



A PYTHON EDA PROJECT

---

## Project Statement Of Work - Crime Data Exploratory Data Analysis

---

*Prepared By: Zharys Consolacion*

Client: Nadine Green

Date: November 10, 2025

## Contents

<b>1 Project Background &amp; Purpose</b>	<b>3</b>
<b>2 Objectives</b>	<b>3</b>
<b>3 Scope</b>	<b>3</b>
3.1 Police forces . . . . .	3
3.2 Metrics and KPIs . . . . .	4
3.3 Time Period . . . . .	4
<b>4 Data Sources</b>	<b>4</b>
<b>5 Methodology</b>	<b>5</b>
5.1 Data Cleaning and Pre-processing . . . . .	5
<b>6 Deliverables</b>	<b>5</b>
<b>7 Timeline</b>	<b>5</b>
<b>8 Success Criteria</b>	<b>6</b>
<b>9 Assumptions &amp; Constraints</b>	<b>6</b>
<b>10 References</b>	<b>8</b>

## 1 Project Background & Purpose

This project is an analysis on the crime in the UK. The stakeholder, Nadine Green has requested this to give insight on which locations are most and least desirable to live in. The purpose is to give insight which will lead to more informed decisions around sales within the real estate company.

The requirements of the project are the following,

- To contain at least 4 police forces
- At least 2 years of data
- Consider crime data/street datasets. Outcome and stop and search are optional
- Exploratory data analysis should consider
  - General analysis between regions
  - Distributions regarding crime type, output, top crimes
  - Crime map
  - Others analysis perspective that you would consider relevant
  - Consider time as a variable
  - Visualisations and Analysis

This project will focus on the police forces serving in and around the M25 region. Therefore, the main purpose will be to provide insight on which areas around the M25 region are more desirable for those that want to live train commuting distance from London. Further specific on the scope of the project to be given within the Scope, section 3, this report.

## 2 Objectives

This analysis aims to answer the following questions,

- What crimes are the most prevalent in each region?
- Which regions have the most crimes reported? Over time?
- Are there any crime hotspots in these regions?

## 3 Scope

### 3.1 Police forces

This analysis will focus on the police forces serving in and around the M25 region. These forces are the following,

- Essex Police
- Kent Police
- Surrey Police
- Hertfordshire Constabulary
- Thames Valley Police

Alternatives to living within London and Greater London want to be given, therefore analysis will omit London's major police force, the Metropolitan, and also City of London Police Force.

### 3.2 Metrics and KPIs

The main metrics/KPIs we will be looking at are the following,

- Crime Rate - Crimes Per 10,000
- Frequency of Stop and search per Region

### 3.3 Time Period

A time period of October 2023 - September 2025 will be used. This is to give insight on what crime has been like in the short term period.

## 4 Data Sources

The main source of data is data.police.uk [1], and the following data files are taken from this site,

- Street-level crime data - A log of crimes happening in the jurisdiction of each force.
- Stop and search data - A log of all stop and searches carried out by each force.
- Force boundaries - Kml files defining the boundaries of the UK in which each of the forces serve

Other data sources outside data.police.uk are used. These include the following,

- Population statistics - the population that each police force serves. This data was found on HMICFRS [2].
- M25 boundary - this data was found on OpenData [3].

## 5 Methodology

Street-level crime and Stop and search data will be taken from data.police.uk, and comes in the form of csv files. The main tool for cleaning and analysis will be Python.

### 5.1 Data Cleaning and Pre-processing

To perform data cleaning and pre-processing, the following steps will be made.

1. Loading and joining of CSVs to form one for Street-level data and one for Stop & Search Data
2. Creating copies of raw-data after joining
3. For each CSV locate NaN values, and where appropriate
  - Drop rows or columns
  - Replace values
4. Locate anomalies

