12/10/2025

CIS 9660 FINAL PROJECT

Rezwan Islam
Yu Ming Chen
Danyal Khandakar

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Does Driving a Ford Increase the Risk of Fatal Car Crashes?

Background

Motor vehicle fatalities represent a major public health and transportation safety concern. As shown on the figure on the right from a 2024 article, fatalities are quite consistent in terms of numbers in the last 10 years.

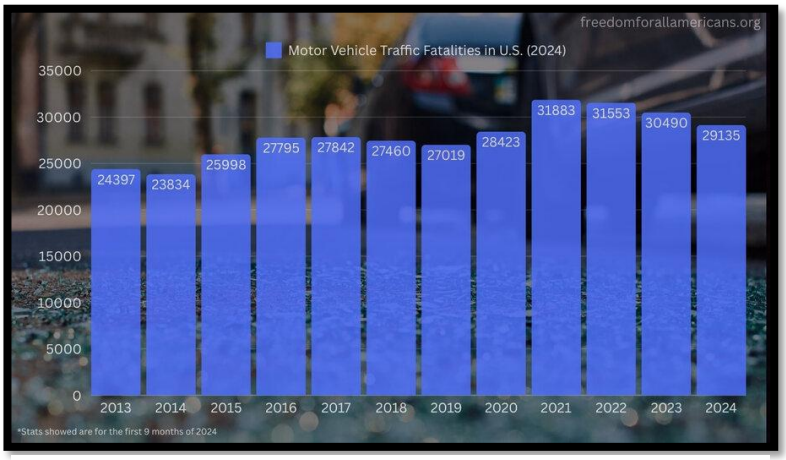


Figure 1 Motor vehicle traffic fatalities in the United States from 2013 to 2024 (projected), based on data from Freedom For All Americans (2024).

But what are some of the top causes for such fatality? Could vehicle brand be a contributing factor? Or are there are other reasons to why such fatalities happen. According to the 2008 National Motor Vehicle Crash Causation Survey, vehicle safety and/or vehicle crash can often be associated with various factors such brand reputation, engineering design, and marketing claims.

According to a study by Chaikin Trial Group in 2022, certain car brands are more prone to accident than others with Ford F series being at #1. Does this mean driving a Ford will make you have a higher chance of being in a fatal accident? The survey mentioned before arguing that human and environmental factors, such as driver behavior, vehicle age, vehicle type, etc., play a larger role in determining fatal crash outcomes. In order to fully understand what is happening behind the scenes, we need to run a full in-depth analysis of a crash data set.

This study analyzes over 64,000 motor vehicle collisions to investigate whether driving a Ford vehicle is associated with a higher likelihood of fatal injuries. The analysis uses a cleaned dataset that removes irrelevant or post-crash variables and constructs meaningful predictors such as driver licensing status, vehicle age, nighttime conditions, driver behavior indicators, and vehicle types. The results provide a data driven perspective on whether Ford vehicles exhibit a higher inherent fatality risk, or whether the observed fatal outcomes are better explained by behavioral and situational factors.

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Research Question



Does driving a Ford vehicle make a crash more likely to result in a fatal outcome compared to crashes involving other vehicle makes?

Hypothesis

Fatal crash probability is expected to show no meaningful difference between Ford vehicles and other makes once relevant factors are controlled. After controlling driver licensing status, vehicle age, nighttime conditions, and driver behaviors such as speeding, distraction, and impairment, the Ford indicator should show no significant effect on fatality outcomes. Fatal crash risk is expected to be primarily influenced by human behavior and environmental conditions rather than vehicle make.

