Table of Contents

[0.1 Task 1 - Standardising raw data sets and data modelling. 2](#_Toc115781041)

[**Source Data Discovery** 2](#_Toc115781042)

[0.2 Dataset : NSW Property Median Value 5](#_Toc115781043)

[0.3 Dataset : NSW\_Stationentrances\_sydney 6](#_Toc115781044)

[0.4 Dimensional Model 7](#_Toc115781045)

[Diagram 8](#_Toc115781046)

[Table Definitions 8](#_Toc115781047)

[Database creation 8](#_Toc115781048)

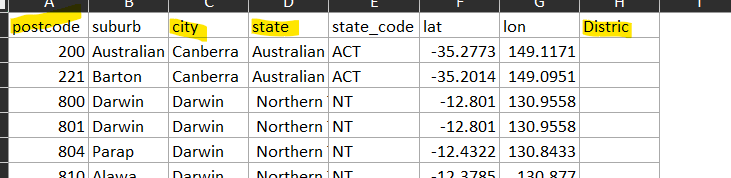
[Create Table Definitions 10](#_Toc115781049)

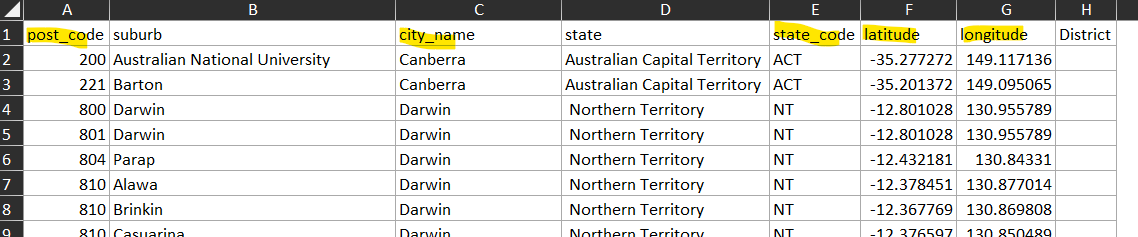
# 0.1 Task 1 - Standardising raw data sets and data modelling.

## **Source Data Discovery**

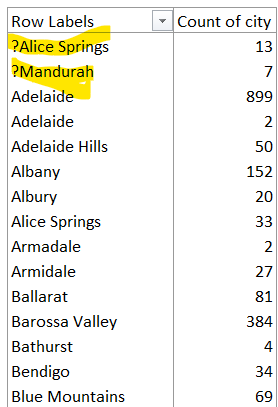
**Dataset : Aus\_SubcityDistrict\_State Data**

(a) Observation : Column Names are not Standardized

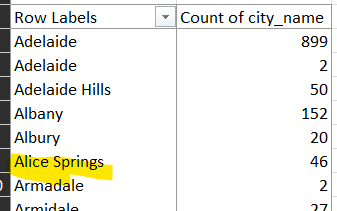


Snippet after Standardization

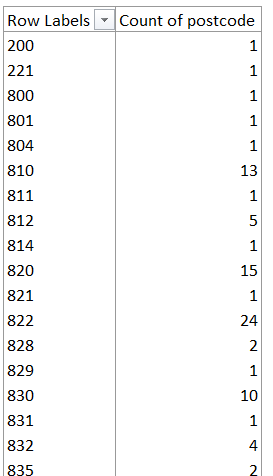
(b) Observation : City Names are having junk values



