MID PROJECT – DATA ANALIST

Motor Vehicle Thefts

Analysis, Visualization and Forecasting of Theft

CREATE BY

Diky Arianto Tarihoran

Data Analyst Enthusiast

A beginner data analysis enthusiast with high enthusiasm, seeking opportunities to learn and grow in data analysis and data science. Currently pursuing Independent Studies with a focus on Data Analysis at Cakap.

TABLE OF CONTENTS

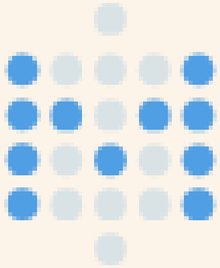
Contents

Intro

PROJECT OVERVIEW

Stolen vehicle data from the New Zealand police department's vehicle of interest database containing 6 months of data. Each record represents a single stolen vehicle, with data on vehicle type, make, year, color, date stolen and region stolen.

TOOLS USED



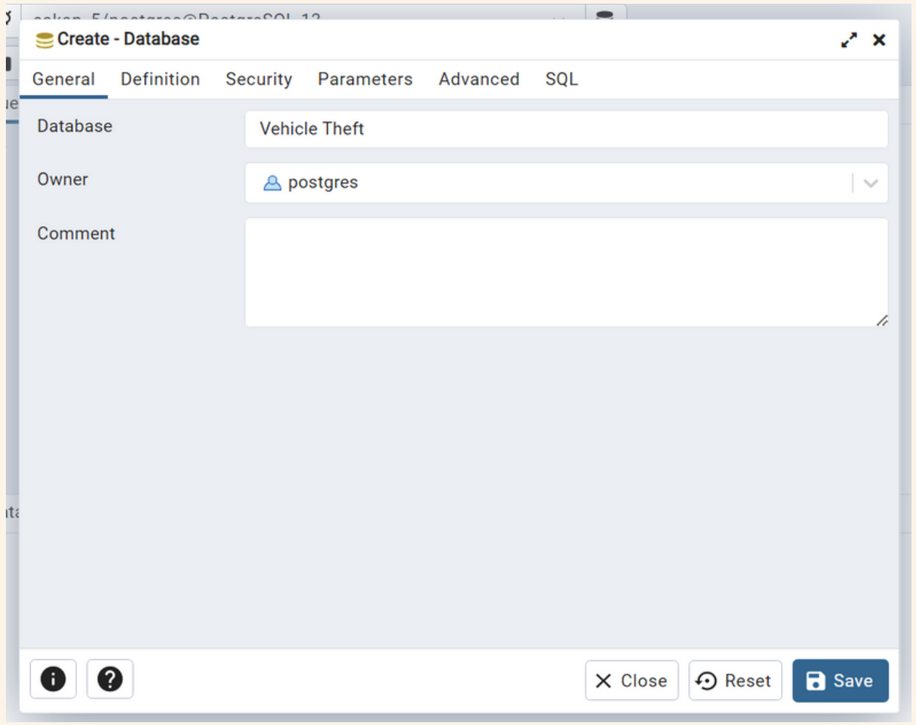
01

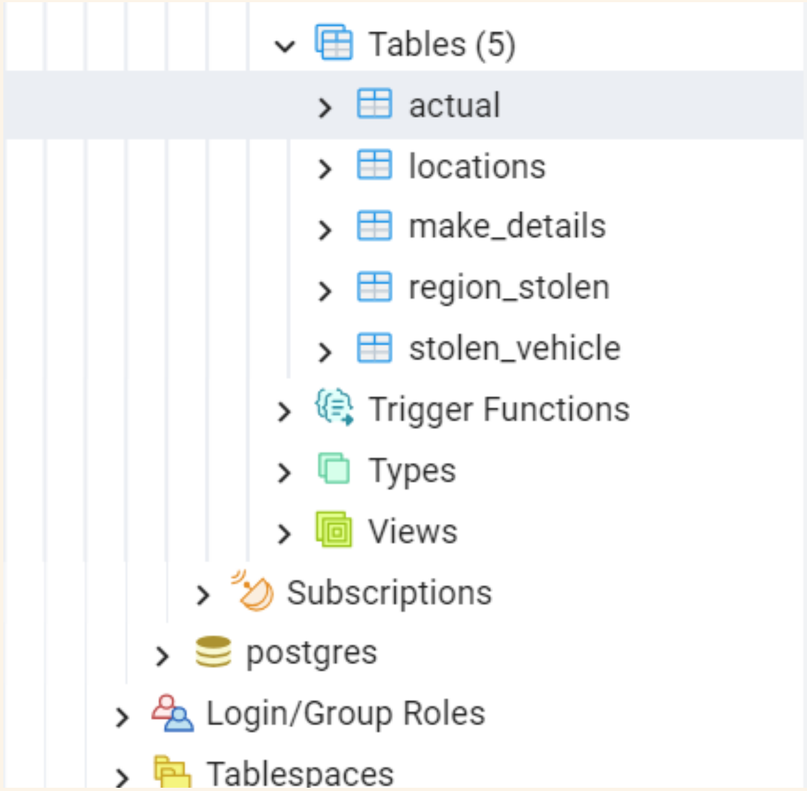
DATA PREPARATION

Chapter 1

DATA IMPORT (SQL)

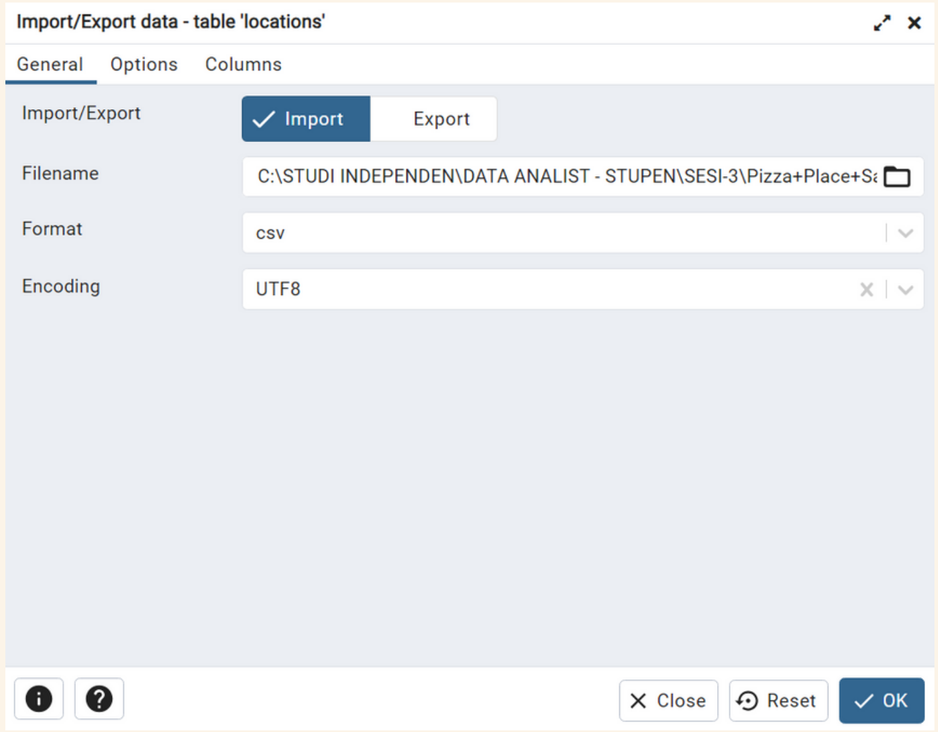
1. CREATE DATABASE





DATA IMPORT (SQL)

2. Import File



Query | Query History

```
1  select * from locations
```

Data Output | Messages | Notifications

	location_id integer	region character varying (100)	country character varying (100)	population integer	density numeric (10,2)
1	101	Northland	New Zealand	201500	16.11
2	102	Auckland	New Zealand	1695200	343.09
3	103	Waikato	New Zealand	513800	21.50
4	104	Bay of Plenty	New Zealand	347700	28.80
5	105	Gisborne	New Zealand	52100	6.21
6	106	Hawke's Bay	New Zealand	182700	12.92
7	107	Taranaki	New Zealand	127300	17.55
8	108	Manawatū-Whanganui	New Zealand	258200	11.62
9	109	Wellington	New Zealand	543500	67.52
10	110	Tasman	New Zealand	58700	6.10
11	111	Nelson	New Zealand	54500	129.15

Total rows: 16 of 16 | Query complete 00:00:00.294

DATA JOINING

There are 3 Tables (separate Datasets) containing data on vehicle thefts that have been cleaned and analyzed previously.