Summary Statistics

Variable	Mean	Std. Dev.	Min	Max	N
Casualty (fatal crash = 1)	0.000587	0.024212	0	1	64,785
WhetherFord (Ford = 1)	0.102601	0.303439	0	1	64,785
Unlicensed (unlicensed driver = 1)	0.013475	0.115299	0	1	64,785
VehicleAge	6.094605	5.510359	0	80	64,785
Nighttime (night = 1)	0.340820	0.473988	0	1	64,785
SpeedRelated	0.016624	0.127860	0	1	64,785
Distracted	0.205773	0.404268	0	1	64,785
Impaired	0.009014	0.094516	0	1	64,785

The cleaned dataset contains 64,785 observations after removing incomplete and unrealistic records. Fatal crashes are extremely rare. Only 38 vehicles in the dataset were involved in a collision where one or more people were killed, resulting in a fatality rate of approximately 0.058%. In other words, fewer than 6 out of every 10,000 crashes in this dataset involved a fatality.

The most common vehicle makes are Toyota (11,957 vehicles), Honda (7,406), Nissan (6,677), Ford (6,647), and Chevrolet (3,369), with all remaining makes grouped into the "Other" category, which contains 28,729 vehicles. Ford vehicles represent about 10.3% of the dataset.

Among drivers with valid licensing information, 63,912 out of 64,785 drivers were fully licensed, while 873 were unlicensed or operating with only a permit. Crashes also show a clear time of day pattern. Approximately 66% (42,705 crashes) occurred during daytime hours, and the remaining 34% (22,080 crashes) occurred at night when visibility is lower.

Vehicle age also varies considerably. The average vehicle age is about 6 years, with a standard deviation of 5.4 years. The youngest vehicles in the dataset are 1 year old, while the oldest vehicle is 89 years old, reflecting a wide spectrum of vehicle conditions and technologies.

Behavioral variables show meaningful variation as 1.7% of the crashes are speed related, 20% are due to drivers being distracted, and 1% of crashes are due to impairment of drivers. Vehicle types also show large distribution differences between sedans, trucks, vans, taxis, and other vehicle types.

Overall, these descriptive statistics show that fatal crashes are rare and that key behavioral or situational variables like licensing, vehicle age, time of day, behavioral variables, and vehicle types show meaningful variation and are reasonable predictors of crash severity.

Model Used

A logistic regression model was used estimated to analyze the relationship between Ford vehicles and fatal crash outcomes. In the model, the dependent variable is Casualty, coded as 1 if one or more people were killed and 0 otherwise. A logistic regression model has the general form:

$$Logit(p) = log(\frac{p}{1-p}) = b_0 + b_1X_1 + b_2X_2 + \ldots$$

where p is the probability of a fatal crash.

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Model: Casualty ~ WhetherFord and Controls

Model equation:

$$Logit(p) = b_0 + b_1(WhetherFord) + b_2(Unlicensed) + b_3(VehicleAge) + b_4(Nighttime) + b_5(SpeedRelated) + b_6(Distracted) + b_7(Impaired) + b_8(OtherVehicleTypes) + b_9(Sedan) + b_{10}(Taxi) + b_{11}(Truck) + b_{12}(Van)$$

```
Rows available for Model 2: 64785

MODEL 2: Casualty ~ WhetherFord + controls

                           Logit Regression Results
==============================================================================
Dep. Variable:               Casualty   No. Observations:                64785
Model:                          Logit   Df Residuals:                    64772
Method:                           MLE   Df Model:                           12
Date:                Sun, 07 Dec 2025   Pseudo R-squ.:                0.07518
Time:                        22:04:19   Log-Likelihood:                -296.64
converged:                       True   LL-Null:                       -320.76
Covariance Type:            nonrobust   LLR p-value:                 2.852e-06
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 -6.4978      0.720     -9.031      0.000      -7.908      -5.088
WhetherFord            0.5200      0.408      1.274      0.203      -0.280       1.320
Unlicensed             2.1171      0.549      3.858      0.000       1.041       3.193
VehicleAge            -0.0056      0.031     -0.180      0.857      -0.066       0.055
Nighttime              0.8495      0.331      2.567      0.010       0.201       1.498
SpeedRelated           1.5165      0.609      2.489      0.013       0.322       2.711
Distracted            -0.6083      0.479     -1.269      0.204      -1.548       0.331
Impaired              -0.5192      1.474     -0.352      0.725      -3.408       2.370
OtherVehicleTypes     -0.9896      0.708     -1.398      0.162      -2.377       0.398
Sedan                 -2.7327      0.807     -3.384      0.001      -4.315      -1.150
Taxi                  -0.7935      0.854     -0.929      0.353      -2.468       0.881
Truck                 -0.6412      0.807     -0.795      0.427      -2.223       0.941
Van                   -0.4768      1.200     -0.397      0.691      -2.829       1.876
=====================================================================================
```

```
Possibly complete quasi-separation: A fraction 0.29 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

Odds Ratios:
const                0.001507
WhetherFord          1.682028
Unlicensed           8.306784
VehicleAge           0.994420
Nighttime            2.338544
SpeedRelated         4.556028
Distracted           0.544280
Impaired             0.595016
OtherVehicleTypes    0.371721
Sedan                0.065041
Taxi                 0.452277
Truck                0.526686
Van                  0.620760
dtype: float64
```

