Data Analysis Final Paper

NABINA KHADKA

26 APRIL, 2023

## Issue Description

Describe the general area or issue you want to investigate in your data analysis project.

With the increase in the price of gas, people have started to opt for more efficient mode of transportation. As USA is designed in a sub-urban style of living, it is not feasible to have a daily commute in the public vehicle. Most people go for driving their private vehicle.

Recently, people have started to go into Electric Vehicles(EVs). However, the old fashioned gas vehicles are still popular but since EVs are gaining more customers, hybrid cars are also being introduced. For my data analysis project, I am trying to gather information about the total EVs registered in Washington State Department of Licensing(DOL). The EVs are of two categories - Battery EVs and Plug-in Hybrid EVs.

## Questions

Define at least two specific questions you would like to attempt to answer.

1. Which brand has the most popularity gained over the year?
2. Estimate the electric range of the most popular vehicle brand.

## Data Source

Identify the data source(s) you used for your analysis. Provide a URL if possible.

The metadata for the data was created on November,2020. It was later updated on April, 2023. The metadata can be found here:

https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld

The publisher of the data is data.wa.gov. The data set was created for public use and its license can be found here.

http://opendatacommons.org/licenses/odbl/1.0/

## Documentation

Provide a link to the documentation for the data or the documentation itself. Is there a data dictionary?

<https://catalog.data.gov/dataset/electric-vehicle-population-data>

This data set 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 data set has 124,716 entries under 17 total columns. There are few columns that are missing so the total accuracy of the data set cannot be determined. In order to get more precise analysis, the NA value columns needs to be removed.

## Description of the Data

Use the tools in R such as str() and summary() to describe the original dataset you imported.

Let’s include the packages that we need

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ purrr 0.3.5  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readr)

Sample\_Data <- read\_csv("Electric\_Vehicle\_Population\_Data.csv")

## Rows: 124716 Columns: 17  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (11): VIN (1-10), County, City, State, Make, Model, Electric Vehicle Typ...  
## dbl (6): Postal Code, Model Year, Electric Range, Base MSRP, Legislative Di...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

str(Sample\_Data)

## spc\_tbl\_ [124,716 × 17] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ VIN (1-10) : chr [1:124716] "5YJ3E1EB4L" "5YJ3E1EA7K" "7JRBR0FL9M" "5YJXCBE21K" ...  
## $ County : chr [1:124716] "Yakima" "San Diego" "Lane" "Yakima" ...  
## $ City : chr [1:124716] "Yakima" "San Diego" "Eugene" "Yakima" ...  
## $ State : chr [1:124716] "WA" "CA" "OR" "WA" ...  
## $ Postal Code : num [1:124716] 98908 92101 97404 98908 98021 ...  
## $ Model Year : num [1:124716] 2020 2019 2021 2019 2017 ...  
## $ Make : chr [1:124716] "TESLA" "TESLA" "VOLVO" "TESLA" ...  
## $ Model : chr [1:124716] "MODEL 3" "MODEL 3" "S60" "MODEL X" ...  
## $ Electric Vehicle Type : chr [1:124716] "Battery Electric Vehicle (BEV)" "Battery Electric Vehicle (BEV)" "Plug-in Hybrid Electric Vehicle (PHEV)" "Battery Electric Vehicle (BEV)" ...  
## $ Clean Alternative Fuel Vehicle (CAFV) Eligibility: chr [1:124716] "Clean Alternative Fuel Vehicle Eligible" "Clean Alternative Fuel Vehicle Eligible" "Not eligible due to low battery range" "Clean Alternative Fuel Vehicle Eligible" ...  
## $ Electric Range : num [1:124716] 322 220 22 289 14 84 215 10 75 26 ...  
## $ Base MSRP : num [1:124716] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Legislative District : num [1:124716] 14 NA NA 14 1 38 23 14 26 26 ...  
## $ DOL Vehicle ID : num [1:124716] 1.27e+08 2.67e+08 1.45e+08 4.77e+08 1.06e+08 ...  
## $ Vehicle Location : chr [1:124716] "POINT (-120.56916 46.58514)" "POINT (-117.16171 32.71568)" "POINT (-123.12802 44.09573)" "POINT (-120.56916 46.58514)" ...  
## $ Electric Utility : chr [1:124716] "PACIFICORP" NA NA "PACIFICORP" ...  
## $ 2020 Census Tract : chr [1:124716] "53077000904" "06073005102" "41039002401" "53077000401" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `VIN (1-10)` = col\_character(),  
## .. County = col\_character(),  
## .. City = col\_character(),  
## .. State = col\_character(),  
## .. `Postal Code` = col\_double(),  
## .. `Model Year` = col\_double(),  
## .. Make = col\_character(),  
## .. Model = col\_character(),  
## .. `Electric Vehicle Type` = col\_character(),  
## .. `Clean Alternative Fuel Vehicle (CAFV) Eligibility` = col\_character(),  
## .. `Electric Range` = col\_double(),  
## .. `Base MSRP` = col\_double(),  
## .. `Legislative District` = col\_double(),  
## .. `DOL Vehicle ID` = col\_double(),  
## .. `Vehicle Location` = col\_character(),  
## .. `Electric Utility` = col\_character(),  
## .. `2020 Census Tract` = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(Sample\_Data)

