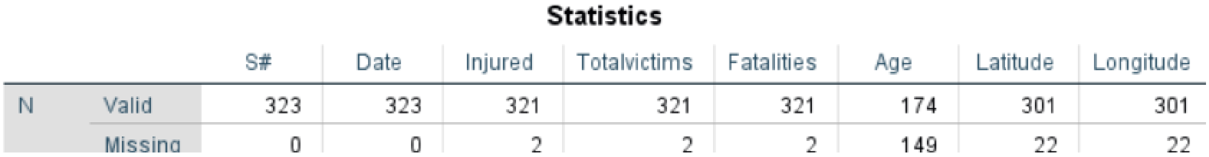
Examination of US Mass Shootings

This report is a deconstruction, examination, and analysis of the US Mass Shooting dataset. This dataset spans 50 years from 1996-2016, consisting of 398 mass shootings including thousands of fatalities in total. A mass shooting is defined in several different ways, but in these events, it seems to be an incident involving multiple victims from firearm violence. The dataset contains abundant information; the date, location, area (e.g. strip club, home, campsite), the target of the attack, the suspected cause, a summary of the event, the gender, race, and information on mental health issues of the attacker, as well as information on the victims; number of fatalities, number of injuries, total victims, as well as how many policemen were killed if any. This examination will first detail the steps taken when cleaning the dataset, then explore the cleaned data to find insights and answer some key questions regarding mass shootings in the US.

Data Cleaning (Q5)

1. Missing Values

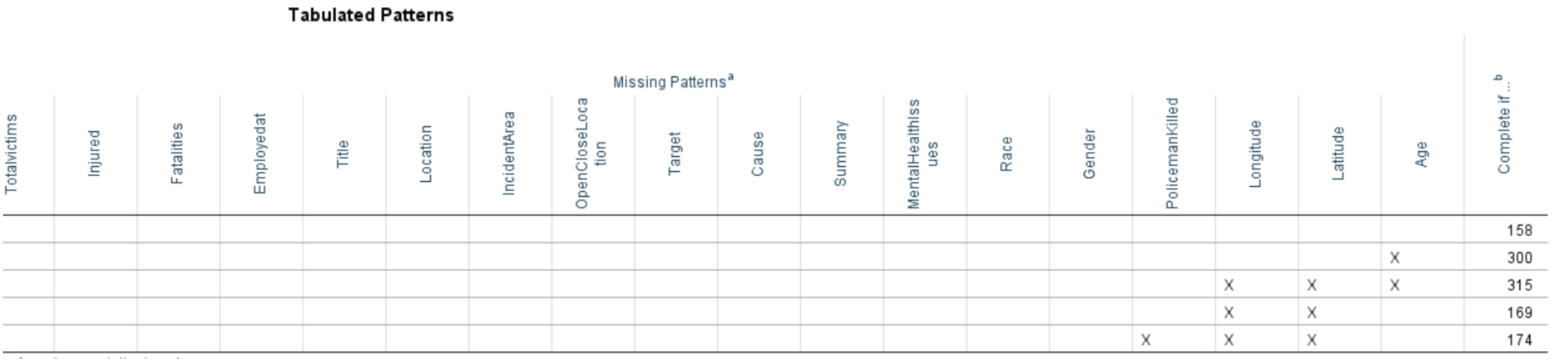
Before we can deal with the missing values of the dataset, we have to see which values are missing, how many, and whether they are systematically missing (MAR: Missing at random), or missing completely at random (MCAR). The best outcome would be for the missing values to be MCAR, as we would then be able to impute using simpler methods.



Statistics

		S#	Date	Injured	Totalvictims	Fatalities	Age	Latitude	Longitude
N	Valid	323	323	321	321	321	174	301	301
	Missing	0	0	2	2	2	149	22	22

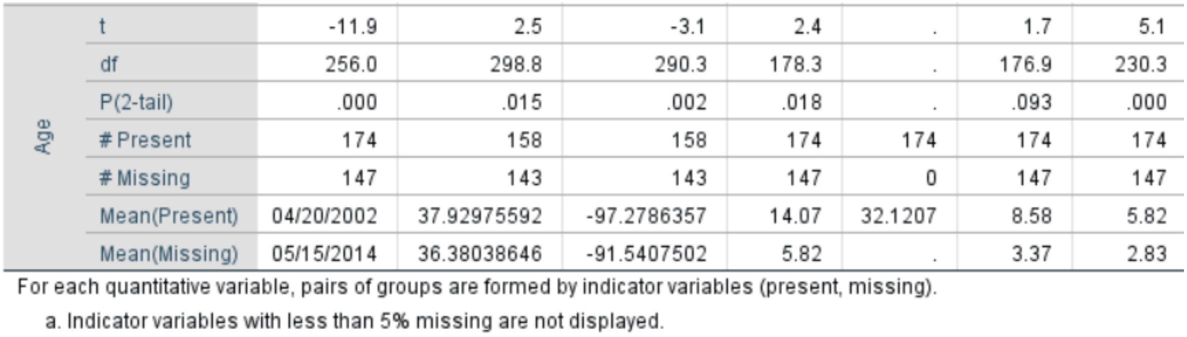
The table above is the output of the frequency of the data, showing the number of missing values for each numerical value. Clearly the variable 'Age' is an issue, as just under 50% of the cases in the dataset are missing the age value. Latitude/Longitude pose minor issues which we will explore later.