Tabels Description

stolen_vehicles:

- vehicle_id: Unique ID of a stolen vehicle
- vehicle_type,: Type of vehicle
- make_id: Matches make_id in the make_details table
- model_year: Model year of vehicle
- vehicle_desc: Description of vehicle
- color: Color of vehicle
- date_stolen: Date the vehicle was stolen (MM/DD/YY)
- location_id: Matches location_id in the locations table

make_details:

- make_id: Unique ID of the make
- make_name: Name of the make
- make_type: Type of make (Standard or Luxury)

locations:

- Unique ID of the region
- region: Name of the region
- country: Country where the region is located
- population: Population of the region
- density: Density of the region (population / km2)

DATA JOINING

```sql
SELECT
    sv.vehicle_type,
    sv.model_year,
    sv.color,
    l.region,
    COUNT(*) AS total_stolen_vehicles
FROM
    stolen_vehicle sv
JOIN
    locations l ON sv.location_id = l.location_id
GROUP BY
    sv.vehicle_type, sv.model_year, sv.color, l.region
ORDER BY
    total_stolen_vehicles DESC
```

vehicle_type character varying (100)	model_year character varying (100)	color character varying (100)	region character varying (100)	total_stolen_vehicles bigint
Trailer	2021	Silver	Auckland	17
Saloon	2005	Silver	Auckland	12
Saloon	2006	Silver	Auckland	12
Trailer	2017	Silver	Auckland	11
Stationwagon	2006	Black	Auckland	11
Saloon	2008	Silver	Auckland	11
Hatchback	2005	Silver	Auckland	11
Stationwagon	2006	Silver	Auckland	11
Trailer	2016	Silver	Canterbury	10
Stationwagon	2004	White	Auckland	10
Saloon	2007	White	Auckland	10
Trailer	2014	Silver	Canterbury	9
Saloon	2006	Black	Auckland	9
Trailer	2012	Silver	Canterbury	9
Trailer	2018	Silver	Auckland	9
Trailer	2019	Silver	Canterbury	9
Saloon	2006	White	Auckland	9
Trailer	2021	Silver	Canterbury	9

DATA IMPORT (JUPYTER)

In [248]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix
from statsmodels.tsa.arima.model import ARIMA
import statsmodels.api as sm
import itertools
```

In [5]:
```
# Ganti path file CSV dengan path yang sesuai
data = pd.read_csv(r"C:\STUDI INDEPENDEN\DATA ANALIST - STUPEN\SESI-3\Motor+Vehicle+Thefts+CSV\CAKAP_5\data_locations.csv")
data.head(30)
```

Out[5]:

	location_id	region	country	population	density
0	101	Northland	New Zealand	201500	16.11
1	102	Auckland	New Zealand	1695200	343.09
2	103	Waikato	New Zealand	513800	21.50
3	104	Bay of Plenty	New Zealand	347700	28.80
4	105	Gisborne	New Zealand	52100	6.21
5	106	Hawke's Bay	New Zealand	182700	12.92
6	107	Taranaki	New Zealand	127300	17.55
7	108	Manawatū-Whanganui	New Zealand	258200	11.62
8	109	Wellington	New Zealand	543500	67.52
9	110	Tasman	New Zealand	58700	6.10
10	111	Nelson	New Zealand	54500	129.15
11	112	Marlborough	New Zealand	51900	4.94
12	113	West Coast	New Zealand	32700	1.41
13	114	Canterbury	New Zealand	655000	14.72
14	115	Otago	New Zealand	246000	7.89
15	116	Southland	New Zealand	102400	3.28

DATA UNDERSTANDING

```
In [6]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   location_id  16 non-null     int64
 1   region       16 non-null     object
 2   country      16 non-null     object
 3   population   16 non-null     int64
 4   density      16 non-null     float64
dtypes: float64(1), int64(2), object(2)
memory usage: 772.0+ bytes
```

In the locations dataset, there are 5 columns containing data for 16 rows. All displayed data types are accurate and do not need to be changed.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4506 entries, 0 to 4505
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   vehicle_id    4506 non-null   int64 
 1   vehicle_type  4506 non-null   object
 2   make_id       4506 non-null   int64 
 3   model_year    4506 non-null   int64 
 4   vehicle_desc  4506 non-null   object
 5   color         4506 non-null   object
 6   date_stolen   4506 non-null   object
 7   location_id   4506 non-null   int64 
dtypes: int64(4), object(4)
memory usage: 281.8+ KB
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4506 entries, 0 to 4505
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   vehicle_id    4506 non-null   int64         
 1   vehicle_type  4506 non-null   object        
 2   make_id       4506 non-null   int64         
 3   model_year    4506 non-null   int64         
 4   vehicle_desc  4506 non-null   object        
 5   color         4506 non-null   object        
 6   date_stolen   4506 non-null   datetime64[ns]
 7   location_id   4506 non-null   int64         
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 281.8+ KB
```

DATA UNDERSTANDING

However, in the stolen_vehicles dataset, there is a data type error. Specifically, in the date_stolen column, the Dtype is object, which should be datetime. Therefore, it needs to be changed using the Python code below.

```
data['date_stolen'] = pd.to_datetime(data['date_stolen'])
```

DATA UNDERSTANDING

1. Handling Missing

```
data.isna().sum()
```

```
vehicle_id       0
vehicle_type     3
make_id          3
model_year       3
vehicle_desc    20
color            3
date_stolen      3
location_id      3
dtype: int64
```

```
data.isna().sum()
```

```
location_id    0
region         0
country        0
population     0
density        0
dtype: int64
```

```
data.isna().sum()
```

```
make_id      0
make_name    0
make_type    0
dtype: int64
```

DATA UNDERSTANDING

2. Cleaning Data Null

```
columns_to_clean = ['vehicle_id', 'vehicle_type', 'make_id', 'model_year', 'vehicle_desc', 'color', 'date_stolen',
                    'location_id']
data = data.dropna(subset=columns_to_clean)
```

```
data.isna().sum()
```

```
vehicle_id      0
vehicle_type    0
make_id         0
model_year      0
vehicle_desc    0
color           0
date_stolen     0
location_id     0
dtype: int64
```

02

SQL QUERY

Chapter 2

SQL ANALISIS

1. What day of the week are vehicles most often and least often stolen?

Query | Query History

```sql
select to_char (date_stolen, 'Day') as day_of_week,
    count(*) as stolen_count
from stolen_vehicle
group by day_of_week
order by stolen_count desc
limit 7
```

Data Output | Messages | Notifications

	day_of_week text	stolen_count bigint
1	Monday	758
2	Tuesday	700
3	Friday	648
4	Wednesday	623
5	Thursday	612
6	Sunday	591
7	Saturday	574

On Mondays, vehicle theft occurs most frequently, with a total of 758 incidents. On Saturdays, however, vehicle theft is the least frequent compared to other days, with a total of 574 incidents.

SQL ANALISIS

2. What types of vehicles are most often and least often stolen? Does this vary by region?

Query | Query History

```sql
SELECT
    sv.vehicle_type,
    l.region,
    COUNT(*) AS total_stolen
FROM
    stolen_vehicle sv
JOIN
    locations l ON sv.location_id = l.location_id
GROUP BY
    sv.vehicle_type, l.region
HAVING
    COUNT(*) = (
        SELECT
            MAX(stolen_count)
        FROM (
```

	vehicle_type character varying (100)	region character varying (100)	total_stolen bigint
1	Other Truck	Hawke's Bay	1
2	Cab and Chassis Only	Wellington	1
3	Light Bus	Northland	1
4	Boat Trailer	Hawke's Bay	1

	vehicle_type character varying (100)	region character varying (100)	total_stolen bigint
1	Saloon	Auckland	326
2	Stationwagon	Auckland	306
3	Hatchback	Auckland	296
4	Stationwagon	Canterbury	164

The Saloon vehicle type is the most frequently stolen, with a total of 326 incidents, while vehicles such as Other Trucks, Light Buses, etc., are the least frequently stolen, with only 1 incident.

SQL ANALISIS

3. What is the average age of the vehicles that are stolen? Does this vary based on the vehicle type?

Query | Query History

```sql
SELECT
    vehicle_type,
    AVG(EXTRACT(YEAR FROM CURRENT_DATE)::INTEGER - CAST(model_year AS INTEGER)) AS average_umur
FROM
    stolen_vehicle
GROUP BY
    vehicle_type
order by average_umur desc
limit 5
```

	vehicle_type character varying (100)	average_umur numeric
1	Special Purpose Vehicle	67.0000000000000000
2	Mobile Home - Light	37.2000000000000000
3	Flat Deck Truck	30.4705882352941176
4	Caravan	28.3250000000000000
5	Other Truck	25.5476190476190476

	vehicle_type character varying (100)	average_umur numeric
1	Mobile Machine	7.0000000000000000
2	Tractor	9.5000000000000000
3	Moped	9.9144385026737968
4	Roadbike	11.8619528619528620
5	Cab and Chassis Only	12.6250000000000000

The average age of stolen vehicles is around 20 years old. The oldest vehicle model stolen is 67 years old, while the youngest model is 7 years old.

