David Lambert
DATA 115 Final
Fall 2020
Deford

• Describe the dataset and why you selected it for this project.
The data set that I used for this project was California Traffic Collision data from SWITRS. This is detailed information about every traffic collision reported to the California Highway Patrol from 2001 to 2020. With over 9 million entries, this dataset was initially quite unwieldy, and so I ended up working with only 250,000 randomly sampled entries for my project. I trusted this dataset for my project because it was expansive, detailed, and had many different qualitative and quantitative descriptors for each entry in its database.

• Describe any processing problems you identified with the data and how you overcame those issues.
A couple of problems that I had run into while processing the state it was of course that it contains 9 million entries which is completely unwieldy for my computer to process effectively in Excel so that needed to initially be sampled down to 250,000 entries by random selection this was done on my behalf by Dr. Deford, and the dataset contained descriptors that I didn't find applicable to my big question, these are descriptors such as the types of vehicles involved in these collisions, the road conditions at the time, the geographic coordinates of every collision and the type of road that the vehicle collided on. so to sort this information out I got rid of all of these extra descriptors and cleaned the data down to only what I needed to answer for my sub-questions, yet one issue, in particular, arose again in that the formatting for the times in which of these collisions happened was not appropriate for migrating into a Jupyter notebook and graphing so they had to do was I had to clean that data to create a column of just the hour that every collision happened.

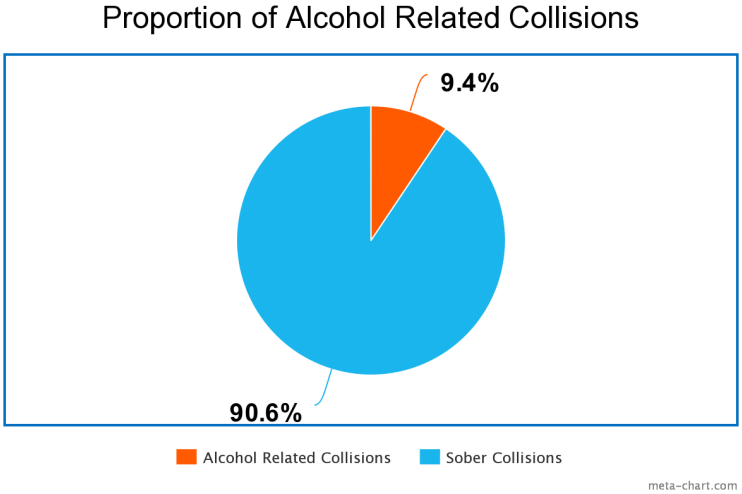
• Describe your 'Big Question' and why the data is a good choice to answer it.
The big question that I attempted to answer with this project was "how do alcohol-related automobile collisions behave as opposed to sober automobile collisions?". this data set is perfect is not the only choice to analyze this sort of information with as it contains many qualitative and quantitative descriptors including a Boolean as to whether or not alcohol was related in any given crash this allowed me to sift and filter through data depending on whether or not there was alcohol involved in any given collision allowing me to create two separate data sets for both alcohol-related and sober collision data.

• Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.
I decided to go about analyzing my big question by breaking it down into smaller sub-questions these sub-questions included "What proportion of crashes are alcohol-related?"," do alcohol-related crashes occur more often at a certain time of day?", and " are alcohol-related crashes any deadlier than sober crashes?". Fortunately for me, I was able to do the majority of my analysis in Excel, it made it very easy to process all of the data in one program however I certainly did utilize python to create visualizations for the final presentation.

