical_dob_df.select('BIN', 'Number', 'Street', 'Postcode', 'BOROUGH').show()
```

SPARK JOB (http://hc02.nyu.cluster:39642/jobs/job?id=10) FINISHED

```
+-------+------+--------------------+--------+-------+
|    BIN|Number|              Street|Postcode|BOROUGH|
+-------+------+--------------------+--------+-------+
|2118801|  2960|      WEBSTER AVENUE|   10458|  BRONX|
|2096812|   100|       DEKRUIF PLACE|   10475|  BRONX|
|2008604|  1898|     HARRISON AVENUE|   10453|  BRONX|
|2007652|  1998|       MORRIS AVENUE|   10453|  BRONX|
|2084155|   565|    WEST  235 STREET|   10463|  BRONX|
|2012264|   606|   EAST FORDHAM ROAD|   10458|  BRONX|
|2103486|   730|CONCOURSE VILLAGE...|   10451|  BRONX|
```

```
|2000391|   345|        BROOK AVENUE|   10454|  BRONX|
|2011594|  4487|        THIRD AVENUE|   10457|  BRONX|
|2001106|   575|       WALTON AVENUE|   10451|  BRONX|
|2119514|    69|         BEACON LANE|   10473|  BRONX|
|2097350|  1400|      PELHAM PARKWAY|   10461|  BRONX|
|2085168|   444|    WEST  259 STREET|   10471|  BRONX|
|2005389|   140|       DEKRUIF PLACE|   10475|  BRONX|
```

Took 0 sec. Last updated by anonymous at December 11 2021, 4:05:48 AM.

```
%pyspark
historical_dob_df.count()
```

SPARK JOB (http://hc02.nyu.cluster:39642/jobs/job?id=11) FINISHED

```
2428526
```

Took 6 sec. Last updated by anonymous at December 11 2021, 4:15:47 AM.

```
%pyspark
historical_dob_df.filter(historical_dob_df['Postcode'].isNull()).filter(historical_dob_df['Latitude'].isNull(
```

SPARK JOB (http://hc02.nyu.cluster:39642/jobs/job?id=25) FINISHED

```
761
```

Took 8 sec. Last updated by anonymous at December 11 2021, 4:56:04 AM.

```
%pyspark
historical_dob_distinct_building_df = historical_dob_df.select('BIN', 'Number', 'Street', 'Postcode', 'BOROU(
```

FINISHED

Took 1 sec. Last updated by anonymous at December 11 2021, 4:17:01 AM.

```
%pyspark
historical_dob_distinct_building_cnt = historical_dob_distinct_building_df.groupby('BIN').count()
```

FINISHED

Took 0 sec. Last updated by anonymous at December 11 2021, 4:42:38 AM.

```
%pyspark
historical_dob_distinct_building_cnt.filter(historical_dob_distinct_building_cnt['count'] > 1).show()
```

SPARK JOB (http://hc02.nyu.cluster:39642/jobs/job?id=14) FINISHED

```
+-------+-----+
|    BIN|count|
+-------+-----+
|1015827|    4|
|1083541|    2|
|1006620|    3|
|1082303|    5|
|1005754|    4|
|1024757|   22|
|4533191|    2|
|2057310|    3|
|3340850|    2|
|1038572|    2|
|4005275|    2|
|5149496|    2|
|2023445|    2|
|2007940|    3|
```

DOB Permit Issuance

Took 11 sec. Last updated by anonymous at December 11 2021, 4:43:18 AM.

