Group 7

Reem Saleh, Hadi Serag Eldin, Christian Janssen, Aisha Almehairi, Peter Filardi

INST327

Final Project Submission

Final Project Report

Introduction:

Our group designed a database based on the electric vehicle population dataset given to us. That dataset had 17 columns and over 100,000 rows. Because of the vast amount of data that we were working with, our group decided to focus on a few different elements including make, model, CAFV (clean alternative fuel vehicle eligibility), and location. We chose to work with the electric vehicle population dataset because of its real-world relevance and importance. It is that same real-world relevance and importance that we sought to keep as we developed our database and created queries. Nowadays, we are seeing a rise in the popularity of electric vehicles due to concerns related to climate change and decreasing carbon emissions.

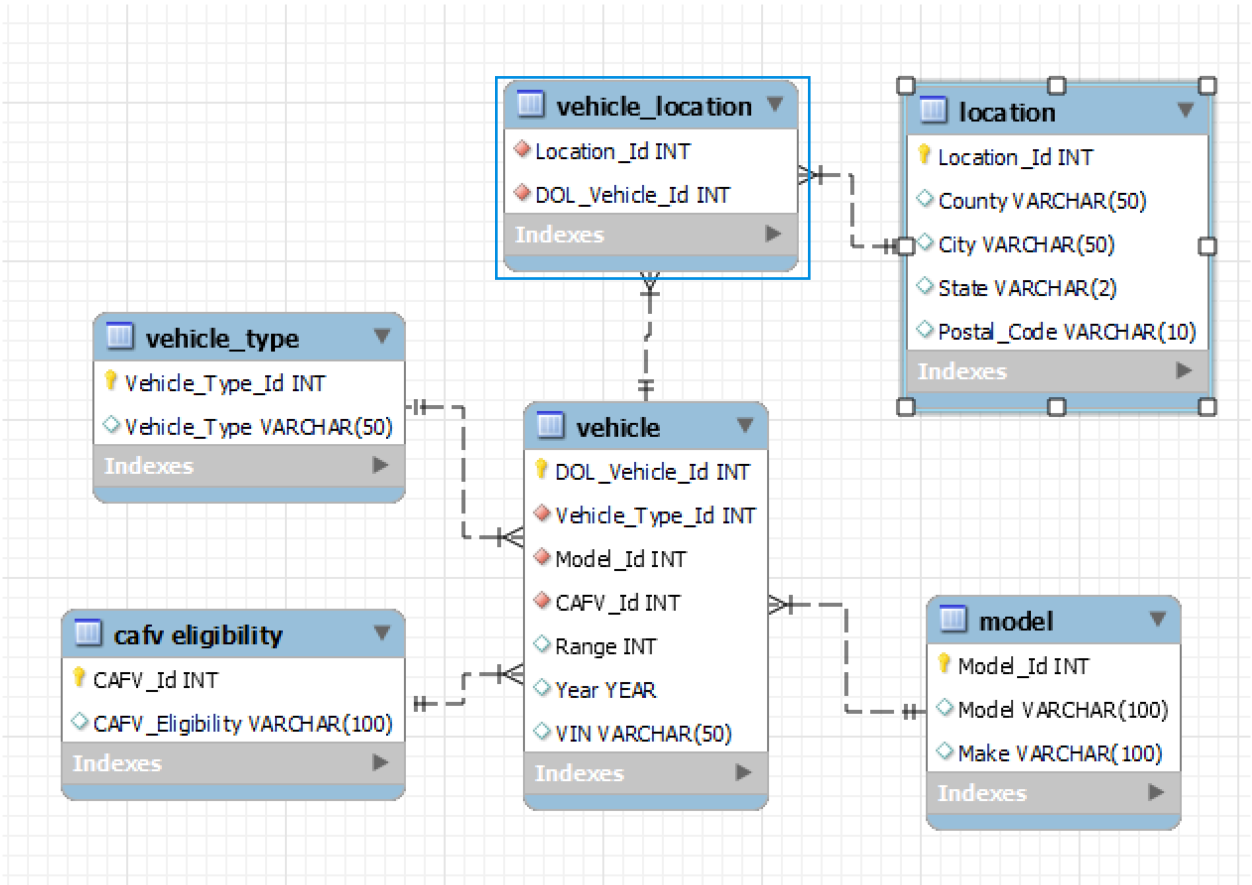
Through the creation of our database, there are several factors we want to examine. We want to track the popularity of electric vehicles by location, specifically cities in Washington state. By knowing which cities have the highest number of electric vehicles, we can then check the socioeconomic demographics in those areas and create a discussion on ethics. Another factor we want to check is the CAFV eligibility of a vehicle to see if electric vehicles are indeed better for the environment. There are a lot of concerns now with climate change, so the database will show if the state of Washington is trending in the direction of becoming more environmentally friendly. We are keeping a broad focus for our database in terms of vehicle make and model. Even though we are still focusing primarily on Tesla since it is the most popular electric vehicle, we are also adding in other electric vehicle manufacturers including Kia and Mini to contrast and compare between the different vehicles and their manufacturers. That will allow us to check popularity trends for each vehicle type.

