data_cleaning

November 15, 2019

[1]:
```
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql import functions as F

import datetime
```

Starting Spark session

[2]:
```
#spark = SparkSession.builder.master('spark://master:7077').appName("Jupyter").
 ↪getOrCreate()

spark = SparkSession.builder.master('local[1]').appName("Jupyter").getOrCreate()
sc = spark.sparkContext
```

[3]:
```
collisions = spark.read.csv('data/accidents.csv', header='true', inferSchema =␣
 ↪True)
collisions.show(2)
```

```
+--------------+-------------+----------+--------+-------------+-------------
---+-------------+----------+-------------+--------------------+-------------
----+--------------+-----------------+-------------+--------+--------------+---
-----+--------------+-------------+--------------+--------------+-------------+-
------------+------------------+-----------------+-----------------+---------
----+----------------------+--------+-----------------+--------------+---------
--------+----+-----------------+-------------------+------------+-------------
+-------------------------+--------------------------+------------------+----
------------------+-------------+--------------+---------------+-------+-------
---+------------+------+--------------+-----------------+---------------+-----
-----------+------------------+-------------+---------------+----------------+-
-----------+--------------+---------------+-------------+----------+----------
---+----------+----------+------------+---------------+---------------+--------
-------+
|    NO_SEQ_COLL|JR_SEMN_ACCDN|
DT_ACCDN|CD_MUNCP|NO_CIVIQ_ACCDN|SFX_NO_CIVQ_ACCDN|BORNE_KM_ACCDN|
RUE_ACCDN|TP_REPRR_ACCDN|       ACCDN_PRES_DE|NB_METRE_DIST_ACCD|CD_GENRE_ACCDN|
CD_SIT_PRTCE_ACCDN|CD_ETAT_SURFC|CD_ECLRM|CD_ENVRN_ACCDN|NO_ROUTE|CD_CATEG_ROUTE
|CD_ETAT_CHASS|CD_ASPCT_ROUTE|CD_LOCLN_ACCDN|CD_POSI_ACCDN|CD_CONFG_ROUTE|CD_ZON
```

```
_TRAVX_ROUTR|CD_PNT_CDRNL_ROUTE|CD_PNT_CDRNL_REPRR|CD_COND_METEO|NB_VEH_IMPLIQUE
S_ACCDN|NB_MORTS|NB_BLESSES_GRAVES|NB_BLESS_LEGERS|          HR_ACCDN|
AN|NB_VICTIMES_TOTAL|              GRAVITE|     REG_ADM|            MRC|nb_automob
ile_camion_leger|nb_camionLourd_tractRoutier|nb_outil_equipement|nb_tous_autobus
_minibus|nb_bicyclette|nb_cyclomoteur|nb_motocyclette|nb_taxi|nb_urgence|nb_moto
neige|nb_VHR|nb_autres_types|nb_veh_non_precise|NB_DECES_PIETON|NB_BLESSES_PIETO
N|NB_VICTIMES_PIETON|NB_DECES_MOTO|NB_BLESSES_MOTO|NB_VICTIMES_MOTO|NB_DECES_VEL
O|NB_BLESSES_VELO|NB_VICTIMES_VELO|VITESSE_AUTOR|      LOC_X|
LOC_Y|LOC_COTE_Q|LOC_COTE_P|LOC_DETACHEE|LOC_IMPRECISION|         LOC_LONG|
LOC_LAT|
+--------------+------------+---------+--------+-------------+-------------
---+--------------+----------+-------------+--------------------+-------------
----+--------------+-----------------+-------------+--------+--------------+---
-----+--------------+------------+--------------+--------------+-------------+-
-------------+------------------+-----------------+-----------------+---------
----+----------------------+--------+-----------------+---------------+---------
--------+----+-----------------+--------------------+-----------+-------------
+--------------------------+--------------------------+-------------------+----
-------------------+-------------+--------------+---------------+-------+-------
---+-----------+------+--------------+-----------------+---------------+-----
-----------+------------------+-------------+---------------+----------------+-
-----------+---------------+----------------+-------------+-----------+--------
---+----------+----------+------------+---------------+----------------+--------
-------+
|SPVM _ 2012 _ 1|           ME|2012/02/01|   66102|          3501|
null|          null|ST CHARLES|             2|                STAT|
null|            31|              null|           16|       1|             1|
null|            21|         null|            11|            33|         null|
4|             null|              null|              null|           11|
2|       0|               0|              0|      Non précisé|2012|
0|Dommages matériel…|Montréal(06)|Montréal (66 )|                          1|
0|                  0|                      0|            0|             0|
0|      0|         0|           0|     0|              0|                 1|
0|                0|                 0|            0|              0|
0|            0|              0|               0|
null|276517.3795|5035127.484|         A|         3|           0|
N|-73.861615900043|45.455504503902|
|SPVM _ 2012 _ 2|           SA|2012/06/02|   66023|          null|
null|          null|      null|          null|COTE VERTU ET AUT 40|
null|            59|              null|           11|       1|             3|
null|            11|         null|            22|            37|         null|
1|             null|              null|              null|           11|
1|       0|               0|              0|0|15:00:00-15:59:00|2012|
0|Dommages matériel…|Montréal(06)|Montréal (66 )|                          1|
0|                  0|                      0|            0|             0|
0|      0|         0|           0|     0|              0|                 0|
0|                0|                 0|            0|              0|
0|            0|              0|               0|         null|
```

```
287913.26|5038666.138|         A|         3|           N|
0|-73.716033473399|45.487715123285|
+--------------+------------+----------+--------+-------------+-------------
---+--------------+----------+-------------+--------------------+-------------
----+--------------+-----------------+------------+--------+-------------+---
-----+--------------+------------+-------------+--------------+-------------+-
-------------+------------------+------------------+------------------+---------
----+----------------------+--------+-----------------+--------------+---------
--------+----+-----------------+--------------------+------------+--------------
+-------------------------+--------------------------+------------------+----
------------------+------------+-------------+---------------+-------+-------
---+------------+------+---------------+-----------------+---------------+-----
-----------+------------------+------------+--------------+---------------+-
-----------+---------------+---------------+------------+-----------+--------
---+----------+----------+-----------+---------------+---------------+--------
-------+
only showing top 2 rows
```