## VIN (1-10) County City State   
## Length:124716 Length:124716 Length:124716 Length:124716   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Postal Code Model Year Make Model   
## Min. : 1730 Min. :1997 Length:124716 Length:124716   
## 1st Qu.:98052 1st Qu.:2018 Class :character Class :character   
## Median :98121 Median :2020 Mode :character Mode :character   
## Mean :98164 Mean :2019   
## 3rd Qu.:98370 3rd Qu.:2022   
## Max. :99701 Max. :2023   
## NA's :2   
## Electric Vehicle Type Clean Alternative Fuel Vehicle (CAFV) Eligibility  
## Length:124716 Length:124716   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##   
## Electric Range Base MSRP Legislative District DOL Vehicle ID   
## Min. : 0.00 Min. : 0 Min. : 1.00 Min. : 4385   
## 1st Qu.: 0.00 1st Qu.: 0 1st Qu.:18.00 1st Qu.:154101503   
## Median : 25.00 Median : 0 Median :34.00 Median :199555765   
## Mean : 79.47 Mean : 1556 Mean :29.66 Mean :204078955   
## 3rd Qu.:200.00 3rd Qu.: 0 3rd Qu.:43.00 3rd Qu.:227516509   
## Max. :337.00 Max. :845000 Max. :49.00 Max. :479254772   
## NA's :297   
## Vehicle Location Electric Utility 2020 Census Tract   
## Length:124716 Length:124716 Length:124716   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

## Cleaning and Preparation

Describe the steps you took to get from your original data set to the final data set you used for your analysis. Include the R code in chunks.

The file has a lot of columns that we do not need. First step would be to remove those columns and then make a data frame with the lesser and useful columns.

data <- Sample\_Data %>%  
 rename("Vehicle" ="Make",  
 "Year" ="Model Year" ,  
 "VehicleType" = "Electric Vehicle Type" ,  
 "Range"="Electric Range") %>%  
 select(County, City , State , Vehicle , Model , Year , VehicleType, Range)

Here, we created a new data set. The summary of new data set is given below.

head(data)