```
%pyspark
historical_dob_distinct_building_df.filter(historical_dob_distinct_building_df['BIN'] == 1024757).show()
```

SPARK JOB FINISHED

```
+-------+------+---------------+--------+---------+
|    BIN|Number|         Street|Postcode|  BOROUGH|
+-------+------+---------------+--------+---------+
|1024757|  1590|        BROADWAY|   10036|MANHATTAN|
|1024757|  1588|        BROADWAY|   10036|MANHATTAN|
|1024757|   708|  SEVENTH AVENUE|   10036|MANHATTAN|
|1024757|   708|        7 AVENUE|   10036|MANHATTAN|
|1024757|     2|     TIME SQUARE|   10036|MANHATTAN|
|1024757|   714|      7TH AVENUE|   10036|MANHATTAN|
|1024757|     2|    TIMES SQUARE|   10036|MANHATTAN|
|1024757|  1592|        BROADWAY|   10036|MANHATTAN|
|1024757|   710|  SEVENTH AVENUE|   10036|MANHATTAN|
|1024757|  1584|        BROADWAY|   10036|MANHATTAN|
|1024757|   704|        7 AVENUE|   10036|MANHATTAN|
|1024757|   714|        7 AVENUE|   10036|MANHATTAN|
|1024757|  1580|        BROADWAY|   10036|MANHATTAN|
|1024757|  1576|        BROADWAY|   10036|MANHATTAN|
```

Took 9 sec. Last updated by anonymous at December 11 2021, 4:44:55 AM.

```
%pyspark
housing_litigations_df.select('LitigationID', 'BuildingID', 'BIN', 'HouseNumber', 'StreetName', 'Zip', 'Bor
```

SPARK JOB FINISHED

```
+------------+----------+-------+-----------+--------------------+-----+----+---------+----------+
|LitigationID|BuildingID|    BIN|HouseNumber|          StreetName|  Zip|Boro| Latitude| Longitude|
+------------+----------+-------+-----------+--------------------+-----+----+---------+----------+
|      299988|      6037|1056401|        784|    AMSTERDAM AVENUE|10025|   1|40.795592|-73.969261|
|       96958|    805376|1085178|       1305|        THIRD AVENUE|10021|   1|40.771509|-73.959304|
|      288066|     43492|1063707|        524|     WEST 184 STREET|10033|   1|40.850261|-73.930470|
|       15892|    116740|2003497|       1055|DR M L KING JR BO...|10452|   2|40.834676|-73.930269|
```

DOB Permit Issuance

```
|      246954|    366832|3079587|        280|     SCHAEFER STREET|11237|   3|40.692319|-73.905219|
|      208468|    265319|3109164|        402|      EAST 46 STREET|11203|   3|40.652120|-73.934341|
|      129962|    186156|3087868|       2869|     ATLANTIC AVENUE|11207|   3|40.677032|-73.888300|
|       28053|    428406|4113860|     144-45|           41 AVENUE|11355|   4|40.760695|-73.820986|
| [illegible]|[illegible]|[illegible]| [illegible]|      CROTONA AVENUE|10457|   2|40.851232|-73.885242|
|      254002|     46172|2012027|       2110|       ARTHUR AVENUE|10457|   2|40.849868|-73.891292|
|      197682|    116660|2014626|       2422|   UNIVERSITY AVENUE|10468|   2|40.863401|-73.904512|
|       42324|    111955|2025351|       1235|    STRATFORD AVENUE|10472|   2|40.830390|-73.875785|
|      244158|    117443|   null|  578 FRONT|     VAN NEST AVENUE|10460|   2|40.842165|-73.868502|
|      170344|    225096|3244549|       3076| CONEY ISLAND AVENUE|11235|   3|40.581118|-73.959892|
```

Took 0 sec. Last updated by anonymous at December 10 2021, 4:17:21 AM.

FINISHED

```
%pyspark
housing_litigations_boro_df = spark.sql("select LitigationID, BuildingID, BIN, HouseNumber, StreetName, Zip
    end as Borough, Latitude, Longitude from housing_litigations")
```

Took 0 sec. Last updated by nbuser at December 12 2021, 5:41:53 AM.

SPARK JOB (http://hc11.nyu.cluster:44088/jobs/job?id=14) FINISHED

```
%pyspark
housing_litigations_boro_df.filter(housing_litigations_boro_df['Zip'].isNull()).filter(housing_litigations_bo
```

```
63
```

Took 1 sec. Last updated by nbuser at December 12 2021, 5:42:55 AM.

SPARK JOB FINISHED

```
%pyspark
housing_litigations_boro_df.select('LitigationID', 'BuildingID', 'BIN', 'HouseNumber', 'StreetName', 'Zip', '
```

