Correlation in Cause of Death: Investigating Unconventional Relationships

M, CONDON, University of Colorado, Boulder
D. GORTHY, University of Colorado, Boulder
B. MELVIN, University of Colorado, Boulder

Conventional fatality studies focus mainly on demographic trends, disease prevalence and temporal trends. The thoroughness of the CDC's data collection allows other, unconventional factors to be considered. This project is comprehensive and will demonstrate every step of a data mining project.

General Terms: Data Mining, Preprocessing, Data Gathering, Trend Prediction, Data Cleaning

Additional Key Words and Phrases: Python, R, Hadoop, Hive

1 ABSTRACT

Conventional analysis of the annual Center for Disease Control and Prevention's dataset focuses primarily on simple totals of each cause of death, race, sex, and other simple variables. Each year's totals, when combined with other year's, shows the trend of these variables over time. This project aimed to examined unconventional measures and their effect on the categories that are traditionally researched. Some of the external, unconventional data includes rare events such as full moon occurrences, above-average solar activity and the S&P 500 movement. Of the unconventional data, we identified certain causes of death to be above the local average. Additionally, we used attributes within the CDC's data to find correlations not traditionally reported.

2 INTRODUCTION

The Center for Disease Control and Prevention (CDC) gathers mortality statistics for the United States on an annual basis. We will begin our project with the data gathered by the CDC in 2014. The CDC mines the data for common causes of death among large demographics and provides a basis for comparison to other years. We will begin by replicating the results given by the CDC to ensure that we establish an accurate baseline to gauge the validity of our results. We will answer questions such as: *Is there a correlation between sex and cause of death?* We will then expand our analysis to estimate the income of an individual by factoring in race, age, marital status and day of death to answer questions like: *Is there a correlation between the day of the week of death and cause of death?* We will then deviate from traditional analysis and search for relationships between factors like day of death, place of death and manner of death. Finally, we will estimate the date of each death and pull in external financial, solar activity and accident datasets. Hopefully, we can answer questions such as: *Does the number of deaths spike in the event of above average solar activity?* The focus of our project will be on questions that CDC and traditional secondary sources do not look to answer.

These questions, especially those that use unconventional data, are not meant to be comprehensive. Rather, they show the importance of considering factors that are not obviously correlated.

3 RELATED WORK

3.1 Center for Disease Control

The work that has been done in previous years from CDC is trends of all deaths from previous years until 2016. There is an in-depth document written by CDC that explains trends such as causes of death associated with a certain demographic. Another focus of the CDC is investigating the mortality rate of and change in mortality of infant and newborn deaths.

Primarily, the CDC focuses on tracking mortality causes and rates temporally for many demographics; for example, CDC first identified an increase in the number of deaths attributed to Alzheimer's in recent years.

3.2 Secondary Analysis

Organizations such as LiveScience take the data provided by the CDC and investigate questions not researched by the primary source. Examples of the work that LiveScience mined from the information were "Top 2 Deadly Diseases", "Top Killers", and "Respiratory Diseases & Attacks". LiveScience, among many other secondary researchers, focus on generating graphics to spread awareness of diseases and other preventable causes of death common in the United States.

4 DATA SET

4.1 Center for Disease Control Data

The CDC's dataset consists of information collected by doctors and other officials upon death. Key components of this set are residential status, education, day of death, age, race, sex, marital status and place of death. Unfortunately, the CDC is not legally allowed to include the exact date, time or location of death, but the attributes provided give enough information to estimate the features that we required.

4.2 External Data

The CDC's dataset does not provide enough information to estimate the geographic location of death, ruling out the ability to include data that requires a location like temperature, precipitation, and firearms ownership. However, we developed a technique to estimate the actual data of death. Using this, we added in supplementary data not traditionally considered in mortality analysis. The list of relevant data includes federal holidays, Friday the 13ths, full moons, sun spot number, mass shootings, and the S&P 500's performance. The purpose of the supplementary data is not to definitively find a definite factor of changes in the death rate, but rather show that interesting relationships can be derived from unconventional sources of data.