Tabulated Patterns

Missing Patterns[a]																		
Totalvictims	Injured	Fatalities	Employedat	Title	Location	IncidentArea	OpenCloseLocation	Target	Cause	Summary	MentalHealthIssues	Race	Gender	PolicemanKilled	Longitude	Latitude	Age	Complete if ...[b]
																		158
																	X	300
															X	X	X	315
															X	X		169
														X	X	X		174

The Tabulated Patterns table above shows how many valid cases (cases with no missing values) there would be if each variable was deleted. We can see that we currently have 158 valid cases, but if we deleted the age variable we would have 300 valid cases, a great increase.



Age	t	-11.9	2.5	-3.1	2.4	.	1.7	5.1
	df	256.0	298.8	290.3	178.3	.	176.9	230.3
	P(2-tail)	.000	.015	.002	.018	.	.093	.000
	# Present	174	158	158	174	174	174	174
	# Missing	147	143	143	147	0	147	147
	Mean(Present)	04/20/2002	37.92975592	-97.2786357	14.07	32.1207	8.58	5.82
	Mean(Missing)	05/15/2014	36.38038646	-91.5407502	5.82	.	3.37	2.83

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

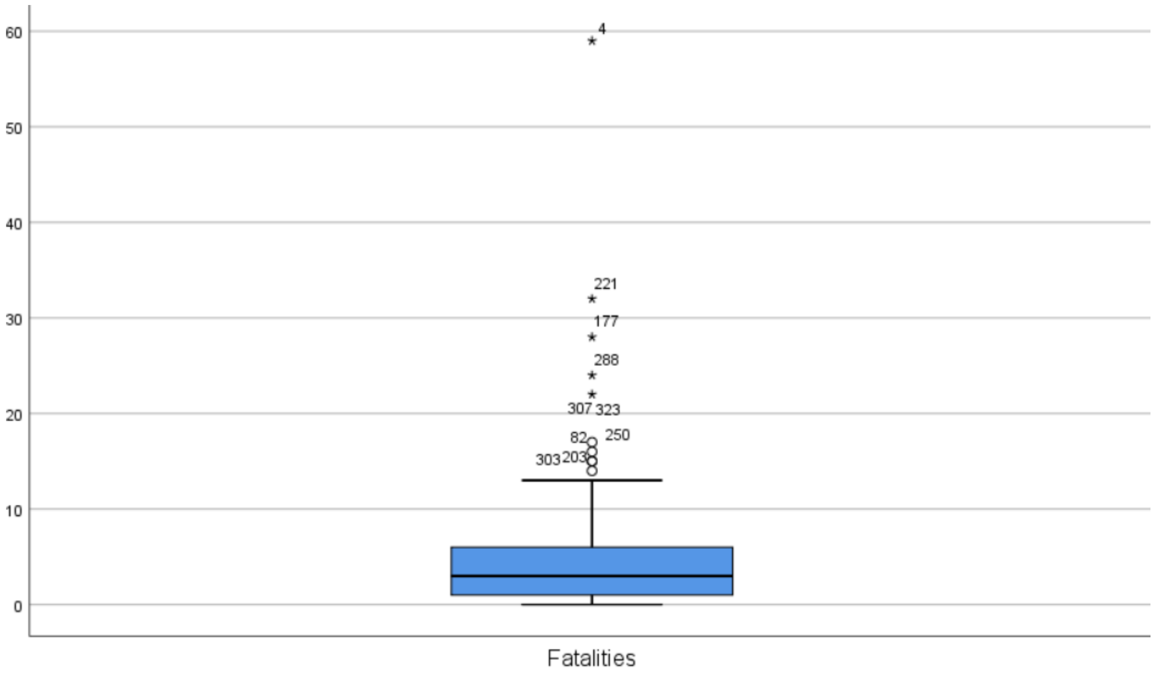
The table above is the output of the Independent-Samples T Test, which I used to test how systematically the variables are missing from the dataset. This particular section is for age against other quantitative variables, and the important row to look at is the P(2-tail). These values are the P values for the T-Test, which tells us how the other variables values compare when subgroups of age are not present and when they are not present, this is shown by the rows mean(present) and mean(missing). The P values for most of the variables are below 0.05 which means Age is problematic with the other variables, so I came to the decision to delete the variable Age. This is simpler than attempting to impute on almost half of the dataset for the variable, and increases the valid cases of the dataset to 300.

Additionally I used the Missing Patterns table to see if there were any cases in particular missing many values. The table is too big to be embedded in this document, but it showed that there were just two cases which were very problematic; 115 and 100, which were both missing variable values each.

