



Codeup Data Science

600 Navarro Street, San Antonio, TX  
78205

cristinajlucin@gmail.com

# Examining Boston Crime: Utilizing Time Series Data to Explore and Predict Fraud Crime With Pandemic Impacts



Cristina Lucin

Codeup Data Science Student

# Agenda



Executive Summary.

Project Description and Initial Thoughts

The Plan

About the Data

Exploration Questions

Modeling

Post-Modeling Analysis

Conclusions and Recommendations

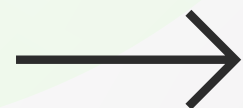
# Executive Summary →



Cristina Lucin  
Codeup Data Science

- This project utilized crime data from Analyze Boston, it contained approximately 650,000 rows, from which 21,000 rows contained fraud data
- The data was explored through time-series analysis, showing that Fraud Crime is not seasonal and trended generally flat before pandemic
- Fraud crime rates spiked during the pandemic, and modeling was not effective for this period
- Post-pandemic Fraud crime rates are significantly lower, and more data is needed to remodel these crimes
- Resampling Fraud data by week or month could improve modeling

# Project Description and Initial Thoughts



[Back to Executive  
Summary](#)


## Project Description

This project was created to examine crime data from the City of Boston. Specifically, this project focuses on examining crime data collected by the City of Boston and examining trends in fraud crime occurrences at several periods. The onset of the COVID pandemic and its impact is seen through the visualization of this data, which leads to further questions about crime data forecasting before and after the pandemic, both within Fraud crimes and crime rates in general.

## Initial Thoughts

Fraud Crimes in Boston will increase after the COVID pandemic begins

# The Plan



- Acquire and clean the data
- Explore data in search of time series patterns for fraud crimes
- Answer the following questions
  - What was the average fraud crime rate in the two years immediately prior to the pandemic?
  - Are certain months more likely to demonstrate greater reported fraud crimes?
  - Do Fraud Crimes follow a seasonal pattern?
  - Is the downtrend seen in 2017 significant?
  - Are pre-pandemic Fraud crime levels equal to Fraud crime levels post-pandemic?
- Develop a Model to predict fraud crimes
- Evaluate models on train and validate data
- Select the best model to use on test data
- Draw Conclusions



# About the Data →

[Back to Agenda Page](#)

## Aquire

- Data aquired from data.boston.gov 'Crime Incident Reports as .CSV files by year
- Crime Incident Reports were collected beginning in 2015, and the data spans to present
- I combined the annual CSV files into one CSV, containing 652,418 rows and 17 features before cleaning
- Each row represents a crime reported to Boston Police (see further explanation re: Unique Incident Numbers)
- Each columns represents a feature of the crime, such as offense type, location, or time the crime occured

