Comprehensive Analysis of Criminal Pattern in Washington, DC, US

GROUP 2

DATASET

- **Source:** Open Data DC which is managed by the Enterprise Data team in the Office of the Chief Technology Officer (OCTO).
- **Number of observations:** 29292
- **Link:** Crime Incidents in 2024

PROJECT 1 SUMMARY

- Examined crime patterns in Washington, D.C., during 2024, analyzing variations across neighborhoods, seasonal fluctuations, and the most frequently reported offenses.

- Investigated the correlation between crime type and time of day, as well as identifying neighborhoods with the highest homicide rates.

- Findings reveal that crime is unevenly distributed, with certain areas experiencing significantly higher rates.

- Statistical analysis confirmed a significant correlation between crime type and time of day, emphasizing the need for time-sensitive law enforcement strategies.

RESEARCH QUESTIONS

1. How do felony counts vary across different police districts and wards in Washington, D.C., during 2024?

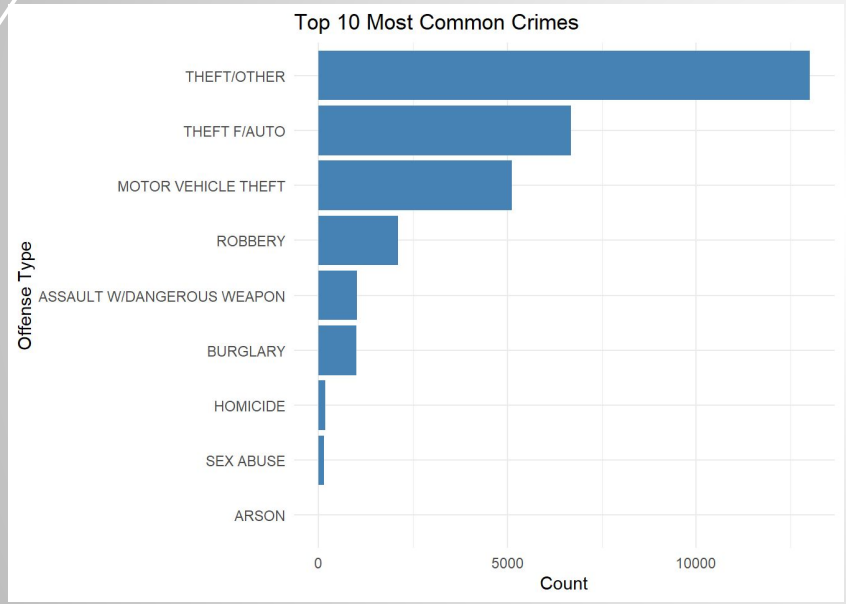
2. How does the type of felony offense vary by police shift in Washington, D.C., in 2024?

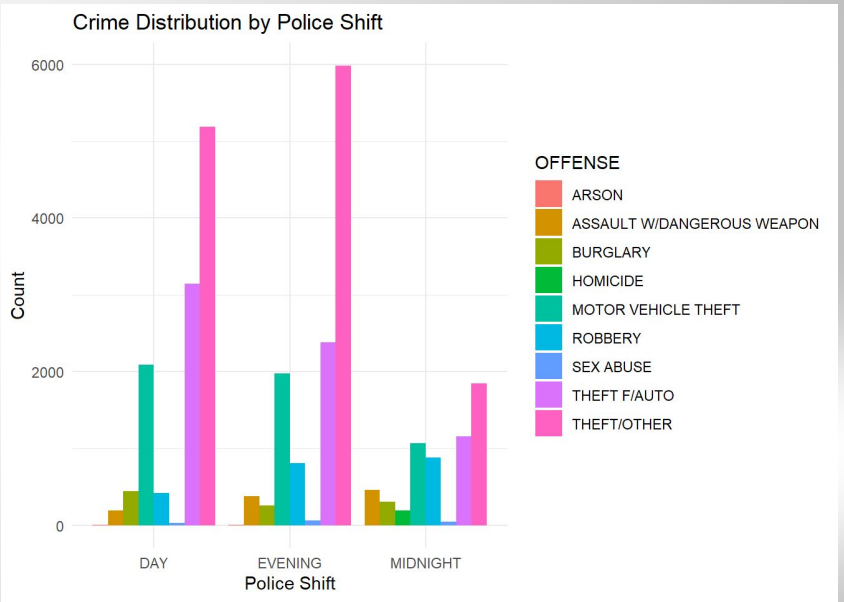
3. Which neighborhood characteristics are most predictive of felony counts in Washington, D.C.?

