DATA WAREHOUSING AND DATA MINING

Instructor:

Dr. Charitha Hettiarachchi

Project Group 5:

Pravalika Pullannagaari

Shanmukha Reddy Tanuboddy

Yuying Pu

Contents

1) Abstract ----------3

2) Project Overview----------3

3) Introduction----------4

a) What is Database?
b) What is Data Warehouse?
c) Difference between Database & Data Warehousing
d) What is Data Mining?
e) What is the need for Data Warehousing?

4) Methodology ----------6

5) Project Data Set ----------6

6) Data Cleansing----------7

a) Data Cleansing Process Used

7) Dimensional Modeling----------9

8) Data Transfer to MS Access ---------- 16

a) Steps to Import file to MS Access
b) Data in MS Access after Import

9) Data Transfer to SQL Server---------- 21

Steps to Import file to MS Access

10) Cube Development and Deployment ---------- 25

Steps to Cube Development and Deployment in Visual Studio

11) Reports ---------- 28

a) Analysis - 1
b) Analysis - 2
c) Analysis - 3
d) Analysis - 4
e) Analysis - 5
f) Analysis - 6

12) Data Mining Queries ---------- 35

a) Query - 1
b) Query - 2
c) Query - 3
d) Query - 4
e) Query - 5
f) Query - 6

13) Conclusion ---------- 44

Abstract

Project Overview:

The global market for electric vehicles has been growing rapidly in recent years, with many automakers expanding their electric vehicle offerings to meet the increasing demand and regulatory requirements. This Electric Vehicle Population Data shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

The global market is increasingly moving toward electric vehicles (EVs) due to a combination of factors that are driving the transition towards more sustainable and environmentally friendly transportation options. Here are some key reasons why the global market is shifting towards electric vehicles:

- **Environmental Concerns:** One of the primary drivers behind the push for electric vehicles is growing awareness of environmental issues, especially climate change. EVs produce zero tailpipe emissions, reducing greenhouse gas emissions and air pollution, making them a cleaner alternative to traditional internal combustion engine vehicles.
- **Government Policies and Incentives:** Many governments around the world are implementing policies and providing incentives to promote electric vehicle adoption. These measures include tax credits, subsidies, reduced registration fees, access to bus lanes, and stricter emissions regulations for traditional vehicles.
- **Advancements in Battery Technology:** Significant advancements in battery technology have resulted in improved energy storage capacity and reduced costs. Lithium-ion batteries, which are commonly used in EVs, have become more efficient and affordable, allowing for longer driving ranges and increased accessibility.
- **Falling Battery Costs:** The cost of lithium-ion batteries has been declining over the years, making electric vehicles more affordable for consumers. As battery costs continue to drop, the price parity between EVs and conventional vehicles is expected to improve further.
- **Innovation and Competition:** The EV market has witnessed increased innovation and competition from both established automakers and new entrants. This competition drives the development of more advanced and appealing electric vehicle models.
- **Public Awareness and Perception:** As public awareness of the benefits of electric vehicles grows, so does their acceptance and desirability. EVs are often seen as a symbol of progress and environmental responsibility, leading to increased consumer interest.
- **Corporate Sustainability Goals:** Many companies and fleet operators are adopting electric vehicles as part of their sustainability initiatives. Switching to EVs helps them reduce their carbon footprint and align with environmental and social responsibility goals.
- **Reduced Operating Costs:** Electric vehicles generally have lower operating costs than conventional vehicles, primarily due to lower energy costs and reduced maintenance requirements. EVs have fewer moving parts, which translates to lower maintenance expenses.
- **Urbanization and Congestion:** The shift towards electric vehicles is also driven by the challenges of urbanization, including air quality issues and traffic congestion. Electric vehicles can help reduce noise and air pollution in densely populated areas.
- **Global Agreements and Targets:** International agreements and targets, such as the Paris Agreement, have put pressure on countries to reduce their carbon emissions. Electrification of transportation, including passenger cars and public transportation, is a significant strategy for achieving these goals.

As these factors continue to converge and strengthen, the global market for electric vehicles is expected to grow further, with more automakers investing in electric vehicle technology and infrastructure to meet the increasing demand for sustainable transportation options.

In this project we choose one of the datasets which was posted on kaggle.com which provides the summary of Electric vehicles production starting from the year 1997 to 2023. The data set contains a total of 17 columns and 1,24,717 rows. We performed a data cleansing on the data set using pandas to ensure that there is no data redundancy and data anomaly . Further the data was segregated into various dimensional tables and fact tables which is the central table that stores the quantitative data for investigation and is frequently denormalized and it works with dimensional table. Later we send this data to Microsoft access and create a star schema. Next step was to create cube and deploy it. This includes SQL Server Management studio, MS Access and Visual Studio 2019. In the next step was to create cube and deploy it. This process includes SQL Server Management Studio, MS Access, and Visual Studio 2019. In MS Access we import each dimension table and fact table and structure a relationship among them and record spared in .mdb configuration and Access 2002-2003 database then it moved to SQL Server Database. To deploy cube, we utilize Visual Studio and make business intelligence project (Analysis Services Multidimensional and Data Mining).

Introduction

What is Database?

A database is an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS). Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system, often shortened to just database.

Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized. Most databases use structured query language (SQL) for writing and querying data.

What is Data Warehouse?

A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources. A Data Warehouse centralizes and consolidates large amounts of data from multiple sources. Organizations can gain useful business insights from their data using their analytical skills to enhance decision-making. Large amounts of data are electronically stored by a company and are intended for analysis and inquiry rather than transaction processing.

A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.

A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users. It is not used for daily operations and transaction processing but used for making decisions.
A Data Warehouse can be viewed as a data system with the following attributes:

- It is a database designed for investigative tasks, using data from various applications.
- It supports a relatively small number of clients with relatively long interactions.
- It includes current and historical data to provide a historical perspective of information.
- Its usage is read-intensive.
- It contains a few large tables.

Difference between Database & Data Warehousing

Data warehouses and databases both act as data storage and management tools. However, there are a few key differences to acknowledge. First, data warehouses have analytical capabilities. They enable companies to make analytical queries that track and record certain variables for business intelligence. In contrast, a database is a simple collection of data in one place. Databases' main purpose is to store data securely and allow users to access it easily.

Organizations often need both databases and data warehouses to manage the massive amounts of data they produce daily. For example, a clothing company may use one database to store customer information and another to track website traffic. They can use a data warehouse to compare both databases on a historical scale to reveal insight into consumer trends.

	Data warehouse	Database
Purpose	Analysis	Reporting
Database	OLAP (online analytical processing)	OLTP (online transactional processing)
Type of collection	Subject-oriented	Application-oriented
Query	Complex analytical queries	Simple transaction queries

Fig. 1: Data Warehouse v.s. Database

What is Data Mining?

Data mining is the process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques. The data can be structured, semi-structured or unstructured, and can be stored in various forms such as databases, data warehouses, and data lakes.

The primary goal of data mining is to discover hidden patterns and relationships in the data that can be used to make informed decisions or predictions. This involves exploring the data using various techniques such as clustering, classification, regression analysis, association rule mining, and anomaly detection.

Data mining has a wide range of applications across various industries, including marketing, finance, healthcare, and telecommunications. For example, in marketing, data mining can be used to identify customer segments and target marketing campaigns, while in healthcare, it can be used to identify risk factors for diseases and develop personalized treatment plans.

The key properties of data mining are:
- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

What is the Need for Data Warehousing?

- Data Warehouse serves as a Single Source of Truth for all the data within the company. Using a Data Warehouse eliminates the following issues:
    - Data quality issues
    - Unstable data in reports
    - Data Inconsistency
    - Low query performance
- Data Warehouse gives the ability to quickly run analysis on huge volumes of datasets.
- If there is any change in the structure of the data available in the operational or transactional Databases. It will not break the business reports running on top of it because they are not directly connected to BI tools or Reporting tools.
- Cloud Data Warehouse (such as Amazon Redshift and Google Big Query) offer an added advantage that you need not invest in them upfront. Instead, you pay as you go as the size of your data increases. You can refer to this article on Amazon Redshift vs Google Big Query for a comparison of the two.
- When companies want to make the data available for all, they will understand the need for Data Warehouse. You can expose the data within the company for analysis. While you do so you can hide certain sensitive information (such as PII – Personally Identifiable Information about your customers, or Partners).
- There is always the need for Data Warehouse as the complexity of queries increases and users need faster query processing. Because the transactional Databases are built to store a store in a normalized form whereas fast query processing can be achieved by denormalized data that is available in Data Warehouse.

Methodology

Steps for creating a Data Warehouse

Project Data Set

The project data set was summary file which contains data for the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

Data Link: https://www.kaggle.com/datasets/utkarshx27/electric-vehicle-population-data

In our actual data set, there are 1,24,717 rows and 17 columns before Data Cleansing. We took 9 columns and achieved 1,24,533 rows after data cleansing.

1) Vehicle ID: The Vehicle Identification Number is a unique alphanumeric code assigned to each vehicle for identification

2) County: The county where the vehicle is registered in Washington State.

3) City: The city where the vehicle is registered in Washington State.

4) State: The state where the vehicle is registered, which is Washington in this case.

5) Model: The specific model or name of the vehicle.

6) Year: The year in which the vehicle was manufactured.

7) Make: The manufacturer or brand of the vehicle.

8) Electric Vehicle Type: Indicates whether the vehicle is a Battery Electric Vehicle (BEV), which runs solely on electricity, or a Plug-in.

9) Clean Alternative Fuel Eligibility: Indicates if the vehicle meets the eligibility criteria for Clean Alternative Fuel Vehicle incentives or benefits.

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process, so you know you are doing it the right way every time.

Data Cleaning Steps & Techniques

Here is a 6-step data cleaning process to make sure your data is ready to go.

- Step 1: Remove irrelevant data
- Step 2: Deduplicate your data
- Step 3: Fix structural errors
- Step 4: Deal with missing data
- Step 5: Filter out data outliers
- Step 6: Validate your data

Characteristics of clean data

- **Validity**: The degree to which the measures conform to defined business rules or constraints
- **Accuracy**: The degree of conformity of a measure to a standard or a true value. Accuracy is very hard to achieve through data-cleansing in the general case because it requires accessing an external source of data that contains the true value: such "gold standard" data is often unavailable.
- **Completeness**: The degree to which all required measures are known. Incompleteness is almost impossible to fix with data cleansing methodology: one cannot infer facts that were not captured when the data in question was initially recorded. (In some contexts, e.g., interview data, it may be possible to fix incompleteness by going back to the original source of data, i.e., re-interviewing the subject, but even this does not guarantee success because of problems of recall - e.g., in an interview to gather data on food consumption, no one is likely to remember exactly what one ate six months ago. In the case of systems that insist certain columns should not be empty, one may work around the problem by designating a value that indicates "unknown" or "missing", but the supplying of default values does not imply that the data has been made complete.)
- **Consistency**: The degree to which a set of measures are equivalent in across systems. Inconsistency occurs when two data items in the data set contradict each other: e.g., a customer is recorded in two different systems as having two different current addresses, and only one of them can be correct. Fixing inconsistency is not always possible: it requires a variety of strategies - e.g., deciding which data were recorded more recently, which data source is likely to be most reliable (the latter knowledge may be specific to a given organization), or simply trying to find the truth by testing both data items (e.g., calling up the customer).
- **Uniformity**: The degree to which a set data measures are specified using the same units of measure in all systems. In datasets pooled from different locales, weight may be recorded either in pounds or kilos and must be converted to a single measure using an arithmetic transformation.

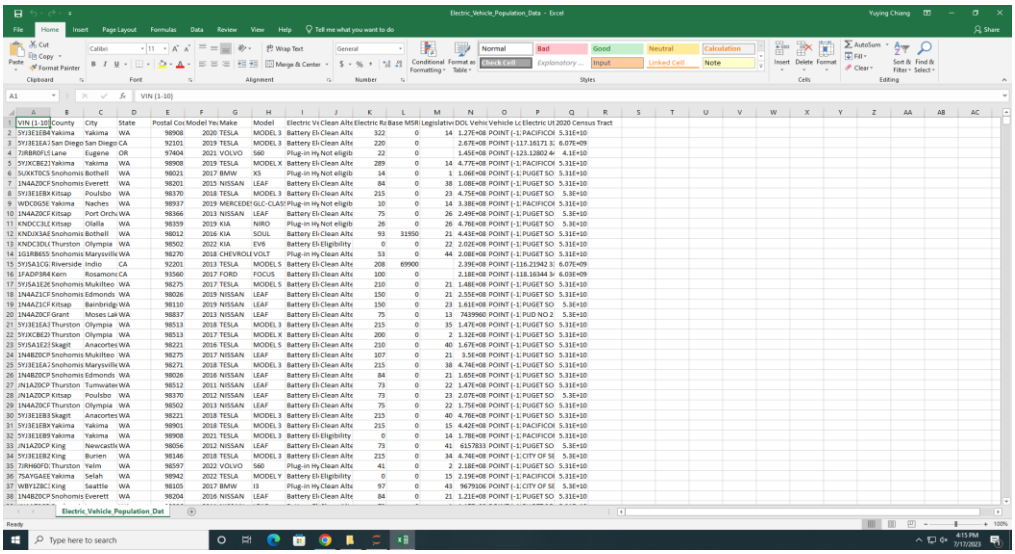


Fig. 2-1: Uncleaned Data

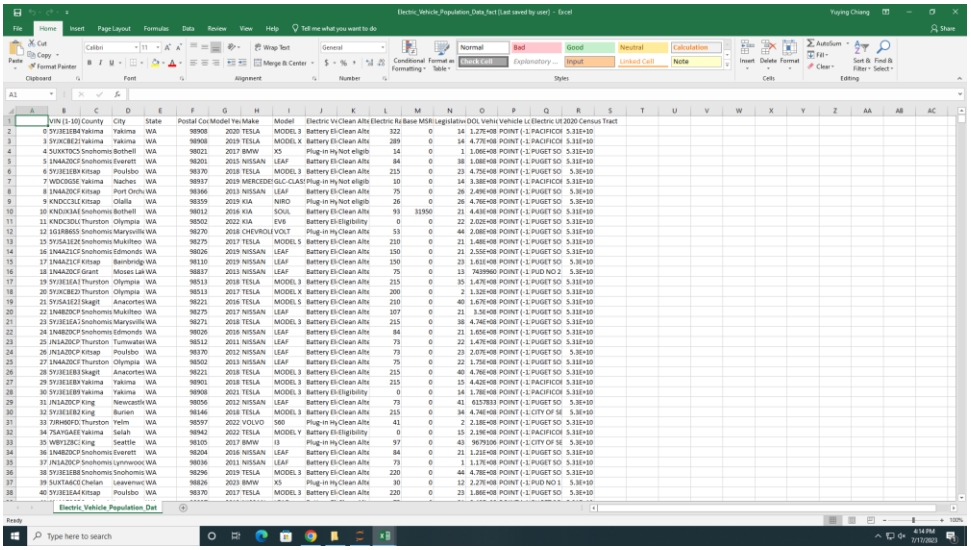


Fig. 2-2: Cleaned Data

Data Cleaning Process Used:

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. Pandas is one of those packages and makes importing and analyzing data much easier. Pandas provide data analysts a way to delete and filter data frame using. drop () method. Rows or columns can be removed using index label or column name using this method.

General Steps:

1) Import the pandas

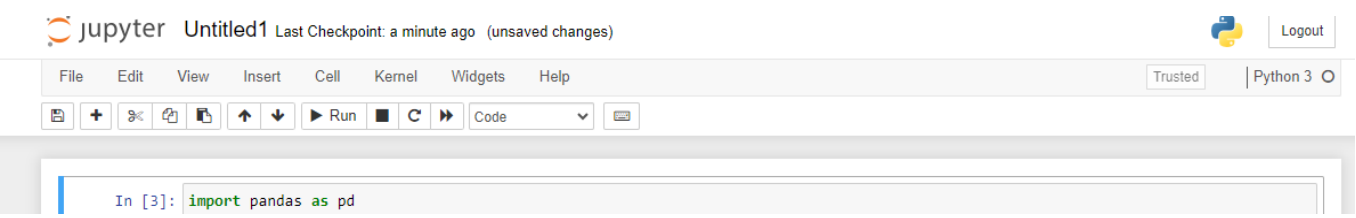


Fig. 3-1: Importing the Pandas

2) Import csv into a Pandas Data Frame object flights = pd.read_csv('filename.csv')

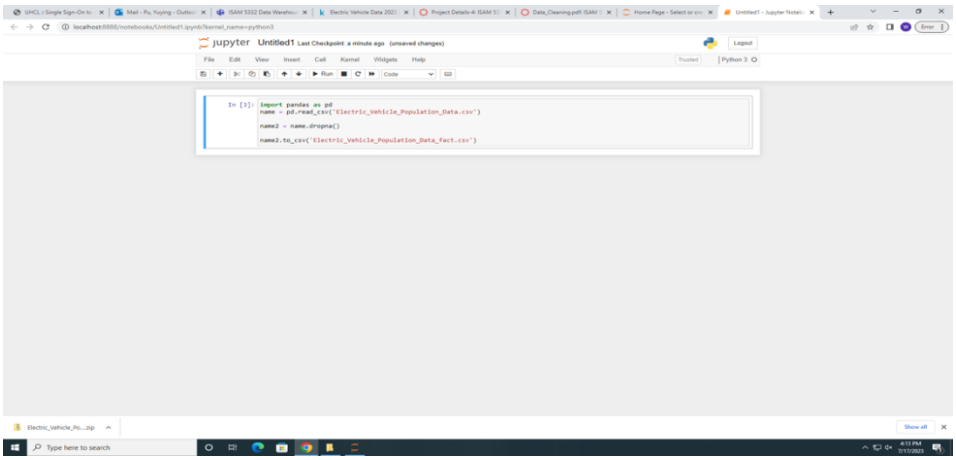


Fig. 3-2: Importing the Csv File

3) (Optional) Check for all null values in your dataset. This will return a Boolean stating if each cell is null. This can take a long time and may not be particularly useful in a very large **dataset.filename.isnull()**

Dimensional Modeling

Dimensional Data Modeling is one of the data modeling techniques used in data warehouse design. The concept of Dimensional Modeling was developed by Ralph Kimball which is comprised of facts and dimension tables. Since the main goal of this modeling is to improve the data retrieval so it is optimized for SELECT OPERATION. The advantage of using this model is that we can store data in such a way that it is easier to store and retrieve the data once stored in a data warehouse. The dimensional model is the data model used by many OLAP systems.