```
+------------+----------+-------+-----------+-------------------+-----+-------------+---------+----------+
|LitigationID|BuildingID|    BIN|HouseNumber|         StreetName|  Zip|      Borough| Latitude| Longitude|
+------------+----------+-------+-----------+-------------------+-----+-------------+---------+----------+
|      131047|     66460|2012965|        786|     EAST 182 STREET|10460|        Bronx|40.848909|-73.884165|
|      161956|     27988|1063950|       1441|   ST NICHOLAS AVENUE|10033|    Manhattan|40.849991|-73.933253|
|      364010|    842702|2114115|       2111|       HUGHES AVENUE|10457|        Bronx|40.849329|-73.890233|
|      258106|     42827|1062752|        538|     WEST 159 STREET|10032|    Manhattan|40.834579|-73.942245|
|       48315|    466695|4451496|     150-10|           71 AVENUE|11367|       Queens|40.729238|-73.817700|
|       32415|    243141|3179168|        800|      EAST 12 STREET|11230|     Brooklyn|40.629530|-73.965065|
|      239944|     59195|2013896|       2301|      CRESTON AVENUE|10468|        Bronx|40.858062|-73.900771|
|      182307|    346877|3029764|        596|     NOSTRAND AVENUE|11216|     Brooklyn|40.678107|-73.949732|
|       41646|    726025|5002483|        247|       CORSON AVENUE|10301|Staten Island|40.637004|-74.086736|
|      238699|    363795|3034421|        332|       ROGERS AVENUE|11225|     Brooklyn|40.665153|-73.953735|
|      125219|     75641|2067469|       4077|        EDSON AVENUE|10466|        Bronx|40.891863|-73.844233|
|      313118|       434|1006490|        215|            1 AVENUE|10003|    Manhattan|40.730517|-73.983183|
|      132792|    352670|3044159|         52|      PATCHEN AVENUE|11221|     Brooklyn|40.690669|-73.927618|
|      241901|     23670|1060119|        657|        LENOX AVENUE|10037|    Manhattan|40.819016|-73.937287|
```

Took 1 sec. Last updated by anonymous at December 10 2021, 4:32:58 AM.

SPARK JOB (http://hc11.nyu.cluster:40268/jobs/job?id=69) FINISHED

```
%pyspark
housing_litigations_boro_df.filter(housing_litigations_boro_df['Zip'] == 10463).show()
```

```
+------------+----------+-------+-----------+------------------+-----+---------+---------+----------+
|LitigationID|BuildingID|    BIN|HouseNumber|        StreetName|  Zip|  Borough| Latitude| Longitude|
```

DOB Permit Issuance

```
+------------+----------+-------+-----------+-------------------+-----+---------+---------+----------+
|      129965|     54954|2084079|       3235|   CAMBRIDGE AVENUE|10463|    Bronx|40.883997|-73.908573|
|       82592|    118583|2083947|       3660|       WALDO AVENUE|10463|    Bronx|40.886851|-73.904392|
|      111668|    108597|2092463|       3605|    SEDGWICK AVENUE|10463|    Bronx|40.881542|-73.896483|
|   [illegible]| [illegible]|[illegible]| [illegible]|KINGSBRIDGE TERRACE|10463|    Bronx|40.872342|-73.903288|
|      373725|     46543|2015948|       3300|      BAILEY AVENUE|10463|    Bronx|40.879454|-73.901253|
|      106442|     89811|2015702|       2765|KINGSBRIDGE TERRACE|10463|    Bronx|40.872517|-73.903252|
|      276716|     89811|2015702|       2765|KINGSBRIDGE TERRACE|10463|    Bronx|40.872517|-73.903252|
|      231703|     25387|1064584|         99| MARBLE HILL AVENUE|10463|Manhattan|40.877059|-73.909248|
|      253339|     89724|2083169|       3120| KINGSBRIDGE AVENUE|10463|    Bronx|40.880369|-73.906014|
|      298527|    879536|2113577|       2726|KINGSBRIDGE TERRACE|10463|    Bronx|40.871667|-73.903513|
|      234483|    108597|2092463|       3605|    SEDGWICK AVENUE|10463|    Bronx|40.881542|-73.896483|
|      148906|    121353|2112572|        231|    WEST 230 STREET|10463|    Bronx|40.877587|-73.907309|
|      258135|     89714|2083150|       3024| KINGSBRIDGE AVENUE|10463|    Bronx|40.878771|-73.907550|
|      375287|    100042|2084093|       3205|      OXFORD AVENUE|10463|    Bronx|40.884053|-73.900517|
```

Took 0 sec. Last updated by anonymous at December 10 2021, 4:59:48 AM.