SQL ANALISIS

4. Which regions have the most and least number of stolen vehicles? What are the characteristics of the regions?

Query | Query History

```sql
SELECT
    l.region, l.population, l.density,
    COUNT(*) AS total_stolen_vehicles
FROM
    stolen_vehicle sv
JOIN
    locations l ON sv.location_id = l.location_id
GROUP BY
     l.region, l.population, l.density
ORDER BY
    total_stolen_vehicles ASC
--  total_stolen_vehicles desc
LIMIT 1;
```

	region character varying (100)	population integer	density numeric (10,2)	total_stolen_vehicles bigint
1	Auckland	1695200	343.09	1620

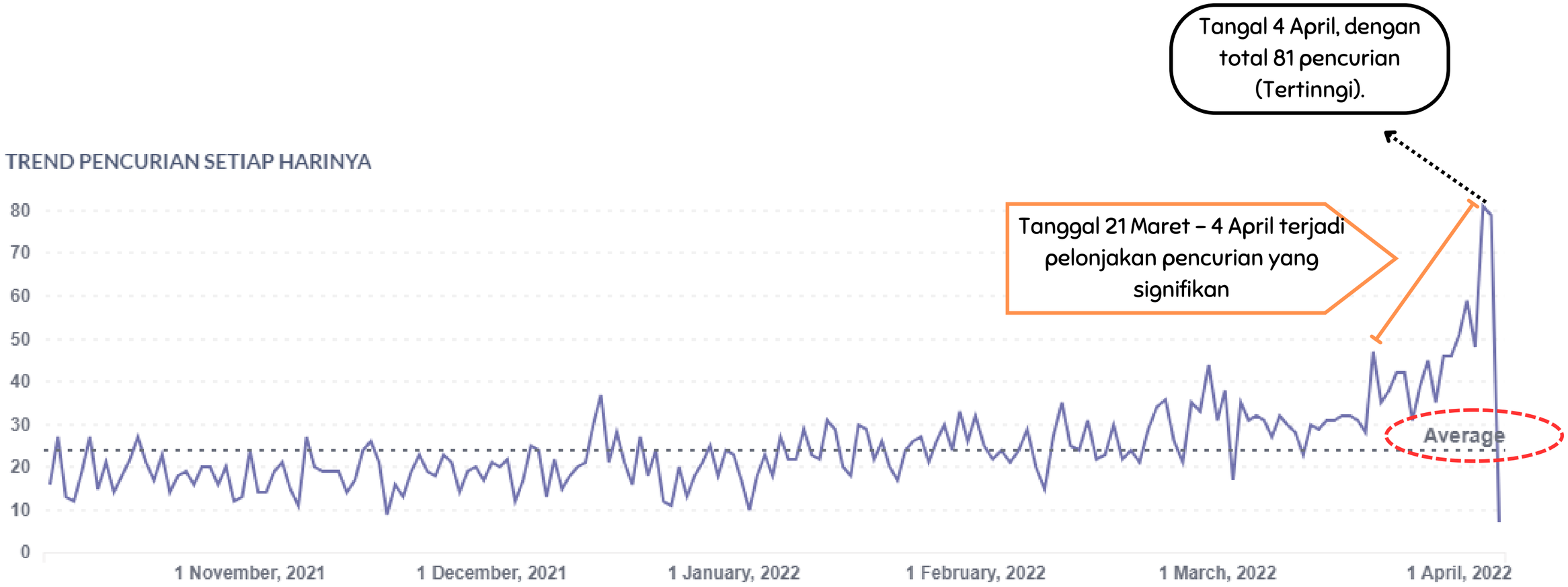
	region character varying (100)	population integer	density numeric (10,2)	total_stolen_vehicles bigint
1	Southland	102400	3.28	26

The region with the highest number of vehicle thefts is Auckland with a total of 1620, while the least is Southland with 26 stolen vehicles.