## # A tibble: 6 × 8  
## County City State Vehicle Model Year VehicleType Range  
## <chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl>  
## 1 Yakima Yakima WA TESLA MODEL 3 2020 Battery Electric Vehicl… 322  
## 2 San Diego San Diego CA TESLA MODEL 3 2019 Battery Electric Vehicl… 220  
## 3 Lane Eugene OR VOLVO S60 2021 Plug-in Hybrid Electric… 22  
## 4 Yakima Yakima WA TESLA MODEL X 2019 Battery Electric Vehicl… 289  
## 5 Snohomish Bothell WA BMW X5 2017 Plug-in Hybrid Electric… 14  
## 6 Snohomish Everett WA NISSAN LEAF 2015 Battery Electric Vehicl… 84

summary(data)

## County City State Vehicle   
## Length:124716 Length:124716 Length:124716 Length:124716   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Model Year VehicleType Range   
## Length:124716 Min. :1997 Length:124716 Min. : 0.00   
## Class :character 1st Qu.:2018 Class :character 1st Qu.: 0.00   
## Mode :character Median :2020 Mode :character Median : 25.00   
## Mean :2019 Mean : 79.47   
## 3rd Qu.:2022 3rd Qu.:200.00   
## Max. :2023 Max. :337.00

Next, we notice that there are missing columns in our data set. We need to remove those NA values in order to get more precise analysis.

data <- data %>%  
 drop\_na()  
  
summary(data)

## County City State Vehicle   
## Length:124533 Length:124533 Length:124533 Length:124533   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Model Year VehicleType Range   
## Length:124533 Min. :1997 Length:124533 Min. : 0.00   
## Class :character 1st Qu.:2018 Class :character 1st Qu.: 0.00   
## Mode :character Median :2020 Mode :character Median : 25.00   
## Mean :2019 Mean : 79.59   
## 3rd Qu.:2022 3rd Qu.:200.00   
## Max. :2023 Max. :337.00

## Final Results

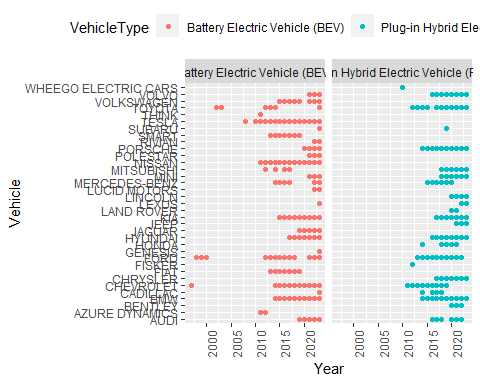
Show how you approached the questions you posed at the beginning. Describe how much you were able to accomplish. There should be both graphical and numerical results produced by R code included in chunks. Explain what you did and what it means.

For the final parts, lets head into the two questions we wrote at the beginning.

## Popularity gained over the year

Let us find out the brands that we have and how their popularity has been growing over the years.

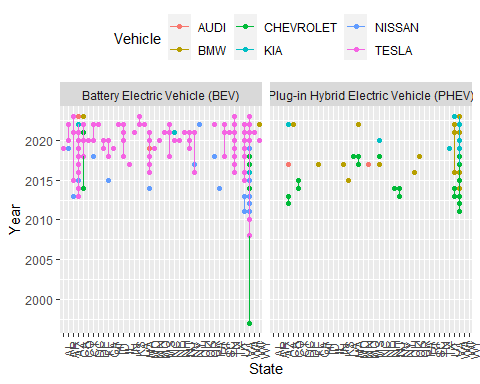
ggplot(data , aes(x = Year , y = Vehicle , color = VehicleType)) +  
 geom\_point() +  
 theme(axis.text.x = element\_text(angle = 90) , legend.position="top") +  
 facet\_grid(~VehicleType)

 Given the data, we can see that over the year, BEVs has gained more popularity. Among the BEVs, Tesla, Nissan, Kia, Chevrolet, BMW and Audi has been most popular among the customers. Next, let’s focus on the state.

