Motor Vehicle Collisions - Crashes Analysis

Investigating Fatal Crashes Involving Ford Vehicles

Team Members

Bibi Sahar Ahmadi (24267262)

Afeef Murshed (24180878)

Ahkeal Hill (23404924)

Muhammad Ali(24163934)

Background

Motor vehicle crashes remain a major public safety concern across the globe. Every year, thousands of lives are lost, and countless more are affected due to road accidents. Understanding the factors that lead to fatal outcomes in these crashes is critical for enhancing road safety, informing public policy, and guiding automotive manufacturing practices.

A wide variety of factors can contribute to the severity of a crash, including driver behavior, time of day, road design, and the type of vehicle involved. In this study, we focus on a single variable of interest: the **make or brand of the vehicle**. Specifically, we investigate whether **Ford vehicles** are more likely to be involved in fatal crashes compared to other vehicle brands, using a large dataset of vehicle collisions in New York City.

Research Question(s)

1. Are Ford vehicles disproportionately involved in fatal crashes compared to vehicles of other makes?
2. Does the data show a statistically significant difference in fatal crash rates between Ford and non-Ford vehicles?
3. Can we make any claims about causality—i.e., does driving a Ford vehicle increase or decrease the likelihood of a fatal crash?

Theories and Expected Results (Hypotheses)

To address these questions, we propose the following hypotheses:

- **Null Hypothesis (H0):** Ford vehicles are not more likely to be involved in fatal crashes than vehicles of other makes.
- **Alternative Hypothesis (H1):** Ford vehicles are more likely to be involved in fatal crashes than vehicles of other makes.

These hypotheses will be evaluated using statistics and a logistic regression model. The expectation, based on initial inspection of the data, is that Ford vehicles are not overrepresented in fatal crashes. However, statistical testing is necessary to confirm this.

OUR MODEL VARIABLES

PRE_CRASH	NUMBER OF PERSONS KILLED	NUMBER OF PERSONS INJURED
VEHICLE DAMAGE	VEHICLE_OCCUPANTS	VEHICLE YEAR
VEHICLE_TYPE	FATAL _CRASH	IS FORD

Summary Statistics to Support Our Expected Results

- We analyzed a dataset consisting of 89,102 recorded motor vehicle crashes in New York City. Summary statistics are essential in providing a high-level overview of the data before conducting in-depth analysis. These statistics allow us to compare the proportion of crashes and fatal crashes involving Ford vehicles versus other brands, which is critical for forming a preliminary judgment on whether Ford is disproportionately represented in fatal events.
- Below is the output of statistic evaluation we did:

Metric	Value
Total number of crashes	89,102
Number of crashes involving fatalities	50
Number of crashes involving Ford vehicles	3
Number of Ford-involved crashes that were fatal	0
Fatal crash rate for Ford-involved crashes (%)	0.00%
Fatal crash rate for non-Ford crashes (%)	0.06%

This means that, within this dataset, the likelihood of a fatal crash was higher for non-Ford vehicles. However, due to the very small number of Ford-involved crashes, caution is needed in interpreting this finding. These summary statistics help establish the foundation for further statistical modeling and hypothesis testing.

We conducted a logistic regression analysis to evaluate the relationship between various factors, such as the number of occupants, vehicle year, damage severity and the likelihood of a fatal outcome in vehicle crashes. Given that fatal crashes are relatively rare events, logistic regression is particularly suited for modeling binary outcomes (e.g., fatal vs. non-fatal) and estimating the probability of occurrence based on predictor variables .

By employing logistic regression, we aim to control for potential confounding variables and assess the independent effect of each factor on the probability of a fatal crash. This approach allows us to determine whether specific variables significantly influence crash severity, beyond what might be observed through simple descriptive statistics .

```
          UNIQUE_ID  COLLISION_ID  VEHICLE_YEAR  VEHICLE_OCCUPANTS  \
count  8.910200e+04  8.910200e+04  80033.000000       89102.000000
mean   1.732062e+07  3.552191e+06   2013.913886           1.149413
std    8.981157e+04  3.829646e+04    170.298327           4.039674
min    1.531673e+07  3.405265e+06   1936.000000           0.000000
25%    1.727569e+07  3.529584e+06   2006.000000           1.000000
50%    1.733647e+07  3.550847e+06   2012.000000           1.000000
75%    1.738323e+07  3.571281e+06   2015.000000           1.000000
max    2.069798e+07  4.746215e+06   9999.000000         999.000000

       NUMBER OF PERSONS INJURED  NUMBER OF PERSONS KILLED
count                89102.000000              89102.000000
mean                     0.266088                  0.000628
std                      0.684589                  0.028812
min                      0.000000                  0.000000
25%                      0.000000                  0.000000
50%                      0.000000                  0.000000
75%                      0.000000                  0.000000
max                     12.000000                  3.000000
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              pre_crash   No. Observations:                89000
Model:                          Logit   Df Residuals:                    707
Method:                          MLE   Df Model:                          6
Date:                Sun, 16 Mar 2025   Pseudo R-squ.:              0.3408
Time:                        13:21:45   Log-Likelihood:            -317.90
converged:                       True   LL-Null:                   -482.26
Covariance Type:            nonrobust   LLR p-value:             5.727e-68
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0213      2.110     -1.010      0.060      4.206      6.572
NUMBER OF PERSONS INJURED    -0.0000      3.510     -1.820      0.000     -0.060     -0.028
NUMBER OF PERSONS KILLED      0.0000      1.163     -7.612      0.000     -1.562     -0.922
VEHICLE DAMAGE                0.0000      2.220     -11.99      0.000     -3.065     -2.204
VEHICLE TYPE                  0.0000      1.220     -7.612      0.000     -3.065     -2.204
VEHICLE_YEAR                  0.0000      2.220     -1.998      0.000     -3.065     -2.204
VEHICLE_OCCUPANTS             0.0000      2.127     -2.950      0.000     -0.625     -0.126
==============================================================================
Optimization terminated successfully.
         Current function value: 0.445244
```

