Final Project (Tesla Death)

Group 4 (CDS 101)

2025-12-12

```
library(readr)
Tesla_Deaths_Deaths_3_ <- read_csv("Tesla_Deaths_Deaths_1_.csv")
```

```
## New names:
## Rows: 296 Columns: 24
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (20): Date, Country, State, Description, Tesla driver, Tesla occupant, O... dbl
## (3): Case #, Year, Deaths lgl (1): Deceased 4
## 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.
## * `` -> `...17`
## * `` -> `...18`
```

```
View(Tesla_Deaths_Deaths_3_)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.1.0
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
```

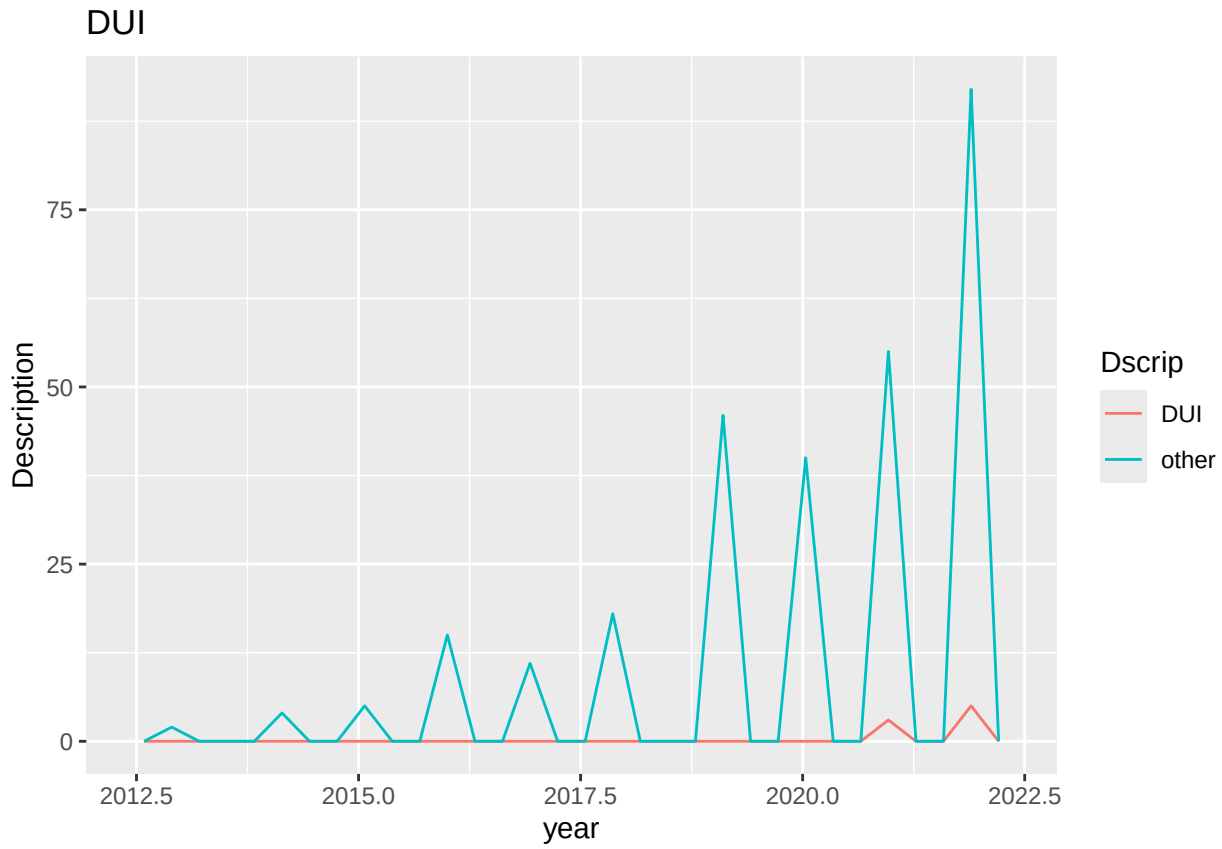
```
New_Tesla <- Tesla_Deaths_Deaths_3_ %>%
mutate(
Dscrip = if_else(
str_detect(Description, "DUI"),
"DUI",
"other"
)
)
```

I used string detect to search for any description case that had the word DUI in it to collect and categorize the DUI from the other Tesla deaths

```
view(New_Tesla)
```

```
New_Tesla %>%
ggplot() +
geom_freqpoly(aes(x = Year, color = Dscrip, )) +
labs(
title = "DUI",
x = "year",
y = "Description"
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We selected the polygraph because the line are easier to read the differences. The colors were separated based on the two outcomes from the the new column. which were Red for the DUI and blue for the other category.