0.0.1 Choosing the columns

[4]:
```
print('Collisions Info: \n')
collisions.columns
```

```
Collisions Info:
```

[4]:
```
['NO_SEQ_COLL',
 'JR_SEMN_ACCDN',
 'DT_ACCDN',
 'CD_MUNCP',
 'NO_CIVIQ_ACCDN',
 'SFX_NO_CIVQ_ACCDN',
 'BORNE_KM_ACCDN',
 'RUE_ACCDN',
 'TP_REPRR_ACCDN',
 'ACCDN_PRES_DE',
 'NB_METRE_DIST_ACCD',
 'CD_GENRE_ACCDN',
 'CD_SIT_PRTCE_ACCDN',
 'CD_ETAT_SURFC',
 'CD_ECLRM',
 'CD_ENVRN_ACCDN',
 'NO_ROUTE',
 'CD_CATEG_ROUTE',
 'CD_ETAT_CHASS',
```

```
'CD_ASPCT_ROUTE',
'CD_LOCLN_ACCDN',
'CD_POSI_ACCDN',
'CD_CONFG_ROUTE',
'CD_ZON_TRAVX_ROUTR',
'CD_PNT_CDRNL_ROUTE',
'CD_PNT_CDRNL_REPRR',
'CD_COND_METEO',
'NB_VEH_IMPLIQUES_ACCDN',
'NB_MORTS',
'NB_BLESSES_GRAVES',
'NB_BLESS_LEGERS',
'HR_ACCDN',
'AN',
'NB_VICTIMES_TOTAL',
'GRAVITE',
'REG_ADM',
'MRC',
'nb_automobile_camion_leger',
'nb_camionLourd_tractRoutier',
'nb_outil_equipement',
'nb_tous_autobus_minibus',
'nb_bicyclette',
'nb_cyclomoteur',
'nb_motocyclette',
'nb_taxi',
'nb_urgence',
'nb_motoneige',
'nb_VHR',
'nb_autres_types',
'nb_veh_non_precise',
'NB_DECES_PIETON',
'NB_BLESSES_PIETON',
'NB_VICTIMES_PIETON',
'NB_DECES_MOTO',
'NB_BLESSES_MOTO',
'NB_VICTIMES_MOTO',
'NB_DECES_VELO',
'NB_BLESSES_VELO',
'NB_VICTIMES_VELO',
'VITESSE_AUTOR',
'LOC_X',
'LOC_Y',
'LOC_COTE_Q',
'LOC_COTE_P',
'LOC_DETACHEE',
'LOC_IMPRECISION',
```

```
 'LOC_LONG',
 'LOC_LAT']
```

[5]:
```
df = collisions.select('NO_SEQ_COLL',  'DT_ACCDN', 'JR_SEMN_ACCDN',\
                       'REG_ADM', 'CD_MUNCP',\
                       'RUE_ACCDN',\
                       'CD_GENRE_ACCDN', 'CD_ETAT_SURFC', 'CD_ECLRM',\
                       'CD_ASPCT_ROUTE', \
                       'CD_CONFG_ROUTE', \
                       'CD_COND_METEO',\
                       'GRAVITE', \
                       'NB_VEH_IMPLIQUES_ACCDN', \
                       'NB_VICTIMES_TOTAL', 'NB_MORTS', \
                       'NB_BLESSES_GRAVES', 'NB_BLESS_LEGERS', \
                       'NB_DECES_PIETON', 'NB_BLESSES_PIETON',␣
↪'NB_VICTIMES_PIETON',\
                       'NB_DECES_MOTO', 'NB_BLESSES_MOTO', 'NB_VICTIMES_MOTO', \
                       'NB_DECES_VELO', 'NB_BLESSES_VELO', 'NB_VICTIMES_VELO',\
                       'nb_automobile_camion_leger',␣
↪'nb_camionLourd_tractRoutier',\
                       'nb_outil_equipement', 'nb_tous_autobus_minibus',␣
↪'nb_bicyclette',\
                       'nb_cyclomoteur', 'nb_motocyclette', 'nb_taxi',␣
↪'nb_urgence',\
                       'nb_motoneige', 'nb_VHR', 'nb_autres_types',␣
↪'nb_veh_non_precise')
```