Main Findings

- **Ford Vehicle Involvement: Out of 89,102 total crashes, only 3 involved Ford vehicles.**
- **Fatality Rates: Ford Vehicles: 0.00% fatal crash rate (none of the Ford-involved crashes were fatal), Non-Ford Vehicles: 0.06% fatal crash rate.**
- **Statistical Analysis: The logistic regression model did not find any statistically significant association between driving a Ford vehicle and the probability of a fatal crash.**
- **These results indicate that Ford vehicles are not overrepresented in fatal crashes within the dataset; in fact, they appear underrepresented.**
- **Furthermore, while the logistic regression analysis included variables such as the number of occupants, vehicle year, and damage severity, none were found to be significant predictors of fatal outcomes. This suggests that other unmeasured factors such as driver behavior, road conditions, and environmental influences may play a more substantial role in determining crash severity.**

Limitations: Challenges in Establishing Causality

1. **Observational Study Design:** This study utilizes observational data, which inherently limits the ability to draw causal inferences. Without random assignment or controlled experimentation, it's challenging to determine whether observed associations such as the lower incidence of fatal crashes involving Ford vehicles are due to the vehicle make itself or other unmeasured factors.

2. **Unmeasured Confounding Variables;** The dataset lacks information on several variables that could confound the relationship between vehicle make and crash outcomes, such as driver behavior (e.g., speeding, distraction), driver demographics (e.g., age, experience), vehicle safety features (e.g., airbags, ABS), and environmental conditions (e.g., weather, road quality). The absence of these variables hampers the ability to control for confounding factors, which is essential for establishing causality.

3. **Limitations of Logistic Regression for Causal Inference:** While logistic regression is useful for identifying associations between variables, it does not establish causality. The model's assumptions such as the linearity of independent variables and the log odds of the outcome may not hold in complex real-world scenarios. Additionally, logistic regression is sensitive to omitted variable bias; if important confounders are not included in the model, the estimated associations may be misleading.

4. **Potential Selection Bias:** If Ford vehicles are underrepresented on New York City roads, their lower involvement in fatal crashes may reflect market share rather than an inherent safety advantage. Without data on the prevalence of different vehicle makes in the driving population, it's challenging to determine whether the observed associations are due to vehicle characteristics or exposure differences.

Conclusion

This study set out to determine whether Ford vehicles are disproportionately involved in fatal vehicle crashes compared to other brands. Based on a dataset of over 89,000 police-reported crashes in New York City, we found no evidence to support the hypothesis that Ford vehicles are more frequently involved in fatal crashes. In fact, only three crashes involved Ford vehicles, and none of those resulted in a fatality yielding a 0.00% fatal crash rate, compared to 0.06% for non-Ford vehicles.

To further examine this relationship, we conducted a logistic regression analysis using crash fatality as the dependent variable and controlling for several independent factors such as vehicle year, number of occupants, and damage severity. The model found no statistically significant association between vehicle make and the likelihood of a fatal crash, suggesting that Ford vehicles were not overrepresented in fatal outcomes.

However, several important limitations prevent us from making strong or causal claims. Most critically, the very small number of Ford-involved crashes limits statistical power, meaning that even one fatal crash could have significantly altered the results. Additionally, the observational nature of the dataset means we can only identify associations, not causes. Unmeasured confounding variables such as driver behavior, safety features, and environmental factors further limit the conclusiveness of our findings.

Therefore, while our results indicate no observed association between Ford vehicles and fatal crashes, this should not be interpreted as evidence that Ford vehicles are inherently safer or more dangerous. Future research should use a larger and more diverse dataset, span a longer timeframe, and incorporate methods that can better assess causal relationships, such as matched sampling or instrumental variable analysis.

Only with more robust and comprehensive data can we make confident, evidence-based statements about the causal impact of vehicle make on crash outcomes.

REFERENCE AND SOURCE:

Motor_Vehicle_Collisions_-_Crashes.csv

Bing.com