```
New_Tesla %>%
  count(Dscrip)
```

```
## # A tibble: 2 x 2
##   Dscrip     n
```

```
##   <chr>  <int>
## 1 DUI        8
## 2 other    288
```

```
New_Tesla %>%
  count(Dscrip) %>%
  mutate(percent = n / sum(n) * 100)
```

```
## # A tibble: 2 x 3
##   Dscrip     n percent
##   <chr>  <int>   <dbl>
## 1 DUI        8    2.70
## 2 other    288   97.3
```

For our project we made a goal to find out how of the of the Tesla deaths were due to DUI. Our original hypothesis was the DUIs would make up 20% of the Tesla deaths.To our surprise the DUI deaths made up 3.14 percent of the Tesla deaths in the years 2013 - 2022. Looking at the trend of the data I would say the numbers may increase over time. Many of the deaths occur towards the later half of the data. I believe this is due to the increase in Tesla owners. Unfortunately we don't have access to that information.

```
New_Tesla2 <- New_Tesla %>%
  mutate(
    Year = as.numeric(Year),
    DUI_num = if_else(Dscrip == "DUI", 1,0),
    Oth_num = if_else(Dscrip == "other", 0,1)
         )
```

```
view(New_Tesla2)
```

```
DUI_Year_model <- lm(DUI_num ~ Year, data = New_Tesla2)
```

```
Oth_Year_model <- lm(Oth_num ~ Year, data = New_Tesla2)
```

```
summary(DUI_Year_model)
```

```
##
## Call:
## lm(formula = DUI_num ~ Year, data = New_Tesla2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04644 -0.04644 -0.02656 -0.01661  0.96350
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.059251   9.150928  -2.192   0.0292 *
## Year          0.009943   0.004530   2.195   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1614 on 294 degrees of freedom
## Multiple R-squared:  0.01612,    Adjusted R-squared:  0.01278
## F-statistic: 4.818 on 1 and 294 DF,  p-value: 0.02895
```

```
library(modelr)
```

```
future_years <- data.frame(
  Year = max(New_Tesla2$Year, na.rm = TRUE) + 1:10
  )
```

I created a new data set and that generated 10 extra years from the New_Tesla data set.

```
fut_predictions <- future_years %>%
add_predictions(DUI_Year_model, var = "DUI_probability") %>%
add_predictions(Oth_Year_model, var = "Other_probability")
```

I attempted several different method to create a predictive model which ended up failing, this method seem to work best for me. I thought about using the information I gained from the previous data set and using that information to create a predictive model.