5 KEY DATA MINING TECHNIQUES

5.1 Data Preprocessing

Death records are standard and strictly monitored; consequently, outliers and missing data do not pose a problem. However, the most important missing attribute for our unconventional analysis is the estimated date of death for each entry. This required research into other studies that had the month of death and the day of the week of death, but did not have the full date. Unfortunately, most datasets with any mention of date has the full date as an attribute.

After finding a couple of estimation techniques that are commonly used in situations when there is not attribute related to date, we ran a couple of trial runs to determine different distributions with which to estimate each date. After settling on a uniform distribution, we ran this estimated each entries' date eight times and choosing the most frequent estimation. The uniform distribution was created by counting the number of days in each month (say the number of Mondays in March) and then assigning a probability to each date. If there are four Mondays in March, then each date that is a Monday in March is assigned a probability of 25%. If eight estimates do not give a definitive estimation, then a date is chosen randomly from the data with the highest estimation count.

Once each entry was assigned an estimated date, we were then able to merge the unconventional datasets with the CDC's dataset. These external datasets either give a daily measure (S&P 500 performance and solar activity data) or a list of infrequent occurrences (Friday the 13th, full moons and mass shooting counts).

Before continuing, we could remove some of the redundant attributes of the CDC's dataset. Some of these include the more general race codes, and the too-specific death codes.

5.2 Data Warehousing

The size of the CDC set is less than 0.1 GB. This required us to split the set up into several small sets to meet Github's storage requirement, but it made calculating totals relatively fast. Because of the amount of time we spent on preprocessing the CDC's dataset and the amount of time needed to analyze the data, we did not spend much time on structuring our program to use a database, either persistent or at runtime. A simple way of storing this data is to create a dictionary in Python upon request. This was especially useful for Bayesian calculations.

5.3 Data Classification/Clustering

We initially started by trying to build a script for creating a decision tree. Due to the nature of the data, we decided that building a process to accept requests to perform a Bayesian analysis on. This proved to be one of the most valuable tools when looking for significant differences between causes of death and other attributes. The process that we used to automate the finding of significant results was to run many tests and calculate differences that may be significant. Results from these runs will be discussed later in *Key Results*.

We considered clustering analysis; however, this only would make sense if we had location data. The CDC does provide this data upon request, but we never received a response for the extended attributes.

Finally, we developed the analysis script in R after developing it in Python to test the processing time difference. The size of the set was small, so the difference in execution time was not significant to our project. However, this is an important consideration if the size of our dataset were to scale to several Terabytes, instead of fractions of a Gigabyte.

6 KEY RESULTS

Of the hundreds of tests that we ran, we took those that found significant differences between probabilities. For example, most of this analysis was performed using the Bayesian technique learned in this class. If one category yielded a significantly higher or lower probability of another event happening, we noted this difference and graphed the results. The most significant results are shown in the *Appendix*.

6.1 Conventional Results

Cause of death with respect to the values of other attributes was a worthwhile and logical starting point for our research. *Graphic 1* shows the probability of the three most prevalent causes of death besides "natural" – homicide, suicide and accident – vary depending on the sex of the individual. The largest difference is the probability that the death is an accident given the sex of an individual – 3.2% for men and 1.8% for women. Men were also more likely to die by suicide – 1.3% for men and 0.4% for women. Additionally, the probability of death by homicide for was 0.5% for men and 0.1% for women. This did align with prevalent homicide and suicide studies.[7]

After finding that the cause of death of yielded several comparisons that had significant differences, we went on to examine the three most common causes of death besides "natural" for ages (separated into 5 and 10 year bins). The results from this analysis, controlling for individuals with an undetermined age, resulted in *Graphic 2.* It is an obvious statement to conclude that the probability of a certain cause of death varies with age; this is especially true considering that the CDC's dataset contains occurrences of all death records, not merely adult records. Again, we replicated the conclusions of results from prominent suicide-prevention groups.[7] Additionally, we drew some less obvious conclusions from the data. The chance of the cause of one's death being attributed to homicide peaks in a person's 20s. However, the chance of death by accident trends upwards and eventually peaks in the 45-54-year range, then declines. The change of an accidental death then increases for those 85 years and older. While this seems an obvious conclusion, our initial research did not find this to be a common conclusion nor a popular focus of research.