Elements of Dimensional Data Model

Facts

Facts are the measurable data elements that represent the business metrics of interest. For example, in a sales data warehouse, the facts might include sales revenue, units sold, and profit margins. Each fact is associated with one or more dimensions, creating a relationship between the fact and the descriptive data.

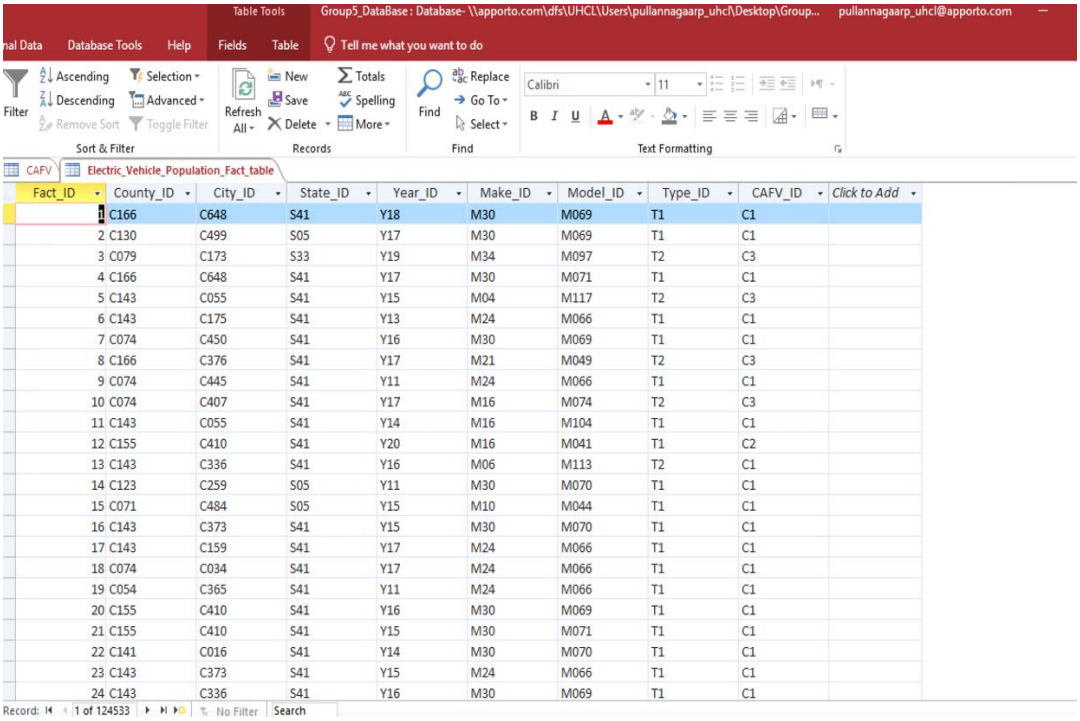
Dimension

Dimensions are the descriptive data elements that are used to categorize or classify the data. For example, in a sales data warehouse, the dimensions might include product, customer, time, and location. Each dimension is made up of a set of attributes that describe the dimension. For example, the product dimension might include attributes such as product name, product category, and product price.

Fact Table:

A Fact Table consists of measurements, metrics or facts. It is located at the center of a Snowflake Schema and in Star Schema surrounded by dimension tables. In Fact, Table we have mainly two types of Keys. Foreign key which is in another dimension table and primary key of a fact table is usually a composite key that is made up of all its foreign key.

Fact_table: Fact_ID,County_ID ,City_ID , State_ID , Year_ID , Make_ID , Type_ID, CAFV_ID



Fact_ID	County_ID	City_ID	State_ID	Year_ID	Make_ID	Model_ID	Type_ID	CAFV_ID	Click to Add
1	C166	C648	S41	Y18	M30	M069	T1	C1	
2	C130	C499	S05	Y17	M30	M069	T1	C1	
3	C079	C173	S33	Y19	M34	M097	T2	C3	
4	C166	C648	S41	Y17	M30	M071	T1	C1	
5	C143	C055	S41	Y15	M04	M117	T2	C3	
6	C143	C175	S41	Y13	M24	M066	T1	C1	
7	C074	C450	S41	Y16	M30	M069	T1	C1	
8	C166	C376	S41	Y17	M21	M049	T2	C3	
9	C074	C445	S41	Y11	M24	M066	T1	C1	
10	C074	C407	S41	Y17	M16	M074	T2	C3	
11	C143	C055	S41	Y14	M16	M104	T1	C1	
12	C155	C410	S41	Y20	M16	M041	T1	C2	
13	C143	C336	S41	Y16	M06	M113	T2	C1	
14	C123	C259	S05	Y11	M30	M070	T1	C1	
15	C071	C484	S05	Y15	M10	M044	T1	C1	
16	C143	C373	S41	Y15	M30	M070	T1	C1	
17	C143	C159	S41	Y17	M24	M066	T1	C1	
18	C074	C034	S41	Y17	M24	M066	T1	C1	
19	C054	C365	S41	Y11	M24	M066	T1	C1	
20	C155	C410	S41	Y16	M30	M069	T1	C1	
21	C155	C410	S41	Y15	M30	M071	T1	C1	
22	C141	C016	S41	Y14	M30	M070	T1	C1	
23	C143	C373	S41	Y15	M24	M066	T1	C1	
24	C143	C336	S41	Y16	M30	M069	T1	C1	

Fig. 4-1: Fact Table

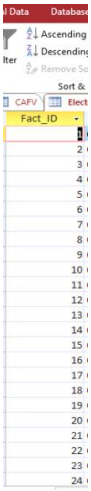


Fig. 4-2: Fact_ID

Dimension Tables:

- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- Dimension tables are denormalized tables.
- The Dimension Attributes are the various columns in a dimension table.
- Dimensions offers descriptive characteristics of the facts with the help of their attributes.
- No set limit set for given for number of dimensions.
- The dimension can also contain one or more hierarchical relationships

County- County_ID, County

City - City_ID, City

State – State_ID, State

Year – Year _ID, Year

Model- Model_ID,Model

Make – Make _ID, Make

Type – Type_ID, Type

CAFV- CAFV_ID, CAFV

County:

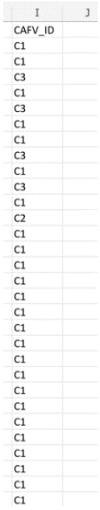
A	B
Country_	County
C001	Adams
C002	Alameda
C003	Alexandria
C004	Allen
C005	Anne Arundel
C006	Arapahoe
C007	Arlington
C008	Asotin
C009	Bartow
C010	Beaufort
C011	Bell
C012	Benton
C013	Bexar
C014	Boulder
C015	Broward
C016	Burlington
C017	Calvert
C018	Camden
C019	Cape May
C020	Carroll
C021	Carteret
C022	Charles
C023	Charleston
C024	Chelan
C025	Chesapeake
C026	Clallam
C027	Clark

Fig. 5-1: Raw Data

CAFV | Electric_Vehicle_Population_Fact_table

Country_ID	County	Click to Add
C001	Adams	
C002	Alameda	
C003	Alexandria	
C004	Allen	
C005	Anne Arundel	
C006	Arapahoe	
C007	Arlington	
C008	Asotin	
C009	Bartow	
C010	Beaufort	
C011	Bell	
C012	Benton	
C013	Bexar	
C014	Boulder	
C015	Broward	
C016	Burlington	
C017	Calvert	
C018	Camden	
C019	Cape May	
C020	Carroll	
C021	Carteret	
C022	Charles	
C023	Charleston	
C024	Chelan	

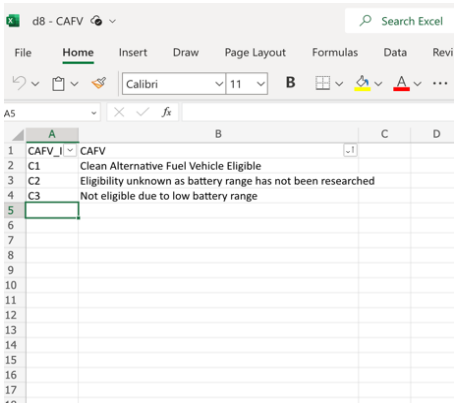
Fig. 5-2: Dimensional Table



I	J
CAFV_ID	
C1	
C1	
C3	
C1	
C3	
C1	
C1	
C3	
C1	
C3	
C1	
C2	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	
C1	

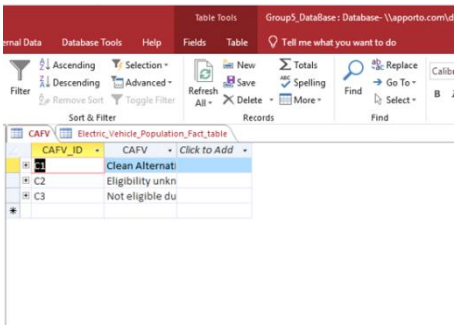
Fig. 5-3: Fact Table

CAFV:



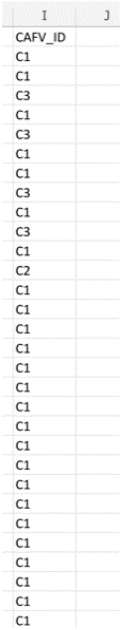
CAFV_I	CAFV
C1	Clean Alternative Fuel Vehicle Eligible
C2	Eligibility unknown as battery range has not been researched
C3	Not eligible due to low battery range

Fig 6-1: Raw data



CAFV_ID	CAFV
C1	Clean Alternati
C2	Eligibility unkn
C3	Not eligible du

Fig 6-2: Dimensional table



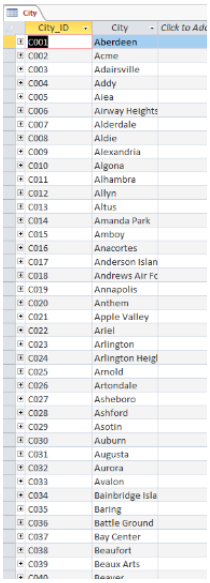
CAFV_ID
C1
C1
C3
C1
C3
C1
C1
C3
C1
C3
C1
C2
C1
C1
C1
C1
C1
C1
C1
C1
C1
C1
C1
C1
C1
C1

Fig. 6-3: Fact Table

City:

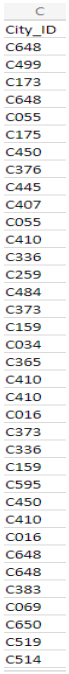
City_ID	City
C001	Aberdeen
C002	Acme
C003	Adairsville
C004	Addy
C005	Aiea
C006	Airway Heights
C007	Alderdale
C008	Aldie
C009	Alexandria
C010	Algona
C011	Alhambra
C012	Allyn
C013	Altus
C014	Amanda Park
C015	Amboy
C016	Anacortes
C017	Anderson Island
C018	Andrews Air Force Base
C019	Annapolis
C020	Anthem
C021	Apple Valley
C022	Ariel
C023	Arlington
C024	Arlington Heights
C025	Arnold
C026	Artondale
C027	Asheboro
C028	Ashford
C029	Asotin
C030	Auburn
C031	Augusta
C032	Aurora
C033	Avalon
C034	Bainbridge Island

Fig. 7-1: Raw Data



City_ID	City
C001	Aberdeen
C002	Acme
C003	Adairsville
C004	Addy
C005	Aiea
C006	Airway Heights
C007	Alderdale
C008	Aldie
C009	Alexandria
C010	Algona
C011	Alhambra
C012	Allyn
C013	Altus
C014	Amanda Park
C015	Amboy
C016	Anacortes
C017	Anderson Islan
C018	Andrews Air Fc
C019	Annapolis
C020	Anthem
C021	Apple Valley
C022	Ariel
C023	Arlington
C024	Arlington Heigl
C025	Arnold
C026	Artondale
C027	Asheboro
C028	Ashford
C029	Asotin
C030	Auburn
C031	Augusta
C032	Aurora
C033	Avalon
C034	Bainbridge Isla
C035	Baring
C036	Battle Ground
C037	Bay Center
C038	Beaufort
C039	Beaux Arts

Fig. 7-2: Dimensional Table



City_ID
C648
C499
C173
C648
C055
C175
C450
C376
C445
C407
C055
C410
C336
C259
C484
C373
C159
C034
C365
C410
C410
C016
C373
C336
C159
C595
C450
C410
C016
C648
C648
C383
C069
C650
C519
C514

Fig. 7-3: Fact_ID

Model:

	A	B
1	Model_ID	Model
2	M001	500
3	M002	918
4	M003	330E
5	M004	530E
6	M005	740E
7	M006	745E
8	M007	745LE
9	M008	A3
10	M009	A7
11	M010	A8 E
12	M011	ACCORD
13	M012	ARIYA
14	M013	AVIATOR
15	M014	B-CLASS
16	M015	BENTAYGA
17	M016	BOLT EUV
18	M017	BOLT EV
19	M018	BZ4X
20	M019	C40
21	M020	CAYENNE
22	M021	C-CLASS
23	M022	CITY
24	M023	CLARITY

Fig. 8-1: Raw Data

City | Model

Model_ID	Model	Click to Add
M001	500	
M002	918	
M003	330E	
M004	530E	
M005	740E	
M006	745E	
M007	745LE	
M008	A3	
M009	A7	
M010	A8 E	
M011	ACCORD	
M012	ARIYA	
M013	AVIATOR	
M014	B-CLASS	
M015	BENTAYGA	
M016	BOLT EUV	
M017	BOLT EV	
M018	BZ4X	
M019	C40	
M020	CAYENNE	
M021	C-CLASS	
M022	CITY	
M023	CLARITY	
M024	C-MAX	
M025	CORSAIR	
M026	COUNTRYMAN	

Fig. 8-2: Dimensional Table

G
Model_ID
M069
M069
M097
M071
M117
M066
M069
M049
M066
M074
M104
M041
M113
M070
M044
M070
M066
M066
M066
M069
M071
M070
M066
M069

Fig. 8-3: Fact _ID

Make:

	A	B
	Make_	Make
	M01	AUDI
	M02	AZURE DYNAMICS
	M03	BENTLEY
	M04	BMW
	M05	CADILLAC
	M06	CHEVROLET
	M07	CHRYSLER
	M08	FIAT
	M09	FISKER
	M10	FORD
	M11	GENESIS
	M12	HONDA
	M13	HYUNDAI
	M14	JAGUAR
	M15	JEEP
	M16	KIA
	M17	LAND ROVER
	M18	LEXUS
	M19	LINCOLN
	M20	LUCID MOTORS
	M21	MERCEDES-BENZ
	M22	MINI
	M23	MITSUBISHI
	M24	NISSAN

Fig. 9-1: Raw Data.

Make

Make_ID	Make	Click to Add
M01	AUDI	
M02	AZURE DYNAM	
M03	BENTLEY	
M04	BMW	
M05	CADILLAC	
M06	CHEVROLET	
M07	CHRYSLER	
M08	FIAT	
M09	FISKER	
M10	FORD	
M11	GENESIS	
M12	HONDA	
M13	HYUNDAI	
M14	JAGUAR	
M15	JEEP	
M16	KIA	
M17	LAND ROVER	
M18	LEXUS	
M19	LINCOLN	

Fig. 9-2: Dimensional Table

F
Make_ID
M30
M30
M34
M30
M04
M24
M30
M21
M24
M16
M16
M16
M06
M30
M10
M30
M24
M24
M24
M30
M30

Fig. 9-3: Fact _ID

Year:

	A	B
	Year_ID	Year
	Y01	1997
	Y02	1998
	Y03	1999
	Y04	2000
	Y05	2002
	Y06	2003
	Y07	2008
	Y08	2010
	Y09	2011
	Y10	2012
	Y11	2013
	Y12	2014
	Y13	2015
	Y14	2016
	Y15	2017
	Y16	2018
	Y17	2019
	Y18	2020
	Y19	2021
	Y20	2022
	Y21	2023

Fig. 10-1: Raw Data

Year

Year_ID	Year	Click to Add
Y01	1997	
Y02	1998	
Y03	1999	
Y04	2000	
Y05	2002	
Y06	2003	
Y07	2008	
Y08	2010	
Y09	2011	
Y10	2012	
Y11	2013	
Y12	2014	
Y13	2015	
Y14	2016	
Y15	2017	
Y16	2018	
Y17	2019	
Y18	2020	
Y19	2021	
Y20	2022	

Fig. 10-2: Dimensional Table

Year_ID
Y18
Y17
Y19
Y17
Y15
Y13
Y16
Y17
Y11
Y17
Y14
Y20
Y16
Y11
Y15
Y15
Y17
Y17
Y11

Fig. 10-3: Fact _ID

State:

A	B
State_I	State
S01	AK
S02	AL
S03	AR
S04	AZ
S05	CA
S06	CO
S07	CT
S08	DC
S09	DE
S10	FL
S11	GA
S12	HI
S13	ID
S14	IL
S15	IN
S16	KS
S17	KY
S18	LA
S19	MA

Fig. 11-1: Raw Data

Year | State

State_ID	State	Click to Add
S01	AK	
S02	AL	
S03	AR	
S04	AZ	
S05	CA	
S06	CO	
S07	CT	
S08	DC	
S09	DE	
S10	FL	
S11	GA	
S12	HI	
S13	ID	
S14	IL	
S15	IN	
S16	KS	
S17	KY	
S18	LA	
S19	MA	
S20	MD	
S21	MN	
S22	MO	

Fig. 11-2: Dimensional Table

D
State_ID
S41
S05
S33
S41
S41
S41
S41
S41
S41
S41
S41
S41
S41
S05
S05
S41
S41
S41
S41
S41
S41
S41

Fig. 11-3: Fact_ID

Type:

A	B
Type_ID	Type
T1	Battery Electric Vehicle (BEV)
T2	Plug-in Hybrid Electric Vehicle (PHEV)

Fig. 12-1: Raw Data

Sort & Filter | Records

Year | State | Type

Type_ID	Type	Click to Add
T1	Battery Electric	
T2	Plug-in Hybrid	

Fig. 12-2: Dimensional Table

H
Type_ID
T1
T1
T2
T1
T2
T1
T1
T2
T1
T2
T1
T1
T2
T1
T1
T1
T1
T1
T1

Fig. 12-3: Fact_ID

Dimensional Hierarchies

A hierarchy is a set of levels having many-to-one relationships between each other, and the set of levels collectively makes up a dimension in the form of a tree (A tree shows a hierarchical relationship) each of the elements of a dimension could be summarized using a hierarchy. The hierarchy is a series of parent-child relationships, typically where a parent member represents the consolidation of the members which are its children. Parent members can be further aggregated as the children of another parent.

