Question No: 1.a

For code based questions -

```
with year_cte as(select substr(Issue_Date,7,4) as year from violation_part)select year,
count(*) from year_cte group by year order by year;
```

```
hive> with year_cte as(select substr(Issue_Date,7,4) as year from violation_part)select year, count(*) from year_cte group by year order by year;
Query ID = hadoop_20221022064526_4e8fcd6e-ffb8-4788-86e4-870038e427d6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666420160923_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 33.94 s
--------------------------------------------------------------------------------
OK
year    _c1
2017    5431903
Time taken: 35.066 seconds, Fetched: 1 row(s)
```

Question No: 1.b

For code based questions -

```
select count(distinct Registration_State)
from violation_part;
```

```
hive> select count(distinct Registration_State)
    > from violation_part;
Query ID = hadoop_20221022064945_50b7b6ee-ab24-4e6e-b4f7-a17a7771e3cf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666420160923_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 31.03 s
--------------------------------------------------------------------------------
OK
_c0
65
```

Question No: 1.c

```
For code based questions -
select count(*)
from violation_part
where Street_Code1 is null
or Street_Code2 is null
or Street_Code3 is null;
```

```
hive> select count(*)
    > from violation_part
    > where Street_Code1 is null
    > or Street_Code2 is null
    > or Street_Code3 is null;
Query ID = hadoop_20221022065650_b4cbfc19-580f-4d88-a3bc-369c92a83bb6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666420160923_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 30.12 s
----------------------------------------------------------------------------------------------
OK
_c0
4
```

Question No: 2.a

For code based questions -

```
with hour_cte as(select HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,'
',Violation_Time,'m'),'mm/dd/yyyy hhmma'))) as hour
from violation_part)
select hour,count(*)
from hour_cte
group by hour
order by hour;
```

```
hive> with hour_cte as(select HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Dat
e,' ',Violation_Time,'m'),'mm/dd/yyyy hhmma'))) as hour
    > from violation_part)
    > select hour,count(*)
    > from hour_cte
    > group by hour
    > order by count(*) desc;
FAILED: SemanticException [Error 10128]: Line 6:9 Not yet supported place for UD
AF 'count'
hive> with hour_cte as(select HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,' ',Violation_Time,'m'),'mm/dd/yyyy hhmma'))) as hour
    > from violation_part)
    > select hour,count(*)
    > from hour_cte
    > group by hour
    > order by hour;
Query ID = hadoop_20221022073114_3eded828-88a0-47d6-8ef0-238ada7c6b5a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666420160923_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 68.90 s
----------------------------------------------------------------------------------------------
```

Question No: 2.a

Output:

```
OK
NULL    62
0       45700
1       46073
2       40313
3       32461
4       14549
5       43158
6       121550
7       270625
8       503841
9       595624
10      489454
11      574625
12      510130
13      549287
14      466070
15      314469
16      295988
17      211173
18      104288
19      26099
20      49224
21      55323
22      42538
23      29279
```

Question No: 2.b

For code based questions -

```
with vio_cte as
(select round(HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,'
',Violation_Time,'m'),'mm/dd/yyyy hhmma')))/4) as time,Violation_Code as vc
from violation_part)
select time, vc,n_violations,vc_rank
from(
select time, vc,count(*) as n_violations,rank() OVER(Partition by time order by count(*)
desc) as vc_rank
from vio_cte
group by time,vc) as x
where vc_rank < 4;
```

```
Time taken: 89.185 seconds, Fetched: 635 row(s)
hive> with vio_cte as
    > (select round(HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,' ',Violation_Time,'m'),'mm/dd/yyyy hhmma')))/4) as time,Violation_Code as vc
    > from violation_part)
    > select time, vc,n_violations,vc_rank
    > from(
    > select time, vc,count(*) as n_violations,rank() OVER(Partition by time order by count(*) desc) as vc_rank
    > from vio_cte
    > group by time,vc) as x
    > where vc_rank < 4;
Query ID = hadoop_20221022091807_2e4db0a3-3634-4431-a89b-cb3de4cb7a1c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666426737285_0004)
```

Question No: 2.b

Output:

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      6          6        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 87.24 s
--------------------------------------------------------------------------------
OK
NULL    21      19      1
NULL    94      15      2
NULL    46      8       3
0.0     21      23038   1
0.0     40      12677   2
0.0     14      9765    3
1.0     40      27366   1
1.0     21      16930   2
1.0     14      13495   3
3.0     36      395976  1
3.0     21      294827  2
3.0     38      226240  3
5.0     7       37587   1
5.0     38      33959   2
5.0     14      24559   3
4.0     38      198713  1
4.0     37      141575  2
4.0     14      129037  3
2.0     21      432722  1
2.0     36      167209  2
2.0     14      147892  3
6.0     7       11657   1
6.0     40      11117   2
6.0     14      10818   3
Time taken: 88.444 seconds, Fetched: 24 row(s)
```

Question No: 2.c