so in order,

David Lambert
DATA 115 Final
Fall 2020
Deford

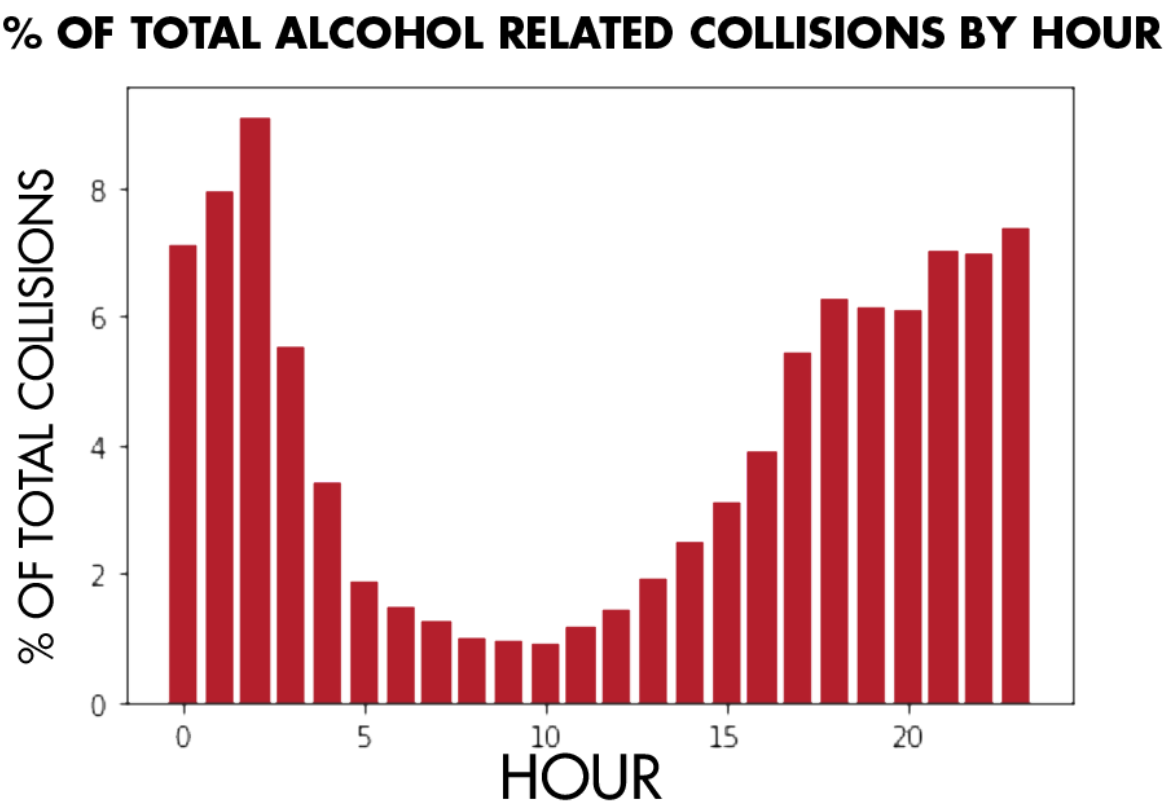
a) *What proportion of crashes are alcohol-related?*
I initially tackled this same question by creating a simple proportion where I took the amount of alcohol-related crashes and divided them by the total number of crashes and this brought me to around 9.4% of all crashes are alcohol-related however as I soon learned this does not mean that at any given time that you get on the road, 1 and 11 people the crash is likely drunk, no, I gave the analysis more thought later and broke it down by hour as you will see to come.



b) *Do alcohol related crashes occur more often at a certain time of day?*

To answer this question I initially made a graph that broke down the number of alcohol-related collisions by the hour in the day so I ended up with a bar graph that showed What proportion of alcohol-related crashes happen at any given hour, this initial analysis however did not incorporate the larger statistic of What proportion of total collisions per hour were alcohol-related that being the other facet of this same question which was to be answered later during my analysis. yet I did come up with this graph that shows alcohol-related crashes to be closely linked to day and night time behavior.

David Lambert
DATA 115 Final
Fall 2020
Deford



c) *are alcohol-related crashes any deadlier than sober crashes?*

During my introductory exploratory analysis of this question, I leaned on the commonly accepted fact that you are more likely to be in a collision while drunk this led me to explore whether or not an alcohol-related collision has a higher lethality rate than a collision that occurred while the driver was sober. So to explore this sub-question I created a new column within my data set wherein I took the number of people involved in the Collision known as the party size in the initial database and divided the number of

David Lambert
DATA 115 Final
Fall 2020
Deford

victims killed over the party size to get a lethality percentage. not much further analysis was needed on this topic as I utilize these lethality percentages for the rest of my presentation to create visualizations that accurately show How likely one is to suffer from Fatal wounds resulting from an alcohol-related accident as opposed to a collision that occurred while sober.

• Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.