The analyst could start at a highly summarized level, such as the total difference between the actual results and the budget, and drill down into the cube to discover which locations, products and periods had produced this difference. The hierarchical structure of dimensions provides the basis for analyzing data through drill down and roll up along its different levels.

Modern software is very useful when designing fact tables, dimension tables, and establishing the relationships between them. There are two types of schemas generally used in a data warehouse. A scheme is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires maintaining a schema. A database uses relational model, while a data warehouse uses Star, Snowflake schema. Each dimension in a star schema is represented with only one-dimension table. This dimension table contains the set of attributes.

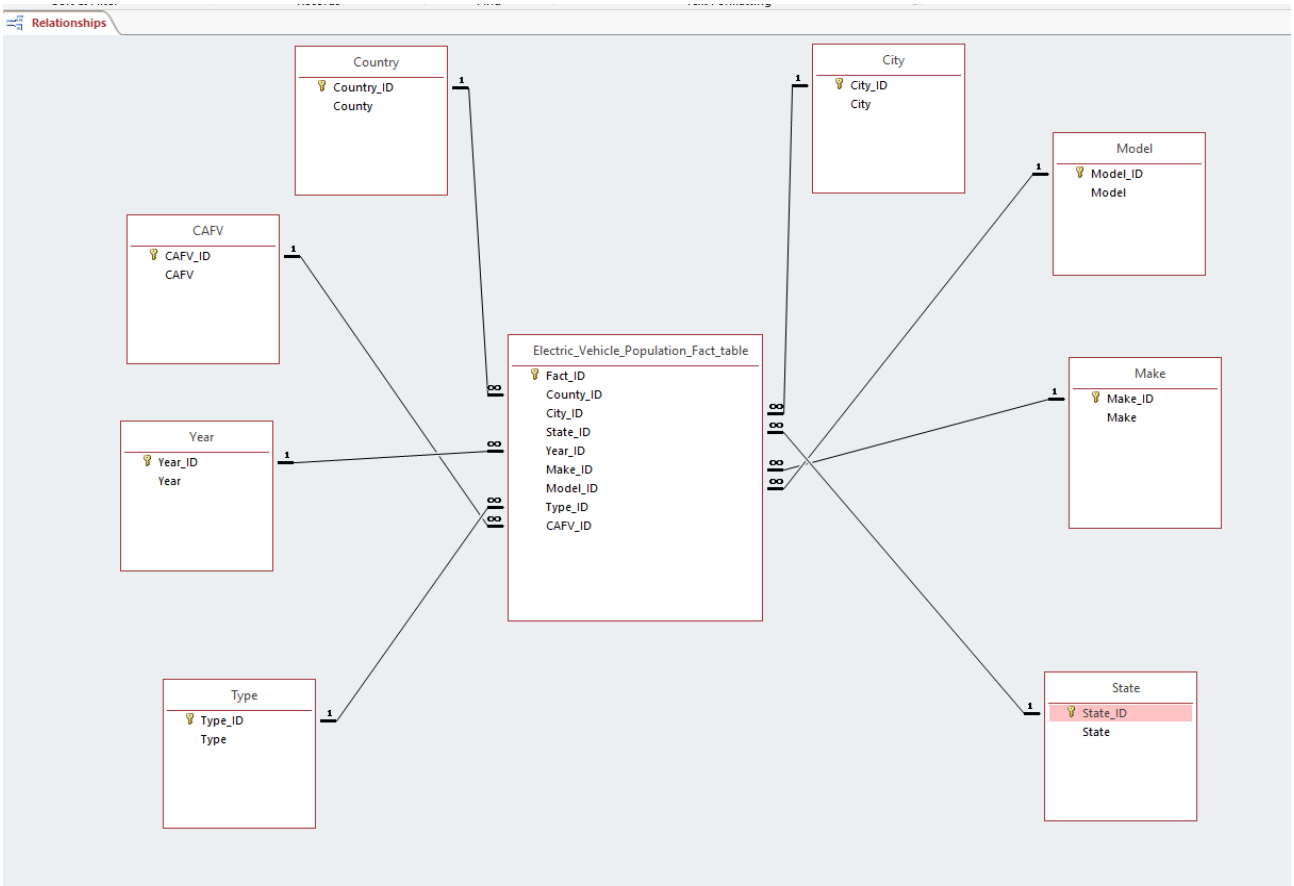


Fig. 13: Star Schema

Data Transfer to MS Access

Referential mapping and STAR schema display are made possible by the data importation into MS Access. The fact table and the dimension tables are linked in this schema. Despite the fact that there are numerous starting points in this data collection, the one-to-many relationship between each dimension and the fact table serves as the STAR schema's foundation rather than the secondary relationships that one would find in a Snowflake schema.

Steps to import file to MS Access:

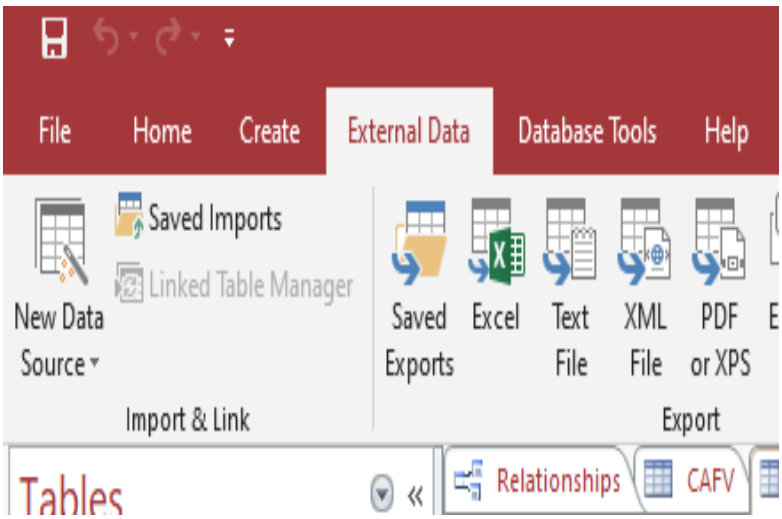


Fig. 14-1: Import File

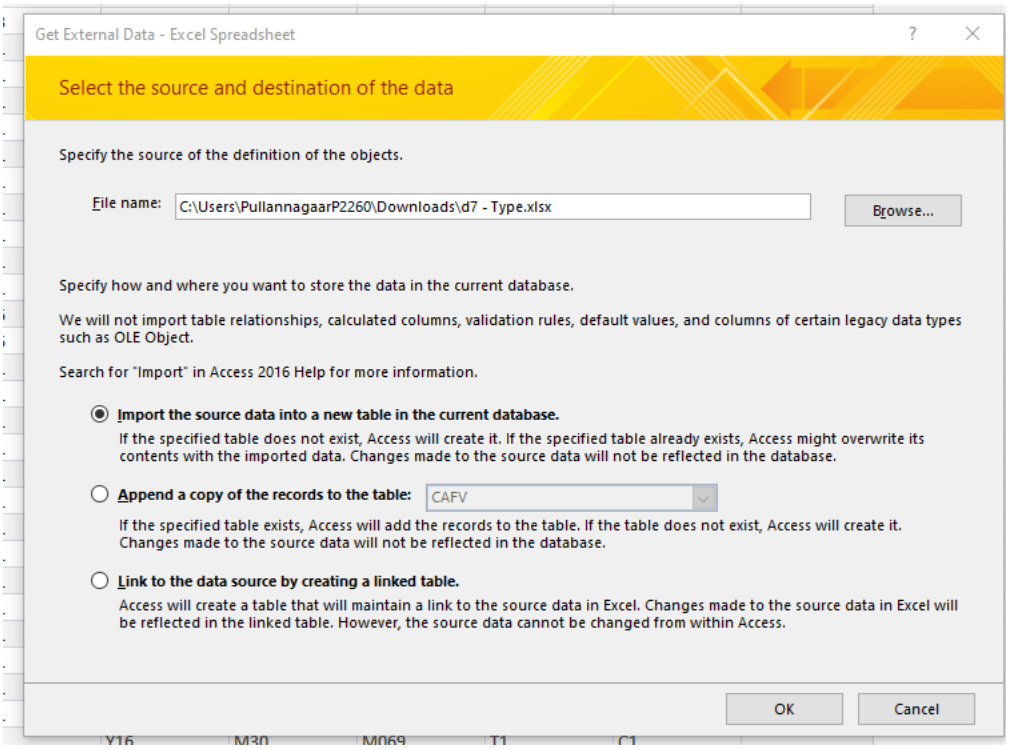


Fig. 14-2: Import the Excel File into a New Table

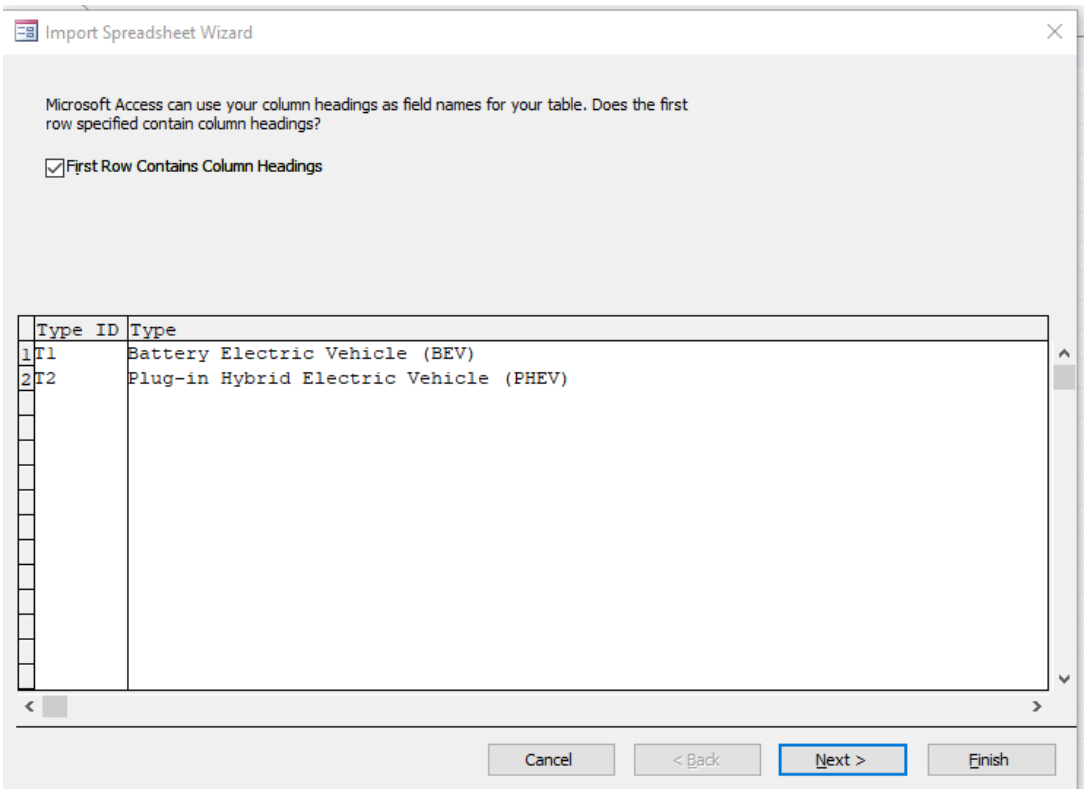


Fig. 14-3: Import Using Import Spreadsheet Wizard

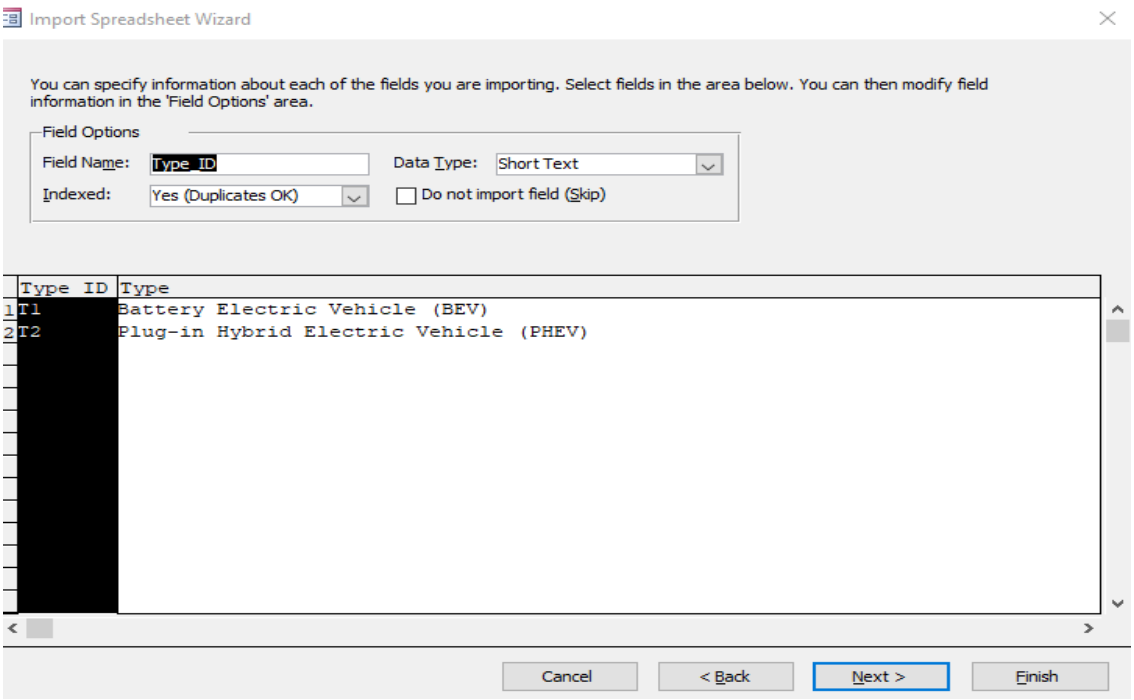


Fig. 14-4: Modify the Field Options

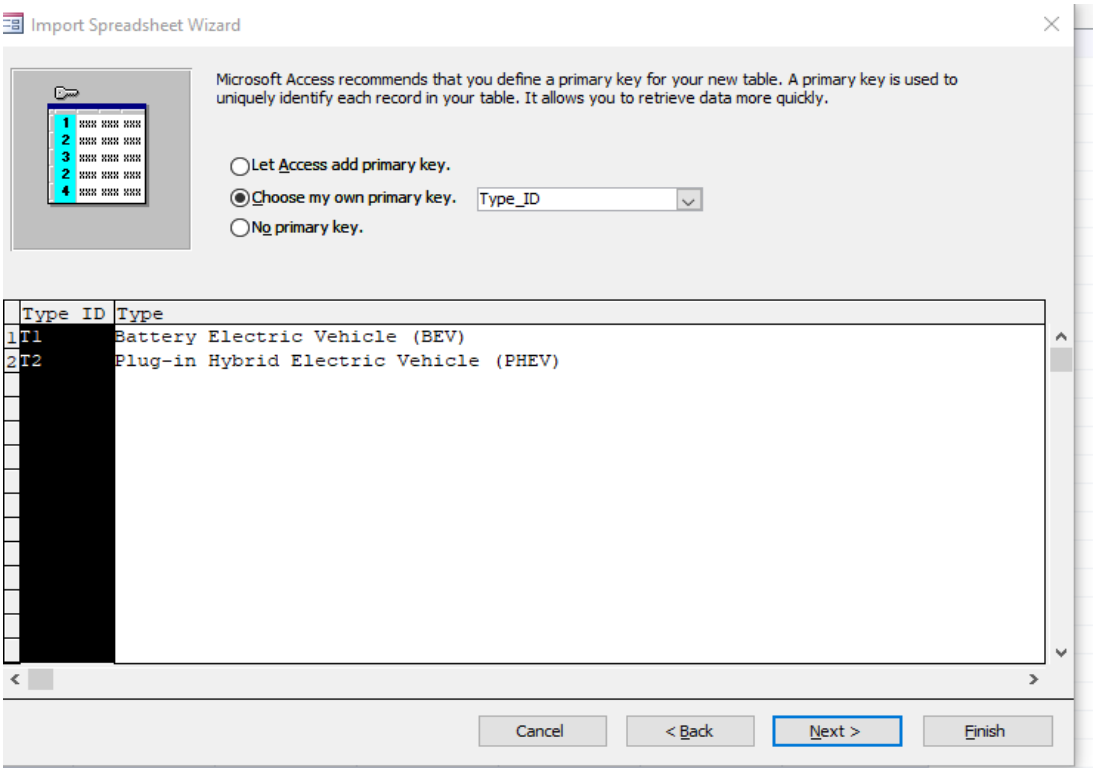


Fig. 14-5: Choose Primary Key

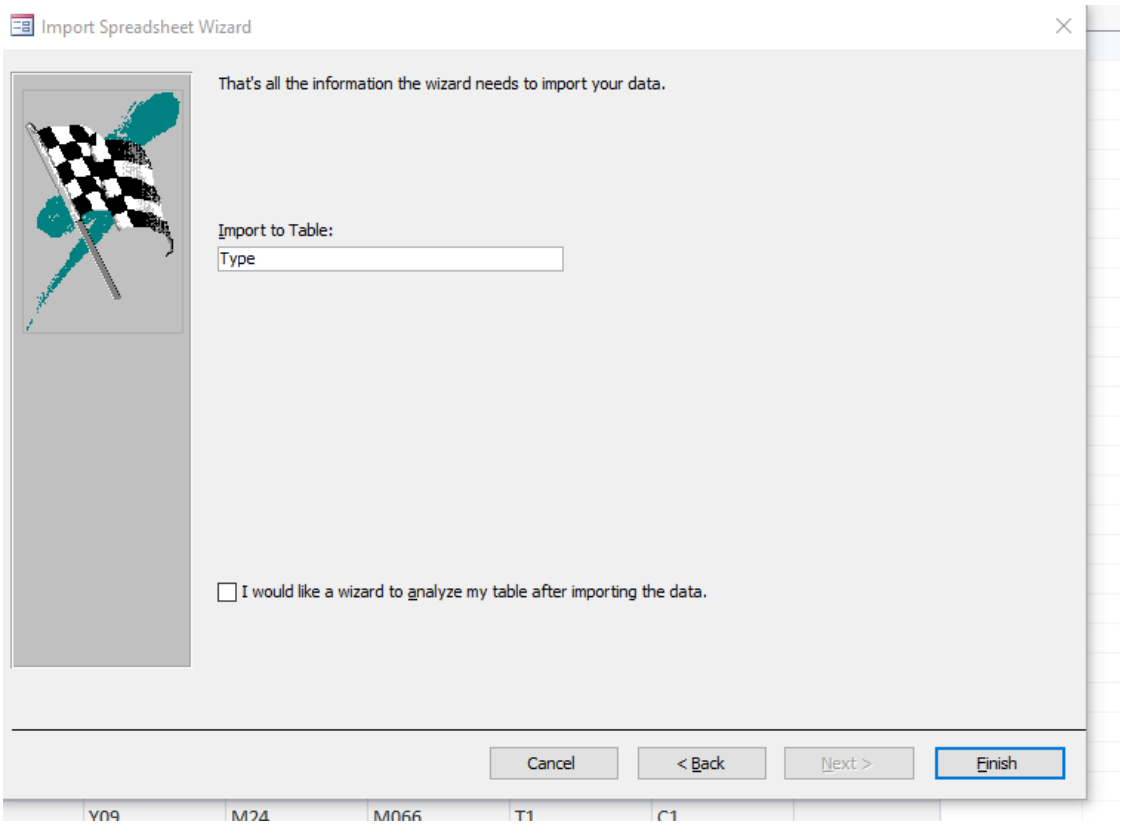


Fig. 14-6: Import to Table

Data in MS Access after import:

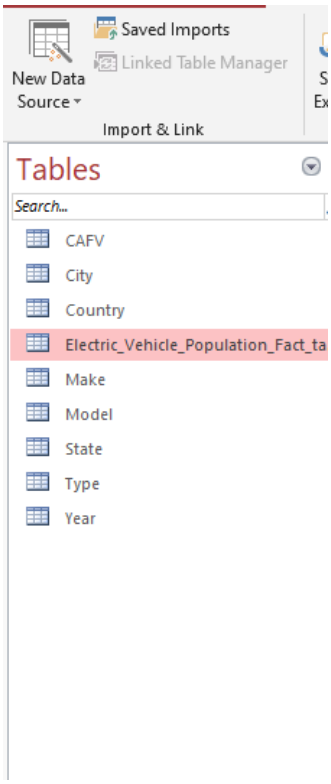
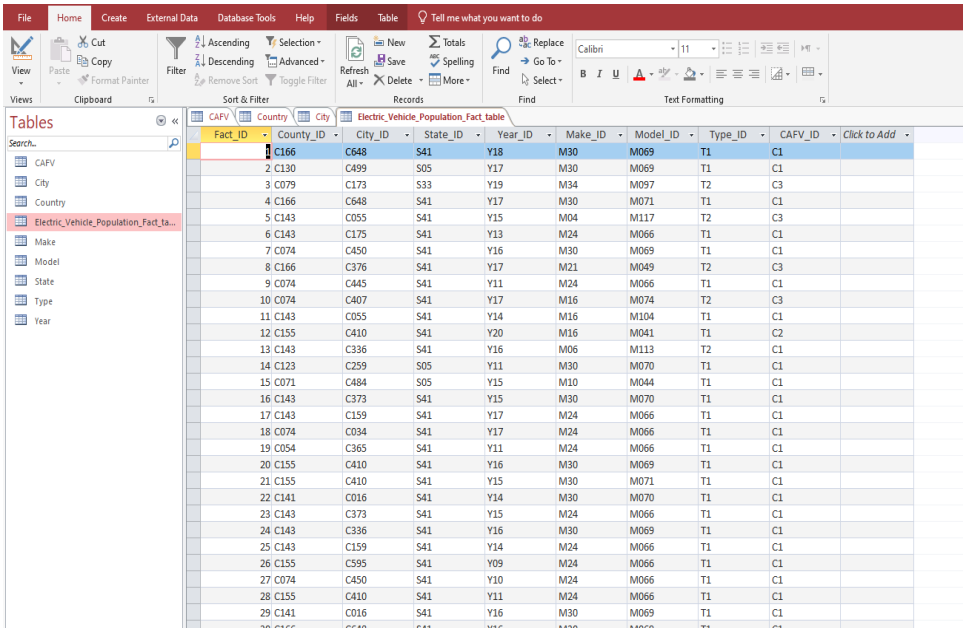


Fig. 15-1: Select the Fact Table



Fact_ID	County_ID	City_ID	State_ID	Year_ID	Make_ID	Model_ID	Type_ID	CAFV_ID	Click to Add
1	C166	C648	S41	Y18	M30	M069	T1	C1	
2	C130	C499	S05	Y17	M30	M069	T1	C1	
3	C079	C173	S33	Y19	M34	M097	T2	C3	
4	C166	C648	S41	Y17	M30	M071	T1	C1	
5	C143	C055	S41	Y15	M04	M117	T2	C3	
6	C143	C175	S41	Y13	M24	M066	T1	C1	
7	C074	C450	S41	Y16	M30	M069	T1	C1	
8	C166	C376	S41	Y17	M21	M049	T2	C3	
9	C074	C445	S41	Y11	M24	M066	T1	C1	
10	C074	C407	S41	Y17	M16	M074	T2	C3	
11	C143	C055	S41	Y14	M16	M104	T1	C1	
12	C155	C410	S41	Y20	M16	M041	T1	C2	
13	C143	C336	S41	Y16	M06	M113	T2	C1	
14	C123	C259	S05	Y11	M30	M070	T1	C1	
15	C071	C484	S05	Y15	M10	M044	T1	C1	
16	C143	C373	S41	Y15	M30	M070	T1	C1	
17	C143	C159	S41	Y17	M24	M066	T1	C1	
18	C074	C034	S41	Y17	M24	M066	T1	C1	
19	C054	C365	S41	Y11	M24	M066	T1	C1	
20	C155	C410	S41	Y16	M30	M069	T1	C1	
21	C155	C410	S41	Y15	M30	M071	T1	C1	
22	C141	C016	S41	Y14	M30	M070	T1	C1	
23	C143	C373	S41	Y15	M24	M066	T1	C1	
24	C143	C336	S41	Y16	M30	M069	T1	C1	
25	C143	C159	S41	Y14	M24	M066	T1	C1	
26	C155	C595	S41	Y09	M24	M066	T1	C1	
27	C074	C450	S41	Y10	M24	M066	T1	C1	
28	C155	C410	S41	Y11	M24	M066	T1	C1	
29	C141	C016	S41	Y16	M30	M069	T1	C1	

Fig. 15-2: The Contents of the Fact Table

Data Transfer to SQL Server

The comprehensive Graphical User Interface (GUI) support and Object Explorer, which let users browse, pick, and interact with any object on the SQL Express Server, are two of the main features of SQL Server Management Studio (SSMS). SSMS is a free download and an integrated environment for administering SQL Express that can interface with other Microsoft products. SSMS is used for database management, database interface, and data warehouse construction. Any SQL infrastructure, whether it be on a local computer (laptop, desktop, server), or in the cloud (Azure Cloud), may be managed using the SSMS, an integrated environment. An excellent tool for understanding databases and data warehousing is SQL Express with SSMS.

SQL Server Management Studio (SSMS) is chosen for its wide variety of features such as its graph generation ability and the script editors that facilitates work on the mining queries. Since the choice for the cube development is through MS Visual Studio and the DB is built on MS access, SSMS also facilitates stability for being the same domain tool, better compatibility and consistent configuration across tools.

Steps to transfer file to SQL Server:

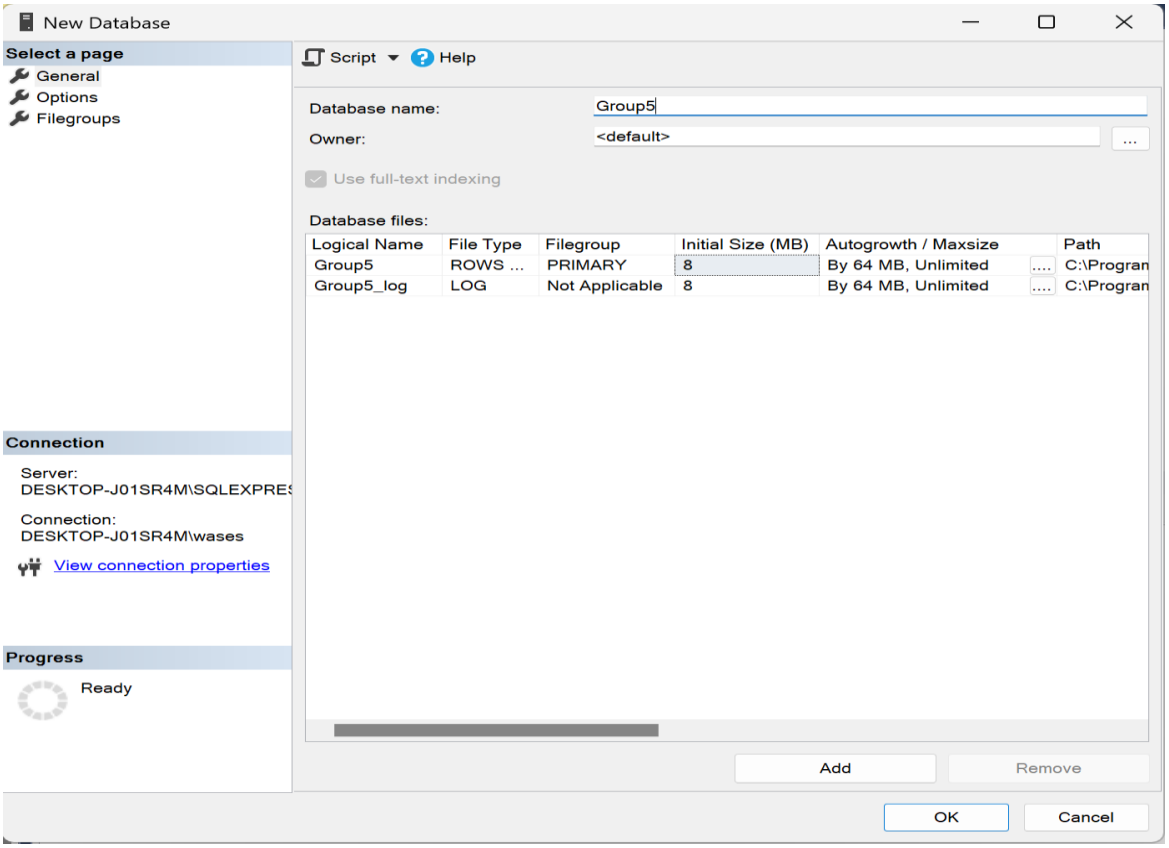


Fig. 16-1: Create a New Database

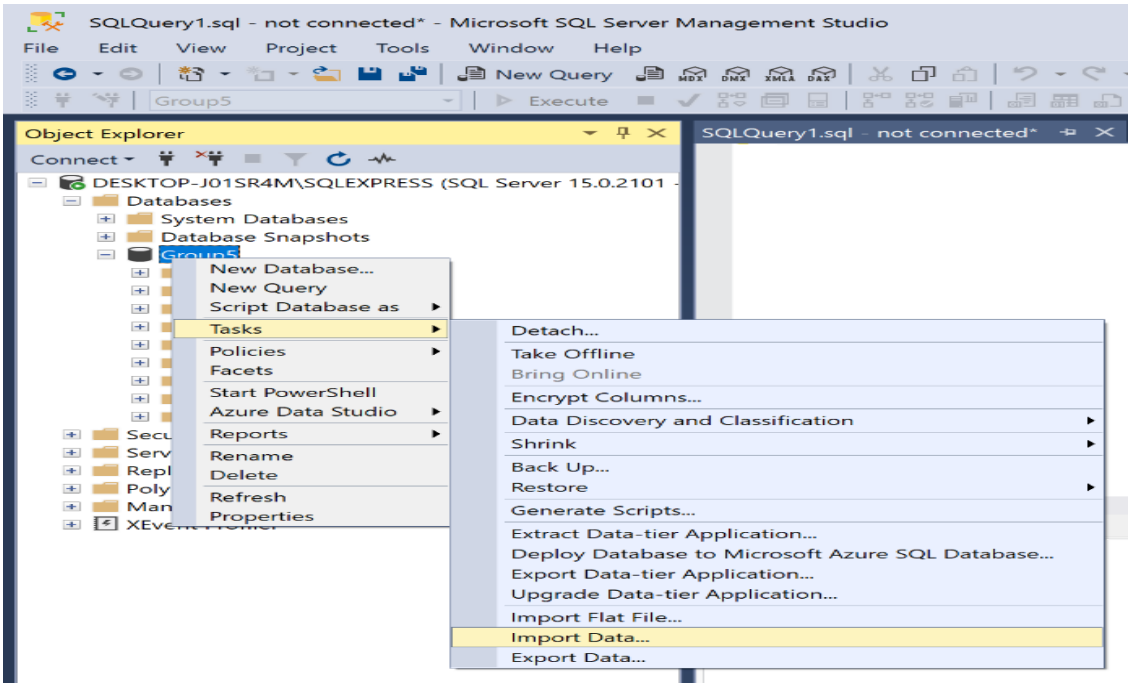


Fig. 16-2: Import Data

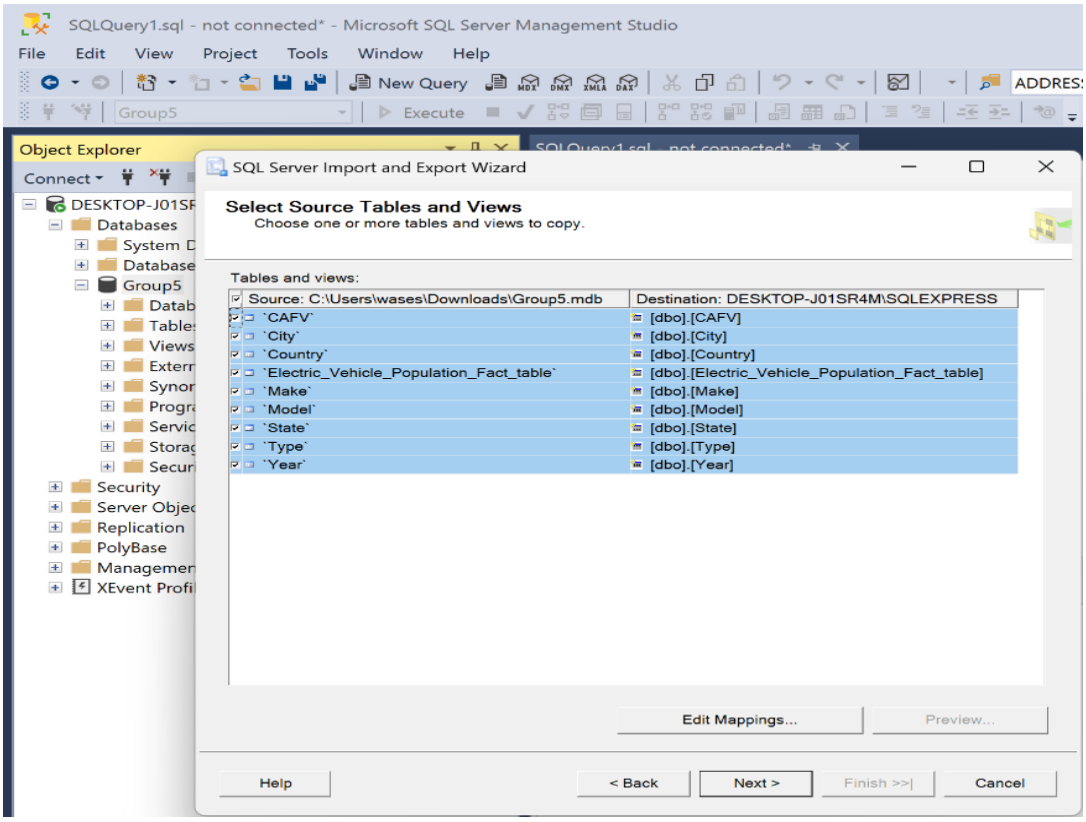


Fig. 16-3: Select Columns to be Imported

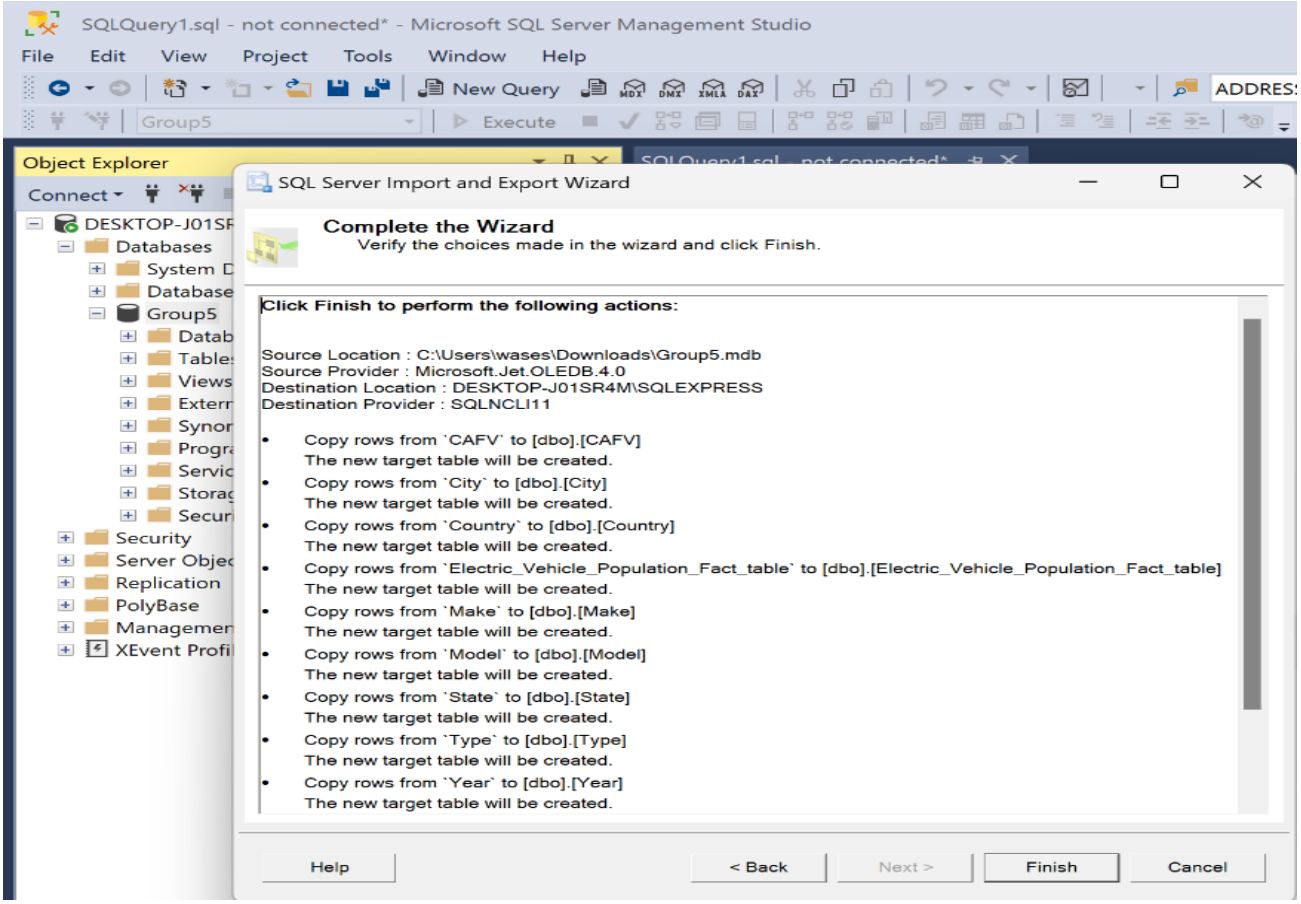


Fig. 16-4: Complete Importing

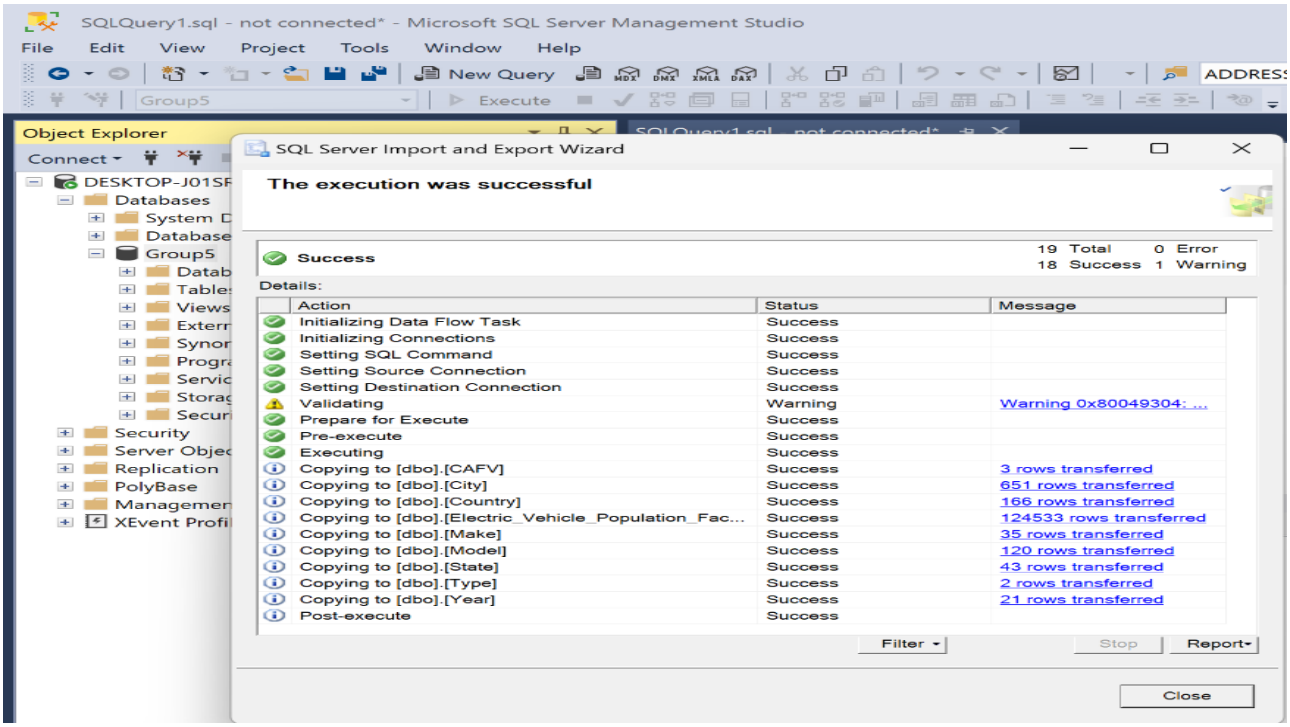


Fig. 16-5: Imported Succeeded

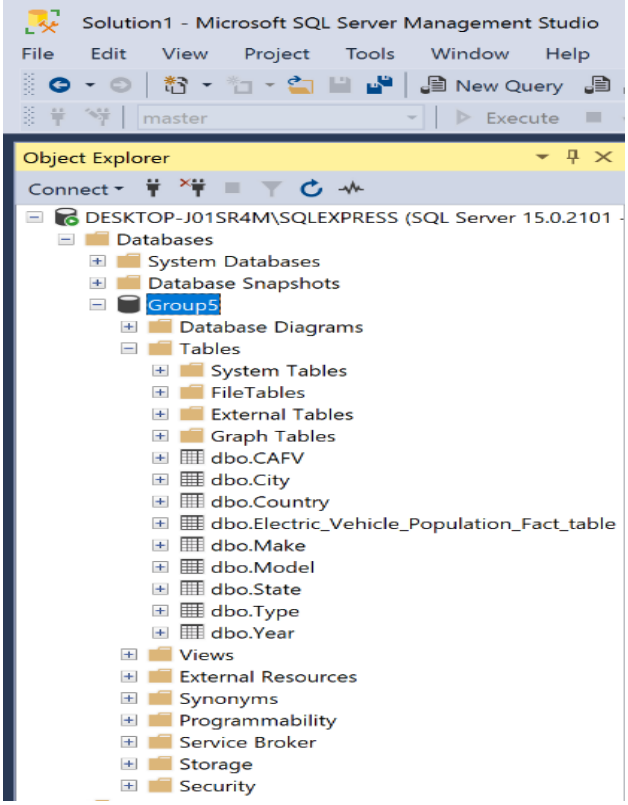


Fig. 16-6: The Tables in Group5

Cube Development and Deployment in Visual Studio

The database is generated from the US Accidents data pulled from Kaggle.com and the data is cleansed , modelled and uploaded to MS access. The MS Access file generated is being taken as DB on MS SQL Server Management Studio for OLAP analysis. This database is being taken as source to generate the cube for Analytical Processing.

Steps to Deployment and Cube development in Visual Studio:

1. Fact table has been created and the necessary dimensions have been identified
2. Data in MS Access is taken as Data-Source to upload into SQL server.
3. The SQL server database is taken as source in the Visual studio analytics project
4. Attribute relationships have been mapped for the respective data-source views
5. Cube has been developed based on the selected dimensions of Weather, City, Bump, County, Sunrise-Sunset, Signal, Junction, Loop, Wind etc

