Team 5: Arianna Lerner, Ellen Seong, Neonidh Singh, Noam Katz, Sean Bryant
INST327-ESG1
Final Project Submission
15 May 2023

Electric Vehicle Final Submission

Introduction

The dataset that we decided to work with and create was regarding electric vehicles. An electric vehicle uses electricity (EVs) as its primary power source, and can be further divided into two categories: Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). Electric Vehicle population data refers to the number of electric vehicles on the roads, including the number of BEVs and PHEVs.

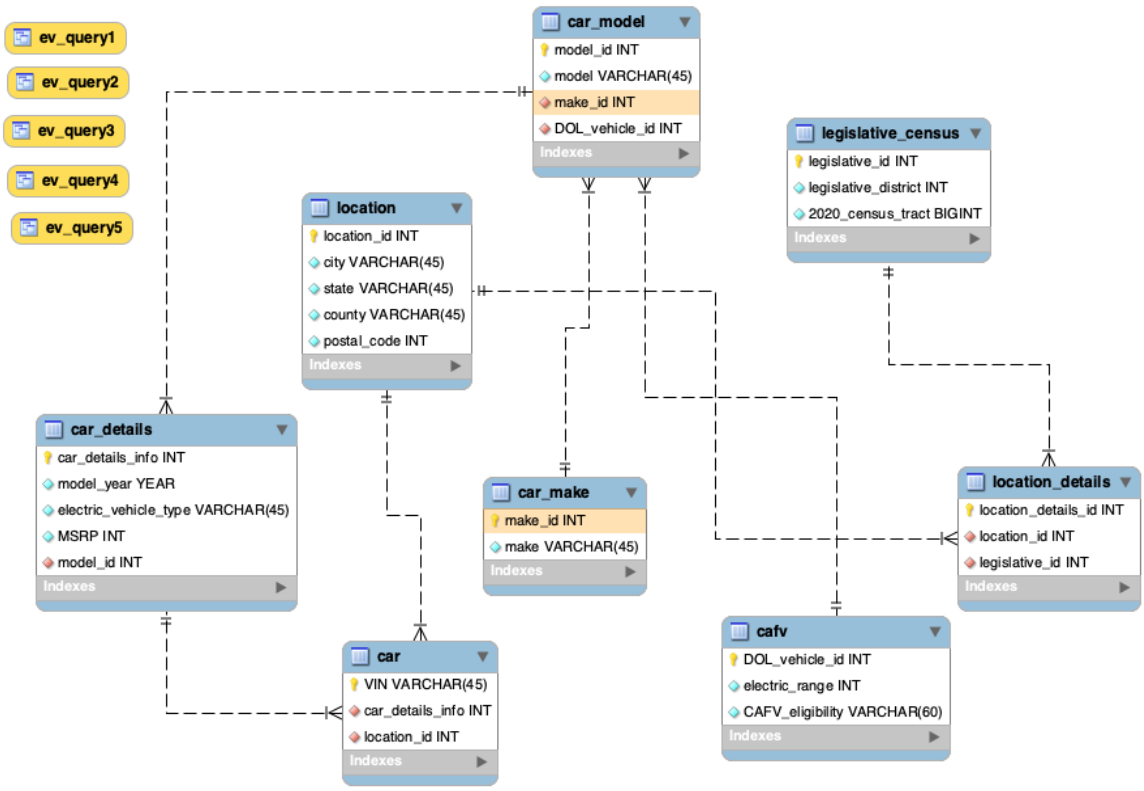
The database takes a look at electric vehicles. Specifically, it presents data on electric vehicles in multiple states such as California, New York, and Virginia. The original database includes the VIN #, address, model year, vehicle make and model, electric vehicle type, the eligibility of the vehicle for Clean Alternative Fuel Vehicle Program (tax exemption), the vehicle's electric range, the base MSRP, legislative district, vehicle location, electric utility, and information on the 2020 census tract.

