Assignment 2

1. Show the busiest site in Ireland.

```
In [2]: vehicle_counter_DF.registerTempTable("transport")
```

```
In [3]: the busiest site in Ireland.
= spark.sql("select cosit as busiest_site, count(*) as count from transport group by cosit order by count(cosit) desc limit 1");
.show()
```

```
+------------+-----+
|busiest_site|count|
+------------+-----+
|        1508|98292|
+------------+-----+
```

```
In [4]: busiest_site.select("busiest_site", "count")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="busiest_site", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table busiest_site(busiest_site int, count int, primar
y key(busiest_site));
cqlsh:assignment2> select * from busiest_site;

 busiest_site | count
--------------+-------
         1508 | 98292

(1 rows)
```

2. Show the average distance between vehicles on all M50 sites.

```
In [26]: # 2.Show the average distance between vehicles on all M50 sites.

distance_between_vehicles = spark.sql("select AVG(gap) as avg_gap from transport where cosit in(1500,1501,1502,1503,1504,1505,15
distance_between_vehicles.show()
```

```
+-----------------+
|          avg_gap|
+-----------------+
|4.274862509505848|
+-----------------+
```

```
In [29]: distance_between_vehicles.select('avg_gap')\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="avg_gap", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table avg_gap(avg_gap float, primary key(avg_gap));
cqlsh:assignment2> select * from avg_gap;

 avg_gap
---------
 4.27486

(1 rows)
```

3. What site has recorded the highest temperature? Show the hour of the day.

```
In [36]: e highest temperature? Show the hour of the day.
         ql("select cosit as site, hour,minute, MAX(temperature) as max_temp from transport group by cosit,hour,minute order by MAX(temper
```

```
+----+----+------+--------+
|site|hour|minute|max_temp|
+----+----+------+--------+
|1015|  13|     4|    12.0|
|1015|  18|    47|    12.0|
|1015|  18|    48|    12.0|
|1015|  12|    32|    12.0|
|1015|  18|    49|    12.0|
+----+----+------+--------+
```

```
In [52]: Maximum_temperature.select("site", "hour", "minute", "max_temp")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="maximum_temperature", keyspace = "assignment2")\
         .save(mode="append")
```

```
cqlsh:assignment2> create table Maximum_temperature(site int, hour int, minute int, max_temp float, primary key(
site));
cqlsh:assignment2> select * from maximum_temperature;

 site | hour | max_temp | minute
------+------+----------+--------

(0 rows)
cqlsh:assignment2> select * from maximum_temperature;

 site | hour | max_temp | minute
------+------+----------+--------
 1015 |   18 |       12 |     49

(1 rows)
```

4. Show total number of WIM sites available in the dataset?

```
In [53]: # 4. Show total number of WIM sites available in the dataset?

         Total_WIM_sites = spark.sql("select COUNT(DISTINCT cosit) as wim_site from transport group where weight IS NOT NULL");
         Total_WIM_sites.show();
```

```
+--------+
|WIM_site|
+--------+
|     325|
+--------+
```

```
In [58]: Total_WIM_sites.select('wim_site')\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="total_wim_sites", keyspace = "assignment2")\
         .save(mode="append")
```

```
cqlsh:assignment2> create table total_wim_sites(wim_site int, primary key(wim_site));
cqlsh:assignment2> select * from total_wim_sites;

 wim_site
----------
      325

(1 rows)
```

5. Compute the average speed for each site on M50.

```
In [67]: he average speed for each site on M50.
park.sql("select cosit, avg(speed) as avg_speed from transport where cosit IN (1500, 1501, 1502, 1503, 1504, 1505, 1506, 1507, 15
w()
```

```
+-----+-----------------+
|cosit|        avg_speed|
+-----+-----------------+
| 1507|95.00087226856925|
| 1506|89.01114291227168|
| 1504|84.95810635538263|
| 1502|84.54840331627523|
| 1501| 83.7040682651283|
| 1505|83.56696008249519|
| 1508|81.37383510356896|
| 1503|79.29442543070944|
| 1012|78.86071923672837|
| 1509|78.19974476773864|
| 1500|76.23903247455863|
+-----+-----------------+
```

```
In [62]: Avg_speed.select("cosit", "avg_speed")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="avg_speed", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table avg_speed(cosit int, avg_speed float, primary key(cosit));
cqlsh:assignment2> select * from avg_speed;

 cosit | avg_speed
-------+-----------
  1500 |  76.23903
  1503 |  79.29443
  1012 |  78.86072
  1501 |  83.70407
  1509 |  78.19975
  1502 |   84.5484
  1508 |  81.37383

(7 rows)
```