```
view(fut_predictions)
```

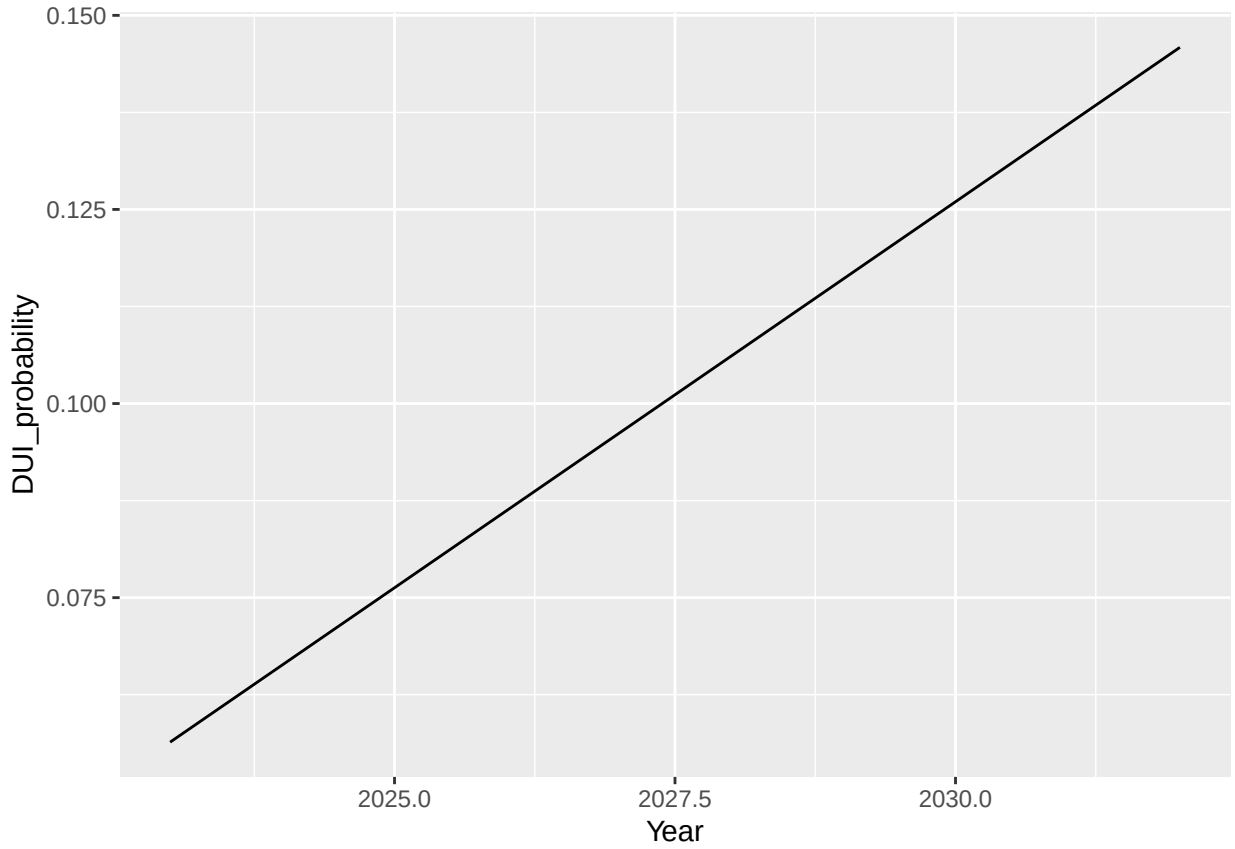
```
fut_predictions <- fut_predictions %>%
  mutate(
    Other_probability = 1 - DUI_probability,
    Predicted_DUI = if_else(DUI_probability > 0.030, "DUI", "other"),
    Predicted_Other = if_else(Other_probability > 0.97, "other", "OTH")
    )
```

I ran into some trouble with this code. When I originally the code. I started by creating the DUI probability but felt the code need more information. So I decided to add the probability of the Other deaths as well. I used the number that I received when we received the percentages of collective DUI and collective other deaths.

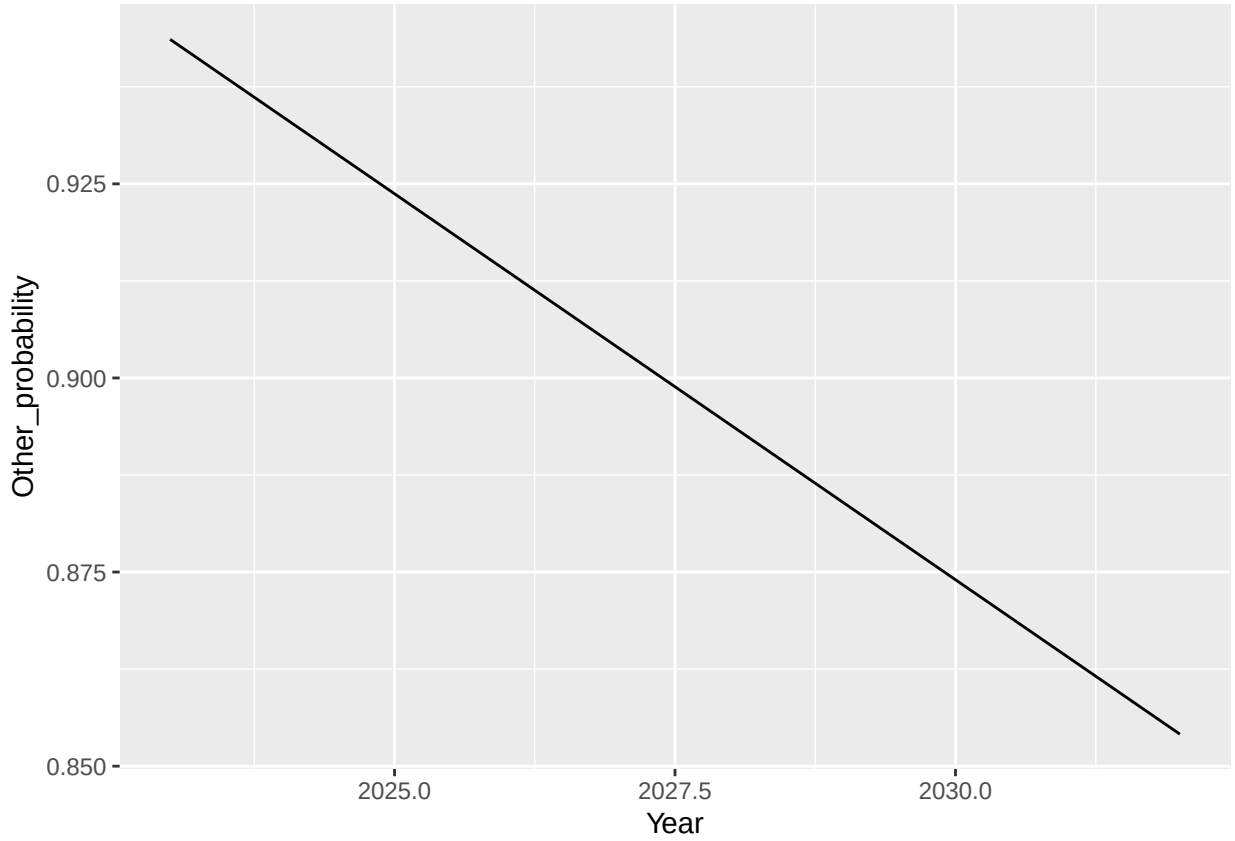
```
fut_predictions
```

