Times Series Analysis of Car Crashes in the United States

Owen Hammond-Lee

Motivation

- Automobile Accidents are the leading cause of death for people in the U.S. aged 1-54
- Understanding what factors lead to these accidents can help optimize policy, for example, when and where to post first responders to mitigate injury and loss of life, or how to predict traffic for travelers by estimating possible delays
- To establish when to put more responders out on major roadways, etc., it is necessary to know how the time of day, day of the week, and season influence accidents

The Data - US Accidents (2016 - 2023)

- Compiled from two academic papers
- Published on Kaggle
- Approximately 7.7 million accident records
- 46 variables including booleans for several weather and road conditions, as well as time, location, road, and severity data
- More accidents reported from later years due to data sourcing
- Prepared for time series analysis by binning accidents by time

Hypothesis

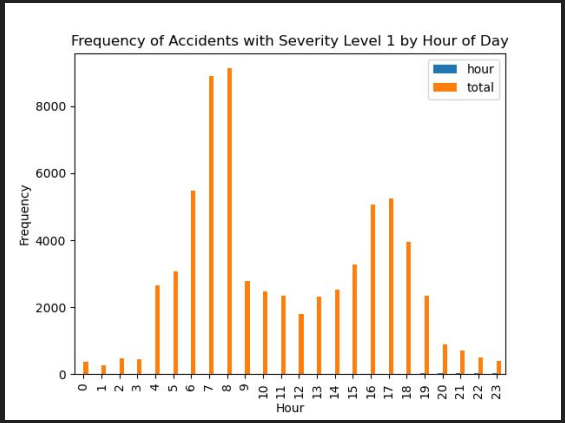
Null Hypothesis: There is not a significant difference in the distributions of number of accidents per hour by severity level

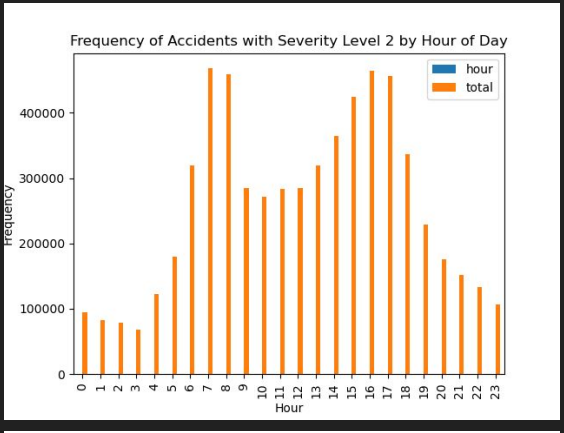
Alternative: There is a significant difference in the distributions of number of accidents per hour by severity level

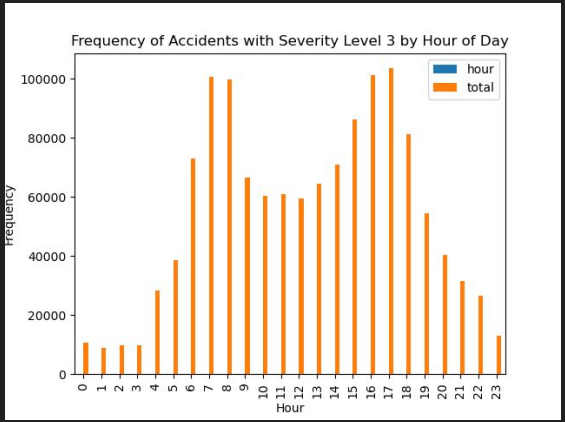
Statistical Test: Chi2 Test using Cramer’s V effect size, with effect size of .3 as the metric for major effect.