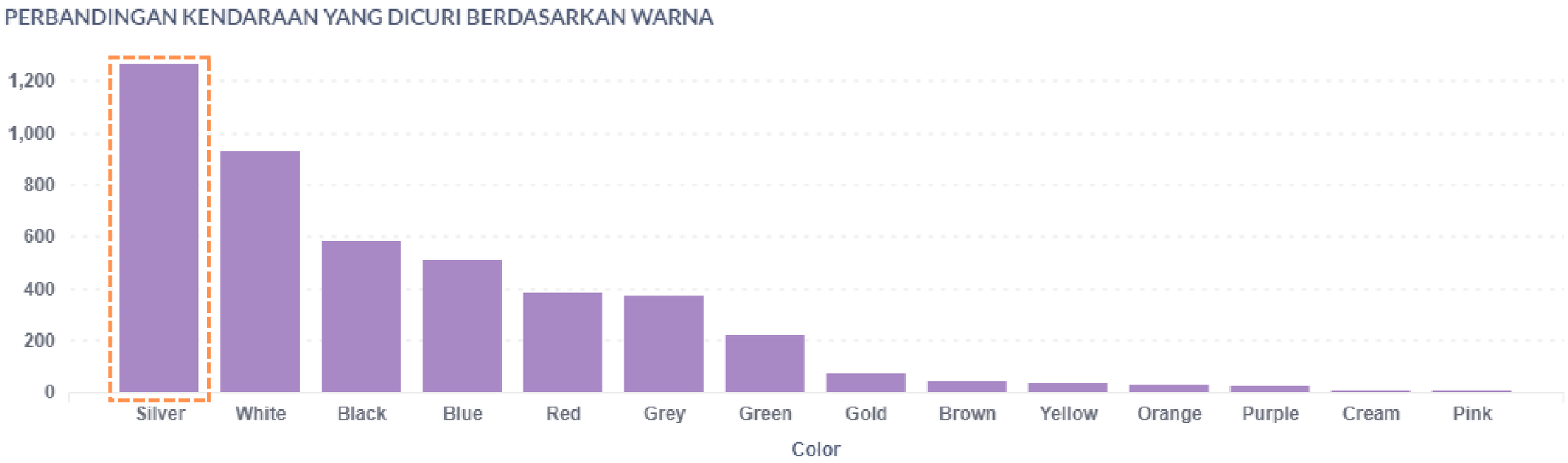
03

DATA EXPLORATION

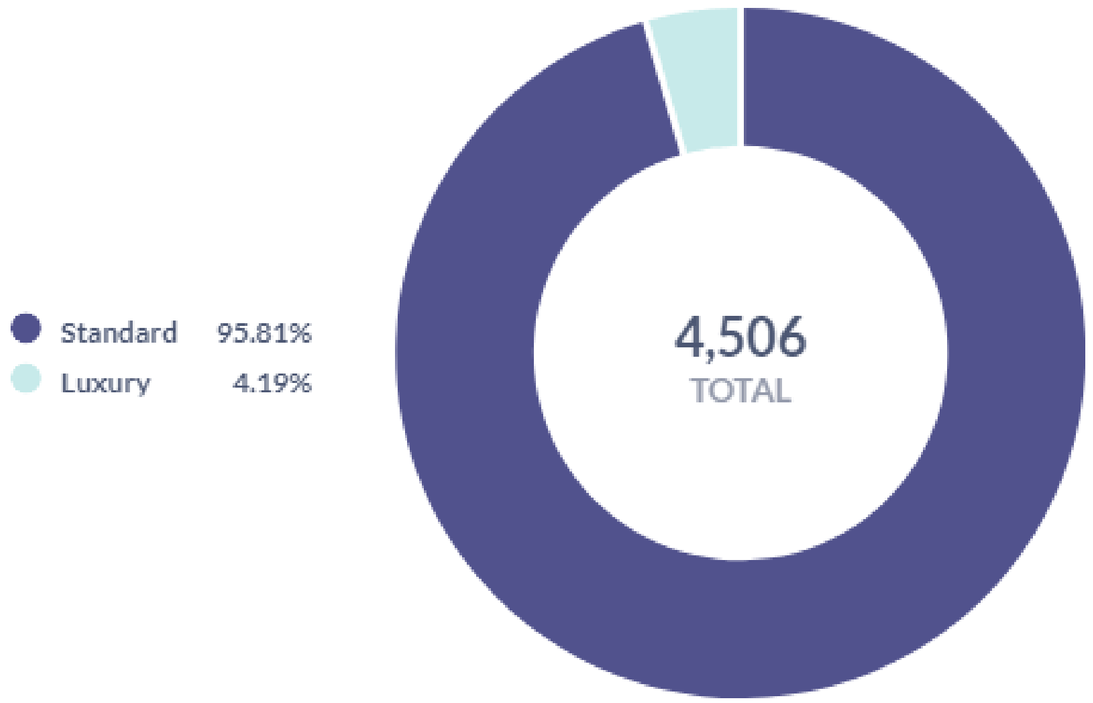
Chapter 3



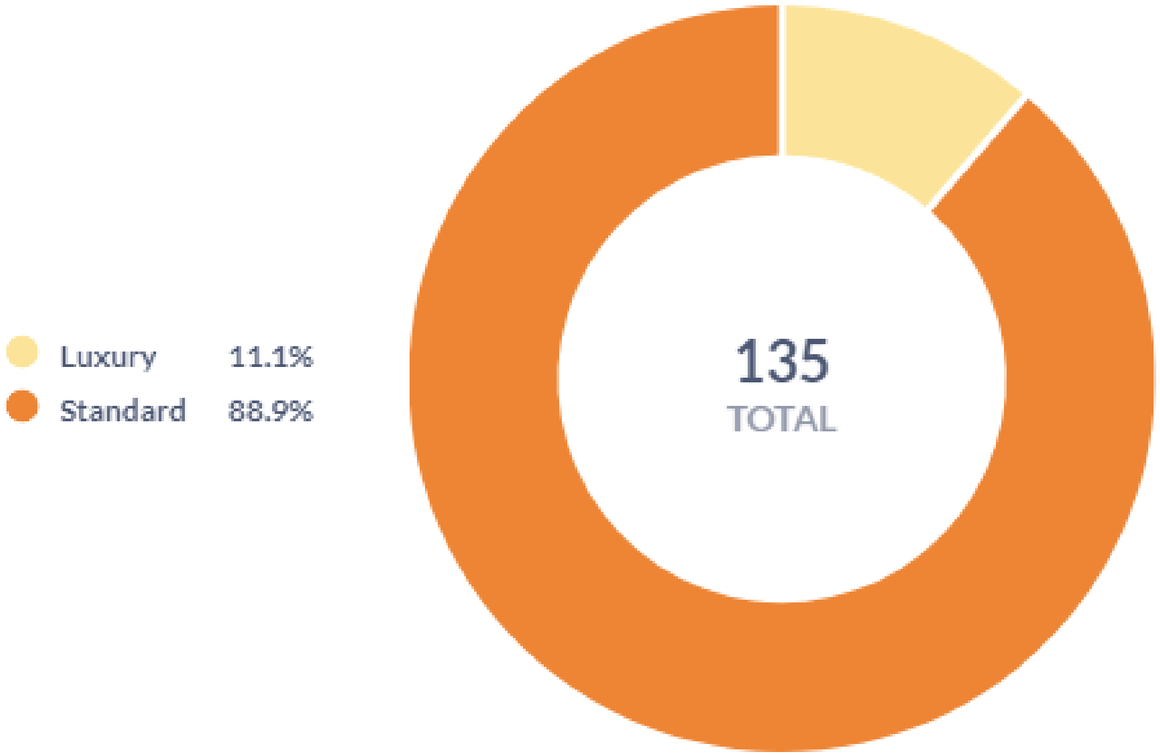
On April 4, 2022, there was a massive theft spree with a total of 81 vehicle thefts. Conversely, on April 6, 2022, there were only 7 reported vehicle thefts, marking the lowest count. On average, there are 24 vehicle thefts per day.



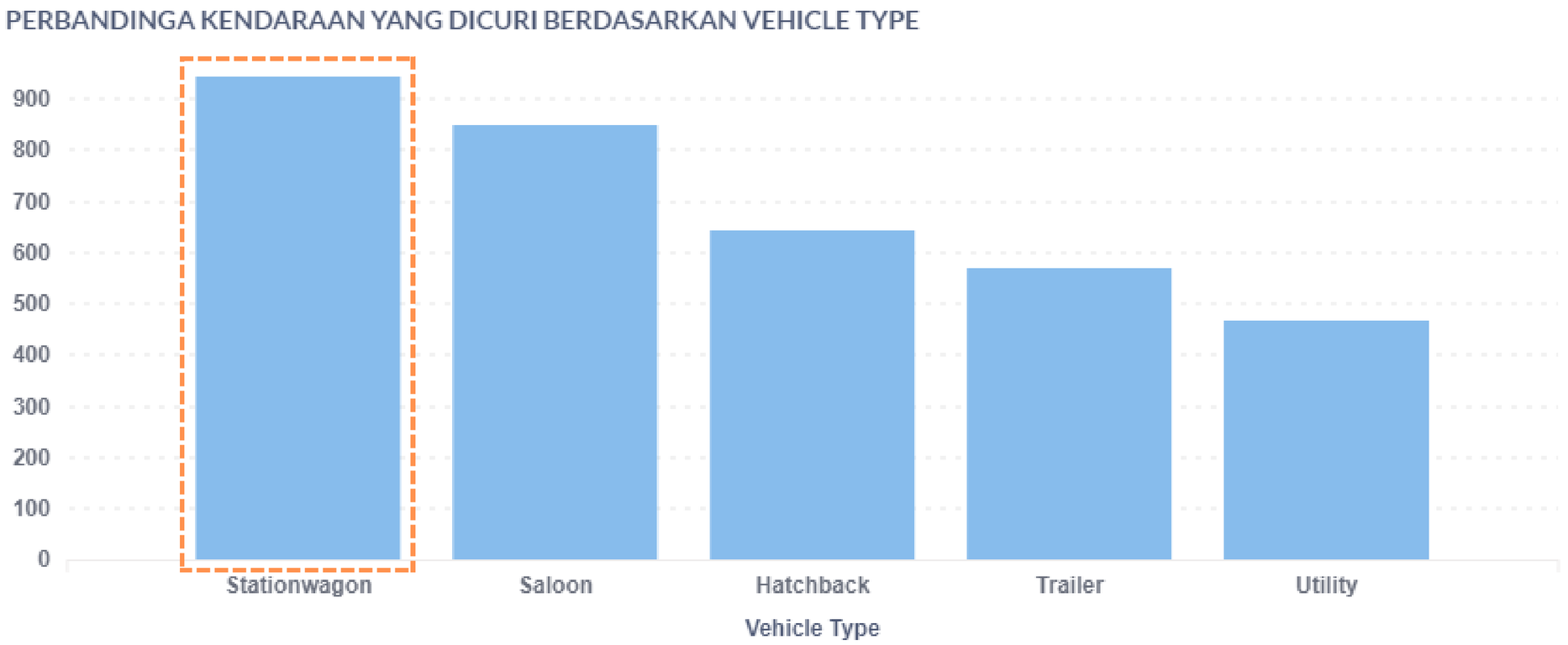
Vehicles in Silver are the most commonly stolen color, with a total of 1267 reported thefts, followed by White vehicles with 931 thefts. On the other hand, Pink vehicles are the least commonly stolen, with only 4 reported thefts.



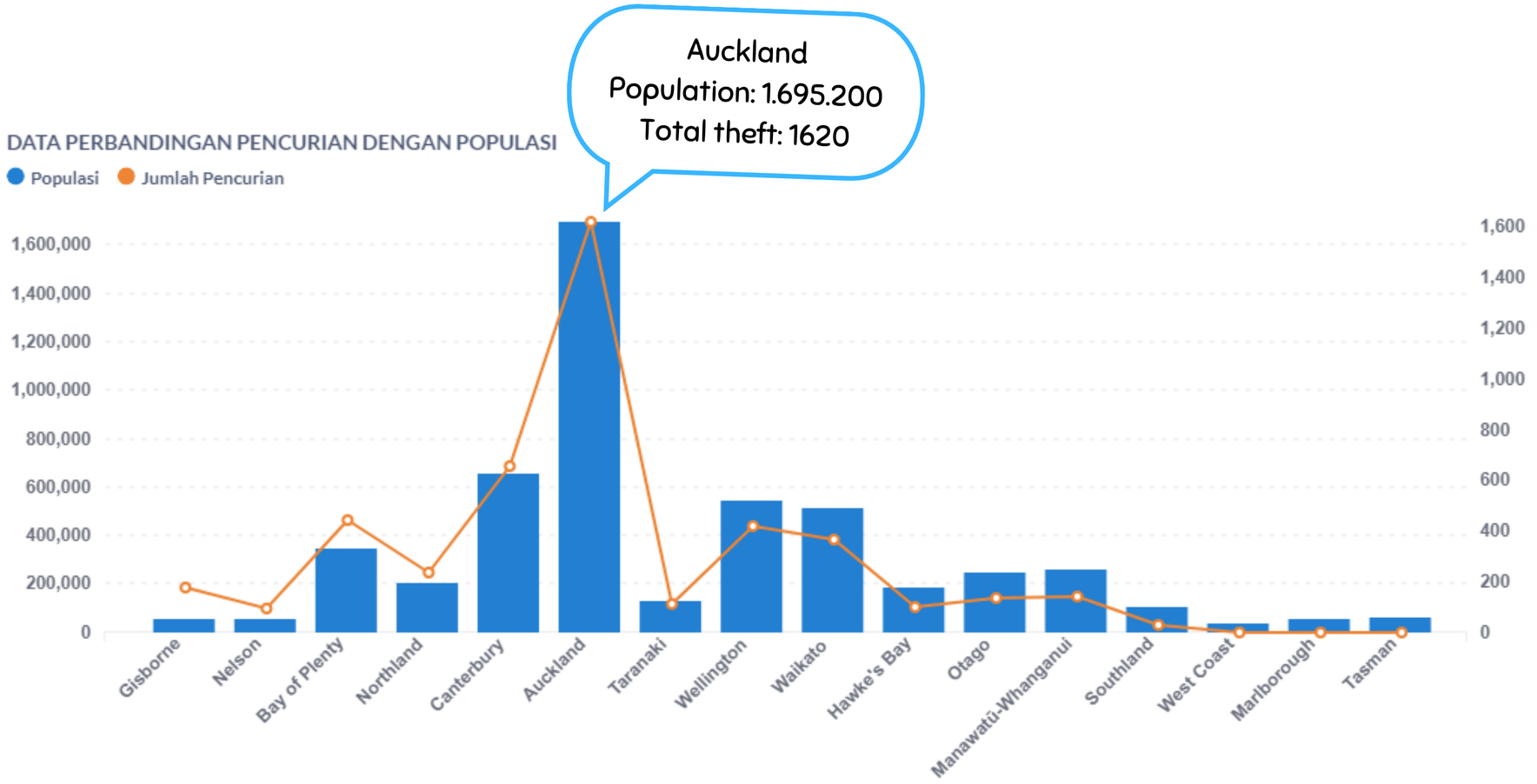
Out of a total of 4,506 stolen vehicles, Standard type vehicles are the most frequently stolen, with 4,317 vehicles reported stolen, while Luxury type vehicles make up 189 of the thefts.