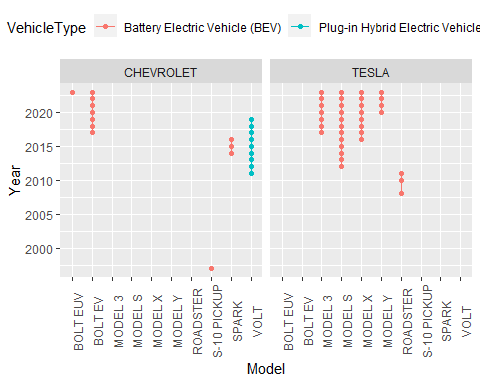
pop <- data %>%  
 filter(Vehicle %in% c('TESLA' ,'NISSAN' ,'KIA' ,'CHEVROLET','BMW','AUDI'))  
   
head(pop)

## # A tibble: 6 × 8  
## County City State Vehicle Model Year VehicleType Range  
## <chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl>  
## 1 Yakima Yakima WA TESLA MODEL 3 2020 Battery Electric Vehicl… 322  
## 2 San Diego San Diego CA TESLA MODEL 3 2019 Battery Electric Vehicl… 220  
## 3 Yakima Yakima WA TESLA MODEL X 2019 Battery Electric Vehicl… 289  
## 4 Snohomish Bothell WA BMW X5 2017 Plug-in Hybrid Electric… 14  
## 5 Snohomish Everett WA NISSAN LEAF 2015 Battery Electric Vehicl… 84  
## 6 Kitsap Poulsbo WA TESLA MODEL 3 2018 Battery Electric Vehicl… 215

ggplot(pop , aes(x = State , y = Year , color = Vehicle)) +  
 geom\_point() +  
 geom\_line() +  
 theme(axis.text.x = element\_text(angle = 90),legend.position="top") +  
 facet\_grid(~VehicleType)

 Let us now see the data in WA State. We can see that BEVs are popular and the brand Tesla and Chevrolet are the most popular. So, let’s make a data frame with that information.

WA\_data <- data %>%  
 filter(Vehicle %in% c('TESLA' ,'CHEVROLET'),  
 State %in% c('WA'))  
  
ggplot(WA\_data , aes(x = Model , y = Year , color = VehicleType)) +  
 geom\_point() +  
 geom\_line() +  
 theme(axis.text.x = element\_text(angle = 90), legend.position="top") +  
 facet\_grid(~Vehicle)



When the data are more analyzed, we can see different models of these brands and the most popular model of each brand.

From the above data, we can see that “TESLA MODEL S” is the most popular EV vehicle in WA State.

## Electric range of the most popular vehicle brand.

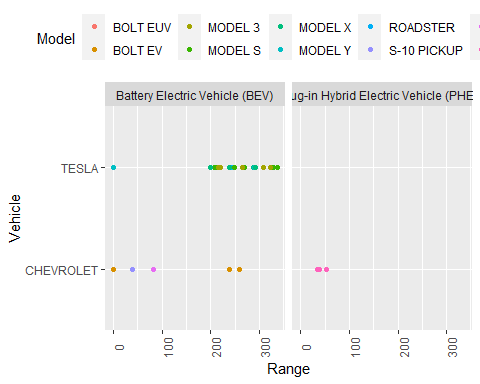
Let us start with the WA state data that we already have.

head(WA\_data)

## # A tibble: 6 × 8  
## County City State Vehicle Model Year VehicleType Range  
## <chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl>  
## 1 Yakima Yakima WA TESLA MODEL 3 2020 Battery Electric Veh… 322  
## 2 Yakima Yakima WA TESLA MODEL X 2019 Battery Electric Veh… 289  
## 3 Kitsap Poulsbo WA TESLA MODEL 3 2018 Battery Electric Veh… 215  
## 4 Snohomish Marysville WA CHEVROLET VOLT 2018 Plug-in Hybrid Elect… 53  
## 5 Snohomish Mukilteo WA TESLA MODEL S 2017 Battery Electric Veh… 210  
## 6 Thurston Olympia WA TESLA MODEL 3 2018 Battery Electric Veh… 215

Let us create a grid with the most popular two vehicles- Tesla and Chevrolet and check out their electric range.

