PROJECT : TRAFFIC COLLISION

April 20, 2019

Tianheng Ma, 2nd Year Computer Science, UCSD
Xinran Wang, 3rd Year Probability and Statistics & Economics, UCSD
Zhao Jin, 1st Year Data Science, UCSD
Table number: 19

INTRODUCTION

These datasets contain vehicle crashes and traffic levels that took place in San Diego. Data includes metadata on collisions and traffic counts.

In this project, we are going to analyze the basic numerical data in the collision dataset, we will also use statistical testing to see if there are more deaths involved in a collision happened on U.S. holiday than a normal day. Moreover, we start combining the two dataset to analyze the relationship between the results of traffic collision and the level of traffic. By using the San Diego Police Department data regarding the police beat, we analyze the relationships between the injured/death number per square miles and the location.

THE DATA

The two raw datasets used in this study are from the city of San Diego's open data portal, https://data.sandiego.gov/, which is the real data relating to San Diego traffic and collisions. The two dataset include metadata on collisions and traffic counts. The vehicle collision data contains the data of all recorded traffic collisions that occurred between January 2017 and April 2019. The traffic level data records the total number of vehicles in a particular date and street between 2005 and 2019.

Dataset	Description	Columns
pd_collisions_datasd.csv	This dataset contains information about each collision	activityNumber activityDate beat StreetNo StreetDir StreetName StreetType CrossStDir CrossStName CrossStType violationSection violationType chargeDescription numberInjured numberKilled hitRunLevel
traffic_counts_dictionary_datasd.csv	This dataset contains 24 hours count about the number of vehicles in a given street	street_name limits northbound_count southbound_count eastbound_count westbound_count total_count count_date

INVESTIGATIONS

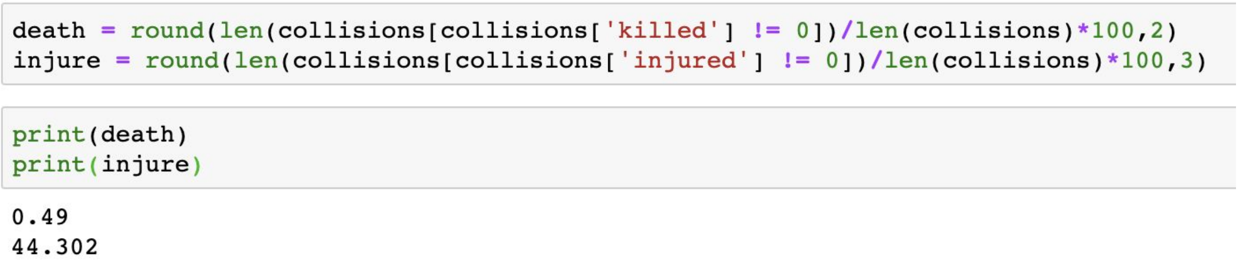
Import Data:

In order to analyze the data, we mainly use pandas as the operator.

Numerical Data Analysis:

Before investigating any categorical data, we start by analyzing the numerical data in `pd_collisions_datasd.csv`, the number of injured and the number of killed in a given collision.

By analyzing, we find that given a collision happens, there are 0.49% chance that the collision causes deaths, and 44.302% chance that the collision causes injuries. Also, we find that the highest number of injuries involves 180 injuries, which happened in December 15th 2017 Ocean view Boulevard crossed at MilBrae street, by the violation of YIELD RIGHT OF WAY TO PEDESTRIANS. The highest number of deaths involves 3 deaths, which happened in December 20th 2018 Village Ridge Road, by the violation of TURNING MOVEMENTS AND REQUIRED SIGNALS.

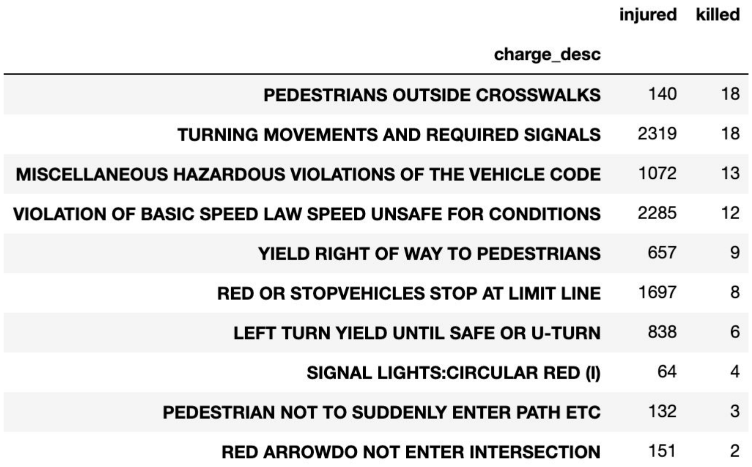


```
death = round(len(collisions[collisions['killed'] != 0])/len(collisions)*100,2)
injure = round(len(collisions[collisions['injured'] != 0])/len(collisions)*100,3)
```

