Hw1

October 13, 2022

[1]:
```
import os
import pyspark

conf = pyspark.SparkConf()
conf.set('spark.ui.proxyBase', '/user/' + os.environ['JUPYTERHUB_USER'] + '/
 ↪proxy/4041')
conf.set('spark.sql.repl.eagerEval.enabled', True)
conf.set('spark.driver.memory','4g')

sc = pyspark.SparkContext(conf=conf)
spark = pyspark.SQLContext.getOrCreate(sc)
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).

22/10/13 05:35:53 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable

/opt/conda/envs/bigdata-fall22/lib/python3.7/site-
packages/pyspark/sql/context.py:159: FutureWarning: Deprecated in 3.0.0. Use
SparkSession.builder.getOrCreate() instead.
  FutureWarning,
```

0.0.1 Q1

[2]:
```
BreadBasket = spark.read.option("inferSchema", "true").option("header", "true").
 ↪csv("shared/hw2/BreadBasket_DMS.csv")
```

[3]:
```
from pyspark.sql.functions import avg, col, desc, hour
BreadBasket.select("Item","Transaction","Time",hour(col("Time")))
```

[3]:
```
+-------------+-----------+-------------------+-----------+
|         Item|Transaction|               Time|hour(Time)|
+-------------+-----------+-------------------+-----------+
|        Bread|          1|2022-10-13 09:58:11|          9|
| Scandinavian|          2|2022-10-13 10:05:34|         10|
```

```
| Scandinavian|          2|2022-10-13 10:05:34|        10|
|Hot chocolate|          3|2022-10-13 10:07:57|        10|
|          Jam|          3|2022-10-13 10:07:57|        10|
|      Cookies|          3|2022-10-13 10:07:57|        10|
|       Muffin|          4|2022-10-13 10:08:41|        10|
|       Coffee|          5|2022-10-13 10:13:03|        10|
|       Pastry|          5|2022-10-13 10:13:03|        10|
|        Bread|          5|2022-10-13 10:13:03|        10|
|    Medialuna|          6|2022-10-13 10:16:55|        10|
|       Pastry|          6|2022-10-13 10:16:55|        10|
|       Muffin|          6|2022-10-13 10:16:55|        10|
|    Medialuna|          7|2022-10-13 10:19:12|        10|
|       Pastry|          7|2022-10-13 10:19:12|        10|
|       Coffee|          7|2022-10-13 10:19:12|        10|
|          Tea|          7|2022-10-13 10:19:12|        10|
|       Pastry|          8|2022-10-13 10:20:51|        10|
|        Bread|          8|2022-10-13 10:20:51|        10|
|        Bread|          9|2022-10-13 10:21:59|        10|
+-------------+-----------+-------------------+----------+
only showing top 20 rows
```

[4]:
```
#Using Python
#BreadBasket_filter.where(.groupby("Item").count().sort(desc("count")).show(5)␣
 ↪Have to add hours


#Using SQL
BreadBasket.createOrReplaceTempView("BreadBasket")
maxSql = spark.sql("""
SELECT Item, count(Transaction)
FROM BreadBasket
WHERE date_part('HOUR', Time) >= 9 and date_part('HOUR', Time) < 23
GROUP BY Item
ORDER BY count(Transaction) DESC
LIMIT 5 """)
maxSql.show()
```

```
[Stage 4:>                                                          (0 + 1) / 1]
```

```
+------+-----------------+
|  Item|count(Transaction)|
+------+-----------------+
|Coffee|             5259|
| Bread|             3151|
|   Tea|             1414|
|  Cake|             1017|
|Pastry|              797|
+------+-----------------+
```

[ ]:

[ ]:

0.0.2 Q2

[5]:
```
Restaurants = spark.read.option("inferSchema", "true").option("header", "true").
↪csv("shared/hw2/Restaurants_in_Durham_County_NC.csv",sep=';')
```

