Database Design and Modeling

Matthew Manik, *Member X, Member Y, Member Z**

Electric Vehicle Analysis

Introduction:

In order to reduce greenhouse gas emissions, Americans are considering buying electric cars as their main mode of transportation. This rise in demand has been reflected in the variety of electric vehicles sold in the market. We chose to address this topic because we believe that the type of car we chose directly impacts our mark on the environment and it is important to address the rise in demand for electric vehicles. We also believe that when buying electric vehicles, buyers also must consider other factors as well in their purchase. These may include what electric vehicles are available to purchase in their region, the price of the vehicle or even the vehicle's electric range. With this in mind, our objective with our database is to make it more accessible for potential buyers and current owners of electric vehicles to understand the environmental impact and competitive value of their cars.

Database Description:

Our team aims to create a database on battery electric vehicles (EVs) registered in the 6th and 7th legislative districts of Washington State from 2016 to 2021. By exploring the different makes, models, and their benefits, we aim to help users make informed decisions about EV purchases. The database will only include vehicles with Clean Alternative Fuel Vehicle Factor (CAVF) eligibility to indicate environmental efficiency and will be valuable to a diverse range of users. Potential buyers in Washington State, specifically in the 6th and 7th legislative districts can assess the popularity and efficiency of recent EV models. Policy advocates and environmentally conscious consumers can evaluate how existing EVs contribute to greenhouse gas reduction.

Additionally, users in other states can also use the Washington State data as a benchmark for assessing EV availability and quality. Economists and policy analysts can leverage the data for market analysis and policy development related to EV adoption and climate change. We will exclude vehicle identification numbers (VINs) and Department of Licensing registration IDs for privacy reasons. Our database will feature the entities (tables) Vehicles, Addresses, Cities,

* Member names blurred for privacy reasons

Counties, Postal Codes, Vehicle Makes, Vehicle Models, Electric Utilities, and Vehicle Electric Utilities. To make our dataset as representative as possible within the table size limits, we included multiple counties and cities, providing options for users in different locations.

Logical Design:

Before normalization we just had a single vehicles table listing each vehicle's county, city, state, electric utility, and electric range, among other things. For our first normal form, we identified two tables vehicles and electric utilities, where the vehicles table had a vehicle_id and electric utilities table had 2 composite primary keys vehicle_id and utlity_id. In our second normal form, we created a linking table for our vehicles table and electric_utilities table. We removed the partial dependencies so that every non-primary-key column in a CPK table depends on the entire primary key. When there were columns that did not depend entirely on the primary key, we moved those columns to another table.

In doing so, we created a linking table for the many-to-many relationship between vehicles and electric utilities. Lastly, the changes reflected in our third normal form represent the current state of our logical design in our current Entity Relationship Diagram. To remove the transitive dependencies, we made new tables for columns with transitive dependencies, in which non-primary-key columns depended on other non-primary-key columns. We also made additional tables with one-to-many relationships to make unique values of attributes more accessible. With this in mind, the resulting tables from the third normal form include postal_codes, vehicle_models, vehicle_makes, addresses, counties, cities and vehicles table.