```
##    Year DUI_probability Other_probability Predicted_DUI Predicted_Other
## 1  2023      0.05638714         0.9436129           DUI             OTH
## 2  2024      0.06633060         0.9336694           DUI             OTH
## 3  2025      0.07627407         0.9237259           DUI             OTH
## 4  2026      0.08621754         0.9137825           DUI             OTH
## 5  2027      0.09616101         0.9038390           DUI             OTH
## 6  2028      0.10610448         0.8938955           DUI             OTH
## 7  2029      0.11604795         0.8839520           DUI             OTH
## 8  2030      0.12599142         0.8740086           DUI             OTH
## 9  2031      0.13593489         0.8640651           DUI             OTH
## 10 2032      0.14587836         0.8541216           DUI             OTH
```

```
ggplot(fut_predictions) +
geom_line(mapping = aes(x = Year, y = DUI_probability))
```




```
ggplot(fut_predictions) +
geom_line(mapping = aes(x = Year, y = Other_probability))
```



After creating visuals for the probability this information seemed very stiff to me. So I decided to add more information to the dataset to try and improve the visual of the model that was created.

```
New_Tesla3 <- New_Tesla2 %>%
  count(Year, Year_Tally = "n")
```

```
New_Tesla2 %>%
  count(Year)
```

```
## # A tibble: 10 x 2
##     Year     n
##    <dbl> <int>
##  1  2013     2
##  2  2014     4
##  3  2015     5
##  4  2016    15
##  5  2017    11
##  6  2018    18
##  7  2019    46
##  8  2020    40
##  9  2021    58
## 10  2022    97
```

```
model_Vol <- lm(n ~ Year, data = New_Tesla3)
summary(model_Vol)
```

```
##
## Call:
## lm(formula = n ~ Year, data = New_Tesla3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.170 -10.233  -1.321   5.464  26.273
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18409.127   3093.068  -5.952 0.000341 ***
## Year             9.139      1.533   5.961 0.000338 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.93 on 8 degrees of freedom
## Multiple R-squared:  0.8162, Adjusted R-squared:  0.7933
## F-statistic: 35.54 on 1 and 8 DF,  p-value: 0.0003377
```

I made a attempt to create a model that would generate numbers of incidents similar to what happened in the original data set but failed to produce results. So after several failed attempts I decided to use a linear prediction model.

```
future3 <- tibble(Year = 2023:2032)

future_predictions_Vol <- future3 %>%
  mutate(predicted_volume = predict(model_Vol, newdata = .))

future_predictions_Vol
```

```
## # A tibble: 10 x 2
##     Year predicted_volume
##    <int>            <dbl>
##  1  2023             79.9
##  2  2024             89.0
##  3  2025             98.1
##  4  2026            107.
##  5  2027            116.
##  6  2028            126.
##  7  2029            135.
##  8  2030            144.
##  9  2031            153.
## 10  2032            162.
```

Looking at the new data set I knew the information was incomplete due to the numbers being decimals. We can't use this numbers for the data set so I rounded the number and placed the data into a new data set.

```
future4 <- tibble(Year = 2023:2032)

future_predictions_Vol2 <- future3 %>%
  mutate(
  predicted_volume = predict(model_Vol, newdata = .),
  predicted_volume = round(predicted_volume)
```

```
  )
future_predictions_Vol2
```

```
## # A tibble: 10 x 2
##     Year predicted_volume
##    <int>            <dbl>
##  1  2023               80
##  2  2024               89
##  3  2025               98
##  4  2026              107
##  5  2027              116
##  6  2028              126
##  7  2029              135
##  8  2030              144
##  9  2031              153
## 10  2032              162
```

```
combined_prediction <- fut_predictions %>%
  left_join(future_predictions_Vol2, by = "Year")
```

After completing the death total predictions I decided to combined data sets to place all the information in one place.

