**Origin of data**

Site Url:

* The crime count is categorized by Region, Crime Type, and Year (spanning from 2013 to 2022), which will be referred to as the "Base Statistics Set."
* The crime count is also broken down by Crime Type, Age, and Year (covering the years 2013 to 2022), and this will be known as the "Age Statistics Set."
* Additionally, the crime count is segmented by Crime Type, Sex, and Year (for the years 2013 to 2022), henceforth called the "Sex Statistics Set."

**Cleaning and Transformation**

1. Converted the provided dataset from an unstructured format into a structured, tabular form using Python, facilitating easier data manipulation and analysis.

2. Eliminated any trailing spaces present in the data to ensure consistency and accuracy in text-based fields.

3. Excluded total count information from the analysis, as it was determined to be extraneous and potentially detrimental to the efficiency of machine learning algorithms.

For the Age statistics set:

• Calculated the percentage representation of each age category within every crime type and year.

For the Sex statistics set:

• Determined the percentage representation of each sex category within every crime type and year.

To prepare the final dataset for machine learning, two key distributions were performed:

• The crime counts were allocated across the dataset based on the percentage contribution of each age category. The resultant dataset includes fields for Age, Region, Crime Type, Year, and the corresponding crime count, making it suited for machine learning models

• Similarly, crime counts were distributed according to the percentage contribution of each sex category. The final dataset structured for machine learning contains information on Sex, Region, Crime Type, Year, and crime count,.

**Exploratory Analysis**

1. EDA for Sex, Region, Crime type, Year, and crime dataset
2. Correlation plot for features

The number of crimes shows a slight negative correlation with the type of crime committed, as well as with the region. Conversely, there is a slight positive correlation between the number of crimes and sex.

A screenshot of a graph

Description automatically generated

1. Density plot for crime count

The distribution of the crime count exhibits a slight rightward skew.