Our database will be useful for car lovers, customers who are considering purchasing an environmentally friendly vehicle, and car manufacturers alike. Our database will examine many factors that are applicable to each of those groups. Car manufacturers may be interested in analyzing the trends with electric vehicles. Consumers and car lovers might be most interested in knowing if a particular vehicle is CAFV eligible both for the tax credit and the environmental benefits. Our database will be able to answer questions many people have about electric vehicles.

Database Description:

The focus of our database is on electric vehicles, the growth of their population overtime, and the areas with the highest EV population. Because Tesla is usually the most popular electric vehicle people purchase, Tesla models are the primary vehicle we are examining. However, in order to have a broader scope in our database and create comparisons between different electric vehicles, we are also including cars made by other manufacturers including Kia and Mini. In terms of area, Washington state is the focus of our database.

Logical Design:



Following the steps of logical design, we defined the structure and organized the data in our database. We created our Entity-Relationship Diagram (ERD) based on several different principles of logical design. We formed our final ERD after going through the normalization process to remove data redundancy and ensure data integrity. We felt that having a lot of tables might be distracting for our database users, and we also didn't want to have too broad of a scope, so we created six tables for our database. The tables represent the elements we wanted to focus on, which include make, model, CAFV eligibility, location, and information pertaining to the vehicle like the year it was manufactured. Our vehicle_location table is the linking table, which connects the vehicle and location tables. The linking table connects the two tables using the Location_ID and DOL_Vehicle_ID.

There were a lot of elements we decided to keep together and combine them into one table. An example of that is our model table, which has a Model ID (INT), Model name (Varchar 100), and Make (Varchar 100). The reason we combined make and model information together is because it will make it easier for database users and viewers to examine the population of electric vehicles by the specific vehicle itself and the vehicle manufacturer as well. It also has the benefit of removing a transitive dependency. The data type we used for the IDs is INT. The IDs served as our primary keys for each table. As a constraint, we put NOT NULL for all primary keys. To simplify the logical design, we used VARCHAR for most of the information in our tables.

We used consistent naming conventions. Every table was named after the information that resides within it so that it is clear to the database users and easy to find the information they need. The vehicle_type, cafv eligibility, and the model table all have a one-to-many relationship with the vehicle table. The location table has a one-to-many relationship with the vehicle table via the linking table vehicle_location.