[6]:
```
df.take(1)
```

[6]:
```
[Row(NO_SEQ_COLL='SPVM _ 2012 _ 1', DT_ACCDN='2012/02/01', JR_SEMN_ACCDN='ME',
REG_ADM='Montréal(06)', CD_MUNCP=66102, RUE_ACCDN='ST CHARLES',
CD_GENRE_ACCDN=31, CD_ETAT_SURFC=16, CD_ECLRM=1, CD_ASPCT_ROUTE=11,
CD_CONFG_ROUTE=4, CD_COND_METEO=11, GRAVITE='Dommages matériels inférieurs au
seuil de rapportage', NB_VEH_IMPLIQUES_ACCDN=2, NB_VICTIMES_TOTAL=0, NB_MORTS=0,
NB_BLESSES_GRAVES=0, NB_BLESS_LEGERS=0, NB_DECES_PIETON=0, NB_BLESSES_PIETON=0,
NB_VICTIMES_PIETON=0, NB_DECES_MOTO=0, NB_BLESSES_MOTO=0, NB_VICTIMES_MOTO=0,
NB_DECES_VELO=0, NB_BLESSES_VELO=0, NB_VICTIMES_VELO=0,
nb_automobile_camion_leger=1, nb_camionLourd_tractRoutier=0,
nb_outil_equipement=0, nb_tous_autobus_minibus=0, nb_bicyclette=0,
nb_cyclomoteur=0, nb_motocyclette=0, nb_taxi=0, nb_urgence=0, nb_motoneige=0,
nb_VHR=0, nb_autres_types=0, nb_veh_non_precise=1)]
```

[7]:
```
df.distinct().count()
```

[7]:
```
171271
```

0.0.2 Rename the columns

[8]:
```
df = df.withColumnRenamed('NO_SEQ_COLL', 'ID')
df = df.withColumnRenamed('DT_ACCDN', 'DATE')
df = df.withColumnRenamed('JR_SEMN_ACCDN', 'WEEK_DAY')
df = df.withColumnRenamed('REG_ADM', 'REG')
df = df.withColumnRenamed('CD_MUNCP', 'MUNCP')
df = df.withColumnRenamed('RUE_ACCDN', 'STREET')
df = df.withColumnRenamed('CD_GENRE_ACCDN', 'TYPE_ACCDN')
df = df.withColumnRenamed('CD_ETAT_SURFC', 'SURFACE')
df = df.withColumnRenamed('CD_ECLRM', 'LIGHT')
df = df.withColumnRenamed('CD_ASPCT_ROUTE', 'STR_ASPCT')
df = df.withColumnRenamed('CD_CONFG_ROUTE', 'STR_CONFIG')
df = df.withColumnRenamed('CD_COND_METEO', 'METEO')
# df = df.withColumnRenamed('VITESSE_AUTOR', 'SPEED')
```

[9]:
```
df.columns
```

[9]:
```
['ID',
 'DATE',
 'WEEK_DAY',
 'REG',
 'MUNCP',
 'STREET',
 'TYPE_ACCDN',
 'SURFACE',
 'LIGHT',
 'STR_ASPCT',
 'STR_CONFIG',
 'METEO',
 'GRAVITE',
 'NB_VEH_IMPLIQUES_ACCDN',
 'NB_VICTIMES_TOTAL',
 'NB_MORTS',
 'NB_BLESSES_GRAVES',
 'NB_BLESS_LEGERS',
 'NB_DECES_PIETON',
 'NB_BLESSES_PIETON',
 'NB_VICTIMES_PIETON',
 'NB_DECES_MOTO',
 'NB_BLESSES_MOTO',
 'NB_VICTIMES_MOTO',
 'NB_DECES_VELO',
 'NB_BLESSES_VELO',
 'NB_VICTIMES_VELO',
 'nb_automobile_camion_leger',
 'nb_camionLourd_tractRoutier',
```

```
 'nb_outil_equipement',
 'nb_tous_autobus_minibus',
 'nb_bicyclette',
 'nb_cyclomoteur',
 'nb_motocyclette',
 'nb_taxi',
 'nb_urgence',
 'nb_motoneige',
 'nb_VHR',
 'nb_autres_types',
 'nb_veh_non_precise']
```

0.0.3 Merging with Municipalities dataset

Dataset is taken from Official site of Municipality of Quebec

[10]:
```
municipalities = spark.read.csv('data/municipalities-1.csv', header='true',␣
 ↪inferSchema = True)
municipalities.show(5)
```

```
+----+--------------------+---------------+--------------------+
|Code| Nom de municipalité|Statut municipal|Date d'incorporation|
+----+--------------------+---------------+--------------------+
|1023|Les Îles-de-la-Ma…|  M-Municipalité| 2002-01-01 00:00:00|
|1042|          Grosse-Île|  M-Municipalité| 2006-01-01 00:00:00|
|2005|               Percé|         V-Ville| 1971-01-01 00:00:00|
|2010|Sainte-Thérèse-de…|  M-Municipalité| 1930-09-06 00:00:00|
|2015|      Grande-Rivière|         V-Ville| 1974-09-21 00:00:00|
+----+--------------------+---------------+--------------------+
only showing top 5 rows
```