[6]:
```
Restaurants
```

[6]:
```
+------+--------------------+--------------------+-----------------+-----------
-+-------------+-----------+-------------+-----------------+------------+-----
-------+-----+--------------------+--------------------+---------+-------------
------+----+--------------+--------------------+---------------+------+-------
--------------+--------------------+
|    ID|        Premise_Name|    Premise_Address1|
Premise_Address2|Premise_City|Premise_State|Premise_Zip|
Premise_Phone|Hours_Of_Operation|Opening_Date|Closing_Date|Seats|
Water|              Sewage|Insp_Freq|      Est_Group_Desc|Risk|Smoking_Allowed|
Type_Description|   Rpt_Area_Desc|Status|Transitional_Type_Desc|
geolocation|
+------+--------------------+--------------------+-----------------+-----------
-+-------------+-----------+-------------+-----------------+------------+-----
-------+-----+--------------------+--------------------+---------+-------------
------+----+--------------+--------------------+---------------+------+-------
--------------+--------------------+
| 56060|    WEST 94TH ST PUB| 4711 HOPE VALLEY RD|          SUITE 6C|
DURHAM|           NC|      27707|(919) 403-0025|              null|  1994-09-01|
null|   60|5 - Municipal/Com…|3 - Municipal/Com…|        4|Full-Service
Rest…|   4|             NO|      1 - Restaurant|    Food Service|ACTIVE|
FOOD|35.9207272, -78.9…|
| 58123|BROOKDALE DURHAM IFS|4434 BEN FRANKLIN…|              null|
DURHAM|           NC|      27704|(919) 479-9966|              null|  2003-10-15|
null|  350|5 - Municipal/Com…|3 - Municipal/Com…|        4|        Nursing
Home|   4|             NO|16 - Institutiona…|    Food Service|ACTIVE|
FOOD|36.0467802, -78.8…|
| 70266|       SMOOTHIE KING|1125 W. NC HWY 54…|              null|
DURHAM|           NC|      27707|(919) 489-7300|              null|  2009-07-09|
null|    7|5 - Municipal/Com…|3 - Municipal/Com…|        2|Fast Food
Restaurant|   2|             NO|      1 - Restaurant|    Food Service|ACTIVE|
FOOD|35.9182655, -78.9…|
| 97837|HAMPTON INN & SUITES|   1542 N GREGSON ST|              null|
```

```
DURHAM|           NC|      27701|(919) 688-8880|              null|  2012-01-09|
null|  100|5 - Municipal/Com…|3 - Municipal/Com…|        2|Full-Service
Rest…|   2|             NO|      1 - Restaurant|    Food Service|ACTIVE|
FOOD|36.0183378, -78.9…|
| 60690|BETTER LIVING CON…|       909 GARCIA ST|              null|
DURHAM|           NC|      27704|(919) 477-5825|              null|  2008-06-02|
null|    6|5 - Municipal/Com…|3 - Municipal/Com…|        1|
null|   0|            N/A|43 - Residential …|Residential Care|ACTIVE|
N/A|36.0556347, -78.9…|
| 60686|     ADVENTURE HOUSE|   4 KIMBROUGH COURT|              null|
DURHAM|           NC|      27703|(919) 957-9097|              null|  2008-06-02|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        1|
null|   0|            N/A|43 - Residential …|Residential Care|ACTIVE|
N/A|35.984012, -78.80…|
| 85252|ANOTHER BEAUTIFUL…|    1309 ANGIER AVE.|              null|
DURHAM|           NC|      27701|(919) 682-5292|              null|  2010-08-23|
null| null|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|             NO|     42 - Child Care|        Day Care|ACTIVE|
N/A|35.9857413, -78.8…|
| 59120|BRIDGES AT SOUTHP…|7304 CALIBRE PARK DR|              null|
DURHAM|           NC|      27707|          null|              null|  1996-04-02|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A|53 - Year-Round S…|  Swimming Pools|ACTIVE|
N/A|35.913596, -78.96…|
| 59124|SHEARTON INN UNIV…|2800 CAMPUS WALK AVE|              null|
DURHAM|           NC|      27705|          null|              null|  1996-04-25|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A|53 - Year-Round S…|  Swimming Pools|ACTIVE|
N/A|36.0111429, -78.9…|
| 59263|     SPA HEALTH CLUB|3419 HILLSBOROUGH RD|              null|
DURHAM|           NC|      27705|          null|              null|  2000-05-30|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A| 55 - Year-Round Spa|  Swimming Pools|ACTIVE|
N/A|36.0184133, -78.9…|
| 58349|KROGER R 381 MEAT…|   3825 S ROXBORO ST|         SUITE 101|
DURHAM|           NC|      27707|(919) 361-0470|              null|  2003-03-03|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        3|Meat and Poultry
…|   3|            N/A|    30 - Meat Market|    Food Service|ACTIVE|
FOOD|35.9495321, -78.9…|
| 58342|LEONE INTERNATION…| 810 FAYETTEVILLE ST|         SUITE 108|
DURHAM|           NC|      27701|(919) 680-0800|              null|  2001-08-17|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        3|
null|   3|             NO|    30 - Meat Market|    Food Service|ACTIVE|
FOOD|35.9852771, -78.8…|
| 57278|  PIZZA HUT DELIVERY|     3808 GUESS ROAD|              null|
DURHAM|           NC|      27705|(919) 477-7377|              null|  1990-07-01|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|Fast Food
```

```
Restaurant|   2|            N/A|     2 - Food Stands|    Food Service|ACTIVE|
FOOD|36.0586094, -78.9…|
| 57190|       CHUBBYS TACOS|    748 NINTH STREET|              null|
DURHAM|           NC|      27705|(919) 286-4499|              null|  2008-04-18|
null|   36|5 - Municipal/Com…|3 - Municipal/Com…|        4|Fast Food
Restaurant|   4|            N/A|      1 - Restaurant|    Food Service|ACTIVE|
FOOD|36.0094173, -78.9…|
| 57585|              SUBWAY|         1058 W CLUB|              null|
DURHAM|           NC|      27701|(919) 286-9681|              null|  1999-01-19|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|Fast Food
Restaurant|   2|             NO|     2 - Food Stands|    Food Service|ACTIVE|
FOOD|36.018111, -78.91…|
| 59132|DUKE UNIVERSITY A…|AQUATIC CENTER WE…|200 FACILITIES CTR|
DURHAM|           NC|      27708|          null|              null|  1996-01-30|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A|53 - Year-Round S…|  Swimming Pools|ACTIVE|
N/A|35.9973617, -78.9…|
| 58444|  DEBORAH'S DAY CARE|       1801 STAGE RD|              null|
DURHAM|           NC|      27703|(919) 596-3636|              null|  1994-08-11|
null|   12|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A|     42 - Child Care|        Day Care|ACTIVE|
N/A|35.946055, -78.84444|
| 59136|EMERALD FOREST AP…|4512 EMERALD FORE…|              null|
DURHAM|           NC|      27713|          null|              null|  1996-01-30|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        2|
null|   0|            N/A|53 - Year-Round S…|  Swimming Pools|ACTIVE|
N/A|35.9039037, -78.9…|
| 59038|    AUBURN SWIM CLUB|1105 PEBBLE CREEK…|              null|
DURHAM|           NC|      27713|          null|              null|  2002-09-18|
null|    0|5 - Municipal/Com…|3 - Municipal/Com…|        1|
null|   0|            N/A|50 - Seasonal Swi…|  Swimming Pools|ACTIVE|
N/A|35.913281, -78.91466|
|178137|DUKE UNIVERSITY W…|    406 CHAPEL DRIVE|      PO BOX 90898|
DURHAM|           NC|      27708|(919) 660-3900|              null|  2016-07-11|
null|  100|5 - Municipal/Com…|3 - Municipal/Com…|        4|Fast Food
Restaurant|   4|            N/A|6 - Edu. Food Ser…|    Food Service|ACTIVE|
FOOD|36.001517, -78.93…|
+------+--------------------+--------------------+------------------+-----------
-+-------------+-----------+-------------+------------------+------------+-----
-------+-----+--------------------+--------------------+---------+-------------
------+----+--------------+--------------------+----------------+------+-------
---------------+--------------------+
only showing top 20 rows
```

[7]:
```
print(Restaurants.select("Id").count())
print("Shows that ID is the primary key as no of rows in df and id count is␣
 ↪same")
```

2472
Shows that ID is the primary key as no of rows in df and id count is same

Method 1

[8]:
```
Restaurants.select("Rpt_Area_Desc").groupby("Rpt_Area_Desc").count().
  ↪sort(desc("count"))
```

[8]:
```
+--------------------+-----+
|       Rpt_Area_Desc|count|
+--------------------+-----+
|        Food Service| 1093|
|      Swimming Pools|  420|
|         Summer Food|  242|
|            Day Care|  173|
|    Residential Care|  154|
|         Mobile Food|  147|
|    School Buildings|   89|
|             Lodging|   62|
|Tattoo Establishm…|   32|
|        Institutions|   30|
|                null|   13|
|      Adult Day Care|    5|
|  Bed&Breakfast Home|    4|
|        Summer Camps|    4|
|   Local Confinement|    2|
|   Bed&Breakfast Inn|    2|
+--------------------+-----+
```

Method 2

[9]:
```
Restaurants.createOrReplaceTempView("Restaurants")
maxSql = spark.sql("""
SELECT Rpt_Area_Desc, count(Id)
FROM Restaurants
GROUP BY Rpt_Area_Desc
ORDER BY count(Id) desc
""")
maxSql.show()
```

```
+--------------------+---------+
|       Rpt_Area_Desc|count(Id)|
+--------------------+---------+
|        Food Service|     1093|
|      Swimming Pools|      420|
|         Summer Food|      242|
|            Day Care|      173|
|    Residential Care|      154|
|         Mobile Food|      147|
```

```
|    School Buildings|       89|
|             Lodging|       62|
|Tattoo Establishm…|       32|
|        Institutions|       30|
|                null|       13|
|      Adult Day Care|        5|
|  Bed&Breakfast Home|        4|
|        Summer Camps|        4|
|   Local Confinement|        2|
|   Bed&Breakfast Inn|        2|
+--------------------+---------+
```

[ ]:

0.0.3 Q3

[10]:
```
Population = spark.read.option("inferSchema", "true").option("header", "true").
  ↪csv("populationbycountry19802010millions.csv")
Population = Population.dropna()
Population
```

```
22/10/13 05:36:19 WARN package: Truncated the string representation of a plan
since it was too large. This behavior can be adjusted by setting
'spark.sql.debug.maxToStringFields'.
```

