NYC Crash Data Analysis

Exploratory analysis of injury risk by brand and time with a logistic regression model

Ian Yoon
10/19/2025

Dataset: 89,102 crash records
Mean injury rate: 18.6% (any reported injury per crash)
Mean fatality rate: 0.056%. (any reported fatality per crash)

Motivating Questions:
When do crashes (and injuries) concentrate by hour and weekday?
Adjusting for time, are some vehicle brands associated with higher/lower odds of injury relative to Ford?(Logistic regression model)

Methods used:
Average Fatality/Injury rate calculation - outcomes/descriptives/insights/

Unadjusted table of injury rates/crash count by automakers - outcomes/descriptives/brandinjurytable/

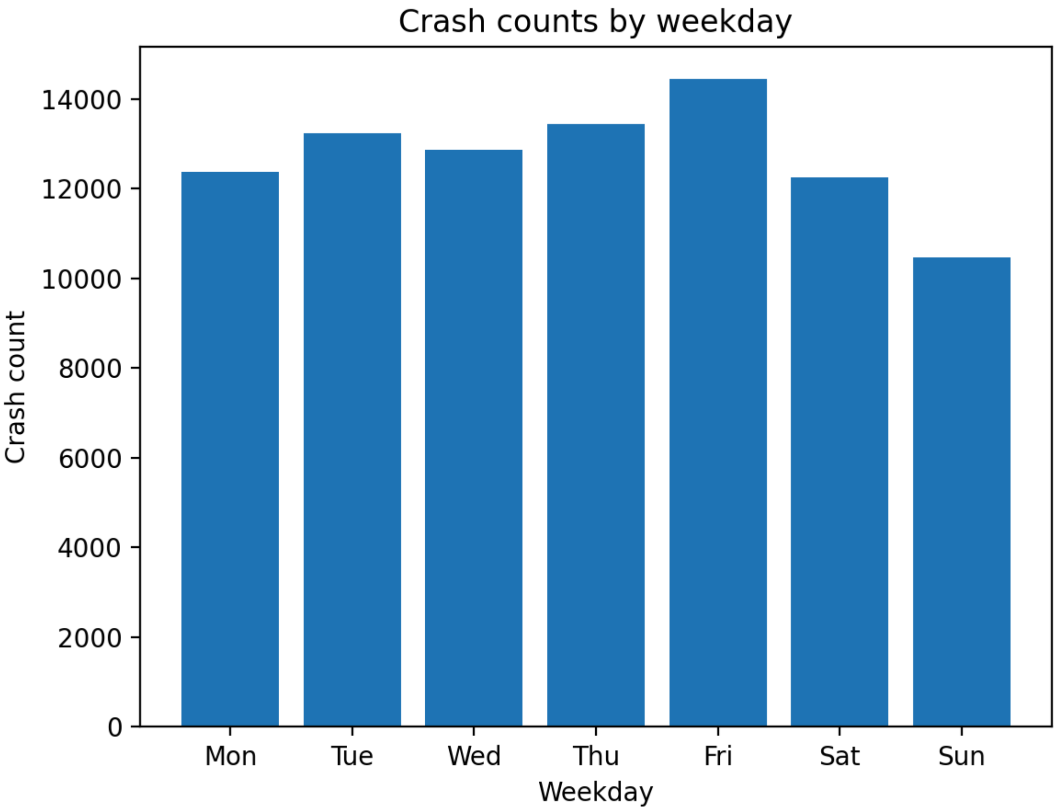
Crosstab of injury rates by VehicleType -> outcomes/descriptives/injuryratescrosstab/