Out of a total of 135 vehicle brands, the Standard type vehicles dominate with 120 brands, while 15 brands are associated with the Luxury type.



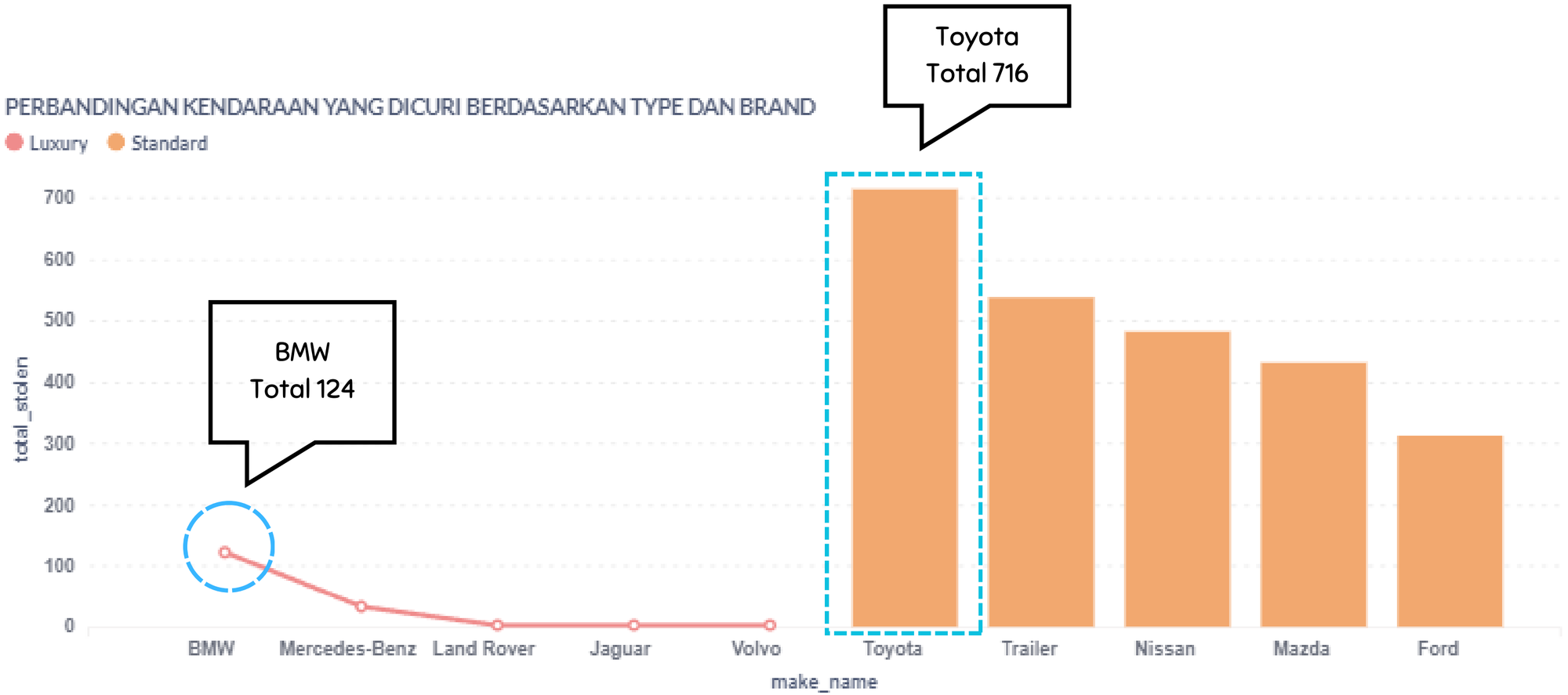
The Stationwagon vehicle type is the most frequently stolen, with a total of 944 incidents, followed by the Saloon type with 850 vehicles. Conversely, the Articulated Truck and Special Purpose Vehicle types are the least stolen, each with only 1 vehicle.



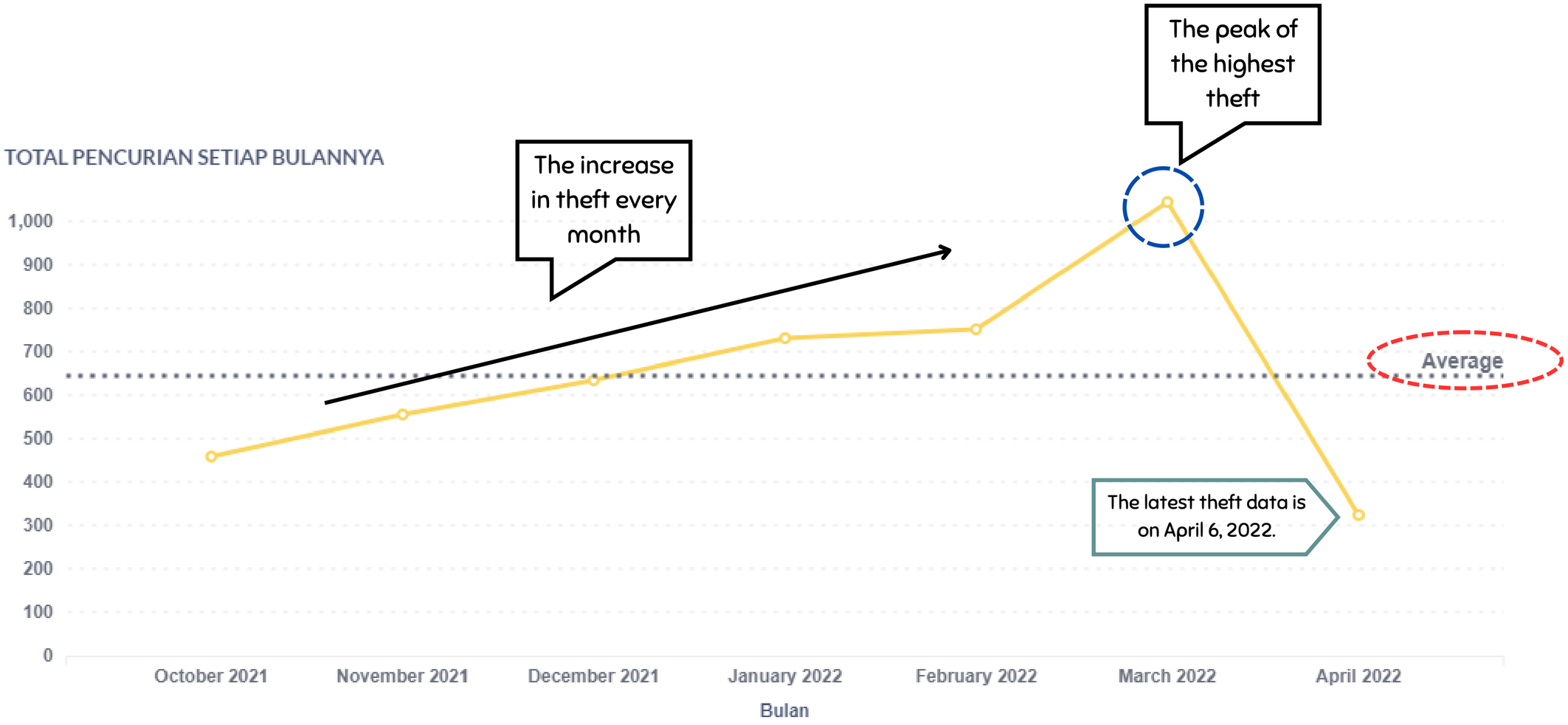
0.98

KORELASI ANTARA POPULASI DAN TOTAL PENCURIAN

The Auckland region has the highest population with a total of 1,695,200, and it also has the highest vehicle thefts with 1620 incidents. Based on the obtained data, the population and total thefts have a high correlation of 0.98, indicating that as the population increases in each region, vehicle thefts also tend to increase.



BMW is the luxury vehicle type with the highest number of thefts, totaling 124 incidents. Meanwhile, Toyota is the standard vehicle type that is most frequently stolen, with a total of 716 thefts.