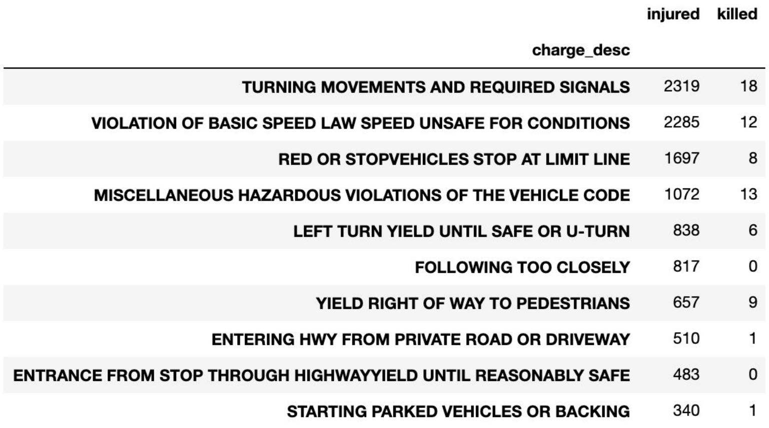
```
print(death)
print(injure)
```

```
0.49
44.302
```

By sorting the number of injured and the number of killed by the collision factor, we find that the top three factors that cause injuries are TURNING MOVEMENTS AND REQUIRED SIGNALS, VIOLATION OF BASIC SPEED LAW SPEED UNSAFE FOR CONDITIONS, and RED OR STOPVEHICLES STOP AT LIMIT LINE; whereas, the top three factors that cause death are PEDESTRIANS OUTSIDE CROSSWALKS, TURNING MOVEMENTS AND REQUIRED SIGNALS, and MISCELLANEOUS HAZARDOUS VIOLATIONS OF THE VEHICLE CODE.

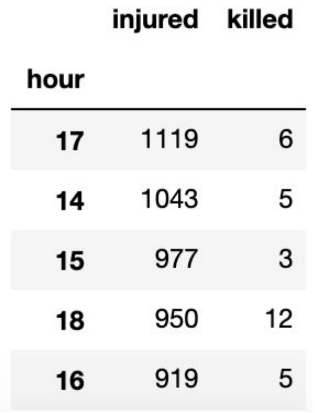


charge_desc	injured	killed
PEDESTRIANS OUTSIDE CROSSWALKS	140	18
TURNING MOVEMENTS AND REQUIRED SIGNALS	2319	18
MISCELLANEOUS HAZARDOUS VIOLATIONS OF THE VEHICLE CODE	1072	13
VIOLATION OF BASIC SPEED LAW SPEED UNSAFE FOR CONDITIONS	2285	12
YIELD RIGHT OF WAY TO PEDESTRIANS	657	9
RED OR STOPVEHICLES STOP AT LIMIT LINE	1697	8
LEFT TURN YIELD UNTIL SAFE OR U-TURN	838	6
SIGNAL LIGHTS:CIRCULAR RED (I)	64	4
PEDESTRIAN NOT TO SUDDENLY ENTER PATH ETC	132	3
RED ARROWDO NOT ENTER INTERSECTION	151	2

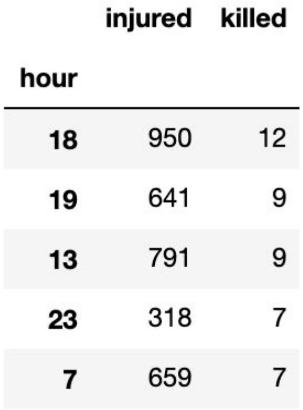


charge_desc	injured	killed
TURNING MOVEMENTS AND REQUIRED SIGNALS	2319	18
VIOLATION OF BASIC SPEED LAW SPEED UNSAFE FOR CONDITIONS	2285	12
RED OR STOPVEHICLES STOP AT LIMIT LINE	1697	8
MISCELLANEOUS HAZARDOUS VIOLATIONS OF THE VEHICLE CODE	1072	13
LEFT TURN YIELD UNTIL SAFE OR U-TURN	838	6
FOLLOWING TOO CLOSELY	817	0
YIELD RIGHT OF WAY TO PEDESTRIANS	657	9
ENTERING HWY FROM PRIVATE ROAD OR DRIVEWAY	510	1
ENTRANCE FROM STOP THROUGH HIGHWAYYIELD UNTIL REASONABLY SAFE	483	0
STARTING PARKED VEHICLES OR BACKING	340	1

By combining the hours that a collision happened with the number of injured and killed, another finding is that the collisions happened around 18:00, 19:00, 1:00, 23:00, and 7:00 cause more death compared to other time interval. However, the most injuries are happened around afternoon from 14:00 to 18:00.