[10]:
```
+--------------------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+
|                 _c0|     1980|     1981|     1982|     1983|     1984|
1985|     1986|     1987|     1988|     1989|     1990|     1991|     1992|
1993|     1994|     1995|     1996|     1997|     1998|     1999|     2000|
2001|     2002|     2003|     2004|     2005|     2006|     2007|     2008|
2009|     2010|
+--------------------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+
|       North America|320.27638|324.44694|328.62014|332.72487|336.72143|340.7481
1|344.89548|349.07829| 353.2939|357.68457|
362.4468|367.70684|373.29069|378.74233| 383.9166|388.97216|
393.9428|398.97205|403.85585|408.60296|
413.3245|417.83236|422.05268|426.06238|430.26938|434.47232|438.82964|
443.3473|447.67394|451.83698|456.59331|
```

```
|             Bermuda|  0.05473|  0.05491|  0.05517|  0.05551|  0.05585|
0.05618|  0.05651|  0.05683|  0.05717|  0.05749|  0.05778|   0.0581|   0.0587|
0.05924|  0.05975|  0.06029|  0.06087|  0.06145|  0.06198|  0.06251|  0.06306|
0.06361|  0.06418|  0.06476|  0.06534|  0.06591|  0.06644|  0.06692|  0.06739|
0.06784|  0.06827|
|              Canada|  24.5933|     24.9|  25.2019|  25.4563|  25.7018|
25.9416|  26.2038|  26.5497|  26.8948|  27.3793|  27.7906|  28.1179| 28.54489|
28.95334| 29.33081| 29.69053| 30.02632|  30.3056| 30.55166| 30.82026| 31.09956|
31.37674| 31.64096| 31.88931| 32.13476| 32.38638| 32.65668| 32.93596|  33.2127|
33.48721| 33.75974|
|           Greenland|  0.05021|  0.05103|  0.05166|  0.05211|  0.05263|
0.05315|  0.05364|   0.0541|  0.05485|  0.05541|  0.05563|  0.05554|  0.05549|
0.05564|  0.05592|  0.05619|  0.05634|  0.05651|  0.05661|   0.0567|  0.05689|
0.05713|  0.05736|  0.05754|   0.0577|  0.05778|  0.05764|  0.05753|  0.05756|
0.0576|  0.05764|
|              Mexico| 68.34748| 69.96926|  71.6409| 73.36288| 75.08014|
76.76723| 78.44243| 80.12249| 81.78182| 83.36684| 84.91365| 86.48803| 88.11103|
89.74914|  91.3379| 92.88035| 94.39858| 95.89515| 97.32506| 98.61691|
99.92662|101.24696|102.47993|103.71806|104.95959| 106.2029|107.44953|108.70089|
109.9554|111.21179|112.46886|
|Saint Pierre and …|  0.00599|  0.00601|  0.00605|  0.00607|  0.00611|
0.00616|  0.00621|  0.00625|  0.00628|  0.00631|  0.00632|  0.00633|  0.00636|
0.00638|   0.0064|   0.0064|  0.00641|  0.00642|  0.00642|  0.00643|  0.00641|
0.00637|  0.00633|  0.00629|  0.00625|   0.0062|  0.00615|   0.0061|  0.00605|
0.006|  0.00594|
|       United States|227.22468|229.46571|231.66446|233.79199| 235.8249| 237.923
8|240.13289|242.28892|244.49898|246.81923|249.62281|252.98094|256.51422|259.9185
9|263.12582|266.27839|269.39428|272.64693| 275.8541|279.04017|282.17196|285.0815
6|287.80391|290.32642|293.04574|295.75315|298.59321|
301.5799|304.37485|307.00655|310.23286|
|Central & South A…|293.05856|299.43033|305.95253|312.51136|318.87955|325.2270
4|331.82291|338.59859|345.44544|352.20471|358.79973|365.15137|371.43224|
377.7438|384.26984|390.75665|397.13002|403.41352|409.62879|415.63607| 421.4539|4
27.24012|433.05116|438.97976|445.01525|451.05504|457.01699|462.89157|468.73872|4
74.53897|480.01228|
|           Antarctica|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|
| Antigua and Barbuda|  0.06855|  0.06826|  0.06801|  0.06562|  0.06447|
0.0644|   0.0644|  0.06442|  0.06443|  0.06439|  0.06416|  0.06403|  0.06518|
0.06633|  0.06749|  0.06871|  0.07006|  0.07145|  0.07278|  0.07402|  0.07535|
0.07673|  0.07794|  0.07907|  0.08019|  0.08128|  0.08234|  0.08343|  0.08452|
0.08563|  0.08675|
|           Argentina|  28.3698| 28.84806| 29.32988| 29.79355| 30.23064|
30.67176| 31.14499| 31.62462| 32.09932| 32.57162| 33.03558| 33.50441| 33.96685|
```

34.40715| 34.84689| 35.27384| 35.68296| 36.10216| 36.51792| 36.92342| 37.33565|
37.69417| 37.99945| 38.33688| 38.74183| 39.18126| 39.61443| 40.04882| 40.482|
40.91358| 41.3432|
| Aruba| --| --| --| --| --|
--| 0.0598| 0.05918| 0.0595| 0.06069| 0.06303| 0.0663| 0.06948|
0.07407| 0.07785| 0.07996| 0.08307| 0.08621| 0.0882| 0.08926| 0.09|
0.09097| 0.09217| 0.09372| 0.09546| 0.09698| 0.0985| 0.10002| 0.10154|
0.10307| 0.10459|
| Bahamas, The| 0.20976| 0.21345| 0.21713| 0.22086| 0.22462|
0.2282| 0.23143| 0.23448| 0.23771| 0.24124| 0.24513| 0.24931| 0.25356|
0.25766| 0.26151| 0.26518| 0.26888| 0.27256| 0.27599| 0.27931| 0.28259|
0.28569| 0.28858| 0.29135| 0.29406| 0.29671| 0.29929| 0.30197| 0.30473|
0.30755| 0.31043|
| Barbados| 0.25197| 0.25236| 0.25348| 0.25485| 0.25611|
0.25725| 0.25827| 0.25912| 0.25995| 0.26109| 0.26226| 0.26334| 0.2646|
0.2657| 0.26663| 0.26767| 0.26881| 0.27006| 0.27129| 0.2725| 0.27368|
0.27491| 0.27622| 0.27755| 0.27882| 0.28004| 0.28121| 0.28236| 0.2835|
0.28459| 0.28565|
| Belize| 0.14442| 0.14921| 0.1533| 0.15685| 0.16081|
0.16556| 0.17124| 0.17635| 0.1814| 0.18643| 0.19087| 0.19575| 0.20082|
0.20609| 0.21155| 0.21717| 0.22297| 0.22895| 0.23513| 0.24148| 0.248|
0.25464| 0.2613| 0.26796| 0.27462| 0.28129| 0.28795| 0.29461| 0.30127|
0.3079| 0.31452|
| Bolivia| 5.4413| 5.54522| 5.64222| 5.73743| 5.83429|
5.93494| 6.04135| 6.15637| 6.28316| 6.42314| 6.5739| 6.73148| 6.89345|
7.05434| 7.21481| 7.37487| 7.5344| 7.69515| 7.8589| 8.02556| 8.1951|
8.36745| 8.54249| 8.71906| 8.89597| 9.07294| 9.24971| 9.42594| 9.60126|
9.77525| 9.94742|
| Brazil|123.01963|125.99213|129.02765|131.96012|134.69947|137.3819
8|140.19628|143.02654|145.87275|148.65864|151.17006|153.58396|156.03206|158.5120
5|161.01706|163.54428|166.08586|168.63874|171.20116|173.76387|176.31962|178.8696
6|181.41759|183.95992|
186.4886|188.99308|191.46901|193.91858|196.34259|198.73927|201.10333|
| Cayman Islands| 0.01708| 0.0179| 0.01852| 0.01909| 0.02002|
0.02085| 0.02144| 0.02207| 0.0245| 0.02507| 0.02636| 0.02751| 0.02868|
0.03001| 0.0313| 0.03249| 0.03368| 0.03487| 0.03606| 0.03725| 0.03844|
0.03962| 0.0408| 0.04199| 0.04316| 0.04434| 0.04552| 0.04669| 0.04786|
0.04904| 0.05021|
| Chile| 11.09372| 11.2823| 11.48711| 11.68662| 11.87977|
12.0678| 12.261| 12.46452| 12.67869| 12.90232| 13.12892| 13.35367| 13.57416|
13.78943| 14.00122| 14.20661| 14.40541| 14.60109| 14.79216| 14.97655| 15.15574|
15.33189| 15.50394| 15.67191| 15.83563| 15.99504| 16.15084| 16.30385| 16.45414|
16.60171| 16.74649|
| Colombia| 26.63129| 27.21489| 27.82604| 28.45499| 29.09546|
29.74762| 30.41039| 31.0853| 31.77087| 32.46085| 33.14725| 33.83221| 34.52032|
35.2056| 35.88762| 36.53183| 37.09791| 37.61994| 38.13259| 38.56386| 38.91035|
39.31245| 39.80495| 40.35102| 40.92215| 41.48778| 42.04625| 42.59732| 43.14111|

```
43.67737| 44.20529|
+-------------------+--------+--------+--------+--------+--------+--------+-------
-+--------+--------+--------+--------+--------+--------+--------+-------
-+--------+--------+--------+--------+--------+--------+--------+-------
-+--------+--------+--------+--------+--------+--------+--------+-------
-+---------+
only showing top 20 rows
```