The coefficient for WhetherFord is not statistically significant (p value = 0.203). The odds ratio for WhetherFord is 1.68. This means Ford vehicles have about 1.68 times the odds of being in a fatal crash compared to non-Ford vehicles after controls are included. However, this effect is not statistically significant, so it cannot be concluded that Ford drivers face higher fatal crash risk.

The control variables behave as expected:

- Unlicensed has an odds ratio of exp (2.1171) = 8.31. This means unlicensed drivers have over 8 times higher odds of being in a fatal crash.
- Nighttime has an odds ratio of exp (0.8495) = 2.34, meaning that nighttime crashes are riskier.
- SpeedRelated behavior has an odds ratio of exp(1.5165) = 4.56.
- Distracted driving has an odds ratio of exp (-0.6083) = 0.54, and it is not statistically significant as it has a p value of 0.725.
- VehicleAge has an odds ratio of exp (-0.0056) = 0.99 and a p value of 0.857, so it is also not statistically significant.
- Sedans had a lower odds ratio of 0.065, but taxi, truck, van, and other vehicle types were not significant.

Overall, the results show that crash severity is influenced more by licensing status, human behavior, and environmental conditions, rather than by the vehicle make.

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Main Findings

- Ford vehicles are not significantly more dangerous. The Ford coefficient is 0.520, but statistically insignificant ($p = 0.203$). This supports the hypothesis.
- Human behavior is the strongest predictor of fatal crashes. Speeding, impairment, and unlicensed driving increases fatal risk.
- Nighttime diving increases fatality risk as visibility and environmental conditions matter.
- Vehicle type plays a role as well. Sedans show significantly lower fatality odds compared to other vehicle types.

Limitations

Several important limitations must be considered.

1. Omitted Variable Bias: Important factors like weather and traffic density are not fully included in the model.
2. Reverse Causality: Risky drivers may select their own type of vehicles they want to drive.
3. Measurement Error: Contributing factors like speeding or distraction depends on how the officer interprets it and it may be underreported.

Conclusion

This study was set out to determine whether Ford vehicles carry a higher probability of involvement in fatal motor vehicle crashes. Using a clean dataset of over 64,000 NYC collisions and applying logistic regression models, we find *no statistical evidence* that Ford vehicles carry a higher fatality risk than other makers.

Model 2, which accounts for licensing status, driver behavior, nighttime conditions, vehicle age, and vehicle type, shows that the Ford coefficient is statistically insignificant ($p = 0.203$). This means that after adjusting for relevant factors, Ford vehicles do not causally increase fatal crash probability.