## 6 Deliverables

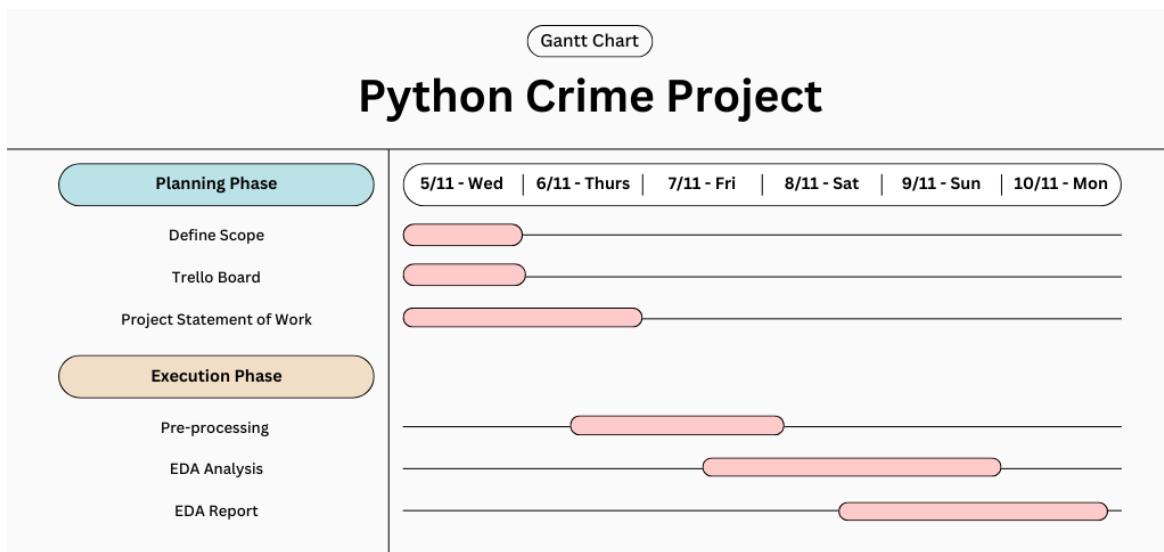
This project will produce the following,

- Planning Phase
  - Project of Work - PDF (this document)
  - Trello Board
- Execution phase
  - Preprocessing Script - jupyter notebook
  - EDA Script - jupyter notebook
  - README.md
- EDA Report - PDF
  - Analysis breakdown
  - Workflow diagram for Preprocessing
  - Trello board information

## 7 Timeline

The timeline of this project can be depicted in figure 1, represented in the form of a gantt chart.

A Trello board to structure the tasks and flow of the project was created [4]. Update of progress to be given in the EDA report.



**Figure 1:** Gantt chart timeline of Python Crime Analysis Project

## 8 Success Criteria

Aligned with the objectives of this project, key indicators of success can be given as,

- A conclusion can be made regarding which regions have the most crimes reported
- A map or visual representation of hotspots delivered, with clear legend and explanation
- A conclusion can be given about which two forces to continue further analysis on

Overall, it is aimed for all objectives to be met, allowing for the Real estate company to make more informed decisions around sales. All deliverables to be handed in by 10/11/2025 11:59pm. Meetings to be held with stakeholder, to ensure satisfaction with the findings.

## 9 Assumptions & Constraints

For this project we make the following assumptions,

- In the time period that this period looks at, the boundaries of the forces do not change
- Assume stop and search practices are performed fairly and uniformly
- Crime events are independent from one another
- Definitions of crimes do not change over this time period

## **9 ASSUMPTIONS & CONSTRAINTS**

---

- Population data is representative in the time period we look at

The following constraints are,

- Longatitude and Latitude data provided is anonymised coordinates about the crime, so is only accurate to the central point of a street, or a commercial premise [1]
- There are higher risk of quality issues for some forces that work as distinct systems, due to more manual handling.
- Different IT systems may leave out a certain column of data, leading to a huge amount of data of one type to be missing
- Population data is only accurate to the 3 significant figures

## **10 References**

- 1.Data.Police.UK. Data downloads — data.police.uk [Internet]. Police.uk. 2024. Available from: <https://data.police.uk/data/>
- 2.HMICFRS. Police forces [Internet]. His Majesty's Inspectorate of Constabulary and Fire & Rescue Services. 2025. Available from: <https://hmicfrs.justiceinspectorates.gov.uk/police-forces/>
- 3.Motorway Boundary Files – Fusion Data Science Open Data [Internet]. Fusion-datasience.com. 2019 [cited 2025 Nov 9]. Available from: <https://opendata.fusiondatasience.com/2019/02/16/motorway-boundary-files/>
- 4.Trello board [Internet]. Trello.com. 2020 [cited 2025 Nov 10]. Available from: <https://trello.com/invite/b/690a57bc5451bf2609711f8f/ATTI2d6c5b1e59bdcd54e88d95d604b91crime-data-exploratory-data-analyst-management>