While the cause of death with respect to day of week, shown in *Graphic 3*, did not vary from day-to-day, it is important to note that accidental deaths peaked on the weekends, and that suicides happened most often on Mondays. This last conclusion was not reported by the source that performed extensive temporal analysis on suicides in the United States.

Finally, *Graphic 4* shows a significantly higher occurrence of homicides for those that are single 0.4% compared to the marital status of the other three categories hovering around 0.1%.

6.2 Unconventional Results

We then moved on to examine results using the supplementary data that we pulled in to expand the breadth of our research.

These tests seemed to produce many fewer results with significant differences that did the traditional tests. *Graphic 5* and *Graphic 6* show some interesting results that we found when looking at suicides on each date. We can conclude that the frequency of suicides peak in August and September, while reaching a low during the winter holiday months. Pulling in the full moon data, we found that most of the days that have a full moon are local maximums relative to their neighboring points. The bar chart shows that it is more likely that someone commits suicide given that it is a full moon (0.7%) compared to when it is not a full moon (0.6%). Considering our estimation technique, it was unclear whether this difference was attributed to chance, was an error in date estimate, or was a true cause of suicides. To test this, we generated a second set and ran estimated the dates using the technique used in the original set. This yielded a different result than the first run and is shown in *Graphic 10*.

Graphic 8 shows the scatter plot of the number of natural deaths compared to the daily sunspot number. This was the comparison that gave the most positive correlation between two sets of data. Although the variation implies that this is not necessarily a meaningful correlation, it is likely that our date estimation dampened any meaningful correlation between sunspot number and the number of natural deaths. It is important to note that this set was cleaned of the outlying sunspot measurements, as shown in *Graphic 7*. Because this distribution is nearly normal, we removed the all the points where the sunspot number did not fall between 50 and 175.

Finally, *Graphic 9* shows the test that we were all most interested in when looking for supplementary data. Common sense would conclude that in the event of a large change in the economy, the suicide rate for that specific day would be largely affected. We assumed that very negative stock performance would correspond with a higher number of suicides, while a bullish market would significantly decrease the number of suicides. The trendline is included to show that this relationship is the opposite; the variation in this test implies that the economic performance on any given day, at least in 2014, did not have any significant effect on the number of deaths, including suicides. An interesting extension to this analysis would be to include several years and isolate only extreme changes. If we could obtain the suicide statistics of New York, Chicago and other financial centers, we could narrow our search to relevant areas of the United States.

7 APPLICATIONS

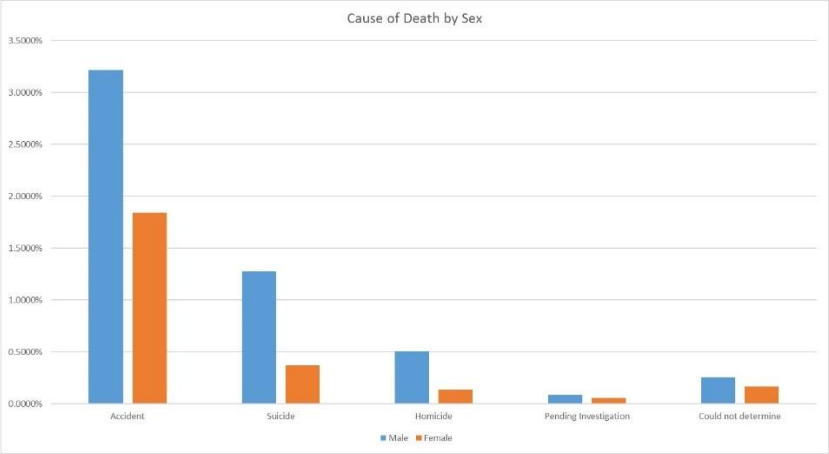
A large portion of our project was learning to identify attributes that are necessary to the project well before the analysis and classification portion. This required a comprehensive understanding of the dataset, the further preprocessing steps, our storage limitations, and the questions that we hoped to answer with our research and analysis. This project was particularly difficult to grasp the best way in which to estimate a necessary attribute – in our case, this attribute was date of death.