```
combined_prediction
```

```
##    Year DUI_probability Other_probability Predicted_DUI Predicted_Other
## 1  2023      0.05638714         0.9436129           DUI             OTH
## 2  2024      0.06633060         0.9336694           DUI             OTH
## 3  2025      0.07627407         0.9237259           DUI             OTH
## 4  2026      0.08621754         0.9137825           DUI             OTH
## 5  2027      0.09616101         0.9038390           DUI             OTH
## 6  2028      0.10610448         0.8938955           DUI             OTH
## 7  2029      0.11604795         0.8839520           DUI             OTH
## 8  2030      0.12599142         0.8740086           DUI             OTH
## 9  2031      0.13593489         0.8640651           DUI             OTH
## 10 2032      0.14587836         0.8541216           DUI             OTH
##    predicted_volume
## 1                80
## 2                89
## 3                98
## 4               107
## 5               116
## 6               126
## 7               135
## 8               144
## 9               153
## 10              162
```

I took a look at the information and felt as though the probability wouldn't be good enough on it's own so I decided to multiply the probability of the DUI and Other by the total number of deaths that year to get the predicted outcome.

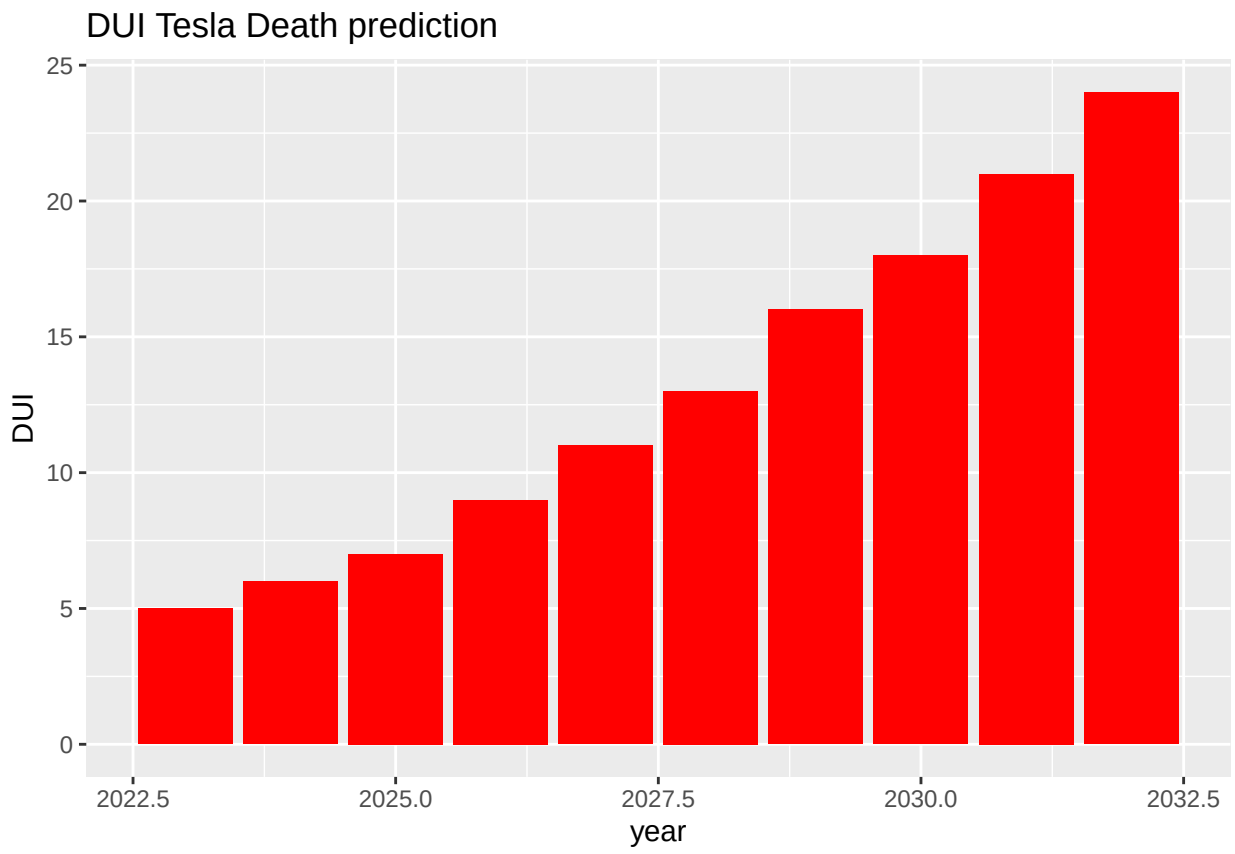
```
combined_prediction2 <- combined_prediction %>%
  mutate(
    DUI_Results = round(DUI_probability * predicted_volume),
    OTH_Results = round(Other_probability * predicted_volume)
  )
```