```
%pyspark
housing_maintenance_code_violations_df.filter(housing_maintenance_code_violations_df['Postcode'].isNull()).fi
```

SPARK JOB (http://hc11.nyu.cluster:44088/jobs/job?id=15) FINISHED

```
3758
```

Took 15 sec. Last updated by nbuser at December 12 2021, 5:44:02 AM.

```
%pyspark
housing_maintenance_code_violations_sample = housing_maintenance_code_violations_df.sample(False, 0.1, 1121
housing_maintenance_code_violations_sample.count()
```

SPARK JOB (http://hc11.nyu.cluster:44088/jobs/job?id=5) FINISHED

```
688358
```

Took 10 sec. Last updated by nbuser at December 12 2021, 5:22:58 AM.

```
%pyspark
housing_maintenance_code_violations_sample.select('ViolationID', 'BuildingID', 'BIN', 'HouseNumber', 'LowHo
.filter(housing_maintenance_code_violations_sample['Postcode'] == 10463)\
.filter(housing_maintenance_code_violations_sample['Borough'] != 'BRONX')\
.show()
```

SPARK JOB FINISHED

```
+----------+---------+-------+----------+-------------+--------------+------------------+--------+-----
----+---------+----------+
|ViolationID|BuildingID|    BIN|HouseNumber|LowHouseNumber|HighHouseNumber|         StreetName|Postcode|  Bor
ough| Latitude| Longitude|
+----------+---------+-------+----------+-------------+--------------+------------------+--------+-----
----+---------+----------+
|   13780863|     28645|1064543|       110|          110|           116|TERRACE VIEW AVENUE|   10463|MANHA
TTAN|40.876444|-73.912829|
|   13827795|     21653|1064651|        26|           26|            28|   FT CHARLES PLACE|   10463|MANHA
TTAN|40.875883|-73.910634|
|   10091660|     25378|1064679|        58|           58|            60| MARBLE HILL AVENUE|   10463|MANHA
TTAN|40.875720|-73.909702|
|   11220171|     28648|1064524|       135|          131|           135|TERRACE VIEW AVENUE|   10463|MANHA
TTAN|40.876439|-73.912854|
|   11604953|     25387|1064584|        99|           99|           101| MARBLE HILL AVENUE|   10463|MANHA
TTAN|40.877059|-73.909248|
|   14040212|     25387|1064584|        99|           99|           101| MARBLE HILL AVENUE|   10463|MANHA
```

DOB Permit Issuance

Took 27 sec. Last updated by anonymous at December 11 2021, 2:18:33 AM.

```
%pyspark
housing_maintenance_code_violations_sample.filter(housing_maintenance_code_violations_sample['Postcode'].isN
```

SPARK JOB FINISHED

```
+----------+---------+----+----------+-------------+--------------+---------------+--------+--------+--
------+---------+
|ViolationID|BuildingID| BIN|HouseNumber|LowHouseNumber|HighHouseNumber|      StreetName|Postcode| Borough|La
titude|Longitude|
+----------+---------+----+----------+-------------+--------------+---------------+--------+--------+--
------+---------+
|   14276447|    967507|null|       532|          532|           532| EAST 142 STREET|    null|   BRONX|
null|     null|
|   10611493|    940715|null|       201|          201|           201| NORTH 11 STREET|    null|BROOKLYN|
null|     null|
|   10612379|    942882|null|     57-37|        57-37|         57-37|VAN DOREN STREET|    null|  QUEENS|
null|     null|
|   14299130|    997630|null|       223|          221|           223|  MILFORD STREET|    null|BROOKLYN|
null|     null|
|   10741476|    956610|null|        91|           91|            91|  CRYSTAL STREET|    null|BROOKLYN|
null|     null|
|   10971774|    955005|null|     90-47|        90-47|         90-47|      198 STREET|    null|  QUEENS|
```

Took 25 sec. Last updated by anonymous at December 11 2021, 9:18:58 PM.