I selected this form of methodology to answer my big question in a way that dissected all facets of the question itself simply asking how alcohol-related collisions behave as opposed to sober collisions has a lot tied into it and implicated. so by breaking this larger question down into more manageable sub-questions for me to analyze we can kind of manifest a Gestalt of the situation surrounding alcohol-related collisions. I thought that the three most pertinent questions to answer here would be “how often alcohol-related collisions happen.”, “when alcohol-related collisions occur.”, and “ are alcohol-related collisions deadlier than sober collisions.”. I believe that altogether this would form a strong picture of how drunk driving and alcohol-related collisions act as opposed to accidents where alcohol was not at play by comparing and contrasting the two types of driving behavior. I think that my analysis chose a detailed image of drunk driving in practicality.

However, this is not to say that my analysis of my big question does not have some drawbacks some of the main cons of utilizing this format to answer a larger question would be in that it can easily appear as though I am dancing around the big question other than answering facets of it and so an alternative analysis approach was considered in that I would Delve deeply into one aspect about the question and highlight many key differences about one statistic as opposed to covering multiple components of what makes alcohol-related collisions different than sober ones.

• Describe your conclusions based on your analysis and support them with analytics on your dataset.

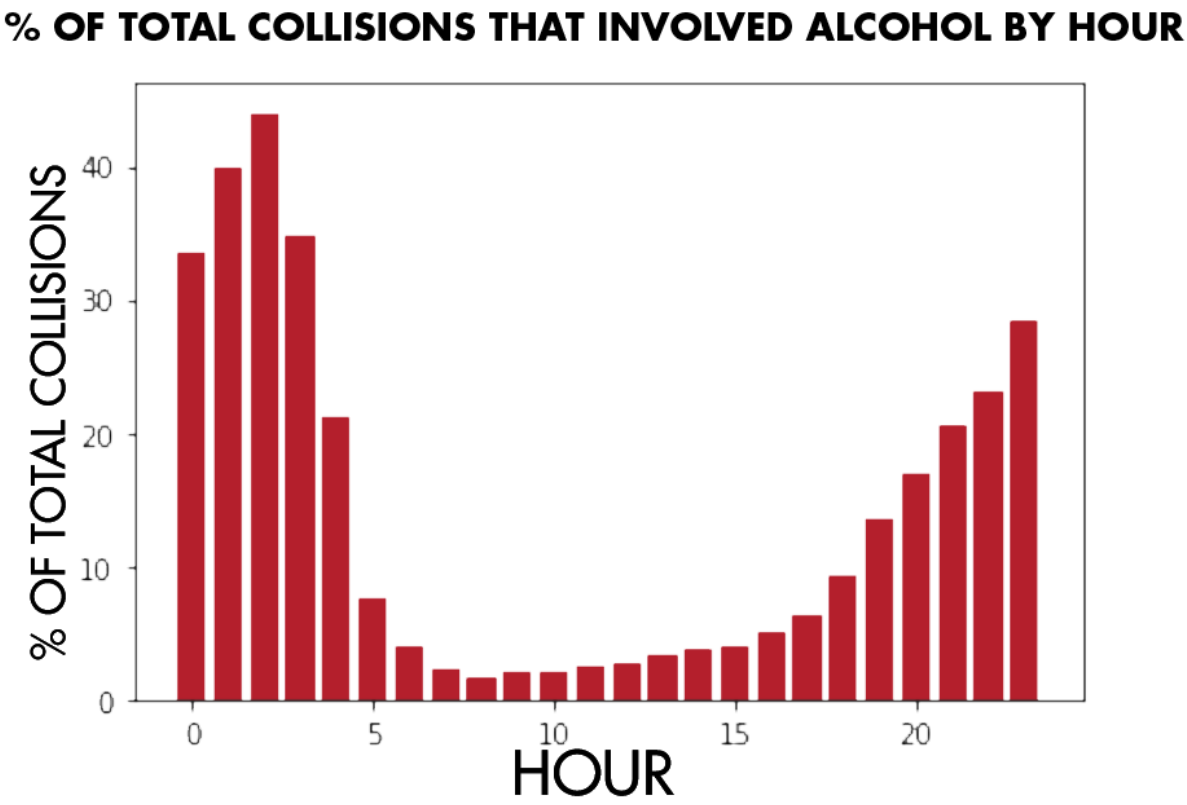
I will be breaking down my conclusions based on my analysis into my three separate sub-questions and their conclusions are as follows.

