BUDT703: Database Management Systems

Professor: Dr. Woei-jyh (Adam) Lee

Last Updated Date: December 6th, 2025

“WMATA(Washington Metropolitan Area Transit Authority) Crime Analysis”

Final Project Report

By

KARY Metro Guards
Youngseo Chang
Rohith Moravaneni
Kurumi Sato
Ajay Vishwanatha

Table of Contents

Ⅰ. Introduction

A. Mission Statement

The mission of this project is to analyze the Washington Metropolitan Area Transit Authority (WMATA) performance records related to crime incidents, customer and employee injuries, and stations. By organizing these records within a structured database and applying analysis, we seek to derive recurring safety trends and patterns across the WMATA systems. The ultimate goal of this project is to generate clear, data-driven insights that can be utilized to improve overall passenger and employee safety throughout the WMATA transit network.

B. Mission Objectives

To achieve this mission, we focus on four specific objectives. First, we analyze how the number of crime incidents varies by time of day and between peak and off-peak hour periods to understand how the level of risk varies across different times of day. This comparison is intended to show whether particular time windows require prioritized attention from WMATA and the Metro Transit Police.

Second, we identify how total crimes and injuries change over time 2020 to 2022. Tracking these trends helps identify how the safety conditions have been changing over the years, including whether incidents are increasing or decreasing, which can be utilized by the WMATA when building future safety initiatives or evaluating their effectiveness.

Third, we determine which stations and rail lines experience the highest frequency of daily incidents. Determining these high-risk locations provides a basis of prioritizing limited safety resources towards the stations within the network with the greatest need, increasing the resource utilization of the WMATA.

Finally, we identify which offense types occur most frequently within the transit system, and how their frequency changes over time. We believe understanding the relative frequency of different offenses will help to highlight the types of incidents that impose burden on the riders, employees and WMATA system operations.

II. Data Processing

Steps:

We began by extracting all relevant sheets related to customer and employee injury from the WMATA Public Records Performance Records Excel file (WMATA, n.d.). We added a primary key to each record of these sheets so that they could be uniquely identified. All calculated or derived attributes were removed.

Next, we used the WMATA station dataset from the DC Open Data portal (Open Data DC, 2012). Only the station name and address fields were retained. A new station ID was created to act as the primary key. The station addresses were geocoded to obtain latitude and longitude coordinates. This was important to match each daily incident to a station. Since each station can be served by multiple Metro lines, a separate Station Line table was created to store these one-to-many relationships.

Python Script to geocode a station and daily incident location to get coordinates:

```
import numpy as np
import pandas as pd
from geopy.geocoders import GoogleV3
import time

# Load data
stations = pd.read_excel(“…”)
unique_locations = pd.read_excel(“…”)   # These are the unique 
locations of the Daily Incident Table

# Initialize geocoder (API KEY)
geolocator = GoogleV3(api_key=”YOUR_API_KEY”)

# Add columns for latitude and longitude
unique_locations["Latitude"] = None
unique_locations["Longitude"] = None

# Geocode each unique address
for i, address in enumerate(unique_locations["location"]):
      try:
      location = geolocator.geocode(address)
      if location:
            unique_locations.at[i, "Latitude"] = location.latitude
           unique_locations.at[i, "Longitude"] = location.longitude
      else:
            print(f"Not found: {address}")
      except Exception as e:
        print(f"Error for {address}: {e}")
    time.sleep(0.1)
```

```
# Save geocoded addresses
unique_locations.to_excel("Incident_unique_geocoded.xlsx",
index=False)
```

The Daily Incident table was created using the monthly blotter sheets from WMATA’s crime statistics (WMATA, n.d.). Python scripts were used to download all available monthly PDFs, convert each PDF into CSV format, and to concatenate the individual CSVs into one combined CSV.