Physical Design:

We used the electric vehicles population dataset to construct our database. That dataset had 17 columns and over 100,000 rows. We decided to narrow down the data that we used to 80 rows, including a lot of Tesla vehicles, and vehicles that were located in Washington state. Our database has six tables, and the reason is so that it does not become too convoluted with unnecessary information. We kept our focus on the essential aspects of electric vehicles. Our target audience includes car lovers and manufacturers alike, so we wanted to make sure our database can answer the questions they might have and offer important insights. We created a vehicle_type table that records what type of EV a vehicle is categorized as, either a Plug-in Hybrid Electric Vehicle (PHEV) or a Battery Electric Vehicle (BEV). The vehicle_location table is a linking table that joins together the vehicle table and location table. This table was made so that users can view a vehicle and all its location information with less clutter. The vehicle table can tell a user core information on each vehicle, including range, year produced, and model ID. The model ID connects each vehicle with the model table, which gives the vehicle's manufacturer and model. The location table has the precise address information of the vehicle. The vehicle_location table includes the Location_ID and DOL_Vehicle_ID as foreign keys, which links together the vehicle and location tables. The location table has address data including county, city, state, and zip code. The other tables we have include information pertaining to data related to the table's name. Lastly, the CAFV Eligibility table categorizes each vehicle based on if it meets the fuel requirement and electric-only range requirement needed to be eligible for Washington State tax exemptions.

Sample Data:

The data we used is sourced from a dataset of BEVs and PHEVs that are currently registered through Washington State Department of Licensing. This is one of the datasets offered through the course. The data is in the form of a CSV file with 17 columns and over 100,000 rows. To create a clear focus for our project, the data was cleaned and narrowed down to 80 rows because we wanted to have no more than 100 rows of data. We looked up the ten most populous cities in the state of Washington and selected rows that were from some of those cities, which is essential to the focus of our database. We made sure to include rows that had mainly Tesla vehicles and only selected vehicles in the state of Washington. The data was separated into one Google sheet for each table we wanted to create and imported into the database. While importing the data, a few minor changes had to be made to the ERD. The data type VARCHAR had to be changed to fit 100 characters, and there was a misspelling of a table name that needed to be corrected.

Views/Queries:

Queries	ReqA	ReqB	ReqC	ReqD	ReqE
View_EVMarketAnalysis	x	x	x		
View_EVDetailedRegistrations	x	x			
View_VehicleTypeComparison	x	x	x	x	

View_TeslaModel Analysis	x	x			x
View_RegionalEV Adoption	x	x			

We created five queries saved as views for our database. While our original questions have changed some since we made our original proposal, we were able to answer many of those original questions as well as some new ones. Although a few of our questions could not be answered due to the limits of the scope of the dataset. The questions we had about EVs outside of Washington could not be answered because the dataset had little data for out-of-state vehicles. Additionally, the eleventh question of our proposal, which was, which county has the highest average income of electric vehicle owners and is there a correlation between income level and electric vehicle adoption, could not be answered. This is because our dataset does not include income information for different areas in Washington, and this would have required us to do outside research and analysis. Outside of these questions, we were successful in finding answers applicable to our dataset. The questions are representative of what users might be curious about related to electric vehicles. The query/view names and the questions they can be used to answer are as follows:

view_evmarketanalysis:

- Who are the top three electric vehicle manufacturers by market share in the state of Maryland?
- Which company is dominating in the manufacturing of electric vehicles?

view_evdetailedregistrations:

- What is the total number of electric vehicles registered in California, and how has this number changed over the past five years?
- How has the electric vehicle population grown over time?

view_vehicletypecomparison:

- Which city in Washington has the highest percentage of battery electric vehicles (BEVs) compared to plug-in hybrid electric vehicles (PHEVs)?
- How many electric vehicles in the state of Washington are eligible for Clean Alternative Fuel Vehicle status, and what percentage of the total EV population do they represent?

view_teslamodelanalysis:

- How many Tesla Model Y, Model X, and Model 3 vehicles are registered in Washington, and what is the average electric range for each model?
- What is the average electric range for Tesla vehicles compared to vehicles from other manufacturers, and how has this affected their market share?
- What is the average electric range for Tesla vehicles compared to vehicles from other manufacturers, and how has this affected their market share?

view_regionalevadoption:

- In which location have electric vehicle incentives or policies been most influential in driving electric vehicle adoption?
- Which company is dominating in the manufacturing of electric vehicles?