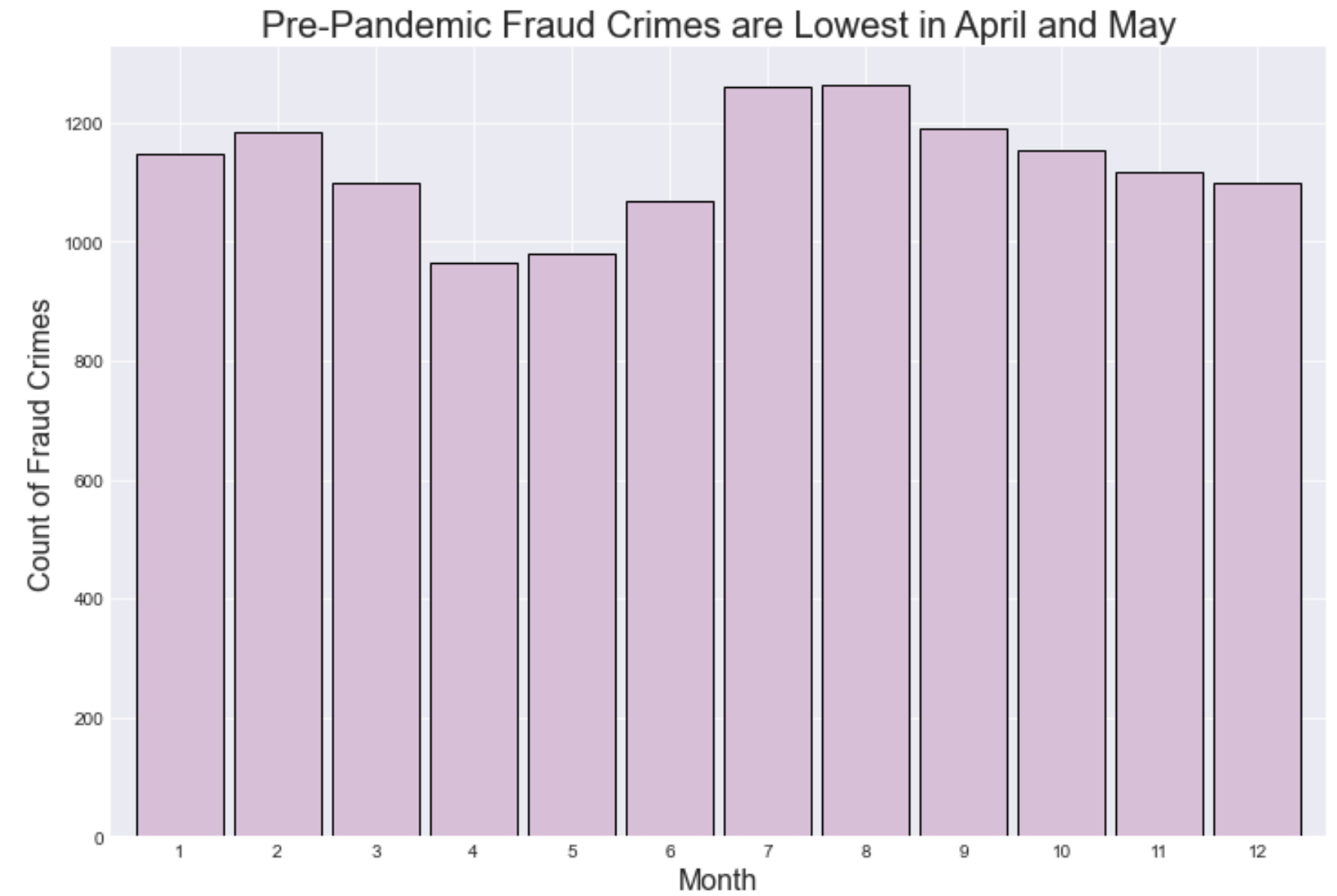