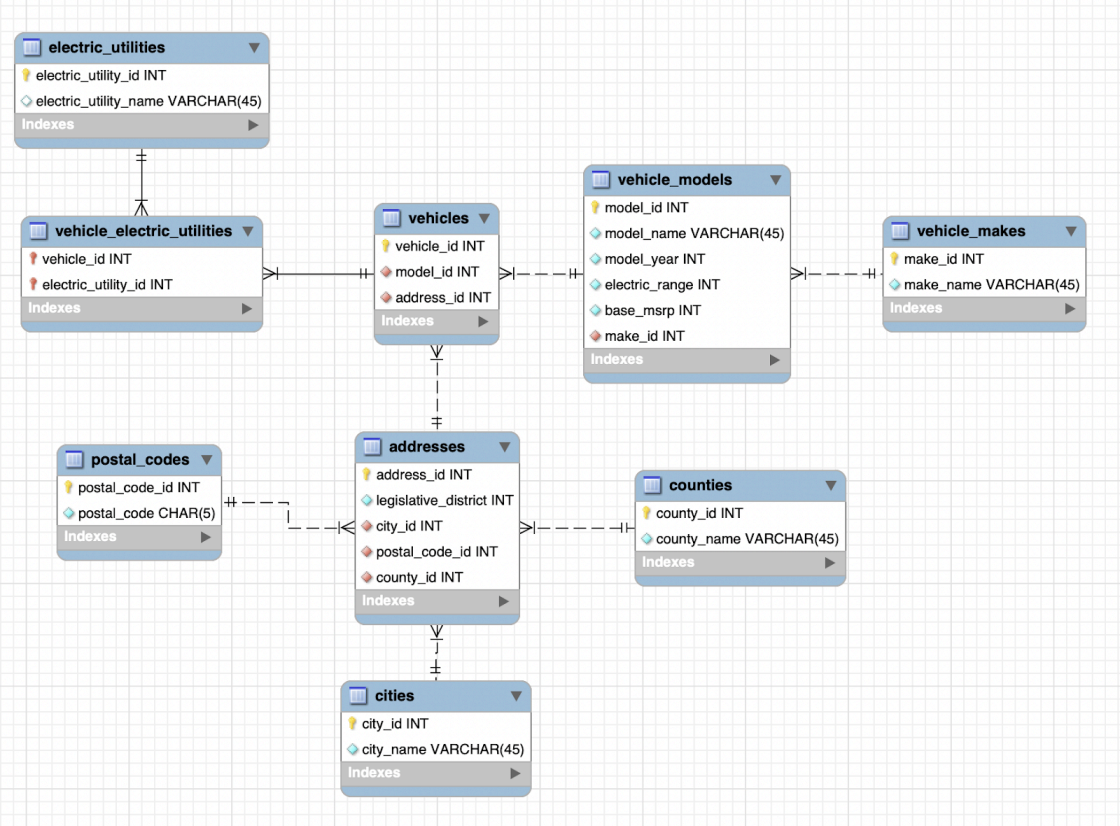

Image 1: Our entity relationship diagram (ERD).

Physical Database:

The main table of our database is the ‘vehicles’ table which contains vehicle id, model id and address id, linked to the ‘vehicle models’ and ‘vehicle makes’ table with a one-to-many relationship. One ‘vehicle makes’ to many ‘vehicle models’ and each ‘vehicle model’ to many ‘vehicles’. The main table is also linked to the ‘electric utilities’ table with ‘vehicle electric utilities’ as the linking table with its corresponding primary key, vehicle id and vehicle utility id on a many-to-many relationship in a one-to-many from ‘vehicle’ to ‘vehicle electric utilities’ then one-to-many from ‘electric utilities’ to ‘vehicle electric utilities’. Lastly, ‘postal codes’, ‘cities’ and ‘countries’ form a one-to-many relationships to ‘addresses’ with their respective primary key and ‘addresses’ to the main table with a one-to-many relationships as well on address id.

Sample Data:

Initially, we wanted to use all values in every selected column in the dataset for the database but after considering the scope of the project, we decided to further filter the data by restricting the entries which to include. For instance, we planned on using only electric powered vehicles even though the dataset included both electric and hybrids vehicles. The dataset also included more than 40 legislative districts in Washington of which only the 6th and 7th legislative districts were used for the database. From the dataset, we also did not include the columns such as DOL, Vehicle Identification Number, Census Tract and Vehicle Location. Our counties table has 5 rows because human population size and the distribution of vehicle makes vary greatly between locations; including more counties would have required us to exclude more vehicle makes or model years to prevent the "vehicles" table from exceeding 500 rows. Our "electric utilities" table has 10 rows for the same reason with the additional factor that most of the electric utilities are specific to certain municipalities. Our remaining 7 tables meet the minimum row requirements. Our linking table "vehicle electric utilities" originally had 1032 rows, so we took a pseudorandom sample of 500 rows with the constraint that every vehicle would have at least one row in the linking table.

Views and Queries:

Query Name	**JOIN (X4)**	**FILTER (X3)**	**AGGREGATE (X2)**	**LINKING (X1)**	**SUB-QUERY (X1)**	**SELECT QUERIES AS VIEWS (X5)**
Query 1	**X**	**X**	**X**			**X**
Query 2	**X**		**X**			**X**
Query 3		**X**	**X**		**X**	
Query 4	**X**	**X**				**X**
Query 5	**X**	**X**	**X**	**X**		**X**
Query 6	**X**	**X**	**X**			**X**

Total	**5**	**5**	**5**	**1**	**1**	**5**

Changes from Original Design:

As previously mentioned, there were many changes that we had made from our initial proposal. The major changes made since then are primarily on the scope of our data. Before we defined our scope of vehicles to both battery and hybrid electric vehicles. However in our design now, we only are interested in battery vehicles. Additionally, we previously included electric vehicles in all legislative districts but now we only include electric vehicles in the 6th and 7th legislative districts in order to limit the amount of vehicles that would be listed in our main vehicles table. Lastly, we changed the design from including vehicles in 2017-2023 to only being interested in the years 2016-2021 because we found that this year's range resulted in greater diversity of car models. In addition, when we initially filtered out our database so that it met the requirements of the assignment, we had a few attributes that we deemed to be insignificant data points upon further review. In our initial proposal, we agreed on only including data points that will be beneficial to a consumer. With our new goal of satisfying the information needs of a buyer, we included attributes such as Average Electric Range as well as Base MSRP which has many NULL values. Even with the NULL values involved, it is important that this data is revealed as it is notable. As far as tables go, we also have removed the Census Tract table as census tracts already are represented by unique numbers.