4. Does the method of crime influence the likelihood of felony offenses occurring in different neighborhoods?

5. What is the relationship between Business Improvement Districts and felony counts?

EXPLORATORY DATA ANALYSIS





Correlation Between Felony Counts & Police Districts/Wards

```
##
##  Pearson's product-moment correlation
##
## data:  crime_data_clean$DISTRICT and crime_data_clean$Is_Felony
## t = 25, df = 29292, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.131 0.153
## sample estimates:
##   cor
## 0.142
```

- **Correlation coefficient (`cor = 0.142`)** – weak positive correlation between the two variables.
- **p-value (`< 2e-16`)** – suggests that the correlation is statistically significant.
- **Confidence interval (`0.131 - 0.153`)** – a range where the true correlation is likely to fall with 95% certainty

- **Weak positive correlation** (`cor = 0.119`) between `WARD` and `Is_Felony`.
- **Correlation coefficient (0.119)** – small positive correlation.
- **p-value (< 2e-16)** – A very small p-value indicates that the correlation is statistically significant.
- **Confidence interval (0.107 - 0.130)** – The true correlation is likely to fall within this range with 95% certainty

```
##
##  Pearson's product-moment correlation
##
## data:  crime_data_clean$WARD and crime_data_clean$Is_Felony
## t = 20, df = 29292, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.107 0.130
## sample estimates:
##   cor
## 0.119
```

Predicting Felony Counts Using Linear Regression.

- Relationship between Is_Felony (the dependent variable) and several predictors (WARD, DISTRICT, METHOD & BID).
- **Intercept (0.9199, p < 2e-16)** – Strong baseline value.
- **District, crime method, and select BID areas** significantly impact whether a crime is classified as a felony
- **Ward does not appear to be a strong predictor** on its own

```
##
## Call:
## lm(formula = Is_Felony ~ WARD + DISTRICT + METHOD + BID, data = crime_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9712 -0.0870 -0.0703 -0.0490  0.9704
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.919913   0.007440  123.64  < 2e-16 ***
## WARD                             0.001416   0.000886    1.60  0.10991
## DISTRICT                         0.005532   0.001089    5.08  3.8e-07 ***
## METHODKNIFE                     -0.106692   0.014089   -7.57  3.8e-14 ***
## METHODOTHERS                    -0.867640   0.005958 -145.63  < 2e-16 ***
## BIDADAMS MORGAN                  0.057100   0.013423    4.25  2.1e-05 ***
## BIDANACOSTIA                     0.049578   0.022920    2.16  0.03054 *
## BIDCAPITOL HILL                  0.011070   0.013383    0.83  0.40816
## BIDCAPITOL RIVERFRONT           -0.019071   0.011089   -1.72  0.08547 .
## BIDDOWNTOWN                     -0.027769   0.007621   -3.64  0.00027 ***
## BIDDUPONT CIRCLE                -0.025181   0.018543   -1.36  0.17448
## BIDFRIENDSHIP HEIGHTS           -0.025479   0.036268   -0.70  0.48236
## BIDGEORGETOWN                   -0.021681   0.013994   -1.55  0.12131
## BIDGOLDEN TRIANGLE              -0.036609   0.013057   -2.80  0.00505 **
## BIDMOUNT VERNON TRIANGLE CID -0.031240   0.018393   -1.70  0.08943 .
## BIDNOMA                         -0.002142   0.010101   -0.21  0.83208
## BIDSOUTHWEST                    -0.017289   0.014377   -1.20  0.22914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.263 on 29277 degrees of freedom
## Multiple R-squared:  0.464,  Adjusted R-squared:  0.464
## F-statistic: 1.58e+03 on 16 and 29277 DF,  p-value: <2e-16
```

Does Felony Offense Occur More During Certain Police Shifts? (Logistic Regression)

```
##
## Call:
## glm(formula = Is_Felony ~ METHOD + SHIFT + NEIGHBORHOOD_CLUSTER,
##     family = binomial, data = crime_data_clean)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.92674    1.91137    1.01     0.31
## METHODKNIFE                    -1.21977    0.17313   -7.05  1.9e-12 ***
## METHODOTHERS                   -5.29231    0.10640  -49.74  < 2e-16 ***
## SHIFTEVENING                    0.06202    0.05442    1.14     0.25
## SHIFTMIDNIGHT                   0.99465    0.05651   17.60  < 2e-16 ***
## NEIGHBORHOOD_CLUSTERCluster 1   0.81220    1.91245    0.42     0.67
## NEIGHBORHOOD_CLUSTERCluster 10 -0.21136    1.94449   -0.11     0.91
## NEIGHBORHOOD_CLUSTERCluster 11 -0.15951    1.92750   -0.08     0.93
## NEIGHBORHOOD_CLUSTERCluster 12  0.00973    1.93105    0.01     1.00
## NEIGHBORHOOD_CLUSTERCluster 13  0.97358    1.93055    0.50     0.61
## NEIGHBORHOOD_CLUSTERCluster 6  -0.05526    1.91244   -0.03     0.98
## NEIGHBORHOOD_CLUSTERCluster 7   0.29879    1.91245    0.16     0.88
## NEIGHBORHOOD_CLUSTERCluster 8  -0.05631    1.91152   -0.03     0.98
## NEIGHBORHOOD_CLUSTERCluster 9   0.23351    1.91695    0.12     0.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25041  on 29293  degrees of freedom
## Residual deviance: 14910  on 29244  degrees of freedom
## AIC: 15010
##
## Number of Fisher Scoring iterations: 10
```

- Crimes committed using a knife are significantly less likely to be classified as felonies.
- Crimes occurring during the midnight shift are significantly more likely to be classified as felonies.
- The reduction in deviance suggests that the model improves upon the baseline.
- This model suggests that crime method and shift play a meaningful role in determining felony classification, while neighborhood cluster does not appear to have a strong independent effect.

ANOVA Test on Crime Method : Does Crime Method Influence Felony Classification?

```
## Analysis of Variance Table
##
## Response: Is_Felony
##              Df Sum Sq Mean Sq F value Pr(>F)
## METHOD        2   1749     874   12541 <2e-16 ***
## Residuals 29291   2042       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

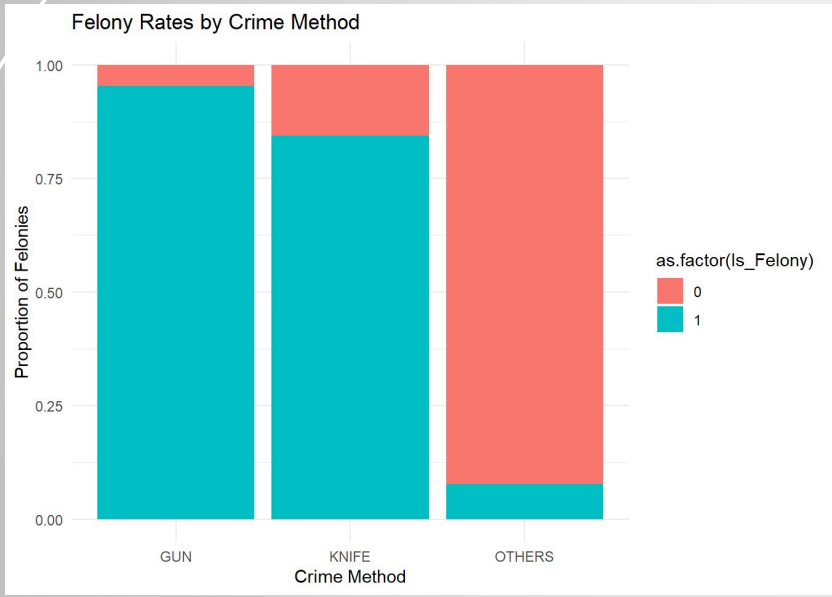
- The p-value is extremely small (`Pr(>F)<2e-16`) - the relationship between crime method and felony classification is highly statistically significant.
- F-value strongly supports that `METHOD` is an important factor affecting felony classification.
- The variance within `METHOD` is 874, suggesting that different crime methods significantly impact on felony classification.

T-Test for Felony Rate Differences : BID vs Non-BID Areas

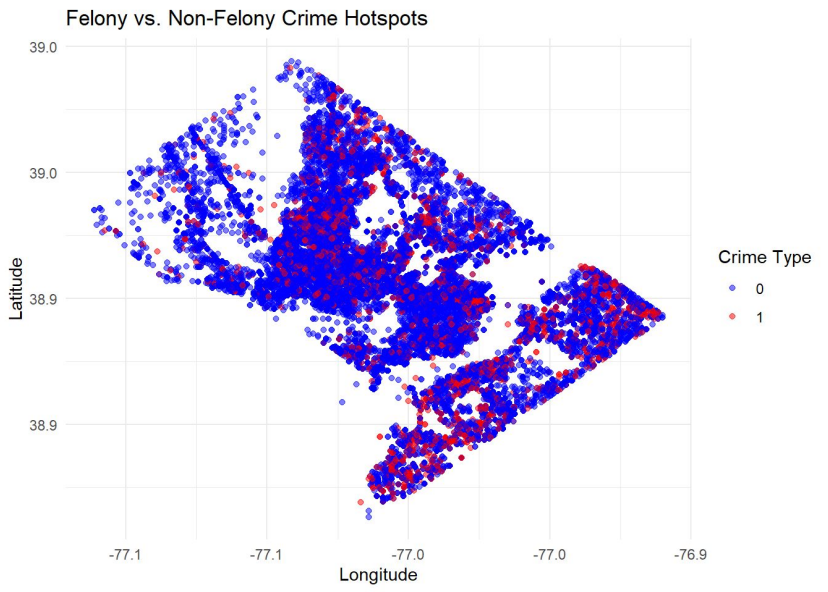
```
##  Two Sample t-test
##
## data:  Is_Felony by BID_presence
## t = -12, df = 29292, p-value <2e-16
## alternative hypothesis: true difference in means between group BID and group Non-BID is not equal to 0
## 95 percent confidence interval:
##  -0.0742 -0.0531
## sample estimates:
##     mean in group BID mean in group Non-BID
##                 0.101                 0.164
```

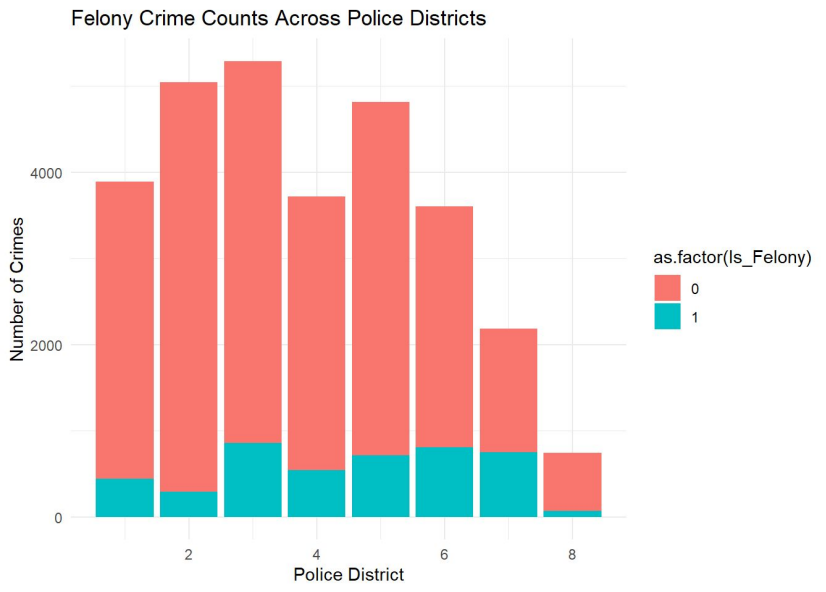
- We tested whether felony rates differ between areas inside vs outside BIDs.
- Mean Felony Rate:
- BID: 10.1% , Non-BID: 16.4%
- T = -12, p < 2e-16
- Conclusion: Statistically significant — Non-BID areas have higher felony rates

Felony Rate Comparison by Method and BID Area



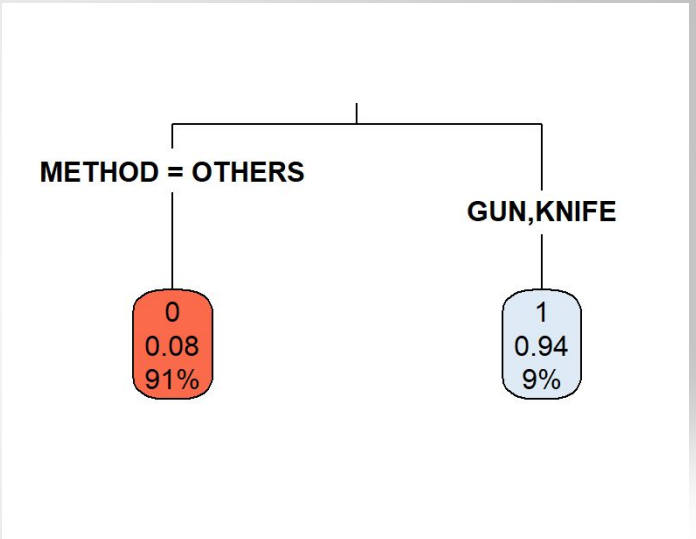
Felony Crime Distribution Map





Classification & Regression Trees(CART)

- **Root Node -** METHOD = OTHERS
- **Left Branch** - non-felony classification (0)
- **Right Branch** - felony classification (`GUN,KNIFE`)
- Crime method strongly influences felony classification.
- Crimes committed using guns or knives are significantly more likely to be classified as felonies.



Cross-validation for decision tree reliability

```
## CART
##
## 29294 samples
##     5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 26365, 26364, 26365, 26365, 26364, 26365, ...
## Resampling results across tuning parameters:
##
##   cp       RMSE   Rsquared  MAE
##   0.00201  0.262  0.468     0.138
##   0.00804  0.263  0.464     0.139
##   0.46018  0.323  0.439     0.212
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.00201.
```

RMSE (`0.262`) – The model's felony classification has an acceptable level of precision.

R-squared (`0.468`) – The model explains 46.8% of the variance in felony classification, indicating moderate predictive power

MAE (`0.138`) – The mean absolute error is relatively low, showing that predictions are generally close to actual values.

Summary of Key Findings

```
##   Total_Felonies Mean_Felonies SD_Felonies Median_Felonies
## 1           4474         0.153        0.36               0
```

- The dataset contains 4,474 felony cases.
- The mean is low, but the standard deviation suggests variability, meaning felony rates may be unevenly distributed across different areas.
- The median of zero highlights that felonies occur only in select locations, rather than being widespread.

Selection and determining the correct model to answer the SMART questions

Analyzed the dataset and identified felony classification was a binary problem.

Tested different models (linear regression, logistic regression and decision trees)

Based on accuracy, interpretability, and cross-validation results, Logistic regression performed well, providing clear insights into how crime method, police shift, and location influence felony likelihood.

Validated the model using 10-fold cross-validation:

- **Accuracy (92.4%) -** The model correctly classifies felony vs. non-felony crimes the vast majority of the time.
- **Sensitivity (True Positive Rate) (99.3%)** - It successfully detects non-felony crimes almost perfectly.
- **Specificity (True Negative Rate) (54.2%)** - The model struggles somewhat with felony classification but still provides useful insights.
- **Balanced Accuracy (76.8%)** - Accounts for the imbalance in felony vs. non-felony crime occurrences.

Interesting interpretations or predictions we can make with our model

- Midnight Shift result in Higher Felony Rate
- Guns leads to Most Predictive Crime Method for Felonies
- Business Improvement Districts (BIDs) results in lower Felony Crime Rates
- Felony Crimes Cluster Differently Across Police Districts
- Neighborhood Characteristics Have Limited Predictive Power

Thank you!