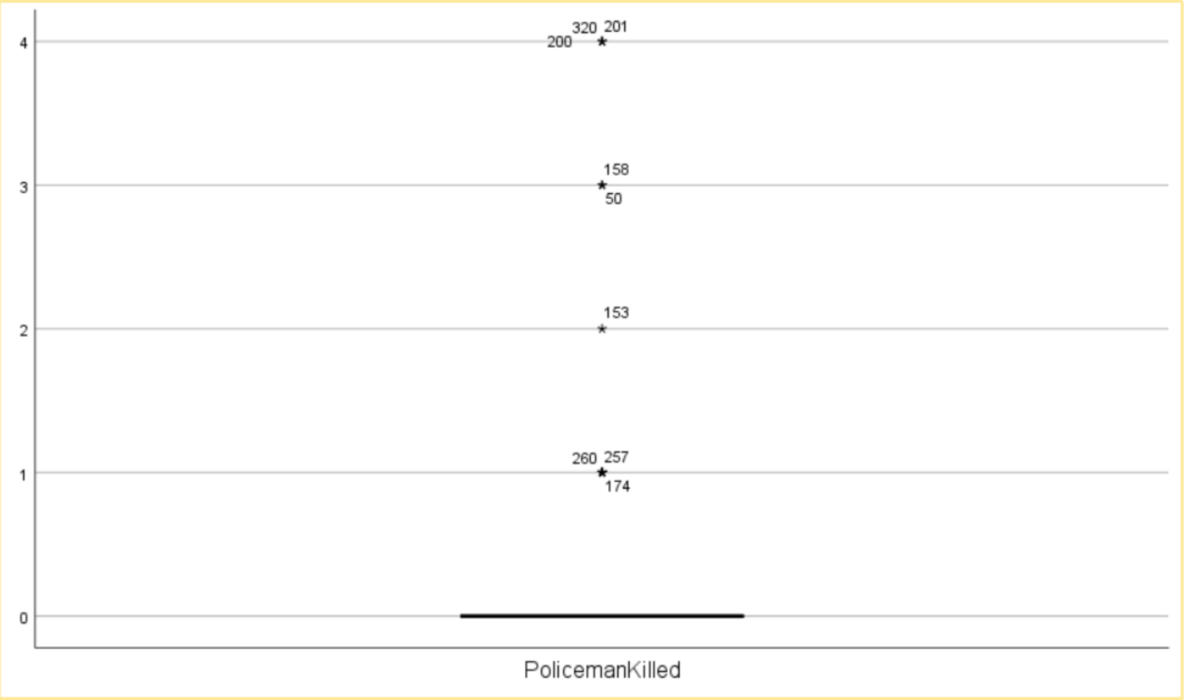
I deleted these two cases, and proceeded to delete the EmployeeDYN variable, a categorical variable, which was missing 79% of the dataset's values. Additionally I deleted the cases with missing values of longitude and latitude, as this was just 20 cases and made all the quantitative values free of missing values.

2. Outliers

When searching for outliers in the data I used boxplots to visualize the variables' data. The common occurrence between multiple boxplots was case 4 for injuries, total victims, and fatalities. Before deleting this outlier I had to think about the context of the problem. Case 4 is an outlier, but in the context of this dataset is it really an outlier? The dataset is of mass shootings in the US and case 4 is the most fatal mass shooting in the US. In this case I decided to delete the case because it was such an outlier relative to other shootings but this should be taken into account with our insights and conclusions based on the data. The boxplot below shows the fatalities and shows how 4 is such an outlier even relative to the other extreme values.



Below is the boxplot for policeman killed which may at first seem like outliers but clearly isn’t when you review the dataset; the difference between one policeman being killed and four being killed isn’t actually a significant difference.

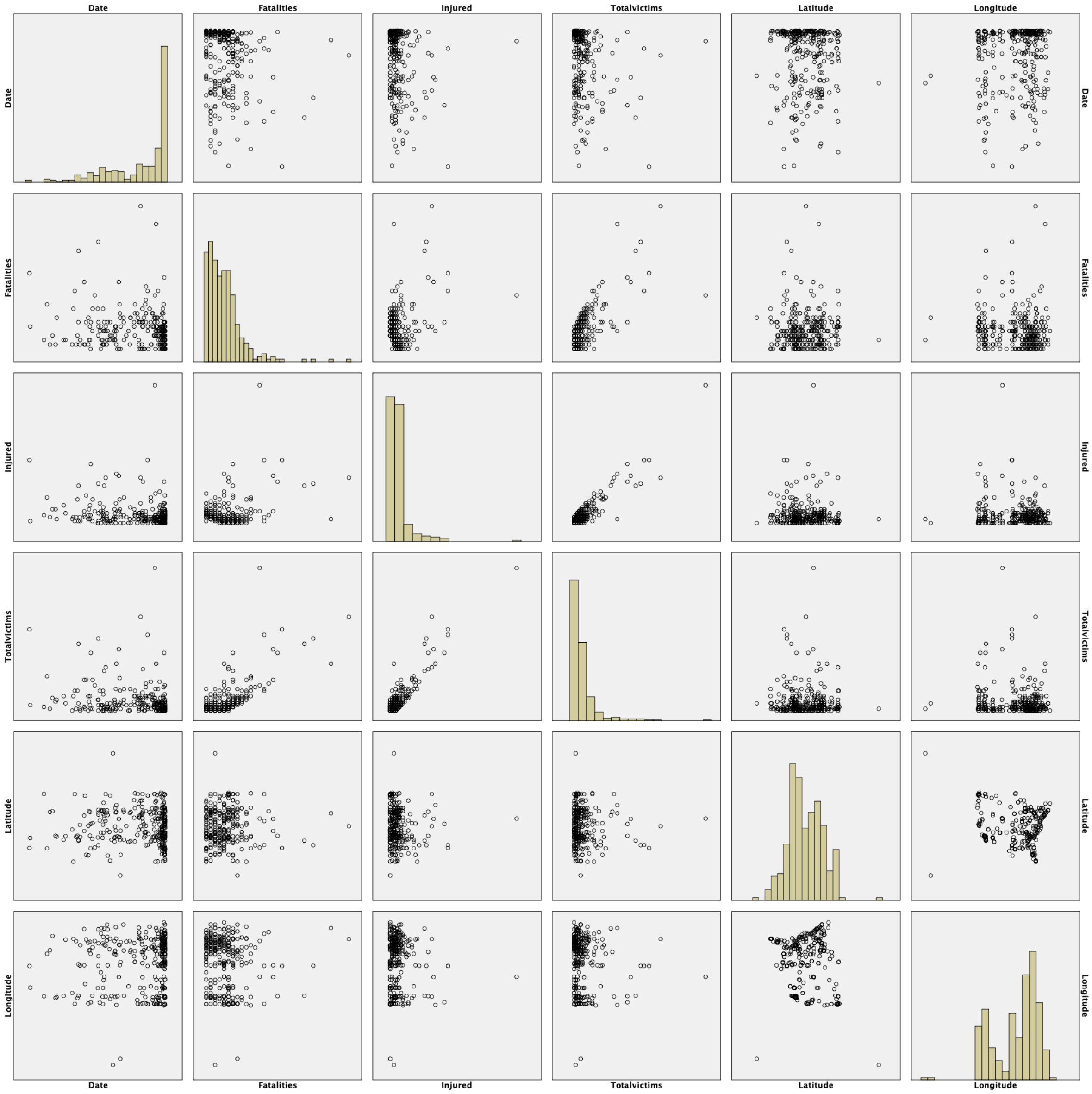


And this is all of the outliers that had potential to be removed, so this part of the examination is finished.

