**Overview**

The purpose of this analysis is to analyze, and draw conclusions from, the Iowa Department of Transportation’s data on car crashes in the state of Iowa. This data set catalogues all reported car crashes in Iowa starting in January 2009, and running until March 2024. We will be seeking to answer the following questions:

1. Has the annual rate of car crashes changed significantly over the years since 2009? Which years have been the most/least deadly?
2. Do more crashes occur during the week than on the weekend?
3. Did car crashes decrease significantly during the year 2020, when COVID-related lockdowns occurred?
4. In which months do the most crashes occur? The least? Which season is the most dangerous to drive in? Are there changes in the distribution of the severity of crashes during the year, in general?

The main dataset for this project comes from the [Iowa Department of Transportation](https://data.iowa.gov/Crashes/Crashes-in-Iowa-by-Severity-and-Day/psra-mit2/about_data), with supplemental population data being brought in from [MacroTrends](https://www.macrotrends.net/global-metrics/states/iowa/population).

Some things to note:

1. In our data analysis, we will often refer to visuals created from an interactive report made using Power BI, which can be found in the GitHub repository for this project as “Interactive Alcohol Report.pbix”.
2. We will also be performing analysis on smaller tables that were queried from the full dataset using MySQL; the SQL scripts used, and the corresponding output tables, can also be found in the GitHub repository.
3. **Car Crashes per Capita Over Time**

We’ll start by analyzing the annual per-capita rate of car crashes in the state of Iowa, looking to see if it has changed significantly over time since 2009. We’ll exclude the data from 2024, since we only have data up to March, and since crash rates vary significantly over the year, it would be difficult to extrapolate an annual rate for 2024. Let’s examine the summary output for this regression given by R:

Call:

lm(formula = Crash\_Rate ~ Year, data = rate\_by\_year)

Residuals:

Min 1Q Median 3Q Max

-0.0019638 -0.0007052 0.0001569 0.0008137 0.0015155

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.881e-02 1.326e-01 0.443 0.665

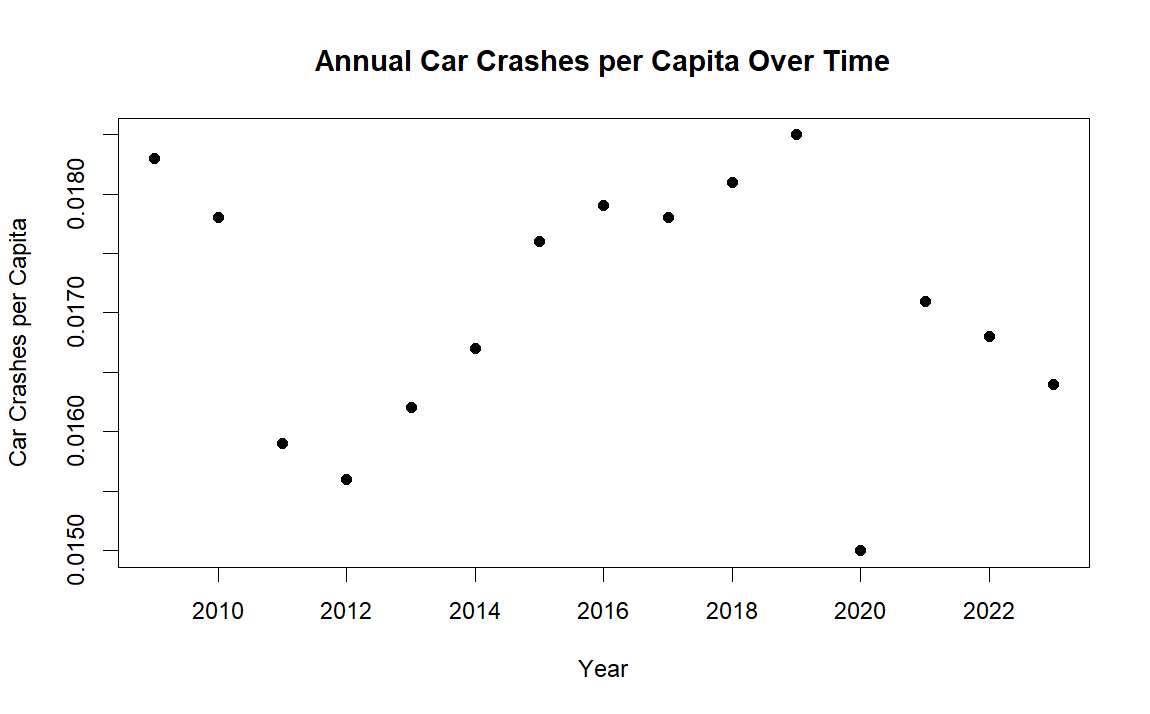
Year -2.071e-05 6.579e-05 -0.315 0.758

Residual standard error: 0.001101 on 13 degrees of freedom

Multiple R-squared: 0.007567, Adjusted R-squared: -0.06877

F-statistic: 0.09912 on 1 and 13 DF, p-value: 0.7579

R gives us a very large p-value of 0.7579 and an R-squared value of only 0.007567; it appears that there is no significant correlation between the year and the per-capita rate of car crashes. To be sure, we’ll examine a scatterplot of the data:



While we can confirm that the annual rate of car crashes did vary significantly during the time period examined, there is no visually discernible pattern within this scatterplot. We can and should take notice that 2020 had the lowest crash rate, being a clear outlier compared to the years around it. We’ll examine this further in section 3.

Using line charts pulled from the interactive report, we can see some interesting trends:

Crash rate over the years:

A graph with a line going up

Description automatically generated

From this chart, we glean that the most dangerous years in terms of car accidents were around 2017-2019, while the least dangerous were 2011-2012 as well as 2020. Now we’ll look at a chart filtered to contain only fatal car accidents:

FATAL crash rate over the years:

A graph with a line

Description automatically generated

From this chart, we see that the *deadliest* years for drivers were 2010 and 2016, while the years 2013-2015 were the least deadly; the years with the most and least crashes overall, are all quite average when the data is narrowed down to only fatal crashes.

1. **Car Crashes During the Week vs. Weekend**

Do car crashes happen more frequently during the week than on the weekend? A quick look at the “By Weekday” page of the interactive report makes this seem likely, since only 24.32% of all car crashes occurred on weekends; if car crashes happened at the same rate on weekends as on weekdays, then we would expect about the total crashes for weekends to make up about 2/7 of the overall total, or about 28.57%.

Let’s do some hypothesis testing using R to see if there really is a significant difference, using the null hypothesis H0: the proportion of crashes occurring on weekends is equal to 2/7, and a one-sided alternative hypothesis. We’ll also check the data for each severity category to see if they all have similar differences. Since we are working with very large samples, we will use simple one-proportion *z*-tests; the data will easily satisfy the independence and large sample size assumptions for such tests.

First, the data for all crashes:

A screenshot of a graph

Description automatically generated

The hypothesis test gives us a p-value less than 2.2 \* 10-16, telling us that this difference is highly statistically significant; car crashes happen significantly more often during the week than on weekends.

Next, the data for property damage crashes:

A screenshot of a graph

Description automatically generated

Like with the full dataset, it appears that property damage crashes occur more during the week than on weekends. This hypothesis test also gives us a p-value less than 2.2 \* 10-16, telling us that this difference is also statistically significant.

Next, the data for crashes that caused minor injuries:

A screenshot of a graph

Description automatically generated

The proportion for this set is only slightly less than our expected proportion of 28.57%; however, due to the very large sample size, the hypothesis test also gives us a p-value less than 2.2 \* 10-16, so we can conclude that this difference is highly significant.

Next, the data for crashes that caused major injuries:

A screenshot of a graph

Description automatically generated

The proportion for this set is quite a bit *greater* than 28.57%. The test of our null hypothesis gives us, again, a p-value that is less than 2.2 \* 10-16, telling us that this difference is highly significant.

Next, the data for fatal crashes:

A screenshot of a graph

Description automatically generated

The proportion for this set is, again, greater than 28.57%. The hypothesis test for this set gives us p = 1.784 \* 10-9, which is still extremely small, telling us that this difference is statistically significant as well.

Finally, we’ll check the data for crashes that had unknown or unverifiable outcomes:

A screenshot of a graph

Description automatically generated

It seems that crashes of this variety occurred less often on weekends than on weekdays. The hypothesis test for this set gives us a p-value that is less than 2.2 \* 10-16, telling us that this difference is also statistically significant.

1. **Car Crashes during 2020**

We noticed in section 1 that 2020 seemed to have a far lower rate of car crashes than the years around it. This is not unexpected, since much of the US population was under COVID-related lockdowns for a large chunk of 2020, which would likely lead to less driving in general. We’ll focus on the apparent drop in the crash rate between 2019 and 2020, which can be seen in the following table from the interactive report:

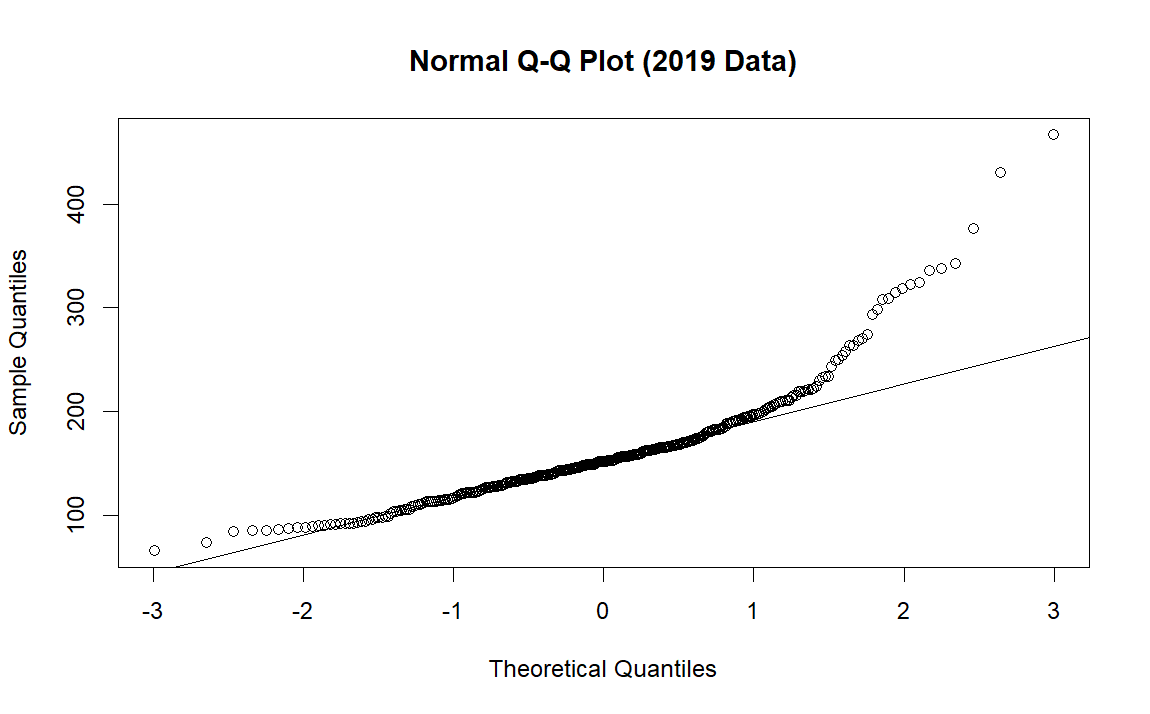
A screenshot of a graph

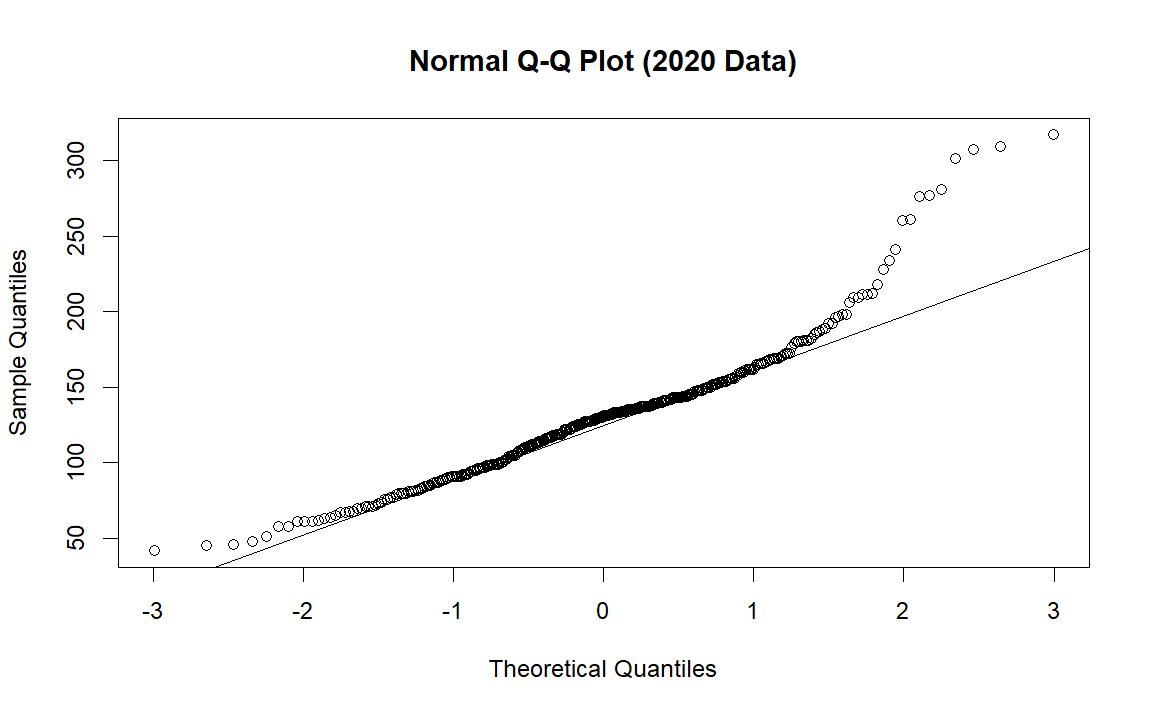
Description automatically generated

Let’s use a hypothesis test to check if the drop in crashes between those years was statistically significant. For a more precisely comparable and testable statistic, we’ll use the average number of crashes per day, instead of per capita. Our null hypothesis is that there was no statistically significant difference in the average crashes per day between 2019 and 2020, while our null hypothesis will be that the 2020 average is significantly lower than 2019.

Before we do the test, let’s check the basic assumptions for a two-sample *t*-test. The data clearly meet the requirements for independence and sampling, but we’ll need to use R to check to make sure the data are normally distributed and have equal variances.

Plots to test normality:





There is a bit of curvature in the plot for both datasets, but most likely not enough to greatly throw off the analysis. We’ll consider this assumption to be met.

To check if the two datasets have equal variance, we’ll perform an *f*-test using R. The test gives us a p-value of 5.338 \* 10-16, telling us that the two sets do not, in fact, have remotely equal variances. This is unfortunate, but the test can still be completed using Welch’s *t*-test in R. We obtain the following summary output:

Welch Two Sample t-test

data: Total\_Crashes by Year

t = 8.3003, df = 700.5, p-value = 5.338e-16

alternative hypothesis: true difference in means between group 2019 and group 2020 is not equal to 0

95 percent confidence interval:

22.59601 36.59773

sample estimates:

mean in group 2019 mean in group 2020

160.4548 130.8579

R gives us a p-value of 5.338 \* 10-16, allowing us to safely reject our null hypothesis and conclude that 2020 did see a significant drop in the annual rate of car crashes.

1. **Car Crashes by Month**

Let’s look at the data for total crashes plotted by the month, to see how the rate of car crashes tends to trend throughout a year:

A graph with a line

Description automatically generated

It appears that car crashes, in general, happen most frequently during the winter months of November to February, with a smaller peak in midsummer and a valley in March and April. Since Iowa has relatively harsh winters, it seems natural that car crashes would happen most frequently during those months where the weather can produce hazardous driving conditions.

Let’s now examine the same measure, broken down by the severity of the crash:

Crashes causing only property damage:

A graph with a line