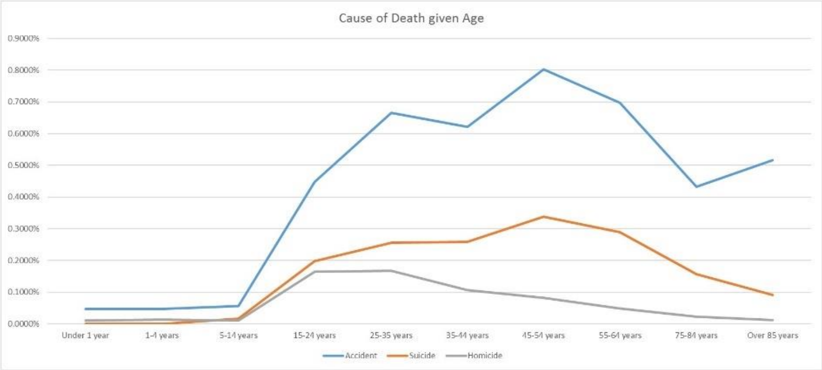
To focus on the former, this task was made easier by realizing that several of the attributes like cause of death are redundant. We dealt with this redundancy by retaining the least specific attribute as well as the most specific attribute. The most specific measure of cause of death yielded populations, especially in the case of rare diseases, had a sample size that was too small. The least specific measurements allowed us to separate data into larger populations that showed a clear trend.

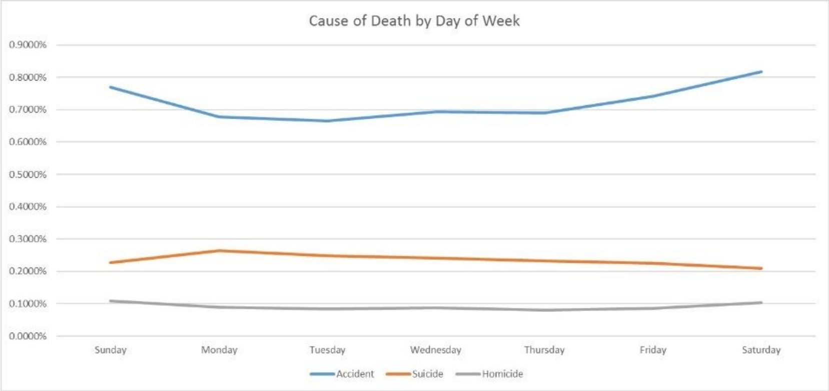
Focusing on the latter, date estimation was very important to the purpose of our project. In the case that the results rely on the best estimation of an attribute, it is important to dedicate a large portion of the project to determining what distribution the attribute should follow and then calculate several samples of the attribute given the distribution assumptions. In our project, we showed that our estimation nearly dampened any correlation that we found. More broadly, this shows the importance of understanding how any assumption with affect the results of an analysis.

Finally, our project showed the importance of considering factors that are not conventionally attributed to affecting a measurement. This point is more philosophical, but examining the unconventional correlations with the conventional can seek to validate common sense or folklore assumptions. In our project, this is demonstrated by disproving the assumption of a correlation between suicide rate and financial performance. On the other hand, this sample yielded a significant difference between the suicide rate and the presence of a full moon. Although our sensitivity analysis concluded that a different estimation of dates yielded a less-certain correlation, this was an interesting result. It also points to the importance of sensitivity analysis when basing conclusions in estimated values.