Testing Assumptions

1. Linearity

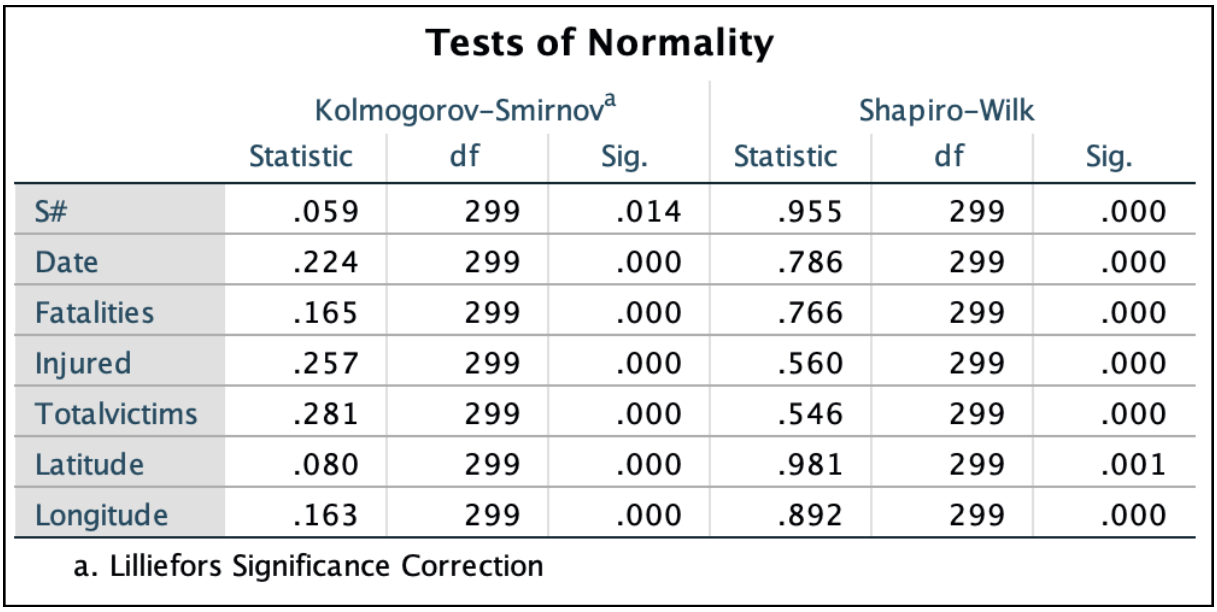
To check for the linearity we created scatter plot graphs that showed the correlation between the independent variables in the data set. The results are shown below:



This is the linearity test for all the independent variables in the data set. As we can observe in the scatter plots, there is linearity in the total victims and injuries graph (3rd row 4th column, or 4th row 3rd column), but then we see some null linearity in the graph that shows latitude and longitude(6th row 5th column).

2. Normality

To check for the normality we used Analyze → Descriptive Stats → Explore, where we chose al the scale variables as the dependent list. Then on went got to plot and once there we selected the normality plots with test and the results obtained are shown below:



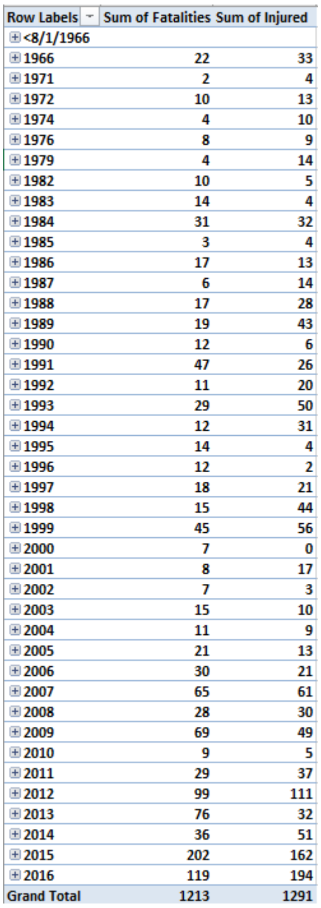
Tests of Normality

	Kolmogorov–Smirnov[a]			Shapiro–Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
S#	.059	299	.014	.955	299	.000
Date	.224	299	.000	.786	299	.000
Fatalities	.165	299	.000	.766	299	.000
Injured	.257	299	.000	.560	299	.000
Totalvictims	.281	299	.000	.546	299	.000
Latitude	.080	299	.000	.981	299	.001
Longitude	.163	299	.000	.892	299	.000

