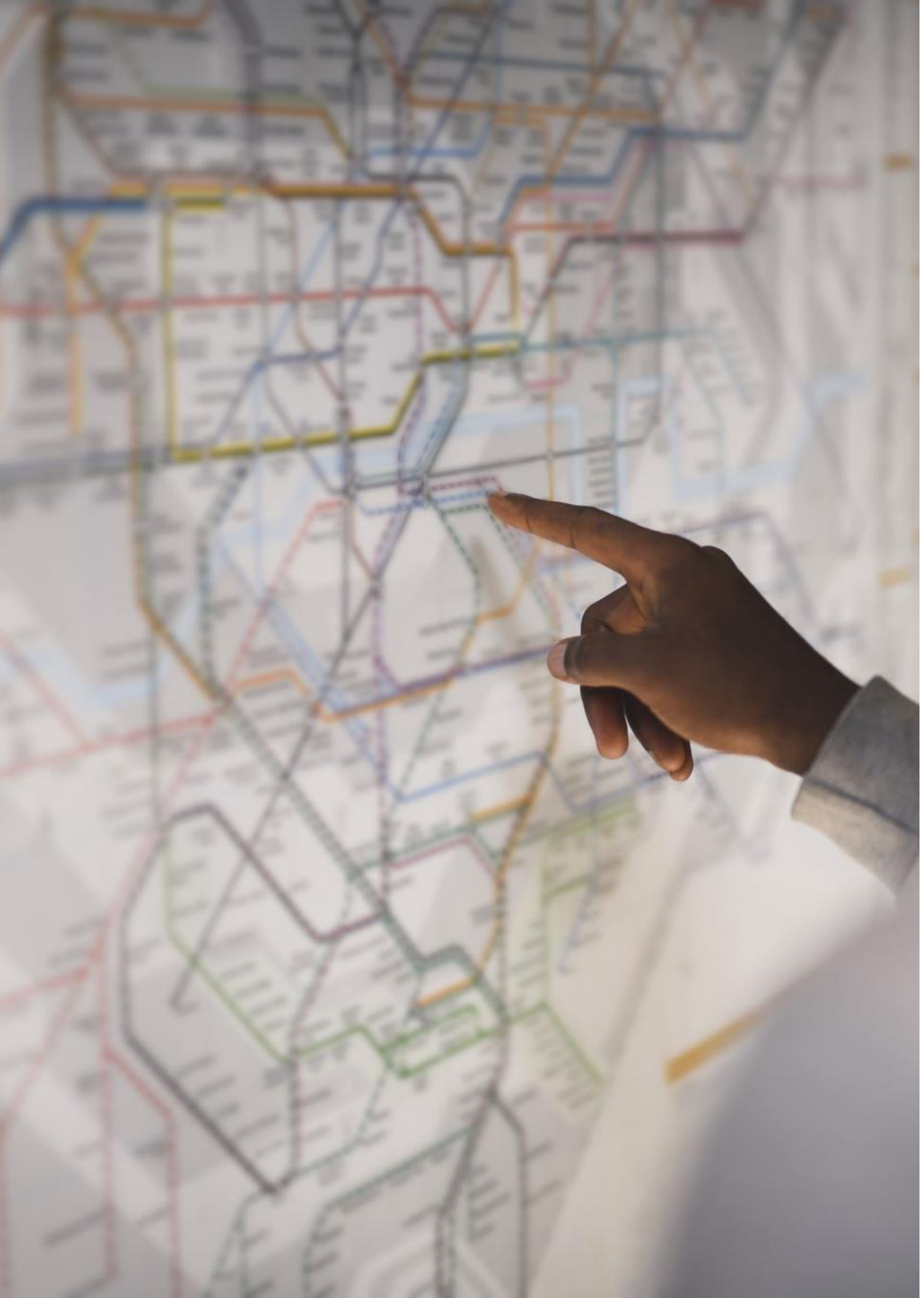
San Francisco Crime Analysis

THE GOAL IS TO UNCOVER CRIME PATTERNS BY TIME, TYPE, AND LOCATION TO SUPPORT DATA-INFORMED PUBLIC SAFETY DECISIONS.