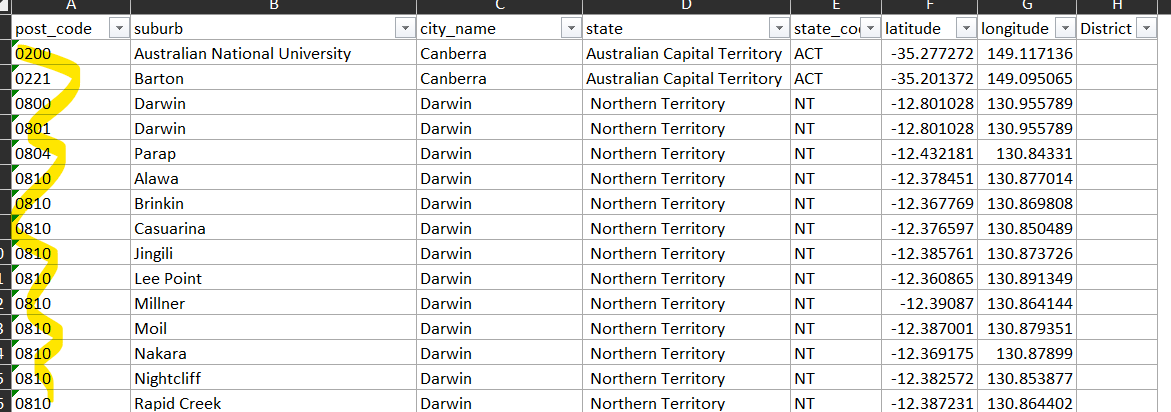
Snippet after standardization



(C) Observation Postcode are having only 3 characters in length

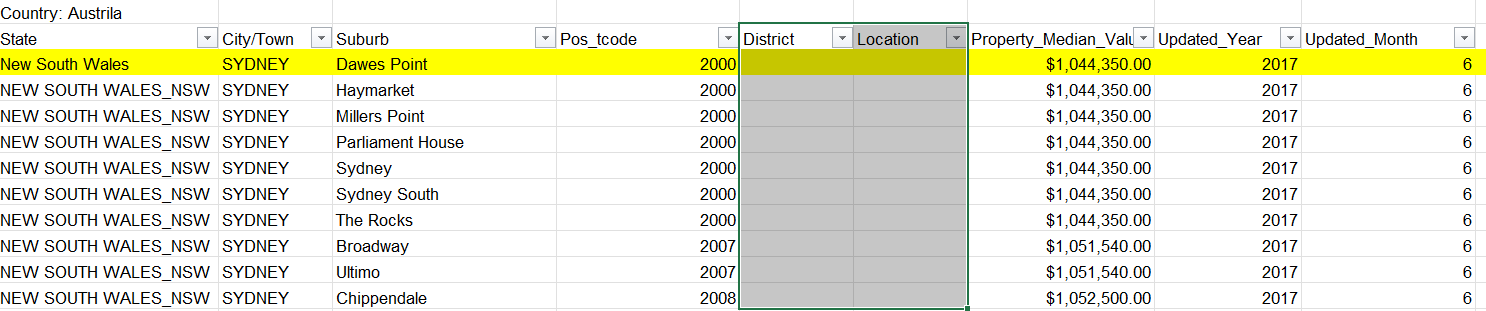


(d) Post Code is standardized to 4 digits using excel formula

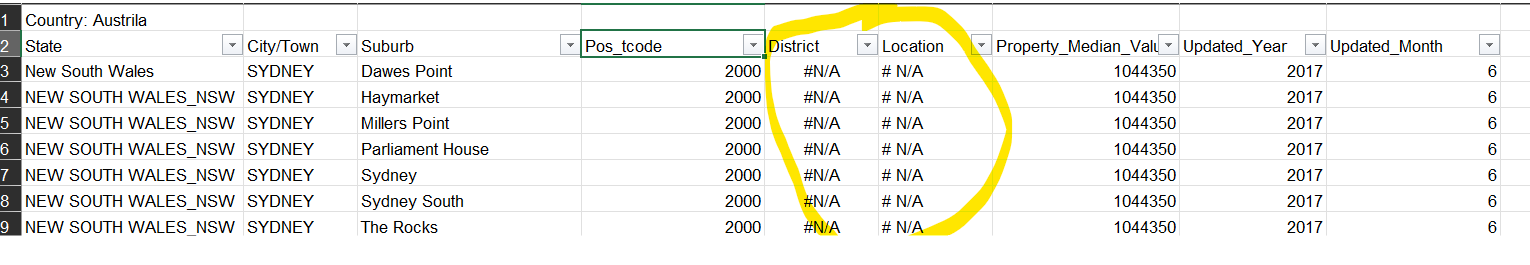