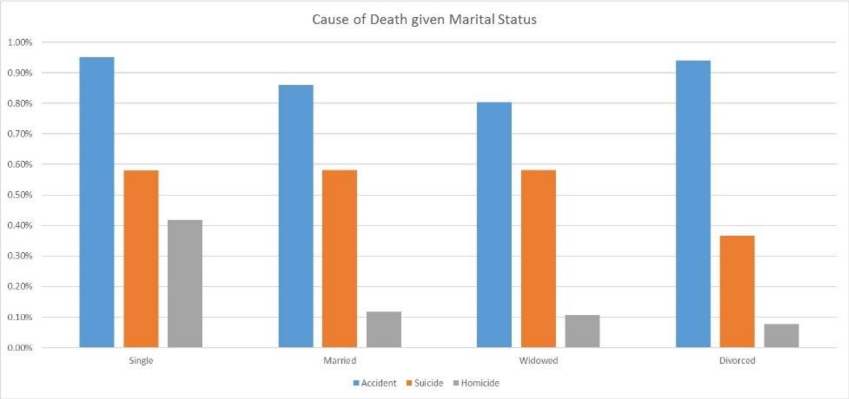
APPENDIX

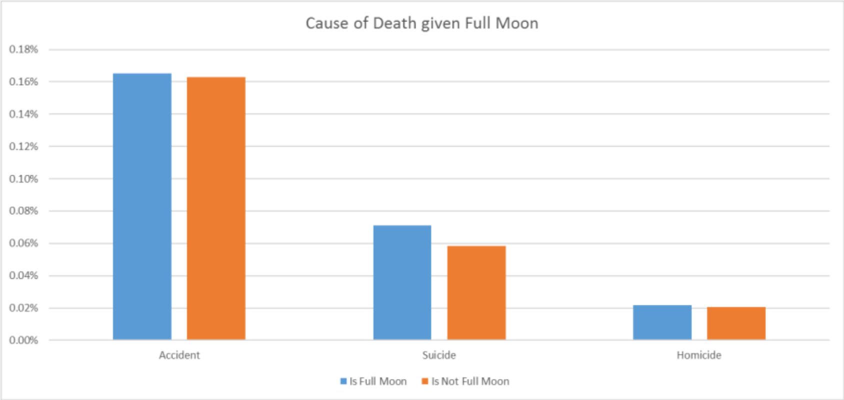
<table>
<tr><td>Graphic 1</td><td>


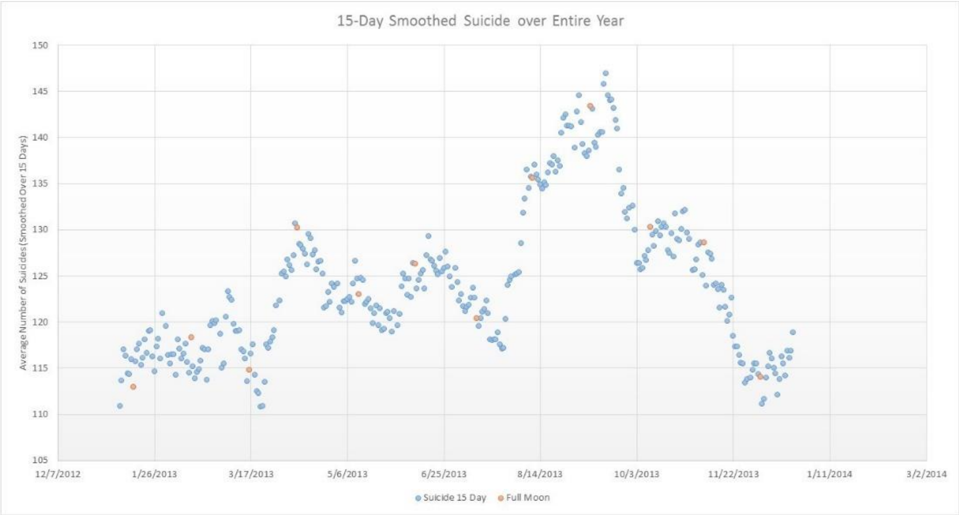
</td></tr>