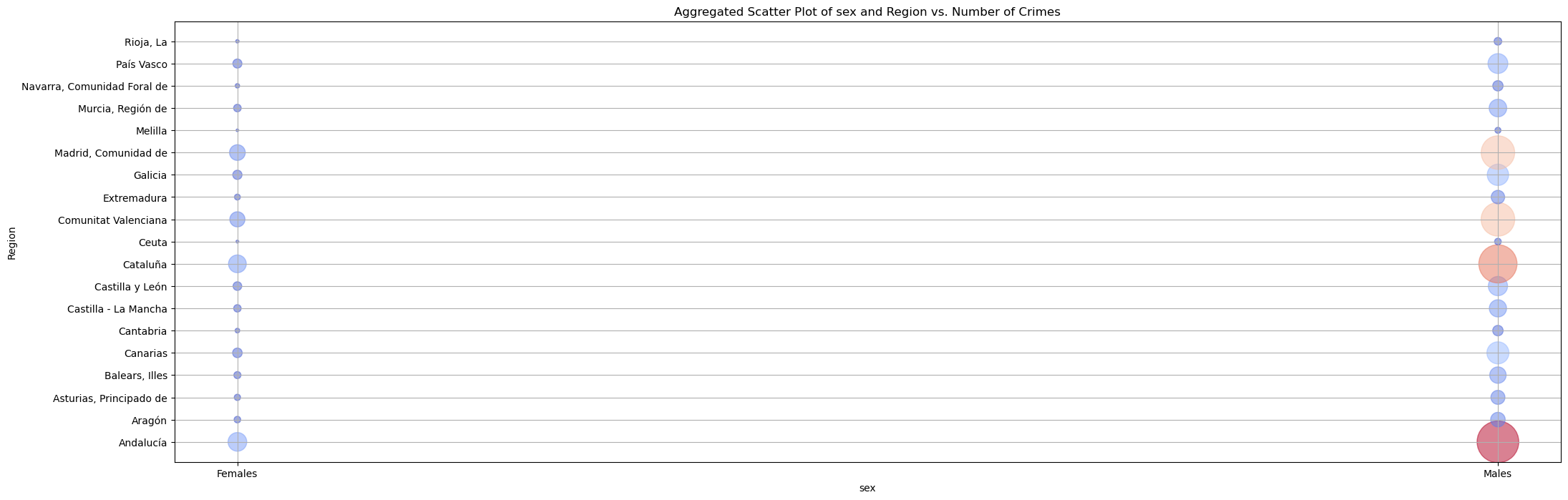
A graph with a line

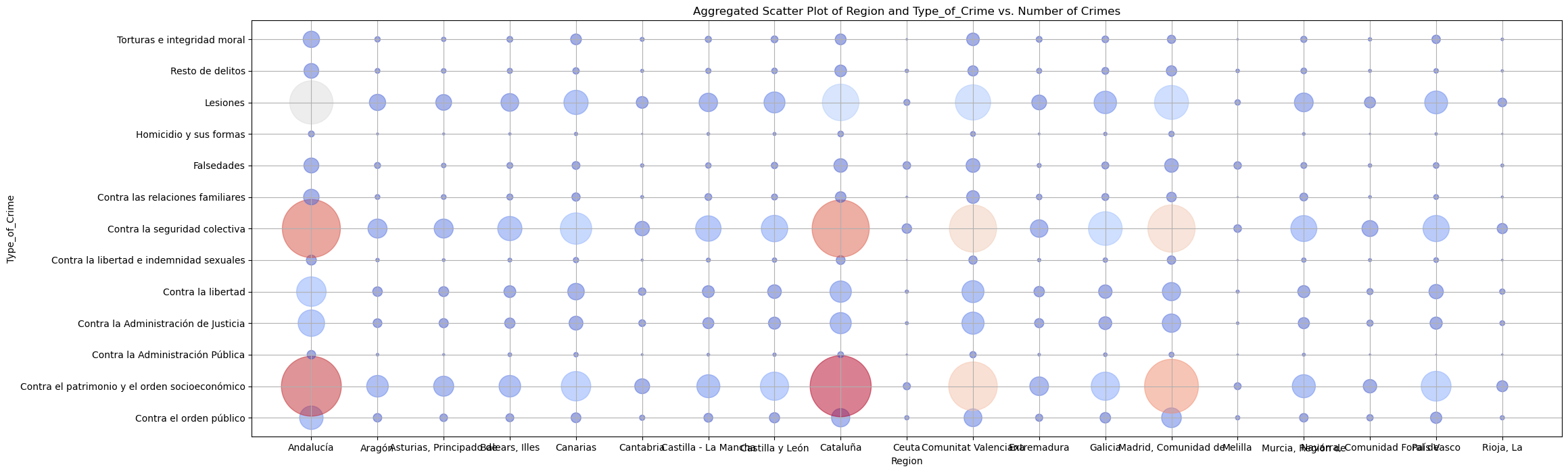
Description automatically generated

1. Correlation plot between sex vs Crime type, sex vs Region, Region vs Crime type

The correlation plot features a more intense and reddish color gradient, reflecting variations in crime count.A black line on a white background

Description automatically generated





1. EDA for Sex, Region, Crime type, Year, and crime dataset

The analysis of the results aligns closely with the outlined steps for modeling, which involve Sex, Region, Crime Type, Year, and Crime Count. So we are skipping in order to avoid repetition of explanation.

**Explanation of models, Measures to adapt model and Visualization**

1. Modelling for Sex, Region, Crime type, Year, and crime dataset
   1. Forecasting crime count by year at total level:

In this instance, the ARIMA model is employed for prediction purposes, with the parameters p, q, and d determined through the use of auto\_arima, resulting in the configuration (0,2,0). The outcome of the Augmented Dickey-Fuller( adfuller) test indicates that the series lacks stationarity.

A graph with a line

Description automatically generated

The forecasted result for three future timesteps are

A black background with white text

Description automatically generated

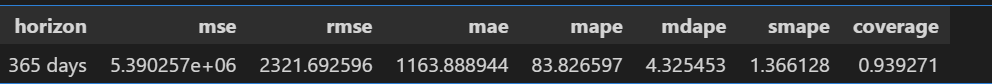
Following are the model summary statistics.

A screenshot of a computer

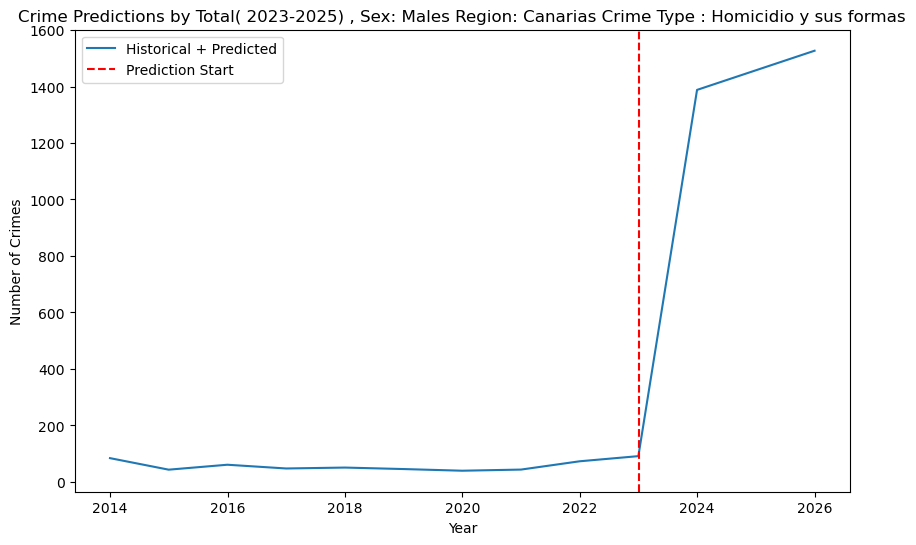
Description automatically generated

* 1. Forecasting crime count by year and other regressors ( sex, region and crime type) using Prophet model

Since there are categorical variables, label encoders are used for transforming those features before passing them to the model. Prophet model model came with a error metrics showing root means squared error as 2321 which shows that the model is not adapted well for this dataset. The reason for this is the smaller datatset which is having less time series components. Adjustment of hyper parameter variables prior\_scale to give ore preference to regressor variables also didn’t make any difference.Since there is no yearly seasonality, the seasonality is turned off for model creation.



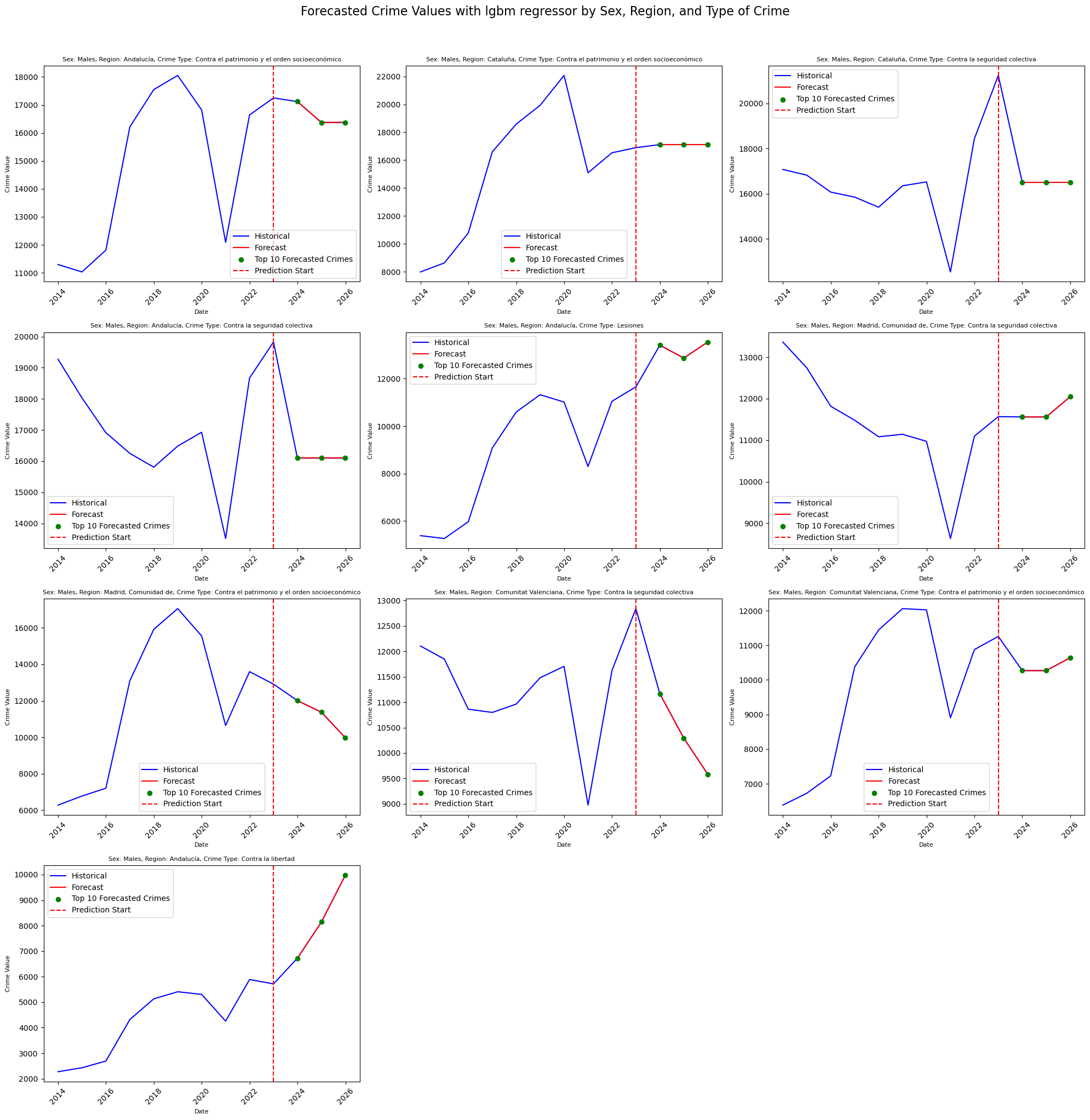
The model's final prediction for a sample category record confirms that it cannot continue any further.



* 1. Forecasting crime count by year and other regressors ( sex, region and crime type) using Light GBM regressor model

The LightGBM regressor is employed for prediction purposes. Given the absence of time series elements in this context, we introduced three lagged features from previous time steps for each record category. This approach ensures the establishment of a connection with past time steps. Additionally, label encoders were used to transform the data prior to its application to the model. The learning rate was adjusted to explore different results, ultimately being fixed at 0.001. The resulting Root Mean Square Error (RMSE) was approximately 511.

Predictions were made across all combinations of sex, region, and crime type for two future steps. The chart below presents the statistical results for the top 10 categories that have the highest crime counts.



* 1. Forecasting crime count by year and other regressors ( sex, region and crime type) using Linear regression model

A Linear Regression model is utilized for making predictions. Due to the lack of time series elements in the dataset, we incorporated three previous time steps as lagged variables for each record category, enabling the establishment of connections based on historical data. Label encoders were also used to preprocess the data before fitting it into the model. The model achieved a Root Mean Square Error (RMSE) of approximately 720. Predictions were executed across various combinations of sex, region, and crime type, extending two steps into the future. The chart provided below displays statistical outcomes for the top 10 categories exhibiting the highest counts of crime.A screenshot of a graph

Description automatically generated

* 1. Forecasting crime count by year and other regressors ( sex, region and crime type) applying ARIMA across individual sex, crime type and Region categories

To evaluate the model's predictive performance, a preliminary analysis was conducted using ARIMA on each category within a single region. The ARIMA model was specified with an order of (1,1,1) for (p,d,q).

A graph of blue and red lines

Description automatically generated

1. Modelling for Age, Region, Crime type, Year, and crime count dataset

The outcome of the analysis is closely aligned with the described modeling process, which involves factors such as Sex, Region, Crime Type, Year, and Crime Count. Skipping the explanation to avoid repetition.

**Final Conclusion:**

Among the models tested, the LightGBM model, which falls under the decision tree category, emerged as the top performer, delivering the best results. In contrast, the Prophet model did not fare as well, showing the least favorable outcomes. The number of records contained within the dataset played a significant role in this scenario, impacting the overall accuracy and effectiveness of the results to a certain degree.

**Further Possibilities**

* Employing Ridge or Lasso regression could enhance the predictive performance and accuracy of the linear regression model.
* Adopting probabilistic modeling with PYMC could substantially expand the dataset. Introducing stochastic variables and incorporating monthly increments may facilitate this expansion.
* Leveraging deep learning approaches, such as LSTM or Transformer models, has the potential to yield improved results compared to current models.