# 0.2 Dataset : NSW Property Median Value

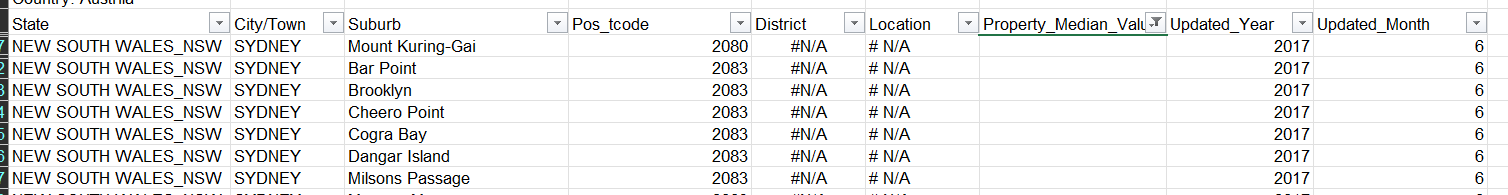
(a) District and location column has Null/Blank (Missing) data.



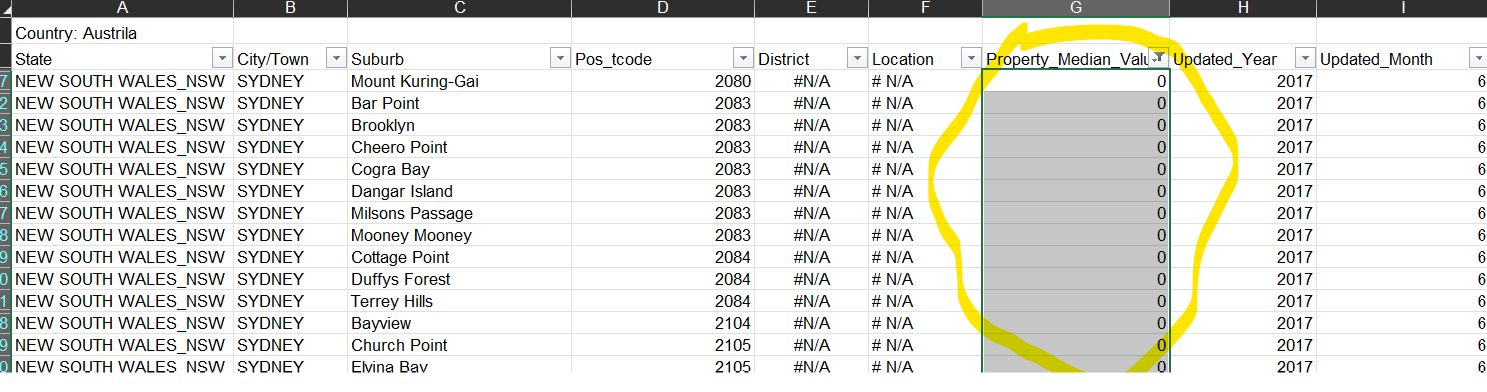
(b) Imputing the missing Data with #N/A.



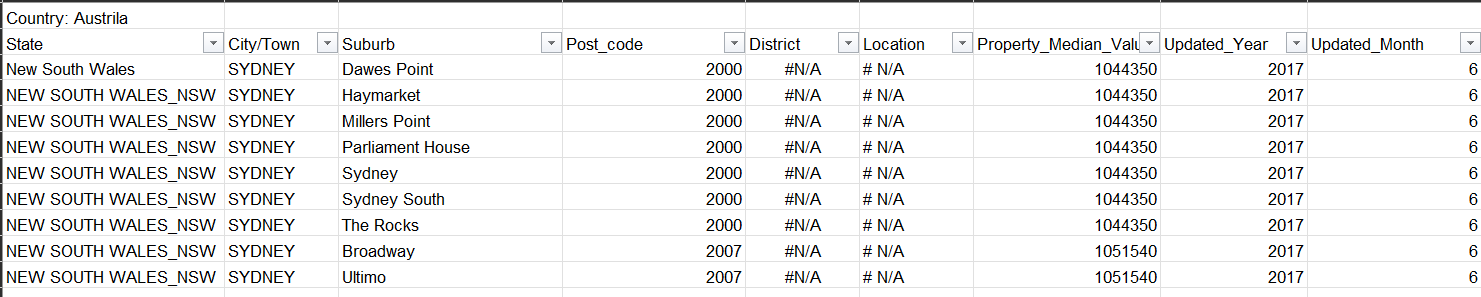
(c) Property Median Values are missing or Nulls