a. Lilliefors Significance Correction

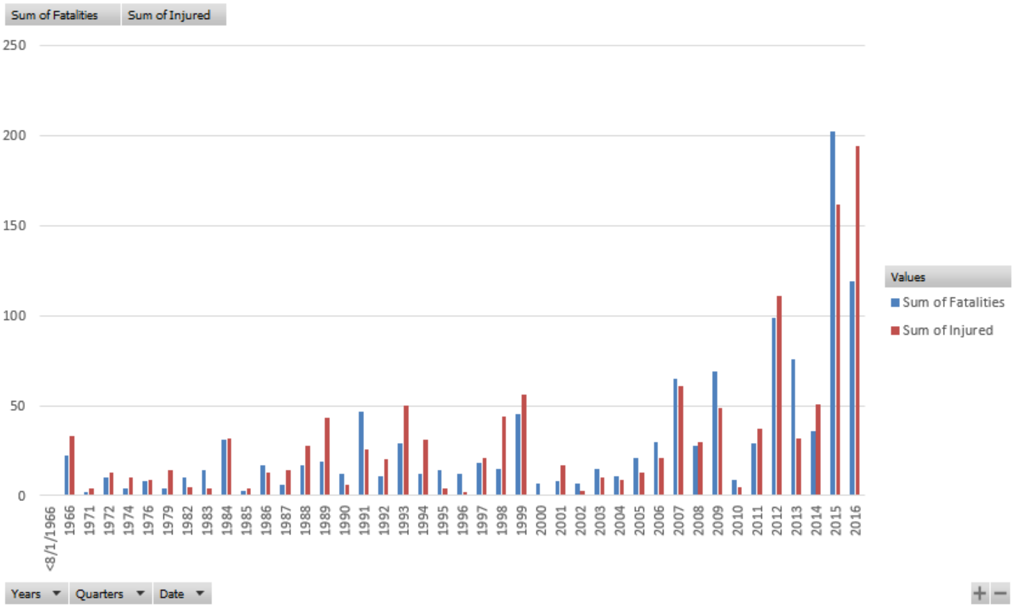
As we can observe on the graph, there is not a high significance level on the Kolmogorov-Smirnov test which means that the data is normal, except the S# which is only the number of the case.

Data Insights



Row Labels	Sum of Fatalities	Sum of Injured
<8/1/1966		
1966	22	33
1971	2	4
1972	10	13
1974	4	10
1976	8	9
1979	4	14
1982	10	5
1983	14	4
1984	31	32
1985	3	4
1986	17	13
1987	6	14
1988	17	28
1989	19	43
1990	12	6
1991	47	26
1992	11	20
1993	29	50
1994	12	31
1995	14	4
1996	12	2
1997	18	21
1998	15	44
1999	45	56
2000	7	0
2001	8	17
2002	7	3
2003	15	10
2004	11	9
2005	21	13
2006	30	21
2007	65	61
2008	28	30
2009	69	49
2010	9	5
2011	29	37
2012	99	111
2013	76	32
2014	36	51
2015	202	162
2016	119	194
Grand Total	1213	1291

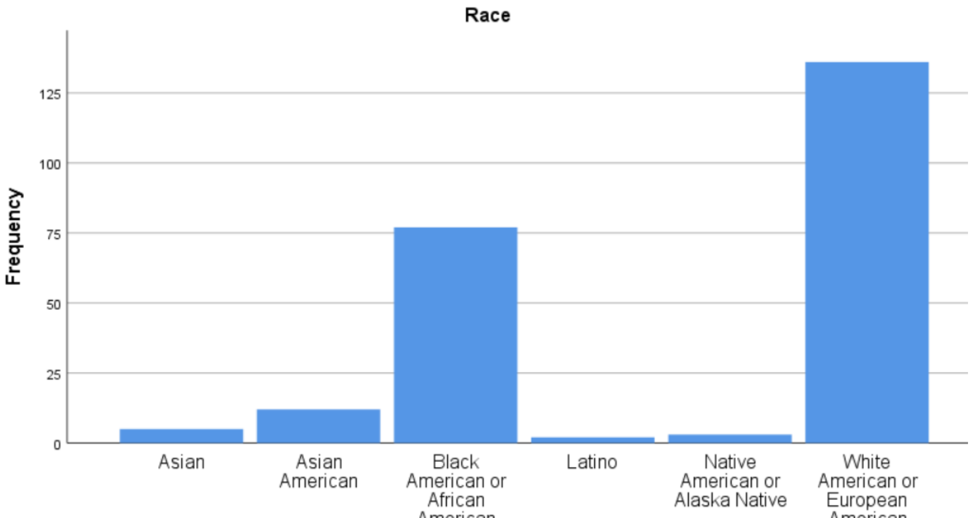
1. How many people were killed and injured per year?