Description automatically generated

Crashes of unknown severity:

A graph on a black background

Description automatically generated

Crashes causing minor injuries:

A graph on a black background

Description automatically generated

Crashes causing major injuries:

A graph on a black background with Wind Wand in the background

Description automatically generated

FATAL crashes during the year:

A graph showing a line

Description automatically generated with medium confidence

Interestingly, it appears that crashes which caused only property damage or had unknown severity tended to follow the trend we saw in the full dataset; however, crashes with injuries or fatalities actually follow an opposite trend; the winter months actually seem to have the least injuries and fatalities, while such crashes occur in far greater numbers during the summer months.

**Conclusion:**

To answer each of our initial questions:

1. There is no visible or statistically verifiable correlation between the year and the overall rate of car crashes. The most dangerous years to drive in were 2017-2019, but the *deadliest* was 2016.
2. Car crashes overall tend to happen more frequently during the week than on weekends, but this trend mainly occurs among minor crashes involving only property damage or minor injuries; crashes with major injuries and fatalities tend to happen more often on weekends than during the week.
3. There was a highly statistically significant drop in the rate of car crashes during the year 2020, as compared to the year 2019.
4. Car crashes overall happen more frequently during the winter months as compared to the rest of the year, but this trend mainly occurs among crashes involving only property damage or unknown severity; crashes with injuries or fatalities see the opposite trend, occurring far more often during the summer months.

Further studies could bring in multiple disciplines to examine the “why” questions raised by this analysis: why do car crashes involving fatalities and injuries tend to follow wildly different trends than crashes limited to property damage? What made 2016 such a deadly year to drive in Iowa? These, and many similar questions, should be answered by proper experimental studies.

**Works Cited:**

1. “Crashes in Iowa by Severity and Day.” *Iowa Open Data*, Iowa Department of Transportation, 26 Mar. 2024, < <https://data.iowa.gov/Crashes/Crashes-in-Iowa-by-Severity-and-Day/psra-mit2/about_data> >. Retrieved 09 Apr. 2024.
2. “Iowa Population 1900-2023.” *Macrotrends*, Macrotrends LLC, < <https://www.macrotrends.net/global-metrics/states/iowa/population> >. Retrieved 12 Apr. 2024.