

arnoldHicran_TermProject

April 26, 2023

MET CS 689

Hicran Arnold

Term Project: Parts Unlimited - EV Data Warehouse

1 Part 1: Are you working on your own or with a partner? If with a partner provide their name. If on your own, just state that this is the case.

I am working alone in this project

2 Part 2: Determine the project scope

- In a short paragraph, describe the topic you wish to explore –an update if any
- Update the five business questions that your data warehouse will answer.

Project Description The topic I wish to explore for my data warehouse project is the management of Parts Unlimited’s EV parts business. This topic was inspired by “The Unicorn Project,” which describes the challenges and opportunities of digital transformation in a large organization. Specifically, I plan to focus on storing and analyzing data related to charging stations, EV products, customer purchases, and geographic location. Parts Unlimited already sells EV parts, and my goal is to improve the organization’s data management, reporting, and analysis capabilities related to this business.

Business Questions:

Business Question 1

What is the distribution of EV charging stations and EV cars across different geographic regions, and how does this relate to the popularity of EVs in those regions?

The Marketing and Sales department of your business can benefit from this query. By understanding the distribution of EV charging stations and EV cars across different geographic regions and how it relates to the popularity of EVs in those region. Additionally, the Sales department can use this information to identify regions with higher demand for EVs and focus their sales efforts in those areas to increase sales and revenue.

Business Question 2

How has the popularity of different EV models and electric vehicle types varied by geographic region over time, and how can Parts Unlimited use this information to target their marketing and sales efforts more effectively

The first question focused on the distribution of EV charging stations and EV cars across different geographic regions and its relation to the popularity of EVs. It was more general and aimed at understanding the overall trends and patterns in the adoption of EVs.

In contrast, the second question focuses specifically on the popularity of different EV models and electric vehicle types in different regions over time. It is more detailed and aimed at understanding how consumer preferences for specific EV models and types have evolved in different regions.

While both queries can help inform marketing and sales efforts, the second query can provide more granular insights into the specific EV models and types that are most popular in different regions, which can help Parts Unlimited make more targeted marketing and sales decisions.

Business Question Q3

How does the weight of EV parts relate to their price, and how does this relationship differ across different manufacturers ?

The Operations and Sales departments of Parts Unlimited can benefit from this query. By understanding how the weight of EV parts relates to their price, and how this relationship differs across different manufacturers, Parts Unlimited can make more informed decisions regarding inventory management, pricing, and supplier selection.

Business Question Q4

What is the average price, weight, and quantity of products sold by Parts Unlimited over time, and how does this vary across different geographic regions in which the products have been manufactured?

The Operations and Sales departments of Parts Unlimited can benefit from this query. By understanding the average price, weight, and quantity of products sold by Parts Unlimited over time and how this varies across different geographic regions in which the products have been manufactured, Parts Unlimited can make more informed decisions regarding inventory management, pricing, and supplier selection.

Business Question Q5

How does the demand for charging stations and the number of vendors operating in different regions relate to the size of the EV market, and what trends can be identified over time?

The Marketing and Sales departments of Parts Unlimited can benefit from this query. By understanding how the demand for charging stations and the number of vendors operating in different regions relates to the size of the EV market and what trends can be identified over time, Parts


Unlimited can make more informed decisions regarding their marketing and sales efforts. Overall, understanding the relationship between the demand for charging stations, the number of vendors, and the size of the EV market can help Parts Unlimited make more strategic decisions regarding their marketing, sales, and product development effort

3 Part 3: Data Sources

Provide two data sources you will be using, for each data source list the number or columns and rows that are in each data source. Provide a header and first 5 rows from each source. - What is the URL or location of the data? - What information does this data provide that will help answer one or more of the above questions? - Do you see any issues in the data that will require transformation.

Dataset Name	Columns	Rows	URL
1-Product Info - EV West	7	187	EV West

The data set provides information about EV parts, manufacturer, weight and price information. The data is not available in the PDF version and requires a Python script to extract and format the information.




Elcon 3.3kW UF CAN Bus Charger with EVCC
The UFC3300 is an excellent low cost solution to h ...