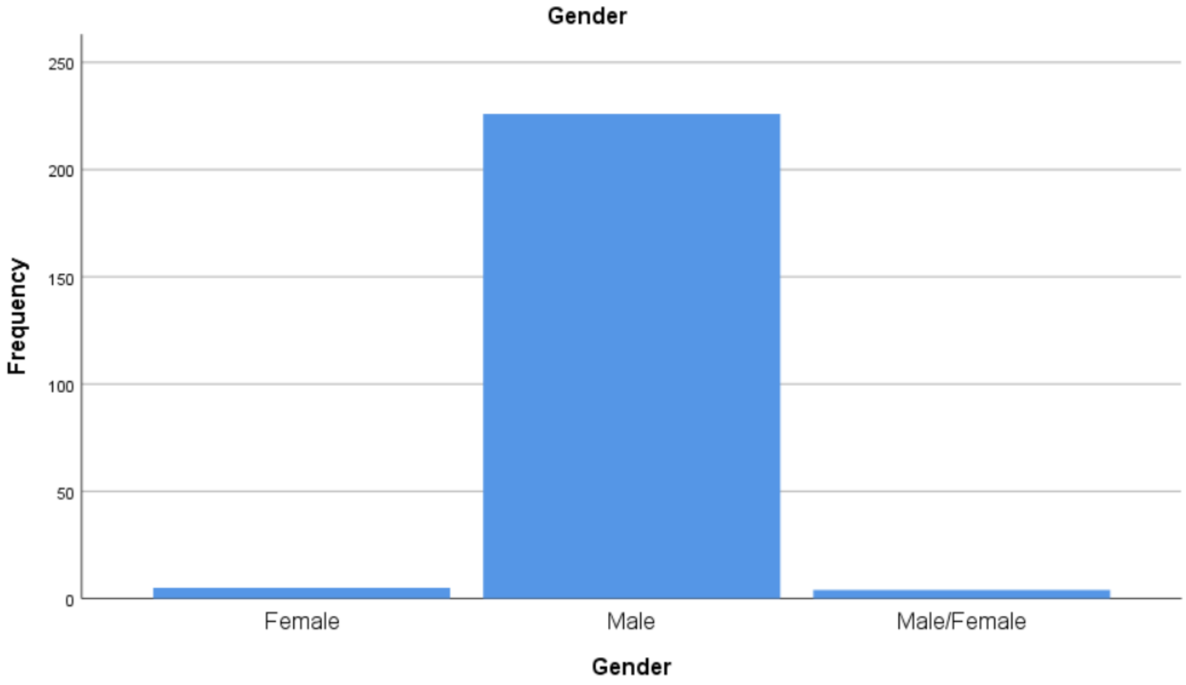
The graph above visualizes the fatalities and injuries per year from the dataset, with the years ascending from left to right. The data shows that the last half-decade has been much worse in terms of fatalities and injuries. Accompanying the graph is a pivot table showing the table version of the same data, as well as a total for all the years. This is with some cases removed from the dataset so the total is much lower than it would have been without cleaning the data via omitting cases with missing values.

2. Is there any correlation between shooter and his/her race, gender?

Every case in this dataset is one shooter, so I will use Analyze → Descriptive Stats → Frequencies to compare the number of shooters of each race and of each gender in order to see the relationship which

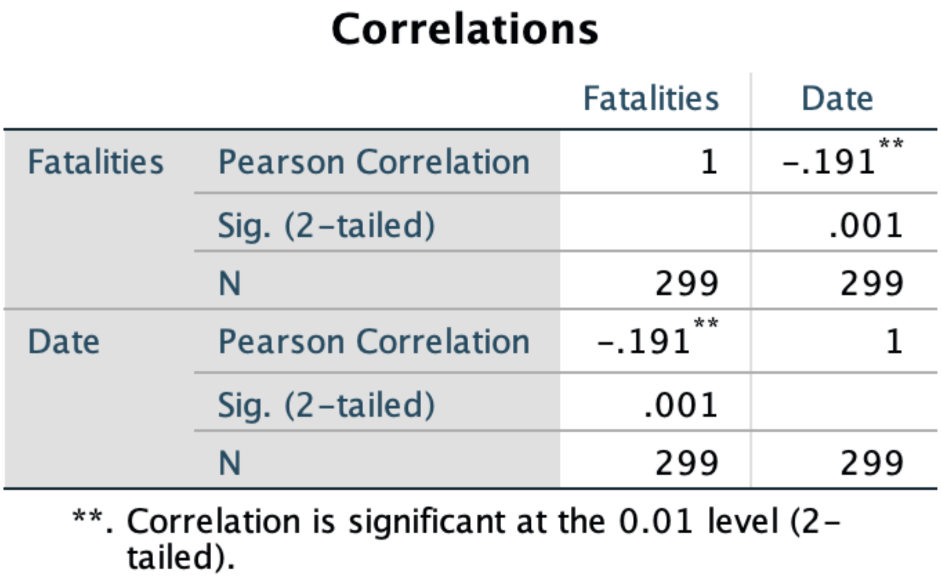


may be present in this dataset. The bar chart of frequency of Race shows That the majority of the mass shootings were committed by White American / Caucasian people. African American is the race with the 2nd highest frequency, which combines with the White Americans to make up virtually all of the shootings, as the rest of the races have very small frequencies. The next bar graph below is for Gender, representing how many shootings were committed by Males and Females. This shows that Males make up almost all of the shootings. There were roughly 6 shootings committed by Females, and fewer by an unknown gender shooter. Overall there is a clear pattern where White Males commit the most mass shootings, with Black Males committing the second most.



3. Any correlation with calendar dates? Do we have more deadly days, weeks or months on average?

For this question we had to do two different correlation analysis, the first of them which didn't require any data transformation consisted of checking the correlation between fatalities and the variable dates. Here are the results:

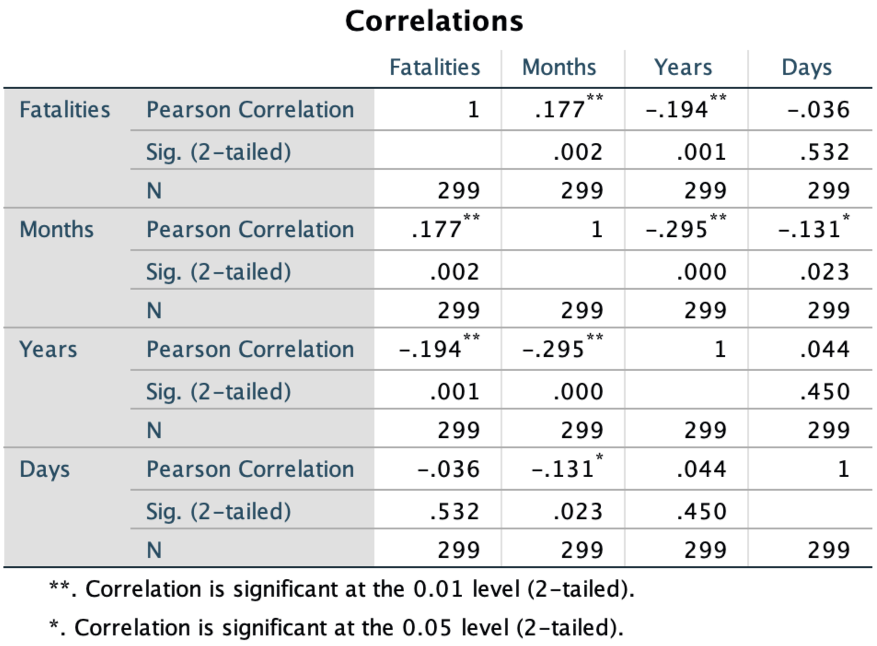


Correlations

		Fatalities	Date
Fatalities	Pearson Correlation	1	-.191**
	Sig. (2-tailed)		.001
	N	299	299
Date	Pearson Correlation	-.191**	1
	Sig. (2-tailed)	.001	
	N	299	299

**. Correlation is significant at the 0.01 level (2-tailed).

As we can see in the graph above, there is a small negative correlation between these two variables. Then for the second question, since the data has the dates with years, months and days we had to create new variables with each of them specifically. We did so by using the Date and Time Wizard tool under transformation. Once there we chose to extract years, then months, and then days to create three new separate variable to see which one had a better correlation with fatalities. The results shown below show us that we have more deadly months on average.



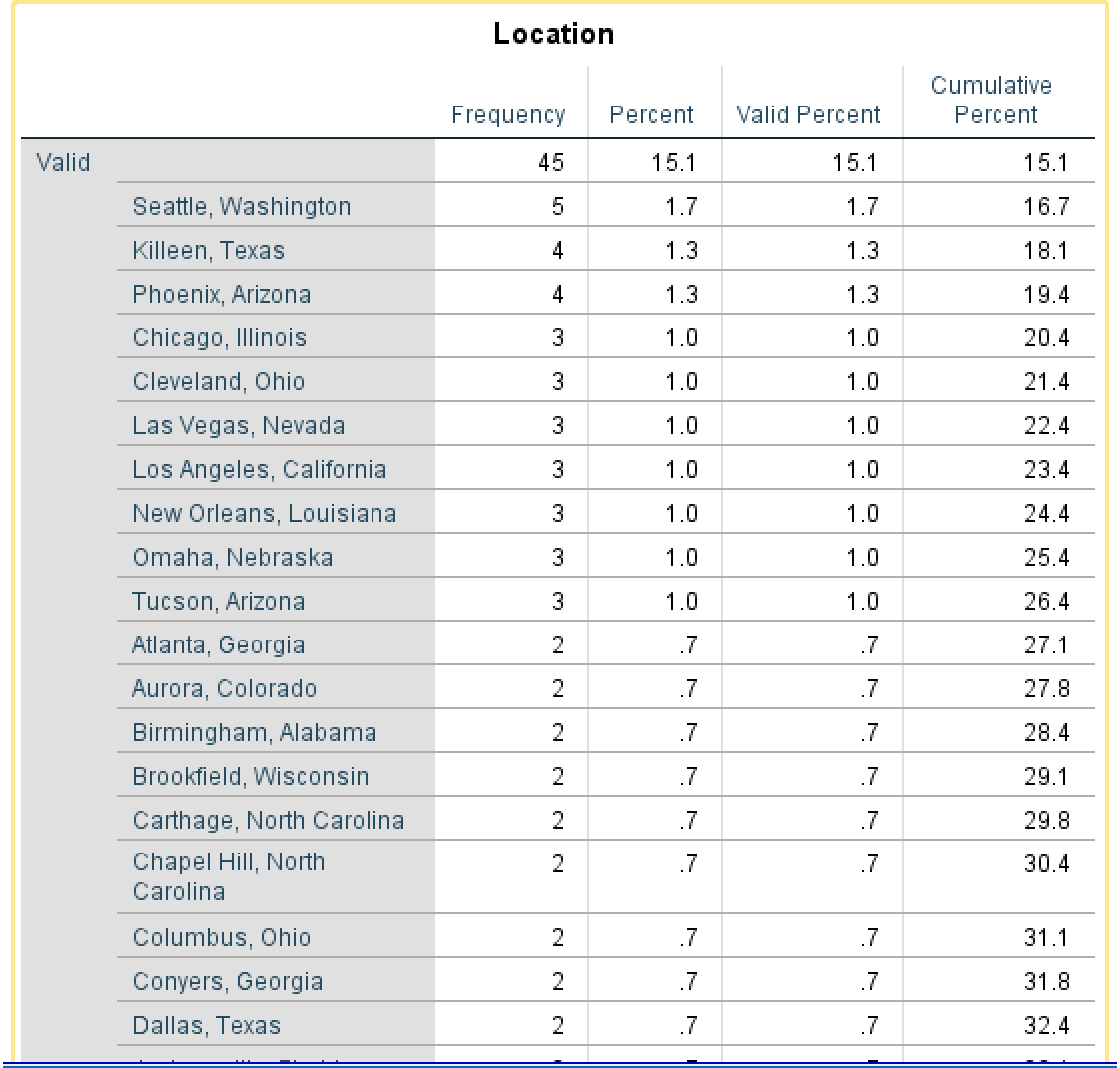
Correlations

		Fatalities	Months	Years	Days
Fatalities	Pearson Correlation	1	.177**	-.194**	-.036
	Sig. (2-tailed)		.002	.001	.532
	N	299	299	299	299
Months	Pearson Correlation	.177**	1	-.295**	-.131*
	Sig. (2-tailed)	.002		.000	.023
	N	299	299	299	299
Years	Pearson Correlation	-.194**	-.295**	1	.044
	Sig. (2-tailed)	.001	.000		.450
	N	299	299	299	299
Days	Pearson Correlation	-.036	-.131*	.044	1
	Sig. (2-tailed)	.532	.023	.450	
	N	299	299	299	299

