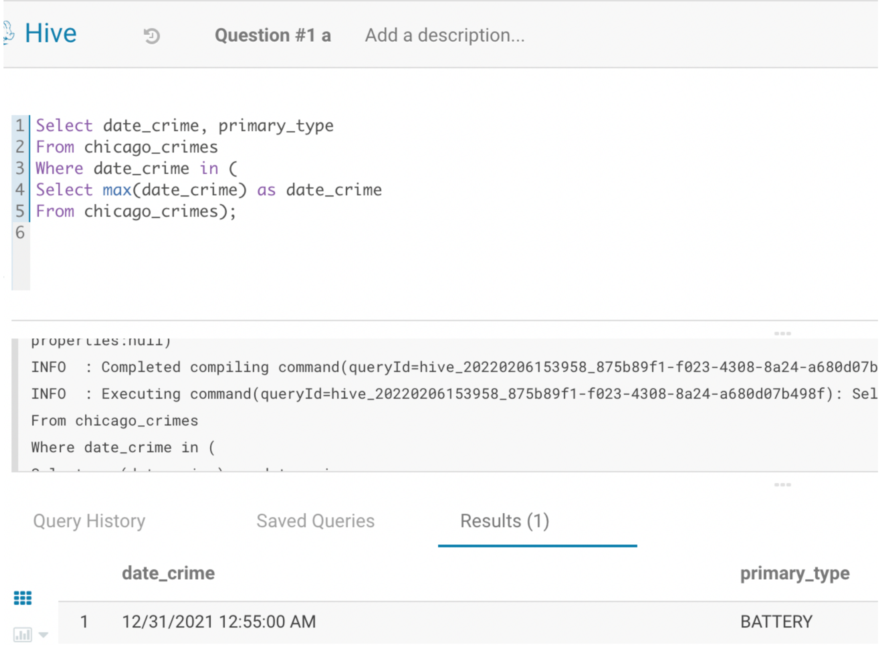
Assignment 2

Fiona Fei

Data Loading:

```
#log into roc
ssh fionafei@hadoop.rcc.uchicago.edu

#used FileZilla to load data to HDFS
wget -qO- https://data.cityofchicago.org/api/views/ijzp-
q8t2/rows.csv?accessType=DOWNLOAD | hdfs dfs -put -
/user/$USER/crime/ChicagoCrime.csv


#create hive table called chicago_crimes
ssh -Y fionafei@hadoop.rcc.uchicago.edu

hive

Create database if not exists fionafei;

#switch to my database
Use fionafei;

#show all tables under my database
Show tables;

#create new table

Create external table chicago_crimes (id string, case_number string,
crime_date string, block string, iucr string, primary_type string,
description string, `location_description` string, arrest string,
domestic string, beat string, district string, ward string,
community_area string, fbi_code string, x_coordinate string,
y_coordinate string, year string, updated_on timestamp, latitude
string, longitude string, location string, location_address string,
location_city string, location_state string, location_zip string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = ",",
   "quoteChar"     = '\"',
   "escapeChar"    = '\\')

STORED AS TEXTFILE
LOCATION '/user/fionafei/crime/'
tblproperties("skip.header.line.count"="1");
```

###########
#Data Manipulation – Hive
#4) What are earliest and most recent dates of the crimes recorded in the dataset and what are the types of those crimes. (Dates might vary based on when you download the dataset)
#most recent date

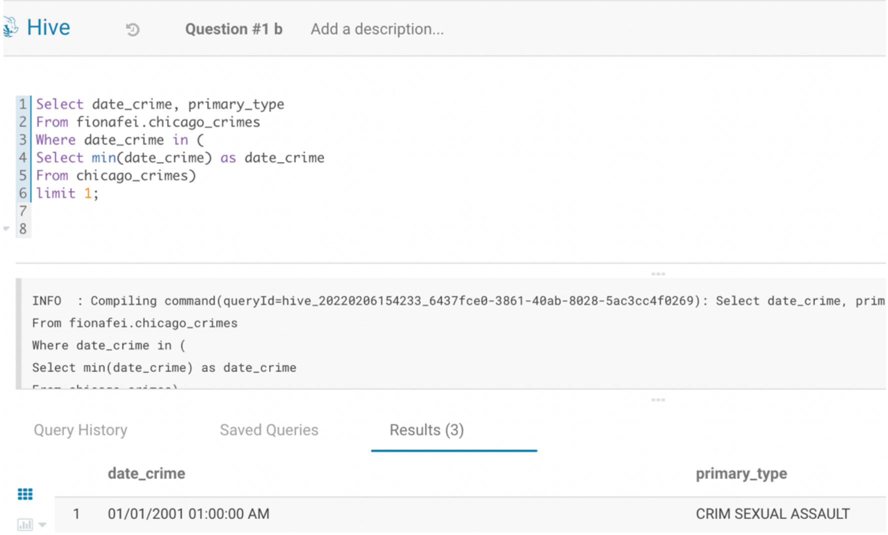


```
Select date_crime, primary_type
From chicago_crimes
Where date_crime in (
Select max(date_crime) as date_crime
From chicago_crimes);
```

```
properties:null)
INFO  : Completed compiling command(queryId=hive_20220206153958_875b89f1-f023-4308-8a24-a680d07b
INFO  : Executing command(queryId=hive_20220206153958_875b89f1-f023-4308-8a24-a680d07b498f): Sel
From chicago_crimes
Where date_crime in (
```

	date_crime	primary_type
1	12/31/2021 12:55:00 AM	BATTERY

#earliest date



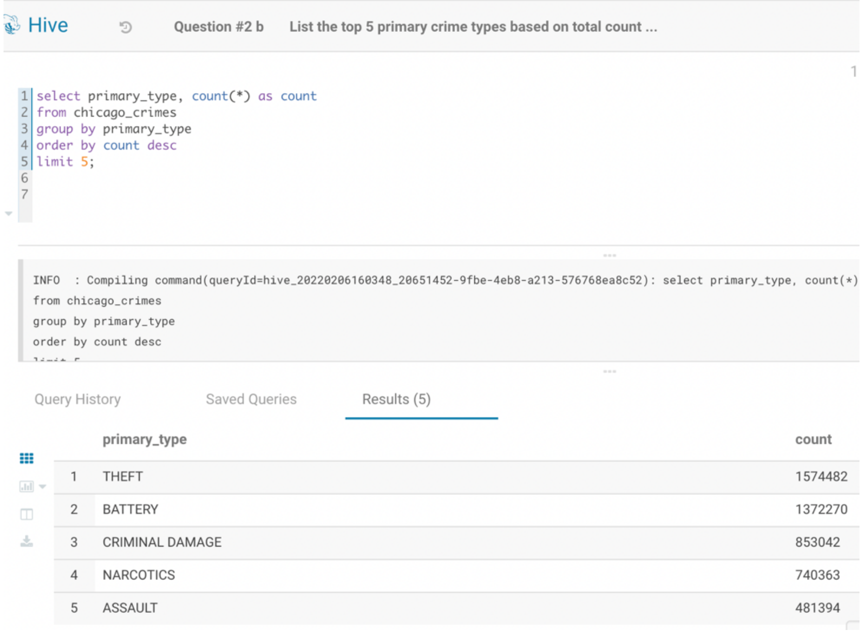
```
Select date_crime, primary_type
From fionafei.chicago_crimes
Where date_crime in (
Select min(date_crime) as date_crime
From chicago_crimes)
limit 1;
```

```
INFO  : Compiling command(queryId=hive_20220206154233_6437fce0-3861-40ab-8028-5ac3cc4f0269): Select date_crime, prim
From fionafei.chicago_crimes
Where date_crime in (
Select min(date_crime) as date_crime
```

	date_crime	primary_type
1	01/01/2001 01:00:00 AM	CRIM SEXUAL ASSAULT

#5 List the top 5 and bottom 5 primary crime types based on total count of occurrences

#Top 5

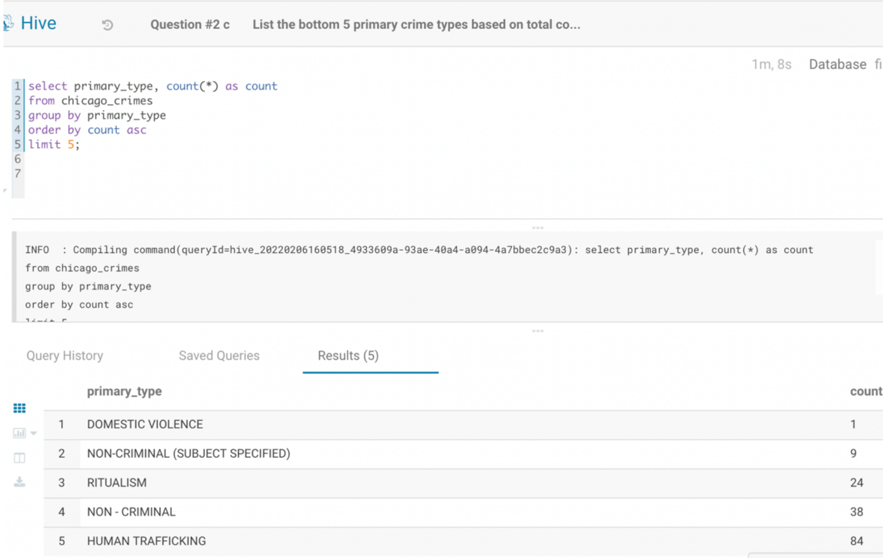


```
select primary_type, count(*) as count
from chicago_crimes
group by primary_type
order by count desc
limit 5;
```

```
INFO  : Compiling command(queryId=hive_20220206160348_20651452-9fbe-4eb8-a213-576768ea8c52): select primary_type, count(*)
from chicago_crimes
group by primary_type
order by count desc
limit 5
```

	primary_type	count
1	THEFT	1574482
2	BATTERY	1372270
3	CRIMINAL DAMAGE	853042
4	NARCOTICS	740363
5	ASSAULT	481394

#Bottom 5

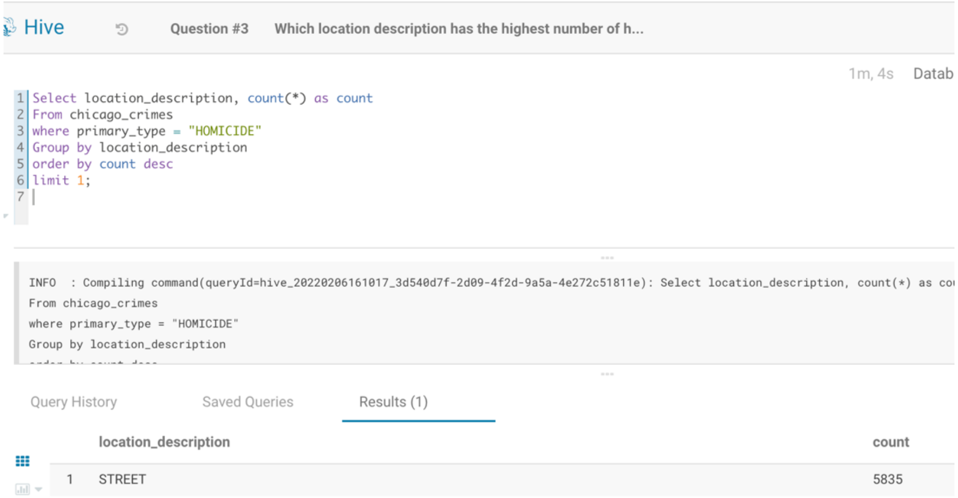


```
select primary_type, count(*) as count
from chicago_crimes
group by primary_type
order by count asc
limit 5;
```

```
INFO  : Compiling command(queryId=hive_20220206160518_4933609a-93ae-40a4-a094-4a7bbec2c9a3): select primary_type, count(*) as count
from chicago_crimes
group by primary_type
order by count asc
limit 5
```

	primary_type	count
1	DOMESTIC VIOLENCE	1
2	NON-CRIMINAL (SUBJECT SPECIFIED)	9
3	RITUALISM	24
4	NON - CRIMINAL	38
5	HUMAN TRAFFICKING	84

#6 which location description has the highest number of homicides associated with it?

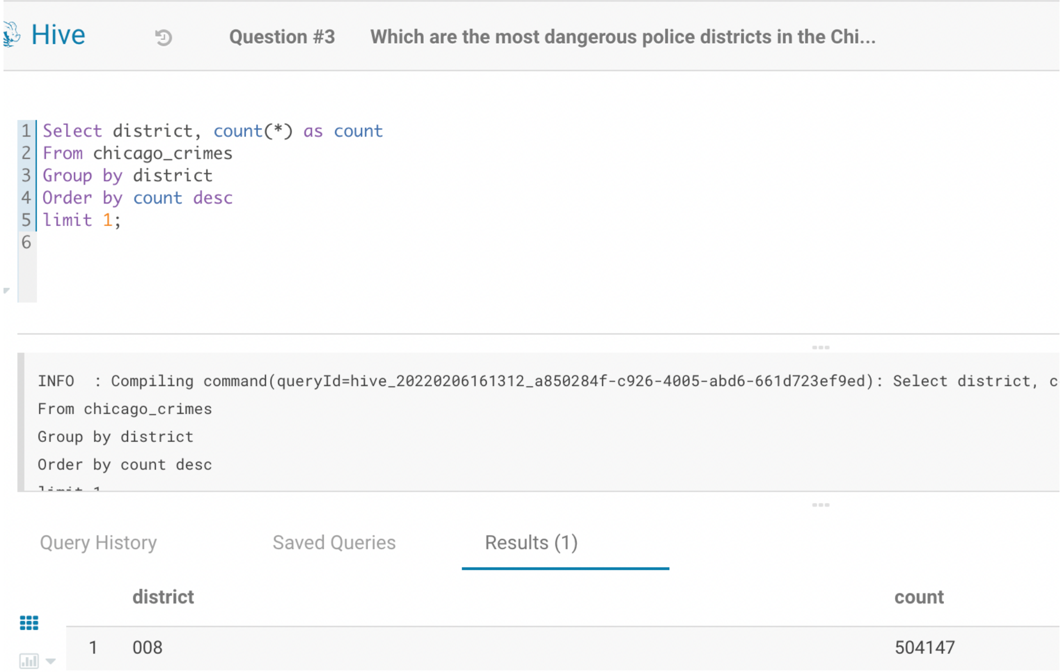


```
Select location_description, count(*) as count
From chicago_crimes
where primary_type = "HOMICIDE"
Group by location_description
order by count desc
limit 1;
```

```
INFO  : Compiling command(queryId=hive_20220206161017_3d540d7f-2d09-4f2d-9a5a-4e272c51811e): Select location_description, count(*) as co
From chicago_crimes
where primary_type = "HOMICIDE"
Group by location_description
order by count desc
```

	location_description	count
1	STREET	5835

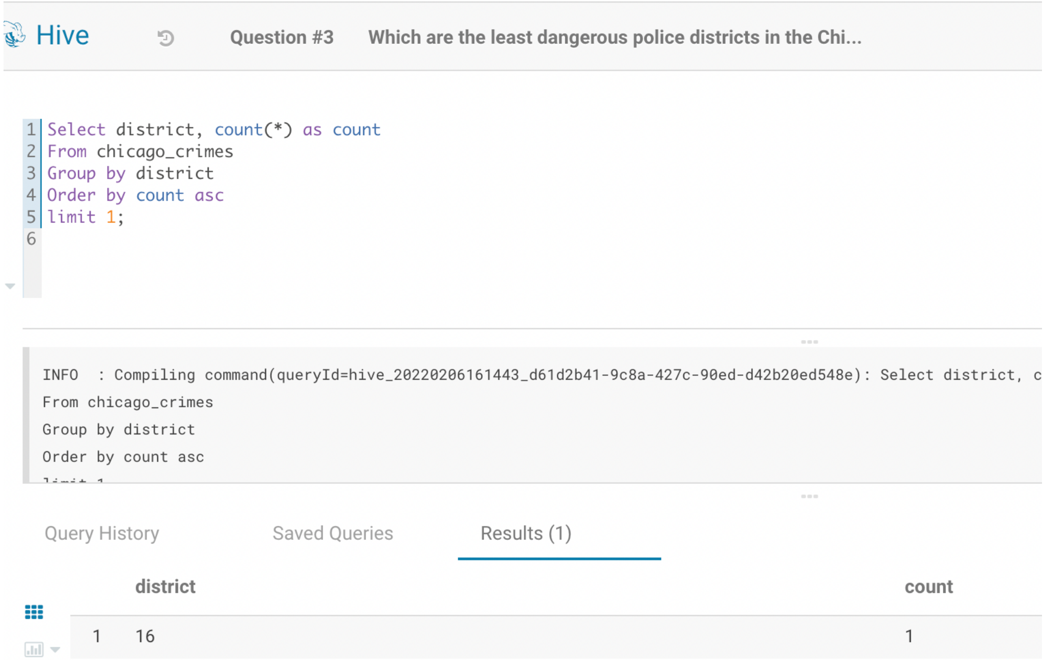
#7 Which are the most dangerous and least dangerous police districts in the Chicago area?



```
Select district, count(*) as count
From chicago_crimes
Group by district
Order by count desc
limit 1;
```

```
INFO  : Compiling command(queryId=hive_20220206161312_a850284f-c926-4005-abd6-661d723ef9ed): Select district, c
From chicago_crimes
Group by district
Order by count desc
limit 1
```

	district	count
1	008	504147



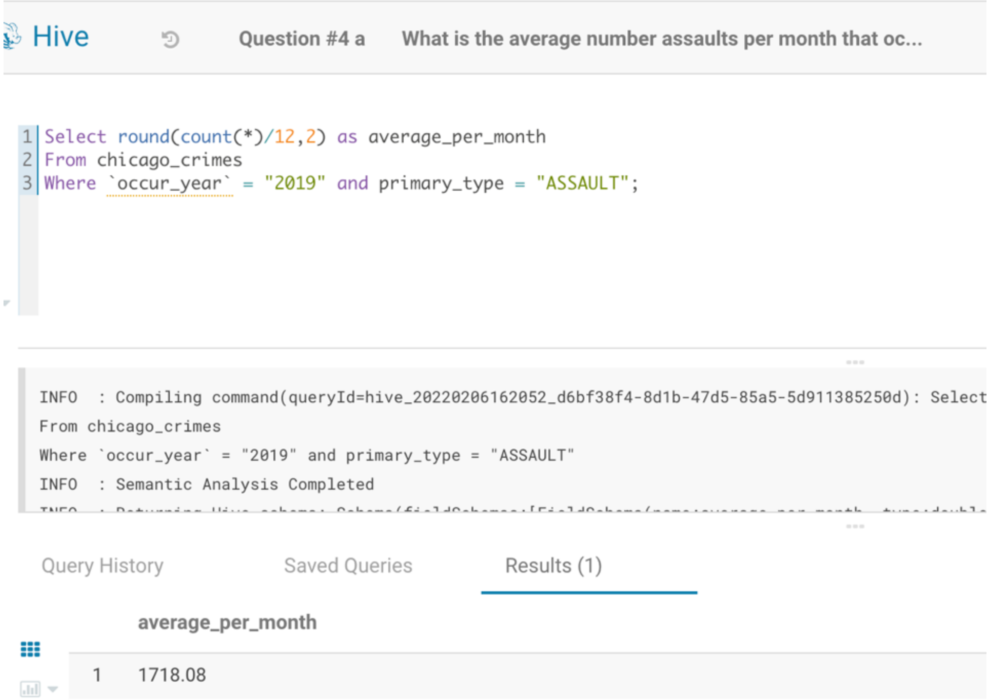
Which are the least dangerous police districts in the Chi...

```
Select district, count(*) as count
From chicago_crimes
Group by district
Order by count asc
limit 1;
```

```
INFO  : Compiling command(queryId=hive_20220206161443_d61d2b41-9c8a-427c-90ed-d42b20ed548e): Select district, c
From chicago_crimes
Group by district
Order by count asc
limit 1
```

	district	count
1	16	1

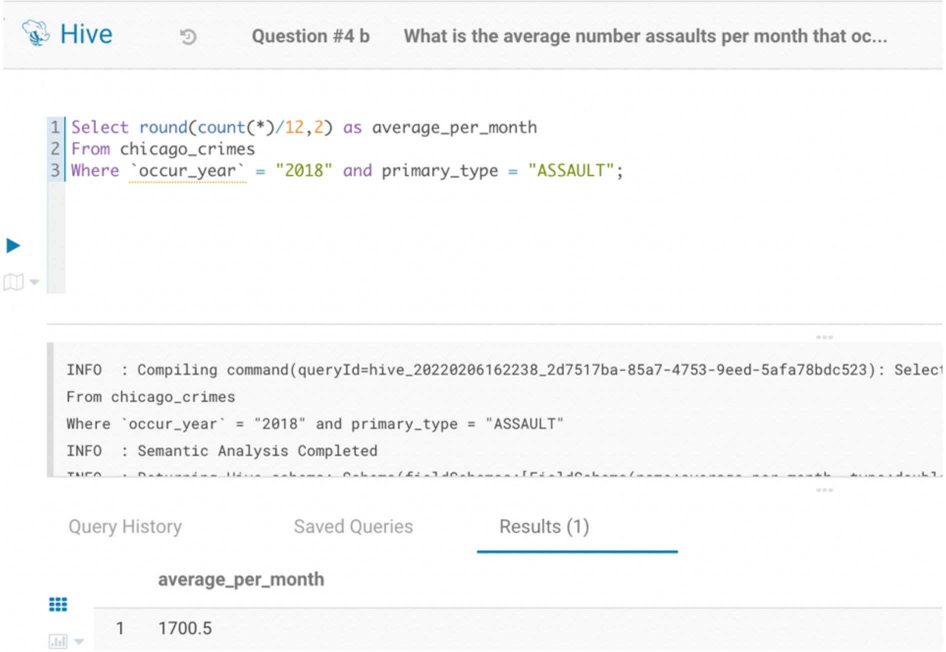
#8 What is the average number assaults per month that occurred in 2019. Has that number increased since the prior period?



```
Select round(count(*)/12,2) as average_per_month
From chicago_crimes
Where `occur_year` = "2019" and primary_type = "ASSAULT";
```

```
INFO  : Compiling command(queryId=hive_20220206162052_d6bf38f4-8d1b-47d5-85a5-5d911385250d): Select
From chicago_crimes
Where `occur_year` = "2019" and primary_type = "ASSAULT"
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:average_per_month, type:double
```

	average_per_month
1	1718.08



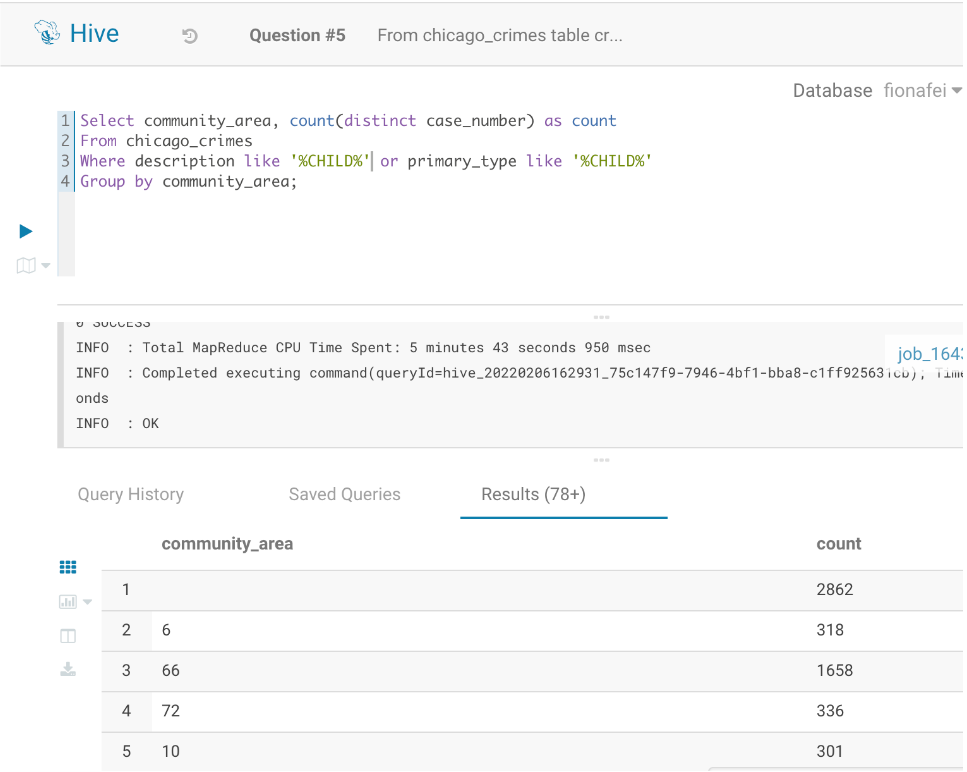
```
Select round(count(*)/12,2) as average_per_month
From chicago_crimes
Where `occur_year` = "2018" and primary_type = "ASSAULT";
```

```
INFO  : Compiling command(queryId=hive_20220206162238_2d7517ba-85a7-4753-9eed-5afa78bdc523): Selec
From chicago_crimes
Where `occur_year` = "2018" and primary_type = "ASSAULT"
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:average_per_month, type:doubl
```

	average_per_month
1	1700.5

The average assaults per month in 2019 is around 1718.08 cases, and the average assaults per month in 2018 is around 1700.5 cases. The average cases per month increased around 18 cases per month.

```
#9 From chicago_crimes table create a smaller (summarized) external
table in Hive (that supports questions 9 and 10) and download this
summarized table to your computer as a CSV file.

#For Question 10
#community area, primary type, count of crimes,
description/primary_type includes “CHILD”


Select community_area, count(distinct case_number) as count
From chicago_crimes
Where description like '%CHILD%' or primary_type like '%CHILD%'
Group by community_area;

#download as csv
```



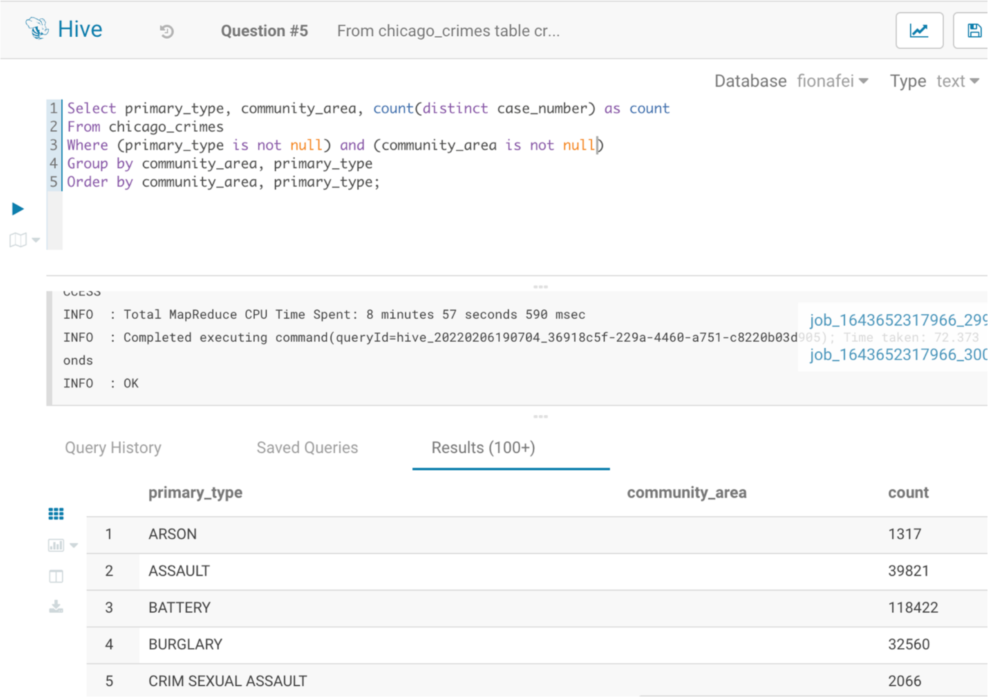
```
Select community_area, count(distinct case_number) as count
From chicago_crimes
Where description like '%CHILD%' or primary_type like '%CHILD%'
Group by community_area;
```

```
0 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 5 minutes 43 seconds 950 msec
INFO  : Completed executing command(queryId=hive_20220206162931_75c147f9-7946-4bf1-bba8-c1ff92563 cb); Time
onds
INFO  : OK
```

job_164

	community_area	count
1		2862
2	6	318
3	66	1658
4	72	336
5	10	301

```
#For Question 11
#primary_type, community_area, count(distinct case_number) as count

Select primary_type, community_area, count(distinct case_number) as
count
From chicago_crimes
Where primary_type is not null and community_area is not null
Group by community_area, primary_type
Order by community_area, primary_type;
```



Database fionafei Type text

```sql
Select primary_type, community_area, count(distinct case_number) as count
From chicago_crimes
Where (primary_type is not null) and (community_area is not null)
Group by community_area, primary_type
Order by community_area, primary_type;
```

```
CCESS
INFO  : Total MapReduce CPU Time Spent: 8 minutes 57 seconds 590 msec
INFO  : Completed executing command(queryId=hive_20220206190704_36918c5f-229a-4460-a751-c8220b03d905); Time taken: 72.373
onds
INFO  : OK
```

job_1643652317966_29

job_1643652317966_30

	primary_type	community_area	count
1	ARSON		1317
2	ASSAULT		39821
3	BATTERY		118422
4	BURGLARY		32560
5	CRIM SEXUAL ASSAULT		2066