Method 1

[11]:
```
from pyspark.sql.functions import udf,lit, cast, col
from pyspark.sql.types import DoubleType
from pyspark.sql.types import IntegerType

data = []


for k in Population.columns[1:]:
    Population = Population.withColumn(k,col(k).cast('double'))

for i,j in zip(Population.columns[1:], Population.columns[2:]):
    Population = Population.
 ↪withColumn("Percentage_increase("+i+","+j+")",((col(j)-col(i))*100)/col(i))
```

[12]:
```
Population = Population.dropna()
```

[13]:
```
from pyspark.sql.functions import udf,lit, cast, col, asc, desc
for i,j in zip(Population.columns[1:], Population.columns[2:]):
    if len(i)==4 and len(j)==4:
        print(j)
        col_name = "Percentage_increase("+i+","+j+")"
        Population = Population.sort(desc(col_name))
        print("MAX",Population.collect()[0]["_c0"], Population.
 ↪collect()[0][col_name])
        Population = Population.sort(asc(col_name))
        print("MIN",Population.collect()[0]["_c0"], Population.
 ↪collect()[0][col_name])
```

```
1981
MAX Western Sahara 12.133182844243787
MIN Afghanistan -9.106330931425992
1982
MAX Western Sahara 11.115105327485804
MIN Afghanistan -8.017227257036874
1983
MAX French Guiana 14.285714285714278
MIN Antigua and Barbuda -3.5141890898397343
```

```
1984
MAX Qatar 10.964057316781224
MIN Antigua and Barbuda -1.7525144772935055
1985
MAX French Guiana 12.499999999999995
MIN Cook Islands -1.4092446448703508
1986
MAX Qatar 8.771732719152874
MIN Netherlands Antilles -24.58781655279631
1987
MAX French Guiana 11.111111111111121
MIN Saint Helena -21.299638989169676
1988
MAX Cayman Islands 11.010421386497516
MIN Mozambique -2.883631837516533
1989
MAX United Arab Emirates 6.119858265290403
MIN Somalia -2.1964965331028314
1990
MAX Djibouti 12.82404791501865
MIN Liberia -12.816300240117076
1991
MAX Jordan 11.273939557210026
MIN Kuwait -55.4531619095637
1992
MAX Kuwait 48.63343882962002
MIN Somalia -5.387440289087448
1993
MAX Afghanistan 13.224594754698687
MIN Bhutan -4.150184489578821
1994
MAX Afghanistan 8.727661664211226
MIN Rwanda -14.363511428676736
1995
MAX Burundi 7.222488903730302
MIN Rwanda -15.871881307134093
1996
MAX Rwanda 19.61417728550077
MIN Montserrat -22.590068159688407
1997
MAX Falkland Islands (Islas Malvinas) 21.49999999999999
MIN Montserrat -25.157232704402517
1998
MAX Liberia 12.01744976042338
MIN Montserrat -43.193277310924366
1999
MAX Falkland Islands (Islas Malvinas) 7.692307692307697
MIN Cook Islands -2.9919447640966608
```

```
2000
MAX Montserrat 16.863905325443792
MIN Cook Islands -3.2621589561091247
2001
MAX Montserrat 7.34177215189872
MIN Cook Islands -3.55610055180871
2002
MAX Montserrat 13.4433962264151
MIN Cook Islands -3.6872218690400547
2003
MAX Afghanistan 5.803891762260126
MIN Montserrat -6.652806652806653
2004
MAX Montserrat 10.467706013363028
MIN Djibouti -4.830771012478634
2005
MAX Liberia 4.7976709085316545
MIN Montserrat -8.669354838709674
2006
MAX Jordan 7.088496587486171
MIN Nauru -4.39560439560439
2007
MAX Jordan 6.764378108744186
MIN Nauru -4.702194357366778
2008
MAX Montserrat 12.638580931263864
MIN Cook Islands -3.3096926713948007
2009
MAX Liberia 4.157111008408977
MIN Cook Islands -3.259983700081494
2010
MAX Niger 3.737166190281749
MIN Cook Islands -3.2013479359730423
```

[ ]:

[ ]:

[ ]:

Method 2

[14]:
```
Population = spark.read.option("inferSchema", "true").option("header", "true").
  ↪csv("populationbycountry19802010millions.csv")
Population.dropna()
```