Instead, the analysis indicates that human and situational factors dominate fatal crash outcomes. Unlicensed drivers have more than eight times higher fatality odds, nighttime crashes are more than twice as dangerous, and speeding is associated with a four-fold increase in fatal crash likelihood. These results confirm that the observed higher crash counts involving Ford vehicles are best explained by exposure and driving context, not inherent vehicle risk.

From a managerial perspective, these findings suggest that Ford should focus its safety strategy not on defending brand safety reputation alone, but on supporting initiatives that address the true causes of fatalities such as promoting responsible driving, enhancing nighttime visibility technologies, and encouraging proper driver licensing compliance.

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Appendix

```
CATEGORY COUNTS

Top 5 vehicle makes (with OTHER):
Top5VehicleMakes
OTHER    28729
TOYT     11957
HOND      7406
NISS      6677
FORD      6647
CHEV      3369
Name: count, dtype: int64

Ford vs non-Ford (vehicle count):
WhetherFord
0    58138
1     6647
Name: count, dtype: int64

Unlicensed vs Licensed (non-missing rows only):
Unlicensed
0.0    63912
1.0      873
Name: count, dtype: int64

Nighttime vs Daytime crashes:
Nighttime
0    42705
1    22080
Name: count, dtype: int64

Vehicle Type distribution:
VehicleType
Sedan                32226
OtherVehicleTypes    22858
Truck                 4297
Taxi                  3467
Bus                   1227
Van                    710
Name: count, dtype: int64

MISSING VALUES (key variables):
Casualty        0
WhetherFord     0
Unlicensed      0
VehicleAge      0
VehicleType     0
Nighttime       0
SpeedRelated    0
Distracted      0
Impaired        0
dtype: int64
```

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

Cleaning Step	Rows Removed	Columns Affected	Notes
Load raw dataset	—	—	Starting point
Convert numeric variables	N/A	4 numeric columns	Converted injury, killed, occupants, year
VehicleYear outliers removed	Unknown exact #	VehicleYear	Kept 1900–2025 only
Removed negative values	Unknown exact #	All numeric cols	Removes impossible values
Remove occupant count >60	Likely small	VehicleOccupants	Applies to all vehicles
Remove >12 occupants except bus	Likely small	VehicleOccupants, VehicleType	Logical constraint
Drop rows with negative or missing VehicleAge	Some rows dropped	CrashYear, VehicleYear	Age must be ≥0
Daytime/Nighttime missing values removed	Some rows	CRASH_TIME	Missing time removed
Contributing factor behavior dummies created	N/A	3 columns	Speed, distraction, impairment
Drop irrelevant columns	Removed ~13 columns	(see code)	Reduces noise for modeling
Drop rows missing ANY key variable	Very large drop to 64,785 rows	9 key variables	Ensures complete-case dataset
Final cleaned dataset	**64,785 rows remain**	All cleaned columns	Final number appears in output