6. Show total number of counts by vehicle class. Order results in descending.

```
In [69]: # 6.Show total number of counts by vehicle class. Order results in descending.

Total_numberof_vehicles = spark.sql("select classname, count(class) as total_vehicle_count from transport group by classname ord
Total_numberof_vehicles.show()
```

```
+---------+-------------------+
|classname|total_vehicle_count|
+---------+-------------------+
|      CAR|            3472965|
|      LGV|             498505|
|  HGV_ART|             216978|
|  HGV_RIG|             135202|
|      BUS|              29347|
|  CARAVAN|              21224|
|    MBIKE|              14682|
|     null|                396|
+---------+-------------------+
```

```
In [74]: Total_numberof_vehicles.select("classname", "total_vehicle_count")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="total_vehicles", keyspace = "assignment2")\
.save(mode="append")
```

```
SyntaxException: line 1:80 no viable alternative at input '(' (... text, total_vehicle_count int, primary_key[(]...)
cqlsh:assignment2> create table total_vehicles(classname text, total_vehicle_count int, primary key(total_vehicle_count));
cqlsh:assignment2> select * from total_vehicles;

 total_vehicle_count | classname
---------------------+-----------
               29347 |       BUS
               14682 |     MBIKE
              498505 |       LGV
                 396 |      null
             3472965 |       CAR
              135202 |   HGV_RIG
               21224 |   CARAVAN
              216978 |   HGV_ART

(8 rows)
```

7. List the top 3 busiest sites on M50.

In [78]:
```
# 7. List the top 3 busiest sites on M50.

Top3_busiest_sites = spark.sql("select cosit, count(*) as vehicle_count from transport where cosit IN (1500, 1501, 1502, 1503, 1
Top3_busiest_sites.show()
```

```
+-----+-------------+
|cosit|vehicle_count|
+-----+-------------+
| 1508|        98292|
| 1502|        89498|
| 1503|        86195|
| 1501|        83205|
| 1509|        78360|
+-----+-------------+
```

In [79]:
```
Top3_busiest_sites.select("cosit", "vehicle_count")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="busiest_sites", keyspace = "assignment2")\
.save(mode="append")
```

```
(8 rows)
cqlsh:assignment2> create table busiest_sites(cosit int, vehicle_count int, primary key(cosit));
cqlsh:assignment2> select * from busiest_sites;

 cosit | vehicle_count
-------+---------------
  1503 |         86195
  1501 |         83205
  1509 |         78360
  1502 |         89498
  1508 |         98292

(5 rows)
```

8. What is the busiest site on M6?

In [84]:
```
# 8.What is the busiest site on M6?

Busiest_site_on_M6 = spark.sql("select cosit, count(*) as count from transport where cosit IN (3601,3602, 3603, 3604, 3605, 2006
Busiest_site_on_M6.show()
```

```
+-----+-----+
|cosit|count|
+-----+-----+
|20062|12387|
+-----+-----+
```

In [88]:
```
Busiest_site_on_M6.select("cosit", "count")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="m6_site", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table m6_site(cosit int, count int, primary key(cosit));
cqlsh:assignment2> select * from m6_site;

 cosit | count
-------+-------
 20062 | 12387

(1 rows)
```

9. What site reports the highest number of HGVs?

```
In [91]: #9. What site reports the highest number of HGVs?

Highest_HGV_site_count = spark.sql("select cosit, count(*) as hgv_count from transport where class = 5 or class = 6 group by cos
Highest_HGV_site_count.show()
```

```
+-----+---------+
|cosit|hgv_count|
+-----+---------+
|  997|    12031|
+-----+---------+
```

```
In [92]: Highest_HGV_site_count.select("cosit", "hgv_count")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="highest_numberof_hgv", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table highest_numberof_hgv(cosit int, hgv_count int, primary key(cosit));
cqlsh:assignment2> select * from highest_numberof_hgv;

 cosit | hgv_count
-------+-----------
   997 |     12031

(1 rows)
```

10. Calculate the total number of vehicles on each site on M7.

```
In [93]: #10. Calculate the total number of vehicles on each site on M7.

Total_numberof_vehicles_on_M7 = spark.sql("select cosit, count(*) as total_vehicles from transport where cosit IN (3703, 3704, 20
Total_numberof_vehicles_on_M7.show()
```

```
+------+--------------+
| cosit|total_vehicles|
+------+--------------+
| 20074|         26331|
| 20075|         26148|
|200713|         25864|
| 20076|         22952|
| 20078|         17977|
| 20077|         17712|
|200722|         17568|
| 20079|         16296|
|200715|         13944|
|200721|         12441|
|200719|          8445|
|200718|          7694|
|200717|          7203|
|200716|          7073|
+------+--------------+
```

```
In [94]: Total_numberof_vehicles_on_M7.select("cosit", "total_vehicles")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="total_vehicles_on_m7", keyspace = "assignment2")\
.save(mode="append")
```

```
cqlsh:assignment2> create table total_vehicles_on_m7(cosit int, total_vehicles int, primary key(cosit));
cqlsh:assignment2> select * from total_vehicles_on_m7;

 cosit  | total_vehicles
--------+----------------
 200713 |          25864
 200722 |          17568
  20075 |          26148
 200721 |          12441
 200719 |           8445
 200715 |          13944
  20078 |          17977
 200716 |           7073
  20079 |          16296
  20074 |          26331
  20076 |          22952
  20077 |          17712
 200718 |           7694
 200717 |           7203

(14 rows)
cqlsh:assignment2>
```