Python Script to download the monthly blotter sheets:

```
import os
import re
import requests
import time
from bs4 import BeautifulSoup
from urllib.parse import urljoin

URL = "https://www.wmata.com/about/transit-police/crime-stats.cfm"
BASE_DIR = "pdfs"
os.makedirs(BASE_DIR, exist_ok=True)

# Match month-day-year ranges like "October 1 - 31, 2020"
month_pattern = re.compile(

r"(January|February|March|April|May|June|July|August|September|Oct
ober|November|December)"
    r"\s+\d+\s*[-–—]\s*\d+,\s*(20\d{2})",
     re.I
)

# Extract actual PDF path from javascript-based links
pdf_js_pattern =
re.compile(r",(/about/transit-police/upload/[^)']+\.pdf)")

# Fetch main page
response = requests.get(URL)
response.raise_for_status()
soup = BeautifulSoup(response.text, "html.parser")
```

```
count = 0

# Loop through all links on the page
for a in soup.find_all("a", href=True):
     text = " ".join(a.get_text(strip=True).split())
     match = month_pattern.search(text)
     if not match:
     continue

     month_name, year = match.group(1), match.group(2)

     # Make year folder inside "pdfs"
     year_folder = os.path.join(BASE_DIR, year)
    os.makedirs(year_folder, exist_ok=True)

     href = a["href"]

     # Convert javascript PDF link into a usable URL
     if href.startswith("javascript:HandleLink"):
        pdf_match = pdf_js_pattern.search(href)
     if pdf_match:
           href = urljoin("https://www.wmata.com",
pdf_match.group(1))
     else:
            print(f"Could not parse PDF URL for: {text}")
            continue
     elif not href.startswith("http"):
     href = urljoin(URL, href)

     # Build a consistent filename: YYYY_MM_Month.pdf
     month_num = time.strptime(month_name, "%B").tm_mon
     filename = f"{year}_{month_num:02d}_{month_name}.pdf"
     filepath = os.path.join(year_folder, filename)

     # Skip if already downloaded
     if os.path.exists(filepath):
        print(f"Skipping (exists): {filename}")
     continue

     # Download the PDF
    print(f"Downloading: {text}")
     try:
     pdf_data = requests.get(href)
        pdf_data.raise_for_status()
     with open(filepath, "wb") as f:
            f.write(pdf_data.content)
     count += 1
     except Exception as e:
        print(f"Failed: {href} ({e})")

print(f"\nDownloaded {count} PDFs into '{BASE_DIR}/' by year.")
```

Python Script to the PDFs convert to CSVs:

```
import os
import tabula

 # folder that contains all year folders
 base_folder = "pdfs"

 # loop through all year folders
 for year_folder in os.listdir(base_folder):
     year_path = os.path.join(base_folder, year_folder)
     if not os.path.isdir(year_path):
     continue

     # make a csv folder for each year
     csv_folder = os.path.join(year_path, "csv")
     os.makedirs(csv_folder, exist_ok=True)

     for file in os.listdir(year_path):
     if file.endswith(".pdf"):
           pdf_path = os.path.join(year_path, file)
           csv_path = os.path.join(csv_folder, file.replace(".pdf",
".csv"))

           try:
                # extract tables and save as CSV
                 tabula.convert_into(pdf_path, csv_path,
output_format="csv", pages="all")
                print(f"Converted: {file}")
           except Exception as e:
                print(f"Failed: {file} — {e}")
```

Python Script to concatenate all the CSVs:

```
import os

# Base folder containing the year subfolders with PDFs and CSVs
base_folder = r"PATH_TO_PDFS_FOLDER"
output_file = r"PATH_TO_OUTPUT\combined_blotter_main.csv"

with open(output_file, "w", encoding="utf-8") as outfile:
     # Loop through each year folder
     for year_folder in os.listdir(base_folder):
        year_path = os.path.join(base_folder, year_folder)
        csv_folder = os.path.join(year_path, "csv")  # monthly
CSVs stored here

     if not os.path.exists(csv_folder):
            continue

     # Append each CSV into one combined file
     for file in os.listdir(csv_folder):
           if file.endswith(".csv"):
                file_path = os.path.join(csv_folder, file)

                with open(file_path, "r", encoding="utf-8") as
infile:
                    outfile.write(infile.read())
                    outfile.write("\n")  # space between merged
files
```

```
print(f"All CSVs have been combined into: {output_file}")
```

Because PDF-to-CSV is often inconsistent, the resulting data required substantial cleaning. All unique offenses were identified from the Daily Incidents, and an Offense table was created to store them. An offense ID was created to uniquely identify each offense type. Later on, we altered the Offense table to add a general offense category to each offense using case statements.

Some daily incident records contained multiple offenses, so a separate Involve table was created to represent the one-to-many relationship between an incident and its offenses. This table stores both the incident ID and the corresponding offense ID for each offense linked to that incident.

Several address fields were incomplete or inconsistently formatted which required manual corrections and standardization. After cleaning, all incident addresses were geocoded to obtain latitude and longitude.

After geocoding all incident addresses, each incident needed to be matched to its nearest WMATA station. We used Python to spatially match each incident to its nearest station. With roughly 11,000 recorded incidents and 98 stations, a brute-force comparison of every pair (over one million checks) would have been very inefficient. Instead, a KDTree was constructed using the latitude–longitude coordinates of all stations for fast nearest-neighbor lookups. Each incident's coordinates were then passed through the KDTree to identify the nearest station and this station was then assigned to the incident record.

Python Script to build a KDTree to match daily incident to its nearest station:

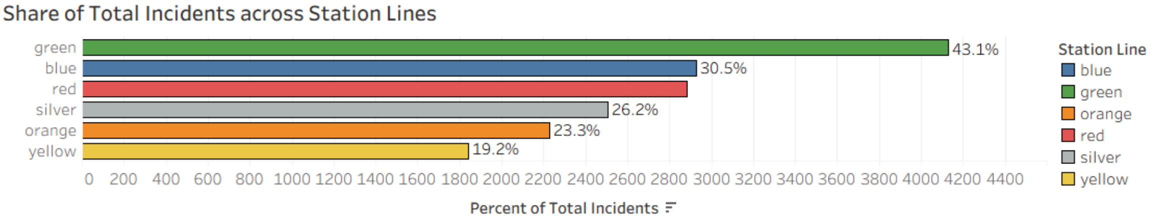
```
import pandas as pd
from scipy.spatial import cKDTree

# Load prepared datasets (
incident_df = pd.read_excel("Daily_Incident.xlsx")
station_df  = pd.read_excel("Stations_with_coords.xlsx")

# Ensure latitude/longitude are numeric
incident_df["Latitude"]  = pd.to_numeric(incident_df["Latitude"],
errors="coerce")
incident_df["Longitude"] = pd.to_numeric(incident_df["Longitude"],
errors="coerce")
```

```
station_df["Latitude"]   = pd.to_numeric(station_df["Latitude"],
errors="coerce")
station_df["Longitude"]  = pd.to_numeric(station_df["Longitude"],
errors="coerce")

# Remove rows with missing coordinates
incident_df = incident_df.dropna(subset=["Latitude",
"Longitude"]).reset_index(drop=True)
station_df  = station_df.dropna(subset=["Latitude",
"Longitude"]).reset_index(drop=True)

print(f"Incidents with valid coordinates: {len(incident_df)}")
print(f"Stations with valid coordinates: {len(station_df)}")

# Convert coordinates to arrays
incident_coords = incident_df[["Latitude",
"Longitude"]].to_numpy()
station_coords  = station_df[["Latitude", "Longitude"]].to_numpy()

# Build KDTree using station coordinates
tree = cKDTree(station_coords)

# Find nearest station for each incident
distances, indices = tree.query(incident_coords, k=1)

incident_df["NearestStationID"] =
station_df.iloc[indices]["StationID"].values

# Convert degree distance to approximate kilometers
incident_df["Distance_km"] = distances * 111

# Save the result
incident_df.to_excel("incident_with_station.xlsx", index=False)
print("Saved nearest-station results to
'incident_with_station.xlsx'")
```

This script adds the nearest station ID to each daily incident with distance in kilometers. Incidents with a distance of greater than 0.5 km were later filtered out in a created view ‘DailyIncidentNearStation’ which was used for analysis.

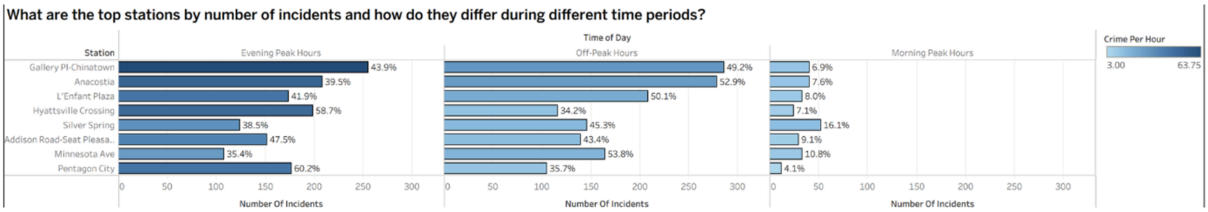
ＩＩI. Business Transactions

1. What is the share of total incidents across all station lines?



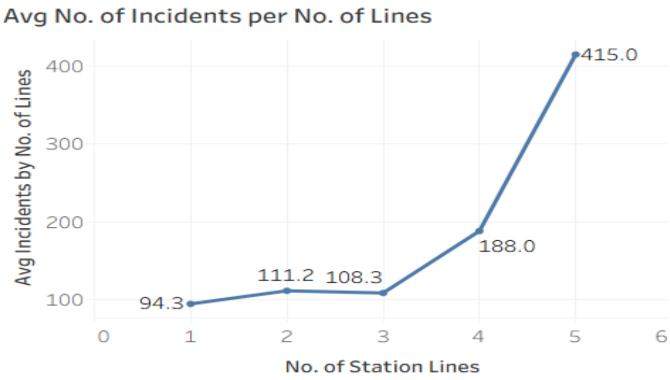
Based on the bar chart, it is shown that the incidents are not spread evenly across the rail lines. The Green line shows the highest number of incidents, with about 43% of the reported incidents happening on the stations of the Green line. The Blue and Red lines follow second, accounting for about 30% of total reported incidents each, while Silver, Orange and Yellow account for smaller shares. This shows that the WMATA’s safety resources should be focused on the lines that account for higher shares of reported incidents, such as the Green line.

2. What are the top stations by number of incidents and how do they differ during different time periods?



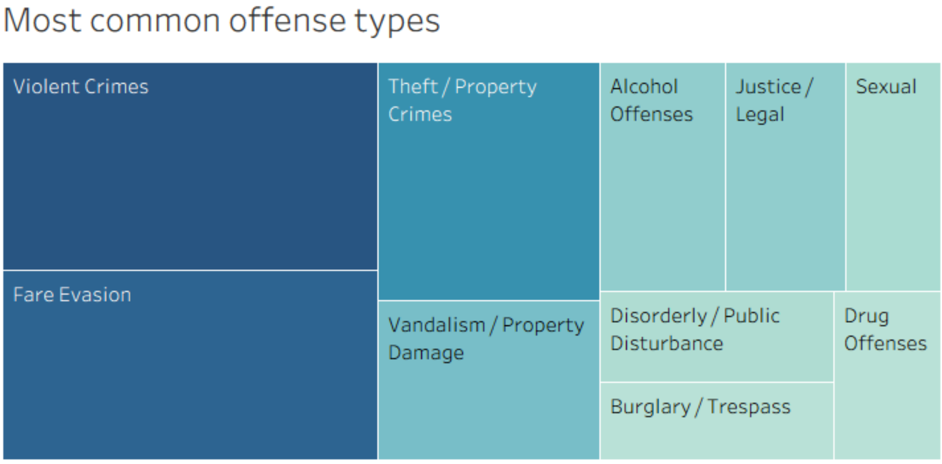
The top stations by the number of incidents are Gallery Place–Chinatown, Anacostia, L’Enfant Plaza, Hyattsville Crossing, Silver Spring, Addison Road–Seat Pleasant, Minnesota Ave, and Pentagon City. Across these stations, most incidents occur during the evening peak and off-peak hours, while morning peak hours accounts for much smaller shares. Stations such as Pentagon City and Hyattsville Crossing are skewed towards evening peak incidents (accounting for 60%), whereas stations such as Anacostia and Minnesota Ave see more incidents during off-peak hours. This analysis identifies specific stations that show a high number of incidents, and shows that evening peaks and off-peak hours should be prioritized for safety. We can also see that in general, there is a higher crime rate for evening-hours compared to off-peak hours even though they share a similar number of crime incidents. This is because the duration of evening-peak hours is much shorter, therefore on a per-hour basis, evening peak-hours are considered more dangerous.

3. Do stations with more lines tend to have more incidents?



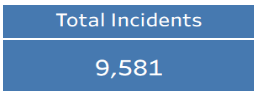
Through this line graph, we can noticeably see that stations with more lines on average have more incidents on average. Single-line stations see about 94 incidents on average, while two-lines and three-lines stations show a slightly-elevated number of incidents at around 108-111 incidents. However, from stations with four lines and five lines, there is a dramatic increase, where stations with four lines see 188 incidents, and stations with five lines see 415. This analysis indicates that multi-line transfer stations account for a higher share of incidents compared to single and double line stations, and should be prioritized in upkeep their safety.

4. What are the Top 10 most frequently occurring offence categories?



Through this chart, we are able to identify the most frequently occurring offence categories. Violent Crimes is the largest category with 2,137 incidents, followed closely by fare evasions with 1,937 incidents. Together with Theft/Property Crimes following third, these three categories account for a large share of all reported crimes. Vandalism / Property Damage, Justice / Legal Obstruction and Alcohol Offenses, make up the second tier in their frequency, each with 700-100 incidents. Lastly, less frequent but still significant, are Sexual Offenses, Disorder / Public Disturbance, Drug Offense, and Burglary / Trespass. Overall, this analysis shows that Violent Crimes, which can pose one of the most significant threats to the safety of riders and employees, should be prioritized when considering safety precautions.

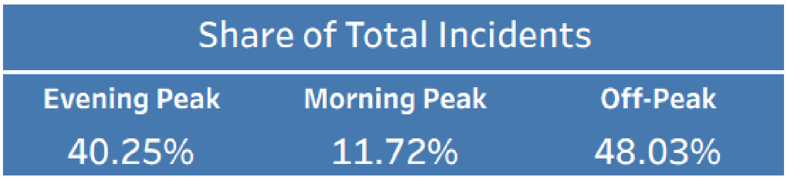
5. What is the total number of incidents?



Total Incidents
9,581

Through this analysis, we can calculate that from 2020 to 2022, the number of crime incidents reported in the WMATA system are 9,581.

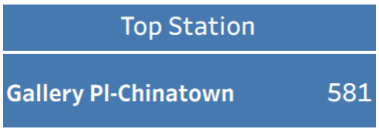
6. What is the share of total incidents by time of the day?



Share of Total Incidents		
Evening Peak	Morning Peak	Off-Peak
40.25%	11.72%	48.03%

Further developing Business Transaction 2, instead of looking only at high-risk stations, we examine how incidents are distributed by time of day across the entire WMATA system. Similar to high-risk stations, only about 11.7% of incidents occur during morning peak hours, while evening peak accounts for roughly 40.3% and off-peak hours account for 48.0% of all incidents. This analysis shows that system-wide, most safety risks arise in the evening and off-peak times. Such findings can be utilized when allocating safety resources across different times of day to ensure efficiency.

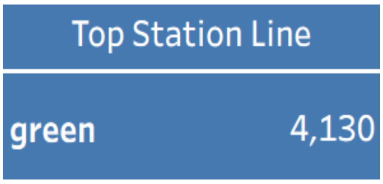
7. What is the station with the most number of incidents?



Top Station	
Gallery Pl-Chinatown	581

Through this analysis, we identify Gallery Place–Chinatown having the most number of reported incidents across the WMATA system, with 581 reported incidents from 2020 to 2022. This finding indicates Gallery Place–Chinatown station's need for safety resources.

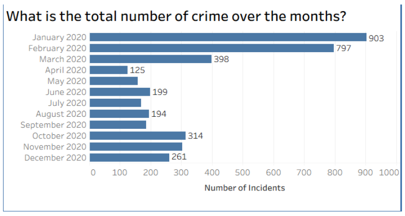
8. What station line has the most number of incidents?



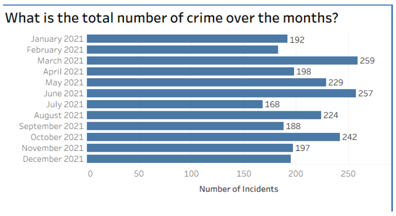
Top Station Line	
green	4,130

Through this analysis, we identify the Green Line having the most number of reported incidents across the WMATA system, with 4,130 reported incidents from 2020 to 2022. This finding reinforces the need to prioritize safety resources on the Green line, compared to other lines in the system.

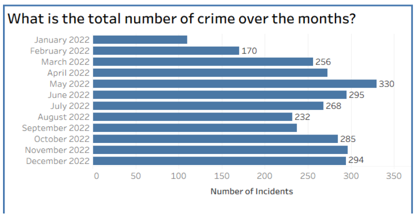
9. What is the total number of crimes over the months?



What is the total number of crime over the months?



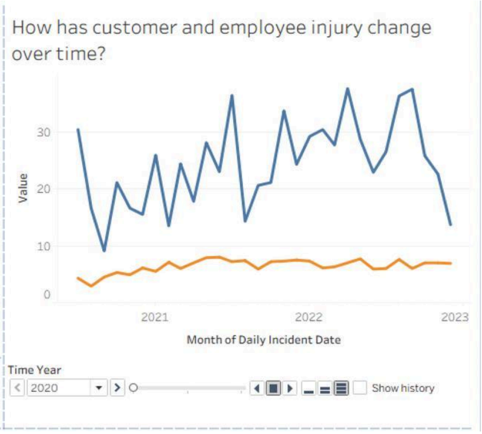
What is the total number of crime over the months?



What is the total number of crime over the months?

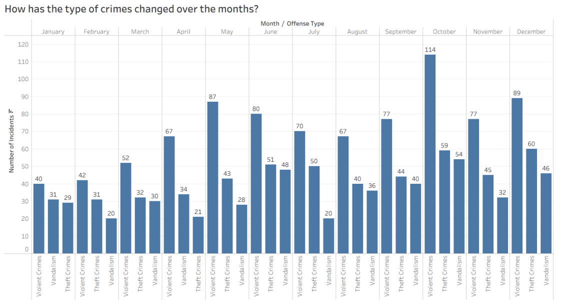
Through this analysis, we can see that the total number of crimes in 2020 dropped significantly towards the beginning of the year. This result can be explained by the fact that the first case of COVID-19 in the DMV area was on March 7th. From around March until the end of 2022 the number of crimes stayed relatively low compared to the larger amount of crimes in early 2020. There is an upward trend of crimes towards the start of 2022, when the number or riders increased with the relaxation of COVID-19 and social distancing regulations.

10. How has the customer and employee injury changed over time?



Through the line charts, we can see that the number of customer injuries is always higher than the number of employee injuries. Such results can be attributed to the higher number of riders compared to employees. The number of customer injuries show higher volatility compared to that of employees, and trends upward from the late 2020 into 2022. In contrast, the employee injury numbers remain lower and relatively stable over the years. This suggests that while both groups face safety challenges, the overall safety conditions are driven by the changes in customer injuries.

11. How have the most frequent types of offenses changed over the months?



Through the bar chart, we can see that the most frequent types of offenses change over the months. In the earlier months of 2020, Fare Evasion is the most dominant type of offense, with Alcohol Offenses and Violent Crimes trailing behind. After the onset of COVID-19, all three categories drop sharply, which can be attributed to the lower ridership. As ridership recovers, Violent Crimes, and Alcohol Offenses begin to make up a larger share of offenses with visible spikes in certain months later in the year. Overall, this analysis shows that both the level of incidents and the leading offense type vary by month, highlighting the need to account for shifting offense trends over time.

I V. Conclusion

A. Insights

1. How do incidents vary by time of day and between peak and off-peak periods?

Based on the results from the analysis done through Business Transactions 2 and 6, we see that incidents are heavily concentrated outside the morning peak hours(6AM to 9AM). Across the entire WMATA system, only about 11.75% of incidents occur during morning peak hours, while evening peak accounts for roughly 40.3% and off-peak hours account for about 48.0% of all incidents. The same pattern appears at the highest-incident stations identified in Business Transaction 2, as high-risk stations demonstrate the same pattern where most incidents occur during evening peak and off-peak periods. As the evening peak covers a shorter time window than off-peak, the rate of incidents per hour is highest during the evening peak hours(3PM to 7PM), which is during the evening commute. In practice, this means that the transit system is at the highest risk on a per-hour basis in the late afternoon and early evening, not in the morning.

2. **How does the total crimes and injuries change over time from 2020 to 2022?**

Based on the results from the analysis done through Business Transactions 5, 9, 10, and 11, we find that from 2020 to 2022 WMATA recorded a total of 9,581 crime incidents. Total crimes drop sharply in the early months of 2020, coinciding with the onset of COVID-19 in the region and the associated drop in ridership, and then remain relatively low before trending upward again in 2022 as the social-distancing restrictions loosen and riders return. Shown in Business Transaction 10, Customer injury rates are consistently higher and more volatile than employee injury rates and show a gradual upward trend from late 2020 into 2022, while employee injuries remain lower and relatively stable. These findings indicate that overall changes in safety conditions during this period are driven primarily by fluctuations in crime and customer injuries rather than by changes in employee injury rates.

3. **Which stations and rail lines have the highest frequency of daily incidents?**

Based on Business Transactions 1, 2, 3, 7, and 8, the occurrence of crime incidents are not evenly distributed across the network's stations and lines. The Green Line accounts for the largest share of incidents, with about 43% of all reported events and 4,130 incidents from 2020 to 2022, making it the highest-risk line in the system. Shown in Business Transaction 2, at the station level, Gallery Place–Chinatown has the most reported incidents, followed by Anacostia, L'Enfant Plaza, Hyattsville Crossing, Silver Spring, Addison Road–Seat Pleasant, Minnesota Ave, and Pentagon City. Business Transaction 3 shows that incident counts rise with the number of lines serving a station, as single-line stations average around 94 incidents, while four-line stations average 188 and the five-line transfer stations average 415. Together, these results confirm that multi-line transfer stations and stations in the Green Lines account for the majority of share of incidents, and therefore should be considered the primary geographic focus for safety efforts.

4. **What are the most frequent offense categories and how do they change over time?**

From Business Transactions 4 and 11, we see that Violent Crimes, Fare Evasion, and Theft/Property Crimes are the three dominant offense categories and account for a large share of all reported incidents. Vandalism/Property Damage, Justice/Legal Obstruction, and Alcohol Offenses form the second tier in terms of frequency, with Sexual Offenses, Disorderly/Public Disturbance, Drug Offenses, and Burglary/Trespass being still meaningful but less frequent. Through Business Transaction 11, we see that the trends of frequent offenses also change over time. In early 2020, Fare Evasion is the leading offense, followed by Alcohol Offenses and Violent Crimes. However, all three major categories dropped sharply after COVID-19 attributed ridership declines. As ridership recovers, Violent Crimes and Alcohol Offenses become more prominent again, with visible spikes in certain months. This demonstrates that both overall incident volume and the leading offense types are dynamic and influenced by broader system and environmental conditions, and suggests that such conditions should be considered when making safety efforts.

B. Recommendations

Taken together, these findings point to several concrete actions for WMATA. First, the analysis from Business Transactions 1, 2, 3, 7, and 8 makes it clear that safety resources should be concentrated where incidents are most frequent rather than spread evenly across the system. In particular, the Green Line and major transfer hubs such as Gallery Place–Chinatown register the highest incident counts and the largest shares of total incidents. WMATA and the Metro Transit Police should conduct targeted reviews to understand the drivers behind the Green Line's elevated incident levels and use those insights to guide operational changes. In the short term, this justifies a higher Metro Transit Police presence and more frequent safety checks on the Green Line and other multi-line stations, while lines and stations with significantly lower incident rates, such as many Yellow Line locations, can be staffed more leanly without ignoring basic safety obligations.

Second, the time-of-day patterns identified in Business Transactions 2 and 6 strongly argue against uniform staffing across the day. Since evening peak and off-peak periods together account for nearly 90 percent of incidents, and evening peak in particular has the highest incident rate per hour, WMATA should shift personnel, patrols, and supervision toward these periods. Rather than treating the "rush hour" as a morning phenomenon, safety planning should regard the late afternoon and evening as the most critical window. Enhancing CCTV monitoring, improving overhead lighting, and increasing visible staff presence during evening peak and key off-peak windows at high-incident stations would better align resources with an actual risk.

Third, the offense patterns from Business Transactions 4 and 11 indicate that WMATA cannot treat all offense types as equally important. Violent Crimes, Fare Evasion, and Theft/Property Crimes drive much of the system's safety burden and should be the focus of targeted strategies. At the same time, the consistent gap between customer and employee injury rates shown in Business Transaction 10 suggests that customer safety initiatives will have the largest impact on overall incident counts.

C. Limitations

One limitation of this analysis is the absence of ridership data for each station. Because we only used raw incident counts, stations with higher passenger volume would naturally appear to have more crime. Having ridership data would have allowed us to normalize crime incident counts and compare stations fairly.

References

Open Data DC. (2012, July 13). *Metro Stations Regional*. Open Data DC. https://opendata.dc.gov/datasets/DCGIS::metro-stations-regional/about

WMATA. (n.d.). *Crime Statistics*. WMATA. Retrieved December 6, 2025, from https://www.wmata.com/about/transit-police/crime-stats.cfm

WMATA. (n.d.). *Public Records Performance Records Excel*. WMATA. Retrieved December 6, 2025, from https://www.wmata.com/about/records/public-records.cfm