Changes from Original Design:

Our database design has changed greatly from the beginning of our project when we submitted the proposal to now, our final submission. As previously stated, some of the questions we were interested in answering evolved as we worked more closely with the dataset. But, the project also changed based on the different project deliverables and feedback we received from the TAs. We originally planned to have seven tables. Those tables include a make table, model table, vehicle type table, CAFV eligibility table, electric range table, location table, and a base table. The base table that we planned to create has since become our vehicle table, which serves as the main table in our ERD and database. The vehicle table has the same data we put into the base table. The number of tables has been reduced to six, and the tables themselves have changed a great deal. We now only have six tables, a model table, cafv_eligibility table, vehicle_type table, vehicle table, vehicle_location table, and a location table. The vehicle_location table is our linking table to connect the vehicle and location tables together. The make table is now combined with the model table. In the beginning we planned to have a separate electric range table, but that has since been included in our vehicle table since the range changes based on the vehicle. In the past we had the range as a part of the model table but realized two of the same models can have different ranges.

Database Ethics Considerations:

There are several considerations related to ethics that we considered when designing our database. We wanted to focus on the impact electric vehicles have on the environment, which influenced our decision to create a separate CAFV eligibility table. It also influenced our decision to make CAFV eligibility a focus of our questions/views. That is related to ethics and environmental justice. Another consideration we had related to ethics is whether we should include the VIN of a vehicle in our database. There is a lot of conflicting information on the internet in terms of whether a VIN is considered PII (Personally Identifiable Information). If it is considered PII, then we would omit it from our database because it would violate data privacy. However, from our search we realized a VIN is sometimes displayed on a vehicle in a visible area, and it isn't necessarily PII since it is not a unique identifier for the database, vehicle (Since the database only gives the first 10 digits of each VIN). That informed our decision to keep a VIN in our database, so that users can more easily identify a vehicle they were interested in exploring more information about. Bias may have been introduced into our database design from the data cleaning process. We had a vast amount of data from the original dataset, which included over 100,000 rows. We cleaned that up a great deal and narrowed it down to only 80 rows. Because we had to manually select which rows to keep in our database, this may have introduced bias.

Lessons Learned:

In terms of teamwork, we had difficulty with time management. We learned how important it is to prioritize this project and not leave it to the last minute. It was difficult because all of us were busy throughout the semester, but we communicated a lot through our GroupMe group chat. On the more technical side we also had some challenges. Some views did not have data included, only the columns were showing up. The evdetailedregistrations view did not have data showing up when we tried to download the views and check if they were correct. The vehicle type comparison chart showed that there were no CAFV eligible vehicles, which was not true, since our dataset showed there were some vehicles that were eligible, so we had to correct that. With data importing, we also had difficulties. We learned that the tables had to be imported in a specific order so that the foreign keys would be created before they were referenced. We were able to resolve that with help from our TA mentor, Claire.

Potential Future Work:

In the future if we were to continue working on this database, we might consider creating a bit of a broader scope by including a larger number of rows from our dataset. We also might create views that are more specific to certain counties or cities in Washington to investigate more aspects of the dataset. For example, we might create a view that checks how many CAFV eligible vehicles are in Tacoma, Washington, in order to then compare that information to a different city in Washington like Seattle. We could also create a view that checks the most common vehicle type registered by zip code.

If given the opportunity to work with this database again on a bigger scale, I think we could derive some really interesting conclusions. In an effort to keep our database accessible for the course, only 80 rows of data were used, but this is just a fraction of the size of the full dataset. With a broader scope we could incorporate a simple random sample of a few thousand rows. This would allow us to aggregate data with a much higher confidence level and make broader conclusions, while reducing bias.