Model : UFC33-CAN
Manufacturer : [Elcon](#)
Weight : 19.00

Price: **\$1,199.00**

[ADD TO CART](#) [Details](#)




Elcon PFC1500 Charger
This item has been discontinued by the manufacture ...

Model : ELC1500
Manufacturer : [Elcon](#)
Weight : 15.00

Price: **\$569.00**

[ADD TO CART](#) [Details](#)



Manzanita Micro PFC 20-XM Charger
This product is no longer available through EV Wes ...

Model : MM PFC20-XM
Manufacturer : [Manzanita](#)
Weight : 16.00

Price: **\$2,390.00**

[ADD TO CART](#) [Details](#)

Dataset Name	Columns	Rows	URL
2-Alternative Fueling Station Locations	65	56,800	Data.gov

Provides information about the current list of charging stations and their locations. There are a lot of missing values that need to be cleaned and lat and long location info needs to be identical with the EV Car Population dataset. Price information is available in a text format and needs to be formatted so it is a number.

EV station Dataset Column IndexNames		EV Car Dataset IndexColumn Names		EV station Dataset Column IndexNames		EV station Dataset Column IndexNames	
1	Fuel Type Code	21	EV Other Info	41	Access Days Time (French)	61	CNG PSI
2	Station Name	22	EV Network	42	BD Blends (French)	62	CNG Vehicle Class
3	Street Address	23	EV Network Web	43	Groups With Access Code (French)	63	LNG Vehicle Class
4	Intersection Directions	24	Geocode Status	44	Hydrogen Is Retail	64	EV On-Site Renewable Source
5	City	25	Latitude	45	Access Code	65	Restricted Access
6	State	26	Longitude	46	Access Detail Code		
7	ZIP	27	Date Last Confirmed	47	Federal Agency Code		
8	Plus4	28	ID	48	Facility Type		
9	Station Phone	29	Updated At	49	CNG Dispenser Num		
10	Status Code	30	Owner Type Code	50	CNG On-Site Renewable Source		
11	Expected Date	31	Federal Agency ID	51	CNG Total Compression Capacity		
12	Groups With Access Code	32	Federal Agency Name	52	CNG Storage Capacity		
13	Access Days Time	33	Open Date	53	LNG On-Site Renewable Source		
14	Cards Accepted	34	Hydrogen Status Link	54	E85 Other Ethanol Blends		
15	BD Blends	35	NG Vehicle Class	55	EV Pricing		
16	NG Fill Type Code	36	LPG Primary	56	EV Pricing (French)		
17	NG PSI	37	E85 Blender Pump	57	LPG Nozzle Types		
18	EV Level1 EVSE Num	38	EV Connector Types	58	Hydrogen Pressures		
19	EV Level2 EVSE Num	39	Country	59	Hydrogen Standards		

EV station Dataset Column IndexNames		EV Car Dataset IndexColumn Names		EV station Dataset Column IndexNames		EV station Dataset Column IndexNames	
20	EV DC Fast Count	40	Intersection Directions (French)	60	CNG Fill Type Code		

```
[1]: import warnings
warnings.simplefilter(action='ignore', category=Warning)
```

```
[2]: import pandas as pd
pd.set_option("display.max_columns",10 )
```

```
[3]: import pandas as pd
ev_station = pd.read_csv('data/alt_fuel_stations.csv')
ev_station.head(5)
```

```
[3]: Fuel Type Code                                Station Name \
0          CNG                                Spire - Montgomery Operations Center
1          CNG                                PS Energy - Atlanta
2          CNG    Metropolitan Atlanta Rapid Transit Authority
3          CNG                                United Parcel Service
4          CNG    Clean Energy - Texas Department of Transportation

      Street Address                                Intersection Directions \
0      2951 Chestnut St                                NaN
1      340 Whitehall St    From I-7585 N, exit 91 to Central Ave, left on...
2      2424 Piedmont Rd NE                                NaN
3      270 Marvin Miller Dr                                NaN
4      7721A Washington St    I-10, Washington Ave exit, 1.5 blocks to the s...

      City ...      CNG PSI CNG Vehicle Class  LNG Vehicle Class \
0  Montgomery ...      3600                MD                NaN
1    Atlanta ...      3600                MD                NaN
2    Atlanta ...      3000                LD                NaN
3    Atlanta ...      3600                HD                NaN
4    Houston ...    3000 3600                MD                NaN

      EV On-Site Renewable Source Restricted Access
0                NaN                NaN
1                NaN                NaN
2                NaN                NaN
3                NaN                NaN
4                NaN                NaN
```

```
[5 rows x 65 columns]
```