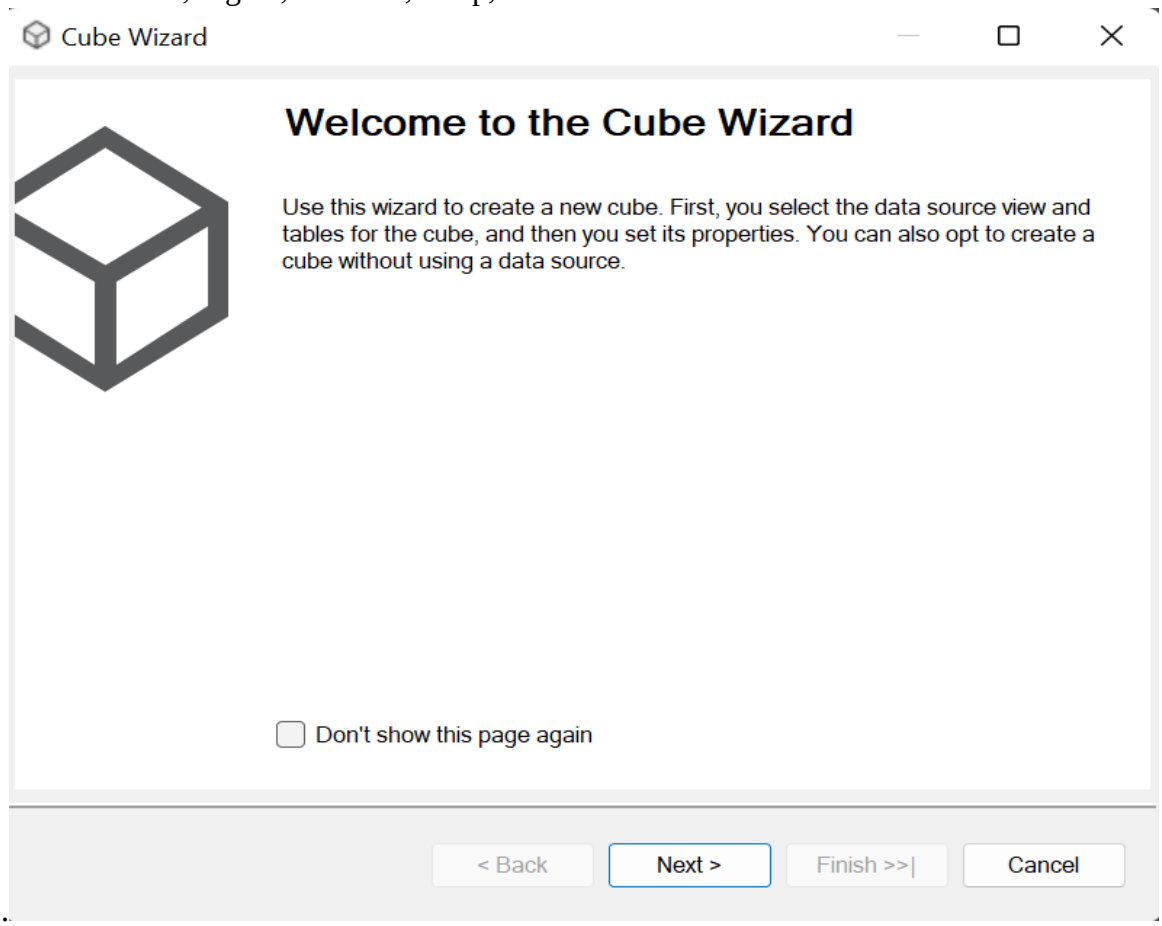


Fig. 17-1: Open Cube Wizard

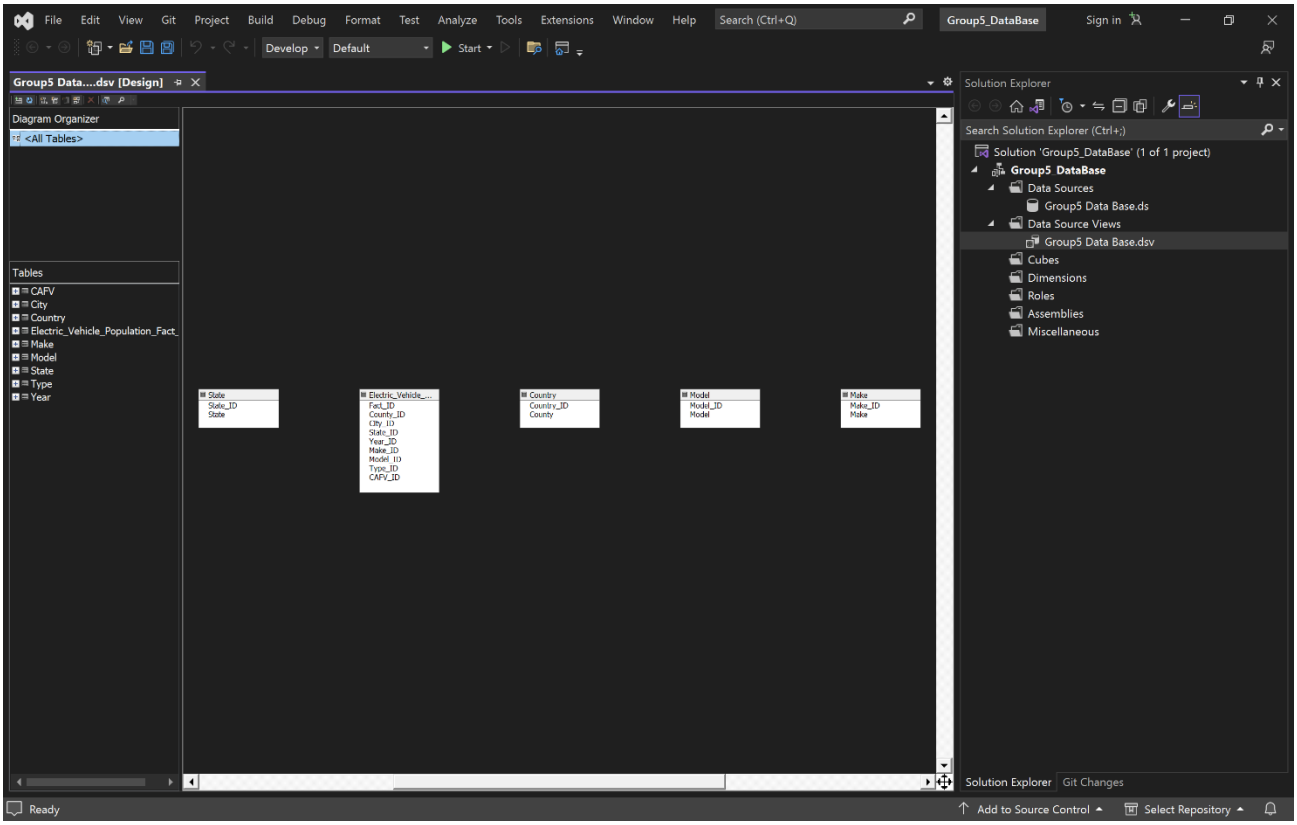


Fig. 17-2: Import Tables

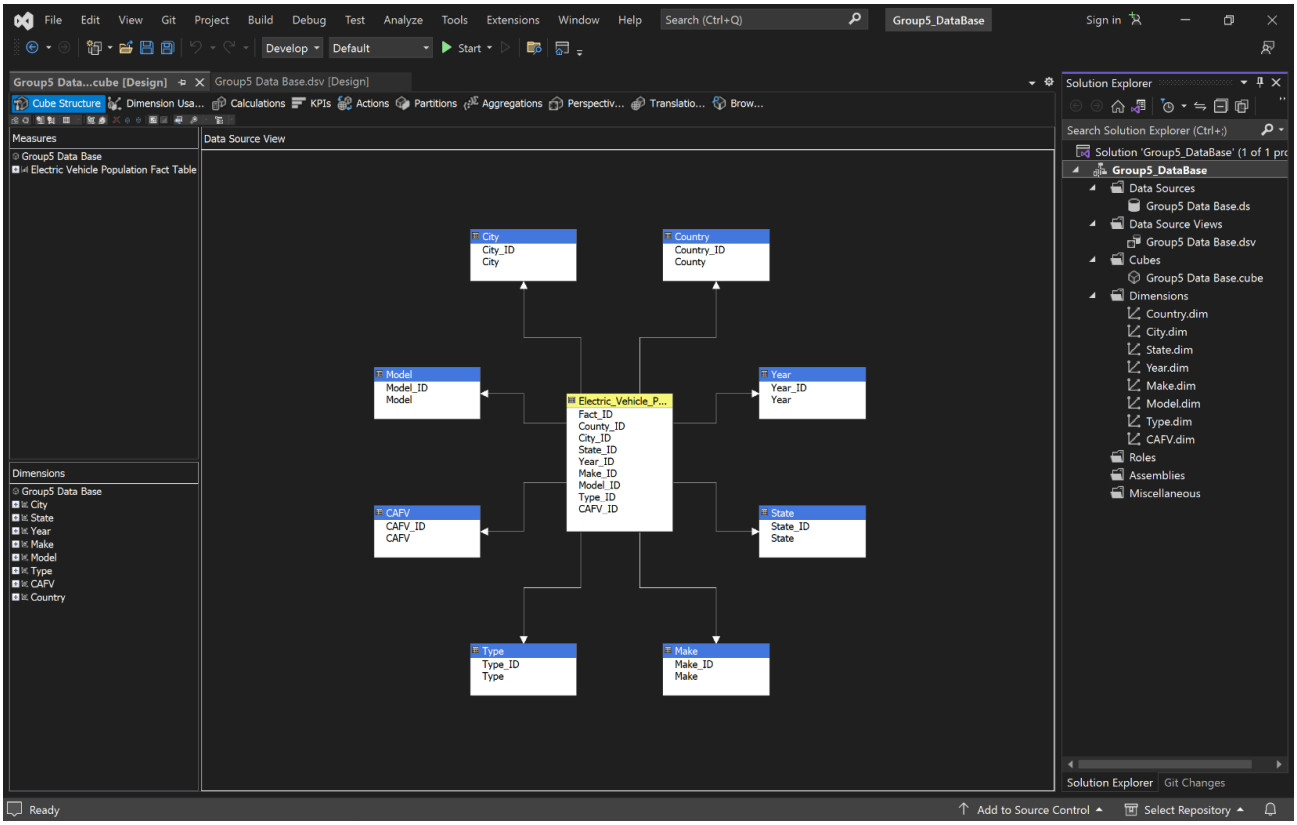


Fig. 17-3: Create Relationships

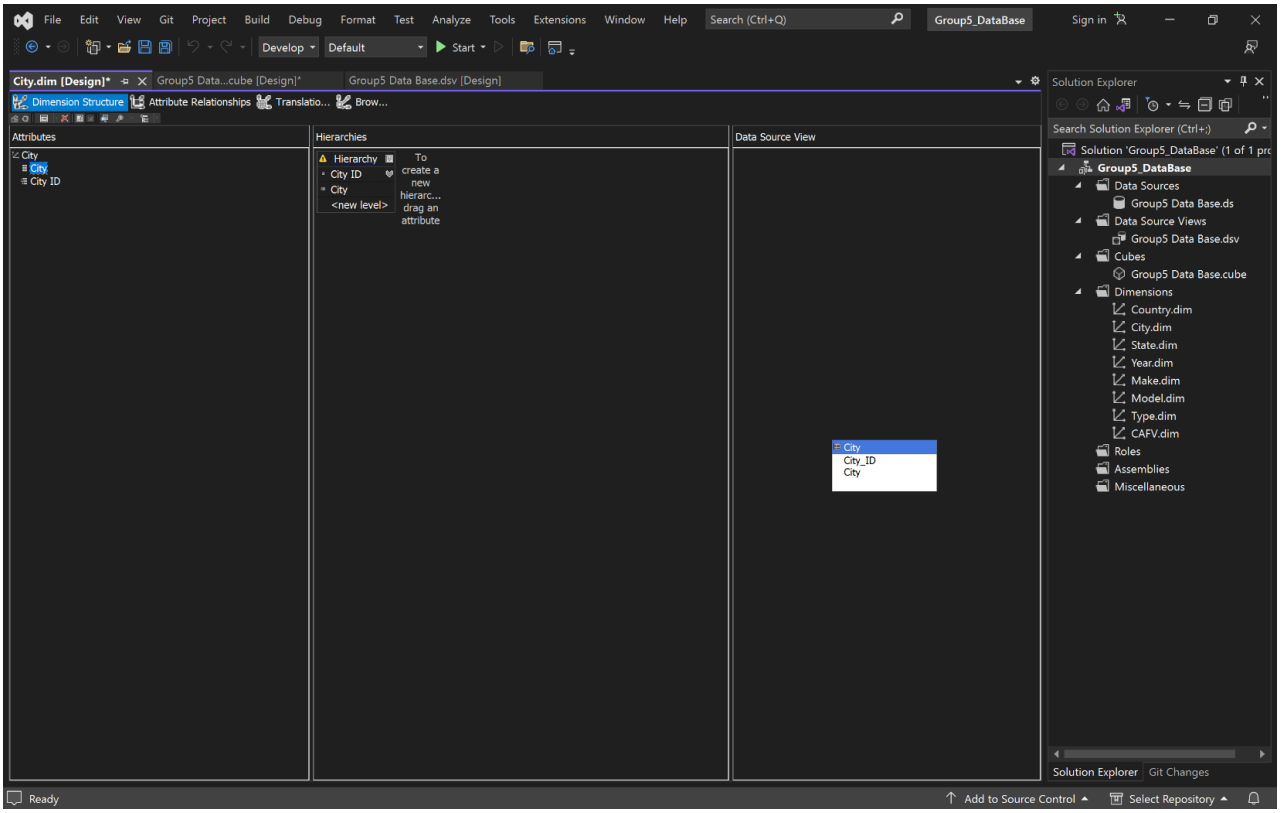


Fig. 17-4: Create Hierarchy

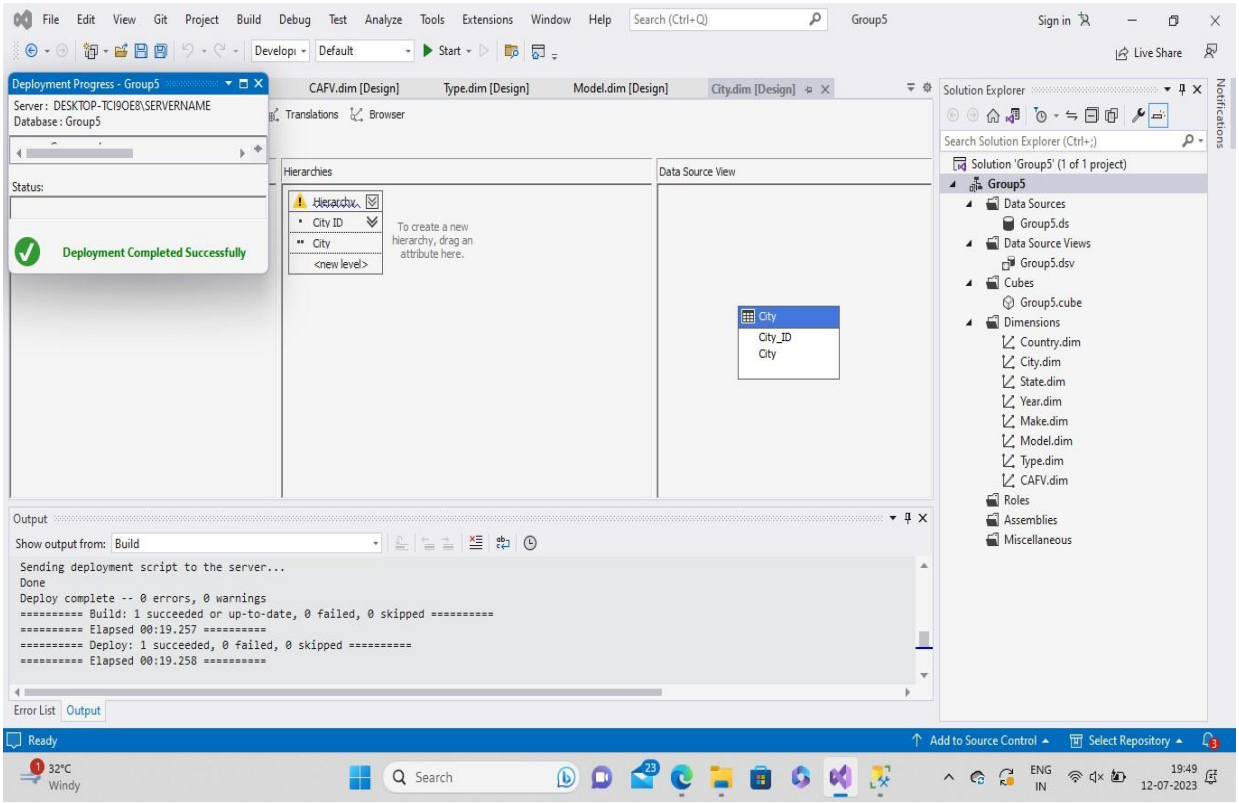


Fig. 17-5: Build and Deploy

Reports:

Analysis 1:

Among the five cities, Bellevue, Redmond, and Kirkland are comparatively close to Seattle where is known as the largest city in Washington. Whereas cities like Vancouver are known for their high population which contributes to the electric Vehicle users.

CITY	NUMBER
Seattle	19851
Bellevue	5705
Vancouver	4043
Redmond	4032
Kirkland	3507

Fig. 18-1: Top 5 Cities Table

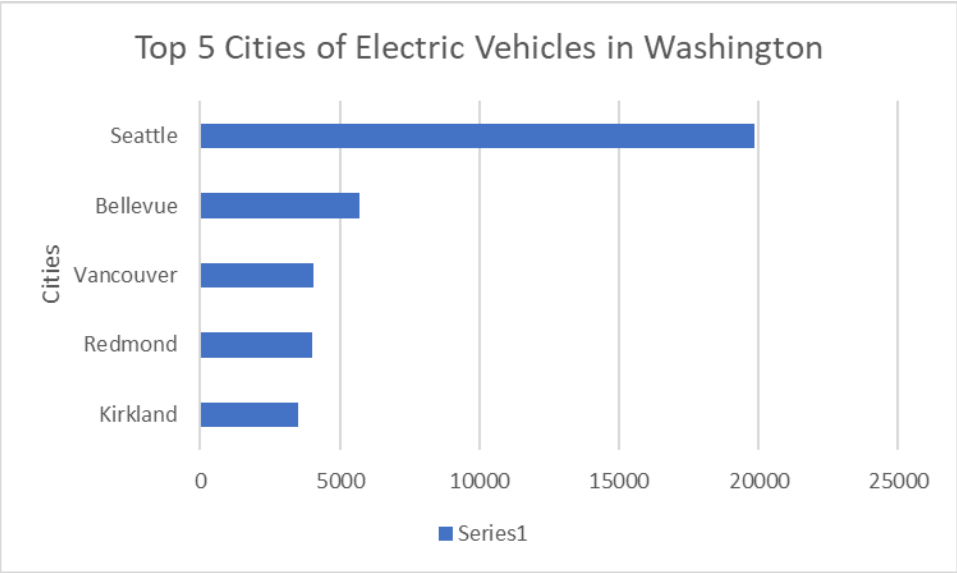


Fig. 18-2: Top 5 Cities Chart

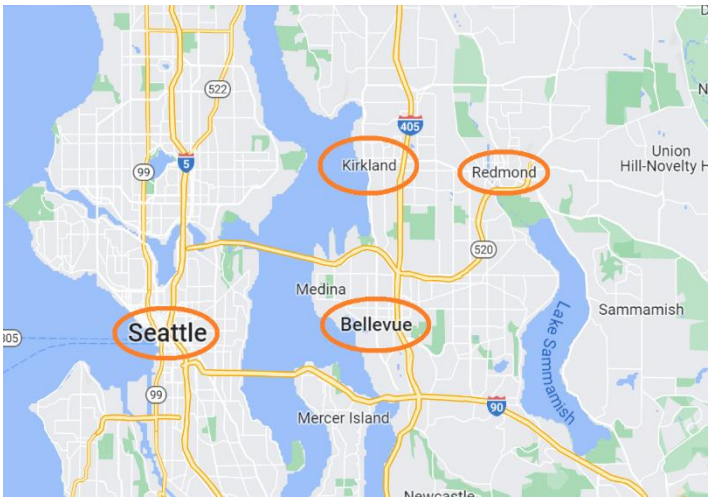


Fig. 18-3: The Cities Close to Seattle

Analysis 2:

The graph below provides us with the information that how electric vehicle population started booming from 1997 to 2022. In the initial years, that is from 1997 to 2012, we can see that there were barely any electric vehicles. We can see from the graph that the number of electric vehicles gradually increased from 2016 to 2022. Where 2022 is recorded as the highest count of electric vehicles, and the number of electric vehicles in 2022 is about 60% increased of the number in 2021.

YEAR	NUMBER
1997	1
1998	1
1999	4
2000	9
2002	2
2003	1
2008	21
2010	24
2011	828
2012	1668
2013	4581
2014	3609
2015	4935
2016	5702
2017	8558
2018	14224
2019	10449
2020	10926
2021	18296
2022	27522

Fig. 19-1: The Number of Electric Vehicles Table

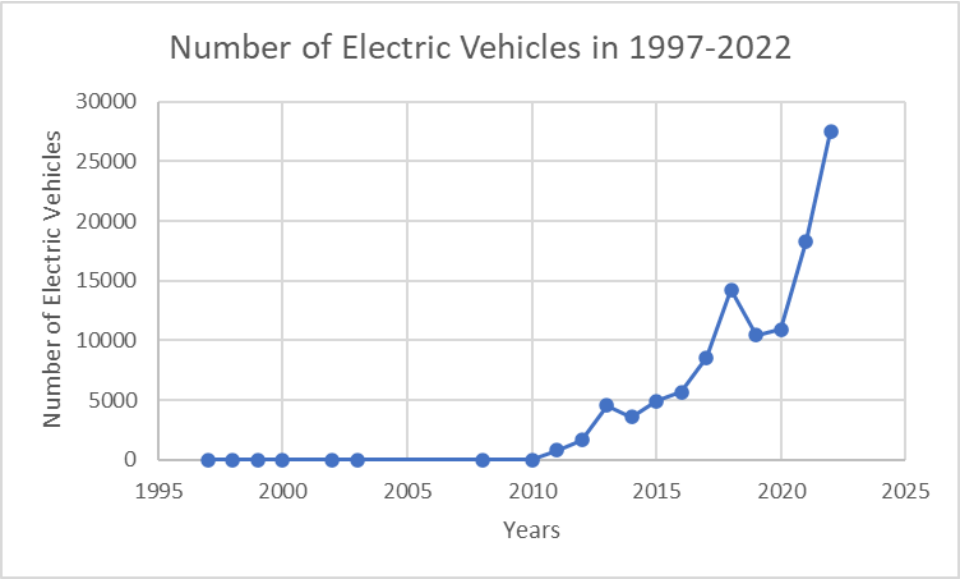


Fig. 19-2: The Number of Electric Vehicles Chart

Analysis 3:

The electric vehicles have been categorized as battery electric and plug-in hybrid electric. The below graph provides us a clear view of the percentages of plug-in hybrid electric vehicles and battery Electric vehicles. The statistical information gives an idea that around 76 % of the buyers are interested in purchasing the battery electric vehicles.

TYPE	NUMBER
Battery Electric Vehicle (BEV)	84411
Plug-in Hybrid Electric Vehicle (PHEV)	26950

Fig. 20-1: Two Type of Electric Vehicles

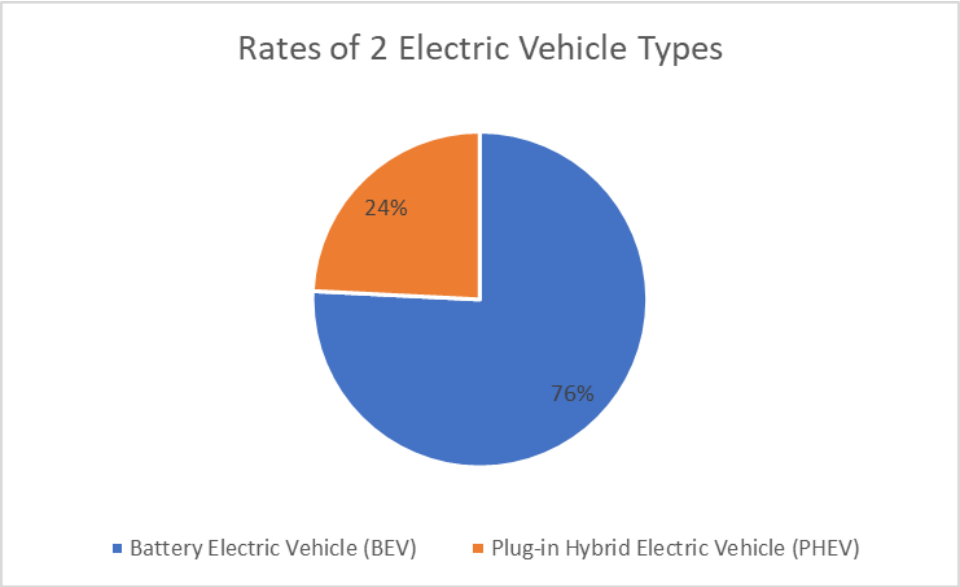


Fig. 20-2: The Rate of Two Type of Electric Vehicles

Battery Electric Vehicle (BEV):

Battery Electric Vehicles are fully electric vehicles that rely only on battery power. Since the vehicles do not produce tailpipe emissions, they are said to be friendly to the environment.