```
%pyspark
housing_maintenance_code_violations_sample.filter(housing_maintenance_code_violations_df['StreetName'] == '
    'Latitude', 'Longitude').show()
```

SPARK JOB FINISHED

```
+----------+---------+----+----------+-------------+--------------+------------------+--------+-------
+--------+---------+
|ViolationID|BuildingID| BIN|HouseNumber|LowHouseNumber|HighHouseNumber|         StreetName|Postcode|Boroug
h|Latitude|Longitude|
+----------+---------+----+----------+-------------+--------------+------------------+--------+-------
+--------+---------+
|   10314587|    850855|null|     109-55|        109-55|         109-61|109-55 VAN WYCK E...|    null| QUEEN
```

```
S|    null|     null|
+----------+----------+----+----------+-------------+--------------+-------------------+--------+-------
+--------+---------+
```

DOB Permit Issuance

Took 22 sec. Last updated by nbuser at December 12 2021, 5:23:29 AM.

```
%pyspark
housing_maintenance_code_violations_sample.select('StreetName').distinct().count()
```

SPARK JOB (http://hc05.nyu.cluster:33536/jobs/job?id=19) FINISHED

```
3295
```

Took 17 sec. Last updated by anonymous at December 11 2021, 9:19:36 PM.

```
%sh
/usr/bin/hadoop fs -rm -r housing_maintenance_code_violations_sample.csv
```

FINISHED

```
21/12/11 18:21:38 INFO fs.TrashPolicyDefault: Moved: 'hdfs://horton.hpc.nyu.edu:8020/user/ky572/housing_maint
enance_code_violations_sample.csv' to trash at: hdfs://horton.hpc.nyu.edu:8020/user/ky572/.Trash/Current/user
/ky572/housing_maintenance_code_violations_sample.csv
```

Took 2 sec. Last updated by anonymous at December 11 2021, 10:21:39 PM.

```
%pyspark
housing_maintenance_code_violations_sample.write.option("header", True).csv('./housing_maintenance_code_viola
```

SPARK JOB (http://hc05.nyu.cluster:33536/jobs/job?id=24) FINISHED

Took 38 sec. Last updated by anonymous at December 11 2021, 10:22:20 PM.

```
%sh
rm housing_maintenance_code_violations_sample.csv
```

FINISHED

Took 3 sec. Last updated by anonymous at December 11 2021, 10:22:46 PM. (outdated)

```
%sh
/usr/bin/hadoop fs -getmerge housing_maintenance_code_violations_sample.csv housing_maintenance_code_violatic
```

READY

```
%python
from openclean.pipeline import stream
from openclean.function.eval.base import Col
from openclean.function.eval.logic import And
from openclean.function.eval.null import IsNotEmpty, IsEmpty
from openclean.operator.map.violations import fd_violations
from openclean.cluster.key import key_collision
from openclean_geo.address.usstreet import USStreetNameKey

def lat_long_boro_fd(ds, latCol, longCol, boroCol):
    data = ds.select([boroCol, latCol, longCol]).where(And(IsNotEmpty(latCol), IsNotEmpty(longCol), IsNotEm
    df = data\
        .select([boroCol, latCol, longCol])\
        .to_df()

    groups = fd_violations(df, lhs=[latCol, longCol], rhs=boroCol)
    return groups

def cleanstreet(street,ds_full):
    streets = ds_full.select(street).distinct()
```

FINISHED

DOB Permit Issuance

Took 3 sec. Last updated by nbuser at December 12 2021, 5:17:28 AM.