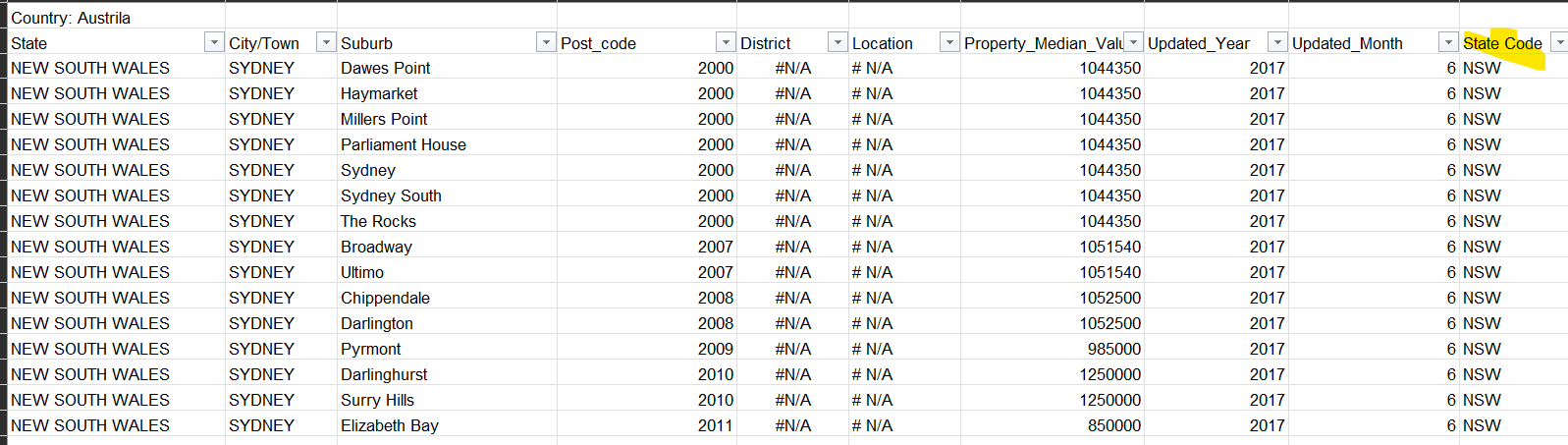
(d) Imputing the missing values to Zero



(e) .State Column values not consistent. State Code and State merged into a single column