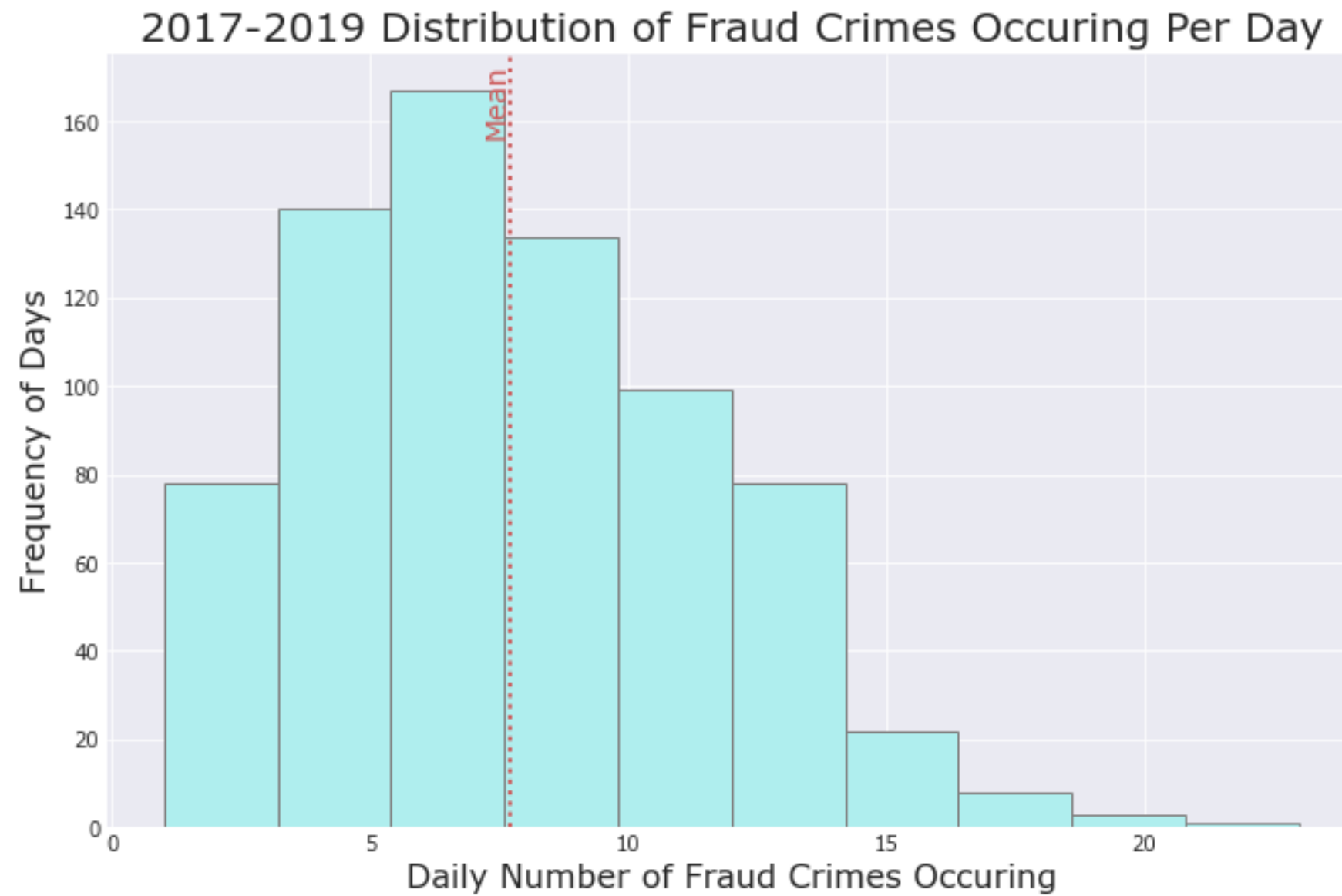
## Prepare