```
%python
housing_maintenance_code_violations_sample_ds = stream('./housing_maintenance_code_violations_sample.csv')
```

FINISHED

Took 0 sec. Last updated by nbuser at December 12 2021, 5:17:38 AM.

```
%python
housing_maintenance_code_violations_sample_ds.where(Col('StreetName') == '27TH STREET').select(['HouseNumber'
```

FINISHED

```
Empty DataFrame
Columns: [HouseNumber, StreetName, Postcode]
Index: []
```

Took 6 sec. Last updated by anonymous at December 11 2021, 10:28:36 PM.

```
%python
cleaned = cleanstreet('StreetName', housing_maintenance_code_violations_sample_ds)
```

FINISHED

Took 6 sec. Last updated by nbuser at December 12 2021, 5:17:48 AM.

```
%python
cleaned.where(IsEmpty('Postcode')).count()
```

FINISHED

```
378
```

Took 8 sec. Last updated by nbuser at December 12 2021, 5:18:01 AM.

```
%python
cleaned.where(IsEmpty('Postcode')).select(['ViolationID', 'BIN', 'HouseNumber', 'StreetName', 'Postcode', 'Bo
```

FINISHED

	ViolationID	BIN	HouseNumber	StreetName	Postcode	Borough
927	10859232		1754	ANTHONY AVENUE		BRONX
1514	11323977		238-11	HILLSIDE AVENUE		QUEENS
6238	13317952		881	ALBANY AVENUE		BROOKLYN
6679	13417672		111-03	38 AVENUE		QUEENS
9041	13681082		235	SKILLMAN STREET		BROOKLYN
11812	14276447		532	EAST 142 STREET		BRONX
11837	14276451		532	EAST 142 STREET		BRONX
12889	13767458		152	SACKMAN STREET		BROOKLYN
12890	13767468		152	SACKMAN STREET		BROOKLYN
12891	13767470		152	SACKMAN STREET		BROOKLYN

Took 0 sec. Last updated by anonymous at December 12 2021, 12:21:48 AM.

```
%python
```

FINISHED

DOB Permit Issuance

Took 19 sec. Last updated by nbuser at December 12 2021, 5:18:34 AM.

FINISHED

```
%python
for key in list(groups.keys())[:10]:
    print(groups.values(key=key, columns='Postcode'))
    print('\n')
```

```
Counter({'11219': 2, '': 1})


Counter({'10468': 2, '': 1})


Counter({'11373': 1, '': 1})


Counter({'': 1, '11226': 1})


Counter({'11219': 3, '': 1})


Counter({'': 1, '11221': 1})
```

Took 0 sec. Last updated by anonymous at December 12 2021, 12:58:53 AM.

FINISHED

```
%python
group_mapping = dict()
for key in groups.keys():
    values = groups.values(key=key, columns='Postcode')
    max=0
    maxValue=""
    for v in values:
        if v is None or len(v) == 0:
            continue
        if max<values[v]:
            max=values[v]
            maxValue=v
    group_mapping[key] = maxValue
```

Took 0 sec. Last updated by nbuser at December 12 2021, 5:18:41 AM.

FINISHED

```
%python
for key in list(group_mapping.keys())[:10]:
    print('{} = {}'.format(key, group_mapping[key]))
```

```
('BROOKLYN', '204', 'LINDEN BOULEVARD') = 11226
('QUEENS', '84-47', 'KNEELAND AVENUE') = 11373
('BROOKLYN', '1064', '57 STREET') = 11219
('BROOKLYN', '1414', '47 STREET') = 11219
('BROOKLYN', '834', 'QUINCY STREET') = 11221
```

```
('BRONX', '2302', 'MORRIS AVENUE') = 10468
```

Took 0 sec. Last updated by nbuser at December 12 2021, 5:18:45 AM.

DOB Permit Issuance

FINISHED

```
%python
from openclean.function.eval.domain import Lookup

cleaned2 = cleaned.where(IsEmpty('Postcode')).update('Postcode', Lookup(columns=['Borough', 'HouseNumber',
```

Took 0 sec. Last updated by nbuser at December 12 2021, 5:18:58 AM.

FINISHED