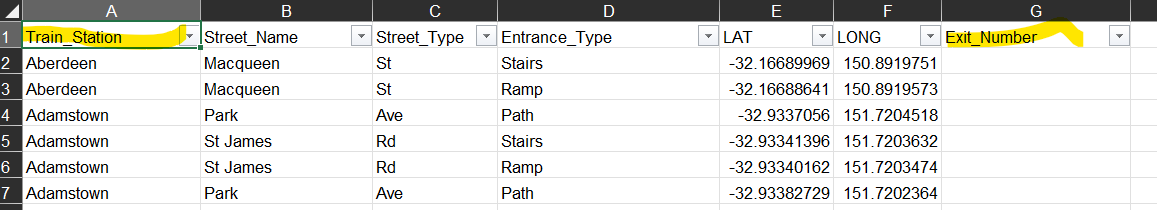


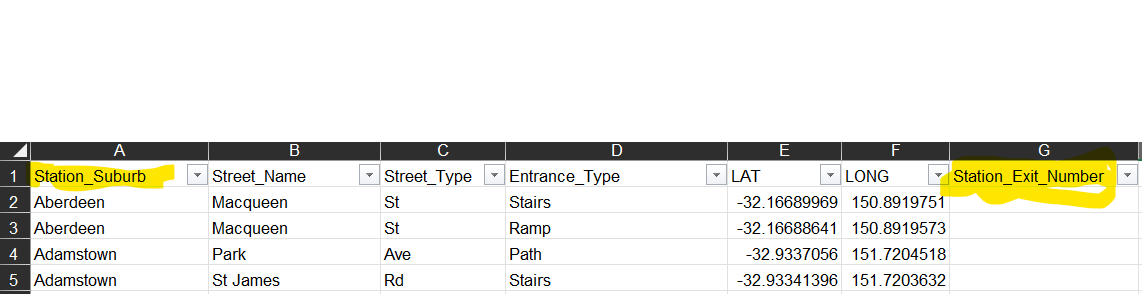
(f) State Code and State Name is split into 2 different columns and data values are standardized



# 0.3 Dataset : NSW\_Stationentrances\_sydney

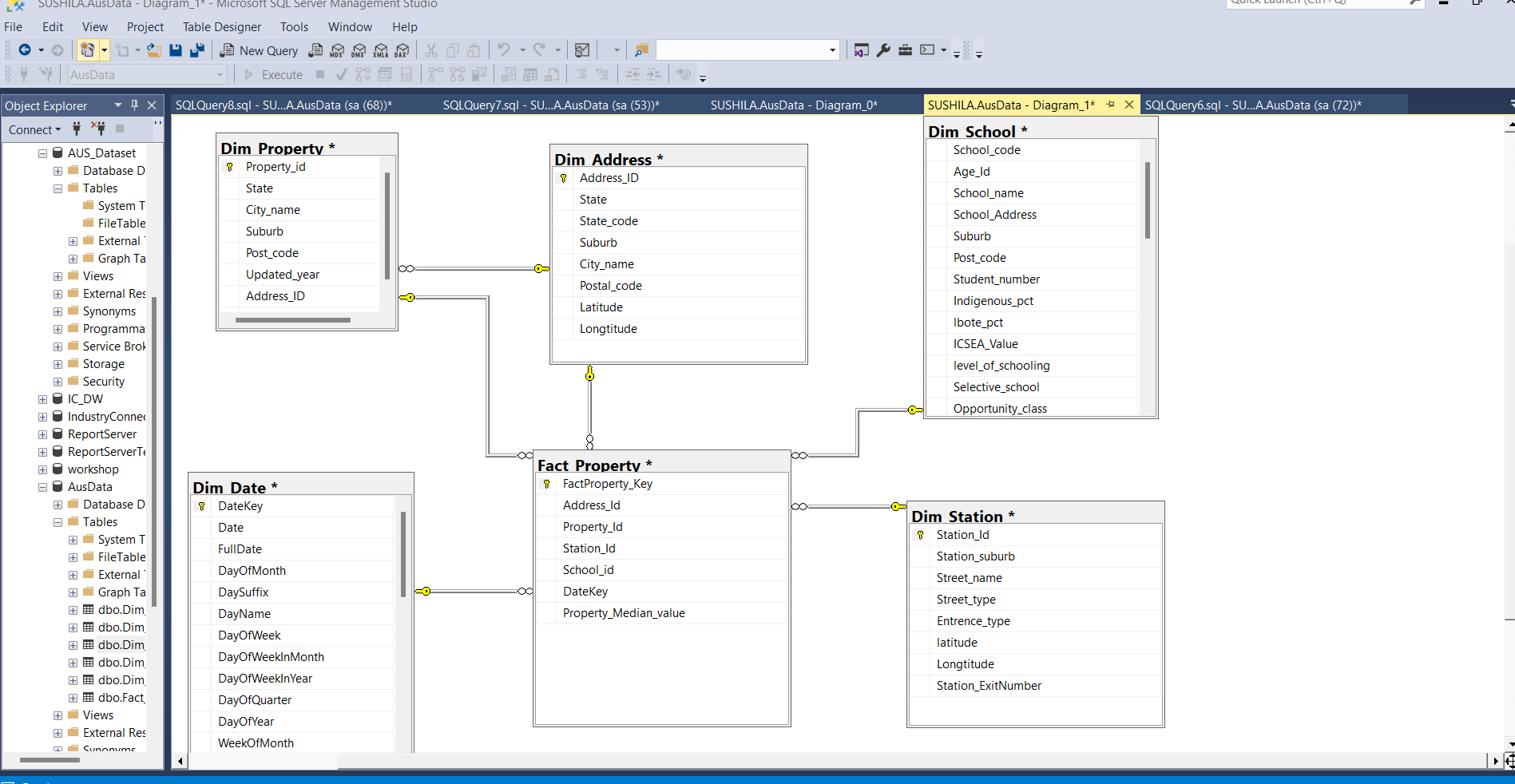
**(a) Column Names are not in a standard form for better understanding.**