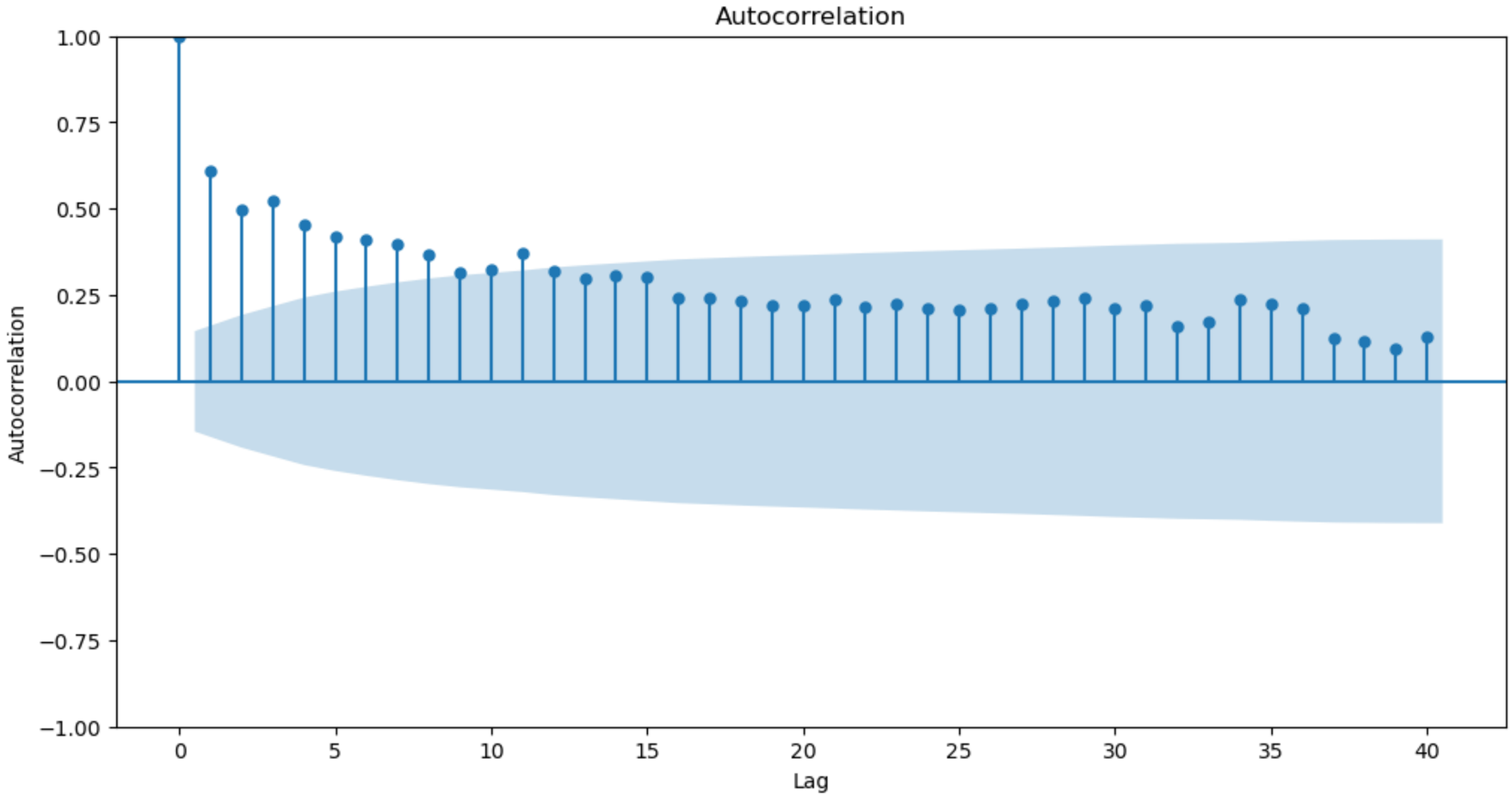


Insight: Vehicle thefts consistently increase every month.

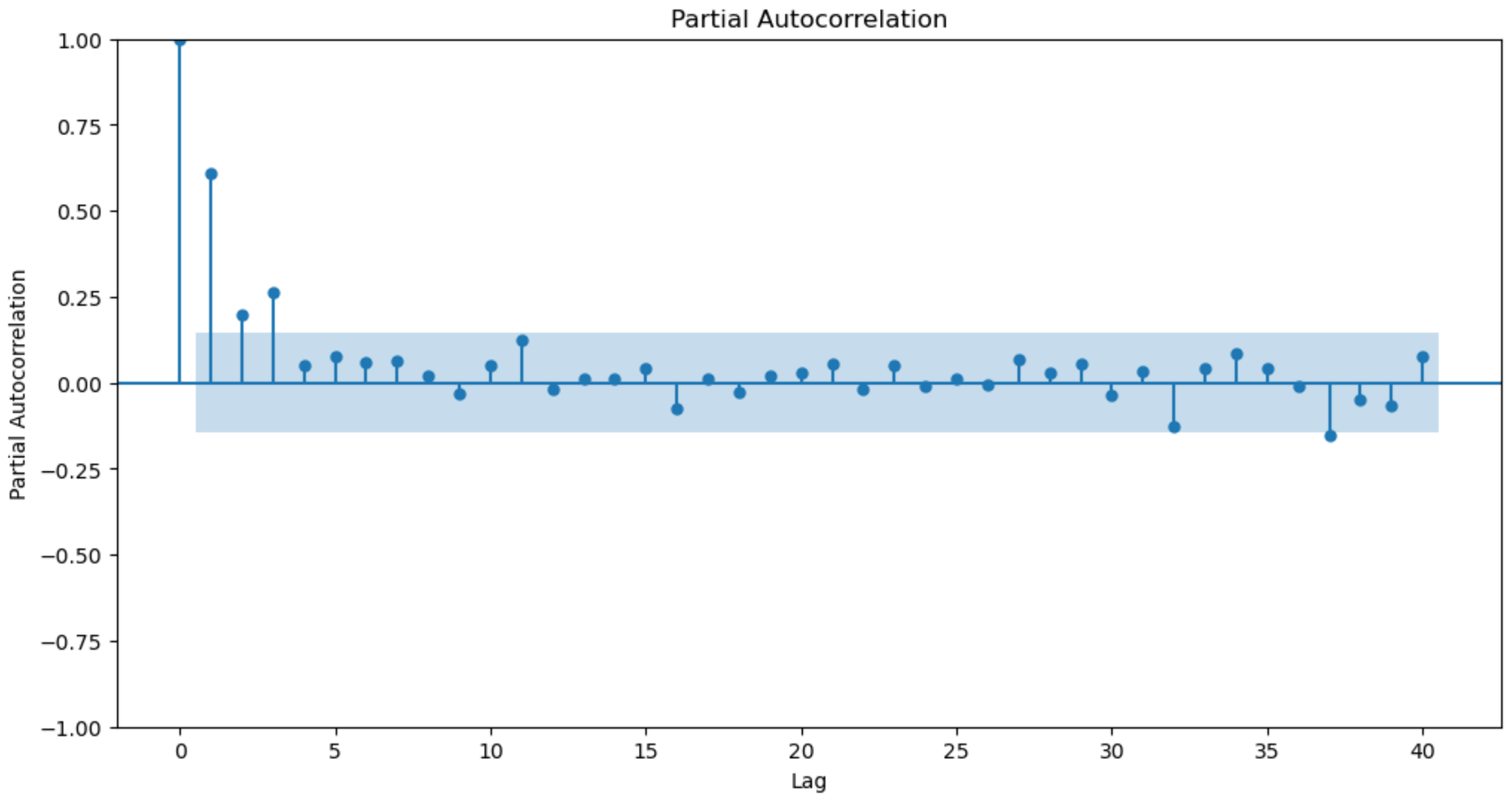
04

FORECASTING AND DASHBOARD

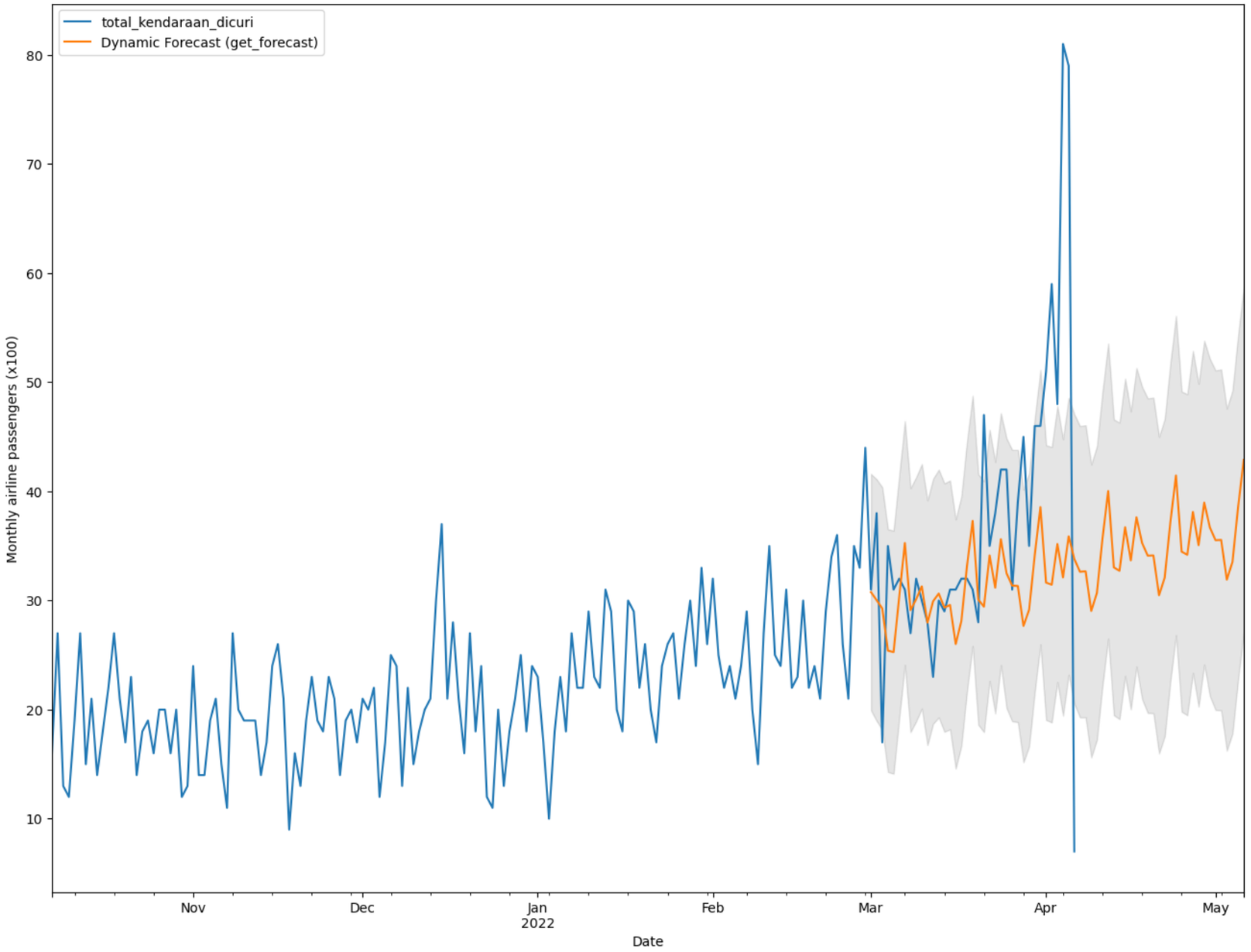
Chapter 4



From the Autocorrelation plot below, it can be concluded that the data has low correlation, as it only extends up to lag 13 without touching the blue shaded area.