```
%python
cleaned2.where(IsEmpty('Postcode')).count()
```

```
372
```

Took 8 sec. Last updated by nbuser at December 12 2021, 5:19:10 AM.

FINISHED

```
%python
cleaned2.where(IsEmpty('Postcode')).select(['ViolationID', 'BIN', 'HouseNumber', 'StreetName', 'Postcode', 'E
```

```
       ViolationID BIN HouseNumber  ...        Borough Latitude Longitude
927        10859232            1754  ...          BRONX
1514       11323977          238-11  ...         QUEENS
6238       13317952             881  ...       BROOKLYN
6679       13417672          111-03  ...         QUEENS
9041       13681082             235  ...       BROOKLYN
...             ...  ..          ...  ...            ...      ...       ...
205019     11606836            2408  ...          BRONX
209011     13843240             441  ...       BROOKLYN
209141     14479940             590  ...  STATEN ISLAND
218166     14524400             114  ...       BROOKLYN
222810     11807813             943  ...          BRONX

[100 rows x 8 columns]
```

Took 3 sec. Last updated by nbuser at December 12 2021, 5:20:15 AM.

FINISHED

```
%python
cleaned.where(And(Col('HouseNumber') == '532', Col('StreetName') == 'EAST 142 STREET', IsNotEmpty('Postcode'
```

```
Empty DataFrame
Columns: [HouseNumber, StreetName, Postcode, Borough, BIN]
Index: []
```

Took 9 sec. Last updated by anonymous at December 12 2021, 1:03:10 AM.

FINISHED

```
%python
cleaned.where(Col('StreetName') == 'VAN WYCK EXPRESSWAY').select(['HouseNumber', 'StreetName', 'Postcode', 'E
```

```
       HouseNumber           StreetName Postcode Borough
127656       95-12  VAN WYCK EXPRESSWAY    11419  QUEENS
252759       95-12  VAN WYCK EXPRESSWAY    11419  QUEENS
466509       87-53  VAN WYCK EXPRESSWAY    11435  QUEENS
```

```
494194       95-12  VAN WYCK EXPRESSWAY    11419  QUEENS
494195       95-12  VAN WYCK EXPRESSWAY    11419  QUEENS
497254       95-18  VAN WYCK EXPRESSWAY    11419  QUEENS
535068      118-19  VAN WYCK EXPRESSWAY    11436  QUEENS
560913       95-12  VAN WYCK EXPRESSWAY    11419  QUEENS
```

DOB Permit Issuance

Took 7 sec. Last updated by nbuser at December 12 2021, 5:19:27 AM.

```
%pyspark
housing_maintenance_code_violations_df.select('ViolationID', 'BuildingID', 'BIN', 'HouseNumber', 'LowHouseN
.filter(housing_maintenance_code_violations_df['Postcode'] == 11232)\
.filter(housing_maintenance_code_violations_df['Borough'] != 'BROOKLYN')\
.show()
```

SPARK JOB FINISHED

```
+-----------+----------+----+-----------+-------------+--------------+-----------+--------+-------+--------
-+----------+
|ViolationID|BuildingID| BIN|HouseNumber|LowHouseNumber|HighHouseNumber| StreetName|Postcode|Borough| Latitud
e| Longitude|
+-----------+----------+----+-----------+-------------+--------------+-----------+--------+-------+--------
-+----------+
|   14283892|    967958|null|      41-32|         41-32|          41-32|27TH STREET|   11232| QUEENS|40.66101
5|-74.001943|
+-----------+----------+----+-----------+-------------+--------------+-----------+--------+-------+--------
-+----------+
```

Took 24 sec. Last updated by anonymous at December 10 2021, 10:13:36 PM.