1. *What proportion of crashes are alcohol-related?*

On average 9.4% of all collisions will involve alcohol this average was derived from the total number of alcohol-related collisions sampled divided by the total number of collisions sampled exact numbers being 23538/250000 as referenced in the alcohol collisions spreadsheet.

David Lambert
DATA 115 Final
Fall 2020
Deford

The proportion of all collisions which are alcohol-related changes given the hour of the day this was graphed by the alcohol collisions by time percentage spreadsheet. The result of that graph was imported into python and ended up looking something like this.



As you can see the vast majority of collisions that involve alcohol occurred from the hours of 7 p.m. to 4 a.m. this is likely due to social drinking behaviors and poor visibility conditions in the nighttime. as you can see at 3 a.m. 44% of all total collisions that occurred within that hour involved alcohol making this potentially the most dangerous time for encountering drunk drivers. Full information Surrounding the formulation of this graph can be found here, where all alcohol collisions and all sober collisions by our have been totaled and turned into a percentage of alcohol-related collisions over the total number of collisions by the hour.

David Lambert
DATA 115 Final
Fall 2020
Deford

Time	AlcCollisio	SoberColli	Alc/Total
0	1664	3297	33.54
1	1858	2797	39.91
2	2129	2709	44
3	1289	2417	34.78
4	802	2989	21.15
5	439	5397	7.52
6	346	8160	4.06
7	301	13184	2.23
8	233	13186	1.73
9	221	10245	2.11
10	212	10059	2.06
11	279	10789	2.52
12	335	12024	2.71
13	447	12763	3.38
14	584	14825	3.78
15	724	17602	3.95
16	917	17171	5.06
17	1272	18503	6.43
18	1471	14331	9.3
19	1437	9091	13.64
20	1422	6927	17.03
21	1639	6362	20.48
22	1633	5442	23.08
23	1721	4314	28.51

a) *Do alcohol related collisions occur more often at a certain time of day?*

The conclusions drawn from the prior question lead directly into the answers for this sub-question. As we know alcohol-related collisions do occur more during the night time when put against sober collisions, however, 6 to answer how alcohol-related collisions are different from sober collisions and so I graft each of these collision types individually. By creating two spreadsheets alcohol collisions by time and sober collisions by time I was able to create two different visualizations of both of these types of data the information used can be found here with sober collisions listed first, followed by the alcohol collision data over time.

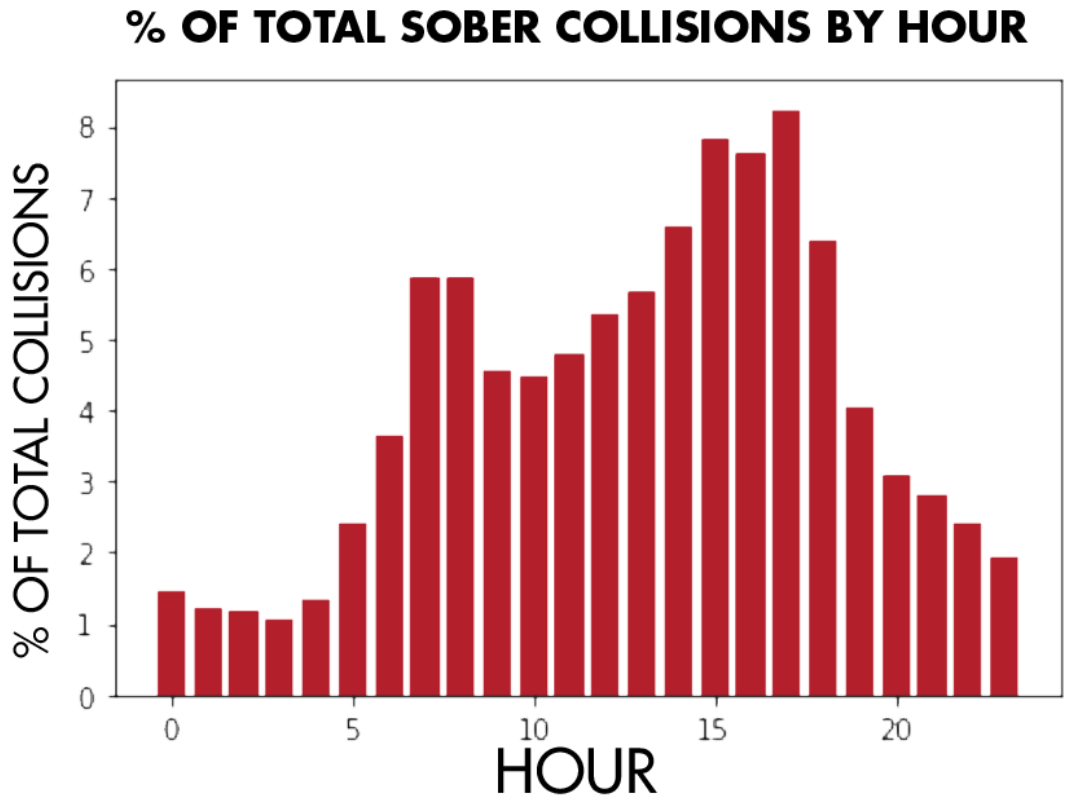
Time	Collisions	CollisionPe
0	3297	1.46
1	2797	1.24
2	2709	1.2
3	2417	1.07
4	2989	1.33
5	5397	2.4
6	8160	3.63
7	13184	5.87
8	13186	5.87
9	10245	4.56
10	10059	4.47
11	10789	4.8
12	12024	5.35
13	12763	5.68
14	14825	6.6
15	17602	7.83
16	17171	7.64
17	18503	8.23
18	14331	6.38
19	9091	4.04
20	6927	3.08
21	6362	2.83
22	5442	2.42
23	4314	1.92