```
combined_prediction2
```

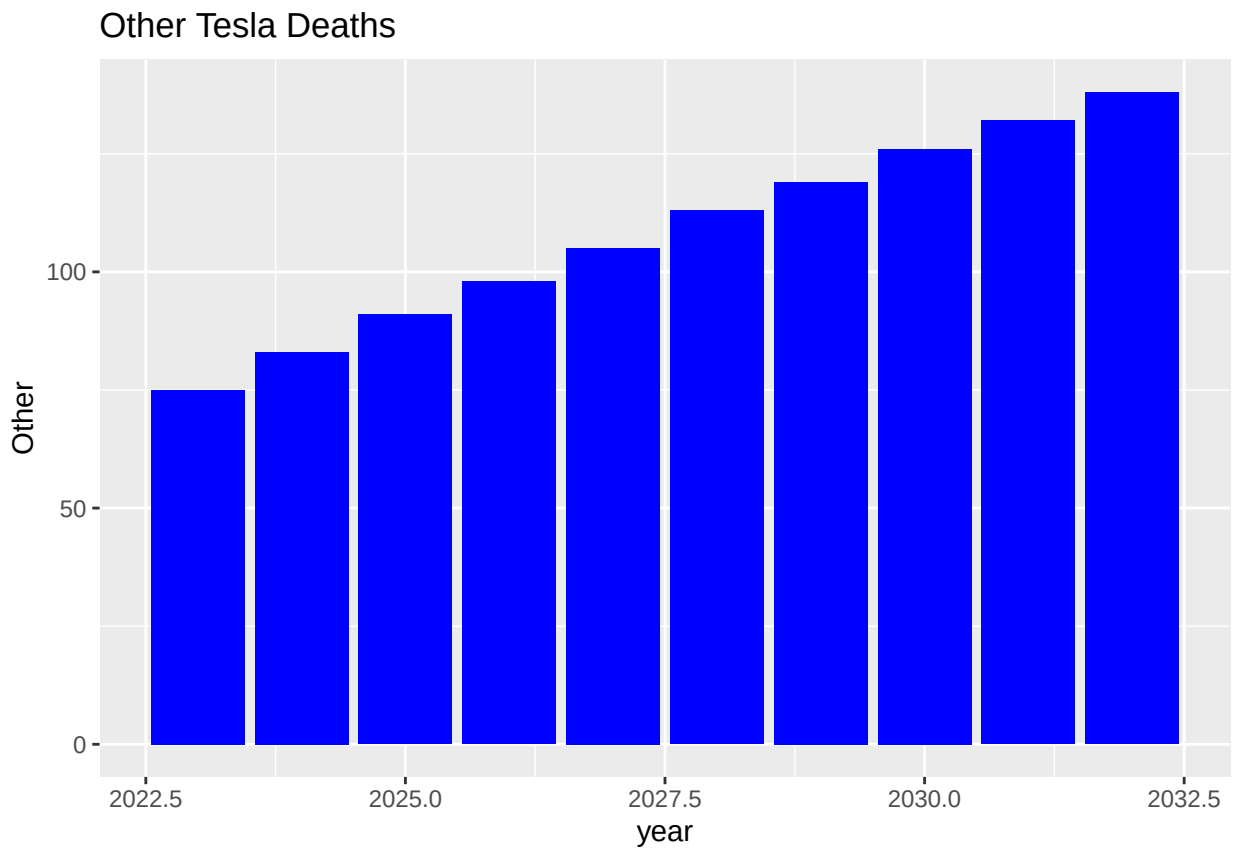
```
##    Year DUI_probability Other_probability Predicted_DUI Predicted_Other
## 1  2023      0.05638714         0.9436129           DUI             OTH
## 2  2024      0.06633060         0.9336694           DUI             OTH
## 3  2025      0.07627407         0.9237259           DUI             OTH
## 4  2026      0.08621754         0.9137825           DUI             OTH
## 5  2027      0.09616101         0.9038390           DUI             OTH
## 6  2028      0.10610448         0.8938955           DUI             OTH
## 7  2029      0.11604795         0.8839520           DUI             OTH
## 8  2030      0.12599142         0.8740086           DUI             OTH
## 9  2031      0.13593489         0.8640651           DUI             OTH
## 10 2032      0.14587836         0.8541216           DUI             OTH
##    predicted_volume DUI_Results OTH_Results
## 1                80           5          75
## 2                89           6          83
## 3                98           7          91
## 4               107           9          98
## 5               116          11         105
## 6               126          13         113
## 7               135          16         119
## 8               144          18         126
## 9               153          21         132
## 10              162          24         138
```

Looking over the information I felt this was good enough to graph.