(b) Given more meaningful names for the column.

# 0.4 Dimensional Model

## Diagram



## Table Definitions

## Database creation

USE [master]

GO

/\*\*\*\*\*\* Object: Database [AusData] Script Date: 4/10/2022 12:54:41 PM \*\*\*\*\*\*/

CREATE DATABASE [AusData]

CONTAINMENT = NONE

ON PRIMARY

( NAME = N'AusData', FILENAME = N'C:\Program Files\Microsoft SQL Server\MSSQL15.MSSQLSERVER\MSSQL\DATA\AusData.mdf' , SIZE = 8192KB , MAXSIZE = UNLIMITED, FILEGROWTH = 65536KB )

LOG ON

( NAME = N'AusData\_log', FILENAME = N'C:\Program Files\Microsoft SQL Server\MSSQL15.MSSQLSERVER\MSSQL\DATA\AusData\_log.ldf' , SIZE = 8192KB , MAXSIZE = 2048GB , FILEGROWTH = 65536KB )

WITH CATALOG\_COLLATION = DATABASE\_DEFAULT

GO

IF (1 = FULLTEXTSERVICEPROPERTY('IsFullTextInstalled'))

begin

EXEC [AusData].[dbo].[sp\_fulltext\_database] @action = 'enable'

end

GO

ALTER DATABASE [AusData] SET ANSI\_NULL\_DEFAULT OFF

GO

ALTER DATABASE [AusData] SET ANSI\_NULLS OFF

GO

ALTER DATABASE [AusData] SET ANSI\_PADDING OFF

GO

ALTER DATABASE [AusData] SET ANSI\_WARNINGS OFF

GO

ALTER DATABASE [AusData] SET ARITHABORT OFF

GO

ALTER DATABASE [AusData] SET AUTO\_CLOSE OFF

GO

ALTER DATABASE [AusData] SET AUTO\_SHRINK OFF

GO

ALTER DATABASE [AusData] SET AUTO\_UPDATE\_STATISTICS ON

GO

ALTER DATABASE [AusData] SET CURSOR\_CLOSE\_ON\_COMMIT OFF

GO

ALTER DATABASE [AusData] SET CURSOR\_DEFAULT GLOBAL

GO

ALTER DATABASE [AusData] SET CONCAT\_NULL\_YIELDS\_NULL OFF

GO

ALTER DATABASE [AusData] SET NUMERIC\_ROUNDABORT OFF

GO

ALTER DATABASE [AusData] SET QUOTED\_IDENTIFIER OFF

GO

ALTER DATABASE [AusData] SET RECURSIVE\_TRIGGERS OFF

GO

ALTER DATABASE [AusData] SET DISABLE\_BROKER

GO

ALTER DATABASE [AusData] SET AUTO\_UPDATE\_STATISTICS\_ASYNC OFF

GO

ALTER DATABASE [AusData] SET DATE\_CORRELATION\_OPTIMIZATION OFF

GO

ALTER DATABASE [AusData] SET TRUSTWORTHY OFF

GO

ALTER DATABASE [AusData] SET ALLOW\_SNAPSHOT\_ISOLATION OFF

GO

ALTER DATABASE [AusData] SET PARAMETERIZATION SIMPLE

GO

ALTER DATABASE [AusData] SET READ\_COMMITTED\_SNAPSHOT OFF

GO

ALTER DATABASE [AusData] SET HONOR\_BROKER\_PRIORITY OFF

GO

ALTER DATABASE [AusData] SET RECOVERY FULL

GO

ALTER DATABASE [AusData] SET MULTI\_USER

GO

ALTER DATABASE [AusData] SET PAGE\_VERIFY CHECKSUM

GO

ALTER DATABASE [AusData] SET DB\_CHAINING OFF

GO

ALTER DATABASE [AusData] SET FILESTREAM( NON\_TRANSACTED\_ACCESS = OFF )

GO

ALTER DATABASE [AusData] SET TARGET\_RECOVERY\_TIME = 60 SECONDS

GO

ALTER DATABASE [AusData] SET DELAYED\_DURABILITY = DISABLED

GO

ALTER DATABASE [AusData] SET ACCELERATED\_DATABASE\_RECOVERY = OFF

GO

ALTER DATABASE [AusData] SET QUERY\_STORE = OFF

GO