Dataset Name	Columns	Rows	URL
3- EV Population Data	17	121,978	Data.gov

This dataset shows EV cars that are currently registered through Washington State Department of Licensing (DOL). Provides information about EV cars, car types, and registered locations. Location fields have to match with the format of the EV Charging dataset. There are some missing fields. Currently, I do not see any other issues related to this dataset.

Index	Electric_Vehicle_Population
1	VIN (1-10)
2	County
3	City
4	State
5	Postal Code
6	Model Year
7	Make
8	Model
9	Electric Vehicle Type
10	Clean Alternative Fuel Vehicle (CAFV) Eligibility
11	Electric Range
12	Base MSRP
13	Legislative District
14	DOL Vehicle ID
15	Vehicle Location
16	Electric Utility
17	2020 Census Tract

```
[4]: ev_pop = pd.read_csv('data/Electric_Vehicle_Population_Data.csv')
ev_pop.head(5)
```

```
[4]: VIN (1-10)    County    City State    Postal Code    ... \
0  5YJ3E1EB2J    Suffolk    Suffolk    VA        23435.0    ...
1  5YJ3E1ECXL    Yakima     Yakima     WA        98908.0    ...
2  WA1LAAGE7M    Yakima     Yakima     WA        98908.0    ...
3  5YJ3E1EA1K    Danville   Danville   VA        24541.0    ...
4  1FADP5CU9E    Norfolk    Norfolk    VA        23518.0    ...

    Legislative District DOL Vehicle ID    Vehicle Location \
0                NaN    476647986    POINT (-76.42443 36.8752)
1                14.0    103490145    POINT (-120.56916 46.58514)
2                14.0    144941534    POINT (-120.56916 46.58514)
3                NaN    168513922    POINT (-79.4172 36.58598)
4                NaN    150749378    POINT (-76.21549 36.92478)
```

Electric Utility 2020 Census Tract

```

0          NaN      5.180008e+10
1    PACIFICORP      5.307700e+10
2    PACIFICORP      5.307700e+10
3          NaN      5.159000e+10
4          NaN      5.171001e+10

```

[5 rows x 17 columns]

Dataset Name	Columns	Rows	URL
4- Manufacture Location	6	25	extracted by online search

Provides information about EV cars. This dataset is clean, currently I do not see any issues.

Index	Manufacture Location
1	company__name
2	latitude
3	longitude
4	city
5	state
6	country

4 Part 4: Dimensions - Review the data and the business questions from part 2.

- What fields (attributes) are in the data that will be used for the dimensions.
- Determine the dimension tables. There should be at least two non-date dimensions and one date dimension for each fact table.
- At least one (non-date) dimension in your design should have a hierarchy.
- What are the attributes that will be tracked via slowly changing dimensions?
- What attributes within the dimensions will need transformation before they are loaded into the dimension, for example it could be to build consistency or any other issues? This is where for example you might build case statements in your code to handle various scenarios. Two to three examples showing some sample data and what you think the transformation will be during your ETL would be helpful here.