David Lambert
DATA 115 Final
Fall 2020
Deford

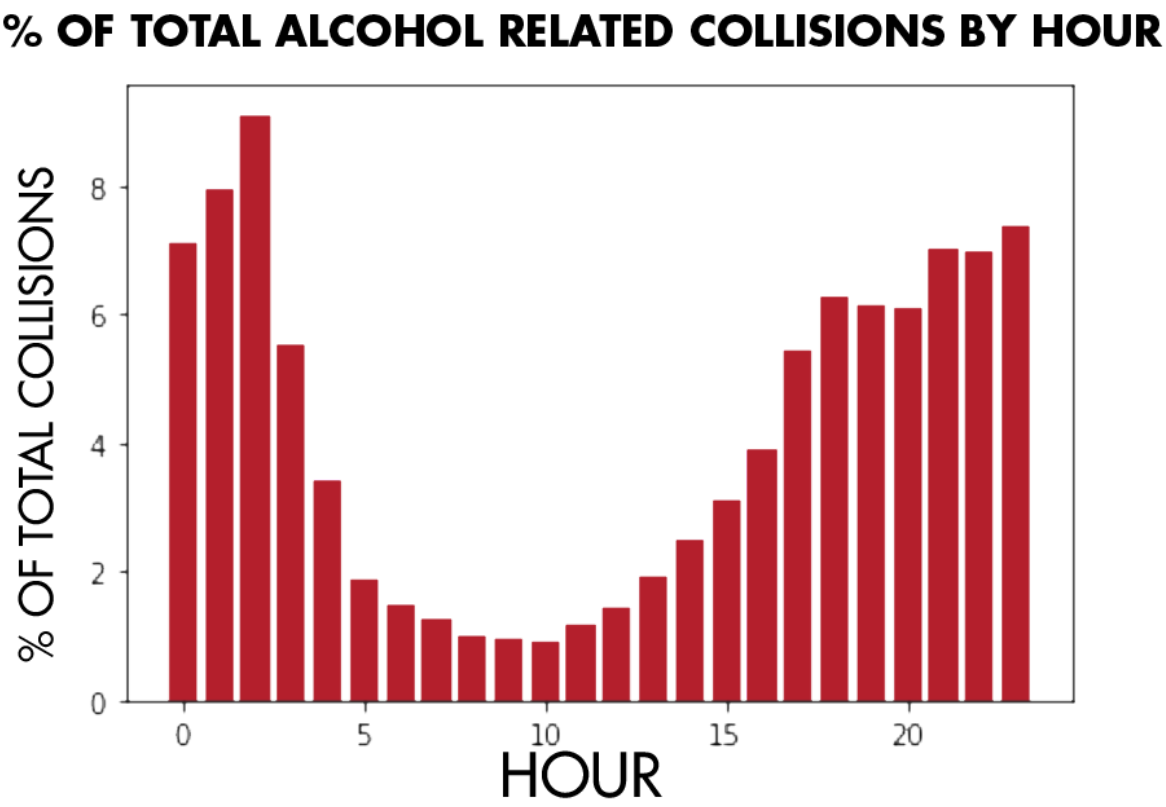
Time	Collisions	CollisionP
0	1664	7.11
1	1858	7.94
2	2129	9.1
3	1289	5.51
4	802	3.43
5	439	1.87
6	346	1.48
7	301	1.28
8	233	0.99
9	221	0.94
10	212	0.9
11	279	1.19
12	335	1.43
13	447	1.91
14	584	2.49
15	724	3.09
16	917	3.92
17	1272	5.44
18	1471	6.29
19	1437	6.14
20	1422	6.08
21	1639	7.01
22	1633	6.98
23	1721	7.36

