Task 6.1.

Data Source

This dataset was externally sourced from the Berlin Police, which every two years publish the "Berlin Crime Atlas", a report on the small-scale distribution of crime in Berlin. This publicly available data was then collected and curated by Danil Zyryanov on Kaggle, who translated the original report and united information pertaining to multiple years in one single dataset (originally, the information was divided by year).

Data Collection

The Berlin Police based their data reports on a regional evaluation of the Police Crime Statistics (PKS) data, which present the absolute number of cases as well as the regionally more comparable frequency figures (cases per 100,000 inhabitants) for the 12 Berlin districts and 138 district regions for 17 different crime areas.

Data Contents

The dataset contains information about small-scale crime in the city of Berlin, divided by districts and locations, from 2012 to 2019. It has 1200 rows and 20 columns (variables):, Year, District, Code, Location, Robbery, Street_robbery, Injury, Agg_assault, Threat, Theft, Car, From_car, Bike, Burglary, Fire, Arson, Damage, Graffiti, Drugs and Local.

Why this data choice?

Berlin being a dangerous city is a recurrent topic of conversation between me and my friends. I have never felt unsafe, even though I usually walk the city alone, but some of my friends seem to disagree and think Berlin is dangerous. Therefore, with this data analysis, I am curious to find

out about crime rates in different parts of Berlin and reach some conclusions regarding this topic.

Data Profile

I performed data wrangling and consistency checks in Python, dropping one column and renaming five columns, but not finding any missing values or duplicates. I also performed some basic descriptive statistics:

	Year	Robbery	Street_robbery	Injury	Agg_assault
count	1.200.000.000	1.200.000.000	1.200.000.000	1.200.000.000	1.200.000.000
mean	2.015.500.000	34.233.333	18.744.167	276.334.167	68.750.000
std	2.292.243	37.093.447	22.171.153	243.697.780	71.113.959
min	2.012.000.000	0.000000	0.000000	0.000000	0.000000
25%	2.013.750.000	10.000.000	5.000.000	108.000.000	22.000.000
50%	2.015.500.000	22.000.000	11.000.000	204.500.000	44.000.000
75%	2.017.250.000	42.000.000	23.000.000	361.000.000	86.000.000
max	2.019.000.000	242.000.000	169.000.000	1.966.000.000	500.000.000

	Threat	Theft	Car_theft	Car_burglary	Bike_theft
count	1.200.000.000	1.200.000.000	1.200.000.000	1.200.000.000	1.200.000.000
mean	92.583.333	1.492.307.500	42.505.833	215.275.000	197.706.667
std	68.455.264	1.364.442.501	28.710.164	150.031.343	178.704.771
min	0.000000	17.000.000	0.000000	1.000.000	0.000000
25%	42.000.000	639.750.000	22.000.000	109.000.000	76.000.000
50%	75.000.000	1.100.000.000	37.000.000	186.000.000	143.000.000
75%	124.000.000	2.019.750.000	57.000.000	291.000.000	286.000.000
max	420.000.000	12.479.000.000	197.000.000	876.000.000	1.288.000.000

	Burglary	Fire	Arson	Damage	Graffiti
count	1.200.000.000	1.200.000.000	1.200.000.000	120.000.000	1.200.000.000
mean	69.489.167	15.990.833	6.281.667	28.158.250	62.884.167
std	57.866.415	12.681.934	5.186.014	20.301.033	62.292.705
min	0.000000	0.000000	0.000000	0.00000	0.000000
25%	28.000.000	7.000.000	3.000.000	13.300.000	20.000.000
50%	59.000.000	13.000.000	5.000.000	24.400.000	45.000.000

75%	96.000.000	22.000.000	9.000.000	38.200.000	87.000.000
max	446.000.000	74.000.000	31.000.000	153.800.000	530.000.000

	Drugs_possession	Antisocial_behavior
count	1.200.000.000	1.200.000.000
mean	97.859.167	662.415.833
std	174.802.343	534.787.220
min	0.000000	10.000.000
25%	18.000.000	269.250.000
50%	40.000.000	553.500.000
75%	86.000.000	870.250.000
max	1.949.000.000	3.813.000.000

Data Types

Variables	Time-variant/-invariant	Structured/Unstructured	Categorical/Quantitative	Changes
Year	Time-invariant	Structured	Categorical	None
District	Time-invariant	Structured	Categorical	None
Code	Time-invariant	Structured	Categorical	Dropped this column
Location	Time-invariant	Structured	Categorical	None
Robbery	Time-variant	Structured	Quantitative	None
Street_robbery	Time-variant	Structured	Quantitative	None
Injury	Time-variant	Structured	Quantitative	None
Agg_assault	Time-variant	Structured	Quantitative	None
Threat	Time-variant	Structured	Quantitative	None
Theft	Time-variant	Structured	Quantitative	None
Car	Time-variant	Structured	Quantitative	Changed name to Car_theft
From_car	Time-variant	Structured	Quantitative	Changed name to car_burglary
Bike	Time-variant	Structured	Quantitative	Changed name to Bike_theft
Burglary	Time-variant	Structured	Quantitative	None
Fire	Time-variant	Structured	Quantitative	None
Arson	Time-variant	Structured	Quantitative	None
Damage	Time-variant	Structured	Quantitative	None
Graffiti	Time-variant	Structured	Quantitative	None
Drugs	Time-variant	Structured	Quantitative	Changed name to Drugs_possession
Local	Time-variant	Structured	Quantitative	Changed name to Antisocial_behavior

Data Limitations and Ethical Concerns

Although the Berlin Police has made data from 2020 to 2023 available, I was unable to incorporate the more recent data into my analysis due to time constraints and limited proficiency in German. As a result, the dataset used for this project contains fewer than 1,500 rows, which may limit the accuracy and depth of the analysis.

It is also important to acknowledge several limitations and ethical considerations associated with this dataset. First, the actual crime rates may be higher than reported figures due to underreporting, a well-documented phenomenon in crime statistics. Some crimes go unreported for various reasons, including fear of reprisal, lack of trust in law enforcement, or the perception that the incident is too minor to report.

Second, issues of racial bias present a significant ethical concern. Both police practices and public reporting can be influenced by implicit or explicit biases. This may result in disproportionately higher crime reports in geographical areas with a higher prevalence of certain racial or ethnic groups. It is critical to emphasize that higher crime reports in these areas do not necessarily indicate a higher incidence of crime but may instead reflect systemic biases in policing and reporting behaviors.

Additionally, I noticed discrepancies between some values in the dataset and the more recent data released by the Berlin Police. These differences suggest that certain data points might have been updated or revised in subsequent reports, which could impact the reliability of the earlier dataset used in this analysis.

Key Questions

- What is the Berlin area with the most small-scale crimes?
- Did crime increase from 2012 to 2019?
- What is the most common type of small crime in Berlin?

- Is there a correlation between the type of crime and specific Berlin districts?
- Are parks the locations with the most crimes?
- Is it less common for crimes to happen in smaller streets?
- Is there any specific type of crime decreasing or increasing?