TOOLS USED: PYSPARK, SQL, TABLEAU, PANDAS, MATPLOTLIB



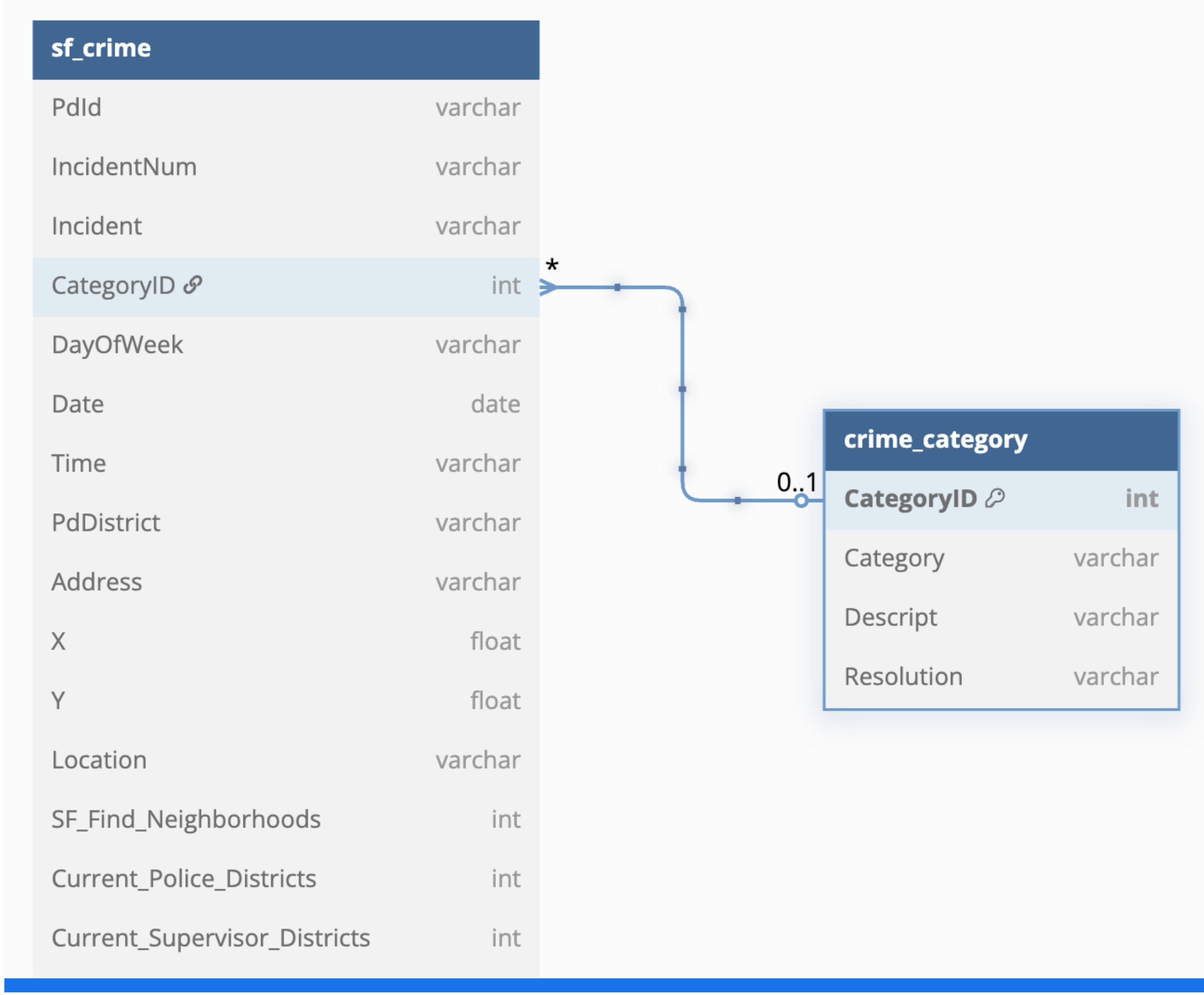
Overview

- This project analyzes San Francisco's public crime data using PySpark and SQL for large-scale processing and Tableau for interactive visualization. The goal is to uncover patterns in crime distribution by time, type, and location to help inform public safety initiatives.
- Tools used: PySpark, Pandas, Tableau, SQL, matplotlib

Data Schema

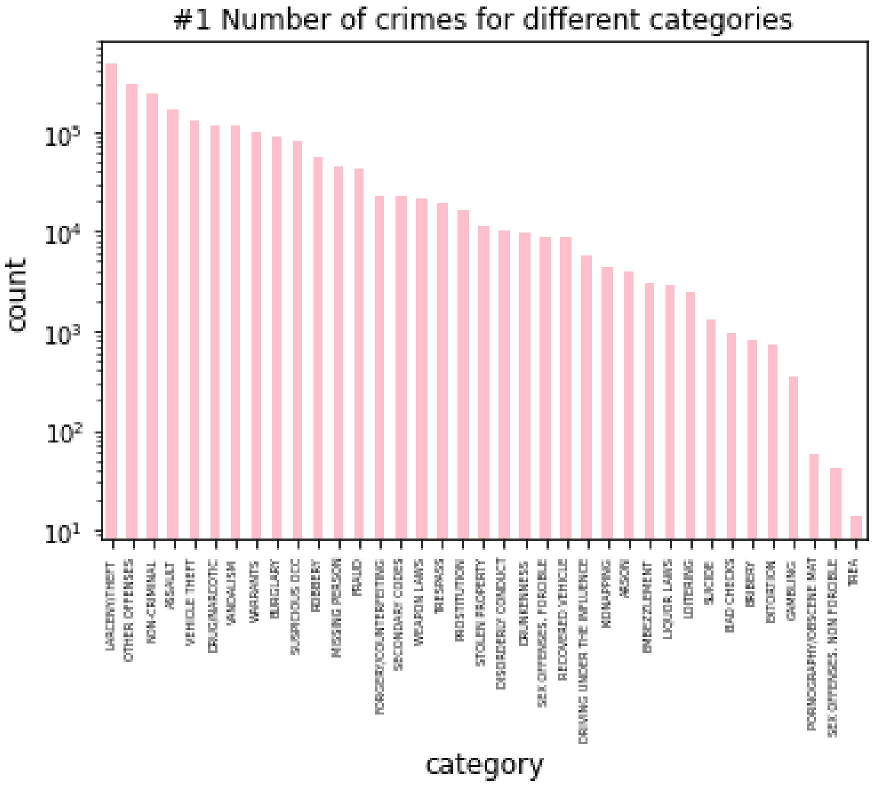
I USED PYSPARK'S SPARKSESSION TO LOAD THE ORIGINAL DATASET, WHICH INCLUDES OVER **5,900 ROWS** AND **33 COLUMNS**. THE DATA WAS LOADED FROM A CSV FILE HOSTED ON DATABRICKS' DBFS.

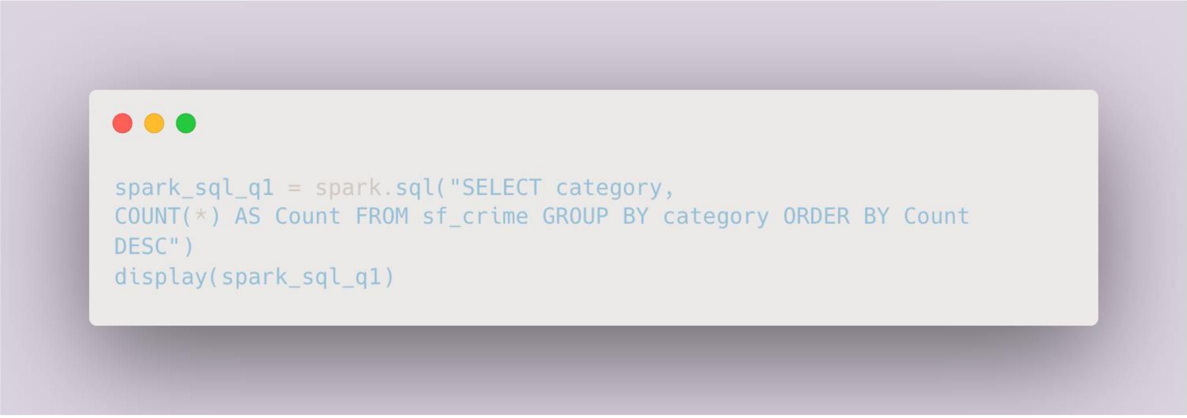
THIS PROJECT USES A CLEANED DATASET OF SAN FRANCISCO CRIME RECORDS. THE DATA WAS STRUCTURED INTO A **RELATIONAL SCHEMA** WITH TWO MAIN TABLES: A **FACT TABLE** SF_CRIME AND A **DIMENSION TABLE** CRIME_CATEGORY.



sf_crime	
PdId	varchar
IncidentNum	varchar
Incident	varchar
CategoryID 🔗	int
DayOfWeek	varchar
Date	date
Time	varchar
PdDistrict	varchar
Address	varchar
X	float
Y	float
Location	varchar
SF_Find_Neighborhoods	int
Current_Police_Districts	int
Current_Supervisor_Districts	int

crime_category	
CategoryID 🔑	**int**
Category	varchar
Descript	varchar
Resolution	varchar





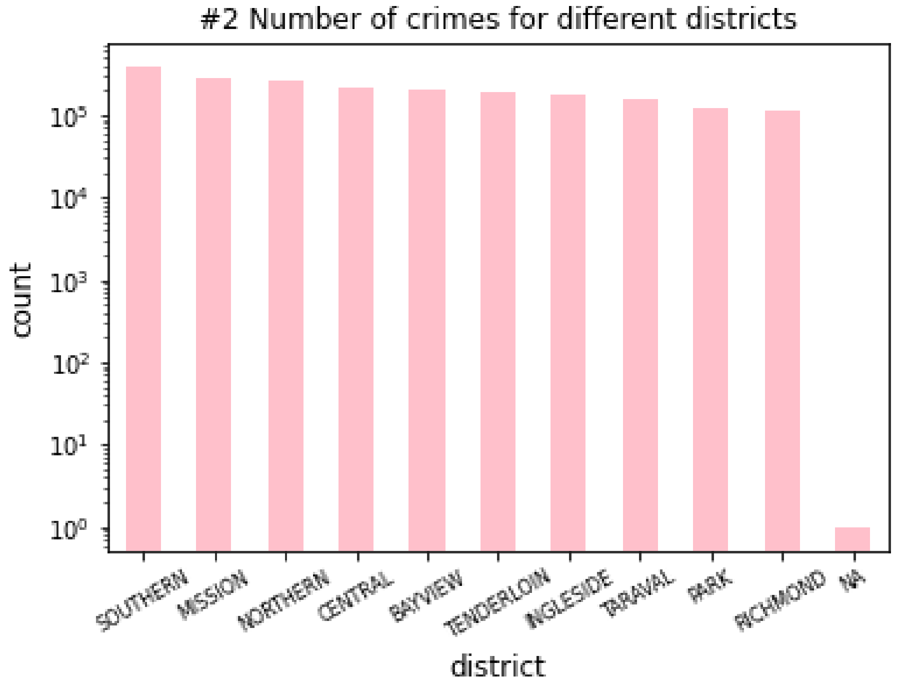
```
spark_sql_q1 = spark.sql("SELECT category,
COUNT(*) AS Count FROM sf_crime GROUP BY category ORDER BY Count
DESC")
display(spark_sql_q1)
```

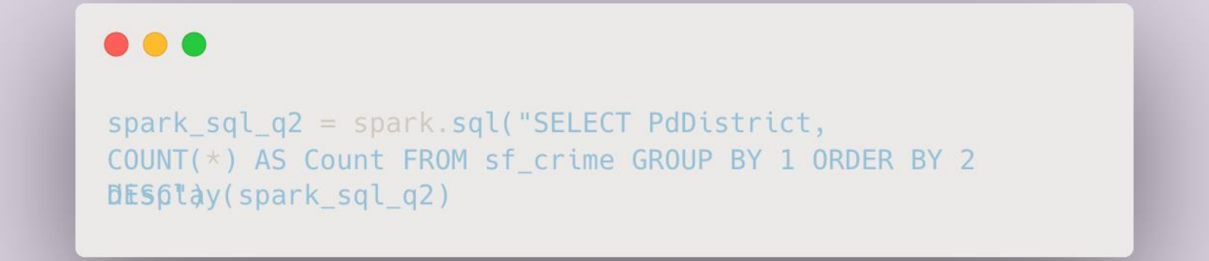
Crime Categories – Frequency Analysis

- This step helps identify which crime types are the most frequent in San Francisco. The results reveal that certain categories such as LARCENY/THEFT, VEHICLE THEFT, and ASSAULT occur more frequently than others, indicating potential areas for increased law enforcement or preventive measures.

Crime Distribution by District

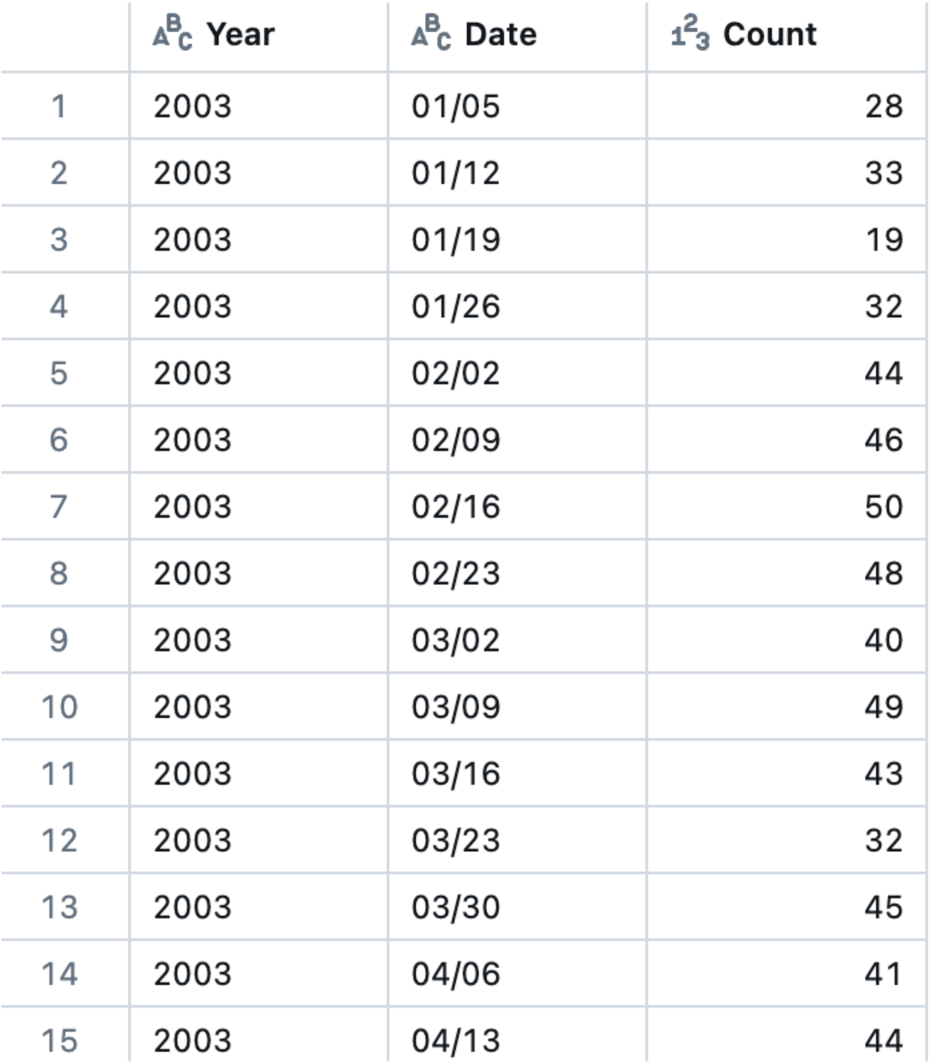
- The chart shows that **Southern**, **Mission**, and **Northern** districts experience the **highest crime volumes**.
  This insight helps target resources to the areas with the greatest need for policing and crime prevention.





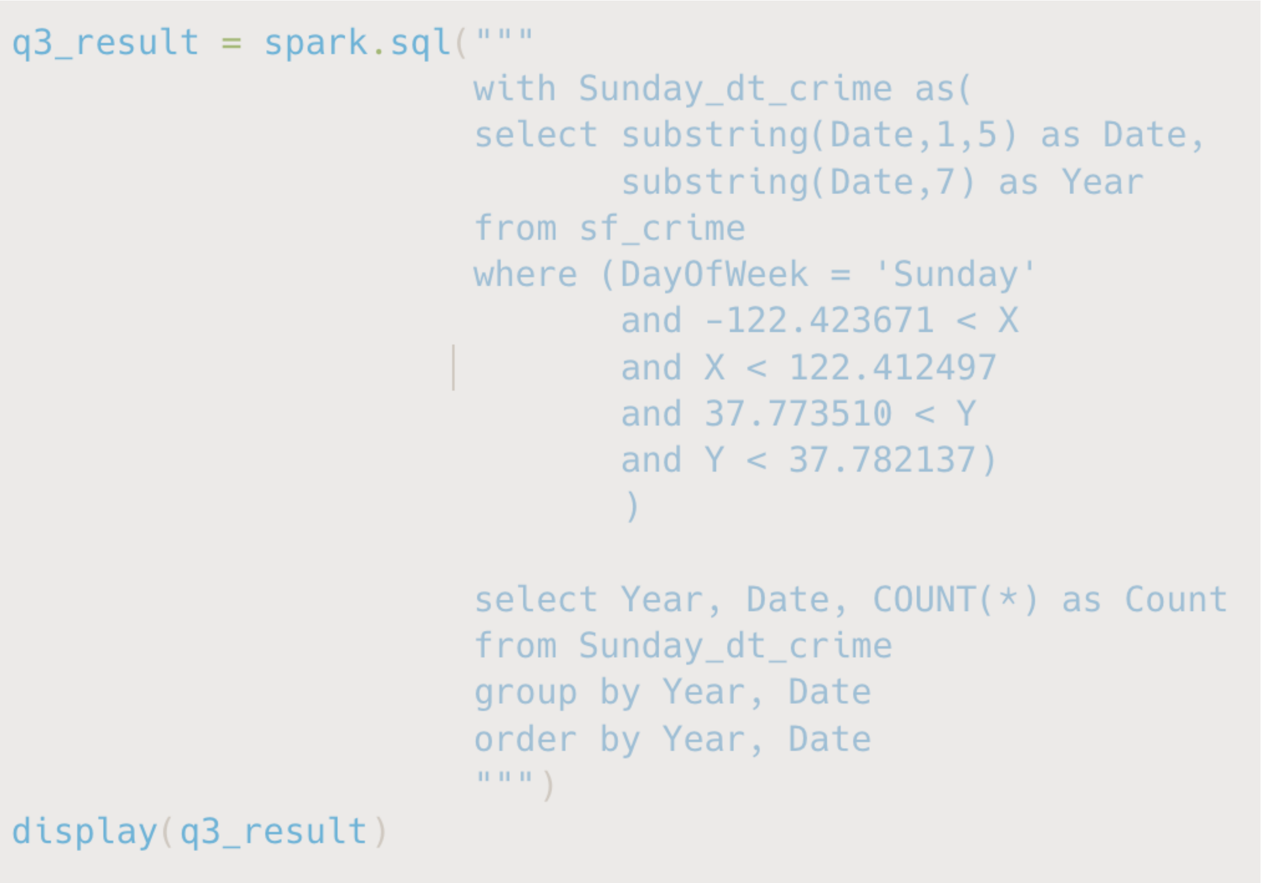
```
spark_sql_q2 = spark.sql("SELECT PdDistrict,
COUNT(*) AS Count FROM sf_crime GROUP BY 1 ORDER BY 2
DESC")
display(spark_sql_q2)
```

Spatiotemporal Analysis: Sunday Crimes in Downtown SF



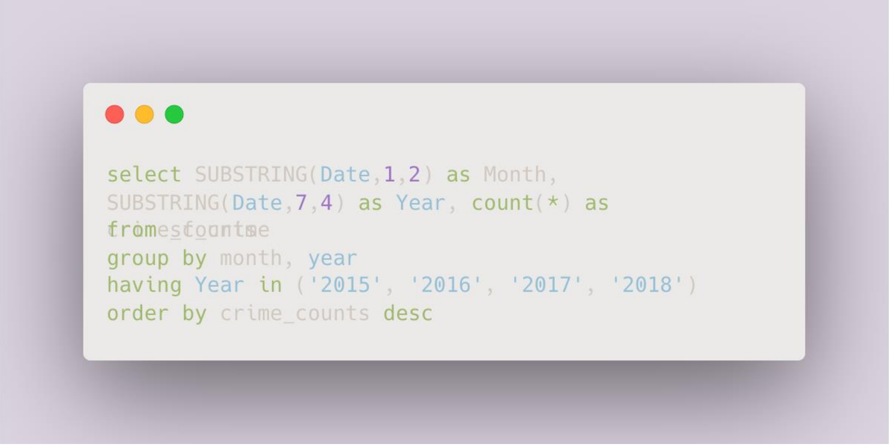
	Year	Date	Count
1	2003	01/05	28
2	2003	01/12	33
3	2003	01/19	19
4	2003	01/26	32
5	2003	02/02	44
6	2003	02/09	46
7	2003	02/16	50
8	2003	02/23	48
9	2003	03/02	40
10	2003	03/09	49
11	2003	03/16	43
12	2003	03/23	32
13	2003	03/30	45
14	2003	04/06	41
15	2003	04/13	44

- This query performs a **spatiotemporal analysis**, focusing on crimes that occurred on **Sundays** within the defined **downtown San Francisco** area.
- I focused on **Sunday incidents** in **downtown San Francisco**, defined by bounding box coordinates, to investigate potential weekend-related crime spikes in high-foot-traffic areas.

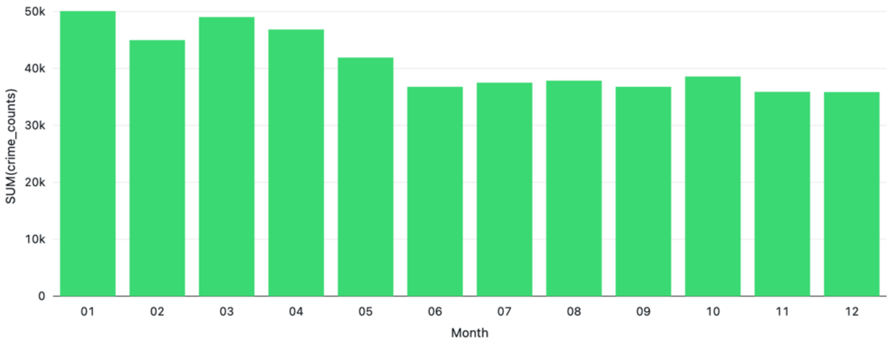


```
q3_result = spark.sql("""
                      with Sunday_dt_crime as(
                      select substring(Date,1,5) as Date,
                             substring(Date,7) as Year
                      from sf_crime
                      where (DayOfWeek = 'Sunday'
                             and -122.423671 < X
                             and X < 122.412497
                             and 37.773510 < Y
                             and Y < 37.782137)
                             )

                      select Year, Date, COUNT(*) as Count
                      from Sunday_dt_crime
                      group by Year, Date
                      order by Year, Date
                      """)
display(q3_result)
```



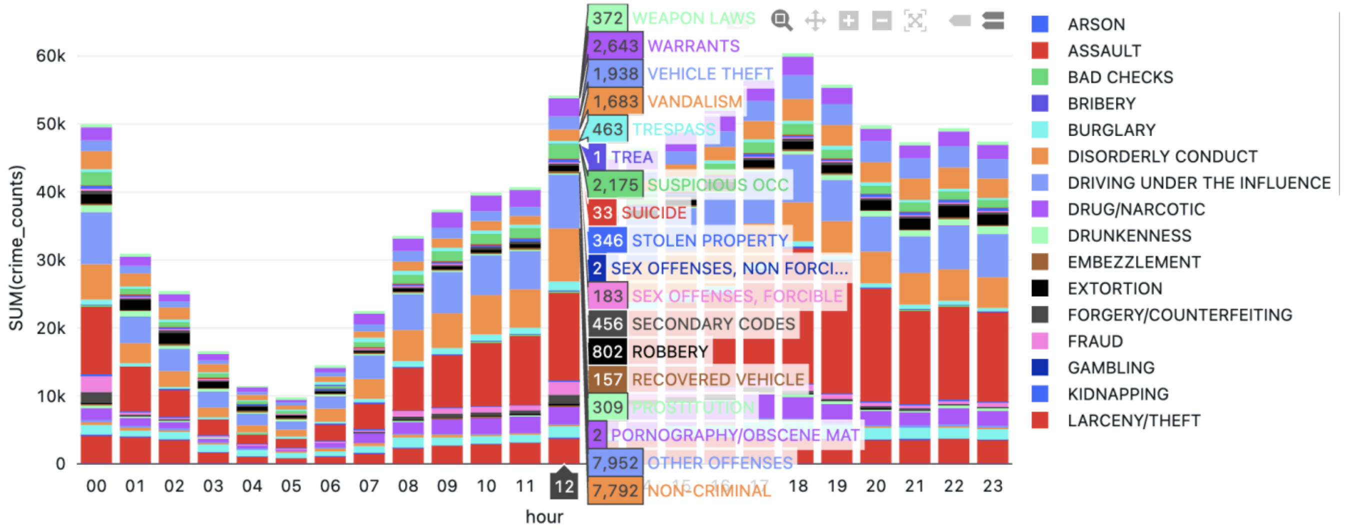
```
select SUBSTRING(Date,1,2) as Month,
SUBSTRING(Date,7,4) as Year, count(*) as
fromes_countse
group by month, year
having Year in ('2015', '2016', '2017', '2018')
order by crime_counts desc
```

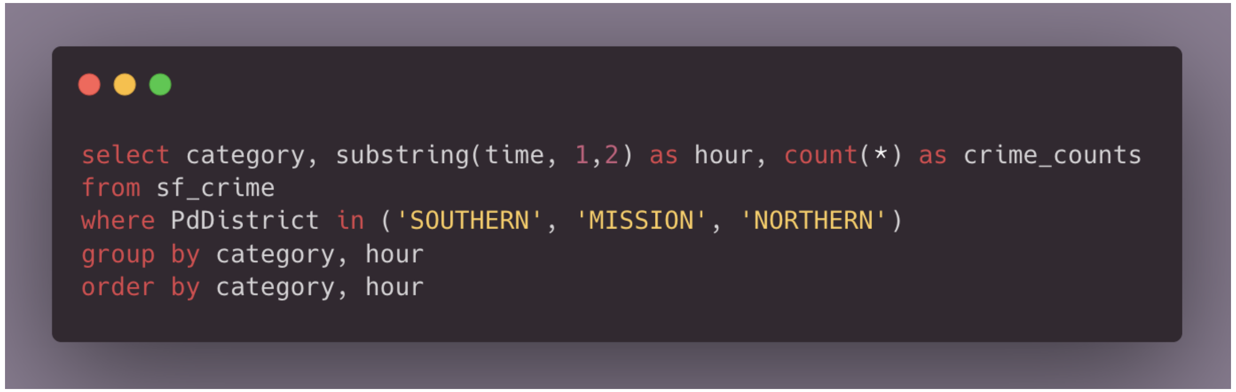


Monthly Crime Trends (2015–2018)

- The data shows a consistent rise in crime in **January and March** across multiple years, possibly linked to post-holiday activity or seasonal social patterns.
  This trend could support **seasonal policing strategies** or further correlation with public events.

Top Crime Districts & Strategic Police Allocation





```
select category, substring(time, 1,2) as hour, count(*) as crime_counts
from sf_crime
where PdDistrict in ('SOUTHERN', 'MISSION', 'NORTHERN')
group by category, hour
order by category, hour
```

- The stacked bar chart displays the **distribution of crime categories across each hour of the day** within the three most dangerous police districts: **SOUTHERN**, **MISSION**, and **NORTHERN**.
- **Midday to early evening (12 PM to 6 PM)** is the **highest-risk window** for criminal activity in SF's most dangerous districts.
- **Theft-related crimes** (Larceny, Vehicle Theft) are the **most prevalent** during peak hours.
- **Police patrols and resources** should be **concentrated in these time slots**, particularly in districts like Southern and Mission.
- **Early morning hours** present an **opportunity for resource reallocation**, as crime rates are minimal.

Key Takeaways & Insights

1. Crime Hotspots Identified

   - The top three most dangerous districts are **Southern**, **Mission**, and **Northern**, each reporting significantly higher incident counts. These findings highlight the importance of localized policing strategies and targeted resource deployment.

2. Peak Crime Times

   - Crime frequency is **lowest between 3 AM and 6 AM** and **peaks from 12 PM to 6 PM**. Theft-related crimes such as **Larceny/Theft**, **Vehicle Theft**, and **Robbery** dominate during peak hours, suggesting a higher need for daytime patrolling.

3. Spatiotemporal Risk Zones

   - Sunday crimes in **downtown SF** (as defined by spatial coordinates) show consistent patterns by year and date. This demonstrates how spatial filtering combined with temporal segmentation can inform targeted prevention strategies.

4. Actionable Recommendations

   - **Increase police patrols** in the top 3 districts during **12 PM–6 PM**, especially focusing on theft-related crimes.
   - **Reduce night shift deployments** in low-crime early morning hours (e.g., 4–6 AM) to optimize resources.
   - **Use data-driven insights** to implement predictive patrol schedules based on location and hour-specific risk.

Thank you

-MINFEI HE