# Proposal

## Fatal Car Accidents ★

Shows data of national fatal car accidents in the year 2023.

Reference: Collected and distributed by the Fatality Analysis Reporting System. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars Data is public and available at: https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/ Documentation at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813706

Specifications: 24MB, but a total of 404MB is available. 37654 observations, 54 variables. Many interesting variables that include the date and time of crash (down to the minute) (int), as well as where the crash occurred (num). There are many other variables that show other details about the crash such as the number of people, vehicles, and pedestrians involved.

Interesting Questions: In which state did the most crashes occur? What month/week/day/hour do the most crashes occur? What factors correlate to a higher number of crashes? How does Weather contribute to crashes?

## The State of the State of the State

This data set consists of common words/phrases used in state of the state speeches by the governors.

The data was collected by analysts at 538, a poll analysis website. 538 created an article analyzing the data (https://fivethirtyeight.com/features/what-americas-governors-are-talking-about/) and also released it publicly via GitHub (https://github.com/fivethirtyeight/data/tree/master/state-of-the-state).

The data is 125 kilobytes and contains 2,224 observations and 8 variables. One of the more interesting variables is the phrase variable, which are characters representing different common words/phrases used in the governors' speeches. Another interesting variable is the total variable, a numeric variable that defines how many speeches that word/phrase was used, and the d_speech and r_speech variables that define how many times that word was used in Democratic governor speeches and Republican governor speeches respectively.

This data set provides us the info to answer questions such as "What issues do Democratic governors focus on? What issues do Republican governors focus on" or "How is the issue of climate changed mentioned in Democratic governors' speeches compared to Republican governors'?"

## College Basketball Dataset

Ranks the top mens college basketball teams from 2013-2025.

(https://barttorvik.com/trank.php?year=2025&sort=&conlimit=#) Provides the link that all data came from, and they use the data from the official Kenpom website. (https://kenpom.com/), also used ESPN - (https://www.espn.com/mens-college-basketball/stats)

Data Set is 370 KB, or 370000 bytes. There are 369 observations in the dataset that looks into the top teams from 2013-2025, but this is only the top 369, there are well over 3000 teams that could be explored which would give us well over 3000 observations. But in just the general dataset there is 369 observations, and 24 different variables. Variables for example are as simple as games won or field goal percentage, or as complex as Adjusted Defensive/Offensive Efficiency, or BARTHAG - which examines how likely it is for a certain team to fare against the average Division 1 basketball team.

This data can answer a large variety of questions when it comes to what statistics might have the most variability, or strength, by seeing which statistics have the most difference between teams. Can also make own rankings lists by using own desired statistics rather than all 24 provided stats. Answering questions such as “How important is defense compared to offense in terms of winning games” or “Which of these teams was ranked high but didn’t end up winning the national championship?”