YEAR	NUMBER
1997	1
1998	1
1999	4
2000	9
2002	2
2003	1
2008	21
2010	21
2011	753
2012	798
2013	2936
2014	1805
2015	3612
2016	3891
2017	4458
2018	9946
2019	8575
2020	9315
2021	14755
2022	23507

Fig. 20-3: The Number of Battery Electric Vehicles Table

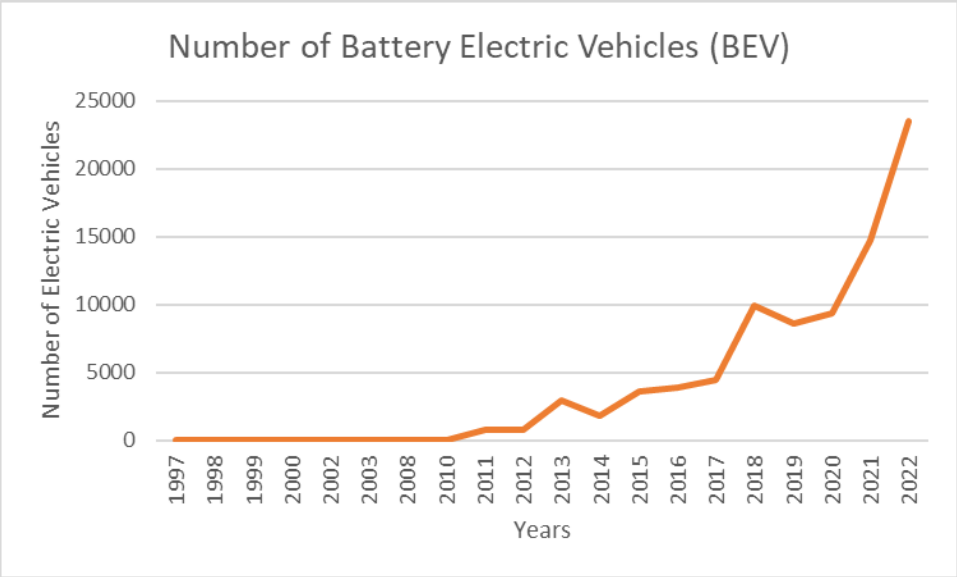


Fig. 20-4: The Number of Battery Electric Vehicles Chart

Plug-in Hybrid Electric Vehicle (PHEV):

Plug-in Hybrid Electric Vehicles combine electric motors with internal engines, and the vehicles allow to be charged by power sources. Drivers can use electric motors to drive for short distances, and then use engines to drive for longer distances.

YEAR	NUMBER
2010	3
2011	75
2012	870
2013	1645
2014	1804
2015	1323
2016	1811
2017	4100
2018	4278
2019	1874
2020	1611
2021	3541
2022	4015

Fig. 20-5: The Number of Plug-in Hybrid Electric Vehicles Table

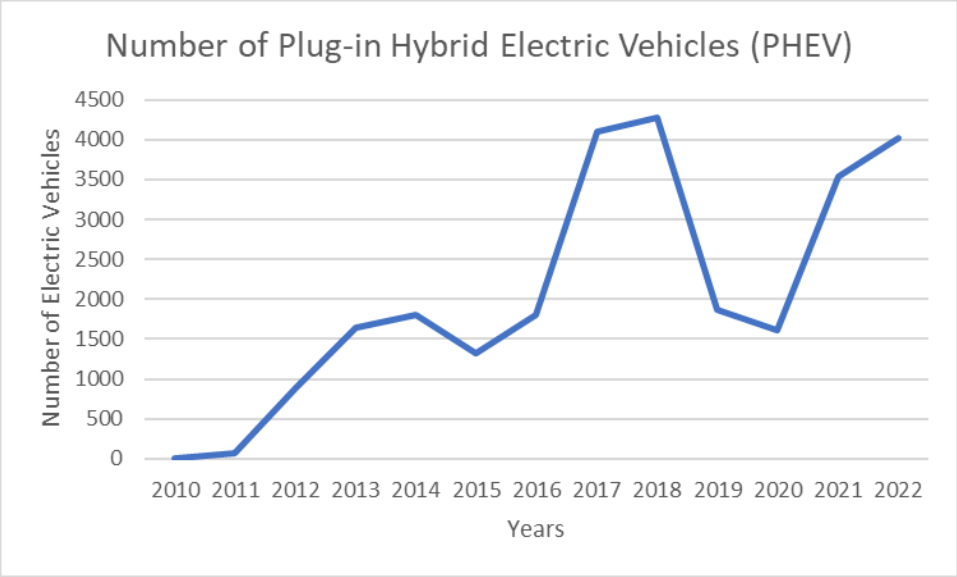


Fig. 20-6: The Number of Plug-in Hybrid Electric Vehicles Chart

Analysis 4:

The below graph describes the top 15 makes of electric vehicles. The statistical information below gives us a conclusion that most buyers purchased the electric vehicles made of Tesla.

MAKE	NUMBER
TESLA	50772
NISSAN	12542
CHEVROLET	9727
FORD	6238
BMW	4588
TOYOTA	4542
KIA	4324
VOLKSWAGEN	2736
AUDI	2344
VOLVO	2223
CHRYSLER	1889
JEEP	1343
RIVIAN	1265
HYUNDAI	1249
PORSCHE	856

Fig. 21-1: Top 15 Makes Table

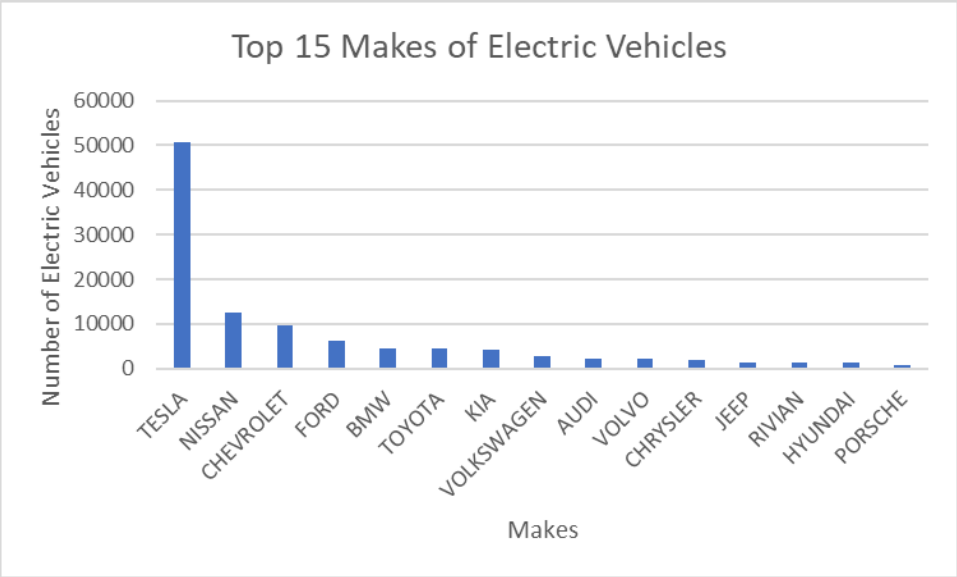


Fig. 21-2: Top 15 Makes Chart

Analysis 5:

The below graph provides various models where Model 3 is recorded as the highest number of electric vehicles. It is worth to mention that Model 3, Model Y, and Model S are made of Tesla.

MODEL	NUMBER
MODEL 3	22673
MODEL Y	16512
LEAF	12542
MODEL S	7266
VOLT	4888

Fig. 22-1: Top 5 Models Table

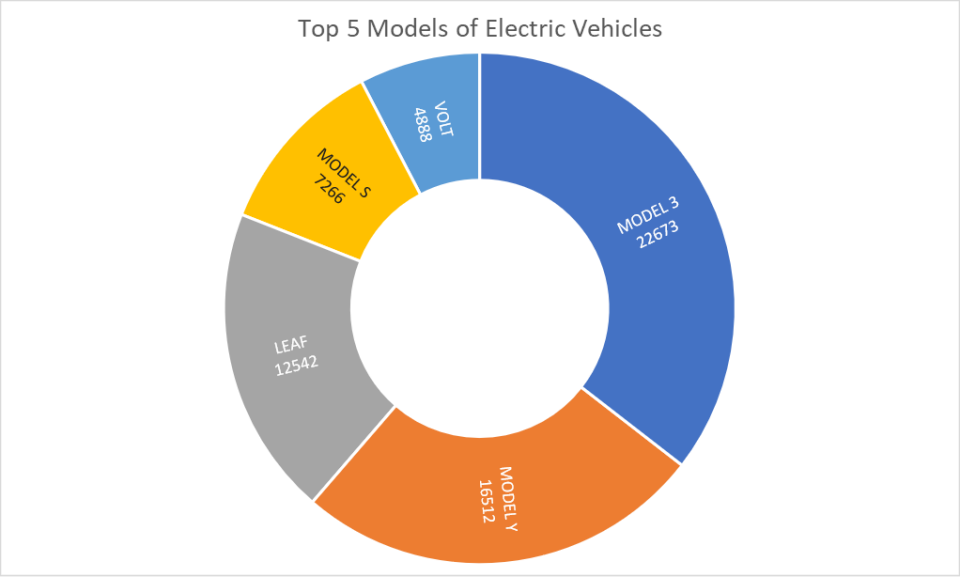


Fig. 22-2: Top 5 Models Chart

Analysis 6:

The below graph provides a detailed rates of clean Alternative Fuel eligibility. The statistical information gives us a conclusion that approximately half of the electric vehicles purchased have a clean alternative Fuel Eligibility.

CAFV	NUMBER
Clean Alternative Fuel Vehicle Eligible	58581
Eligibility unknown as battery range has not been researched	37832
Not eligible due to low battery range	14948

Fig. 23-1: Clean Alternative Fuel Vehicle Eligible Table

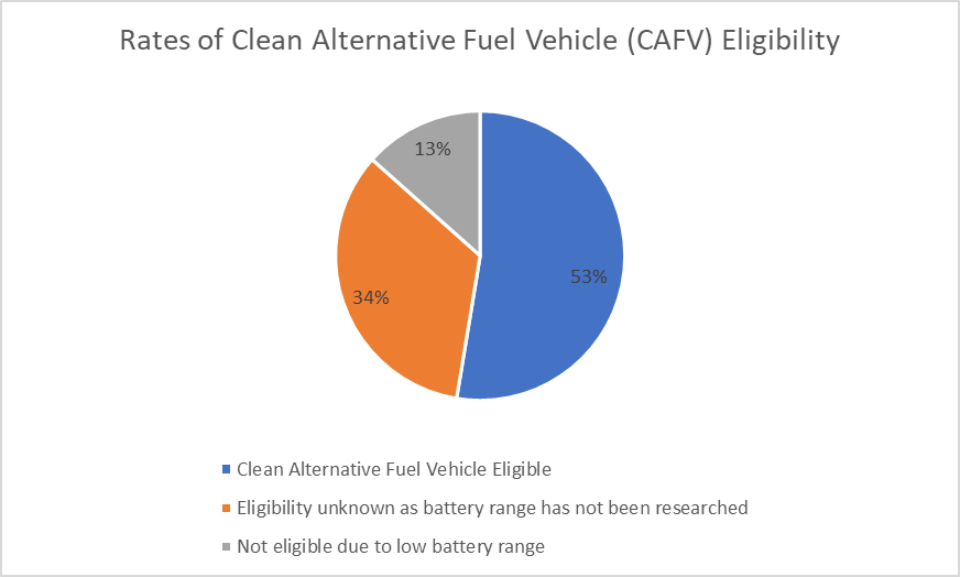


Fig. 23-2: Clean Alternative Fuel Vehicle Eligible Chart

Data Mining Queries

Data Mining Queries can be useful for below purposes according to Microsoft-

· Apply the model to new data, to make single or multiple predictions. You can provide input values as parameters, or in a batch.

· Get a statistical summary of the data used for training.

· Extract patterns and rules, or generate a profile of the typical case representing a pattern in the model.

· Extract regression formulas and other calculations that explain patterns.

· Get the cases that fit a particular pattern.

· Retrieve details about individual cases used in the model, including data not used in analysis

Query 1 - City:

SELECT TOP 5 CITY, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN STATE S

ON E.State_ID = S.State_ID

JOIN CITY C

ON E.City_ID = C.City_ID

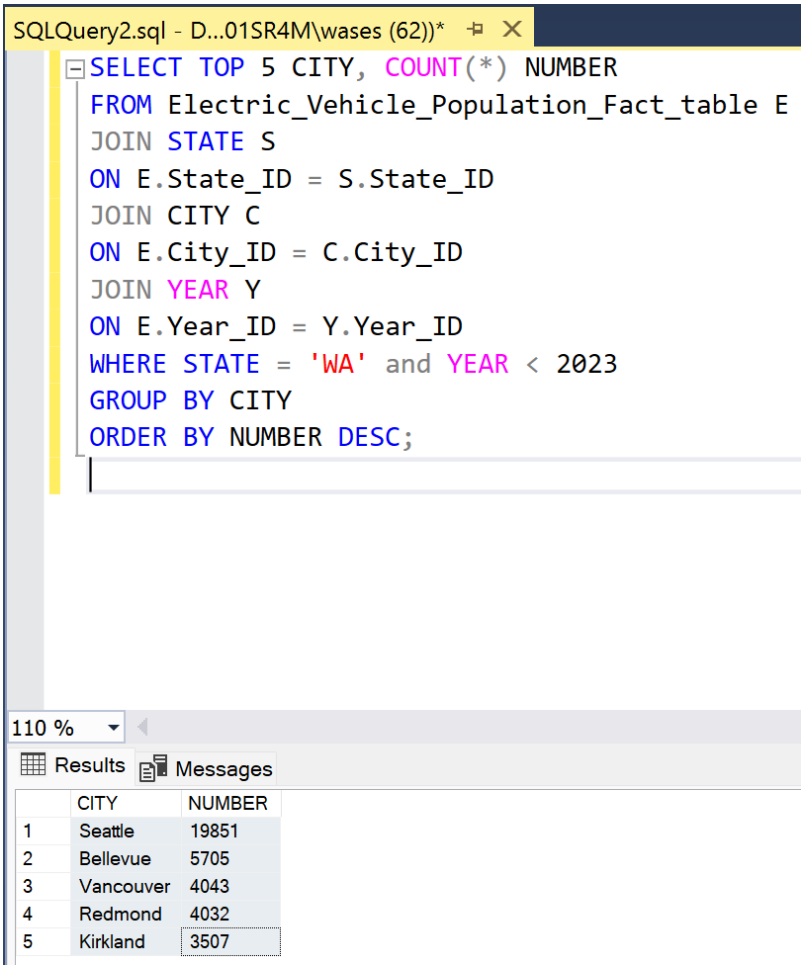
JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE STATE = 'WA' AND YEAR < 2023

GROUP BY CITY

ORDER BY NUMBER DESC;



	CITY	NUMBER
1	Seattle	19851
2	Bellevue	5705
3	Vancouver	4043
4	Redmond	4032
5	Kirkland	3507

Fig. 24-1: The Query of City

Query 2 - Year:

SELECT YEAR, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

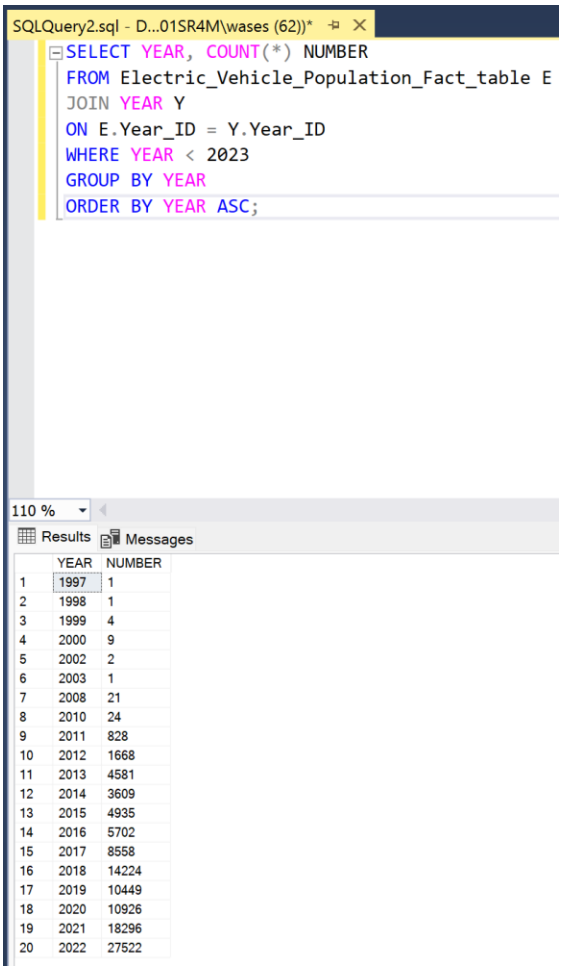
JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE YEAR < 2023