- Renamed columns to improve readability
- Checked that column data types were appropriate, changed data types when necessary
- Removed white space from values in object columns
- Checked for null values in the data, imputing where appropriate (location null values were not addressed for this iteration)
- Outliers were not removed in this iteration
- Utilized string functions to clean column data information

**\*\*From the 652,418 rows, 21,247 were crimes categorized as fraud\*\***

# Exploration Questions →

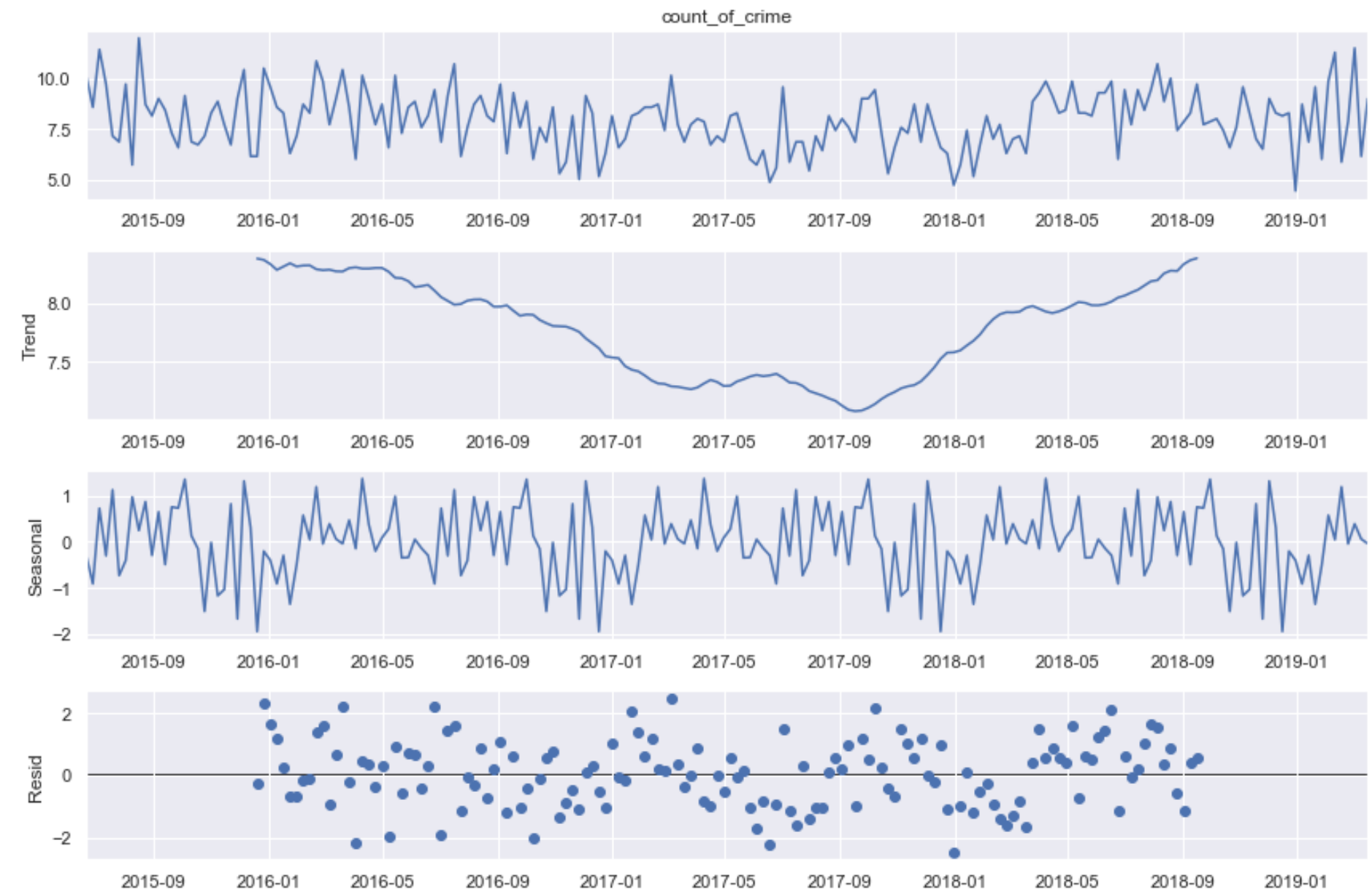
[Back to Agenda Page](#)



# Exploration Questions →

[Back to Agenda Page](#)