[11]:
```
df = df.join(municipalities, df.MUNCP == municipalities.Code)
df.take(1)
```

[11]:
```
[Row(ID='SPVM _ 2012 _ 1', DATE='2012/02/01', WEEK_DAY='ME', REG='Montréal(06)',
MUNCP=66102, STREET='ST CHARLES', TYPE_ACCDN=31, SURFACE=16, LIGHT=1,
STR_ASPCT=11, STR_CONFIG=4, METEO=11, GRAVITE='Dommages matériels inférieurs au
seuil de rapportage', NB_VEH_IMPLIQUES_ACCDN=2, NB_VICTIMES_TOTAL=0, NB_MORTS=0,
NB_BLESSES_GRAVES=0, NB_BLESS_LEGERS=0, NB_DECES_PIETON=0, NB_BLESSES_PIETON=0,
NB_VICTIMES_PIETON=0, NB_DECES_MOTO=0, NB_BLESSES_MOTO=0, NB_VICTIMES_MOTO=0,
NB_DECES_VELO=0, NB_BLESSES_VELO=0, NB_VICTIMES_VELO=0,
nb_automobile_camion_leger=1, nb_camionLourd_tractRoutier=0,
nb_outil_equipement=0, nb_tous_autobus_minibus=0, nb_bicyclette=0,
nb_cyclomoteur=0, nb_motocyclette=0, nb_taxi=0, nb_urgence=0, nb_motoneige=0,
nb_VHR=0, nb_autres_types=0, nb_veh_non_precise=1, Code=66102, Nom de
```

```
municipalité='Kirkland', Statut municipal='V-Ville', Date
d'incorporation=datetime.datetime(2006, 1, 1, 0, 0))]
```

[12]:
```
df.count()
```

[12]:
```
171266
```

[13]:
```
df = df.withColumn('MUNCP', df["Nom de municipalité"])
df = df.drop("Code", "Nom de municipalité", "Statut municipal", "Date␣
↪d'incorporation")
df.take(1)
```

[13]:
```
[Row(ID='SPVM _ 2012 _ 1', DATE='2012/02/01', WEEK_DAY='ME', REG='Montréal(06)',
MUNCP='Kirkland', STREET='ST CHARLES', TYPE_ACCDN=31, SURFACE=16, LIGHT=1,
STR_ASPCT=11, STR_CONFIG=4, METEO=11, GRAVITE='Dommages matériels inférieurs au
seuil de rapportage', NB_VEH_IMPLIQUES_ACCDN=2, NB_VICTIMES_TOTAL=0, NB_MORTS=0,
NB_BLESSES_GRAVES=0, NB_BLESS_LEGERS=0, NB_DECES_PIETON=0, NB_BLESSES_PIETON=0,
NB_VICTIMES_PIETON=0, NB_DECES_MOTO=0, NB_BLESSES_MOTO=0, NB_VICTIMES_MOTO=0,
NB_DECES_VELO=0, NB_BLESSES_VELO=0, NB_VICTIMES_VELO=0,
nb_automobile_camion_leger=1, nb_camionLourd_tractRoutier=0,
nb_outil_equipement=0, nb_tous_autobus_minibus=0, nb_bicyclette=0,
nb_cyclomoteur=0, nb_motocyclette=0, nb_taxi=0, nb_urgence=0, nb_motoneige=0,
nb_VHR=0, nb_autres_types=0, nb_veh_non_precise=1)]
```

[15]:
```
df.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- DATE: string (nullable = true)
 |-- WEEK_DAY: string (nullable = true)
 |-- REG: string (nullable = true)
 |-- MUNCP: string (nullable = true)
 |-- STREET: string (nullable = true)
 |-- TYPE_ACCDN: integer (nullable = true)
 |-- SURFACE: integer (nullable = true)
 |-- LIGHT: integer (nullable = true)
 |-- STR_ASPCT: integer (nullable = true)
 |-- STR_CONFIG: integer (nullable = true)
 |-- METEO: integer (nullable = true)
 |-- GRAVITE: string (nullable = true)
 |-- NB_VEH_IMPLIQUES_ACCDN: integer (nullable = true)
 |-- NB_VICTIMES_TOTAL: integer (nullable = true)
 |-- NB_MORTS: integer (nullable = true)
 |-- NB_BLESSES_GRAVES: integer (nullable = true)
 |-- NB_BLESS_LEGERS: integer (nullable = true)
 |-- NB_DECES_PIETON: integer (nullable = true)
 |-- NB_BLESSES_PIETON: integer (nullable = true)
```

```
 |-- NB_VICTIMES_PIETON: integer (nullable = true)
 |-- NB_DECES_MOTO: integer (nullable = true)
 |-- NB_BLESSES_MOTO: integer (nullable = true)
 |-- NB_VICTIMES_MOTO: integer (nullable = true)
 |-- NB_DECES_VELO: integer (nullable = true)
 |-- NB_BLESSES_VELO: integer (nullable = true)
 |-- NB_VICTIMES_VELO: integer (nullable = true)
 |-- nb_automobile_camion_leger: integer (nullable = true)
 |-- nb_camionLourd_tractRoutier: integer (nullable = true)
 |-- nb_outil_equipement: integer (nullable = true)
 |-- nb_tous_autobus_minibus: integer (nullable = true)
 |-- nb_bicyclette: integer (nullable = true)
 |-- nb_cyclomoteur: integer (nullable = true)
 |-- nb_motocyclette: integer (nullable = true)
 |-- nb_taxi: integer (nullable = true)
 |-- nb_urgence: integer (nullable = true)
 |-- nb_motoneige: integer (nullable = true)
 |-- nb_VHR: integer (nullable = true)
 |-- nb_autres_types: integer (nullable = true)
 |-- nb_veh_non_precise: integer (nullable = true)
```