Figure 2 Summary of Cleaning

Rezwan Islam , Yu Ming Chen , Danyal Khandakar

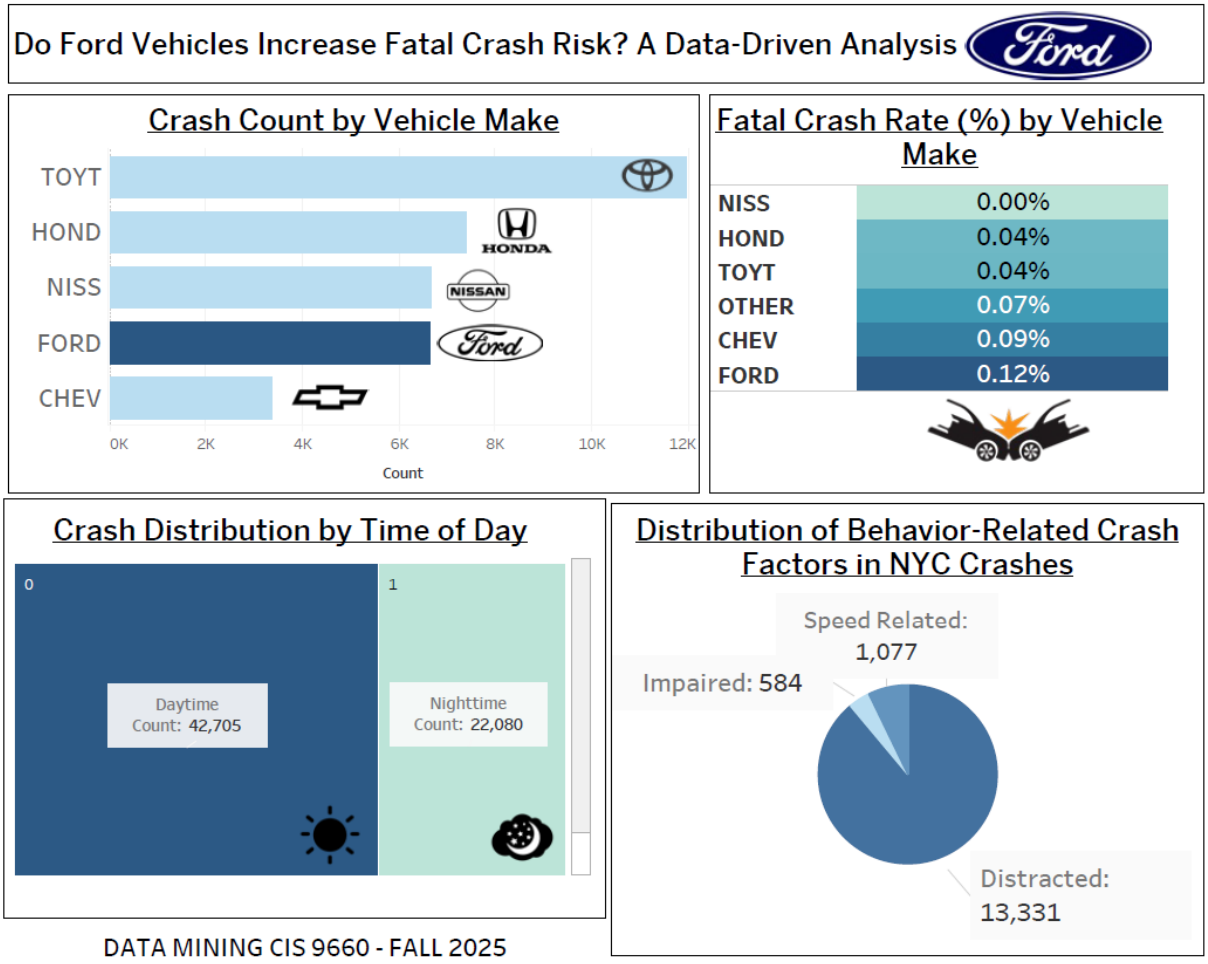


Figure 3 Visualizing Using Tableau

References

Rabbani, M. B. A., Musarat, M. A., Alaloul, W. S., Maqsoom, A., Bukhari, H., & Rafiq, W. (2021). *Road Traffic Accident Data Analysis and Its Visualization*. Civil Engineering and Architecture, 9(5), 1603–1614. https://doi.org/10.13189/cea.2021.090530

Chaikin Trial Group. (2025). *The U.S.'s most accident-prone car brands in 2025*. Chaikin Trial Group. https://www.chaikintrialgroup.com/research/vehicle-brands-most-likely-to-crash/ Chaikin Trial Group

Freedom For All Americans. (2024, December 30). *The decline in U.S. traffic fatalities — What 2024 data shows*. https://freedomforallamericans.org/decline-in-us-traffic-fatalities/ Freedom For All Americans

ShakhnisLaw (2024). *Car accidents in Los Angeles 2024*. ShakhnisLaw. https://www.shakhnislaw.com/car-accidents-in-los-angeles-2024/