[14]:
```
+--------------------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+
|                 _c0|     1980|     1981|     1982|     1983|     1984|
1985|     1986|     1987|     1988|     1989|     1990|     1991|     1992|
1993|     1994|     1995|     1996|     1997|     1998|     1999|     2000|
2001|     2002|     2003|     2004|     2005|     2006|     2007|     2008|
2009|     2010|
+--------------------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+
|       North America|320.27638|324.44694|328.62014|332.72487|336.72143|340.7481
1|344.89548|349.07829| 353.2939|357.68457|
362.4468|367.70684|373.29069|378.74233| 383.9166|388.97216|
393.9428|398.97205|403.85585|408.60296|
413.3245|417.83236|422.05268|426.06238|430.26938|434.47232|438.82964|
443.3473|447.67394|451.83698|456.59331|
|             Bermuda|  0.05473|  0.05491|  0.05517|  0.05551|  0.05585|
0.05618|  0.05651|  0.05683|  0.05717|  0.05749|  0.05778|   0.0581|   0.0587|
0.05924|  0.05975|  0.06029|  0.06087|  0.06145|  0.06198|  0.06251|  0.06306|
0.06361|  0.06418|  0.06476|  0.06534|  0.06591|  0.06644|  0.06692|  0.06739|
0.06784|  0.06827|
|              Canada|  24.5933|     24.9|  25.2019|  25.4563|  25.7018|
25.9416|  26.2038|  26.5497|  26.8948|  27.3793|  27.7906|  28.1179| 28.54489|
28.95334| 29.33081| 29.69053| 30.02632|  30.3056| 30.55166| 30.82026| 31.09956|
31.37674| 31.64096| 31.88931| 32.13476| 32.38638| 32.65668| 32.93596|  33.2127|
33.48721| 33.75974|
|           Greenland|  0.05021|  0.05103|  0.05166|  0.05211|  0.05263|
0.05315|  0.05364|   0.0541|  0.05485|  0.05541|  0.05563|  0.05554|  0.05549|
0.05564|  0.05592|  0.05619|  0.05634|  0.05651|  0.05661|   0.0567|  0.05689|
0.05713|  0.05736|  0.05754|   0.0577|  0.05778|  0.05764|  0.05753|  0.05756|
0.0576|  0.05764|
|              Mexico| 68.34748| 69.96926|  71.6409| 73.36288| 75.08014|
76.76723| 78.44243| 80.12249| 81.78182| 83.36684| 84.91365| 86.48803| 88.11103|
89.74914|  91.3379| 92.88035| 94.39858| 95.89515| 97.32506| 98.61691|
99.92662|101.24696|102.47993|103.71806|104.95959| 106.2029|107.44953|108.70089|
109.9554|111.21179|112.46886|
|Saint Pierre and …|  0.00599|  0.00601|  0.00605|  0.00607|  0.00611|
0.00616|  0.00621|  0.00625|  0.00628|  0.00631|  0.00632|  0.00633|  0.00636|
0.00638|   0.0064|   0.0064|  0.00641|  0.00642|  0.00642|  0.00643|  0.00641|
0.00637|  0.00633|  0.00629|  0.00625|   0.0062|  0.00615|   0.0061|  0.00605|
0.006|  0.00594|
|       United States|227.22468|229.46571|231.66446|233.79199| 235.8249| 237.923
```

```
8|240.13289|242.28892|244.49898|246.81923|249.62281|252.98094|256.51422|259.9185
9|263.12582|266.27839|269.39428|272.64693| 275.8541|279.04017|282.17196|285.0815
6|287.80391|290.32642|293.04574|295.75315|298.59321|
301.5799|304.37485|307.00655|310.23286|
|Central & South A…|293.05856|299.43033|305.95253|312.51136|318.87955|325.2270
4|331.82291|338.59859|345.44544|352.20471|358.79973|365.15137|371.43224|
377.7438|384.26984|390.75665|397.13002|403.41352|409.62879|415.63607| 421.4539|4
27.24012|433.05116|438.97976|445.01525|451.05504|457.01699|462.89157|468.73872|4
74.53897|480.01228|
|          Antarctica|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|       NA|       NA|       NA|       NA|       NA|       NA|
NA|       NA|
| Antigua and Barbuda|  0.06855|  0.06826|  0.06801|  0.06562|  0.06447|
0.0644|   0.0644|  0.06442|  0.06443|  0.06439|  0.06416|  0.06403|  0.06518|
0.06633|  0.06749|  0.06871|  0.07006|  0.07145|  0.07278|  0.07402|  0.07535|
0.07673|  0.07794|  0.07907|  0.08019|  0.08128|  0.08234|  0.08343|  0.08452|
0.08563|  0.08675|
|           Argentina|  28.3698| 28.84806| 29.32988| 29.79355| 30.23064|
30.67176| 31.14499| 31.62462| 32.09932| 32.57162| 33.03558| 33.50441| 33.96685|
34.40715| 34.84689| 35.27384| 35.68296| 36.10216| 36.51792| 36.92342| 37.33565|
37.69417| 37.99945| 38.33688| 38.74183| 39.18126| 39.61443| 40.04882|   40.482|
40.91358|  41.3432|
|               Aruba|       --|       --|       --|       --|       --|
--|   0.0598|  0.05918|   0.0595|  0.06069|  0.06303|   0.0663|  0.06948|
0.07407|  0.07785|  0.07996|  0.08307|  0.08621|   0.0882|  0.08926|     0.09|
0.09097|  0.09217|  0.09372|  0.09546|  0.09698|   0.0985|  0.10002|  0.10154|
0.10307|  0.10459|
|        Bahamas, The|  0.20976|  0.21345|  0.21713|  0.22086|  0.22462|
0.2282|  0.23143|  0.23448|  0.23771|  0.24124|  0.24513|  0.24931|  0.25356|
0.25766|  0.26151|  0.26518|  0.26888|  0.27256|  0.27599|  0.27931|  0.28259|
0.28569|  0.28858|  0.29135|  0.29406|  0.29671|  0.29929|  0.30197|  0.30473|
0.30755|  0.31043|
|           Barbados|  0.25197|  0.25236|  0.25348|  0.25485|  0.25611|
0.25725|  0.25827|  0.25912|  0.25995|  0.26109|  0.26226|  0.26334|   0.2646|
0.2657|  0.26663|  0.26767|  0.26881|  0.27006|  0.27129|   0.2725|  0.27368|
0.27491|  0.27622|  0.27755|  0.27882|  0.28004|  0.28121|  0.28236|   0.2835|
0.28459|  0.28565|
|             Belize|  0.14442|  0.14921|   0.1533|  0.15685|  0.16081|
0.16556|  0.17124|  0.17635|   0.1814|  0.18643|  0.19087|  0.19575|  0.20082|
0.20609|  0.21155|  0.21717|  0.22297|  0.22895|  0.23513|  0.24148|    0.248|
0.25464|   0.2613|  0.26796|  0.27462|  0.28129|  0.28795|  0.29461|  0.30127|
0.3079|  0.31452|
|            Bolivia|   5.4413|  5.54522|  5.64222|  5.73743|  5.83429|
5.93494|  6.04135|  6.15637|  6.28316|  6.42314|   6.5739|  6.73148|  6.89345|
7.05434|  7.21481|  7.37487|   7.5344|  7.69515|   7.8589|  8.02556|   8.1951|
```

```
8.36745|  8.54249|  8.71906|  8.89597|  9.07294|  9.24971|  9.42594|  9.60126|
9.77525|  9.94742|
|             Brazil|123.01963|125.99213|129.02765|131.96012|134.69947|137.3819
8|140.19628|143.02654|145.87275|148.65864|151.17006|153.58396|156.03206|158.5120
5|161.01706|163.54428|166.08586|168.63874|171.20116|173.76387|176.31962|178.8696
6|181.41759|183.95992|
186.4886|188.99308|191.46901|193.91858|196.34259|198.73927|201.10333|
|      Cayman Islands|  0.01708|   0.0179|  0.01852|  0.01909|  0.02002|
0.02085|  0.02144|  0.02207|   0.0245|  0.02507|  0.02636|  0.02751|  0.02868|
0.03001|   0.0313|  0.03249|  0.03368|  0.03487|  0.03606|  0.03725|  0.03844|
0.03962|   0.0408|  0.04199|  0.04316|  0.04434|  0.04552|  0.04669|  0.04786|
0.04904|  0.05021|
|              Chile| 11.09372|  11.2823| 11.48711| 11.68662| 11.87977|
12.0678|   12.261| 12.46452| 12.67869| 12.90232| 13.12892| 13.35367| 13.57416|
13.78943| 14.00122| 14.20661| 14.40541| 14.60109| 14.79216| 14.97655| 15.15574|
15.33189| 15.50394| 15.67191| 15.83563| 15.99504| 16.15084| 16.30385| 16.45414|
16.60171| 16.74649|
|           Colombia| 26.63129| 27.21489| 27.82604| 28.45499| 29.09546|
29.74762| 30.41039|  31.0853| 31.77087| 32.46085| 33.14725| 33.83221| 34.52032|
35.2056| 35.88762| 36.53183| 37.09791| 37.61994| 38.13259| 38.56386| 38.91035|
39.31245| 39.80495| 40.35102| 40.92215| 41.48778| 42.04625| 42.59732| 43.14111|
43.67737| 44.20529|
+--------------------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+---------+---------+---------+---------+---------+---------+--------
-+---------+
only showing top 20 rows
```

