Prediction of US Car Accidents: A Time Series Approach

Group 2

- Jeebina Adhikary
- Karthik Elaprolu
- Prashanth Haranagiri
- Niriha Kommu
- Kriston Pal
- Rakesh Velaga

Agenda

INTRODUCTION

DATA COLLECTION

OBJECTIVE

METHOD OF FORECAST

EXPECTED OUTPUT

Introduction

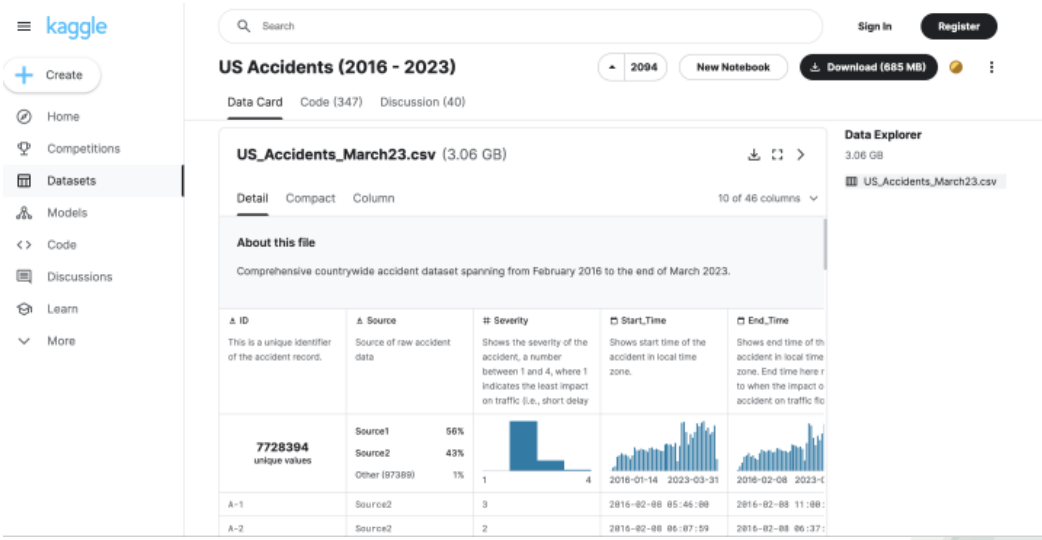
- This is a nationwide car accident dataset encompassing 49 states in the United States.
- The accident data were gathered between February 2016 and March 2023
- The collection presently comprises 7.7 million accident records.
- Forecasting will be based on the existing records
- Since the dataset is very huge, we only took one state (Connecticut)
- The dataset was not a time series data initially.
- We converted the dataset into time series.

Data Collection

- The dataset for this project was obtained from the following website:

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

- Data from 2016 to 2022 is taken from the dataset



Objective

The major goal of this research is to evaluate historical data on car accidents in the United States using time series forecasting techniques, notably Linear Regression and ARIMA models and see how effective are these model.

The overarching goal is to gain actionable insights into the patterns, seasonality, and influencing factors associated with accidents. By achieving this objective, the project aims to:

- Develop Accurate Forecasts
- Identify Seasonal Trends
- Evaluate Model Performance
- Explore Influencing Factors
- Facilitate Decision-Making
- Enhance Road Safety Strategies
- Communicate Findings Effectively

Method of Forecast

Linear Regression and ARIMA(Autoregressive Integrated Moving Average)

These are statistical analysis models that uses time series data to either better understand the data set or to predict future trends using the latest existing data.

Why Linear Regression?

- Linear Regression is employed to capture and model the linear relationship between time and the number of car accidents.
- This model is effective when there is a clear trend or pattern in the data that can be approximated by a straight line.
- Linear Regression is straightforward, interpretable, and provides a good baseline for understanding the overall trend in the time series data.

Why ARIMA?