<tr><td>Graphic 2</td><td>


</td></tr>
<tr><td>Graphic 3</td><td>


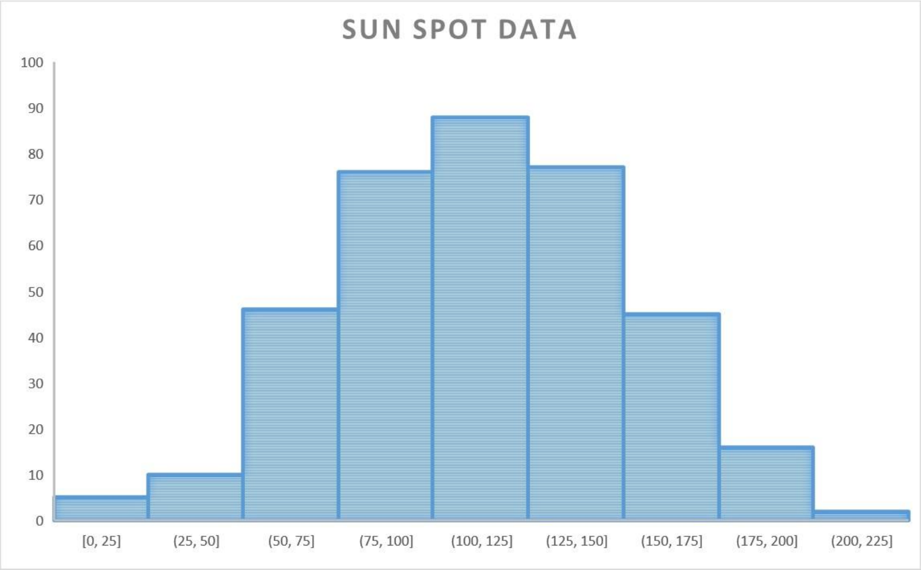
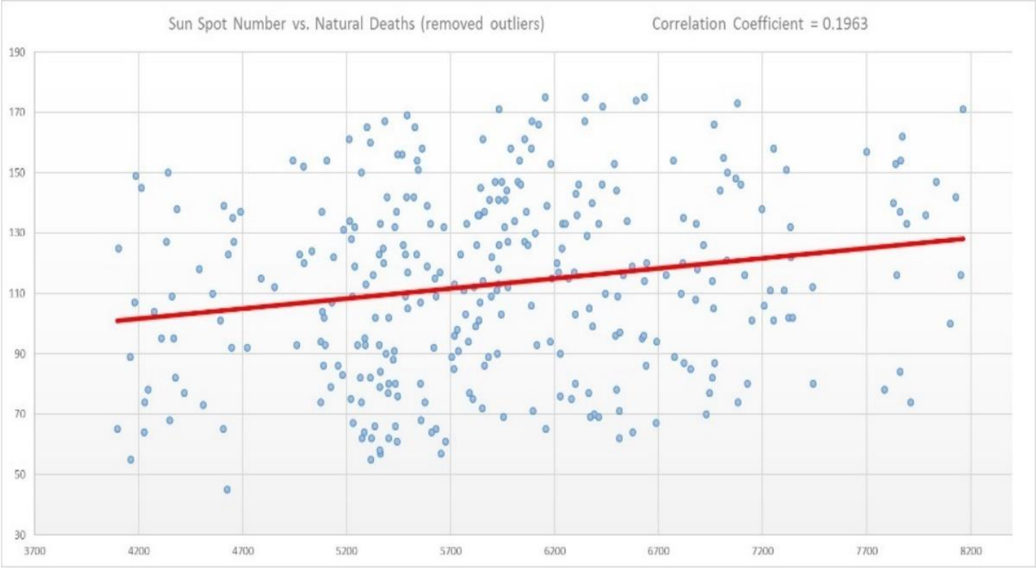
</td></tr>
</table>