Based on the Partial Autocorrelation plot below, it can be concluded that the data has low correlation, as it only extends up to lag 5 without touching the blue shaded area.



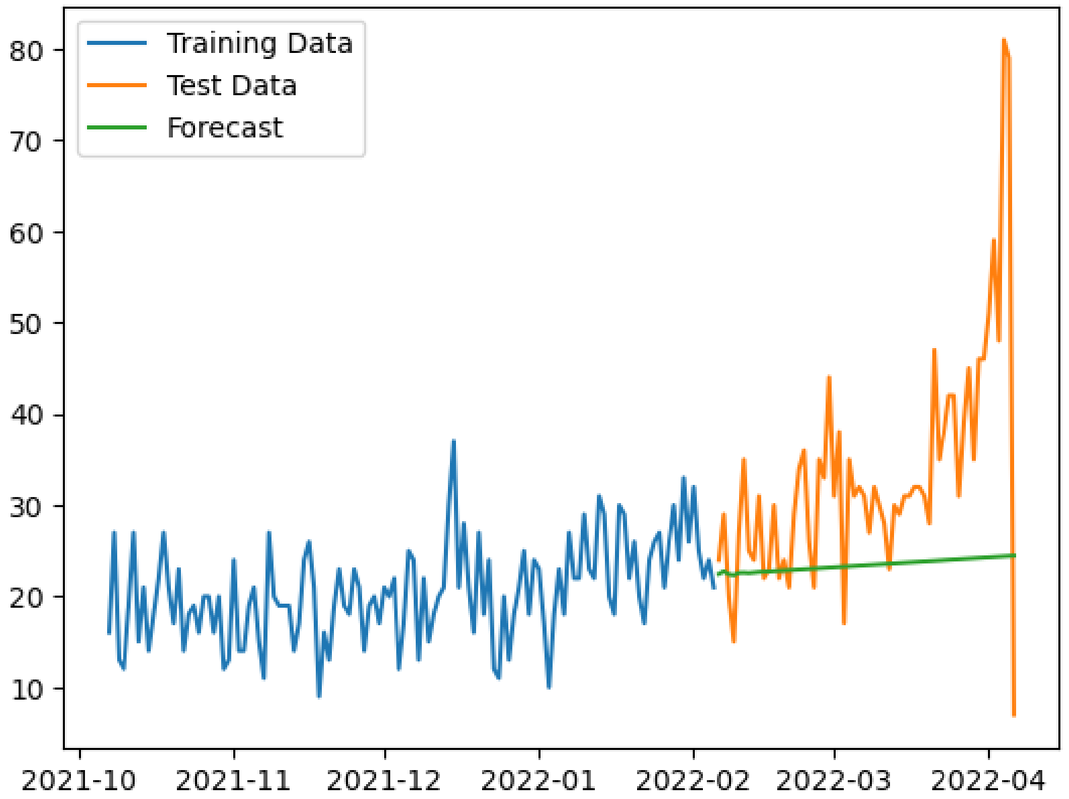
Garis Actual

Garid Prediksi

Based on the prediction plot on the side, the yellow line represents the forecast of thefts in the future. The trend appears to increase slightly each month, although not significantly.

```python
# Train-test split
train = data.iloc[:-60]  # Ambil data historis sebagai data latih
test = data.iloc[-60:]   # Ambil 30 hari terakhir sebagai data uji


model = ARIMA(train, order=(3,2,1))  # Contoh order ARIMA (p,d,q)
model_fit = model.fit()
# Forecast
forecast = model_fit.forecast(steps=60)  # Prediksi 30 hari ke depan
# Visualisasi prediksi
plt.plot(train.index, train, label='Training Data')
plt.plot(test.index, test, label='Test Data')
plt.plot(test.index, forecast, label='Forecast')
plt.legend()
plt.show()
```



CONCLUSION

1. The implementation of the ARIMA model for predicting vehicle thefts seems to be effective.
2. There's an increase in thefts in the future, although not significantly.
3. The data exhibits a low correlation, as indicated by the lag autocorrelation at lag 13 and the Partial Autocorrelation at lag 5.

POWER BI DASHBOARD

Dashboard Motor Vehicle Thefts

Dataset berasal dari Departemen Polisi New Zealand yang berisikan data pencurian kendaraan selama 6 bulan

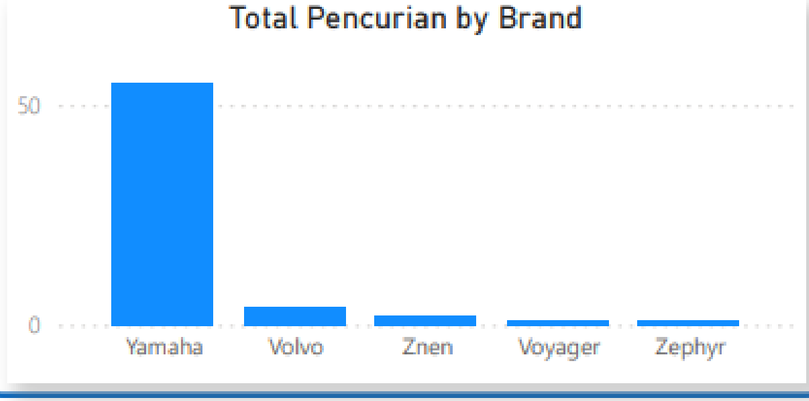
Total Penduduk	Jumlah Region	Brand Kendaraan
5M	16	138