[15]:
```
from pyspark.sql.functions import expr
expression = "stack(31, '1980', `1980`, '1981', `1981`, '1982', `1982`, '1983',␣
 ↪`1983`, '1984', `1984`, '1985', `1985`, '1986', `1986`, '1987', `1987`,␣
 ↪'1988', `1988`, '1989', `1989`, '1990', `1990`, '1991', `1991`, '1992',␣
 ↪`1992`, '1993', `1993`, '1994', `1994`, '1995', `1995`, '1996', `1996`,␣
 ↪'1997', `1997`, '1998', `1998`, '1999', `1999`, '2000', `2000`, '2001',␣
 ↪`2001`, '2002', `2002`, '2003', `2003`, '2004', `2004`, '2005', `2005`,␣
 ↪'2006', `2006`, '2007', `2007`, '2008', `2008`, '2009', `2009`, '2010',␣
 ↪`2010`) as (Year, Population)"
Population_unpivot = Population.select("_c0", expr(expression))
Population_unpivot.show()
```

```
+-------------+----+----------+
|          _c0|Year|Population|
+-------------+----+----------+
|North America|1980| 320.27638|
|North America|1981| 324.44694|
|North America|1982| 328.62014|
```

```
|North America|1983| 332.72487|
|North America|1984| 336.72143|
|North America|1985| 340.74811|
|North America|1986| 344.89548|
|North America|1987| 349.07829|
|North America|1988|  353.2939|
|North America|1989| 357.68457|
|North America|1990|  362.4468|
|North America|1991| 367.70684|
|North America|1992| 373.29069|
|North America|1993| 378.74233|
|North America|1994|  383.9166|
|North America|1995| 388.97216|
|North America|1996|  393.9428|
|North America|1997| 398.97205|
|North America|1998| 403.85585|
|North America|1999| 408.60296|
+-------------+----+----------+
only showing top 20 rows
```

[16]:
```
from pyspark.sql.window import Window
windowSpec  = Window.partitionBy("_c0").orderBy("Year")
```

[17]:
```
from pyspark.sql.functions import lag
Population_unpivot_lag = Population_unpivot.
  ↪withColumn('LastYear_Population',lag(Population_unpivot['Population']).
  ↪over(windowSpec))

Population_unpivot_lag.dropna()
```

[17]:
```
+-----------+----+----------+------------------+
|        _c0|Year|Population|LastYear_Population|
+-----------+----+----------+------------------+
|Afghanistan|1981|  13.67368|           15.0436|
|Afghanistan|1982|  12.57743|          13.67368|
|Afghanistan|1983|  12.43058|          12.57743|
|Afghanistan|1984|  12.75384|          12.43058|
|Afghanistan|1985|  13.09371|          12.75384|
|Afghanistan|1986|  13.08496|          13.09371|
|Afghanistan|1987|  12.99949|          13.08496|
|Afghanistan|1988|  13.09382|          12.99949|
|Afghanistan|1989|  13.40714|          13.09382|
|Afghanistan|1990|  13.44937|          13.40714|
|Afghanistan|1991|  13.52717|          13.44937|
|Afghanistan|1992|   14.8552|          13.52717|
|Afghanistan|1993|  16.81974|           14.8552|
```

```
|Afghanistan|1994|  18.28771|           16.81974|
|Afghanistan|1995|  19.22594|           18.28771|
|Afghanistan|1996|  19.86465|           19.22594|
|Afghanistan|1997|  20.49022|           19.86465|
|Afghanistan|1998|  21.13212|           20.49022|
|Afghanistan|1999|  21.82157|           21.13212|
|Afghanistan|2000|  22.02095|           21.82157|
+-----------+----+----------+-------------------+
only showing top 20 rows
```

[18]:
```
from pyspark.sql.types import FloatType

def percentage_increase(a,b):
    return ((b-a)*100)/a


percentage_increase_udf = udf(percentage_increase, FloatType())
Population_unpivot_percent_growth = Population_unpivot_lag.
 ↪withColumn("Percentage_increase",  ((Population_unpivot_lag['Population'] -␣
 ↪Population_unpivot_lag['LastYear_Population']) /␣
 ↪Population_unpivot_lag['LastYear_Population'])*100)
```

[19]:
```
Population_unpivot_percent_growth = Population_unpivot_percent_growth.dropna()
```

[20]:
```
from pyspark.sql.functions import col, dense_rank, desc, asc

max_increase_window = Window.partitionBy("Year").
 ↪orderBy(desc("Percentage_increase"))
max_increase = Population_unpivot_percent_growth.withColumn("Max", dense_rank().
 ↪over(max_increase_window)).where(col("Max") == 1)


max_increase = max_increase.withColumn("Type", lit("Max"))

min_increase_window = Window.partitionBy("Year").
 ↪orderBy(asc("Percentage_increase"))
min_increase = Population_unpivot_percent_growth.withColumn("Min", dense_rank().
 ↪over(min_increase_window)).where(col("Min") == 1)
min_increase = min_increase.withColumn("Type", lit("Min"))

min_increase.union(max_increase).select("_c0", "Year","Type",␣
 ↪"Percentage_increase").orderBy("Year").show()
min_increase.union(max_increase).select("_c0", "Year","Type",␣
 ↪"Percentage_increase").orderBy(asc("Year"),desc("Type")).show()
```

```
+-------------------+----+----+-------------------+
```

```
|                 _c0|Year|Type|Percentage_increase|
+--------------------+----+----+------------------+
|      Western Sahara|1981| Max| 12.133182844243787|
|         Afghanistan|1981| Min| -9.106330931425992|
|      Western Sahara|1982| Max| 11.115105327485802|
|         Afghanistan|1982| Min| -8.017227257036874|
| Antigua and Barbuda|1983| Min|-3.5141890898397343|
|       French Guiana|1983| Max| 14.285714285714276|
|               Qatar|1984| Max| 10.964057316781224|
| Antigua and Barbuda|1984| Min|-1.7525144772935055|
|       French Guiana|1985| Max| 12.499999999999993|
|        Cook Islands|1985| Min|-1.4092446448703508|
|               Qatar|1986| Max|  8.771732719152874|
|Netherlands Antilles|1986| Min| -24.58781655279631|
|       French Guiana|1987| Max| 11.111111111111121|
|        Saint Helena|1987| Min|-21.299638989169676|
|          Mozambique|1988| Min|-2.8836318375165324|
|      Cayman Islands|1988| Max| 11.010421386497516|
|             Somalia|1989| Min|-2.1964965331028314|
|United Arab Emirates|1989| Max|  6.119858265290403|
|            Djibouti|1990| Max|  12.82404791501865|
|             Liberia|1990| Min|-12.816300240117076|
+--------------------+----+----+------------------+
only showing top 20 rows

+--------------------+----+----+------------------+
|                 _c0|Year|Type|Percentage_increase|
+--------------------+----+----+------------------+
|         Afghanistan|1981| Min| -9.106330931425992|
|      Western Sahara|1981| Max| 12.133182844243787|
|         Afghanistan|1982| Min| -8.017227257036874|
|      Western Sahara|1982| Max| 11.115105327485802|
| Antigua and Barbuda|1983| Min|-3.5141890898397343|
|       French Guiana|1983| Max| 14.285714285714276|
| Antigua and Barbuda|1984| Min|-1.7525144772935055|
|               Qatar|1984| Max| 10.964057316781224|
|        Cook Islands|1985| Min|-1.4092446448703508|
|       French Guiana|1985| Max| 12.499999999999993|
|Netherlands Antilles|1986| Min| -24.58781655279631|
|               Qatar|1986| Max|  8.771732719152874|
|        Saint Helena|1987| Min|-21.299638989169676|
|       French Guiana|1987| Max| 11.111111111111121|
|          Mozambique|1988| Min|-2.8836318375165324|
|      Cayman Islands|1988| Max| 11.010421386497516|
|             Somalia|1989| Min|-2.1964965331028314|
|United Arab Emirates|1989| Max|  6.119858265290403|
|             Liberia|1990| Min|-12.816300240117076|
|            Djibouti|1990| Max|  12.82404791501865|
```

```
+--------------------+----+----+------------------+
only showing top 20 rows
```