[16]:
```
df.select('MUNCP', 'REG').groupby('REG', 'MUNCP').count().sort('REG',␣
↪accending=True).show(300)
```

```
+--------------------+--------------------+------+
|                 REG|               MUNCP| count|
+--------------------+--------------------+------+
|Abitibi-Témiscami…|            Val-d'Or|     1|
|Abitibi-Témiscami…|         Ville-Marie|     3|
|Abitibi-Témiscami…|                Amos|     1|
|Bas-Saint-Laurent…|       La Rédemption|     1|
|Capitale-National…|  L'Isle-aux-Coudres|     1|
|Capitale-National…|              Québec|     4|
|Capitale-National…|     Baie-Saint-Paul|     1|
|Centre-du-Québec(17)|     Saint-Wenceslas|     2|
|Centre-du-Québec(17)|       Drummondville|     1|
|Centre-du-Québec(17)|        Plessisville|     3|
|Centre-du-Québec(17)|             Nicolet|     1|
|Chaudière-Appalac…|           Val-Alain|     1|
|Chaudière-Appalac…|               Lévis|     1|
|Chaudière-Appalac…|      Thetford Mines|     1|
|          Estrie(05)|           Stornoway|     1|
|          Estrie(05)|              Weedon|     1|
|          Estrie(05)|          Sherbrooke|     5|
|          Estrie(05)|            Dixville|     1|
|Gaspésie/-Îles-de…|           Paspébiac|     1|
|Gaspésie/-Îles-de…|Les Îles-de-la-Ma…|     1|
```

Lanaudière(14)	Saint-Sulpice	3
Lanaudière(14)	Saint-Alexis	11
Lanaudière(14)	Repentigny	3
Lanaudière(14)	Saint-Pierre	1
Lanaudière(14)	L'Assomption	1
Lanaudière(14)	Terrebonne	1
Lanaudière(14)	L'Épiphanie	2
Lanaudière(14)	Saint-Paul	1
Laurentides(15)	Saint-Jérôme	1
Laurentides(15)	Mirabel	1
Laurentides(15)	Labelle	1
Laurentides(15)	Rivière-Rouge	1
Laval(13)	Laval	8
Mauricie(04)	Shawinigan	1
Mauricie(04)	Trois-Rivières	3
Montréal(06)	Dorval	2370
Montréal(06)	Montréal	155084
Montréal(06)	Senneville	33
Montréal(06)	Côte-Saint-Luc	1328
Montréal(06)	Montréal-Est	479
Montréal(06)	Dollard-Des Ormeaux	2362
Montréal(06)	L'Île-Dorval	6
Montréal(06)	Baie-D'Urfé	120
Montréal(06)	Beaconsfield	585
Montréal(06)	Pointe-Claire	3561
Montréal(06)	Kirkland	1198
Montréal(06)	Westmount	1441
Montréal(06)	Mont-Royal	1704
Montréal(06)	Montréal-Ouest	179
Montréal(06)	Sainte-Anne-de-Be…	427
Montréal(06)	Hampstead	295
Montérégie(16)	Delson	1
Montérégie(16)	Sorel-Tracy	1
Montérégie(16)	Sainte-Julie	2
Montérégie(16)	Saint-Césaire	1
Montérégie(16)	Beauharnois	1
Montérégie(16)	Lacolle	1
Montérégie(16)	Saint-Simon	1
Montérégie(16)	Vaudreuil-Dorion	1
Montérégie(16)	Saint-Michel	1
Montérégie(16)	Varennes	1
Montérégie(16)	Saint-Bruno-de-Mo…	1
Montérégie(16)	Saint-Zotique	2
Montérégie(16)	Salaberry-de-Vall…	1
Montérégie(16)	Brossard	1
Outaouais(07)	Mansfield-et-Pont…	1
Outaouais(07)	Gatineau	3
Outaouais(07)	Papineauville	1

```
|       Outaouais(07)|              Pontiac|     1|
|Saguenay/-Lac-Sai…|   L'Anse-Saint-Jean|     1|
+--------------------+--------------------+------+
```