In the below tables I am showing each data sets by their corresponding tables and their scd types and their data transformation information summary

EV poulation Data set				
Field Name	Description	Transformation Info	Table Name	SCD Type Info
ev_car_pop_id	this is a pk field assign to the unique each records	not in the original data/ has to be populated	ev_car_populat	PK
dol_vehicle_Id	unique id for each vehicle	no transfor- mation needed	ev_car_populat	id type3
VIN	unique id for each make model	no transfor- mation needed	ev_car_populat	id type3
make	ev brand	no transfor- mation needed	ev_car_populat	id type3
vechicle_model	brand model	no transfor- mation needed	ev_car_populat	id type3
current_reg_status	shows if the registration in current (scd type 2)	not in the original data/ has to be populated	ev_car_populat	id type3(scd field)
reg_status_prev	previous registration status	not in the original data/ has to be populated	ev_car_populat	id type3(scd field)
reg_sta_updated	when the registration has been updated	not in the original data/ has to be populated	ev_car_populat	id type3(scd field)
location_id	unique location id for each car	not in the original data/ has to be populated	location_dim	scd type3
city	city name that car registered	no transfor- mation needed	location_dim	scd type3
region	the region that city falls under to	not in the original data/ has to be populated	location_dim	scd type3
country	the country that car registred	not in the original data/ has to be populated	location_dim	scd type3

EV poulation Data set					
latitude	car registered address lat info	Point field , I have split into lat and long	location_dim	scd type3	
longitude	car registered address lat info	Point field , I have split into lat and long	location_dim	scd type3	
city_lat	city lat info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical- lypopulated by city name	location_dim	scd type3	
city_long	city long info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical- lypopulated by city name	location_dim	scd type3	
latitude_prev	scd field keeping change in the lat field	not in the original data/ has to be populated	location_dim	scd type3(scd field)	
longitude_prev	scd field keeping change in the long field	not in the original data/ has to be populated	location_dim	scd type3(scd field)	

EV poulation Data set					
updated_at	scd field tracking adjustment date	not in the original data/ has to be populated	location_dim	scd type3(scd field)	
date_id	unique id for registration date	not in the original data/ has to be populatedI have used make year to create this data	date_dim	scd type 1	
day	registration date	not in the original data/ has to be populatedI have used make year to create this data	date_dim	scd type 1	
month	registration month	not in the original data/ has to be populatedI have used make year to create this data	date_dim	scd type 1	
year	registration year	not in the original data/ has to be populatedI have used make year to create this data	date_dim	scd type 1	

EV poulation Data set					
date	registration year	not in the original data/ has to be populatedI have used make year to create this data as01/01/make year	date_dim	scd type 1	

below is the charging station dataset with the final data fields information

Field Name	Description	Transformation Info	Table Name	SCD Type Info
charging_station_dim_id	unique id for each charging stations	not in the original data/ has to be populated	ev_charging_stations	PK
charging_station_unique_id	unique id for each charging stations- source	no transfor- mation needed	ev_charging_stations	type 2
sk_ev_dim_id	surrogate key to keep track of scd 2 change	not in the original data/ has to be populated		SK type2(scd field)
station_name	charging station name	no transfor- mation needed	ev_charging_stations	type 2
status_code	temporarily closed, open	transformed from the code (P,T) to show more readable format	ev_charging_stations	type 2
access_code	Private or Public	no transfor- mation needed	ev_charging_stations	type 2
currency	charging station price currency	not in the original data/ has to be populated	ev_charging_stations	type 2

Field Name	Description	Transformation Info	Table Name	SCD Type Info
row_effective_date	scd type 2 shows when the first record created	not in the original data/ has to be populated	ev_charging_station	scd type2(scd field)
row_expiration_date	scd type 2 shows if this record has been adjusted and new recorded added	not in the original data/ has to be populated	ev_charging_station	scd type2(scd field)
row_status	scd type 2 shows if this record still effective	not in the original data/ has to be populated	ev_charging_station	scd type2(scd field)
location_id	unique location id for each charging station recorded	not in the original data/ has to be populated	location_dim	PK
city	city name of the charging station	no transfor- mation needed	location_dim	scd type3
region	the region that city falls under to	not in the original data/ has to be populated	location_dim	scd type3
country	country name	no transfor- mation needed	location_dim	scd type3
latitude	charging station address lat info	no transfor- mation needed	location_dim	scd type3
longitude	charging station address long info	no transfor- mation needed	location_dim	scd type3