These two spreadsheets were transformed into, separated values files and exported into python for graphing and further analysis by this method I was able to create the following visualisations.

David Lambert
DATA 115 Final
Fall 2020
Deford



David Lambert
DATA 115 Final
Fall 2020
Deford



As you can likely tell these two graphs while identical, both exploring the percentage of total collisions by the hour based on whether or not they involve alcohol you will notice that sober collisions occur in an almost inverse fashion as opposed to when alcohol-related collisions occur. this could have been inferred from the first bar graph that I showed where one can tell got more alcohol-related collisions occur during the night time in proportion to sober collisions, however, when we break these down by subgroup alcohol-related in sober collisions we can tell that the majority of sober collisions occur during the rush hour 5 p.m. when most people are on the road, as opposed to alcohol-related collisions which occur most at 3 a.m. when feasibly only drunks are on the road. So in conclusion it's not that more collisions happen at night it's that more drunks have accidents at night.

b) *Are alcohol-related collisions any deadlier than sober collisions?*

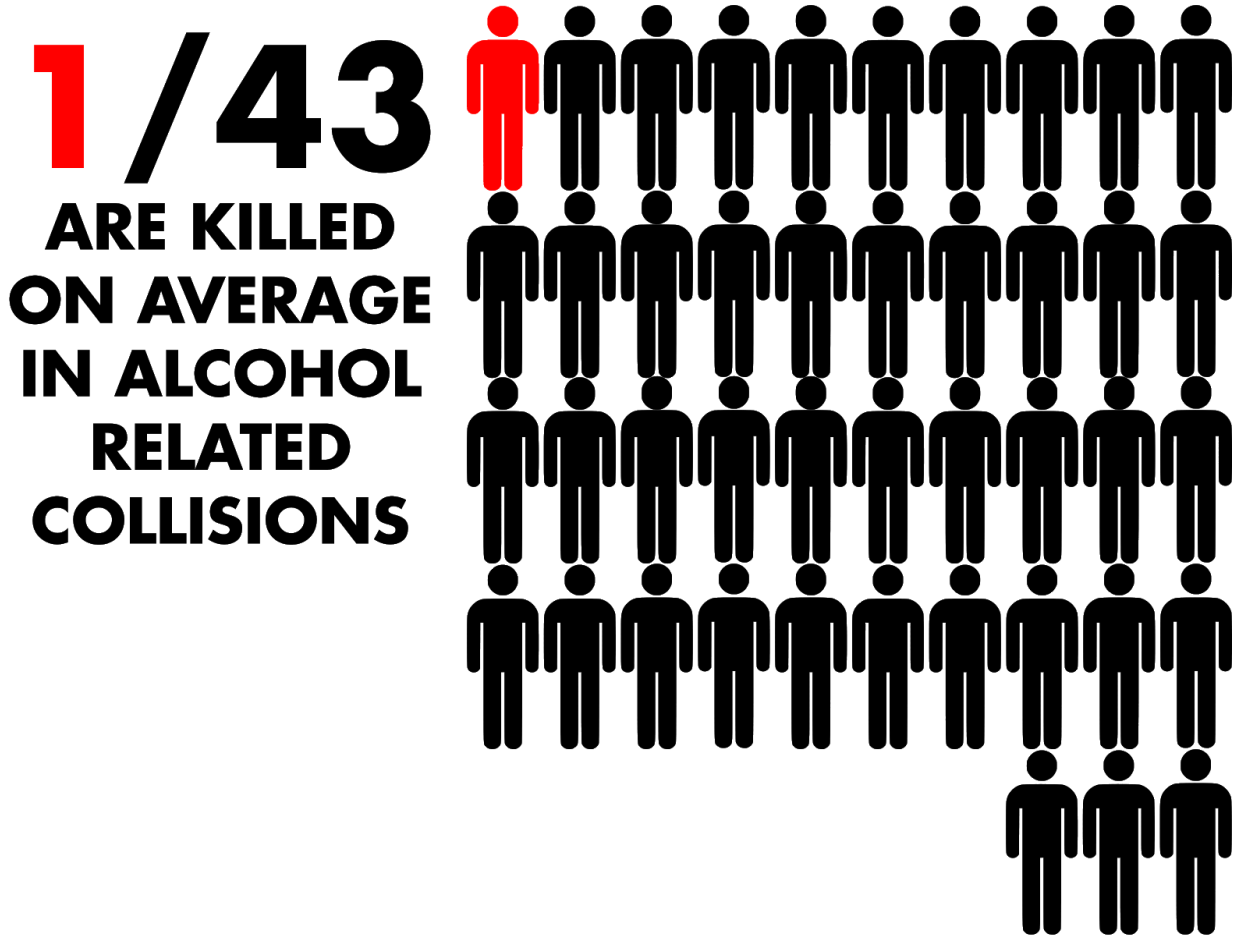
As I stated before, not much further analysis went into this sub-question after I got the initial result which told me, yes alcohol-related collisions are far deadlier than sober collisions this information was garnered by taking the party counts or the number of people involved with any given collision and dividing it by the number of victims killed in that accident and then sorting the