```
%pyspark

housing_maintenance_code_violations_df.select('ViolationID', 'BuildingID', 'BIN', 'HouseNumber', 'LowHouseN
    (housing_maintenance_code_violations_df['BuildingID'] != 1).show(n=600)
```

SPARK JOB FINISHED

```
+-----------+----------+----+-----------+-------------+--------------+--------------------+--------+-------
------+--------+---------+
|ViolationID|BuildingID| BIN|HouseNumber|LowHouseNumber|HighHouseNumber|          StreetName|Postcode|      B
orough|Latitude|Longitude|
+-----------+----------+----+-----------+-------------+--------------+--------------------+--------+-------
------+--------+---------+
|   14265577|    904119|null|        273|          273|           273|     WEST 122 STREET|    null|    MAN
HATTAN|    null|     null|
|   10521517|    958166|null|         42|           42|            42|     SWEETBROOK ROAD|    null|STATEN
ISLAND|    null|     null|
|   10594623|    956415|null|      27-16|         27-16|         27-16|           18 STREET|    null|
QUEENS|    null|     null|
|   10623685|    803906|null|       2411|          2411|          2411|FREDERICK DOUGAL ...|   10027|    MAN
HATTAN|    null|     null|
|   10694011|    936460|null|       1431|          1431|          1431|       ZEREGA AVENUE|    null|
BRONX|    null|     null|
|   10694022|    936460|null|       1431|          1431|          1431|       ZEREGA AVENUE|    null|
```

Took 8 sec. Last updated by anonymous at December 10 2021, 10:35:14 PM. (outdated)

```
%pyspark
housing_maintenance_code_violations_df.filter(housing_maintenance_code_violations_df['StreetName'] == '72 STR
```

SPARK JOB (http://hc11.nyu.cluster:44088/jobs/job?id=16) FINISHED

DOB Permit Issuance

```
+-----------+----------+-------+-----------+-------------+--------------+----------+--------+--------+-----
----+----------+
|ViolationID|BuildingID|    BIN|HouseNumber|LowHouseNumber|HighHouseNumber|StreetName|Postcode| Borough| Lati
tude|Longitude|
+-----------+----------+-------+-----------+-------------+--------------+----------+--------+--------+-----
----+----------+
|   10970791|    890887|4536785|      47-05|        47-05|         47-05| 72 STREET|   11377|  QUEENS|40.73
8861|-73.892238|
|   11545234|    890887|4536785|      47-05|        47-05|         47-05| 72 STREET|   11377|  QUEENS|40.73
8861|-73.892238|
|   14531260|    809275|3337150|        190|          190|           198| 72 STREET|   11209|BROOKLYN|40.63
4294|-74.030261|
|   12136761|    890887|4536785|      47-05|        47-05|         47-05| 72 STREET|   11377|  QUEENS|40.73
8861|-73.892238|
|   12660175|    169543|3146999|        871|          871|           871| 72 STREET|   11228|BROOKLYN|40.62
7756|-74.015102|
|   12716984|    890887|4536785|      47-05|        47-05|         47-05| 72 STREET|   11377|  QUEENS|40.73
```

Took 1 sec. Last updated by nbuser at December 12 2021, 7:17:43 AM.

```
%pyspark
housing_maintenance_code_violations_df.filter(housing_maintenance_code_violations_df['BuildingID'] == 1).sele
```

SPARK JOB (http://hc13.nyu.cluster:39590/jobs/job?id=13) FINISHED

```
+-----------+----------+----+-----------+-------------+--------------+--------------------+--------+-------
--+--------+---------+
|ViolationID|BuildingID| BIN|HouseNumber|LowHouseNumber|HighHouseNumber|          StreetName|Postcode|  Borou
gh|Latitude|Longitude|
+-----------+----------+----+-----------+-------------+--------------+--------------------+--------+-------
--+--------+---------+
|   10526625|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
AN|    null|     null|
|   11491071|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
AN|    null|     null|
|   11491082|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
AN|    null|     null|
|   11491177|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
AN|    null|     null|
|   11491148|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
AN|    null|     null|
|   11550294|         1|null|        401|          401|           403|ST GEORGE'S CRESCENT|   10016|MANHATT
```

Took 1 sec. Last updated by anonymous at December 11 2021, 2:45:02 AM.

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance

DOB Permit Issuance