```
combined_prediction2 %>%
ggplot(aes(x = Year, y = DUI_Results)) +
geom_col(fill ="red") +
labs(
title = "DUI Tesla Death prediction",
x = "year",
y = "DUI"
)
```




```
combined_prediction2 %>%
ggplot(aes(x = Year, y = OTH_Results)) +
geom_col(fill ="blue") +
labs(
title = "Other Tesla Deaths",
x = "year",
y = "Other"
)
```




```
combined_graph <- combined_prediction2 %>%
select(Year, DUI_Results, OTH_Results) %>%
  pivot_longer(
    cols = c(DUI_Results, OTH_Results),
    names_to = "Results",
    values_to = "Deaths"
  )
```

I combined the death results into one column to use as a comparison graph.

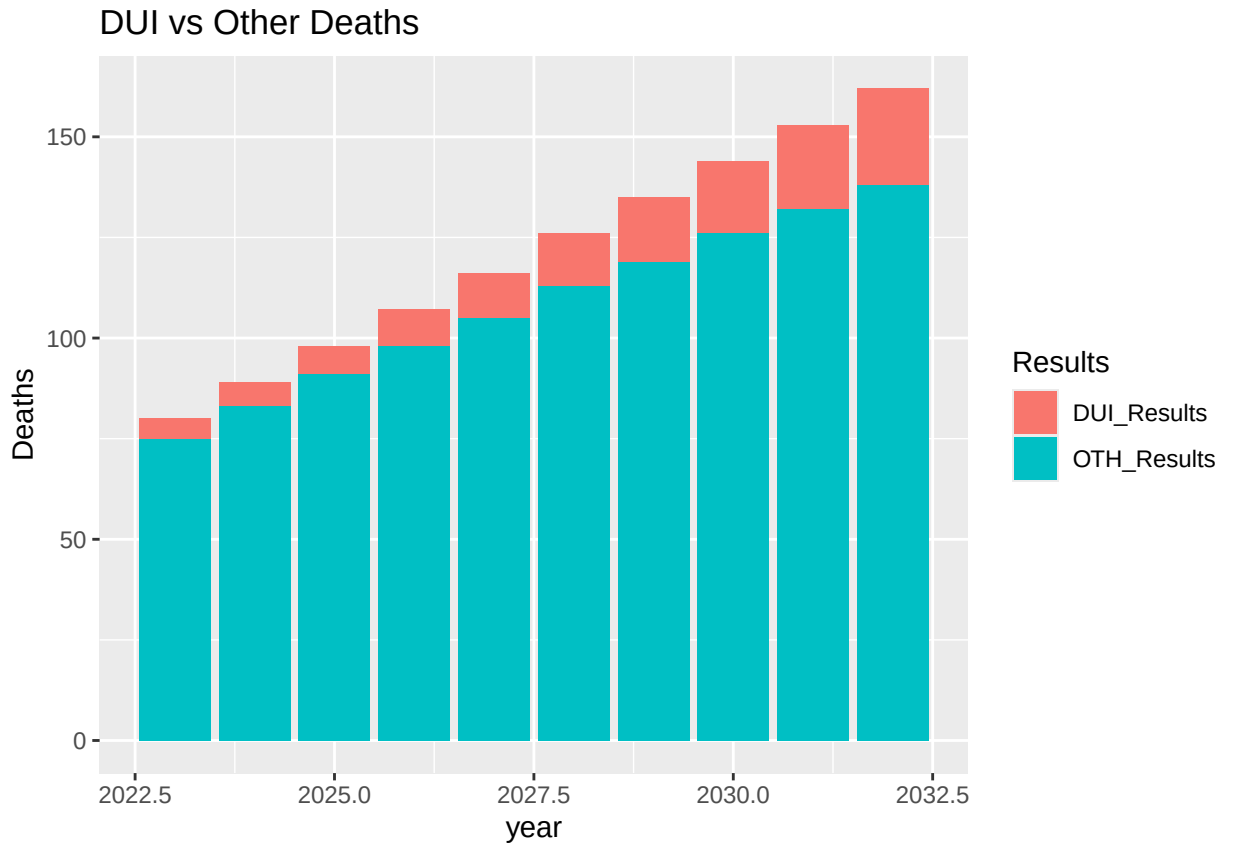
```
combined_graph
```