David Lambert
DATA 115 Final
Fall 2020
Deford

data between alcohol-related collisions in non-alcohol-related collisions to create a lethality percentage for both types. Further analysis could be done on this topic as to whether or not drivers involved in alcohol-related collisions have more passengers with them at the time of the accident however that is neither here nor there surrounding lethality exactly.

By taking the average of my lethality column calculated by the way I mentioned above, I was able to create an average lethality for both alcohol-related collisions and sober collisions by taking this lethality percentage and Dividing 100 by it I was able to create a statistic used for my visualizations wherein with an average lethality of 2.33% for alcohol-related collisions in an average lethality of .45% for sober collisions I calculated that one in 222 people involved with a sober collision will suffer fatal injuries, as opposed to 1 in 43 people involved in an alcohol-related collision suffering the same fatal injuries. in conclusion, yes alcohol-related collisions are far deadlier than non-alcohol-related collisions, by a factor of 5.17.

David Lambert
DATA 115 Final
Fall 2020
Deford



All in all the culmination of all the sub questions allow me to paint a picture of alcohol-related crashes being deadlier than sober crashes, occurring at nighttime as opposed to sober crashes occurring during the day, Alcohol-related crashes being linked more to time of day rather than number of cars on the road, and making up approximately 1 out of every 11 total collisions.

• Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

There's one statistic that I would have liked to describe being that of whether one is more likely to get in a collision while under the influence of alcohol as opposed to driving sober. However, Collision data is only reported by motorists who collide; we do not track statistics about drivers who do not get into accidents or get pulled over, and therefore knowing how many drunk drivers are on the road at any given time will ultimately always remain a mystery. that they would have been essential in performing this analysis as the total number of drunk drivers on the road at any given time would be the denominator to the total number of alcohol-related collisions.

David Lambert
DATA 115 Final
Fall 2020
Deford

SOURCE:
https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs
This dataset included all Collision data reported to the California Highway Patrol from 2001 to 2020 it was instrumental in the formulation of this final project and served as the main data set for my analysis, within this set was included many descriptors of every collision and both qualitative and quantitative data surrounding facets of each accident all as reported to the CHP.

GITHUB APPENDIX: (Includes all documentation & materials)

https://github.com/yugledorf/DATA-115-Group-Final