#Remove Duplicate Records

CREATE TABLE StolenVehicles\_Dataset.cleaned\_stolenvehicles as

SELECT

DISTINCT VehicleDesc,

VehicleType,

VehicleModel,

ModelYear,

color,

DateStolen,

Location

FROM StolenVehicles\_Dataset.stolenvehicles

## Remove null values

Select \*

From StolenVehicles\_Dataset.cleaned\_stolenvehicles

Where VehicleType is not null and VehicleDesc is not null

## USE a temporary table to store cleaned (non-duplicate & non null values)

WITH Cleaned\_Stolen\_VehicleDB as (

    SELECT  \*

    From StolenVehicles\_Dataset.cleaned\_stolenvehicles

    WHERE VehicleType is not null and VehicleDesc is not null

    )

## Check Vehicle Model for typos or categorizing mistakes

SELECT

    VehicleModel,

    COUNT(VehicleModel) as VModel\_count

FROM Cleaned\_Stolen\_VehicleDB

GROUP BY VehicleModel

HAVING VModel\_count <=2

/\*

Based on results of above query, we now have 73 rows of vehicle models which appeared up to two times in this 4+K dataset.

Further investigation shall take place to ensure there are no typos or miscategorization

Following are the typos and miscategorization found:

\*/

## Change Landrover to Land Rover

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Land Rover' WHERE vehiclemodel='Landrover'

## Change Rover to Land Rover

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Land Rover' WHERE vehiclemodel='Rover'

## Change Mitsubishio Fuso to Fuso

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Fuso' WHERE vehiclemodel='Mitsubishio Fuso'

## Change Nissan Diesel to Nissan

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Nissan' WHERE vehiclemodel='Nissan Diesel'

## Change Toyota Lexus to Lexus

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Lexus' WHERE vehiclemodel='Toyota Lexus'

## Change Mobile Machine to Takeuchi

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET vehiclemodel='Takeuchi' WHERE vehiclemodel='Mobile Machine'

## Update FERGUSSON to Ferguson

UPDATE StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET VehicleDesc='Ferguson' WHERE VehicleDesc='FERGUSSON'

## USE a temporary table to store cleaned (non-duplicate & non null values)

with Cleaned\_Stolen\_VehicleDB as (

    SELECT  \*

    From StolenVehicles\_Dataset.cleaned\_stolenvehicles

    WHERE VehicleType is not null and VehicleDesc is not null

    )

## Check Model Year Ranges

Select MIN(ModelYear) as Oldest\_model, MAX(ModelYear) as Newest\_model

From Cleaned\_Stolen\_VehicleDB

Graphical user interface, application

Description automatically generated with medium confidence

## USE a temporary table to store cleaned (non-duplicate & non null values)

WITH Cleaned\_Stolen\_VehicleDB as (

    SELECT  \*

    From StolenVehicles\_Dataset.cleaned\_stolenvehicles

    WHERE VehicleType is not null and VehicleDesc is not null

    )

## Check report date range (6 months)

SELECT MIN(DateStolen) as Report\_Start\_Date, MAX(DateStolen) as Report\_End\_Date

FROM Cleaned\_Stolen\_VehicleDB

Graphical user interface, application

Description automatically generated

## USE a temporary table to store cleaned (non-duplicate & non null values)

WITH Cleaned\_Stolen\_VehicleDB as (

    SELECT  \*

    From StolenVehicles\_Dataset.cleaned\_stolenvehicles

    WHERE VehicleType is not null and VehicleDesc is not null

    )

## Compare Model Year to Year Stolen

SELECT distinct stolen\_date,

VehicleDesc,

VehicleType,

VehicleModel,

ModelYear,

color,

Location

From Cleaned\_Stolen\_VehicleDB,

    (SELECT

    distinct extract(year FROM DateStolen) as stolen\_date

    FROM  StolenVehicles\_Dataset.cleaned\_stolenvehicles) as year\_of\_vehicle\_theft

WHERE ModelYear>stolen\_date

/\*

Above query is checking if the theft date is more recent than the car model year.

This is to validate that data entered in this dataset is reliable and valid.

The results of this query included 19 theft records where 2022 car models were stolen in 2021.

This is fine because newer car models are typically intoduced for sale up to 1 year in advance.

\*/

Graphical user interface, application

Description automatically generated

## USE a temporary table to store cleaned (non-duplicate & non null values)

WITH Cleaned\_Stolen\_VehicleDB as (

    SELECT  \*

    From StolenVehicles\_Dataset.cleaned\_stolenvehicles

    WHERE VehicleType is not null and VehicleDesc is not null

    )

## Check location names

SELECT DISTINCT location,

FROM Cleaned\_Stolen\_VehicleDB

/\*

Decided to replace forward slash with an underscore for easy readability.

\*/

## Update Counties/Manukau to Counties\_Manukau

UPDATE  StolenVehicles\_Dataset.cleaned\_stolenvehicles

SET location='Counties\_Manukau' where location='Counties/Manukau'