ALTER DATABASE [AusData] SET READ\_WRITE

GO

## Create Table Definitions

**Dim\_Address**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Dim\_Address] Script Date: 4/10/2022 12:56:08 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Dim\_Address](

[Address\_ID] [int] NOT NULL,

[State] [nvarchar](10) NULL,

[State\_code] [nvarchar](10) NULL,

[Suburb] [nvarchar](50) NULL,

[City\_name] [nvarchar](50) NULL,

[Postal\_code] [nvarchar](50) NULL,

[Latitude] [nvarchar](50) NULL,

[Longtitude] [nvarchar](50) NULL,

CONSTRAINT [PK\_Dim\_Address] PRIMARY KEY CLUSTERED

(

[Address\_ID] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO

**Dim\_Date**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Dim\_Date] Script Date: 4/10/2022 12:56:45 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Dim\_Date](

[DateKey] [int] NOT NULL,

[Date] [datetime] NULL,

[FullDate] [char](10) NULL,

[DayOfMonth] [varchar](2) NULL,

[DaySuffix] [varchar](4) NULL,

[DayName] [varchar](9) NULL,

[DayOfWeek] [char](1) NULL,

[DayOfWeekInMonth] [varchar](2) NULL,

[DayOfWeekInYear] [varchar](2) NULL,

[DayOfQuarter] [varchar](3) NULL,

[DayOfYear] [varchar](3) NULL,

[WeekOfMonth] [varchar](1) NULL,

[WeekOfQuarter] [varchar](2) NULL,

[WeekOfYear] [varchar](2) NULL,

[Month] [varchar](2) NULL,

[MonthName] [varchar](9) NULL,

[MonthOfQuarter] [varchar](2) NULL,

[Quarter] [char](1) NULL,

[QuarterName] [varchar](9) NULL,

[Year] [char](4) NULL,

[YearName] [char](7) NULL,

[MonthYear] [char](10) NULL,

[MMYYYY] [char](6) NULL,

[FirstDayOfMonth] [date] NULL,

[LastDayOfMonth] [date] NULL,

[FirstDayOfQuarter] [date] NULL,

[LastDayOfQuarter] [date] NULL,

[FirstDayOfYear] [date] NULL,

[LastDayOfYear] [date] NULL,

[IsHoliday] [bit] NULL,

[IsWeekday] [bit] NULL,

[HolidayName] [varchar](50) NULL,

PRIMARY KEY CLUSTERED

(

[DateKey] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO

**Dim\_Property**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Dim\_Property] Script Date: 4/10/2022 12:57:32 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Dim\_Property](

[Property\_id] [int] NOT NULL,

[Address\_id] [int] NULL,

[State] [nvarchar](50) NULL,

[City\_name] [nvarchar](50) NULL,

[Suburb] [nvarchar](50) NULL,

[Post\_code] [nvarchar](50) NULL,

[updated\_year] [nvarchar](50) NULL,

[updated\_month] [nvarchar](50) NULL,

CONSTRAINT [PK\_Dim\_Property] PRIMARY KEY CLUSTERED

(

[Property\_id] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO

**Dim\_School**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Dim\_School] Script Date: 4/10/2022 12:58:19 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Dim\_School](