[ ]:

[ ]:

Q4

[21]:
```
df = spark.read.text("shared/hw2/romeo-juliet-pg1777.txt")
```

[22]:
```
df
```

[22]:
```
+--------------------+
|               value|
+--------------------+
|                    |
|This Etext file i…|
|cooperation with …|
|Future and Shakes…|
|Etexts that are N…|
|                    |
|*This Etext has c…|
|                    |
|<<THIS ELECTRONIC…|
|SHAKESPEARE IS CO…|
|PROVIDED BY PROJE…|
|MACHINE READABLE …|
|(1) ARE FOR YOUR …|
|DISTRIBUTED OR US…|
|DISTRIBUTION INCL…|
|TIME OR FOR MEMBE…|
|                    |
|*Project Gutenber…|
|in the presentati…|
|for your reading …|
+--------------------+
only showing top 20 rows
```

[23]:
```
from pyspark.sql.functions import regexp_replace, regexp_extract, lower,␣
↪explode, split, col, desc
df = df.withColumn("all_lower",lower(col('value')))
```

[24]:
```
df1 = df.withColumn("parsed",regexp_replace(col("all_lower"),'[^a-z0-9]',' '))
```

[25]:
```
df1.select(explode(split(col("parsed"),' ')).alias("exploding")).
  ↪filter(col("exploding")!='').groupby("exploding").count().sort(desc("count"))
```

[25]:
```
+---------+-----+
|exploding|count|
+---------+-----+
|      and|  780|
|      the|  749|
|        i|  658|
|       to|  616|
|        a|  485|
|       of|  474|
|       is|  389|
|     that|  373|
|       my|  360|
|       in|  329|
|      you|  327|
|        s|  294|
|     thou|  278|
|      not|  275|
|      for|  268|
|     with|  268|
|       me|  265|
|     this|  258|
|       it|  237|
|        d|  236|
+---------+-----+
only showing top 20 rows
```

0.0.4 Q5

[26]:
```
json_reader = spark.read.format("json").load("shared/hw2/
  ↪durham-nc-foreclosure-2006-2016.json")
json_reader
```

[26]:
```
+--------------------+--------------------+--------------------+----------------
----+--------------------+
|           datasetid|              fields|            geometry|
record_timestamp|            recordid|
+--------------------+--------------------+--------------------+----------------
----+--------------------+
|foreclosure-2006-…|{217 E CORPORATIO…|{[-78.8922549,
36…|2017-03-06T12:41:…|629979c85b1cc68c1…|
|foreclosure-2006-…|{401 N QUEEN ST, …|{[-78.895396,
35…|2017-03-06T12:41:…|e3cce8bbc3c9b804c…|
|foreclosure-2006-…|{403 N QUEEN ST, …|{[-78.8950321,
35…|2017-03-06T12:41:…|311559ebfeffe7ebc…|
```

```
|foreclosure-2006-…|{918 GILBERT ST, …|{[-78.8873774,
35…|2017-03-06T12:41:…|7ec0761bd385bab8a…|
|foreclosure-2006-…|{721 LIBERTY ST, …|{[-78.888343,
35…|2017-03-06T12:41:…|c81ae2921ffca8125…|
|foreclosure-2006-…|{729 HOPKINS ST, …|{[-78.888092,
35…|2017-03-06T12:41:…|ae17ea44c5918fd2d…|
|foreclosure-2006-…|{1302 E MAIN ST, …|{[-78.886681,
35…|2017-03-06T12:41:…|33bcfab5aa69ed55e…|
|foreclosure-2006-…|{402 CLAY ST, [35…|{[-78.8806365,
35…|2017-03-06T12:41:…|0322e07438201c6cf…|
|foreclosure-2006-…|{1516 LATHROP ST,…|{[-78.874621,
35…|2017-03-06T12:41:…|f88a039e24c4182df…|
|foreclosure-2006-…|{2604 E MAIN ST, …|{[-78.869642,
35…|2017-03-06T12:41:…|54b29375c19aff597…|
|foreclosure-2006-…|{209 NELSON ST, […|{[-78.9041979,
35…|2017-03-06T12:41:…|1644733ddb6c7f3b1…|
|foreclosure-2006-…|{2721 ATLANTIC ST…|{[-78.9060606,
35…|2017-03-06T12:41:…|97d53f9a363445bd3…|
|foreclosure-2006-…|{518 RED OAK AVE,…|{[-78.903483,
35…|2017-03-06T12:41:…|64e85ba0cbf6272c1…|
|foreclosure-2006-…|{ROXBORO ST, [36…|{[-78.8960511,
36…|2017-03-06T12:41:…|be817e36d57bd79b5…|
|foreclosure-2006-…|{1420 WABASH ST, …|{[-78.890249,
35…|2017-03-06T12:41:…|65d1396789609f295…|
|foreclosure-2006-…|{500 POTTER ST, […|{[-78.889404,
35…|2017-03-06T12:41:…|e4a0a58d755c8a9eb…|
|foreclosure-2006-…|{2820 ANGIER AVE,…|{[-78.8692657,
35…|2017-03-06T12:41:…|a778f7cda928028c8…|
|foreclosure-2006-…|{2822 ANGIER AVE,…|{[-78.868701,
35…|2017-03-06T12:41:…|aad612d17ffbddb4e…|
|foreclosure-2006-…|{515 BACON ST, [3…|{[-78.882454,
35…|2017-03-06T12:41:…|e4c35ab6d0af22b85…|
|foreclosure-2006-…|{418 SOWELL ST, […|{[-78.8873413,
35…|2017-03-06T12:41:…|65606f3b8531c5e7a…|
+-------------------+-------------------+--------------------+---------------
----+--------------------+
only showing top 20 rows
```

[27]:
```
Restaurants = spark.read.option("inferSchema", "true").option("header", "true").
  ↪csv("shared/hw2/Restaurants_in_Durham_County_NC.csv",sep=';')
Restaurants.columns
```

[27]:
```
['ID',
 'Premise_Name',
 'Premise_Address1',
 'Premise_Address2',
 'Premise_City',
```

```
 'Premise_State',
 'Premise_Zip',
 'Premise_Phone',
 'Hours_Of_Operation',
 'Opening_Date',
 'Closing_Date',
 'Seats',
 'Water',
 'Sewage',
 'Insp_Freq',
 'Est_Group_Desc',
 'Risk',
 'Smoking_Allowed',
 'Type_Description',
 'Rpt_Area_Desc',
 'Status',
 'Transitional_Type_Desc',
 'geolocation']
```