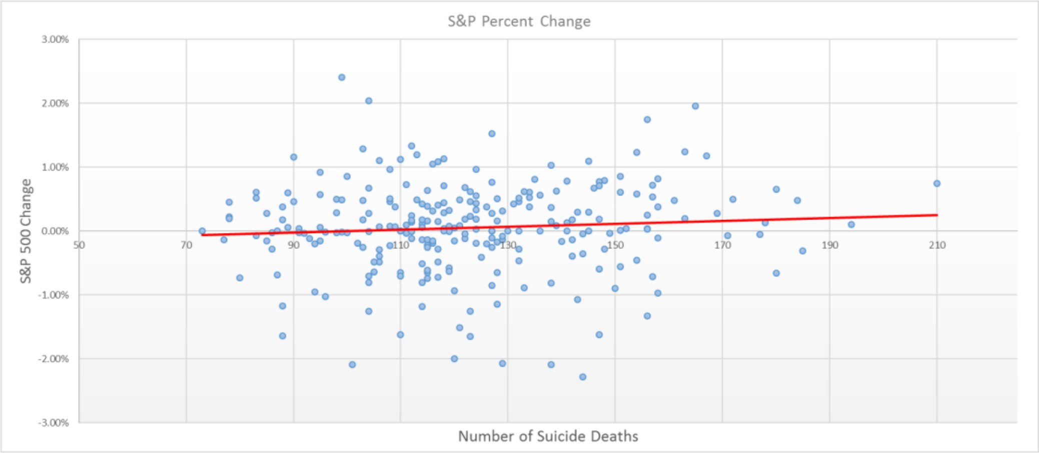
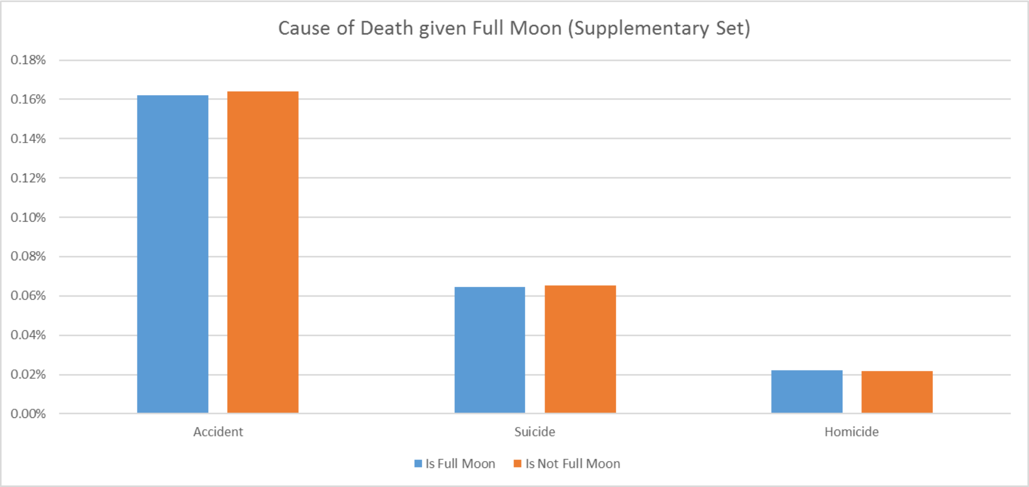
<table>
<tr>
<td><b>Graphic 4</b></td>
<td>


</td>
</tr>
<tr>
<td><b>Graphic 5</b></td>
<td>


</td>
</tr>
<tr>
<td><b>Graphic 6</b></td>
<td>


</td>
</tr>
</table>

Graphic 7	
Graphic 8	

Graphic 9	
Graphic 10	

REFERENCES

[1] "Death in the United States." Kaggle. Center for Disease Control and Prevention, 2016. Web. 2 Mar. 2017.
[2] DeNavas-Walt, Carmen, and Bernadette D. Proctor. "Income and Poverty in the United States: 2014." Census.gov. United States Census Bureau, Sept. 2014. Web. 2 Mar. 2017.
[3] Geggel, Laura . "The Odds of Dying." LiveScience. N.p., 9 Feb. 2016. Web. 2 Mar. 2017.
[4] "Historical Treasury Rates." U.S. Department of Treasury, n.d. Web. 2 Mar. 2017.
[5] "National Solar Radiation Data Base." National Renewable Energy Laboratory, n.d. Web. 2 Mar. 2017. <https://maps.nrel.gov/nsrdb-viewer/>.
[6] "S&P 500 (^GSPC)." Yahoo Finance, n.d. Web. 2 Mar. 2017. <https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC>.
[7] "Suicide Statistics." American Foundation for Suicide Prevention, Web. 28 Apr. 2017. < https://afsp.org/about-suicide/suicide-statistics/>.