Jenis Kendaraan

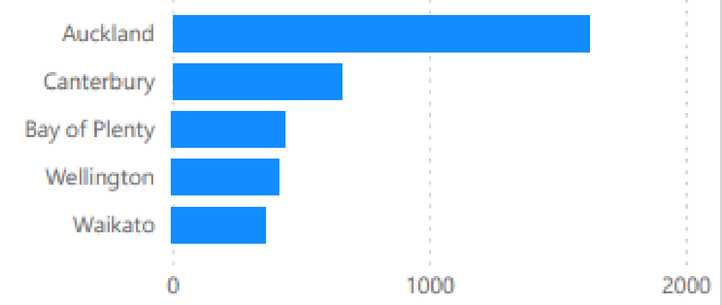
Luxury | Standard

Trend Pencurian Kendaraan

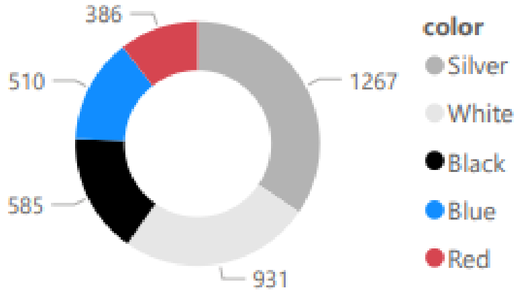
Total Pencurian by Brand



Total Pencurian by Region



Total Pencurian by Color



METABASE DASHBOARD

Dashboard Motor Vehicle Thefts

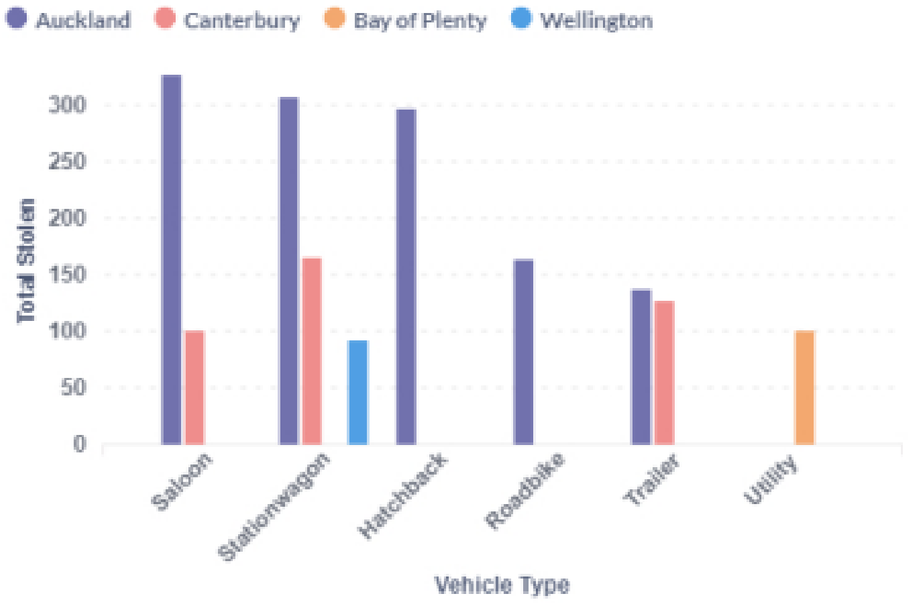
5,123,200
Jumlah Penduduk

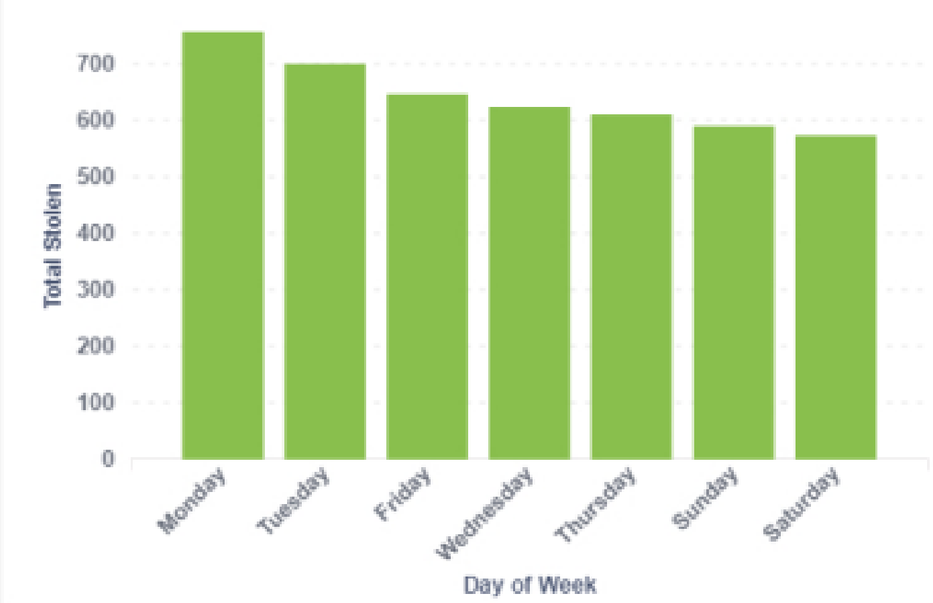
16
Jumlah Region

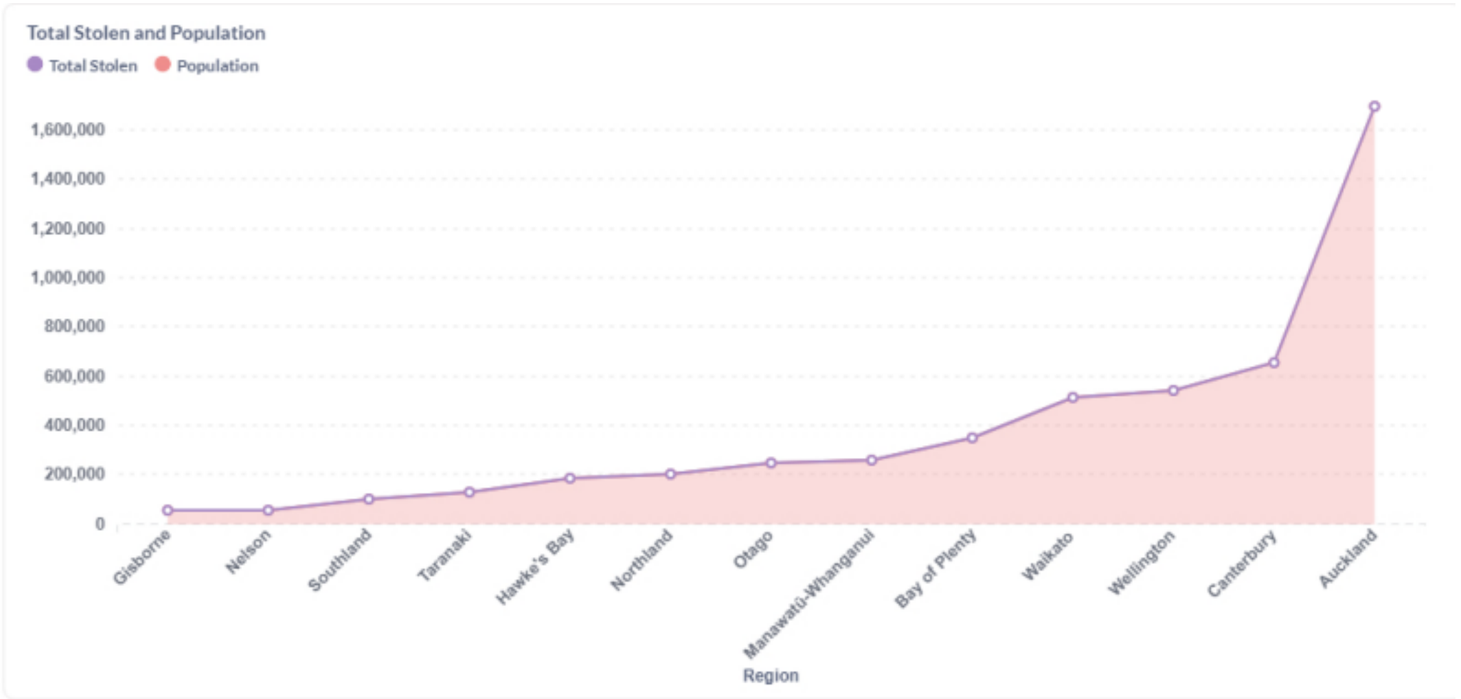
4,506
Total Stolen

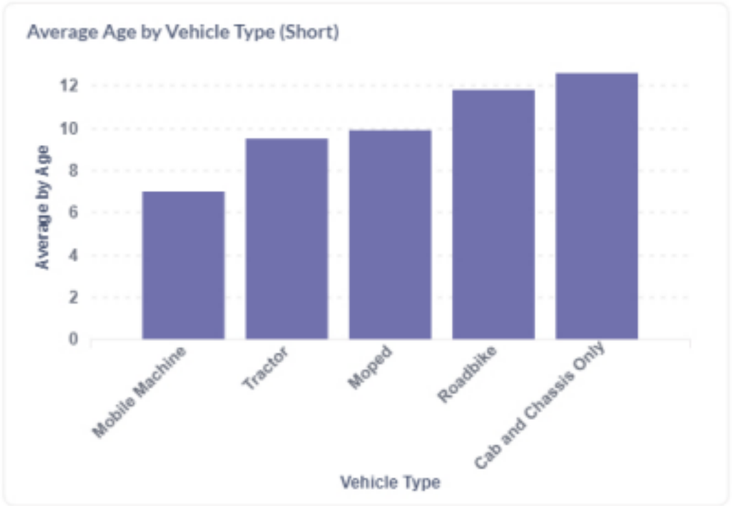
138
Total Brand Vehicle

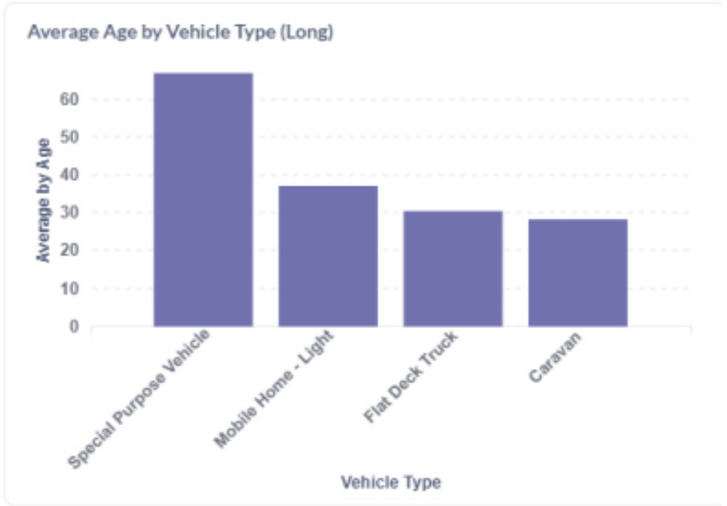
98%
Correlation by Population and Total Stolen






Total Stolen and Population
Total Stolen
Population
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Gisborne
Nelson
Southland
Taranaki
Hawke's Bay
Northland
Otago
Manawatū-Whanganui
Bay of Plenty
Waikato
Wellington
Canterbury
Auckland
Region


Average Age by Vehicle Type (Short)
12
10
8
6
4
2
0
Average by Age
Mobile Machine
Tractor
Moped
Roadbike
Cab and Chassis Only
Vehicle Type


Average Age by Vehicle Type (Long)
60
50
40
30
20
10
0
Average by Age
Special Purpose Vehicle
Mobile Home - Light
Flat Deck Truck
Caravan
Vehicle Type

Thank You

By Diky Arianto Tarihoran