[School\_Id] [int] IDENTITY(1,1) NOT NULL,

[School\_code] [nvarchar](50) NULL,

[Age\_Id] [nvarchar](50) NULL,

[School\_name] [nvarchar](50) NULL,

[School\_Address] [nvarchar](50) NULL,

[Suburb] [nvarchar](50) NULL,

[Post\_code] [nvarchar](50) NULL,

[Student\_number] [nvarchar](50) NULL,

[Indigenous\_pct] [nvarchar](50) NULL,

[lbote\_pct] [nvarchar](50) NULL,

[ICSEA\_Value] [nvarchar](50) NULL,

[level\_of\_schooling] [nvarchar](50) NULL,

[Selective\_school] [nvarchar](50) NULL,

[Opportunity\_class] [nvarchar](50) NULL,

[School\_specialty\_type] [nvarchar](50) NULL,

[School\_subtype] [nvarchar](50) NULL,

[Support\_classes] [nvarchar](50) NULL,

[Preschool\_ind] [nvarchar](50) NULL,

[Distance\_education] [nvarchar](50) NULL,

[Intensive\_english\_centre] [nvarchar](50) NULL,

[School\_gender] [nvarchar](10) NULL,

[phone] [nvarchar](10) NULL,

[school\_email] [nvarchar](50) NULL,

[fax] [nchar](10) NULL,

[late\_opening\_school] [nvarchar](50) NULL,

[date\_1st\_teacher] [nvarchar](50) NULL,

[lga] [nchar](10) NULL,

[electorate] [nvarchar](50) NULL,

[fed\_electorate] [nvarchar](50) NULL,

[operational\_directorate] [nvarchar](50) NULL,

[principal\_network] [nvarchar](50) NULL,

[facs\_district] [nvarchar](50) NULL,

[local\_health\_district] [nvarchar](50) NULL,

[aecg\_region] [nvarchar](50) NULL,

[ASGS\_remoteness] [nvarchar](50) NULL,

[latitude] [nvarchar](50) NULL,

[longitude] [nvarchar](50) NULL,

[date\_extracted] [nvarchar](50) NULL,

CONSTRAINT [PK\_Dim\_School] PRIMARY KEY CLUSTERED

(

[School\_Id] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO

**Dim\_Station**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Dim\_Station] Script Date: 4/10/2022 12:59:07 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Dim\_Station](

[Station\_Id] [int] NOT NULL,

[Station\_suburb] [nvarchar](50) NULL,

[Street\_name] [nvarchar](50) NULL,

[Street\_type] [nvarchar](50) NULL,

[Entrence\_type] [nvarchar](50) NULL,

[latitude] [nvarchar](50) NULL,

[Longtitude] [nvarchar](50) NULL,

[Station\_ExitNumber] [nchar](10) NULL,

CONSTRAINT [PK\_Dim\_Station] PRIMARY KEY CLUSTERED

(

[Station\_Id] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO

**Fact\_Property**

USE [AusData]

GO

/\*\*\*\*\*\* Object: Table [dbo].[Fact\_Property] Script Date: 4/10/2022 12:59:57 PM \*\*\*\*\*\*/

SET ANSI\_NULLS ON

GO

SET QUOTED\_IDENTIFIER ON

GO

CREATE TABLE [dbo].[Fact\_Property](

[FactProperty\_Key] [int] NOT NULL,

[Address\_Id] [int] NOT NULL,

[Property\_Id] [int] NOT NULL,

[Station\_Id] [int] NOT NULL,

[School\_id] [int] NOT NULL,

[DateKey] [int] NOT NULL,

[MedianvalueID] [int] NOT NULL,

[Property\_Median\_value] [int] NULL,

CONSTRAINT [PK\_Fact\_Property] PRIMARY KEY CLUSTERED

(

[FactProperty\_Key] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]

) ON [PRIMARY]

GO