We can see that it makes sense to work with Greater Montreal area only

[17]:
```
df = df.filter(df['REG'] == "Montréal(06)")
df.count()
```

[17]:
```
171172
```

[18]:
```
df.select('MUNCP', 'REG').groupby('REG', 'MUNCP').count().sort('REG',␣
↪accending=True).show(300)
```

```
+------------+--------------------+------+
|         REG|               MUNCP| count|
+------------+--------------------+------+
|Montréal(06)|              Dorval|  2370|
|Montréal(06)|          Mont-Royal|  1704|
|Montréal(06)|        L'Île-Dorval|     6|
|Montréal(06)|            Montréal|155084|
|Montréal(06)|           Hampstead|   295|
|Montréal(06)| Dollard-Des Ormeaux|  2362|
|Montréal(06)|       Pointe-Claire|  3561|
|Montréal(06)|          Senneville|    33|
|Montréal(06)|           Westmount|  1441|
|Montréal(06)|      Côte-Saint-Luc|  1328|
|Montréal(06)|        Montréal-Est|   479|
|Montréal(06)|         Baie-D'Urfé|   120|
|Montréal(06)|        Beaconsfield|   585|
|Montréal(06)|      Montréal-Ouest|   179|
|Montréal(06)|Sainte-Anne-de-Be…|   427|
|Montréal(06)|            Kirkland|  1198|
+------------+--------------------+------+
```

We can also see that most of the time instead of entering the municipal codes, a general code of Montreal (66023) was used. We will probably gather them all in one group later.

0.0.4 Dealing with NULLs

[19]:
```
df.count()
```

[19]:
```
171172
```

[20]:
```
df.na.drop().count()
```

[20]: 128647

It's not enough to just drop rows which include nulls. Let's look closer into each column and see what can we do with nulls

Checking nulls for each column:

[21]:
```
for i in df.columns:
    print(i, ': ', df.filter(F.isnull(i)).count())
```

```
ID :  0
DATE :  0
WEEK_DAY :  0
REG :  0
MUNCP :  0
STREET :  10878
TYPE_ACCDN :  9038
SURFACE :  11265
LIGHT :  11398
STR_ASPCT :  8589
STR_CONFIG :  18720
METEO :  11915
GRAVITE :  0
NB_VEH_IMPLIQUES_ACCDN :  0
NB_VICTIMES_TOTAL :  0
NB_MORTS :  0
NB_BLESSES_GRAVES :  0
NB_BLESS_LEGERS :  0
NB_DECES_PIETON :  0
NB_BLESSES_PIETON :  0
NB_VICTIMES_PIETON :  0
NB_DECES_MOTO :  0
NB_BLESSES_MOTO :  0
NB_VICTIMES_MOTO :  0
NB_DECES_VELO :  0
NB_BLESSES_VELO :  0
NB_VICTIMES_VELO :  0
nb_automobile_camion_leger :  0
nb_camionLourd_tractRoutier :  0
nb_outil_equipement :  0
nb_tous_autobus_minibus :  0
nb_bicyclette :  0
nb_cyclomoteur :  0
nb_motocyclette :  0
nb_taxi :  0
nb_urgence :  0
nb_motoneige :  0
nb_VHR :  0
```

```
nb_autres_types :  0
nb_veh_non_precise :  0
```

A *STREET* column contains 10895 nulls. We will remove them from the dataset.

As for the other columns, we can replace nulls with 0 and still use then in our analysis as these columns are informative and the numbers we have there are pointing at the specific state of weather or road condition.

Let's for now drop all the nulls

[22]:
```
df = df.na.drop()
df.count()
```

[22]: 128647

[23]:
```
df.select('MUNCP', 'REG').groupby('REG', 'MUNCP').count().sort('REG',␣
↪accending=True).show(300)
```

```
+------------+--------------------+------+
|         REG|               MUNCP| count|
+------------+--------------------+------+
|Montréal(06)|              Dorval|  1687|
|Montréal(06)|          Mont-Royal|  1340|
|Montréal(06)|        L'Île-Dorval|     6|
|Montréal(06)|            Montréal|117195|
|Montréal(06)|           Hampstead|   248|
|Montréal(06)| Dollard-Des Ormeaux|  1605|
|Montréal(06)|       Pointe-Claire|  2355|
|Montréal(06)|          Senneville|    24|
|Montréal(06)|           Westmount|  1116|
|Montréal(06)|      Côte-Saint-Luc|  1013|
|Montréal(06)|        Montréal-Est|   362|
|Montréal(06)|         Baie-D'Urfé|    86|
|Montréal(06)|        Beaconsfield|   401|
|Montréal(06)|      Montréal-Ouest|   140|
|Montréal(06)|Sainte-Anne-de-Be…|   315|
|Montréal(06)|            Kirkland|   754|
+------------+--------------------+------+
```