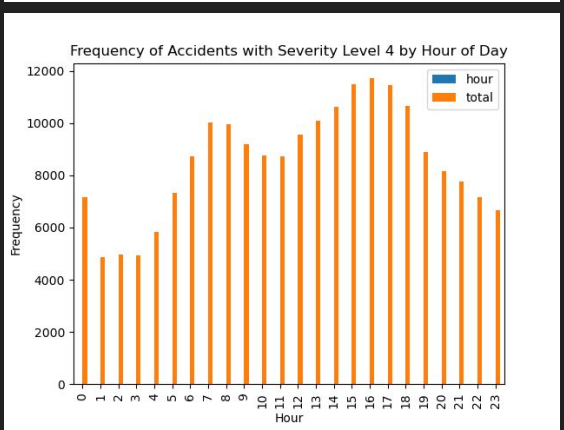
Variations: Data by hour of day, and by hour of day and day of week

Frequency of Accidents By Hour





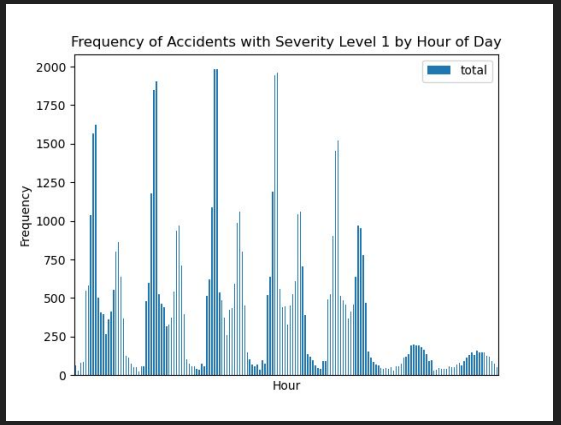


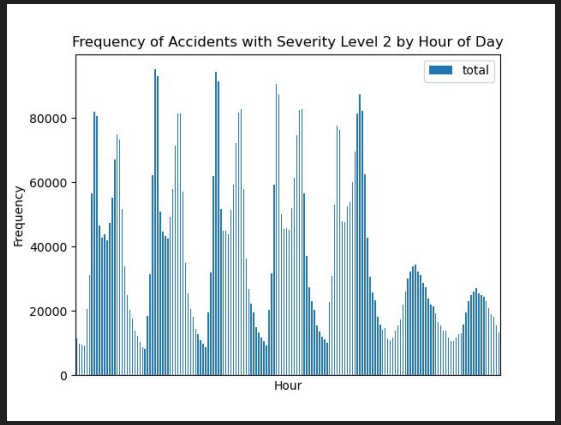


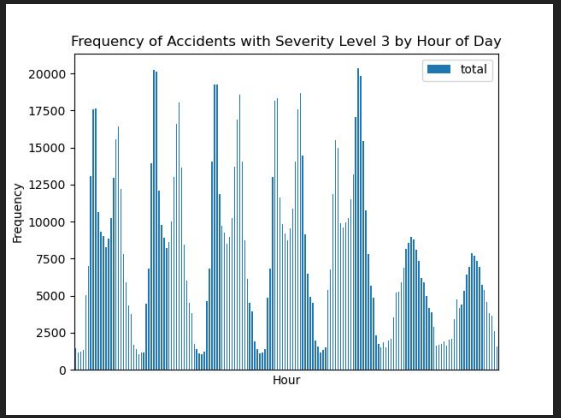
Chi2 Value: 62998.438368136274

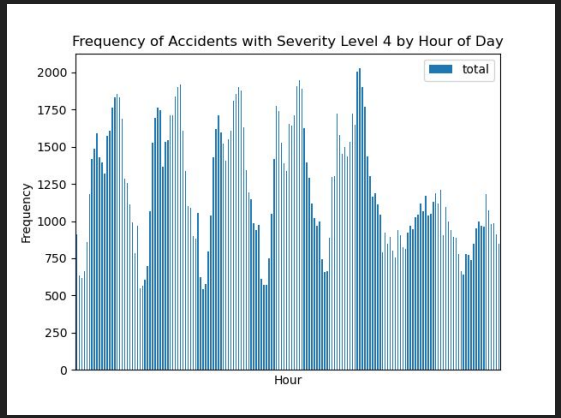
Effect Size: 0.15637988595891023

Frequency of Accidents By Hour and Day of Week





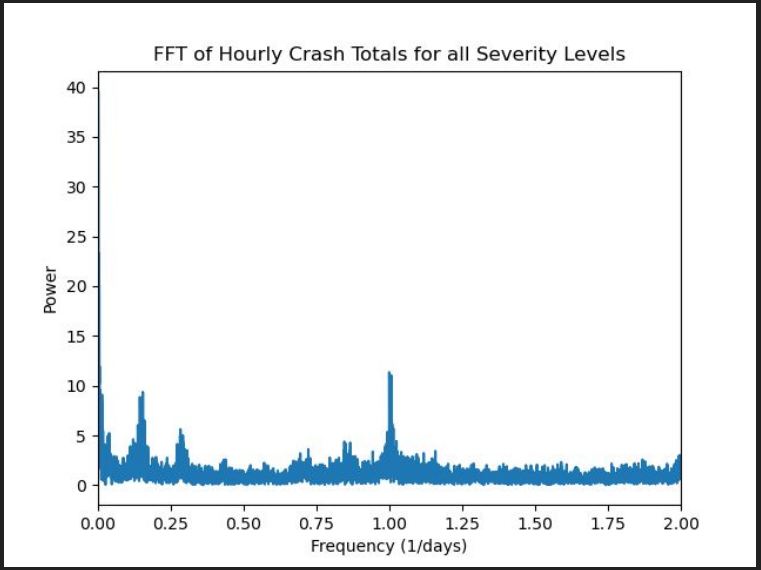




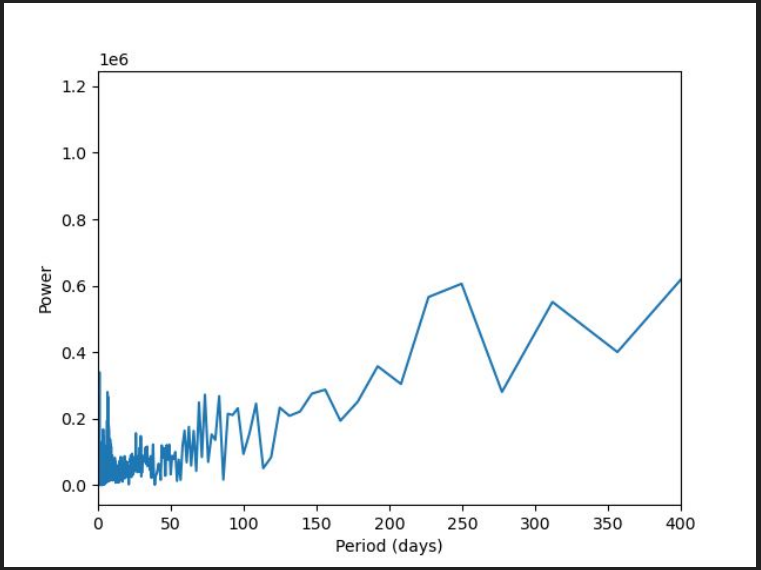
Chi2 Value: 88046.20855135124

Effect Size: 0.1848720808817592

Spectral Analysis



Major Peak: 1/day
Minor Peak: .15/day



Noise in low frequencies from long term trends in recorded results

Discussion

- Failure to reject null hypothesis, though we see small to medium effect size of severity in the differences in distribution.
- More consistent reporting or more advanced detrending is needed for longer term seasonality analysis
- Uncertainty in variability within severity levels

Further Analysis

-Traffic Dataset to cross compare with reporting and reduce long term noise

-Logistic Regression on multiple variables (weather, type of road, etc.)

-Seasonality Analysis

-Nonhomogenous Poisson Process Model

Sources

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Ng, Tin Lok James, and Andrew Zammit-Mangion. Non-Homogeneous Poisson Process Intensity Modeling and Estimation Using Measure Transport. arXiv:2007.00248, arXiv, 10 Feb. 2022. arXiv.org, http://arxiv.org/abs/2007.00248.