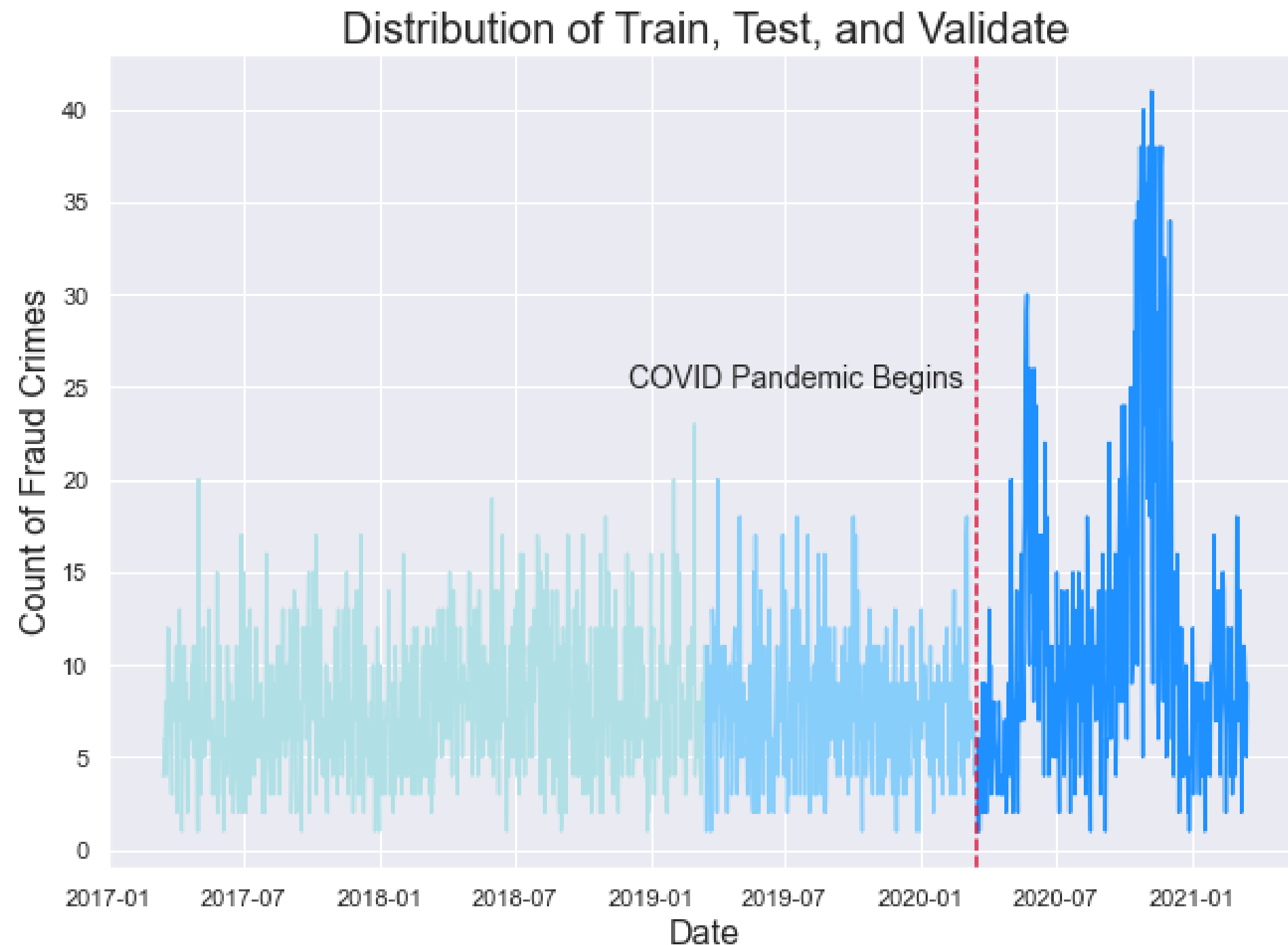
Some Seasonality is Present, with a Trend Down in 2017





# Modeling →

[Back to Agenda Page](#)



- I utilized RMSE as the evaluation metric
- I created five different model types with various hyperparameter configurations
- Models were evaluated on validation data
- The model that performed the best was evaluated on test data
- Baseline for modeling was chosen as Simple Average

# Modeling →

[Back to Agenda Page](#)

	model_name	train_score	validate_score
0	simple_average	4	3
1	30 d moving_average	4	4
2	90 d moving_average	4	4

	model_type	target_var	rmse
0	Holts	count_of_crime	4.0
1	Holts seaonal add	count_of_crime	4.0