Field Name	Description	Transformation Info	Table Name	SCD Type Info
city_lat	city lat info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical-lypopulated by city name	location_dim	scd type3
city_long	city long info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical-lypopulated by city name	location_dim	scd type3
latitude_prev	scd field keeping change in the lat field	not in the original data/ has to be populated	location_dim	scd type3(scd field)
longitude_prev	scd field keeping change in the long field	not in the original data/ has to be populated	location_dim	scd type3(scd field)
updated_at	scd field tracking adjustment date	not in the original data/ has to be populated	location_dim	scd type3(scd field)
date_id	unique id for registration date	not in the original data/ has to be populated	date_dim	PK

Field Name	Description	Transformation Info	Table Name	SCD Type Info
day	charging station open date	transformed from the open date field in source data	date_dim	scd type 1
month	charging station open month	transformed from the open date field in source data	date_dim	scd type 1
year	charging station open year	transformed from the open date field in source data	date_dim	scd type 1
date	charging station open year	charging station open date	date_dim	scd type 1

EvWest data set

Field Name	Description	Transformation Info	Table Name	SCD Type Info
manufacturer id	unique id for each manufacturer	not in the original data/ has to be populated	manufacturer	PK
sk manufacturer	surrogate key to keep track of scd 2 change	not in the original data/ has to be populated	manufacturer	SK SCD type2(scd field)
manufacturer_name	manufacturer station name	pdf extraction transformation	manufacturer	scd type 2
manu_row_effective_date	scd2 tracking : record creatation date	not in the original data/ has to be populated	manufacturer	scd type2(scd field)

EvWest data-set				
manu_row_expiration	scd type 2 tracking : record expiration date track	not in the original data/ has to be populated	manufacturer	scd type2(scd field)
manu_row_ind	scd type 2 tracking : indicated if the record is in use or there is an adjustment (expired/active)	not in the original data/ has to be populated	manufacturer	scd type2(scd field)
product_id	unique id for each product	not in the original data/ has to be populated	Products	PK
product_code	unique code for each product type	pdf extraction transformation	Products	scd type 3
product_name	product type name	pdf extraction transformation	Products	scd type 3
product_description	long description of the product	pdf extraction transformation	Products	scd type 3
product_category	product category	pdf extraction transformation	Products	scd type 3
product_name_history	scd type 3: tracks of the name change in the pro- ductPrevious name of the product	not in the original data/ has to be populated	Products	scd type3(scd field)
product_updated	scd type 3: tracks of the name change in the pro- ducttracks of the adjustment date	not in the original data/ has to be populated	Products	scd type3(scd field)

EvWest data-set					
location_id	unique location id for each charging station recorded	not in the original data/ has to be populated	location_dim	PK	
city	city name that car registered	no transformation needed	location_dim	scd type3	
region	the region that city falls under to	not in the original data/ has to be populated	location_dim	scd type3	
country	the country that car registred	no transformation needed	location_dim	scd type3	
latitude	charging station address lat info	no transformation needed	location_dim	scd type3	
longitude	charging station address long info	no transformation needed	location_dim	scd type3	
city_lat	city lat info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical-lypopulated by city name	location_dim	scd type3	

EvWest data-set				
city_long	city long info	this was not in the original I pulled via api call but I realised that Tablaue already has city lat and long info can be automatical-lypopulated by city name	location_dim	scd type3
latitude_prev	scd field keeping change in the lat field	not in the original data/ has to be populated	location_dim	scd type3(scd field)
longitude_prev	scd field keeping change in the long field	not in the original data/ has to be populated	location_dim	scd type3(scd field)
updated_at	scd field tracking adjustment date	not in the original data/ has to be populated	location_dim	scd type3(scd field)
date_id	unique id for manufacturer vendor since date	Not in the original data/ has to be populated	date_dim	PK
day	manufacturer vendor since date -day	not in the original data/ has to be populated	date_dim	scd type 1
month	manufacturer vendor since date -month	not in the original data/ has to be populated	date_dim	scd type 1
year	manufacturer vendor since date -year	not in the original data/ has to be populated	date_dim	scd type 1