Database Ethics Considerations:

There are potential biases in our database design. By including only battery electric vehicles, our database design may not account for those who can afford only cheaper options such as hybrid vehicles rather than battery electric vehicles. Additionally, we have included vehicles only from the state of Washington, so our database is less useful for people in different parts of the United States. However, given the constraints of the project, we do believe that our design is somewhat inclusive. To maximize diversity within the constraints, we included 34 cities in Washington. Additionally, the model years of our cars range from 2016 to 2021. Though we do not include recent years, this 6 year span allows for a diverse range of vehicles to be included within the design. We included 15 different car makes such as Audi, Tesla, Nissan, and Ford. The dataset originates from the Research and Analysis Office of the Washington State

Department of Licensing, and the copyright license of the dataset is the Open Data Commons Open Database License (Data.WA.gov). The license allows us to modify and display the dataset freely (Open Knowledge Foundation). Therefore, we are not infringing on the Washington state government's copyright. We removed the DOL Vehicle ID and the Vehicle Identification Number (VIN) to reduce the risk that a user would link the information about a vehicle to the real vehicle. The original dataset metadata do not explicitly declare that all DOL Vehicle IDs were truncated (Data.WA.gov). Additionally, although the VIN numbers were truncated to 10 digits (Unrau), both DOL Vehicle ID and VIN were unnecessary to the purpose of our database. For that reason, privacy concerns did not have a significant impact on our database design. Database ethics was not very relevant when it came to our project topic due to the fact that the data was not too revealing of information and was publicly accessible data from the state of Washington. As stated earlier, privacy was the number one ethical obstacle that our group faced. With there being information such as Vehicle Identification Number (VIN), we made sure that this information was not revealed.

In conclusion, while our project may not have involved overly sensitive information, our emphasis on database ethics was crucial in establishing a foundation of responsible data management. By addressing the privacy concerns associated with seemingly innocuous data elements like VINs, we demonstrated a commitment to comprehensive ethical practices, acknowledging the interconnected nature of data and its potential impact on privacy and security. This approach not only ensures the integrity of our project in the present but also positions it ethically for the evolving landscape of data use and regulation.

Lessons Learned:

Since the creation of the proposal and logical design, our group was able to make great progress in designing and implementing our database while responding to the various feedback given to us. When we defined the scope in the proposal, we originally had more than a thousand rows of unique vehicle cars. When given feedback to limit the main table to less than 500 rows, our group had a meeting over Zoom to decide how to filter the data. Using Excel, we used the filter feature on various columns to limit our scope to smaller counties or cities, but that did not work. Eventually, our team member, David, thought to limit the geographic location by limiting to just the 6th and 7th legislative districts. This allowed us to have less than 500 rows for the

main vehicles table but also maintain a diverse number of car models and car makes in various cities. Additionally, In our initial proposal, we had included the DOL information but since DOL was already unique and we already had vehicle_id, we were suggested to remove it. Throughout our proposal and top down logical design, we had not considered any many-to-many relationships at all. We had group meetings where we had decided to not include a linking table because we could not identify any. However, when we attended office hours we were suggested to define a linking table for vehicles to electric utilities. Based on this feedback, in our project's logical design we defined a many to many relationship for vehicles and electric utilities. Another suggestion we were also given is to change the datatype for the model year of the vehicle to be an INT instead of YEAR. Initially we did year because we thought that the model year corresponded to the time datatype year. However, we were instructed to change it to INT as this data type would allow us to use mathematical calculations when we conduct our future analyses.

Potential Future Work:

Our database is biased in the sense that it does not include every relevant piece of information necessary to make a sound decision on whether or not to purchase a specific vehicle model over another in other locations since the sample data mainly focuses on a few districts in Washington. Not only was the sample dataset very limited to fit the scope of the project, conditions and reasons for the popularity of those models in Washington may not exactly apply to other places. The data obtained from the impact of having electric vehicles of certain models in a few districts of Washington can't be used for proof that the same result will be achieved in other places after only experimenting in one location. To address this sense of biases in the database under consideration that the scope of an INST 327 project no longer applies, we can expand the location sampled to cover an entire state instead of a few districts in a state and increase the range of year for sample vehicles from 2010 to present instead of 2016-2021. From there, we can replicate the schema under a different name for all 50 states so that each state has its own database created with a sample dataset from that specific state obtained by surveying the market and the population of that state. In addition, we will include hybrid vehicle models in the database and also make a price table in each database to store the prices for quick sorting and other calculations.

References:

Data.WA.gov. (2022). *Electric Vehicle Population Data*. https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2

Open Knowledge Foundation. (n.d.). *Open Data Commons Open Database License (ODbL) v1.0*. Open Data Commons, https://opendatacommons.org/licenses/odbl/1-0/

Unrau, Jason. (2016). *How to Read a VIN (Vehicle Identification Number)*. YourMechanic, https://www.yourmechanic.com/article/how-to-decode-a-vin-vehicle-identification-number-by-jason-unrau