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

4. What cities and states are more prone to such attacks?



Location

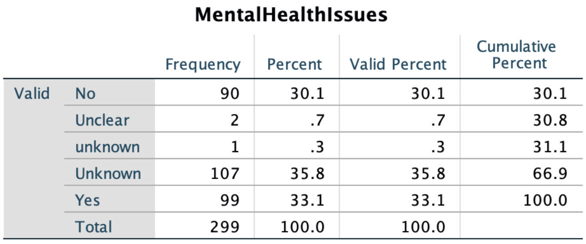
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		45	15.1	15.1	15.1
	Seattle, Washington	5	1.7	1.7	16.7
	Killeen, Texas	4	1.3	1.3	18.1
	Phoenix, Arizona	4	1.3	1.3	19.4
	Chicago, Illinois	3	1.0	1.0	20.4
	Cleveland, Ohio	3	1.0	1.0	21.4
	Las Vegas, Nevada	3	1.0	1.0	22.4
	Los Angeles, California	3	1.0	1.0	23.4
	New Orleans, Louisiana	3	1.0	1.0	24.4
	Omaha, Nebraska	3	1.0	1.0	25.4
	Tucson, Arizona	3	1.0	1.0	26.4
	Atlanta, Georgia	2	.7	.7	27.1
	Aurora, Colorado	2	.7	.7	27.8
	Birmingham, Alabama	2	.7	.7	28.4
	Brookfield, Wisconsin	2	.7	.7	29.1
	Carthage, North Carolina	2	.7	.7	29.8
	Chapel Hill, North Carolina	2	.7	.7	30.4
	Columbus, Ohio	2	.7	.7	31.1
	Conyers, Georgia	2	.7	.7	31.8
	Dallas, Texas	2	.7	.7	32.4

.

Analyze → descriptive stats → frequencies → variable = location → format → order by = descending counts → multiple variables = organize outputs by variables. This is the process that was made to obtain the data in the chart from above. Seattle is the only location that has 5, which makes it the city that is most prone to these attacks. Followed by Killeen and Phoenix that are the only two cities that have a frequency of 2.

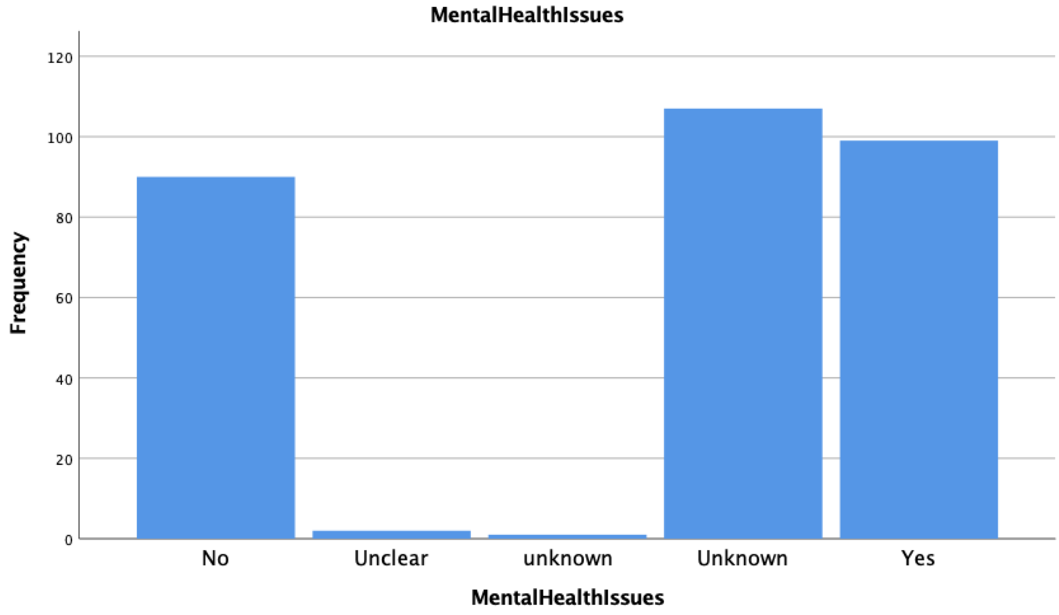
5. How many shooters have some kind of mental health problem?

Again, every case in this data set is one shooter, therefore we will use Analyze → Descriptive Stats → Frequencies to check whether those had any mental health issues. The results obtained showed that the highest percentage of shooters was unknown whether they had any mental problems or not, although the second highest was that they did have with only a frequency of 9 lower, which is a lower 3.1%. The rest of them where either that they didn't have any problem with a 30.1% or that it was unclear with only .7%.



MentalHealthIssues

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	90	30.1	30.1	30.1
	Unclear	2	.7	.7	30.8
	unknown	1	.3	.3	31.1
	Unknown	107	35.8	35.8	66.9
	Yes	99	33.1	33.1	100.0
	Total	299	100.0	100.0	



7. Can you model the number of killed (and/or injured people) based on the data you collect?

No, since most of the variables are string variables you cannot use them as predictive variables, the only numerical variables that you have are your dependent variables and date.A model needs independent variables that are going to be used for predictions, but you lack those in this dataset.

Furthermore, if you give a numeric value to variables like gender and try to use that variable as a predictive variable is going to be useless since the value that they have has no real meaning. Maybe the only way to create a model with this data set would be trying to give a value to variables like mental health and use that variable as the dependent variable. And try to predict that variable with the numeric variables.