GROUP BY YEAR

ORDER BY YEAR ASC;



	YEAR	NUMBER
1	1997	1
2	1998	1
3	1999	4
4	2000	9
5	2002	2
6	2003	1
7	2008	21
8	2010	24
9	2011	828
10	2012	1668
11	2013	4581
12	2014	3609
13	2015	4935
14	2016	5702
15	2017	8558
16	2018	14224
17	2019	10449
18	2020	10926
19	2021	18296
20	2022	27522

Fig. 24-2: The Query of Year

Query 3 - Type:

SELECT TYPE, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN TYPE T

ON E.Type_ID = T.Type_ID

JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE YEAR < 2023

GROUP BY TYPE;

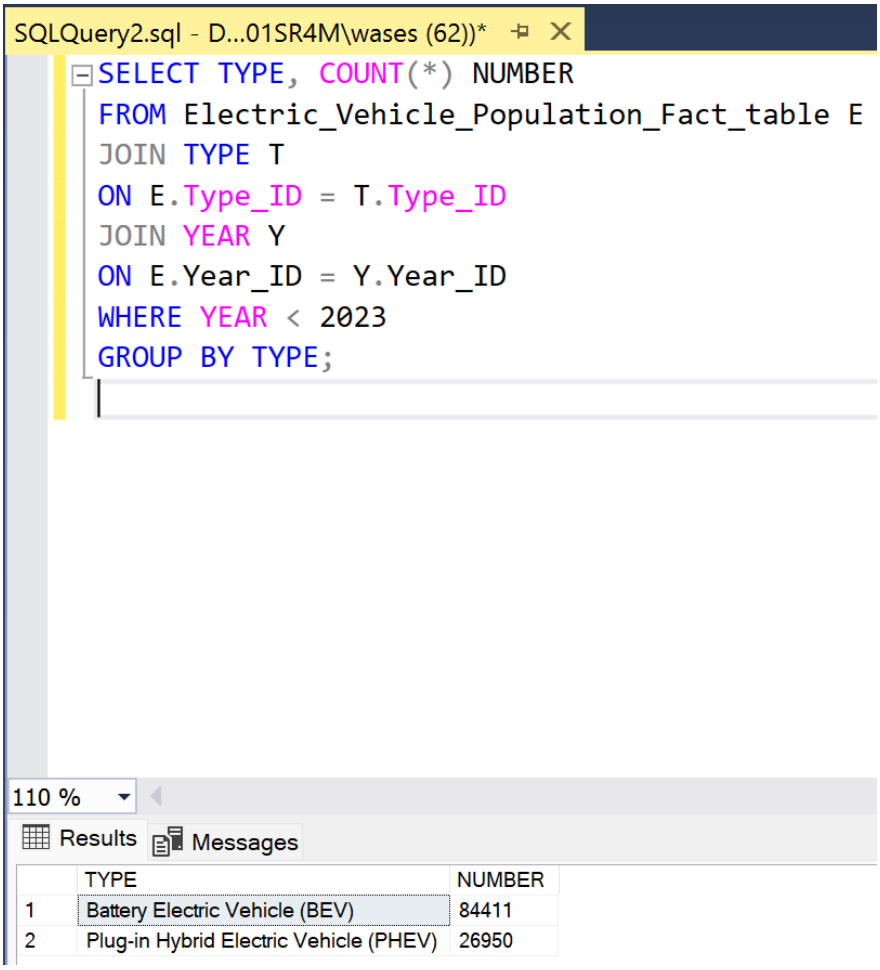


Fig. 24-3: The Query of Type

Battery Electric Vehicle (BEV):

SELECT YEAR, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN TYPE T

ON E.Type_ID = T.Type_ID

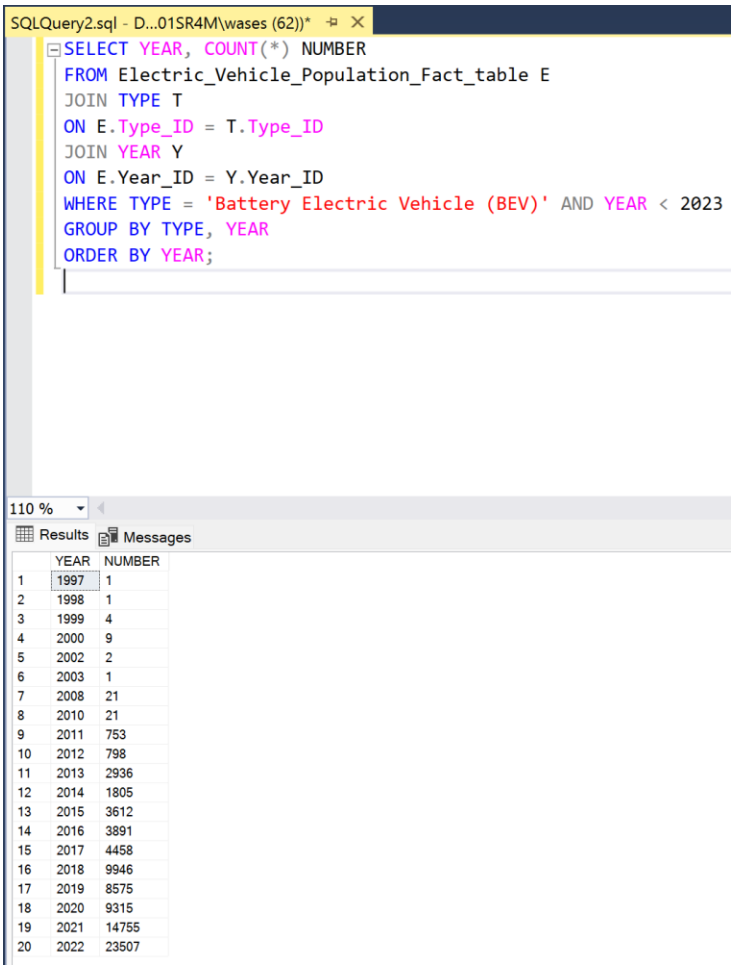
JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE TYPE = 'Battery Electric Vehicle (BEV)' AND YEAR < 2023

GROUP BY TYPE, YEAR

ORDER BY YEAR;



	YEAR	NUMBER
1	1997	1
2	1998	1
3	1999	4
4	2000	9
5	2002	2
6	2003	1
7	2008	21
8	2010	21
9	2011	753
10	2012	798
11	2013	2936
12	2014	1805
13	2015	3612
14	2016	3891
15	2017	4458
16	2018	9946
17	2019	8575
18	2020	9315
19	2021	14755
20	2022	23507

Fig. 24-4: The Query of Type - BEV

Plug-in Hybrid Electric Vehicle (PHEV):

SELECT YEAR, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN TYPE T

ON E.Type_ID = T.Type_ID

JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE TYPE = 'Plug-in Hybrid Electric Vehicle (PHEV)' AND YEAR < 2023

GROUP BY TYPE, YEAR

ORDER BY YEAR;

SQLQuery2.sql - D...01SR4M\wases (62))*

```sql
SELECT YEAR, COUNT(*) NUMBER
FROM Electric_Vehicle_Population_Fact_table E
JOIN TYPE T
ON E.Type_ID = T.Type_ID
JOIN YEAR Y
ON E.Year_ID = Y.Year_ID
WHERE TYPE = 'Plug-in Hybrid Electric Vehicle (PHEV)' AND YEAR < 2023
GROUP BY TYPE, YEAR
ORDER BY YEAR;
```

110 %

Results | Messages

	YEAR	NUMBER
1	2010	3
2	2011	75
3	2012	870
4	2013	1645
5	2014	1804
6	2015	1323
7	2016	1811
8	2017	4100
9	2018	4278
10	2019	1874
11	2020	1611
12	2021	3541
13	2022	4015

Fig. 24-5: The Query of Type - PHEV

Query 4 - Make:

SELECT TOP 15 MAKE, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN MAKE M

ON E.Make_ID = M.Make_ID

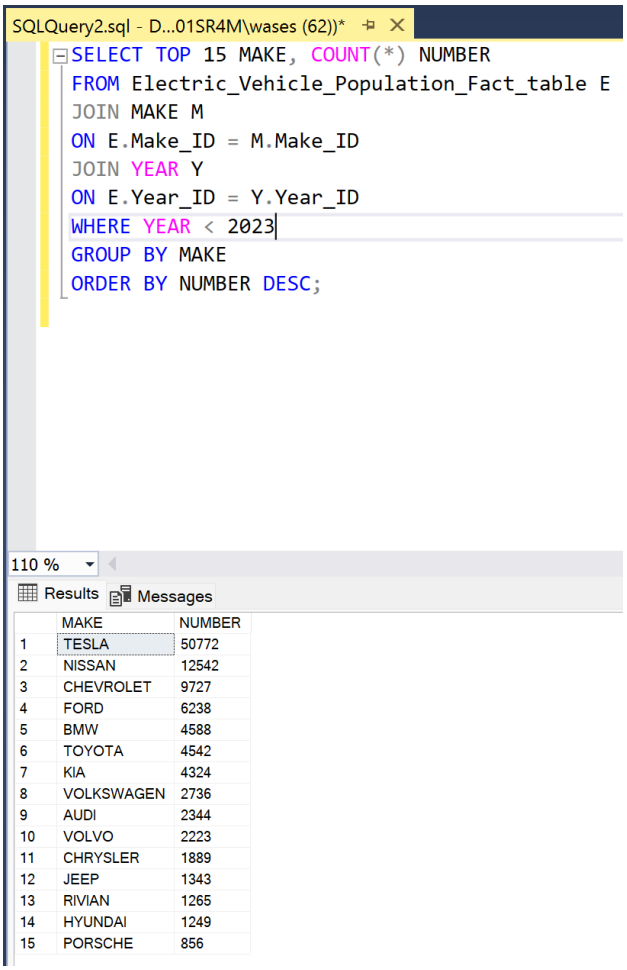
JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE YEAR < 2023

GROUP BY MAKE

ORDER BY NUMBER DESC;

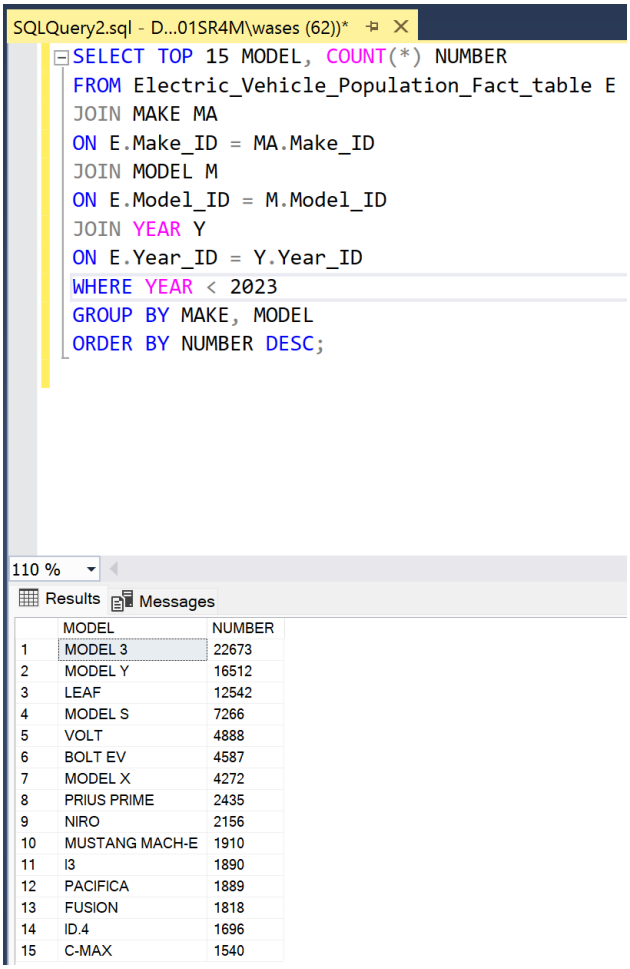


	MAKE	NUMBER
1	TESLA	50772
2	NISSAN	12542
3	CHEVROLET	9727
4	FORD	6238
5	BMW	4588
6	TOYOTA	4542
7	KIA	4324
8	VOLKSWAGEN	2736
9	AUDI	2344
10	VOLVO	2223
11	CHRYSLER	1889
12	JEEP	1343
13	RIVIAN	1265
14	HYUNDAI	1249
15	PORSCHE	856

Fig. 24-6: The Query of Make

Query 5 - Model:

```
SELECT TOP 5 MODEL, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN MAKE MA

ON E.Make_ID = MA.Make_ID

JOIN MODEL M

ON E.Model_ID = M.Model_ID

JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE YEAR < 2023

GROUP BY MAKE, MODEL

ORDER BY NUMBER DESC;
```

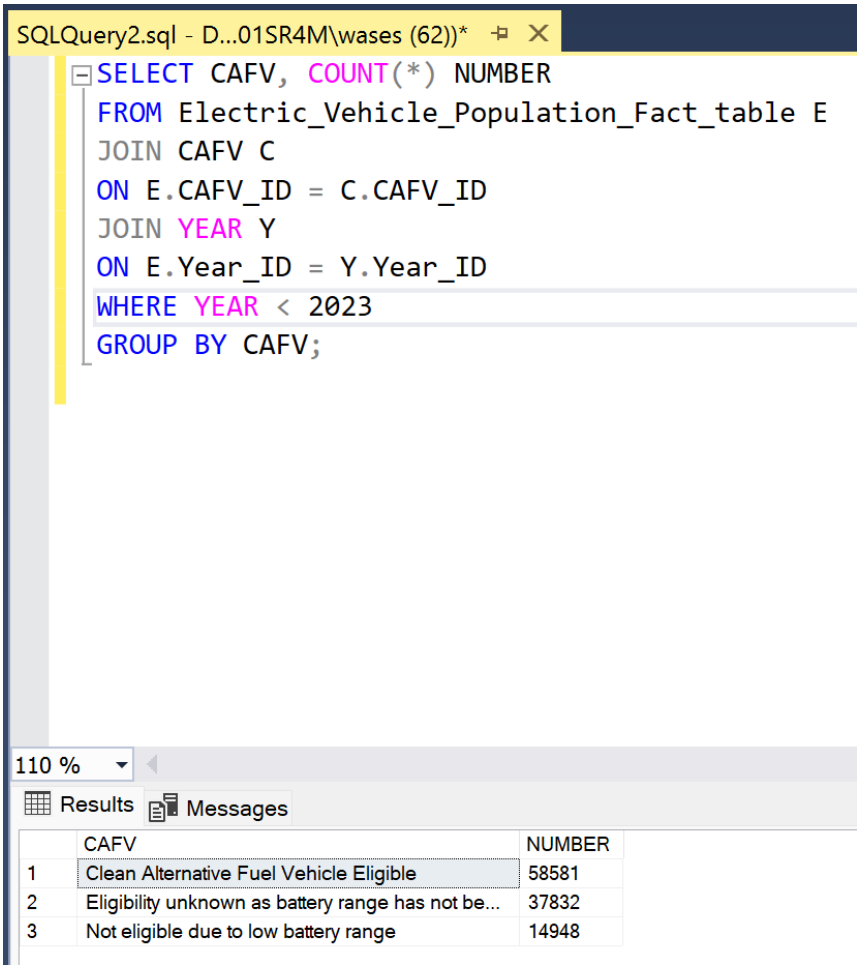


	MODEL	NUMBER
1	MODEL 3	22673
2	MODEL Y	16512
3	LEAF	12542
4	MODEL S	7266
5	VOLT	4888
6	BOLT EV	4587
7	MODEL X	4272
8	PRIUS PRIME	2435
9	NIRO	2156
10	MUSTANG MACH-E	1910
11	I3	1890
12	PACIFICA	1889
13	FUSION	1818
14	ID.4	1696
15	C-MAX	1540

Fig. 24-7: The Query of Model

Query 6 - CAFV:

```
SELECT CAFV, COUNT(*) NUMBER

FROM Electric_Vehicle_Population_Fact_table E

JOIN CAFV C

ON E.CAFV_ID = C.CAFV_ID

JOIN YEAR Y

ON E.Year_ID = Y.Year_ID

WHERE YEAR < 2023

GROUP BY CAFV;
```



	CAFV	NUMBER
1	Clean Alternative Fuel Vehicle Eligible	58581
2	Eligibility unknown as battery range has not be...	37832
3	Not eligible due to low battery range	14948

Fig. 24-8: The Query of Clean Alternative Fuel Vehicle Eligible

Conclusion:

The dataset we used for this project was posted on Kaggle.com and provides the annual production statistics for electric vehicles for the period 1997–2023. There are several columns that support the production of electric vehicles, including the type, model, and make and year.

Based on our research, we conclude that the majority of electric vehicle production occurred between 2015 and 2022. According to our research, Seattle has the most electric vehicles. About 70 percent of the electric vehicles produced are battery-powered, and around 50 percent of them are classified as clean Alternative Fuel Vehicles.

We also saw that Tesla produced a large number of different models of electric vehicles. Among all electric vehicles, the Model 3 is the most popular.

References:

https://www.kaggle.com/datasets/utkarshx27/electric-vehicle-population-data

https://www.google.com/maps

https://worldpopulationreview.com/states/cities/washington