EvWest data-set				
date	manufacturer vendor since date	not in the original data/ has to be populated	date_dim	scd type 1

5 Part 5: Facts – Review the data and the business questions from step 1.

What measurements are in the data that will be used for the fact tables? What measures will you be calculating (i.e. using an aggregate function, or some other transformation – recall as an example some of the aggregation you did in assignment 1A)

manufacture-facts: This table is keeping track of manufacturers and manufacturer products that the company sells. The measurements are weight,price

Field Name	Description	Transformation Info	SCD Type Info
manu_fact_id	unique id for each manufacturer fact record	not in the original data/ has to be populated	PK
location_id	foreign key to location info	extracted by merge	FK
date_dim	foreign key to date info	extracted by merge	FK
manufacture_id	foreign key to location manufacturer table connection	extracted by merge	FK
product_id	foreign key to location product table connection	extracted by merge	FK
price	price of the product for each manufacturer	pdf extraction	measurement
weight	weight of the product for each manufacturer	pdf extraction	measurement
quantity	quantity of the product for each manufacturer	not in the original data/ has to be populated	measurement

below is my second fact ev_charging_facts snapshot table tracks of snapshot of the charging stations quantity and the ev population

Field Name	Description	Transformation Info	SCD Type Info
ev_char_fact_id	unique id for each charging fact records	not in the original data/has to be populated	PK
location_id	fk to location info	extracted by merge	FK
date_dim	fk to date info	extracted by merge	FK
charging_station_id	fk to location manufacturer table connection	extracted by merge	FK
ev_car_pop_id	fk to location ev population table connection	extracted by merge	FK
vehichle_pop_quantity	quantiy of cars	pdf extraction	measurement
ev_cs_quantity	quantity of charging stations	pdf extraction	measurement
ev_price	price of the charging stations	not in the original data/has to be populated	measurement

cumulative fact table: this tracks of the overall quantities , this is a bridge of operations and research tables

Field Name	Description	Transformation Info	SCD Type Info
manu_fact_cumu_id	unique id for each charging fact records	not in the original data/ has to be populated	PK
month_id	month (time dimension in month grain)	extracted by merge	FK
location_id	fk to location info	extracted by merge	FK
manufacturer_id	fk to manufacturer info	extracted by merge	FK
total_product_quntity	total product quantity for that month	aggregated with etl	measurement
total_ev_charging_stations	total charging stations	aggregated with etl	measurement

Field Name	Description	Transformation Info	SCD Type Info
total_ev_population	total ev car population	aggregated with etl	measurement

6 Part 7: Provide code and screenshots of loading your data into staging/data frame

For Question 7-9 : I used pandas df to load data and before transformation and again I load the data with the help of the pandas sql load function. Please see my etl jupyter notebook for more details. I am including a three etl file one is for charging station facts and the other one for manufacturer facts. I grouped etl for each documents for each fact table

7 Part 8: Provide code and screenshots of transforming the data. Perhaps you are adjusting for consistency of data or calculating aggregates.

For Question 7-9 : I used pandas df to load data and before transformation and again I load the data with the help of the pandas sql load function. Please see my etl jupyter notebook for more details. I am including a three etl file one is for charging station facts and the other one for manufacturer facts. I grouped etl for each documents for each fact table

8 Part 9: Provide code and screenshots of loading the data into the two dimensions and the fact. At this time, you do not need to worry about maintenance of slowly changing dimensions, the focus is on the initial data load. If you are loading into SCD2 or SCD3, make sure to show the SCD maintenance attributes populated.

I have included a separate notebook for this transformation and breakdown the process for each scd type and showed examples

8.1 Part 11: Outline one of the business questions you outlined in part 2 (can be same as part 10) and answer it with a Tableau or PowerBI visualization. In a single sentence explain why you selected this particular visualization to answer this question.

Please see my query notebook for my business questions. I answered each business questions with sql aggregated queries. Please also see my presentation documents for visualizations for those questions