Bar charts of crash count per Hour/Weekday -> outcomes/figs/

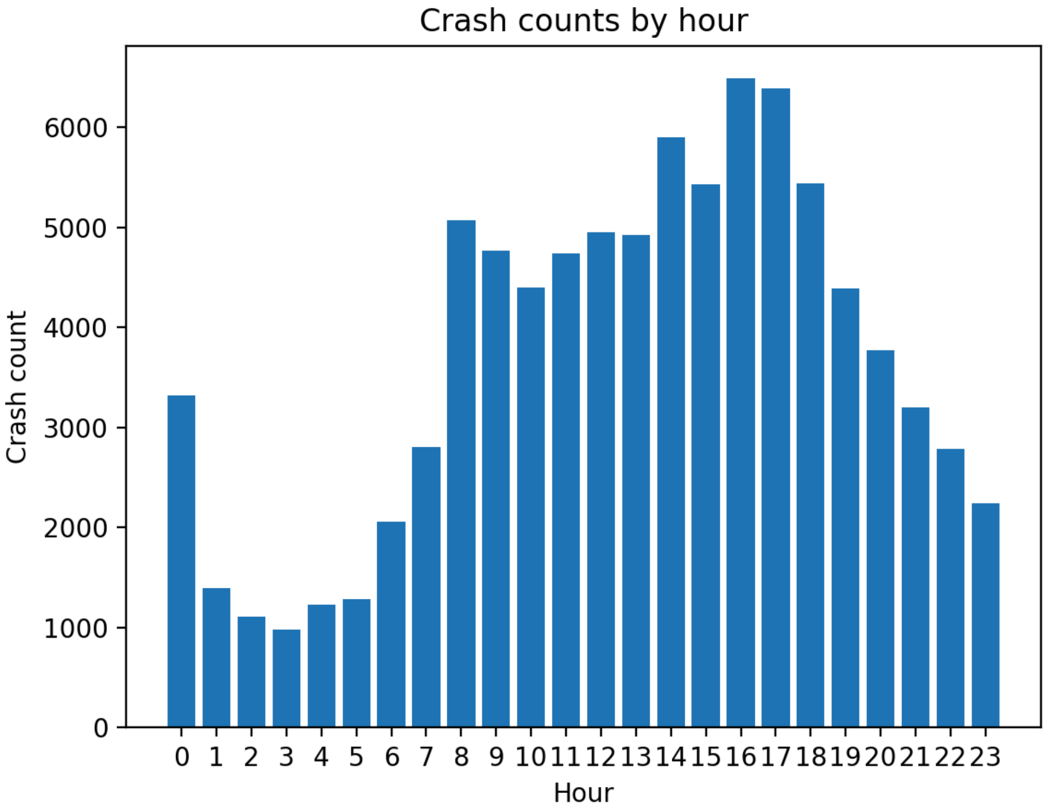
Logit Regression model: Dependent: any_injury, Predictors: vehicle_make, weekend, hour. Odds ratio table with confidence intervals. -> outcomes\logitreg

Descriptives results:
Day/Time of Day relation to crashes(pg2):
Weekday crash counts peak on Fridays and somewhat rise through the weekdays, and weekends are substantially lower than weekdays. This likely reflects less commuters and not less risk while driving.

Furthermore time of day spikes at 5-6pm at peak commuter times, with a abnormal bump at 12am, potentially due to tired commuters, low light visibility, and people returning home under the influence.



Crash counts by weekday



Crash counts by hour

Logit Regression results:

Below are all results that were statistically significant based on confidence intervals excluding 1.00 and p-values being under 0.05
Higher than Ford (CI = confidence interval, OR = Odds Ratio)

- LINCOLN: OR = 1.32, (CI[1.14, 1.54]), p=0.000
- KIA: OR = 1.28, (CI [1.09, 1.52]), p=0.003
- CHRY: OR = 1.24, (CI [1.07, 1.43]), p=0.003
- NISSAN: OR = 1.19, (CI [1.10, 1.29]), p=0.000
- HYUNDAI: OR = 1.14, (CI [1.02, 1.28]), p=0.017

Lower than Ford (INTL = Navistar Heavy duty truck brand, FRHT = Freightliner)

- INTL: OR = 0.39, (CI [0.29, 0.51]), p=0.000
- FRHT: OR = 0.43, (CI [0.33, 0.56]), p=0.000
- AUDI: OR = 0.66, (CI [0.54, 0.82]), p=0.000
- GMC: OR = 0.79, (CI [0.67, 0.94]), p=0.008
- VOLKSWAGEN: OR = 0.80, (CI [0.67, 0.94]), p=0.008
- MERCEDES: OR = 0.88, (CI [0.78, 0.99]), p=0.039

Time:
Weekend: OR = 1.051, (CI [1.007, 1.097]), p=0.021
Hour: OR = 1.004, (CI [1.000, 1.007]), p=0.028

Logit Insights:
Time values have relatively small impacts(+5% on weekends and +0.4% per hour) within the analysis while certain brands such as LINCOLN and KIA are associated with higher odds of injury given a crash. However given that control values were limited on the model, and that fit is somewhat low(pseudo-R² ≈ 0.004; LLR p < 0.001), safety/injury risk cannot be confirmed off of this study.

Conclusions:

Crashes concentrate in weekday PM peaks, and the logit model shows very small effects from time on injury risk (+5% on weekends; +0.4% per hour). Brand associated differences in injury odds were much higher given a crash. Truck makes (INTL/FRHT) show 0.4x injury odds compared to Ford, while certain brands(Liuncoln, KIA) were around 1.3x more likely to involve an injury then Ford. Descriptives also showed insights on vehicle classes, such as motorcycles having the highest injury rate per crash by a large margin(0.528). Since the model only controls for hour and weekend at the moment we lack outside context on other large factors such as vehicle class.