For the columns that we chose to exclude as a final interpretation, we began with using only the state of California in order to scale down the size of the dataset. This decision allowed for the dataset to go from 112,635 rows to less than 100 rows. We will exclude the vehicle location column, since that information is overly specific and is redundant to us since we have good enough vehicle location information from the address columns provided.

Database Description

Logical Design

-Picture of ERD (png) on next page



Physical Database

- The creation of our physical database is done and will be attached. Our database is made up of a total of 8 tables. Our database also includes several one-to-many relationships and has 2 linking tables.

Sample Data

- We decided to have the 2020 Census Tract as a BIGINT.
- We decided to only show data for the state of California.
- We kept electric range as INT because it was a whole number and not decimals.
- We made sure that Postal Code did not include the ZF option to avoid SQL adding or removing zeros when the data was being imported.
- Model Year was put as YEAR and not INT.

Sample data from the table location: (on next page)

location_id	city	state	county	postal_code
1	Monterey	CA	Monterey	93940
2	San Diego	CA	San Diego	92101
3	Coronado	CA	San Diego	92118
4	Sacramento	CA	Sacramento	95817
5	Chula Vista	CA	San Diego	91913
6	Aliso Viejo	CA	Orange	92656
7	Roseville	CA	Placer	95747
8	Ridgecrest	CA	Kern	93555
9	Berkeley	CA	Alameda	94710
10	San Diego	CA	San Diego	92101

Views / Queries

VIEW NAME	REQ A	REQ B	REQ C	REQ D	REQ E
ev_query1	**x**	**x**	**x**		
ev_query2	**x**	**x**	**x**	**x**	
ev_query3	**x**	**x**		**x**	
ev_query4	**x**				
ev_query5	**x**	**x**			**x**

Changes from Original Design

Some changes from the original design that we decided to make included only cars that were manufactured in California. This decision came about when we were getting ready to import all of our data into the newly created database but noticed that there were over 100 thousand rows of data. For this reason, we decided to cut the amount of data which gave us only 76 rows of data. This change still satisfied the requirements of having at least 15 rows of data for normal tables, and 30 rows for linking tables.

The original database featured vehicles from 2011 to 2022. From the original proposal, with 112,635 total rows, we wanted to cut the vehicles that were made before 2019 so we can cover the last 4 years of the data. However, after deciding to include only cars manufactured in California, using years over 2019 became unnecessary because the rows from our dataset already

cut down to 76 rows. Therefore, we decided to keep all the years that were covered in our dataset.

Also, we originally excluded the census tract column from what we were going to include in our database, but later on decided that it was actually useful and decided to include it in our database.

Lastly, we were going to remove records that had a NULL or zero as a value in them since they posed no risk or relationship to the database. However, after importing our data and looking at how many records we would have, in order to satisfy the requirement of having 15-30 rows, we decided to keep these records and create more rows as other columns in the rows had valuable data even if only one or 2 columns were NULL or zero.

Database Ethics Considerations

Some of the data elements that we observed in our dataset that might lead to compromising of privacy are the VIN number, the DOL vehicle ID, and the vehicle location. In terms of the VIN number, for the most part this piece of information doesn't hold much potential for being used in an unethical way. That being said, there is still a chance, even if a small one, that the VIN number could be used for nefarious reasons, so as a team we thought that it should still be acknowledged. For example, the VIN number could be used to try and sell a car under a clean VIN if the original VIN has accidents or other problems related to it. A VIN number is similar to a social security number but for a vehicle. If this identification number gets into the wrong hands, scammers will use it for fraudulent purposes such as filing duplicate paperwork and registering multiple cars under that one identification number.

The vehicle location can be a privacy-related issue because it shows the state, city, county, and postal code of where the vehicle is located. The DOL vehicle ID is also a privacy concern because someone could look up a DOL vehicle ID in the Washington state "Electric Vehicle Title and Registration" database and find a cars matching VIN number, which we already established in the previous paragraph could be taken advantage of by scammers.

Lessons Learned

One lesson we learned is that normalization can be much more complicated than we originally thought. We had to ask for help from multiple TAs and go to the professor more than

once in order to finally get our tables normalized.