0.0.5 Arranging the types

[24]:
```
df.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- DATE: string (nullable = true)
 |-- WEEK_DAY: string (nullable = true)
```

```
 |-- REG: string (nullable = true)
 |-- MUNCP: string (nullable = true)
 |-- STREET: string (nullable = true)
 |-- TYPE_ACCDN: integer (nullable = true)
 |-- SURFACE: integer (nullable = true)
 |-- LIGHT: integer (nullable = true)
 |-- STR_ASPCT: integer (nullable = true)
 |-- STR_CONFIG: integer (nullable = true)
 |-- METEO: integer (nullable = true)
 |-- GRAVITE: string (nullable = true)
 |-- NB_VEH_IMPLIQUES_ACCDN: integer (nullable = true)
 |-- NB_VICTIMES_TOTAL: integer (nullable = true)
 |-- NB_MORTS: integer (nullable = true)
 |-- NB_BLESSES_GRAVES: integer (nullable = true)
 |-- NB_BLESS_LEGERS: integer (nullable = true)
 |-- NB_DECES_PIETON: integer (nullable = true)
 |-- NB_BLESSES_PIETON: integer (nullable = true)
 |-- NB_VICTIMES_PIETON: integer (nullable = true)
 |-- NB_DECES_MOTO: integer (nullable = true)
 |-- NB_BLESSES_MOTO: integer (nullable = true)
 |-- NB_VICTIMES_MOTO: integer (nullable = true)
 |-- NB_DECES_VELO: integer (nullable = true)
 |-- NB_BLESSES_VELO: integer (nullable = true)
 |-- NB_VICTIMES_VELO: integer (nullable = true)
 |-- nb_automobile_camion_leger: integer (nullable = true)
 |-- nb_camionLourd_tractRoutier: integer (nullable = true)
 |-- nb_outil_equipement: integer (nullable = true)
 |-- nb_tous_autobus_minibus: integer (nullable = true)
 |-- nb_bicyclette: integer (nullable = true)
 |-- nb_cyclomoteur: integer (nullable = true)
 |-- nb_motocyclette: integer (nullable = true)
 |-- nb_taxi: integer (nullable = true)
 |-- nb_urgence: integer (nullable = true)
 |-- nb_motoneige: integer (nullable = true)
 |-- nb_VHR: integer (nullable = true)
 |-- nb_autres_types: integer (nullable = true)
 |-- nb_veh_non_precise: integer (nullable = true)
```

Checking the dates

[25]:
```
df.select('DATE', 'MUNCP', 'REG').groupby('DATE', 'REG', 'MUNCP').count().
 ↪sort('DATE', accending=True).show(50)
```

```
+----------+------------+--------------------+-----+
|      DATE|         REG|               MUNCP|count|
+----------+------------+--------------------+-----+
|2012/01/01|Montréal(06)|        Beaconsfield|    1|
```

```
|2012/01/01|Montréal(06)| Dollard-Des Ormeaux|    2|
|2012/01/01|Montréal(06)|            Montréal|   37|
|2012/01/02|Montréal(06)|            Montréal|   21|
|2012/01/02|Montréal(06)|           Kirkland|    2|
|2012/01/02|Montréal(06)|      Pointe-Claire|    1|
|2012/01/03|Montréal(06)|           Kirkland|    1|
|2012/01/03|Montréal(06)|             Dorval|    2|
|2012/01/03|Montréal(06)|         Mont-Royal|    2|
|2012/01/03|Montréal(06)|        Baie-D'Urfé|    1|
|2012/01/03|Montréal(06)|      Pointe-Claire|    2|
|2012/01/03|Montréal(06)|     Côte-Saint-Luc|    2|
|2012/01/03|Montréal(06)|       Montréal-Est|    1|
|2012/01/03|Montréal(06)|            Montréal|   29|
|2012/01/04|Montréal(06)| Dollard-Des Ormeaux|    2|
|2012/01/04|Montréal(06)|      Pointe-Claire|    1|
|2012/01/04|Montréal(06)|            Montréal|   39|
|2012/01/04|Montréal(06)|          Westmount|    1|
|2012/01/05|Montréal(06)|             Dorval|    1|
|2012/01/05|Montréal(06)|          Hampstead|    1|
|2012/01/05|Montréal(06)|           Kirkland|    1|
|2012/01/05|Montréal(06)|            Montréal|   56|
|2012/01/06|Montréal(06)|         Mont-Royal|    1|
|2012/01/06|Montréal(06)|      Pointe-Claire|    2|
|2012/01/06|Montréal(06)| Dollard-Des Ormeaux|    2|
|2012/01/06|Montréal(06)|            Montréal|   35|
|2012/01/07|Montréal(06)|      Pointe-Claire|    2|
|2012/01/07|Montréal(06)|       Beaconsfield|    1|
|2012/01/07|Montréal(06)|             Dorval|    1|
|2012/01/07|Montréal(06)| Dollard-Des Ormeaux|    2|
|2012/01/07|Montréal(06)|          Hampstead|    1|
|2012/01/07|Montréal(06)|          Westmount|    1|
|2012/01/07|Montréal(06)|     Côte-Saint-Luc|    1|
|2012/01/07|Montréal(06)|            Montréal|   36|
|2012/01/08|Montréal(06)|      Pointe-Claire|    1|
|2012/01/08|Montréal(06)|Sainte-Anne-de-Be…|    1|
|2012/01/08|Montréal(06)|            Montréal|   34|
|2012/01/09|Montréal(06)|      Pointe-Claire|    1|
|2012/01/09|Montréal(06)|            Montréal|   41|
|2012/01/09|Montréal(06)|         Mont-Royal|    1|
|2012/01/10|Montréal(06)|      Pointe-Claire|    2|
|2012/01/10|Montréal(06)| Dollard-Des Ormeaux|    1|
|2012/01/10|Montréal(06)|             Dorval|    1|
|2012/01/10|Montréal(06)|            Montréal|   44|
|2012/01/11|Montréal(06)| Dollard-Des Ormeaux|    1|
|2012/01/11|Montréal(06)|       Beaconsfield|    2|
|2012/01/11|Montréal(06)|            Montréal|   45|
|2012/01/12|Montréal(06)|Sainte-Anne-de-Be…|    1|
|2012/01/12|Montréal(06)|     Côte-Saint-Luc|    2|
```

```
|2012/01/12|Montréal(06)|              Dorval|    3|
+----------+------------+--------------------+-----+
only showing top 50 rows
```