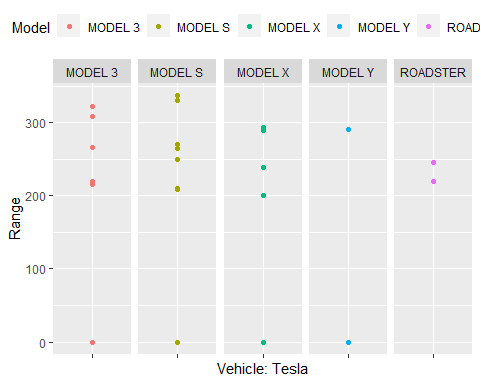
ggplot(WA\_data , aes(x = Range , y = Vehicle , color = Model)) +  
 geom\_point() +  
 theme(axis.text.x = element\_text(angle = 90) , legend.position="top") +  
 facet\_grid(~VehicleType)

 Now, let’s focus into BEVs and the brand Tesla as we can already see that TESLA EV is the most popular brand in WA State.

Tesla <- WA\_data %>%  
 filter(Vehicle %in% c('TESLA'),  
 State %in% c('WA') ,  
 VehicleType %in% c('Battery Electric Vehicle (BEV)'))  
  
head(Tesla)

## # A tibble: 6 × 8  
## County City State Vehicle Model Year VehicleType Range  
## <chr> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl>  
## 1 Yakima Yakima WA TESLA MODEL 3 2020 Battery Electric Vehicle… 322  
## 2 Yakima Yakima WA TESLA MODEL X 2019 Battery Electric Vehicle… 289  
## 3 Kitsap Poulsbo WA TESLA MODEL 3 2018 Battery Electric Vehicle… 215  
## 4 Snohomish Mukilteo WA TESLA MODEL S 2017 Battery Electric Vehicle… 210  
## 5 Thurston Olympia WA TESLA MODEL 3 2018 Battery Electric Vehicle… 215  
## 6 Thurston Olympia WA TESLA MODEL X 2017 Battery Electric Vehicle… 200

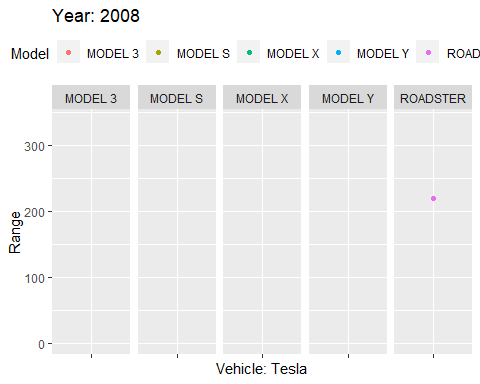
ggplot(Tesla , aes(x = Vehicle , y = Range , color = Model)) +  
 geom\_point() +  
 facet\_grid(~Model)+  
 labs(x = "Vehicle: Tesla", y = "Range") +  
 theme(axis.text.x=element\_blank(),legend.position="top")



Let’s animate it for fun.

library(gganimate)

ggplot(Tesla , aes(x = Vehicle , y = Range , color = Model)) +  
 geom\_point() +  
 facet\_grid(~Model)+  
 theme(axis.text.x=element\_blank(),legend.position="top")+  
 transition\_time(Year) +  
 labs(title = "Year: {as.integer(frame\_time)}" , x = "Vehicle: Tesla", y = "Range")+  
 shadow\_wake(wake\_length = 0.1, alpha = FALSE)

 ## Conclusion

Here, we can see that over the year, the electric range has reached above 300 by 2023. We can see the growth over the year. Still, TESLA MODEL S has gained the most popularity.

## Link to Rpubs

<https://rpubs.com/NabinaKhadka/1034893>

## Link to the video

<https://youtu.be/yLqNmy8Eebs>