Another lesson we learned is how important it is to be detail oriented in the database building and data import process. We kept running into problems when importing our data into the database that stemmed from our csv files not matching the table format 100%. Almost every import we did was not successful on the first try and required troubleshooting, ranging from simple to extensive, in order to successfully populate our database's tables with the proper information.

A third lesson we learned is how dealing with larger amounts of data can make database building and management significantly more complicated. This might seem like an obvious presumption, and our group was well aware that a larger dataset would be harder to work with, but the extent to which it complicates the data import process was not revealed to us until we started the data import process and ran into roadblocks at every step of the way. The data imports would have taken us hours longer to complete if we had not filtered our sample data to a more reasonable quantity.

A final lesson we learned is that it's important to leave time for revisions. There were numerous parts of our project that required multiple partial or total rehauls, such as the normalization part of the project. This caused parts of our project to take several days longer than expected to complete. Our group members went above and beyond to make time beyond what we expected to work on the necessary revisions so that our final product was adequate for the project requirements, but we definitely learned that it's important to take this unexpected extra time into consideration.

Potential Future Work

- One change we could make in the future to improve our database would be to add more states other than just California.
    - We decided to focus only on electric vehicles from California because the sheer amount of data we would have had to include in the database would have been massive if we included some of the other states. Choosing vehicles just from California wasn't overly difficult because there were 76 vehicles in that case. If we had more time then we could try and expand the database further to include more states, so that way the database could give some contrast between electric

vehicles in different states. You could figure out information like, which state has the most electric vehicles with a model year of 2022, or which state has the least hybrids, etc. Information like this and many other useful pieces of information could be gleaned from adding in electric vehicles from other states, so if we had additional time in the future then this could be a priority.

- Find a way to keep our dataset up to date
    - One potential change we would make to our database if we were to work on it a bit more would be to find a way to keep the data that is entered to our database more up to date. Maybe we could find a way to import data on a monthly basis so that new vehicles manufactured in California throughout the year can be added and so forth. This way the database would stay up to date and have a live view of the EV environment in today's world at least for the state of California.
- One extension of our database could be to include tax credit information for ease of viewing
    - One extension that our team was thinking to make for our database would be to include tax credit information based on vehicle makes and models. There are limits and specific requirements that need to be met for certain cars/SUVs to make them eligible for both state and federal taxes so to have that information alongside as a column on our database would be very helpful to buyers and sellers of EV's.

Appendix

Appendix A: Introduction

The Introduction section includes a description of the dataset we used with how we finalized the dataset to use for the final project.

Appendix B: Data Description

The Data Description section includes four parts: logical design, physical database, sample data, and views/queries.

B-1: Logical Design

The Logical Design section includes an up-to-date picture of ERD and representation of the final structure of our database.

B-2: Physical Database

The Physical Database section includes the actual project database that we worked on MySQL workbench. It is attached separately and will show our understanding of physical design concepts and methods. It also includes a brief description of our current database design.

B-3: Sample Data

The Sample Data section will show our understanding of how information about real-world entities can be kept as data in our database. We include an example picture that shows at least 15 sample records in each of non-join tables.

B-4: Views / Queries

The Views / Queries section will have an actual file of queries we have been working on. Also, we will include a table showing which queries satisfy which requirements.

Appendix C: Changes from Original Design

The Changes from Original Design section include all the changes we have made from

the original proposal. There were few changes made in the process of project design, however; we included the finalized version of changes we made compared to the original proposal.

Appendix D: Database Ethics Considerations

The Database Ethics Considerations section includes some of the concerns we can face with using information in the database. We listed privacy concerns based on using identification number and location information of the vehicles.

Appendix E: Lessons Learned

The Lessons Learned section includes what we learned from this final project as we worked on it from beginning to end. We include some of the challenges encountered as well as some accomplishments we had after going through those challenges.

Appendix F: Potential Future Work

The Potential Future Work section includes what changes we can make in the future as well as some work that could have been done to make the final project easier.