[26]:
```
df = df.withColumn('DATE', F.from_unixtime(F.unix_timestamp('DATE', 'yyyy/MM/
↪dd')))
df = df.withColumn('DATE', df['DATE'].cast(DateType()))
```

[27]:
```
df.dtypes
```

[27]:
```
[('ID', 'string'),
 ('DATE', 'date'),
 ('WEEK_DAY', 'string'),
 ('REG', 'string'),
 ('MUNCP', 'string'),
 ('STREET', 'string'),
 ('TYPE_ACCDN', 'int'),
 ('SURFACE', 'int'),
 ('LIGHT', 'int'),
 ('STR_ASPCT', 'int'),
 ('STR_CONFIG', 'int'),
 ('METEO', 'int'),
 ('GRAVITE', 'string'),
 ('NB_VEH_IMPLIQUES_ACCDN', 'int'),
 ('NB_VICTIMES_TOTAL', 'int'),
 ('NB_MORTS', 'int'),
 ('NB_BLESSES_GRAVES', 'int'),
 ('NB_BLESS_LEGERS', 'int'),
 ('NB_DECES_PIETON', 'int'),
 ('NB_BLESSES_PIETON', 'int'),
 ('NB_VICTIMES_PIETON', 'int'),
 ('NB_DECES_MOTO', 'int'),
 ('NB_BLESSES_MOTO', 'int'),
 ('NB_VICTIMES_MOTO', 'int'),
 ('NB_DECES_VELO', 'int'),
 ('NB_BLESSES_VELO', 'int'),
 ('NB_VICTIMES_VELO', 'int'),
 ('nb_automobile_camion_leger', 'int'),
 ('nb_camionLourd_tractRoutier', 'int'),
 ('nb_outil_equipement', 'int'),
 ('nb_tous_autobus_minibus', 'int'),
 ('nb_bicyclette', 'int'),
 ('nb_cyclomoteur', 'int'),
 ('nb_motocyclette', 'int'),
 ('nb_taxi', 'int'),
 ('nb_urgence', 'int'),
```

```
 ('nb_motoneige', 'int'),
 ('nb_VHR', 'int'),
 ('nb_autres_types', 'int'),
 ('nb_veh_non_precise', 'int')]
```

0.0.6 Checking general tendencies

Meteo for each day

[28]:
```
df.select('DATE', 'METEO').groupby('DATE', 'METEO').count().sort('DATE',␣
↪accending=True).show(50)
```

```
+----------+-----+-----+
|      DATE|METEO|count|
+----------+-----+-----+
|2012-01-01|   14|    8|
|2012-01-01|   17|    2|
|2012-01-01|   99|    1|
|2012-01-01|   11|   12|
|2012-01-01|   13|    1|
|2012-01-01|   12|   16|
|2012-01-02|   11|   16|
|2012-01-02|   14|    2|
|2012-01-02|   12|    6|
|2012-01-03|   17|    1|
|2012-01-03|   12|    2|
|2012-01-03|   11|   37|
|2012-01-04|   17|   14|
|2012-01-04|   11|   20|
|2012-01-04|   12|    9|
|2012-01-05|   12|   12|
|2012-01-05|   11|   40|
|2012-01-05|   17|    7|
|2012-01-06|   11|   11|
|2012-01-06|   12|   18|
|2012-01-06|   17|   10|
|2012-01-06|   13|    1|
|2012-01-07|   14|    8|
|2012-01-07|   11|   12|
|2012-01-07|   17|    6|
|2012-01-07|   12|   15|
|2012-01-07|   13|    1|
|2012-01-07|   19|    3|
|2012-01-08|   17|    1|
|2012-01-08|   99|    1|
|2012-01-08|   12|    9|
|2012-01-08|   11|   25|
```

```
|2012-01-09|   99|    1|
|2012-01-09|   17|    2|
|2012-01-09|   11|   22|
|2012-01-09|   12|   18|
|2012-01-10|   19|    1|
|2012-01-10|   17|    1|
|2012-01-10|   14|    1|
|2012-01-10|   99|    1|
|2012-01-10|   11|   37|
|2012-01-10|   12|    7|
|2012-01-11|   17|    3|
|2012-01-11|   12|   14|
|2012-01-11|   11|   31|
|2012-01-12|   12|    4|
|2012-01-12|   11|    6|
|2012-01-12|   99|    2|
|2012-01-12|   19|    1|
|2012-01-12|   18|   22|
+----------+-----+-----+
only showing top 50 rows
```

Meteo options selected for the collisions in one day

[37]:
```
df[df.DATE == '2012-06-12'].select('DATE', 'METEO').groupby('METEO').count().
↪show()
```

```
+-----+-----+
|METEO|count|
+-----+-----+
|   12|    8|
|   15|    4|
|   11|   20|
|   14|   24|
+-----+-----+
```

[30]:
```
df.count()
```

[30]:
```
128647
```

[31]:
```
df.write.csv(path='data/accidents_new.csv', header="true")
```

[32]:
```
sc.stop()
```