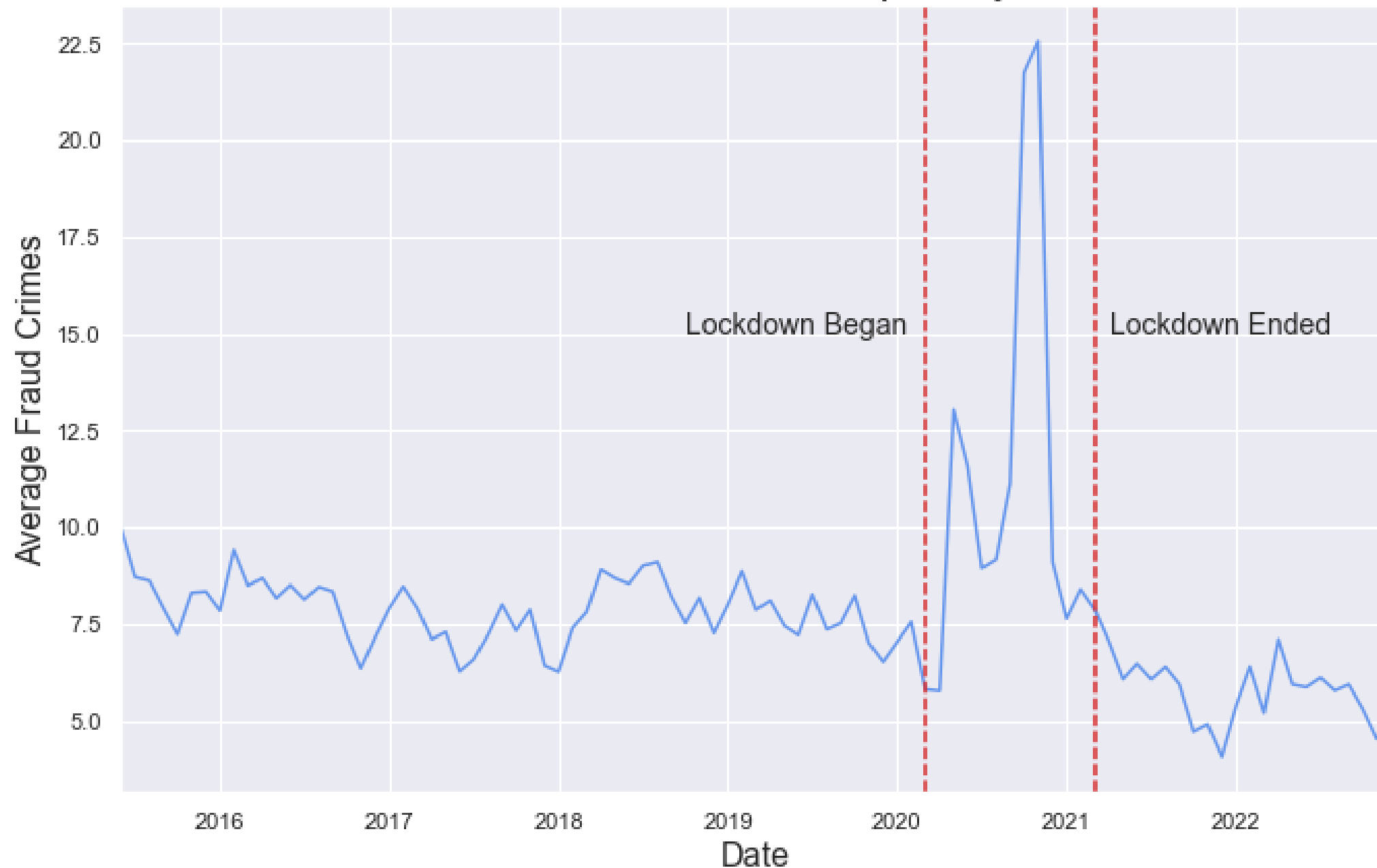
	model_name	train_score	validate_score	test_score
0	simple_average	4	3	9

- All models performed similarly on train, and validate
- Simple Average model, the baseline, was chosen for test
- Simple Average performed very poorly on test data, with a RMSE score of 9 (compared with 3 on validation data)

# Post-Modeling Analysis →

[Back to Agenda Page](#)

Mean Fraud Crimes Resampled by Month



- Our model for predicting fraud crimes using a simple average was equivalent to the baseline
- Fraud crimes spiked during the pandemic
- This visualization shows that since lockdown restrictions were eased in Boston in the spring of 2021, Fraud crimes appears to be decreasing
- Statistical testing showed that this change was significant

# Conclusions and Recommendations



Cristina Lucin  
Codeup Data Science



- When aggregating fraud crimes by day, there was low seasonality in this data
- Pre-pandemic, a simple average model was the best model for predicting fraud crime
- A simple average model performed extremely poorly on test data, which was the first year of COVID. In Boston, COVID restrictions meant strict lockdowns for the early stages of the pandemic
- Though lockdown restrictions have been completely lifted, the fraud crime rate has significantly decreased in Boston
- Resampling fraud crimes by week or month to get a clearer picture of seasonality
- Creating a new model for post-pandemic data, excluding lockdown periods due to severe outliers

[Back to Project  
Description](#)





Codeup Data Science



# Thank you for your time!

For further information about this project, check out my github:

<https://github.com/cristinalucin>

Find me on LinkedIn: <https://www.linkedin.com/in/cristina-lucin/>

[Back to Agenda Page](#)