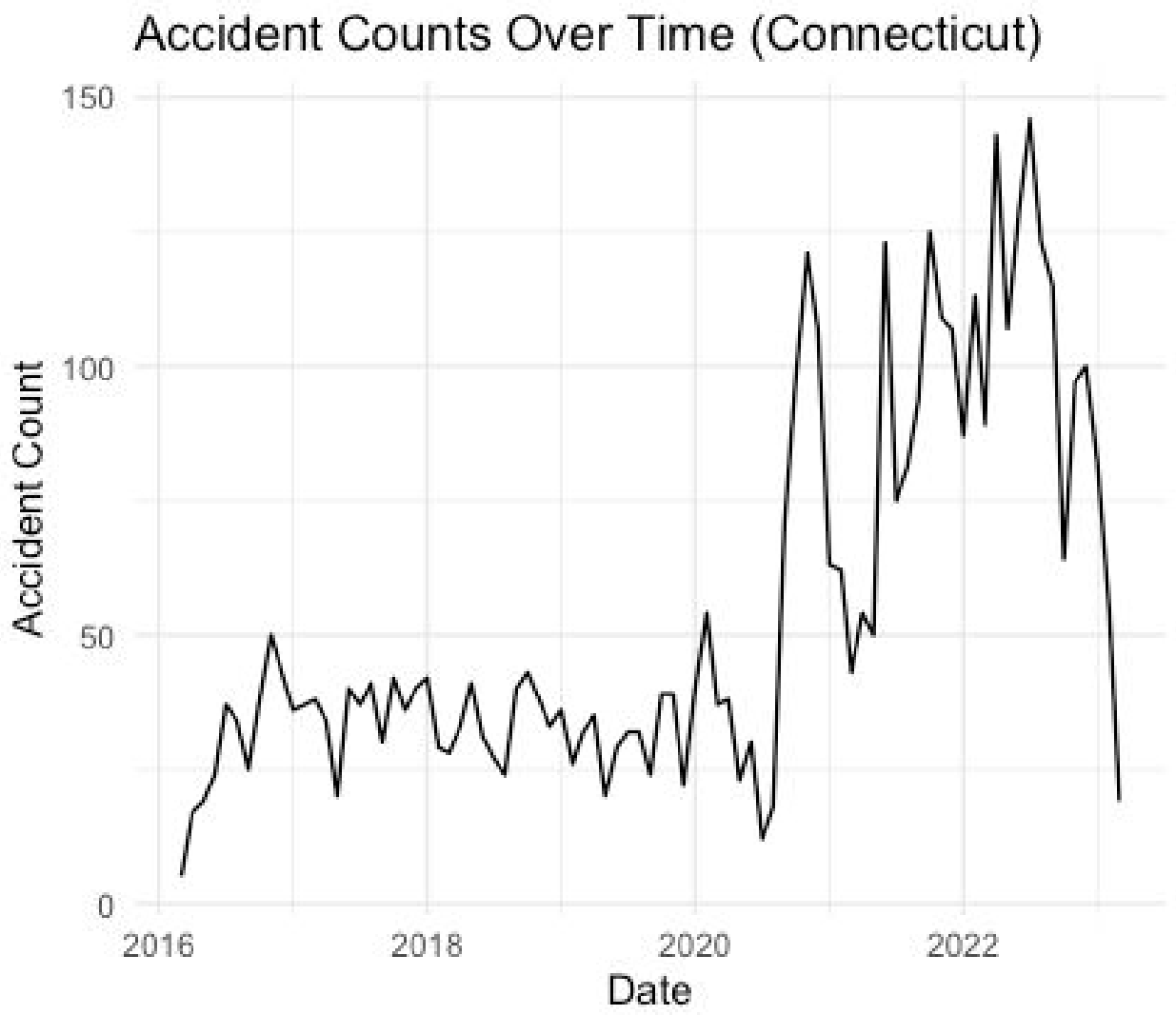
- ARIMA models are suitable for capturing more complex patterns in time series data, including seasonality and autocorrelation.
- They are adept at handling non-linear trends and fluctuations in the data.
- ARIMA models are flexible and can adapt to different patterns in the data.
- They are particularly effective when the time series exhibits a stationary behavior after differencing.

Variable

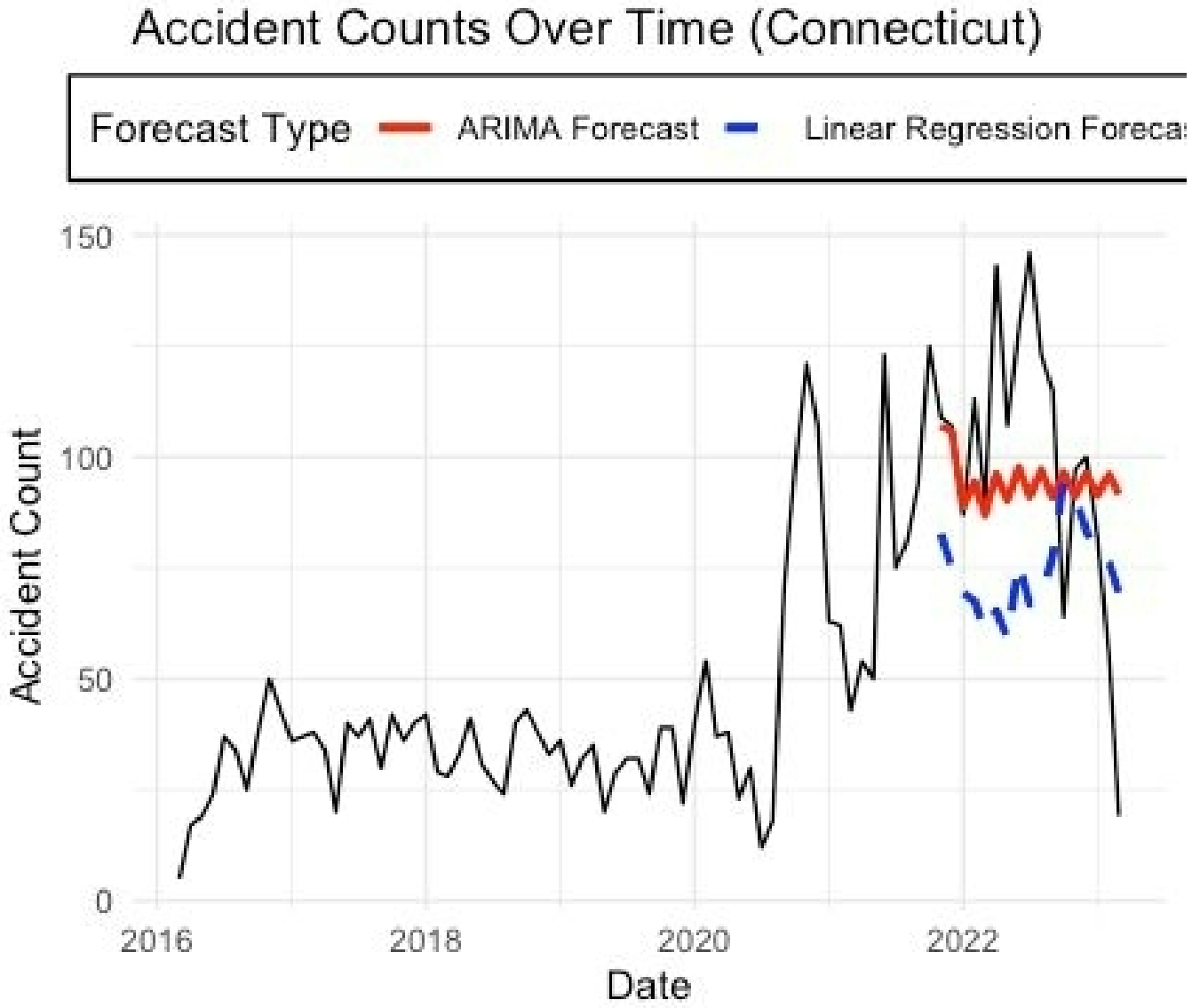
Accident Counts is used as the variable to forecast the car accidents in CT using ARIMA and Linear Regression.

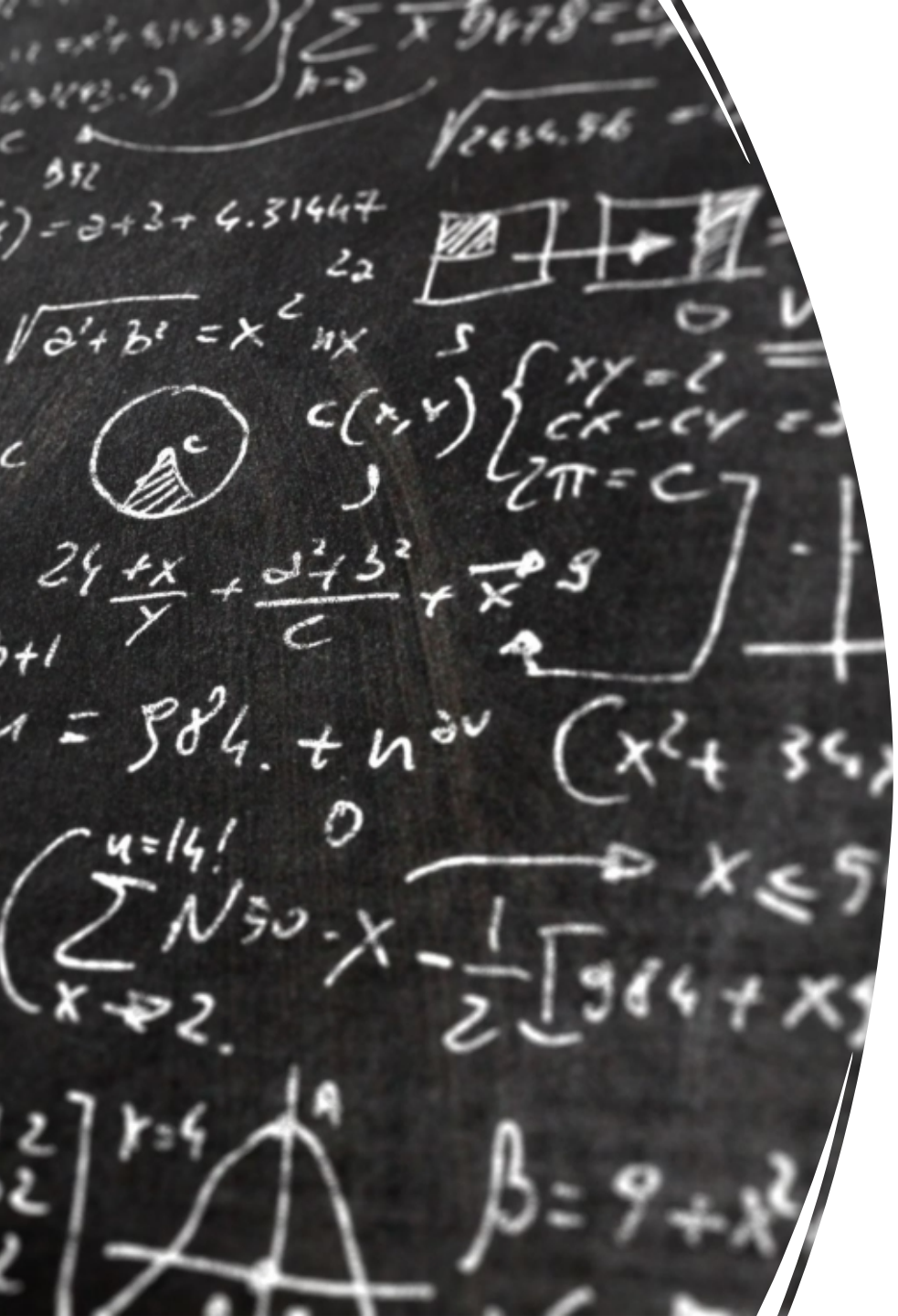
Understanding the Accident Trend and Seasonality in CT

- It has an increasing trend.
- We find the seasonality as there is a peak around December in almost every year



Forecasting using Linear Regression and ARIMA model





MAE value for Linear Regression and ARIMA model

LR_MAE = 36.90907

ARIMA_MAE = 23.00974

Conclusion

- The higher MAE for Linear Regression suggests that this model, on average, has larger prediction errors compared to the ARIMA model.
- Hence, the ARIMA model is performing better in predicting accident counts in CT than the Linear Regression model.
- Additional features, model tuning or a more sophisticated modeling approach might be explored to improve predictive performance.


Thank You