```
For code based questions -
with vio_cte as
(select round(HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,'
',Violation_Time,'m'),'mm/dd/yyyy hhmma')))/4) as time_period ,Violation_Code as vc
from violation_part)
select *
from
(select a.vc, time_period, rank() over(partition by a.vc order by sum(time_period) desc) as
vc_time_period_rank from
( select Violation_Code as vc , rank() over(partition by null order by count(*) desc) as
rc_rank
from violation_part
group by Violation_Code) a
left outer join
(select time_period , vc,
count(*) as rc_time_period_count
from vio_cte
group by time_period,vc) b
on (a.vc = b.vc)
where rc_rank < 4
group by a.vc, time_period
) c
where vc_time_period_rank < 4;
```

Question No: 2.c

Output1:

```
hive> with vio_cte as
    > (select round(HOUR(FROM_UNIXTIME(UNIX_TIMESTAMP(concat(Issue_Date,' ',Violation_Time,'m'),'mm/dd/yyyy hhmma')))/4) as time_period ,Violation_Code as vc

    > from violation_part)
    > select *
    > from
    > (select a.vc, time_period, rank() over(partition by a.vc order by sum(time_period) desc) as vc_time_period_rank from
    > ( select Violation_Code as vc , rank() over(partition by null order by count(*) desc) as rc_rank
    > from violation_part
    > group by Violation_Code) a
    > left outer join
    > (select time_period , vc,
    > count(*) as rc_time_period_count
    > from vio_cte
    > group by time_period,vc) b
    > on (a.vc = b.vc)
    > where rc_rank < 4
    > group by a.vc, time_period
    > ) c
    > where vc_time_period_rank < 4;
Query ID = hadoop_20221022104727_61936f96-179d-4468-a7fa-65aa03a46c7b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666434000867_0002)
```

Question No: 2.c

Output2:

```
Query ID = hadoop_20221022104727_61936f96-179d-4468-a7fa-65aa03a46c7b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666434000867_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Map 7 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 6 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 8 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 08/08  [==========================>>] 100%  ELAPSED TIME: 102.06 s
----------------------------------------------------------------------------------------------
OK
36      4.0     1
36      3.0     2
36      2.0     3
21      6.0     1
21      5.0     2
21      4.0     3
38      6.0     1
38      5.0     2
38      4.0     3
```

Question No:
2.d.1

```
For code based questions -
select case
when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
end as season, count(*)
from violation_part
group by case
when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
end;
```

Question No: 2.d.1

Output:

```
hive> select case
    > when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
    > when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
    > when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
    > when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
    > end as season, count(*)
    > from violation_part
    > group by case
    > when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
    > when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
    > when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
    > when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
    > end;
Query ID = hadoop_20221022114654_558520be-3e17-425c-84c6-adea2d41dffc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666438388589_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 37.13 s
--------------------------------------------------------------------------------
OK
season  _c1
fall    979
spring  2873380
summer  852864
winter  1704680
```

Question No:
2.d.2

For code based questions -

```
select * from
(select *, rank() over(partition by season order by vc_count desc) as vc_rank
from( select case
when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
end as season,Violation_Code, count(*) as vc_count
from violation_part
group by case
when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
end, Violation_code
) a
) b
where vc_rank < 4;
```

Question No: 2.d.1

Output1:

```
hive> select * from
    > (select *, rank() over(partition by season order by vc_count desc) as vc_rank
    > from( select case
    > when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
    > when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
    > when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
    > when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
    > end as season,Violation_Code, count(*) as vc_count
    > from violation_part
    > group by case
    > when substr(Issue_Date,1,2) in ('03','04','05') then 'spring'
    > when substr(Issue_Date,1,2) in ('06','07','08') then 'summer'
    > when substr(Issue_Date,1,2) in ('09','10','11') then 'fall'
    > when substr(Issue_Date,1,2) in ('12','01','02') then 'winter'
    > end, Violation_code
    > ) a
    > ) b
    > where vc_rank < 4;
Query ID = hadoop_20221022115451_ebc66297-d1ed-4a35-98f7-0b286038953b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666438388589_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
```

Question No: 2.d.1

Output 2:

```
Query ID = hadoop_20221022115451_ebc66297-d1ed-4a35-98f7-0b286038953b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1666438388589_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      6          6        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 41.09 s
----------------------------------------------------------------------------------------------
OK
b.season        b.violation_code        b.vc_count      b.vc_rank
fall    46      231     1
fall    21      128     2
fall    40      116     3
spring  21      402424  1
spring  36      344834  2
spring  38      271167  3
summer  21      127350  1
summer  36      96663   2
summer  38      83518   3
winter  21      238180  1
winter  36      221268  2
winter  38      187386  3
```