[28]:
```
Restaurant_active = Restaurants.filter((Restaurants.Status=="ACTIVE") &␣
 ↪(Restaurants.Rpt_Area_Desc=="Food Service"))
```

[29]:
```
Restaurant_active.groupby(col("Status")).count()
```

[29]:
```
+------+-----+
|Status|count|
+------+-----+
|ACTIVE| 1093|
+------+-----+
```

[30]:
```
Restaurant_subset = Restaurant_active.select("id","geolocation","Premise_City")
Restaurant_subset = Restaurant_subset.na.drop(subset=["geolocation"])
Restaurant_two = Restaurant_subset.select("*")
Restaurant_one = Restaurant_subset.
 ↪withColumnRenamed("geolocation","geolocation_x")
Restaurant_one = Restaurant_one.withColumnRenamed("id","id_x")
Restaurant_one = Restaurant_one.
 ↪withColumnRenamed("Premise_City","Premise_City_x")
```

[31]:
```
Restaurant_two
```

[31]:
```
+------+--------------------+------------+
|    id|         geolocation|Premise_City|
+------+--------------------+------------+
| 56060|35.9207272, -78.9…|      DURHAM|
| 58123|36.0467802, -78.8…|      DURHAM|
| 70266|35.9182655, -78.9…|      DURHAM|
```

```
| 97837|36.0183378, -78.9…|      DURHAM|
| 58349|35.9495321, -78.9…|      DURHAM|
| 58342|35.9852771, -78.8…|      DURHAM|
| 57278|36.0586094, -78.9…|      DURHAM|
| 57190|36.0094173, -78.9…|      DURHAM|
| 57585|36.018111, -78.91…|      DURHAM|
|178137|36.001517, -78.93…|      DURHAM|
|181419|35.9550003, -78.9…|      DURHAM|
|179950|35.9950533, -78.9…|      DURHAM|
| 57048|36.0507859, -78.9…|      DURHAM|
|179953|35.9950533, -78.9…|      DURHAM|
|180462|35.9953688, -78.9…|      DURHAM|
|188410|36.1027227, -78.8…|      DURHAM|
|178997|35.9965963, -78.9…|      DURHAM|
|179334|35.9965963, -78.9…|      DURHAM|
| 58147|35.9094306, -79.0…| CHAPEL HILL|
|189532|35.9205016, -78.8…| MORRISVILLE|
+------+--------------------+------------+
only showing top 20 rows
```

[32]:
```
Restaurant_one = Restaurant_one.join(Restaurant_two, ((Restaurant_one["id_x"]!
 ↪=Restaurant_two["id"]) &␣
 ↪(Restaurant_one["Premise_City_x"]==Restaurant_two["Premise_City"])), "cross")
Restaurant_one
```

[32]:
```
+-----+--------------------+-------------+------+--------------------+---------
---+
| id_x|       geolocation_x|Premise_City_x|    id|
geolocation|Premise_City|
+-----+--------------------+-------------+------+--------------------+---------
---+
|56060|35.9207272, -78.9…|        DURHAM| 80309|35.8822361, -78.8…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 57300|36.013017, -78.93…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 57189|36.0075308, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 56774|35.9968356, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 56881|35.9454822, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|188176|36.0113259, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|181953|35.8806773, -78.8…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 55881|36.030452, -78.92…|
DURHAM|
```

```
|56060|35.9207272, -78.9…|        DURHAM| 55627|35.8938411, -78.8…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 56093|36.0553803, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|161512|36.003665, -78.87…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 80349|35.9519154, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 57227|35.9207272, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 69061|36.0037149, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM| 57097|36.0095555, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|164257|36.009727, -78.92…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|161457|35.8987439, -78.8…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|156320|35.8821059, -78.8…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|179466|35.9979158, -78.9…|
DURHAM|
|56060|35.9207272, -78.9…|        DURHAM|170345|35.8991451, -78.8…|
DURHAM|
+-----+--------------------+-------------+------+--------------------+---------
---+
only showing top 20 rows
```

[33]:
```
from haversine import haversine, Unit
def haversine_distance(a,b):
    a_lis = list(a.split(","))
    li = [float(x) for x in a_lis]

    b_lis = list(b.split(","))
    lj = [float(x) for x in b_lis]

    distance = haversine(li, lj)

    return distance

haversine_distance_udf = udf(haversine_distance,DoubleType())
```

[34]:
```
'''
a = Restaurant_one.select("geolocation").collect()[0]
print(a)
b = Restaurant_one.select("geolocation_x").collect()[0]
print(b)
```

```
haversine_distance(a[0],b[0])
'''
```

[34]:
```
'\na = Restaurant_one.select("geolocation").collect()[0]\nprint(a)\nb = Restaura
nt_one.select("geolocation_x").collect()[0]\nprint(b)\nhaversine_distance(a[0],b
[0])\n'
```

[ ]:
```
```

[35]:
```
Restaurant_one
```

[35]:
```
DataFrame[id_x: string, geolocation_x: string, Premise_City_x: string, id:
string, geolocation: string, Premise_City: string]
```

[36]:
```
Restaurant_one.withColumn("distance",␣
↪haversine_distance_udf(col("geolocation_x"),col("geolocation"))).
↪sort("distance").show()
```

```
[Stage 438:>                                                        (0 + 1) / 1]

+-----+------------------+-------------+------+------------------+---------
---+--------+
| id_x|       geolocation_x|Premise_City_x|    id|
geolocation|Premise_City|distance|
+-----+------------------+-------------+------+------------------+---------
---+--------+
|70266|35.9182655, -78.9…|        DURHAM|170325|35.9182655, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 55761|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 57227|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 56907|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 57663|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 58333|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 77025|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 70486|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM|170061|35.9207272, -78.9…|
DURHAM|     0.0|
|56060|35.9207272, -78.9…|        DURHAM| 57003|35.9207272, -78.9…|
DURHAM|     0.0|
|70266|35.9182655, -78.9…|        DURHAM|173158|35.9182655, -78.9…|
DURHAM|     0.0|
```

```
|70266|35.9182655, -78.9…|        DURHAM|160739|35.9182655, -78.9…|
DURHAM|     0.0|
|70266|35.9182655, -78.9…|        DURHAM| 63586|35.9182655, -78.9…|
DURHAM|     0.0|
|70266|35.9182655, -78.9…|        DURHAM|181352|35.9182655, -78.9…|
DURHAM|     0.0|
|70266|35.9182655, -78.9…|        DURHAM| 63622|35.9182655, -78.9…|
DURHAM|     0.0|
|58349|35.9495321, -78.9…|        DURHAM| 56744|35.9495321, -78.9…|
DURHAM|     0.0|
|57278|36.0586094, -78.9…|        DURHAM|170846|36.0586094, -78.9…|
DURHAM|     0.0|
|57278|36.0586094, -78.9…|        DURHAM|170284|36.0586094, -78.9…|
DURHAM|     0.0|
|57278|36.0586094, -78.9…|        DURHAM| 57661|36.0586094, -78.9…|
DURHAM|     0.0|
|57278|36.0586094, -78.9…|        DURHAM|169256|36.0586094, -78.9…|
DURHAM|     0.0|
+-----+--------------------+--------------+------+--------------------+---------
---+--------+
only showing top 20 rows
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: