Analysis of the Factors Contributing to Changes in Crime Rates and Property Values in New York City

BIG DATA GROUP PROJECT

C-SUITE EMPLOYEES:

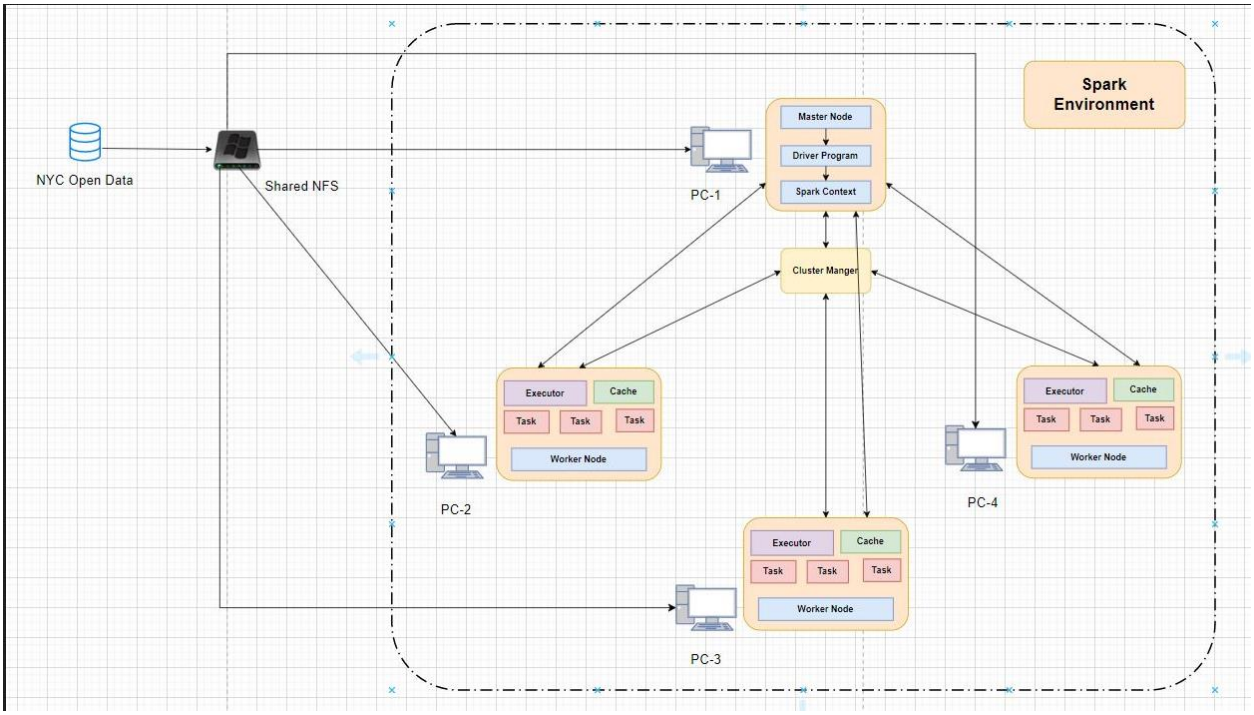
Satya Vinod Kantipudi - SXK210365 (CEO- Chief Executive Office)

Anurag S S S M - AXS210475 (CTO – Chief Technical Officer)

Sriya Meghana Nandam – SMN210004 (CIO – Chief Information Officer)

Lekhashree Jeezula - LXJ220002 (COO – Chief Operating Officer)

ARCHITECTURE



Distributed Computing System: The architecture you described is a distributed computing system. This means that the processing of data is distributed across multiple nodes in the system to improve performance and efficiency.

Network File System (NFS): The system uses a network file system (NFS) to load data onto the system. NFS is a protocol that allows a user to access files over a network as if they were stored on the local machine.

Master-Slave Architecture: The distributed computing system consists of a master node and one or more slave nodes. The master node is responsible for coordinating the processing of data and distributing tasks to the slave nodes.The slave nodes perform the computations on the data.

Data Loading: To start the data processing, the master node loads the data from the NFS server into memory. The data is partitioned into smaller chunks that can be processed by the slave nodes in parallel.

Parallel Processing: The slave nodes are equipped with processing power and memory to perform computations on the data. The master node assigns the data chunks to the slave nodes, which perform computations on the data in parallel.

Results Combination: Once the slave nodes have processed their assigned data chunks, they send the results back to the master node. The master node combines the results from each of the slave nodes and presents a final output to the user.

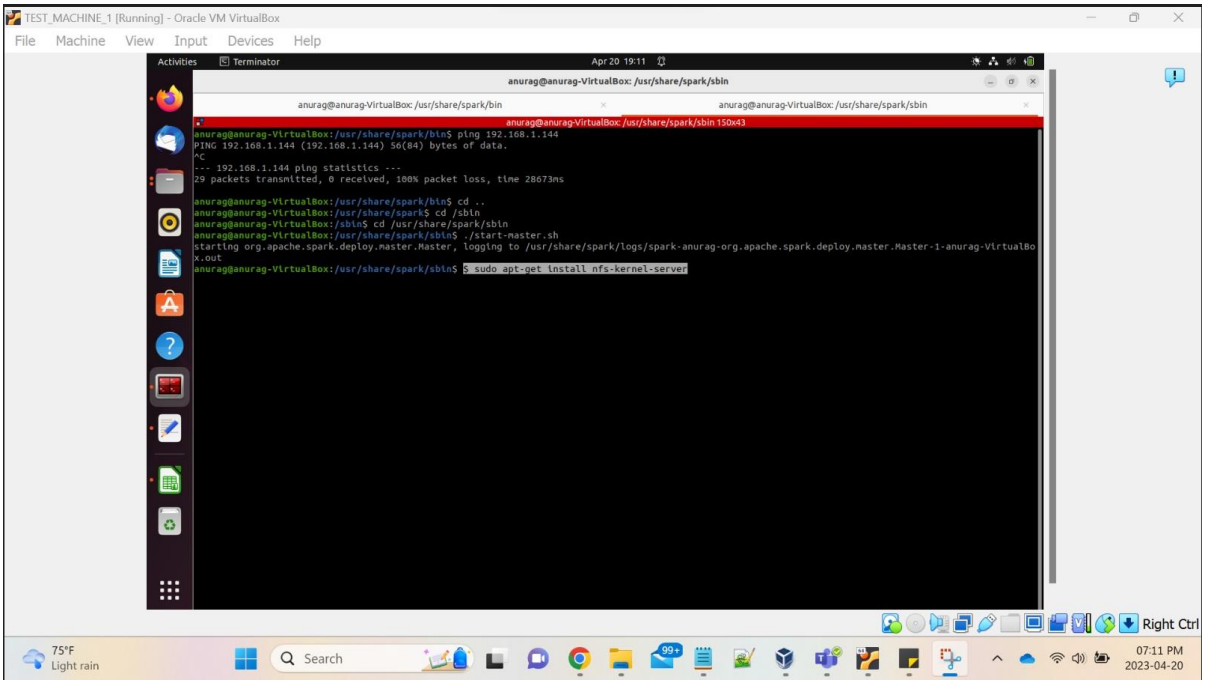
Performance Improvement: The architecture you described is often used in big data processing systems such as Hadoop and Apache Spark. By distributing the workload across multiple nodes, the processing time can be reduced significantly, improving performance and efficiency.

Network File System (NFS):

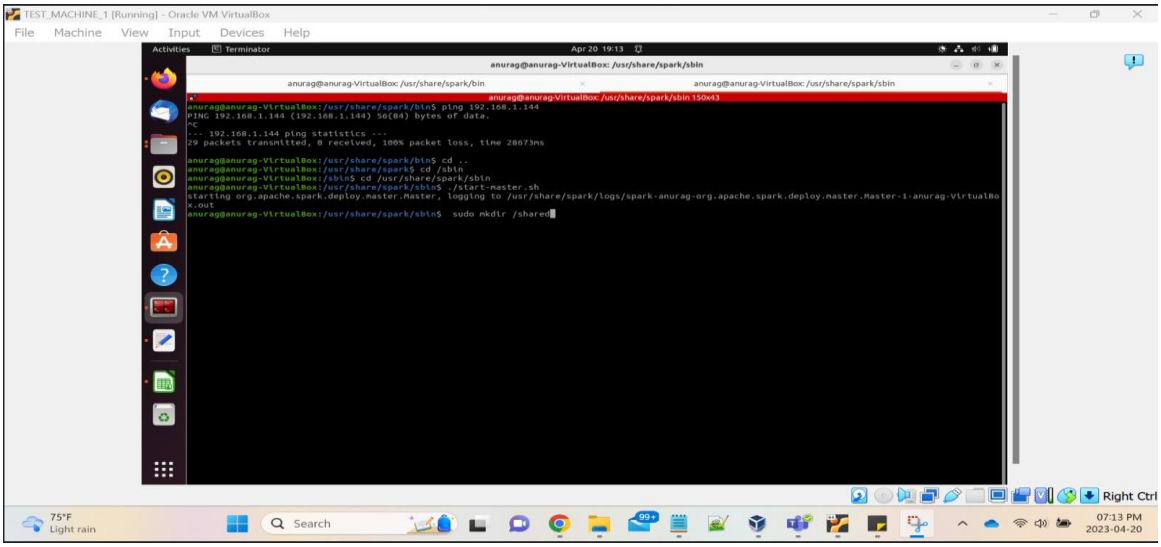
Steps to install NFS on both master nodes and worker nodes:

Below are the NFS screenshots :-

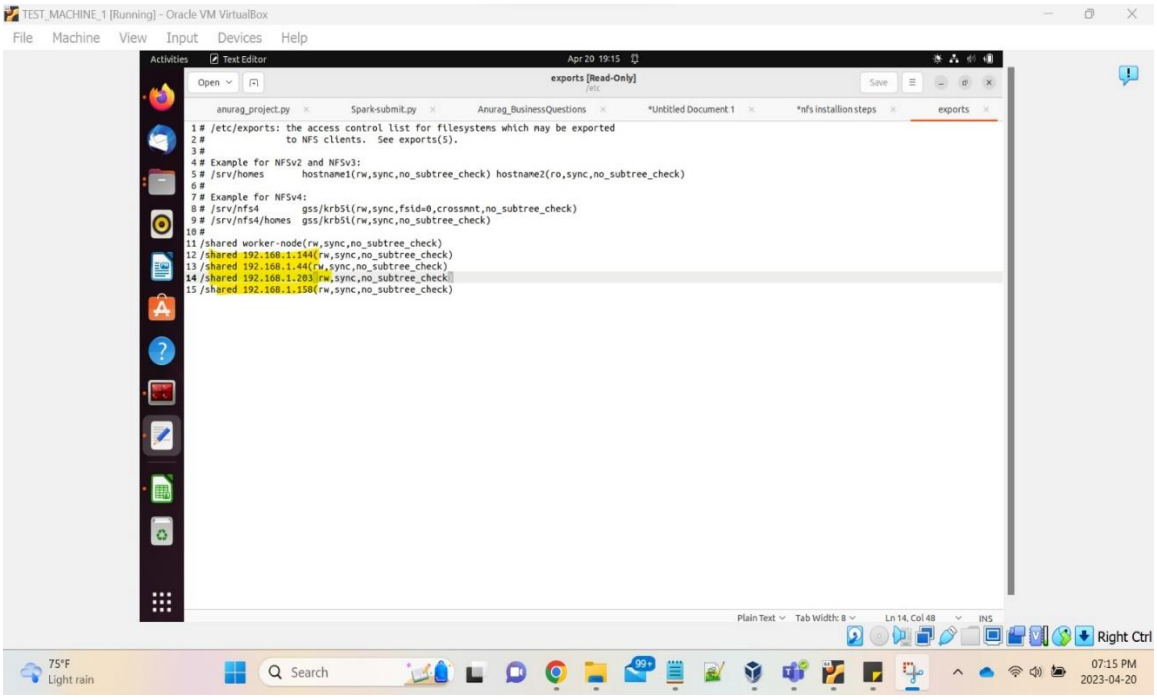
1. **Master node NFS installation (Anurag)**



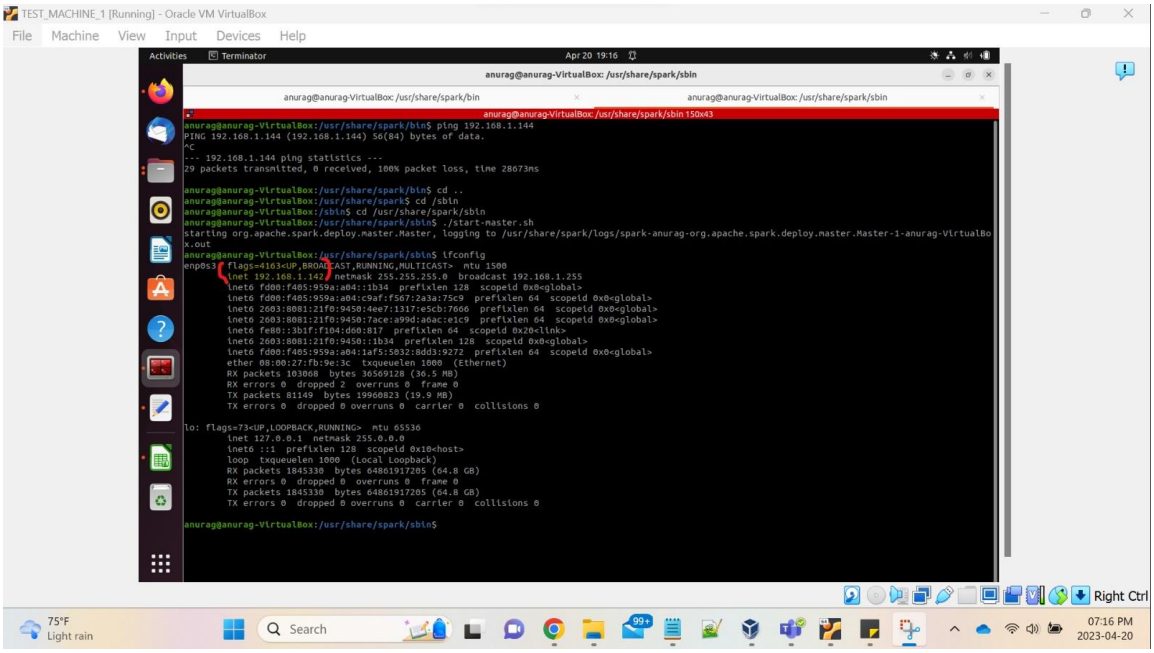
2. **Creating a shared folder in the master node**



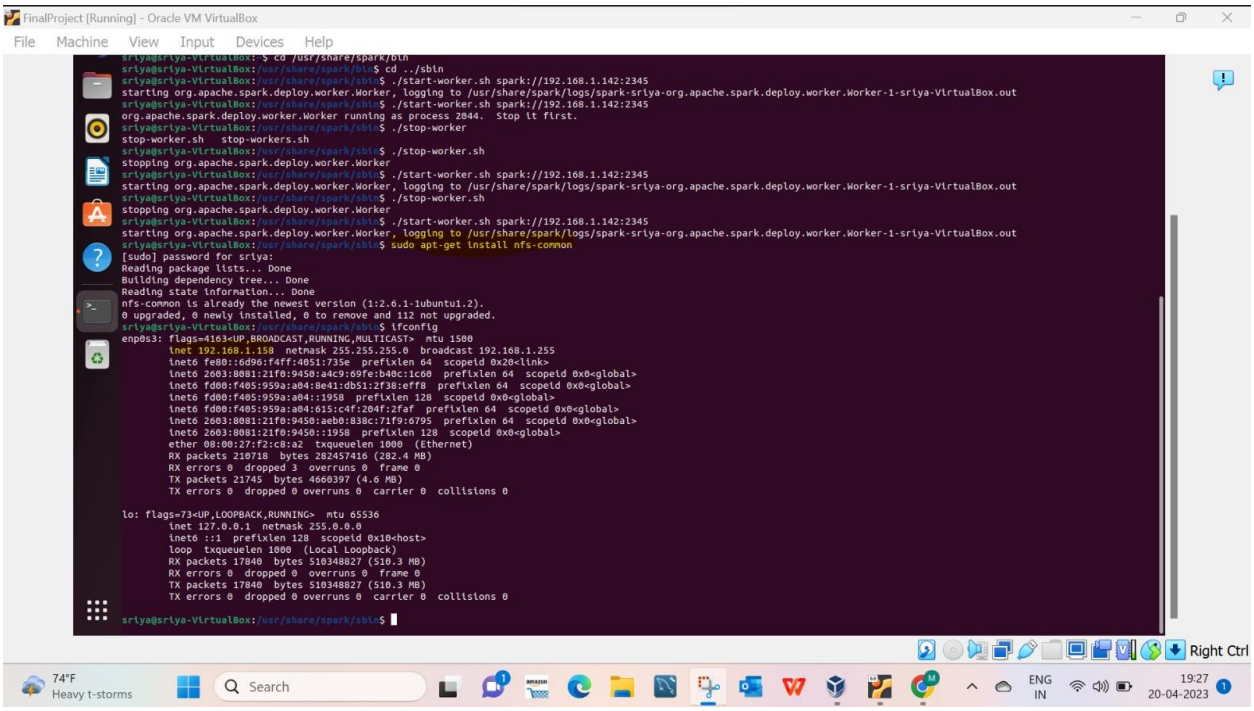
3. **Adding worker node ip's in the /etc/export file in order to give the worker nodes access to the shared folder.**



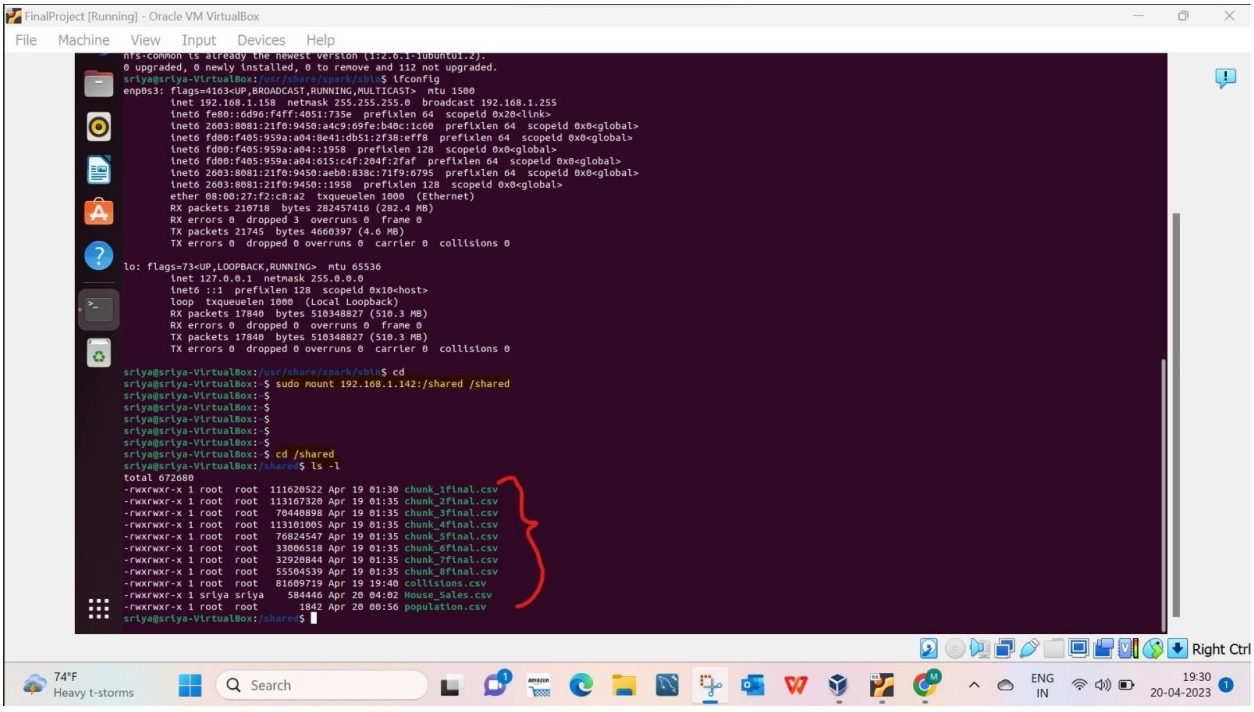
4. **IP address of the master node 192.168.1.142**



5. **Worker node nfs-common installation (192.168.1.158)**



6. **Accessing shared folder by mounting the master /shared location to the worker shared location.**



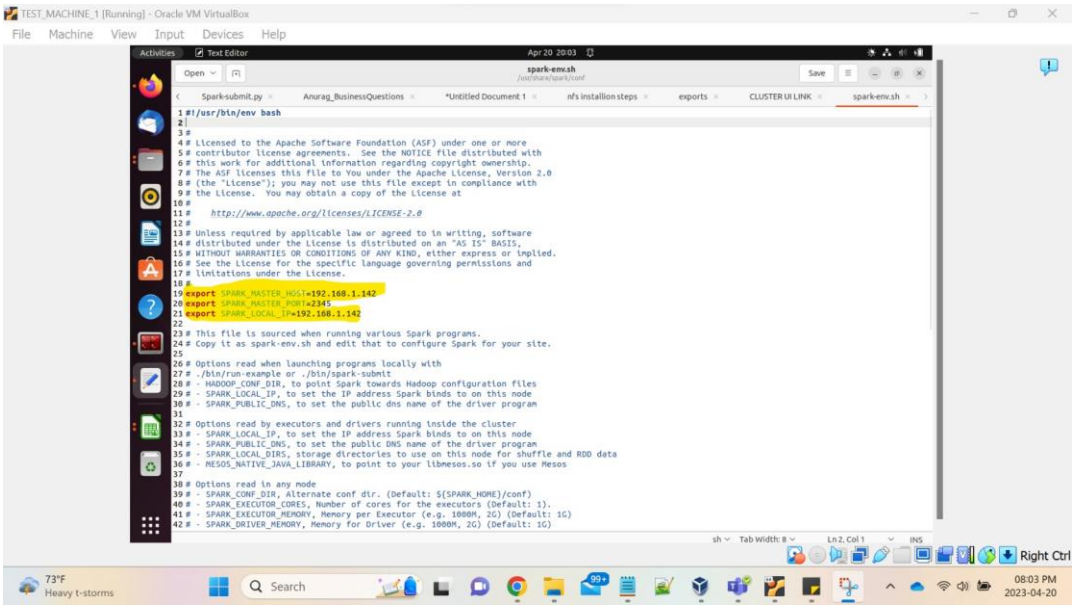
We have repeated this step for all the worker nodes like 192.168.1.144(Vinod) 192.168.1.44(Lekha), 192.168.1.158(Meghana). Now all the worker nodes have access to the shared folder from where the data is ingested into our program.

CLUSTER ARCHITECTURE:

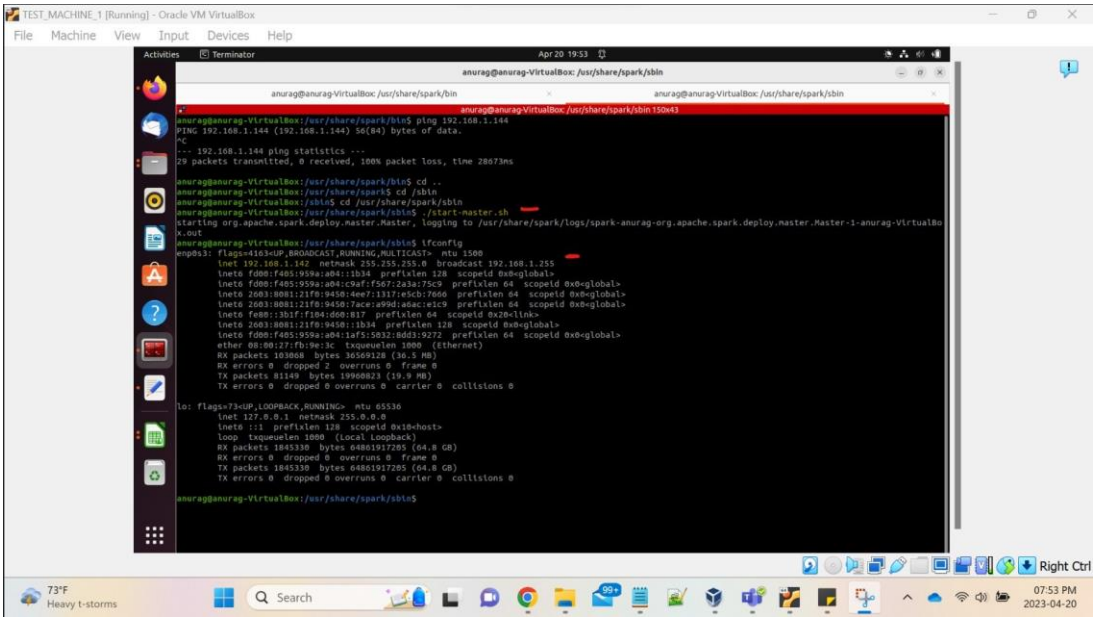
Steps performed to enable cluster processing:

Step 1:- We have set up the SPARK MASTER NODE and SPARK MASTER PORT configuration in /conf/spark-env.sh.

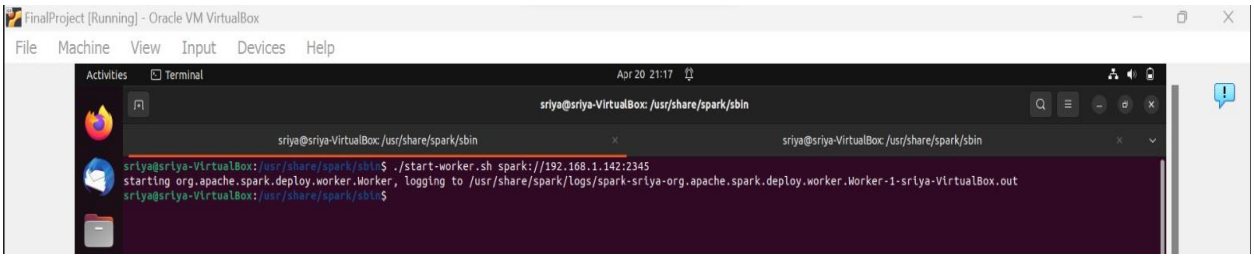
This configuration is set for all the Worker nodes to specify who the master node is and which port we need to use to connect to the master node.



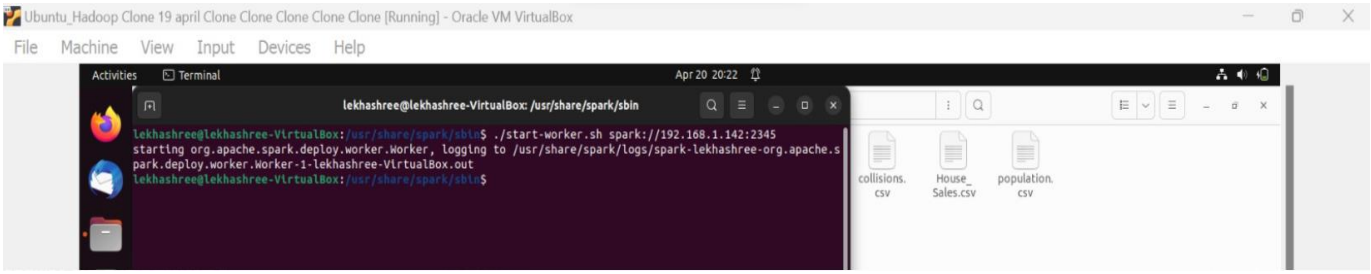
Step 2:- After adding the conf properties, we will have to run the Master node. using above command.



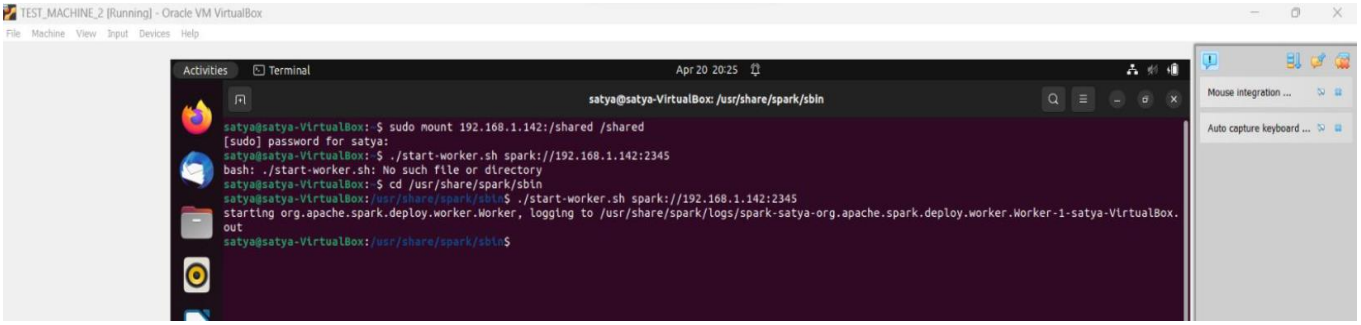
Step 3 – Start Worker nodes



Worker node 192.162.1.158(Meghana) started.

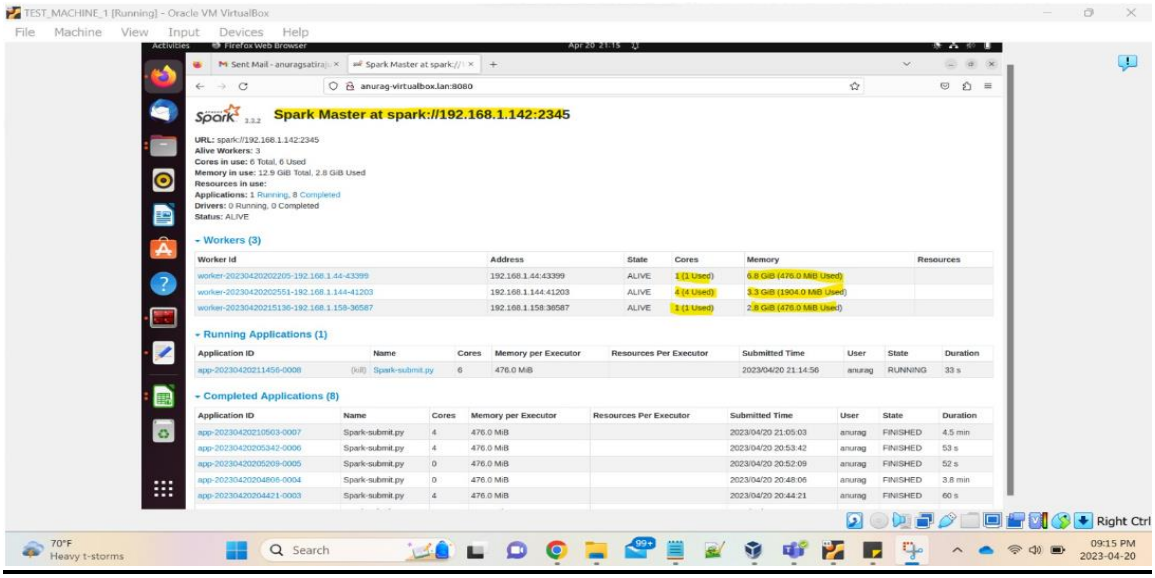


Worker node 192.168.1.44 (Lekha) started.

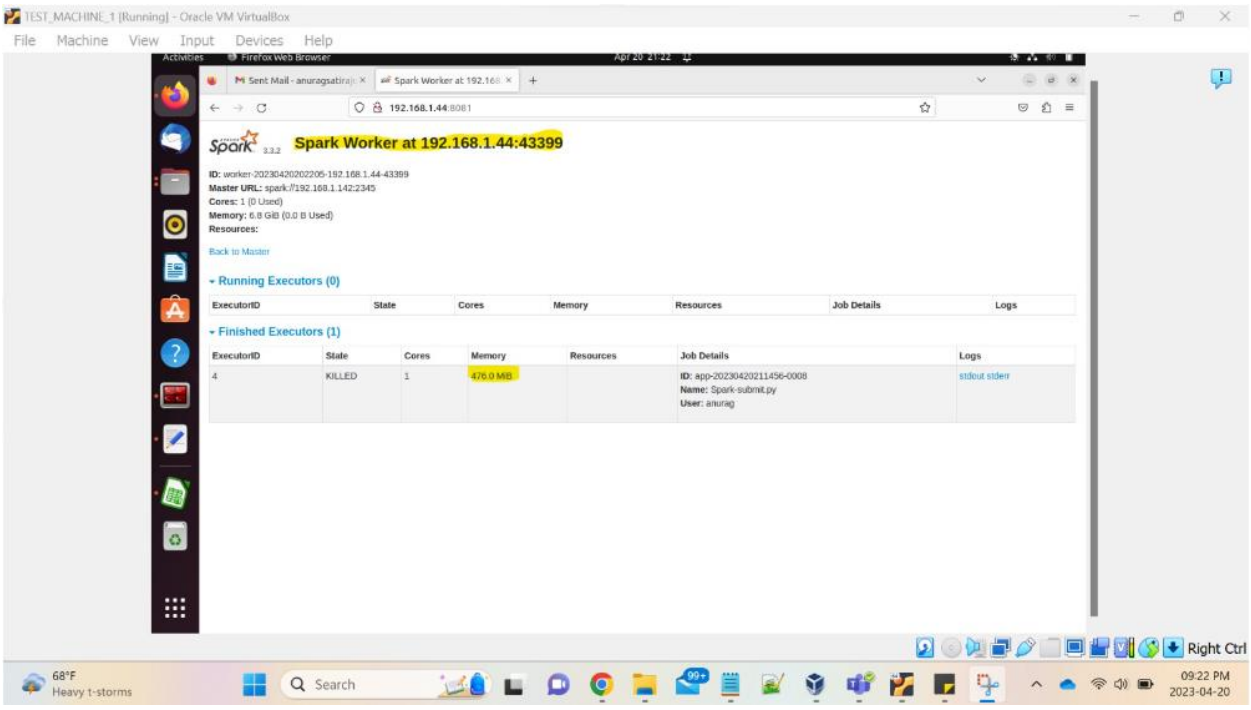


Worker node 192.168.1.144(Vinod) started.

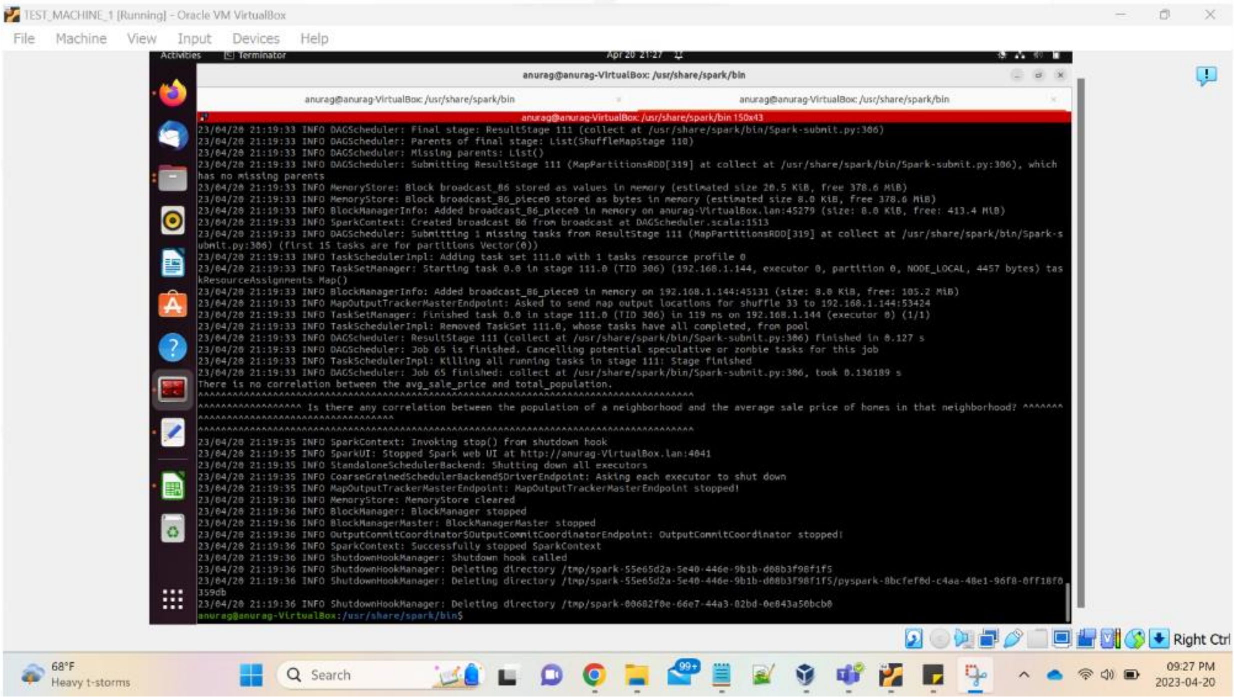
CLUSTER PROCESSING:



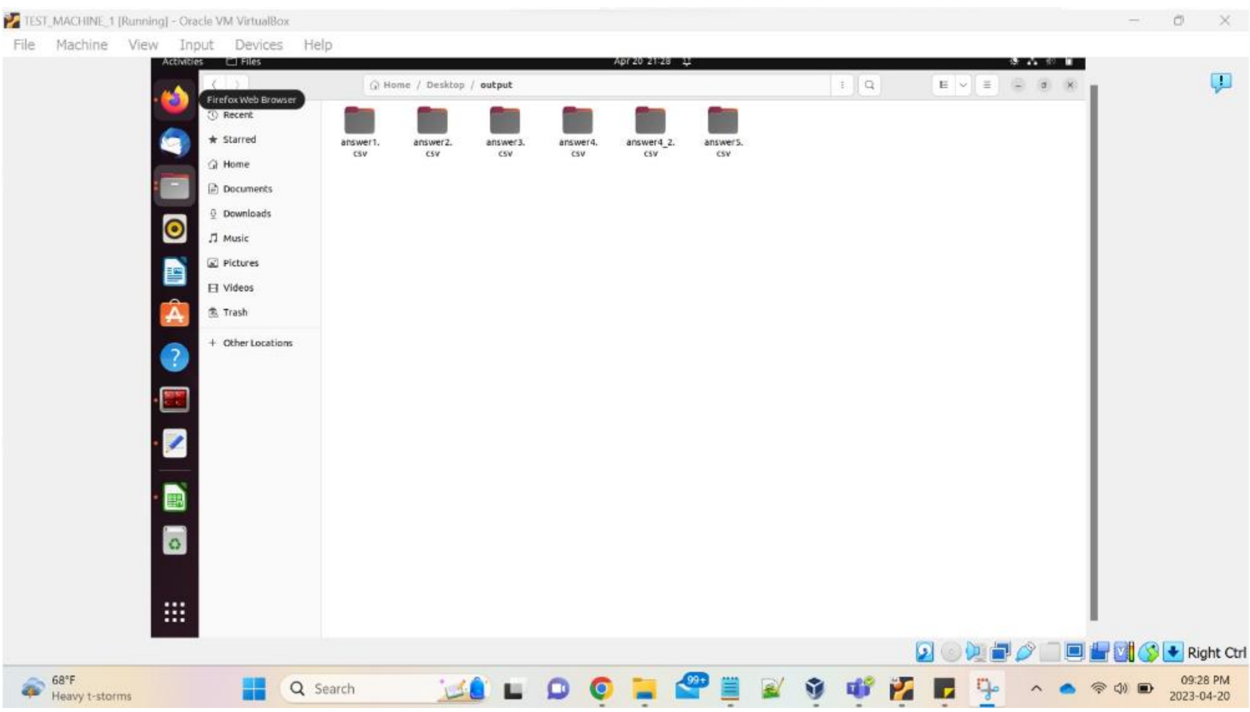
This is the Master Web UI where it shows the Number of worker nodes involved in the process and amount of memory used by each node for processing the data.



Log - one among the user node that shows the execution and the memory utilized for processing of the data.



Pyspark shell - background execution of the Spark-submit.py code.



The output file is written in this location. The result from the data analysis is written into a csv file and stored here.

BUSINESS QUESTIONS

1. ***How can we correlate the population of a neighborhood,crime rate and the average sale price of homes in that neighborhood?***

Business Conclusion:

There may be a correlation between the population of a neighborhood and the average sale price of homes in that neighborhood, although other factors may also play a role in determining home prices.

In general, areas with larger populations may have higher demand for housing, which could drive up prices. Additionally, neighborhoods with higher populations may be more attractive to businesses, leading to more jobs and economic activity in the area, which could also contribute to higher home prices.

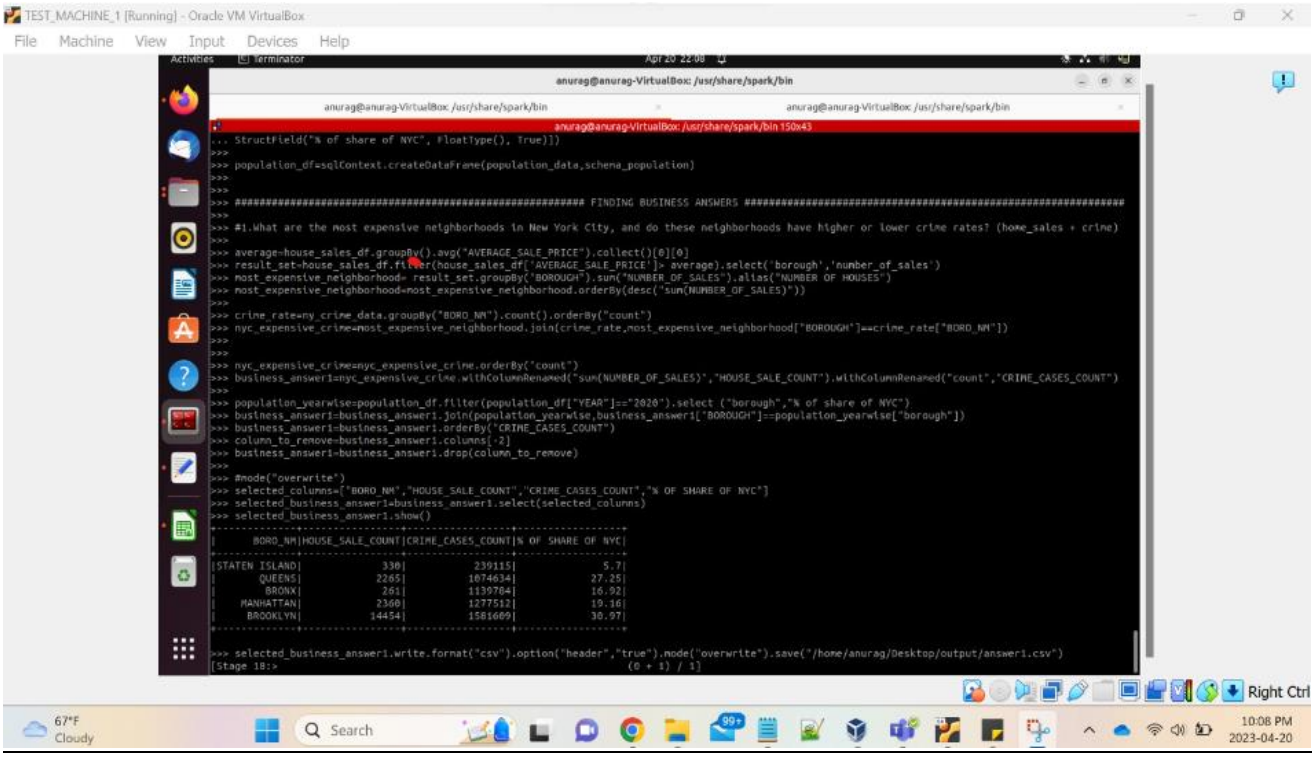
However, it is important to note that other factors, such as the quality of schools, crime rates, and access to amenities such as parks and shopping centers, can also play a significant role in determining home prices. As such, it is important to consider a variety of factors when analyzing the relationship between population and home prices.

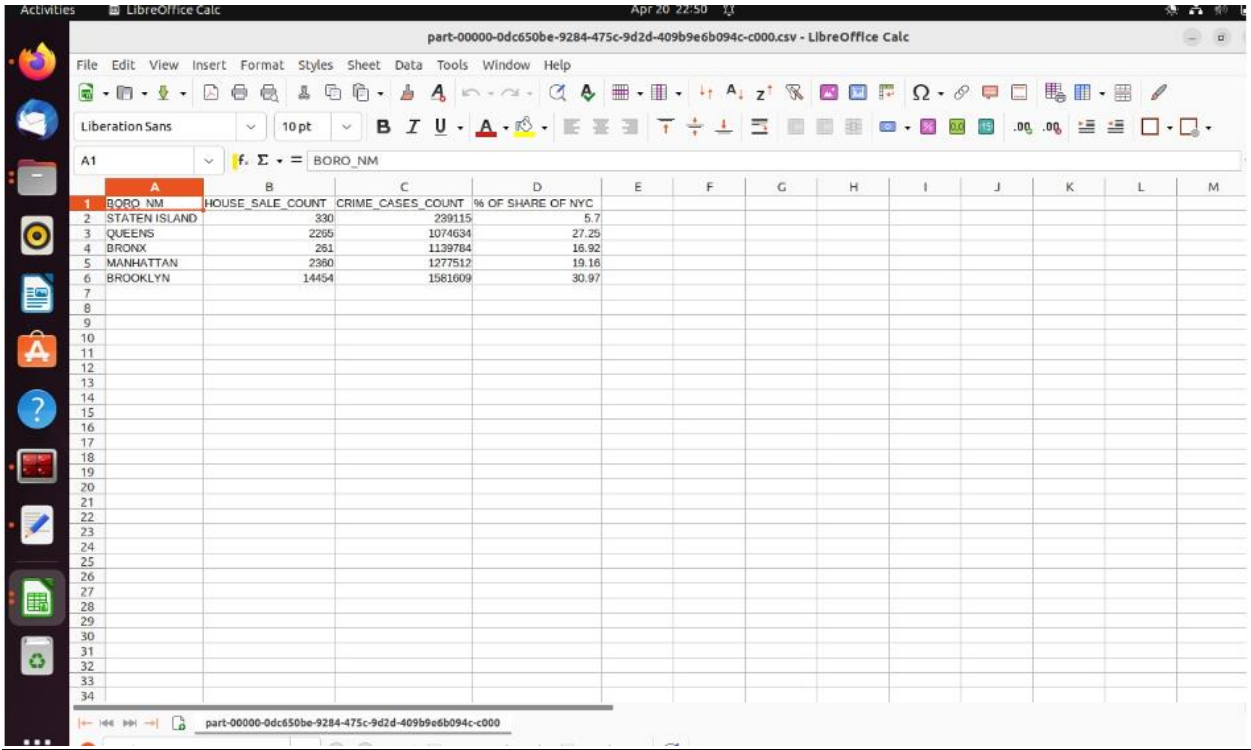
To determine whether there is a correlation between population and home prices in a particular neighborhood, you could analyze data on population and home sales in the area. One approach might be to plot population against home sale prices on a scatter plot, and then calculate the correlation coefficient between the two variables. A positive correlation coefficient would suggest that as population increases, home sale prices also tend to increase.

Overall, understanding the relationship between population and home prices can be useful for businesses and individuals who are looking to invest in real estate or develop new properties. By analyzing data on population and home prices, businesses can gain insights into the factors that influence demand for housing in a particular area and make more informed decisions about where to invest resources.

We can prevent this from happening by installing cameras, increased patrolling, increasing penalties for violent crime etc.

Output:


TEST_MACHINE_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
anurag@anurag-VirtualBox: /usr/share/spark/bin
>>> StructField("% of share of NYC", FloatType(), True)])
>>> population_df=sqlContext.createDataFrame(population_data,schena_population)
>>> ########## FINDING BUSINESS ANSWERS ##########
>>> #1.What are the most expensive neighborhoods in New York City, and do these neighborhoods have higher or lower crime rates? (home_sales + crime)
>>> average=house_sales_df.groupBy().avg("AVERAGE_SALE_PRICE").collect()[0][0]
>>> result_set=house_sales_df.filter(house_sales_df['AVERAGE_SALE_PRICE']> average).select('borough','number_of_sales')
>>> most_expensive_neighborhood= result_set.groupBy("BOROUGH").sum("NUMBER_OF_SALES").alias("NUMBER OF HOUSES")
>>> most_expensive_neighborhood=most_expensive_neighborhood.orderBy(desc('sum(NUMBER_OF_SALES)'))
>>> crime_rate=ny_crime_data.groupBy("BORO_NM").count().orderBy("count")
>>> nyc_expensive_crime=most_expensive_neighborhood.join(crime_rate,most_expensive_neighborhood['BOROUGH']==crime_rate["BORO_NM"])
>>> nyc_expensive_crime=nyc_expensive_crime.orderBy('count')
>>> business_answer1=nyc_expensive_crime.withColumnRenamed("sum(NUMBER_OF_SALES)","HOUSE_SALE_COUNT").withColumnRenamed("count","CRIME_CASES_COUNT")
>>> population_yearwise=population_df.filter(population_df["YEAR"]=="2020").select ("borough","% of share of NYC")
>>> business_answer1=business_answer1.join(population_yearwise,business_answer1['BOROUGH']==population_yearwise["borough"])
>>> business_answer1=business_answer1.orderBy('CRIME_CASES_COUNT')
>>> column_to_remove=business_answer1.columns[-2]
>>> business_answer1=business_answer1.drop(column_to_remove)
>>> #mode("overwrite")
>>> selected_columns=["BORO_NM","HOUSE_SALE_COUNT","CRIME_CASES_COUNT","% OF SHARE OF NYC"]
>>> selected_business_answer1=business_answer1.select(selected_columns)
>>> selected_business_answer1.show()
+-------------+---------------+----------------+----------------+
| BORO_NM|HOUSE_SALE_COUNT|CRIME_CASES_COUNT|% OF SHARE OF NYC|
+-------------+---------------+----------------+----------------+
|STATEN ISLAND| 330| 239115| 5.7|
QUEENS	2265	1074634	27.25
BRONX	261	1139704	16.92
MANHATTAN	2360	1277512	19.16
BROOKLYN	14454	1581609	30.97
+-------------+---------------+----------------+----------------+
>>> selected_business_answer1.write.format("csv").option("header","true").mode("overwrite").save("/home/anurag/Desktop/output/answer1.csv")
[Stage 18:> (0 + 1) / 1]
Right Ctrl
67°F Cloudy
Search
10:08 PM 2023-04-20


Activities
LibreOffice Calc
Apr 20 22:50
part-00000-0dc650be-9284-475c-9d2d-409b9e6b094c-c000.csv - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Liberation Sans
10 pt
A1
BORO_NM
BORO_NM HOUSE_SALE_COUNT CRIME_CASES_COUNT % OF SHARE OF NYC
STATEN ISLAND 330 239115 5.7
QUEENS 2265 1074634 27.25
BRONX 261 1139784 16.92
MANHATTAN 2360 1277512 19.16
BROOKLYN 14454 1581609 30.97
part-00000-0dc650be-9284-475c-9d2d-409b9e6b094c-c000

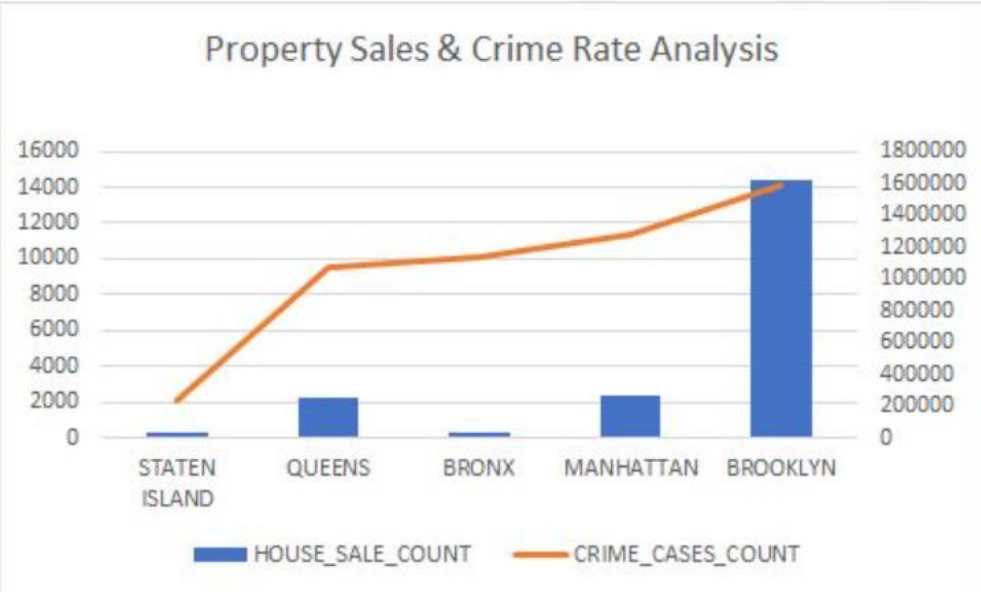
SQL Code:

```
average=house_sales_df.groupBy().avg("AVERAGE_SALE_PRICE").collect()[0][0]
result_set=house_sales_df.filter(house_sales_df['AVERAGE_SALE_PRICE']>
average).select('borough','number_of_sales')
most_expensive_neighborhood=result_set.groupBy("BOROUGH").sum("NUMBER_OF_S
ALES").alias("NUMBER OF HOUSES")
most_expensive_neighborhood=most_expensive_neighborhood.orderBy(desc("sum(
NUMBER_OF_SALES)"))

crime_rate=ny_crime_data.groupBy("BORO_NM").count().orderBy("count")
nyc_expensive_crime=most_expensive_neighborhood.join(crime_rate,most_expen
sive_neighborhood["BOROUGH"]==crime_rate["BORO_NM"])


nyc_expensive_crime=nyc_expensive_crime.orderBy("count")
business_answer1=nyc_expensive_crime.withColumnRenamed("sum(NUMBER_OF_SALE
S)","HOUSE_SALE_COUNT").withColumnRenamed("count","CRIME_CASES_COUNT")

population_yearwise=population_df.filter(population_df["YEAR"]=="2020").se
lect ("borough","% of share of NYC")
business_answer1=business_answer1.join(population_yearwise,business_answer
1["BOROUGH"]==population_yearwise["borough"])
business_answer1=business_answer1.orderBy("CRIME_CASES_COUNT")
column_to_remove=business_answer1.columns[-2]
business_answer1=business_answer1.drop(column_to_remove)

#mode("overwrite")
selected_columns=["BORO_NM","HOUSE_SALE_COUNT","CRIME_CASES_COUNT","% OF
SHARE OF NYC"]
selected_business_answer1=business_answer1.select(selected_columns)
selected_business_answer1.show()
```

Code Output Explanation:

Based on the analysis of New York house sales data and crime data, it can be concluded that the neighborhoods with the highest number of house sales in New York City tend to have higher crime rates than other neighborhoods, and are typically the most densely populated areas. Specifically, the analysis found that Brooklyn had the highest number of both house sales and crime cases followed by Manhattan.

Visualization:

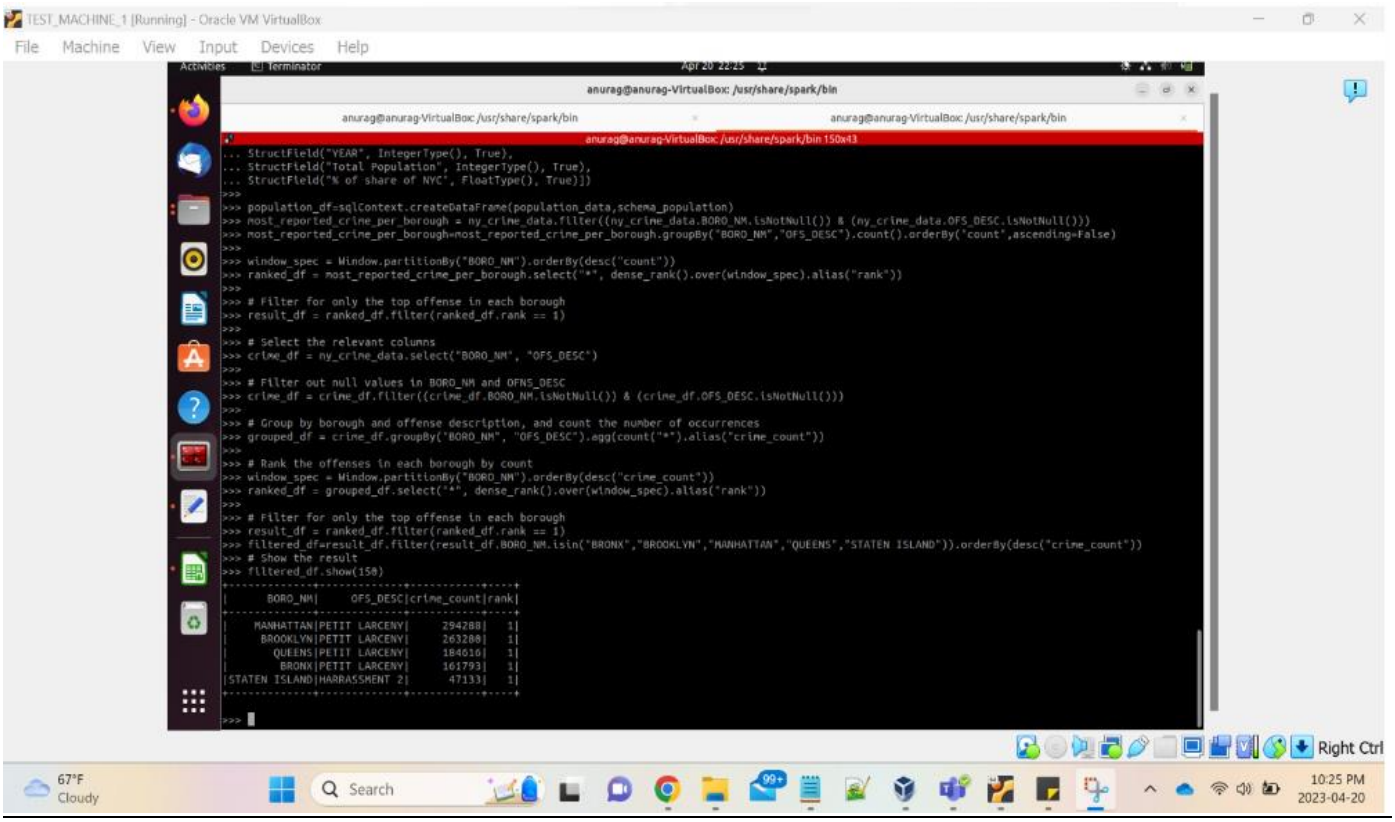


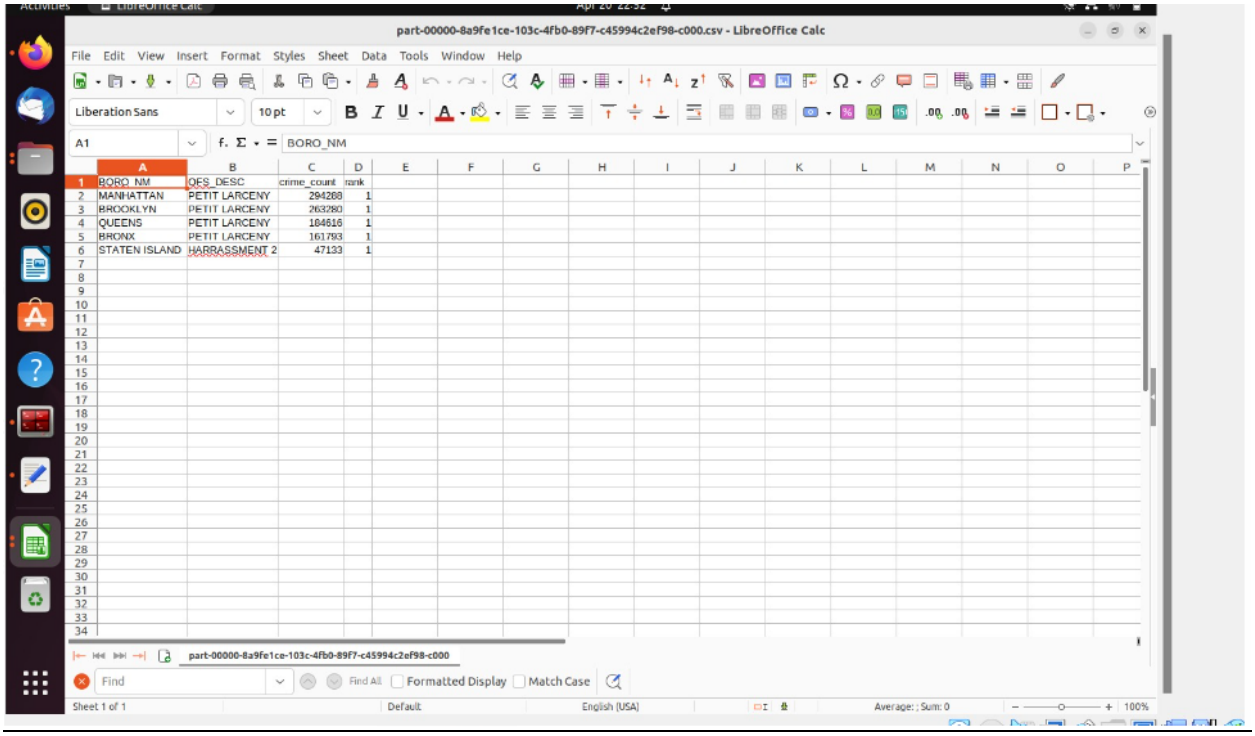
2. ***What type of crimes are most commonly reported in each borough, and are there any patterns or trends that can be identified?***

Business Conclusion:

Petit larceny and harassment are two of the most commonly committed crimes in several boroughs. To tackle these issues, several strategies can be employed. Firstly, increasing law enforcement presence in areas where these crimes are prevalent can act as a deterrent and lead to more arrests and prosecutions. Secondly, the installation of surveillance cameras in public spaces such as parking lots and parks can deter criminals and provide valuable evidence for law enforcement. Additionally, the use of social media can help raise awareness of these issues among the public. Ensuring proper lighting and visibility in public spaces is also essential in reducing crime, as poor lighting can make it easier for criminals to commit petit larceny and harassment. Addressing underlying social and economic issues, such as poverty and substance abuse, can also contribute to reducing crime in the long run. However, it is important to note that reducing crime requires a comprehensive and sustained effort from law enforcement, community members, and other stakeholders.

Output:





SQL Code:

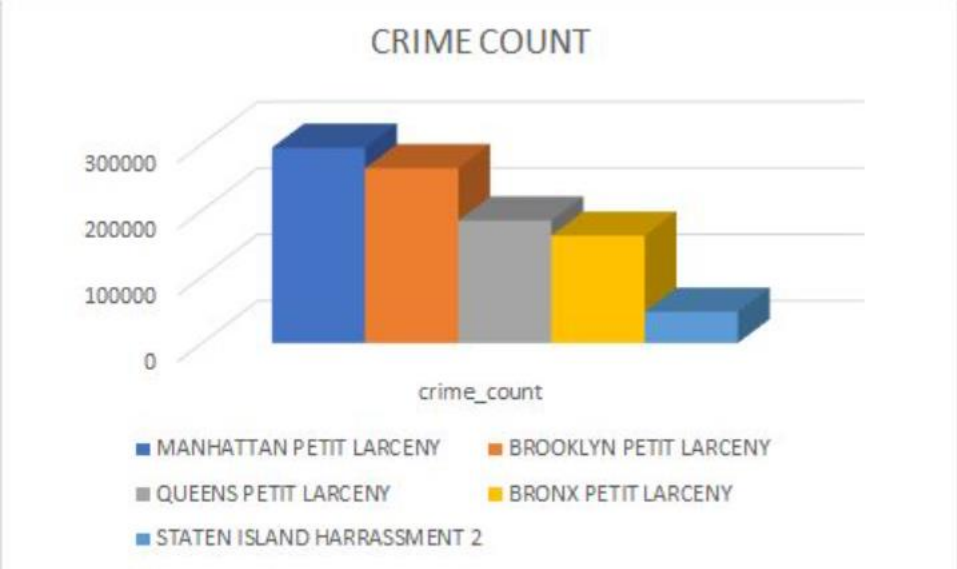
```
most_reported_crime_per_borough =
ny_crime_data.filter((ny_crime_data.BORO_NM.isNotNull()) &
(ny_crime_data.OFS_DESC.isNotNull()))
most_reported_crime_per_borough=most_reported_crime_per_borough.groupBy("B
ORO_NM","OFS_DESC").count().orderBy("count",ascending=False)

window_spec = Window.partitionBy("BORO_NM").orderBy(desc("count"))
ranked_df = most_reported_crime_per_borough.select("*",
dense_rank().over(window_spec).alias("rank"))

# Filter for only the top offense in each borough
result_df = ranked_df.filter(ranked_df.rank == 1)

# Select the relevant columns
crime_df = ny_crime_data.select("BORO_NM", "OFS_DESC")

# Filter out null values in BORO_NM and OFNS_DESC
crime_df = crime_df.filter((crime_df.BORO_NM.isNotNull()) &
(crime_df.OFS_DESC.isNotNull()))

# Group by borough and offense description, and count the number of
occurrences
grouped_df = crime_df.groupBy("BORO_NM",
"OFS_DESC").agg(count("*").alias("crime_count"))

# Rank the offenses in each borough by count
window_spec = Window.partitionBy("BORO_NM").orderBy(desc("crime_count"))
```

```
ranked_df = grouped_df.select("*",
dense_rank().over(window_spec).alias("rank"))

# Filter for only the top offense in each borough
result_df = ranked_df.filter(ranked_df.rank == 1)
filtered_df=result_df.filter(result_df.BORO_NM.isin("BRONX","BROOKLYN","MA
NHATTAN","QUEENS","STATEN ISLAND")).orderBy(desc("crime_count"))
```

Code Output Explanation:

According to the analysis of sales and crime data in New York, it can be inferred that the Manhattan neighborhoods with the highest crime rates have predominantly reported incidents of petit larceny, with a total count of over 250,000 reported crimes. This trend is followed by Brooklyn, Queens, and the Bronx. In contrast, the highest recorded incidents of harassment were reported in Staten Island.

Visualization:

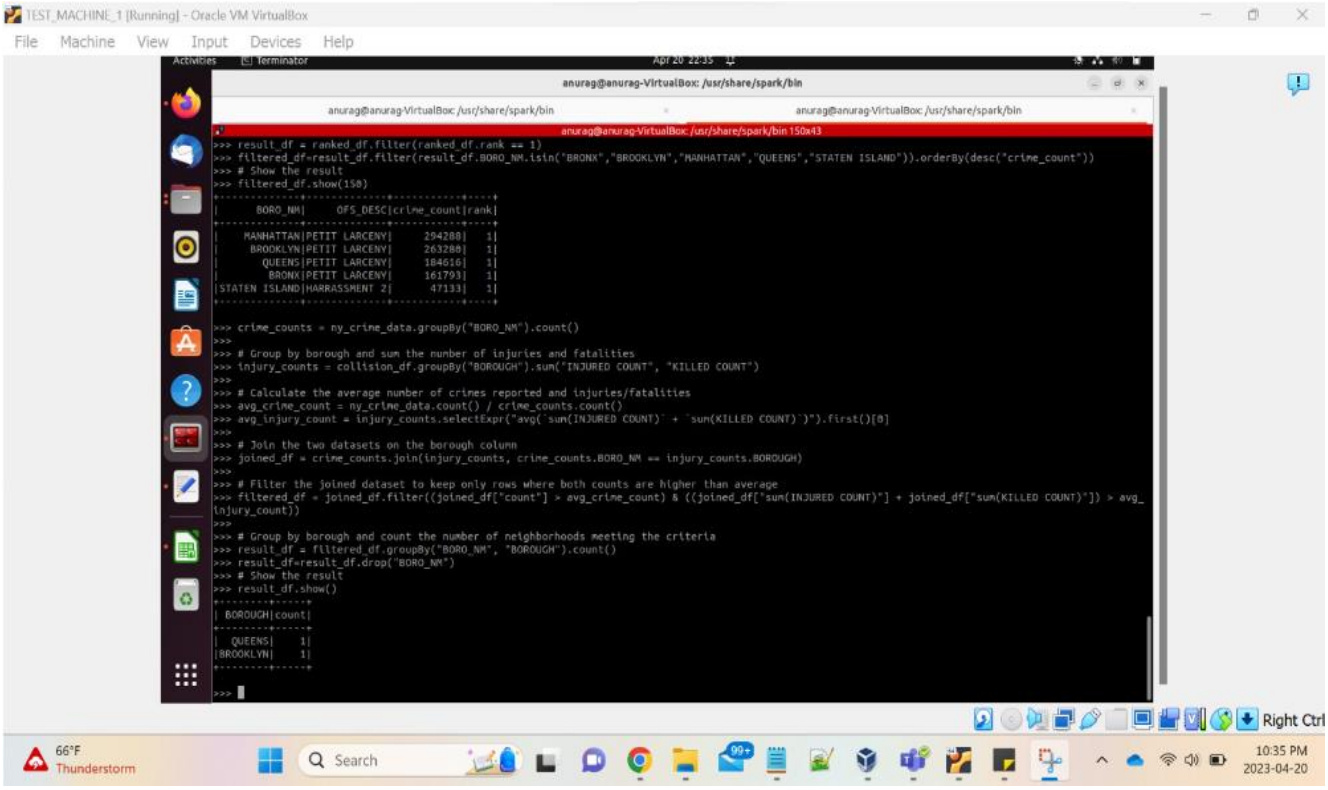


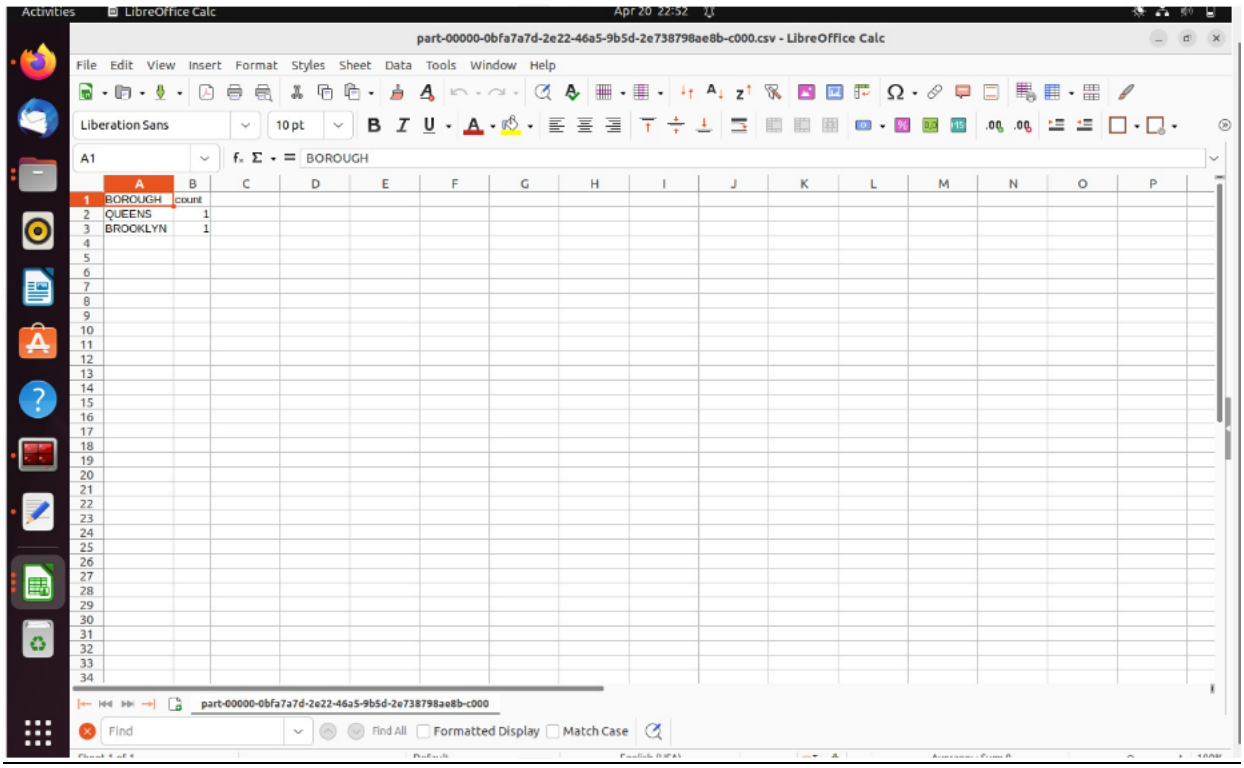
3. ***Are there any specific neighborhoods in New York City where both the number of crimes reported and the number of motor vehicle accidents where there is either injury or person killed are higher than average?***

Business Conclusion:

In neighborhoods with high rates of motor vehicle accidents and crime, businesses may need to adjust their operations to mitigate risk and address safety concerns. Insurance companies may need to increase premiums or adjust policies, while real estate companies may have an opportunity to revitalize distressed areas and improve safety. Car rental companies may need to offer additional safety features or reassurances, and businesses relying on transportation and logistics may need to invest in additional safety training and equipment. Technology companies specializing in safety and security may see increased demand, and emergency response services may also experience increased demand in these areas.

Output:





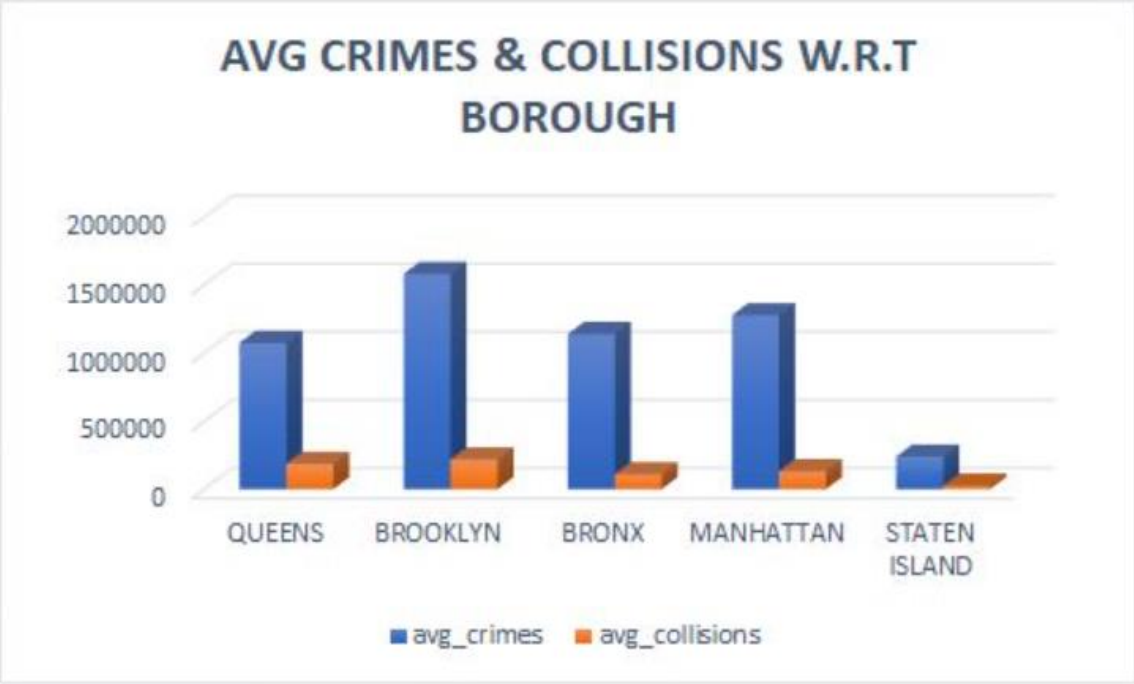
SQL Code:

```
crime_counts = ny_crime_data.groupBy("BORO_NM").count()

# Group by borough and sum the number of injuries and fatalities
injury_counts = collision_df.groupBy("BOROUGH").sum("INJURED COUNT",
"KILLED COUNT")

# Calculate the average number of crimes reported and injuries/fatalities
avg_crime_count = ny_crime_data.count() / crime_counts.count()
avg_injury_count = injury_counts.selectExpr("avg(`sum(INJURED COUNT)` +
`sum(KILLED COUNT)`)").first()[0]

# Join the two datasets on the borough column
joined_df = crime_counts.join(injury_counts, crime_counts.BORO_NM ==
injury_counts.BOROUGH)

# Filter the joined dataset to keep only rows where both counts are higher
than average
filtered_df = joined_df.filter((joined_df["count"] > avg_crime_count) &
((joined_df["sum(INJURED COUNT)"] + joined_df["sum(KILLED COUNT)"]) >
avg_injury_count))

# Group by borough and count the number of neighborhoods meeting the
criteria
result_df = filtered_df.groupBy("BORO_NM", "BOROUGH").count()
result_df=result_df.drop("BORO_NM")
# Show the result
result_df.show()
```

Code Output Explanation:

The analytics indicate that the incidence of reported crimes and motor vehicle accidents resulting in injury or fatality are above the average in neighborhoods located in Queens and Brooklyn, New York City.

Visualization:

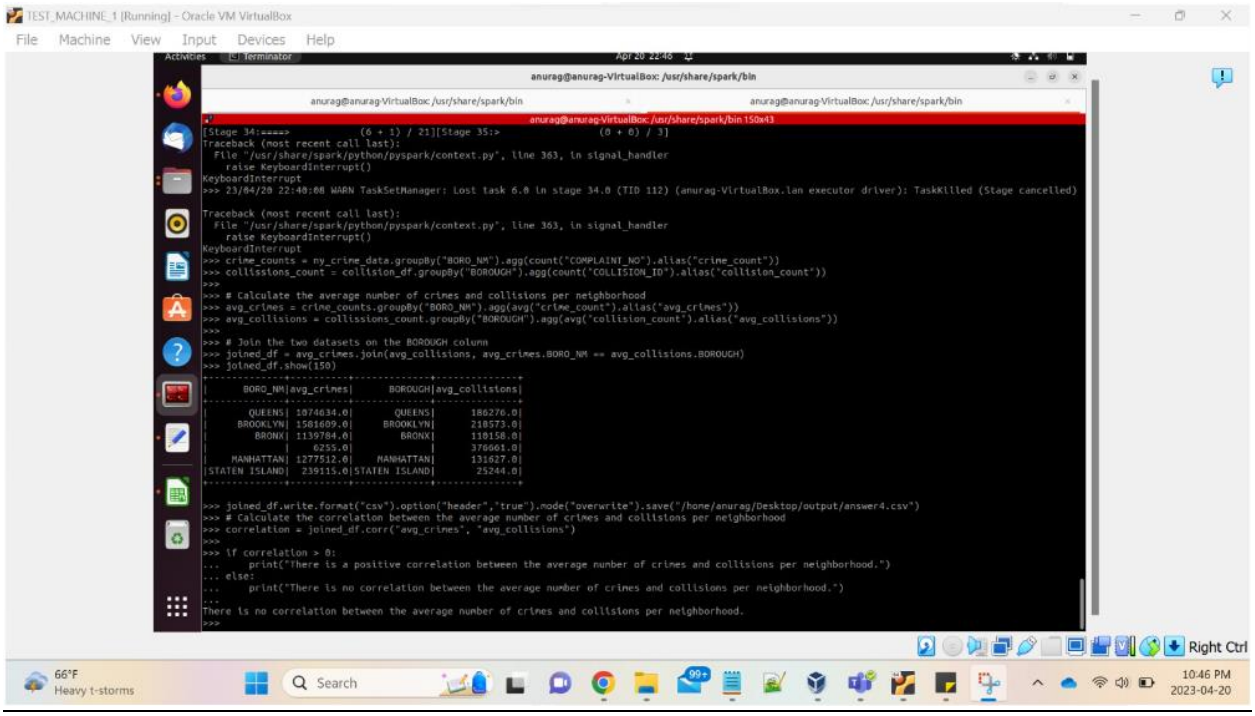


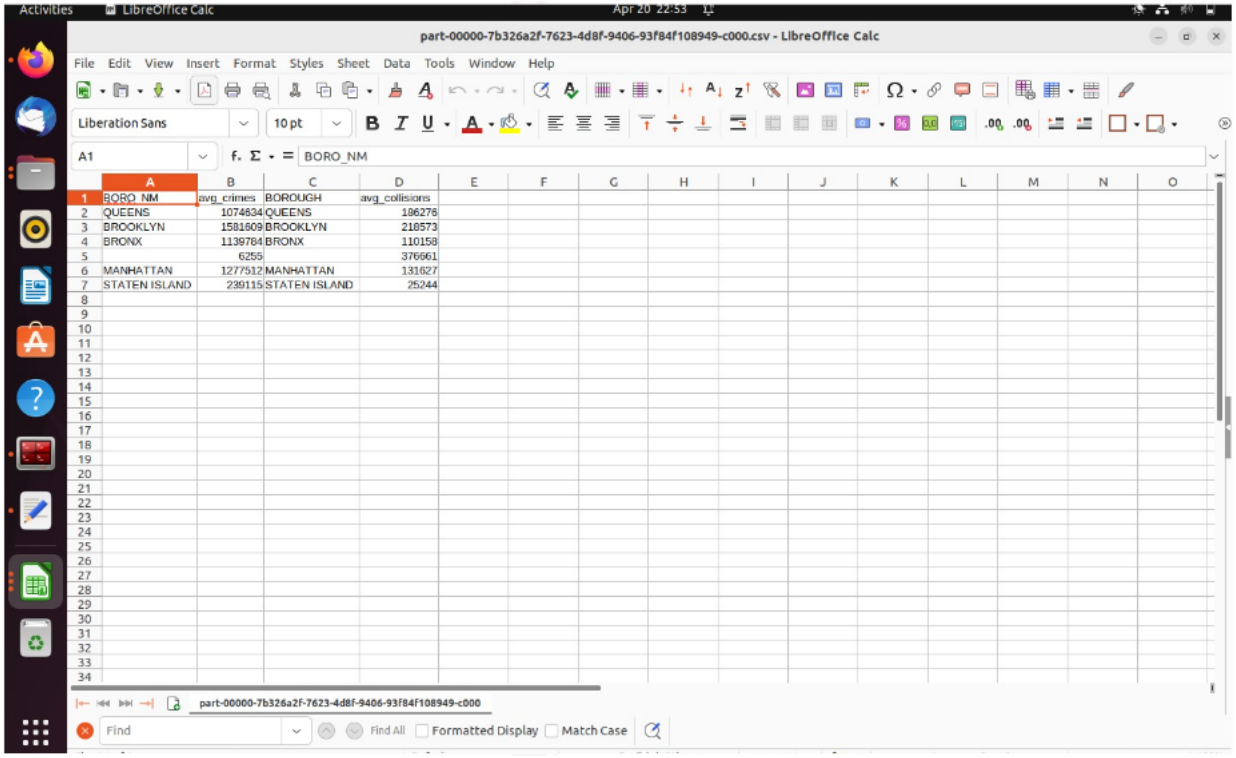
4. ***IS there any relation between crime and the accidents happening?***

Business Conclusion:

Companies in the automotive industry, such as car dealerships, auto parts stores, and safety equipment manufacturers, may benefit from increased demand for safety equipment in areas with high numbers of motor vehicle collisions. Real estate companies may need to invest in security upgrades to make their properties more attractive to potential tenants or buyers in high-crime and high-accident areas. Transportation companies may need to adjust their routes to avoid problem areas, and companies with fleets of vehicles may need to invest in additional safety training and equipment. Retail stores may need to implement additional security measures to deter theft and ensure safety. High-crime areas may also have higher rates of accidents due to poor road conditions or a lack of infrastructure.

Output:





SQL Code:

```
crime_counts =
ny_crime_data.groupBy("BORO_NM").agg(count("COMPLAINT_NO").alias("crime_co
unt"))
collissions_count =
collision_df.groupBy("BOROUGH").agg(count("COLLISION_ID").alias("collision
_count"))
# Calculate the average number of crimes and collisions per neighborhood
avg_crimes =
crime_counts.groupBy("BORO_NM").agg(avg("crime_count").alias("avg_crimes")
)
avg_collisions =
collissions_count.groupBy("BOROUGH").agg(avg("collision_count").alias("avg
_collisions"))
# Join the two datasets on the BOROUGH column
joined_df = avg_crimes.join(avg_collisions, avg_crimes.BORO_NM ==
avg_collisions.BOROUGH)
joined_df.write.format("csv").option("header","true").mode("overwrite").sa
ve("/home/anurag/Desktop/output/answer4.csv")
# Calculate the correlation between the average number of crimes and
collisions per neighborhood
correlation = joined_df.corr("avg_crimes", "avg_collisions")
if correlation > 0:
    print("There is a positive correlation between the average number of
crimes and collisions per neighborhood.")
else:
    print("There is no correlation between the average number of crimes
and collisions per neighborhood.")
```

Code Output Explanation:

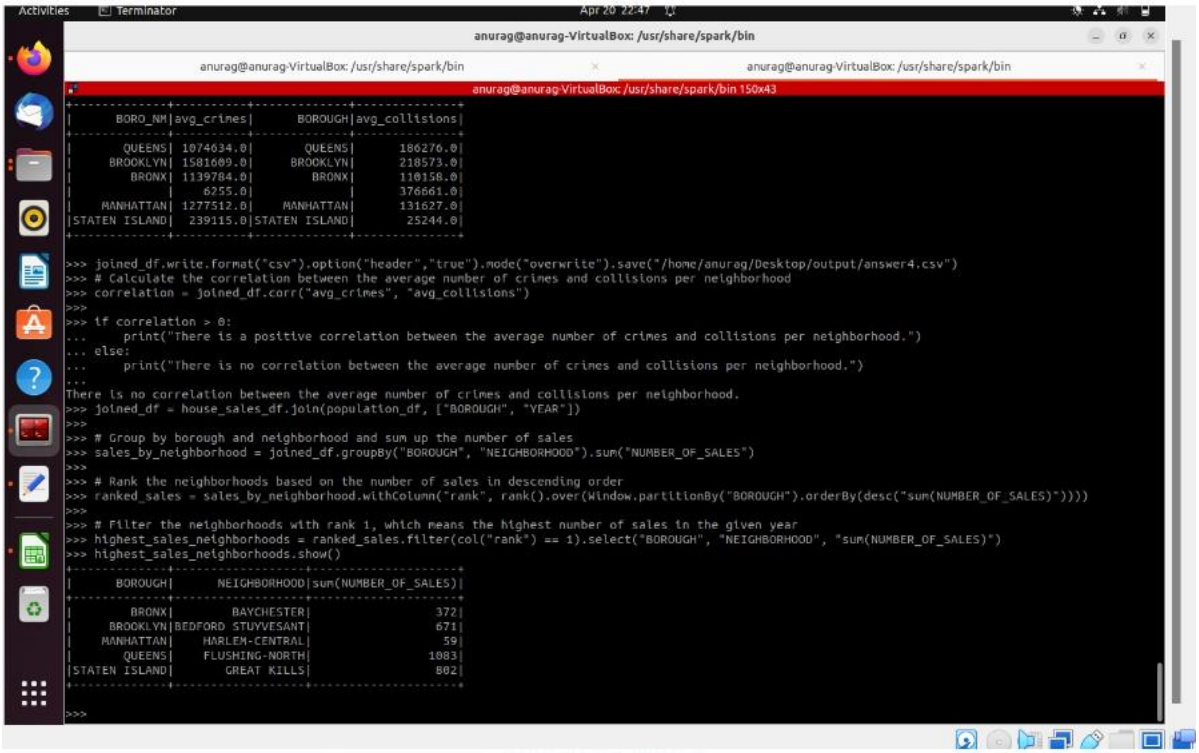
Based on our analysis, we have found that there is no correlation between the average number of motor vehicle collisions and the frequency of certain types of crimes such as theft or assault in specific areas. Manhattan had been recorded to have highest average number of crimes, followed by the Bronx, whereas Brooklyn had the second-highest average number of collisions, and the Bronx was fourth on the list. Therefore, we can conclude that there is no direct relationship between the occurrence of crimes and motor vehicle collisions in a given area.

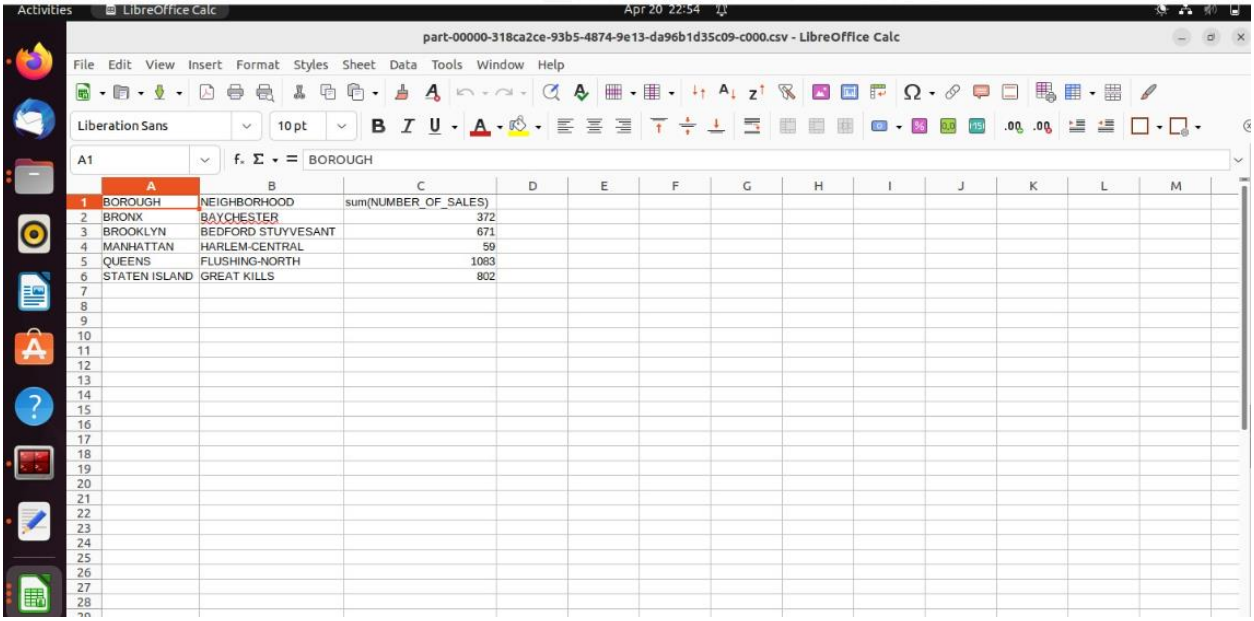
5.***Which neighborhood in each borough had the highest number of sales ?***

Business Conclusion:

The high number of property sales in a given neighborhood can be influenced by various factors such as location, affordability, housing quality, market conditions, effective marketing and advertising, investment and development, and demographic changes. Neighborhoods located in desirable areas close to employment hubs or popular attractions are often in high demand, as well as affordable neighborhoods that offer good value for money. Housing quality, favorable market conditions, effective marketing, and advertising campaigns can also attract potential buyers. Investment and development efforts, as well as demographic changes, such as an influx of young professionals or retirees, can further drive up demand and lead to increased property sales in a neighborhood.

Output:





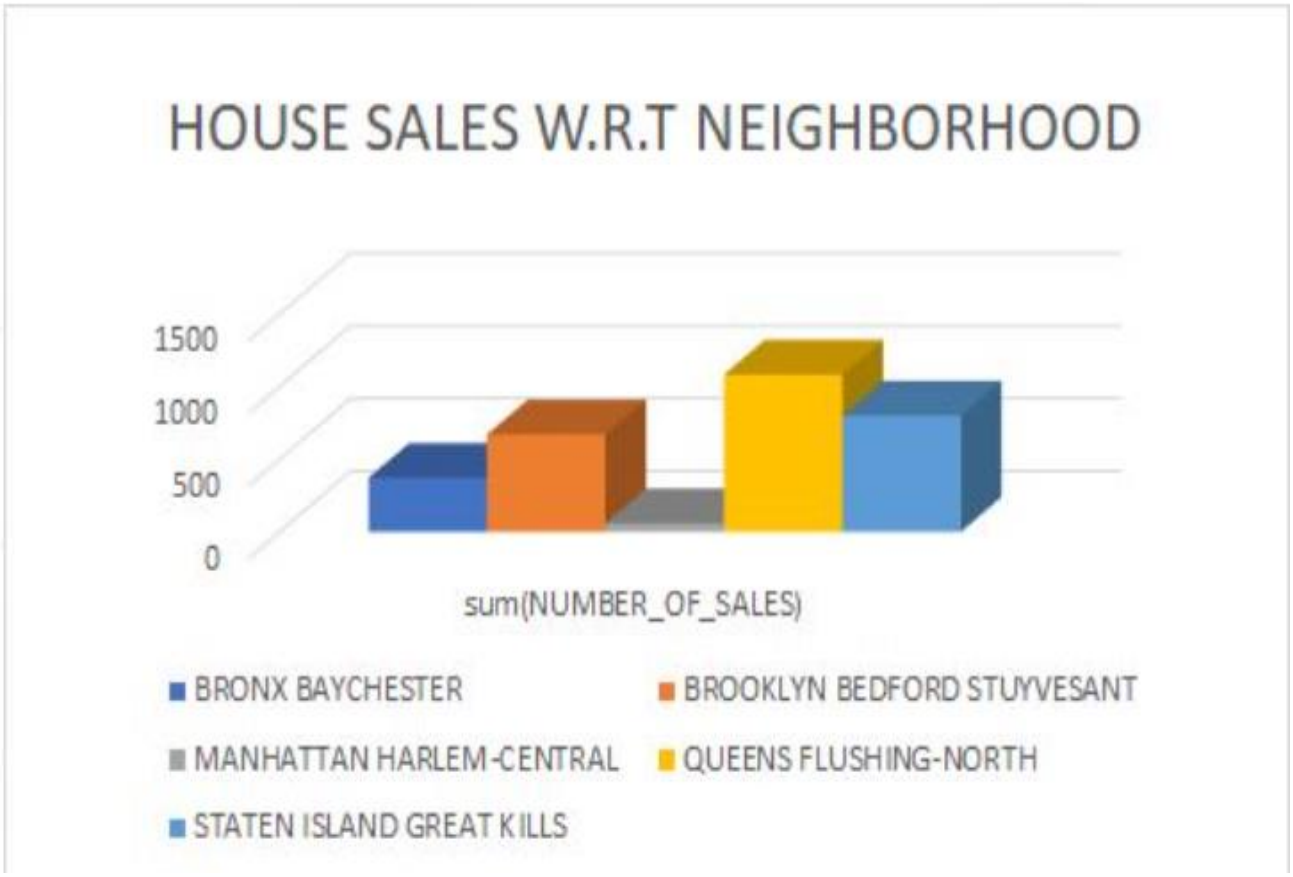
SQL Code:

```
# Join sales_data and population_data dataframes based on borough and year
joined_df = house_sales_df.join(population_df, ["BOROUGH", "YEAR"])

# Group by borough and neighborhood and sum up the number of sales
sales_by_neighborhood = joined_df.groupBy("BOROUGH",
"NEIGHBORHOOD").sum("NUMBER_OF_SALES")

# Rank the neighborhoods based on the number of sales in descending order
ranked_sales = sales_by_neighborhood.withColumn("rank",
rank().over(Window.partitionBy("BOROUGH").orderBy(desc("sum(NUMBER_OF_SALE
S)"))))

# Filter the neighborhoods with rank 1, which means the highest number of
sales in the given year
highest_sales_neighborhoods = ranked_sales.filter(col("rank") ==
1).select("BOROUGH", "NEIGHBORHOOD", "sum(NUMBER_OF_SALES)")
highest_sales_neighborhoods.show()
```

Code Output Explanation:

The analysis revealed that Flushing North neighborhood in Queens had the highest number of sales, with a total of approximately 1083. Great Kills neighborhood in Staten Island ranked second with around 802 sales. On the other hand, Harlem Central in Manhattan had the least number of sales, recording only 59 in the given year.

Visualization:

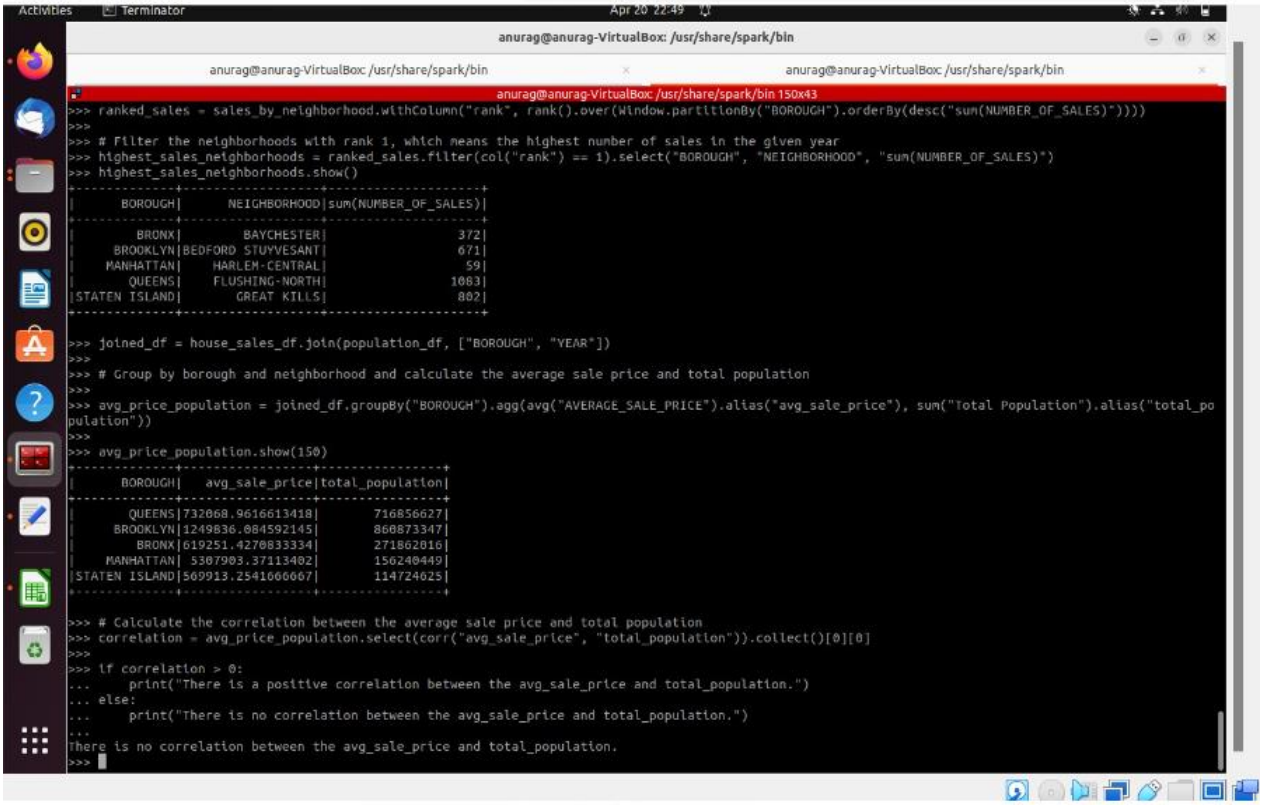


6. ***Is there any correlation between the population of a neighborhood and the average sale price of homes in that neighborhood?***

Business Conclusion:

The correlation between the population of a neighborhood and the average sale price of homes in that neighborhood can have several business implications. Real estate developers and investors can use this information to make informed decisions about where to focus their efforts, such as identifying neighborhoods with a growing population and potential for increased property values. It can also inform pricing strategies for properties in different neighborhoods, with higher population areas potentially commanding higher prices. Real estate agents can use this information to tailor their marketing and outreach efforts to attract potential buyers in areas with a specific population size. Additionally, businesses such as retail stores and restaurants can use this information to identify areas with a growing population to establish their presence and attract customers. Overall, understanding the correlation between population and home prices can inform various business decisions within the real estate and retail industries.

Output:



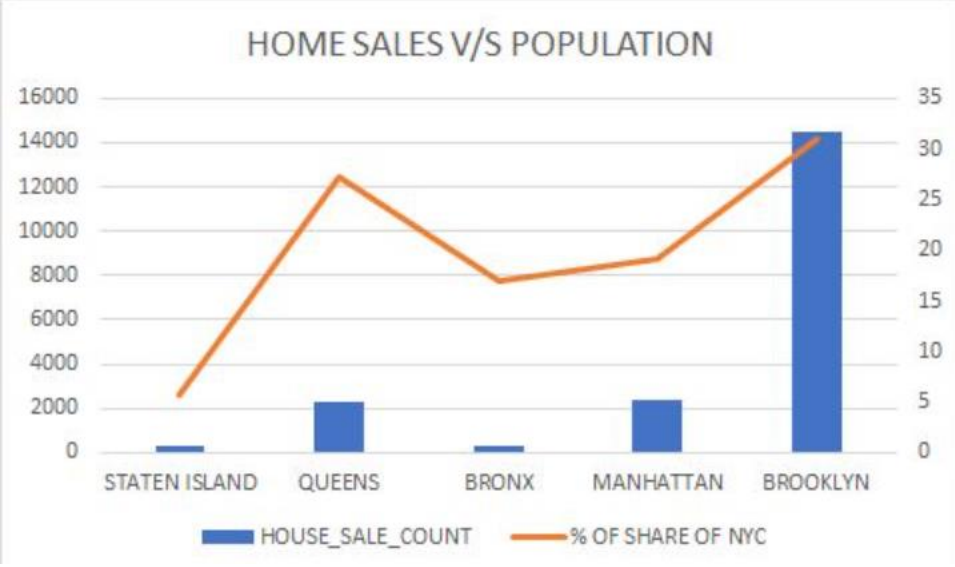
SQL Code:

```
joined_df = house_sales_df.join(population_df, ["BOROUGH", "YEAR"])

# Group by borough and neighborhood and calculate the average sale price
and total population

avg_price_population=joined_df.groupBy("BOROUGH").agg(avg("AVERAGE_SALE_PR
ICE").alias("avg_sale_price"), sum("Total
Population").alias("total_population"))

avg_price_population.show(150)
# Calculate the correlation between the average sale price and total
population
correlation = avg_price_population.select(corr("avg_sale_price",
"total_population")).collect()[0][0]

if correlation > 0:
    print("There is a positive correlation between the avg_sale_price and
total_population.")
else:
    print("There is no correlation between the avg_sale_price and
total_population.")
```

Code Output Explanation:

Based on the analysis, it can be concluded that there is no correlation or association between the population of a neighborhood and the average sale price of homes in that neighborhood.

Visualization:



REFERENCES

NEW YORK HOME SALES DATA: https://data.cityofnewyork.us/City-Government/DOF-Summary-of-Neighborhood-Sales-by-Neighborhood-/5ebm-myj7 (0.00578GB)

NYPD COMPLAINT DATA: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i (2.45GB)

MOTOR VEHICAL COLLISIONS DATA: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 (0.21GB)

NEW YORK POPULATION BY BOROUGH: https://data.cityofnewyork.us/City-Government/New-York-City-Population-by-Borough-1950-2040/xywu-7bv9 (2KB)

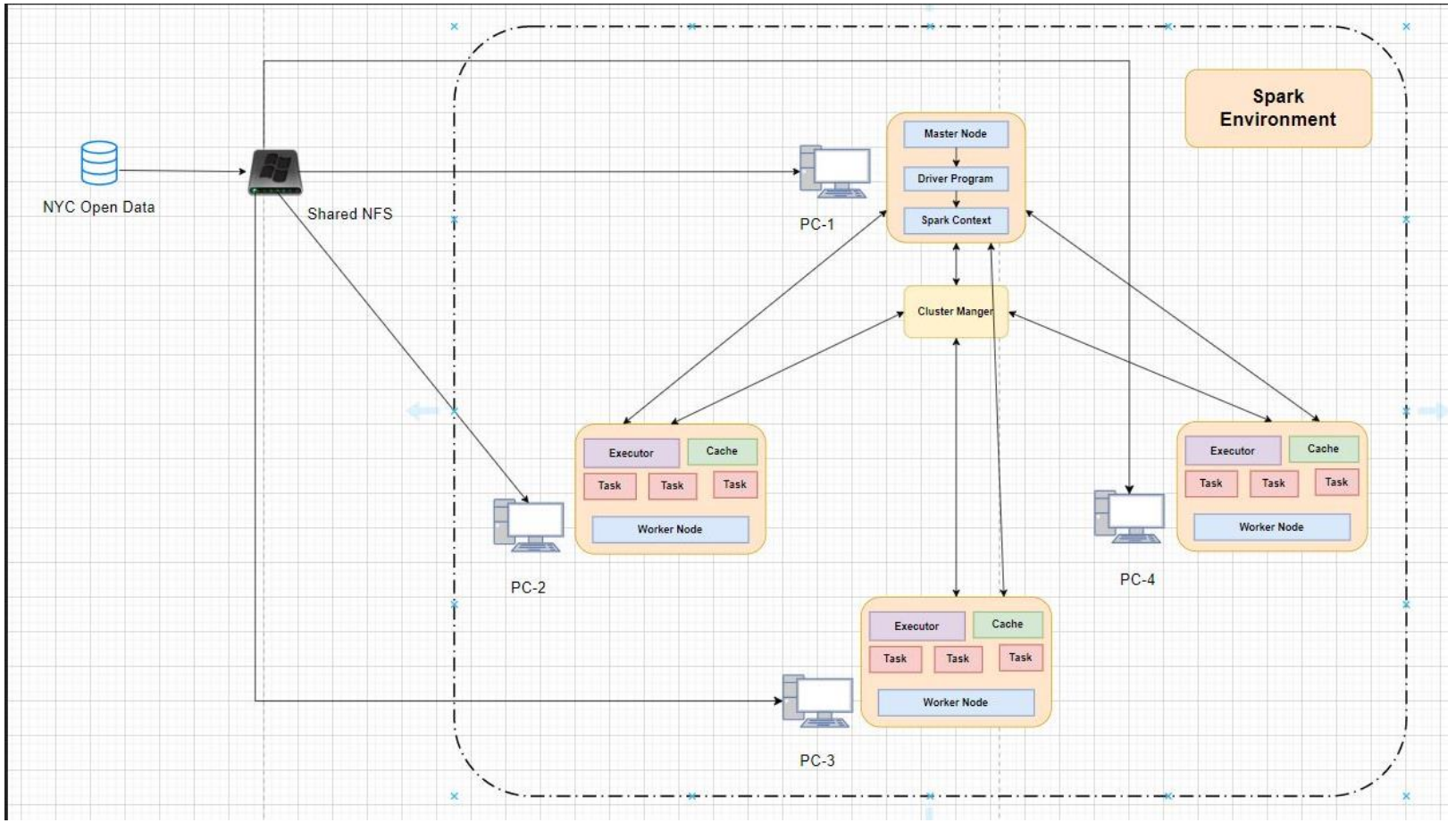
Total Dataset size = 2.7GB (approx..)

NextGen Analytical Solutions

Analysis of the Factors Contributing to Changes in Crime Rates and Property Values in New York City

Anurag S S S M
Lekhashree J
Meghana N
Satya Vinod K

ARCHITECTURE



- **Distributed Computing System**
- **Network File System (NFS)**
- **Master-Slave Architecture**
- **Data Loading**
- **Parallel Processing**

NFS

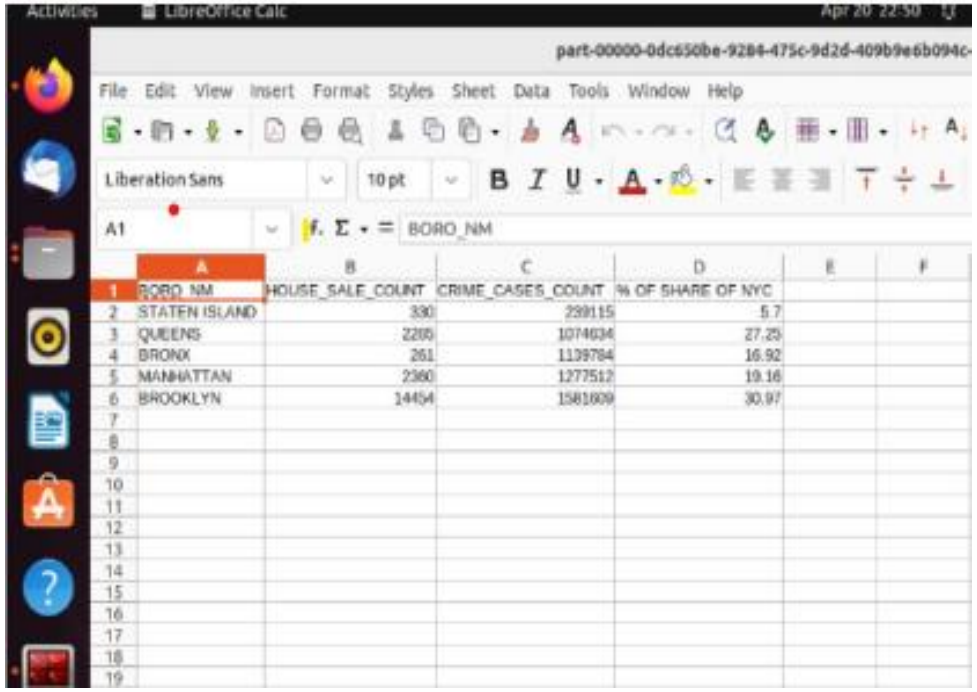
NFS (Network File System) is a distributed file system protocol that enables multiple users to access and share files over a network. It provides a scalable and centralized storage solution for big data applications, allowing data to be stored on a central server and accessed by multiple clients. NFS can improve the performance of big data applications by providing high-speed access to data, and it can handle large volumes of data transfers with minimal latency. Overall, NFS is a valuable tool for managing and storing big data.

Cluster architecture

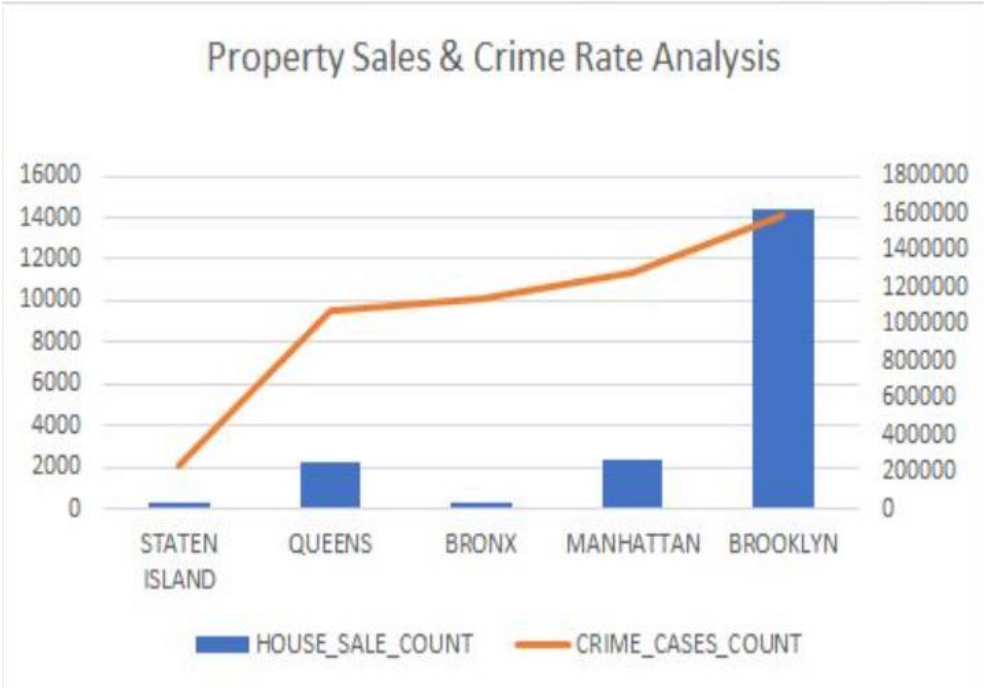
Cluster architecture is commonly used in large-scale computing systems, such as big data processing, scientific computing, and high-performance computing. It allows for efficient processing of large volumes of data by dividing the workload among multiple nodes, enabling parallel processing and reducing the time it takes to process data.

1. How can we correlate the population of a neighborhood,crime rate and the average sale price of homes in that neighborhood?

The correlation between population, crime rate, and home prices in a neighborhood can affect several businesses. Real estate and homebuilding businesses can identify profitable investment opportunities based on this data, while retail businesses can benefit from areas with higher populations and purchasing power. Security system and neighborhood watch businesses can benefit from areas with higher crime rates, while other businesses may suffer if an area becomes less desirable to customers.

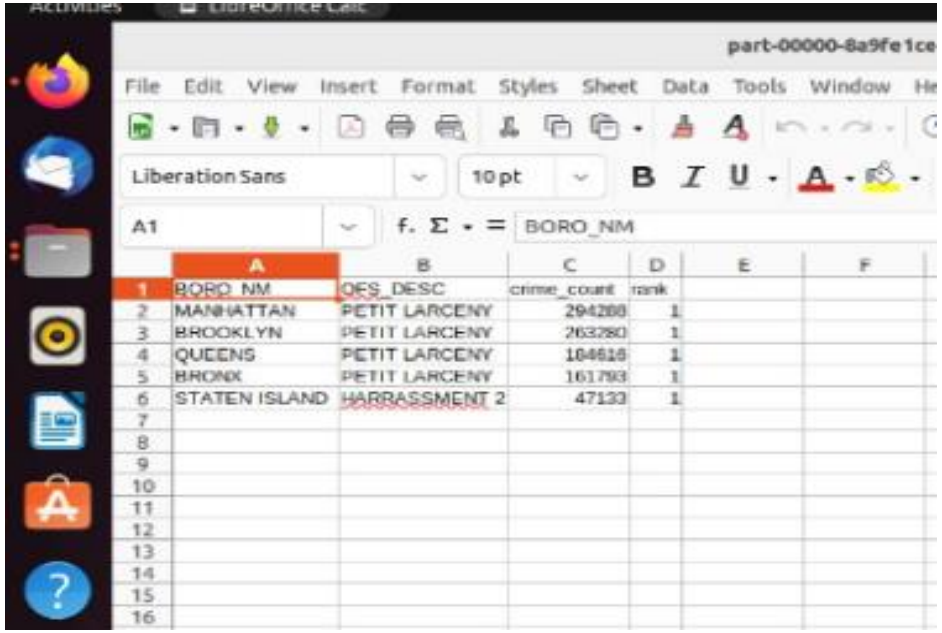


	A	B	C	D	E	F
1	BORO_NM	HOUSE_SALE_COUNT	CRIME_CASES_COUNT	% OF SHARE OF NYC		
2	STATEN ISLAND	330	299115	5.7		
3	QUEENS	2265	1074634	27.25		
4	BRONX	261	1139784	16.92		
5	MANHATTAN	2360	1277512	19.16		
6	BROOKLYN	14454	1581609	30.97		

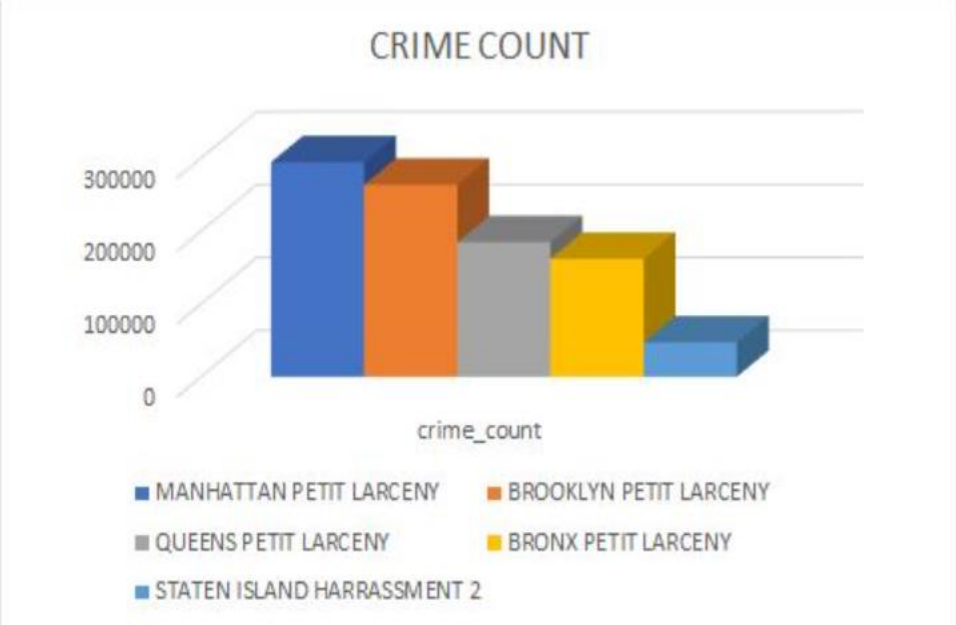


2.What type of crimes are most commonly reported in each borough, and are there any patterns or trends that can be identified?

Businesses operating in areas with high crime rates such as petit larceny and harassment may need to implement safety measures to protect their customers and employees. This could include increasing security personnel or installing surveillance cameras. They may also face financial implications such as higher insurance premiums and loss of revenue. Technology companies specializing in safety and security may see increased demand. Addressing the underlying causes of crime may require a collaborative effort from businesses, law enforcement, and community members.

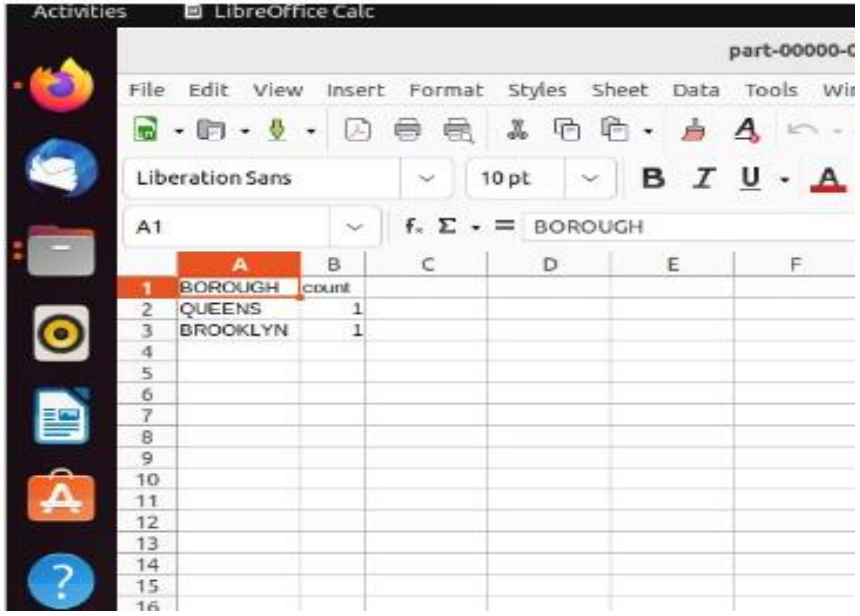


	A	B	C	D
1	BORO_NM	OFS_DESC	crime_count	rank
2	MANHATTAN	PETIT LARCENY	294288	1
3	BROOKLYN	PETIT LARCENY	263280	1
4	QUEENS	PETIT LARCENY	184618	1
5	BRONX	PETIT LARCENY	161793	1
6	STATEN ISLAND	HARRASSMENT 2	47133	1



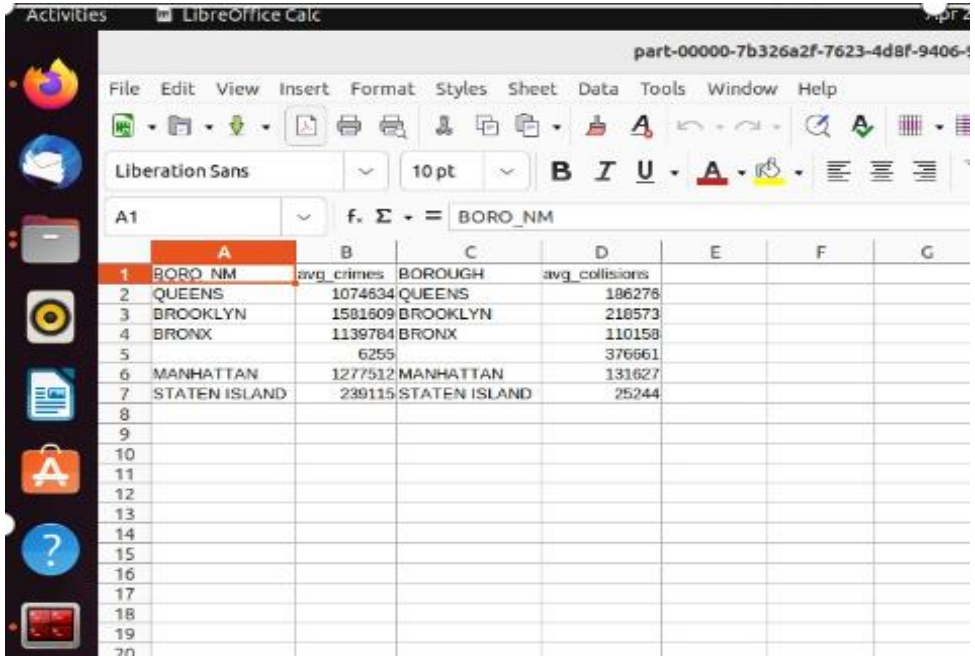
3.Are there any specific neighborhoods in New York City where both the number of crimes reported and the number of motor vehicle accidents where there is either injury or person killed are higher than average?

Businesses operating in high-crime and high-accident areas should prioritize safety to mitigate risks and address safety concerns. This may involve offering additional safety features, investing in safety training and equipment, and partnering with technology companies specializing in safety and security. Revitalization of distressed areas and improvement of safety by real estate companies may provide opportunities, while emergency response services and insurance companies may see increased demand.

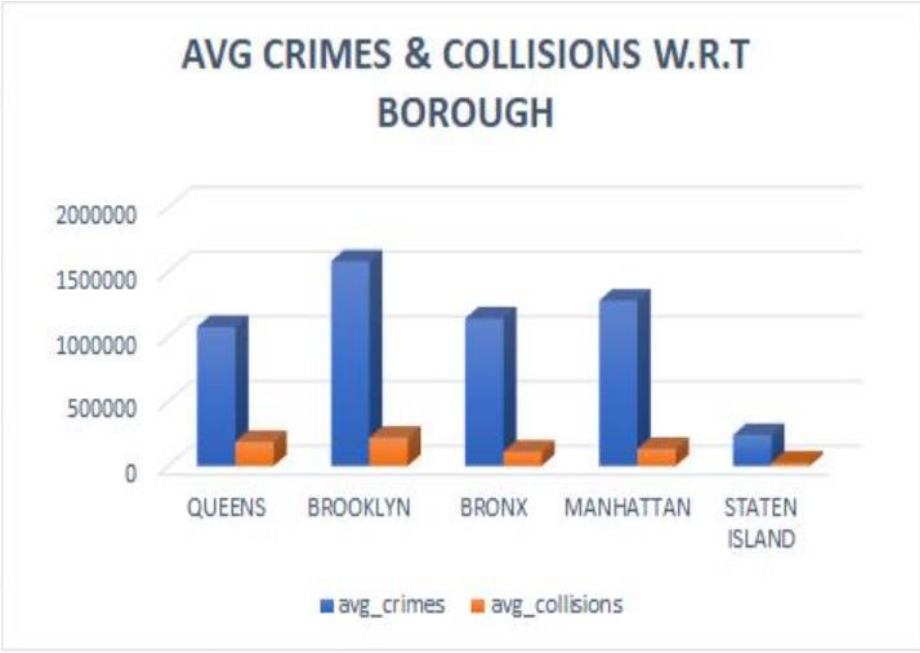


4. Is there any relation between crime and the accidents happening?

Businesses in the automotive industry may experience higher demand for safety equipment in areas with high rates of motor vehicle collisions, while real estate companies may need to invest in security upgrades to attract tenants or buyers in high-crime areas. Transportation companies may need to adjust their operations for safety, and retail stores may need to implement additional security measures. Poor road conditions and inadequate infrastructure in high-crime areas may also affect businesses relying on transportation and logistics.

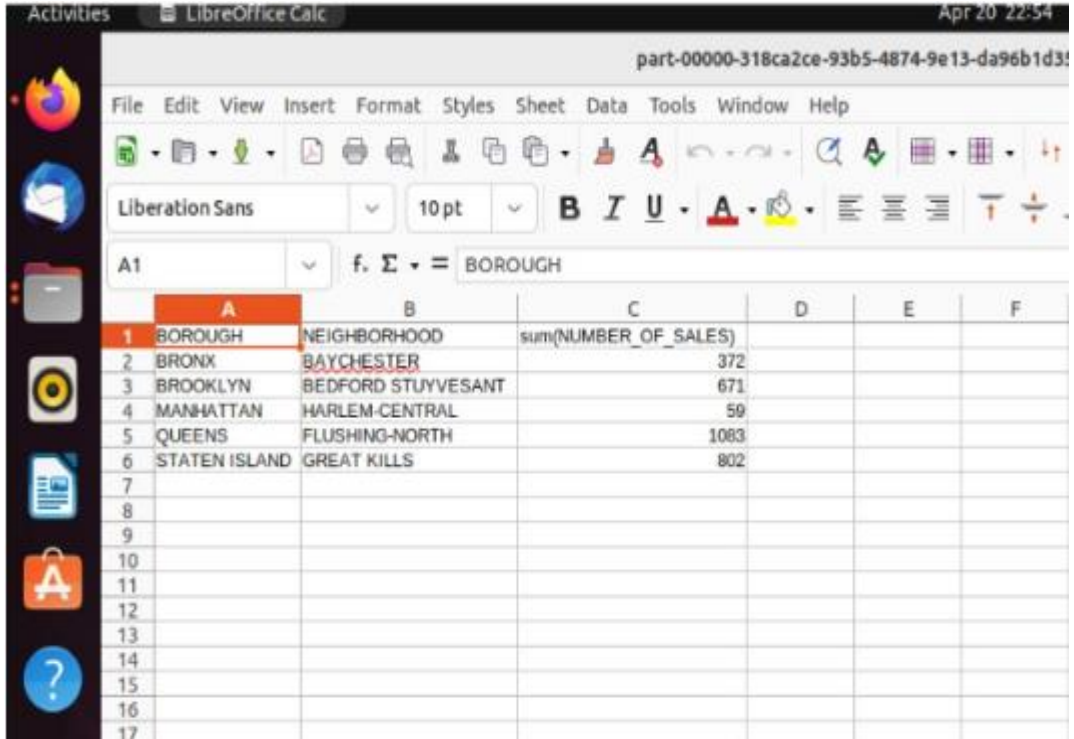


	A	B	C	D
1	BORO_NM	avg_crimes	BOROUGH	avg_collisions
2	QUEENS	1074634	QUEENS	186278
3	BROOKLYN	1581609	BROOKLYN	218573
4	BRONX	1139784	BRONX	110158
5		6255		376661
6	MANHATTAN	1277512	MANHATTAN	131627
7	STATEN ISLAND	239115	STATEN ISLAND	25244

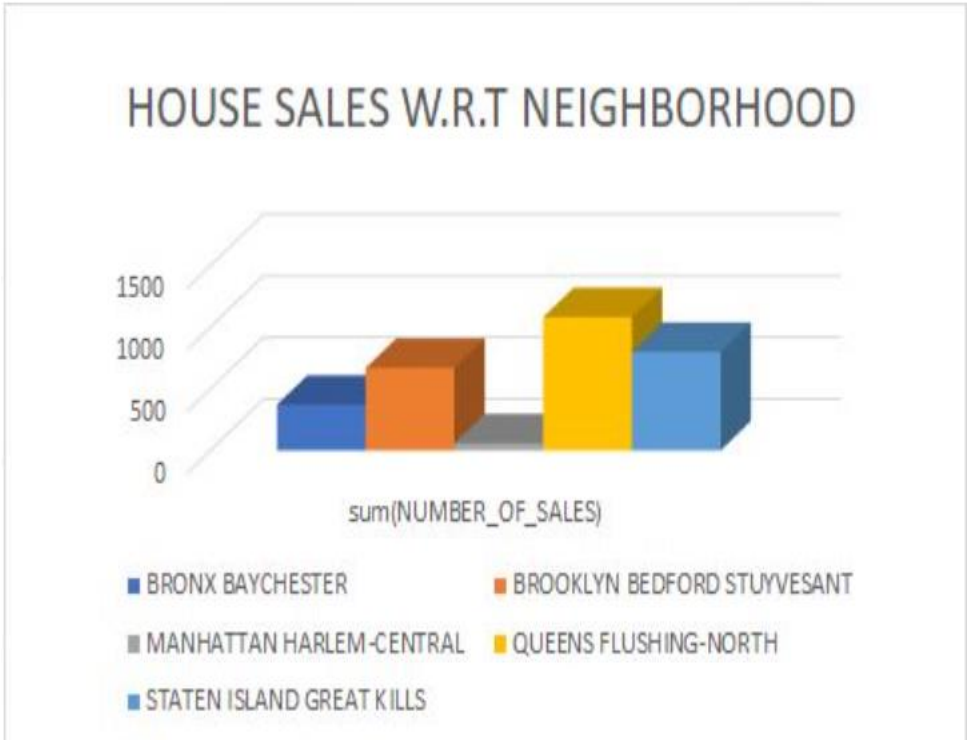


5. Which neighborhood in each borough had the highest number of sales?

Real estate businesses can benefit from identifying the factors that influence the number of property sales in a neighborhood. These include location, affordability, housing quality, market conditions, effective marketing, investment and development, and demographic changes. Neighborhoods located in desirable areas, offering good value for money, and experiencing investment and development efforts or demographic changes are often in high demand, leading to increased property sales. Businesses can develop strategies to target potential buyers based on these factors.



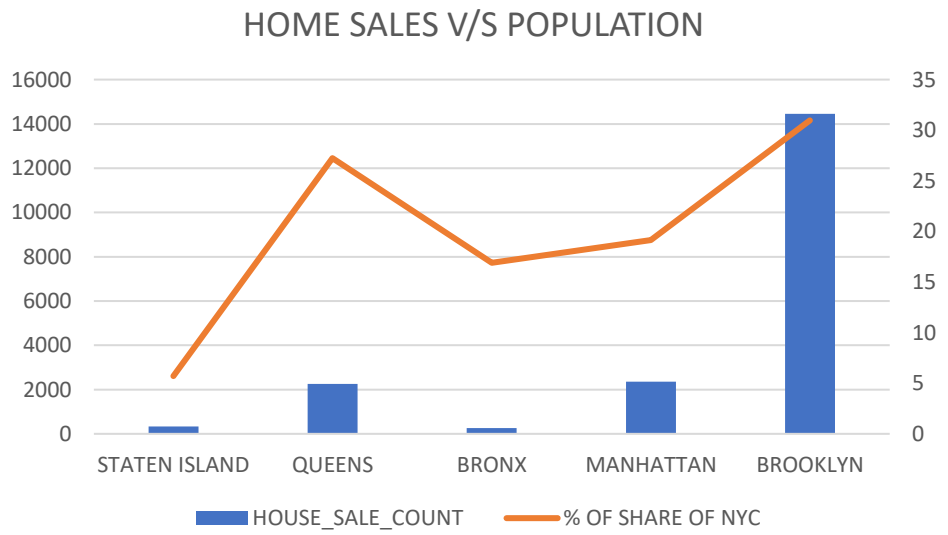
	A	B	C	D	E	F
1	BOROUGH	NEIGHBORHOOD	sum(NUMBER_OF_SALES)			
2	BRONX	BAYCHESTER	372			
3	BROOKLYN	BEDFORD STUYVESANT	671			
4	MANHATTAN	HARLEM-CENTRAL	59			
5	QUEENS	FLUSHING-NORTH	1083			
6	STATEN ISLAND	GREAT KILLS	802			



6.Is there any correlation between the population of a neighborhood and the average sale price of homes in that neighborhood?

The relationship between neighborhood population and home sale prices can impact the decisions of businesses in the real estate and retail industries. Real estate investors and developers can focus on neighborhoods with a growing population and potential for increased property values. Real estate agents can tailor marketing strategies to areas with a specific population size. Retail businesses can identify areas with a growing population to attract customers.

	A	B	C
1	BORO_NM	HOUSE_SALE_COUNT	% OF SHARE OF NYC
2	STATEN ISLAND	330	5.7
3	QUEENS	2265	27.25
4	BRONX	261	16.92
5	MANHATTAN	2360	19.16
6	BROOKLYN	14454	30.97



REFERENCES

- **NEW YORK HOME SALES DATA:** https://data.cityofnewyork.us/City-Government/DOF-Summary-of-Neighborhood-Sales-by-Neighborhood-/5ebm-myj7 (0.00578GB)
- **NYPD COMPLAINT DATA:** https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i (2.45GB)
- **MOTOR VEHICAL COLLISIONS DATA:** https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 (0.21GB)
- **NEW YORK POPULATION BY BOROUGH:** https://data.cityofnewyork.us/City-Government/New-York-City-Population-by-Borough-1950-2040/xywu-7bv9 (2KB)

Total Dataset size = 2.7GB (approx..)

Thank you