```
## # A tibble: 20 x 3
##     Year Results     Deaths
##    <dbl> <chr>        <dbl>
##  1  2023 DUI_Results      5
##  2  2023 OTH_Results     75
##  3  2024 DUI_Results      6
##  4  2024 OTH_Results     83
##  5  2025 DUI_Results      7
##  6  2025 OTH_Results     91
##  7  2026 DUI_Results      9
##  8  2026 OTH_Results     98
##  9  2027 DUI_Results     11
## 10  2027 OTH_Results    105
```

```
## 11  2028 DUI_Results     13
## 12  2028 OTH_Results    113
## 13  2029 DUI_Results     16
## 14  2029 OTH_Results    119
## 15  2030 DUI_Results     18
## 16  2030 OTH_Results    126
## 17  2031 DUI_Results     21
## 18  2031 OTH_Results    132
## 19  2032 DUI_Results     24
## 20  2032 OTH_Results    138
```

The graph below shows the total deaths which parts are DUI compared to the other deaths. I actually wanted to have the DUI at the bottom and the other at the top, however I got burnout from the chunks of code and left as is.

```
combined_graph %>%
ggplot() +
geom_col(aes(x = Year, y = Deaths, fill = Results, )) +
labs(
title = "DUI vs Other Deaths",
x = "year",
y = "Deaths"
)
```



```
Pred_Summary <- combined_graph %>%
group_by(Year) %>%
  summarize(
    Total_Deaths = sum(Deaths),
    DUI_Deaths = sum(Deaths[Results == "DUI_Results"]),
    Other_Deaths = sum(Deaths[Results == "OTH_Results"]),
  )
```

```
Pred_Summary
```

```
## # A tibble: 10 x 4
##     Year Total_Deaths DUI_Deaths Other_Deaths
##    <dbl>        <dbl>      <dbl>        <dbl>
##  1  2023           80          5           75
##  2  2024           89          6           83
##  3  2025           98          7           91
##  4  2026          107          9           98
##  5  2027          116         11          105
##  6  2028          126         13          113
##  7  2029          135         16          119
##  8  2030          144         18          126
##  9  2031          153         21          132
## 10  2032          162         24          138
```

```
Pred_Summary2 <- Pred_Summary %>%
  mutate(DUI_percent = DUI_Deaths / Total_Deaths * 100) %>%
  mutate(OTH_percent = Other_Deaths / Total_Deaths * 100)
```

```
Pred_Summary2
```

```
## # A tibble: 10 x 6
##     Year Total_Deaths DUI_Deaths Other_Deaths DUI_percent OTH_percent
##    <dbl>        <dbl>      <dbl>        <dbl>       <dbl>       <dbl>
##  1  2023           80          5           75        6.25        93.8
##  2  2024           89          6           83        6.74        93.3
##  3  2025           98          7           91        7.14        92.9
##  4  2026          107          9           98        8.41        91.6
##  5  2027          116         11          105        9.48        90.5
##  6  2028          126         13          113       10.3         89.7
##  7  2029          135         16          119       11.9         88.1
##  8  2030          144         18          126       12.5         87.5
##  9  2031          153         21          132       13.7         86.3
## 10  2032          162         24          138       14.8         85.2
```

The previous 4 chunks of code were actually an error that I made but this results seemed interesting so I decided to keep it. I meant to get a summary of the collective DUI deaths and Other deaths.

```
Pred_Summary3 <- Pred_Summary2 %>%
  summarise(
  SumTDeath = sum(Total_Deaths),
  SumTDUI = sum(DUI_Deaths),
  SumTOTH = sum(Other_Deaths)
  )
```

I finished the code and found out according to our prediction the DUI would go up to 14 - 15% of Tesla deaths with in the range of 2023 - 2032.

```
Pred_Summary3%>%
mutate(T_DUI_percent = SumTDUI / SumTDeath) %>%
mutate(T_OTH_percent = SumTOTH / SumTDeath)
```

```
## # A tibble: 1 x 5
##   SumTDeath SumTDUI SumTOTH T_DUI_percent T_OTH_percent
##       <dbl>   <dbl>   <dbl>         <dbl>         <dbl>
## 1      1210     130    1080         0.107         0.893
```