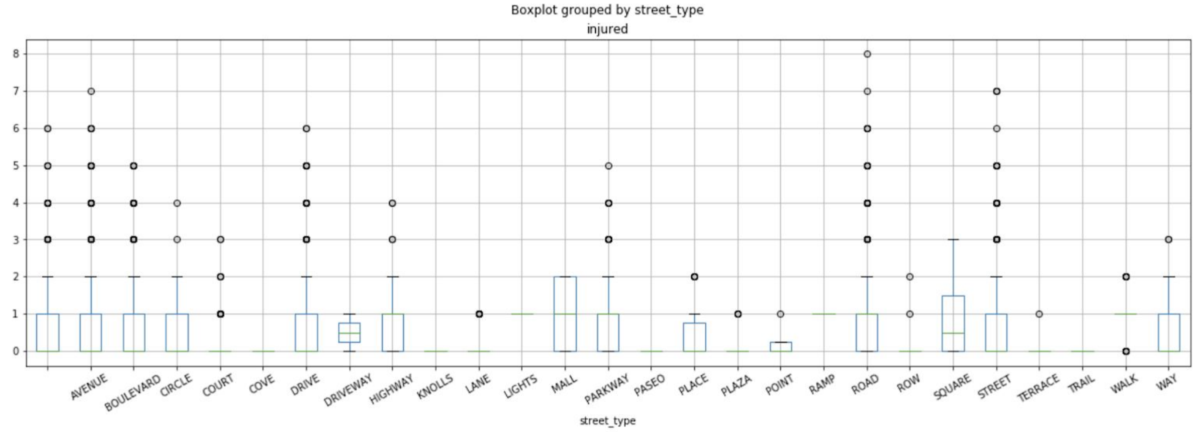
hour	injured	killed
17	1119	6
14	1043	5
15	977	3
18	950	12
16	919	5



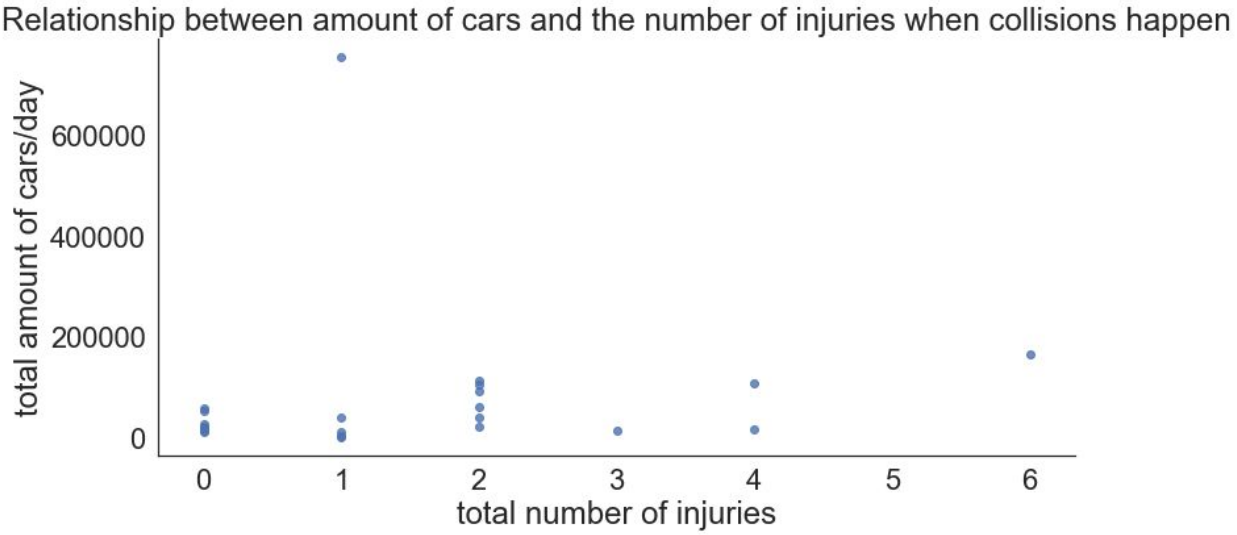
hour	injured	killed
18	950	12
19	641	9
13	791	9
23	318	7
7	659	7

Boxplot grouped by street_type and injured

This boxplot describes the number of injures with respect to the type of street. Generally looking, more injuries are happened at square; however, outliers exist.



The following scatterplot is an illustration of how the number of injuries when a collision happens relates to the total number of vehicles passed through that street on that day in a 24-hour time frame.



Death in Collision between Holiday and non-Holiday

Question: Statistical testing to see if there are more deaths involved in a collision happened on U.S. holiday than a normal day.

Hypothesis testing 1

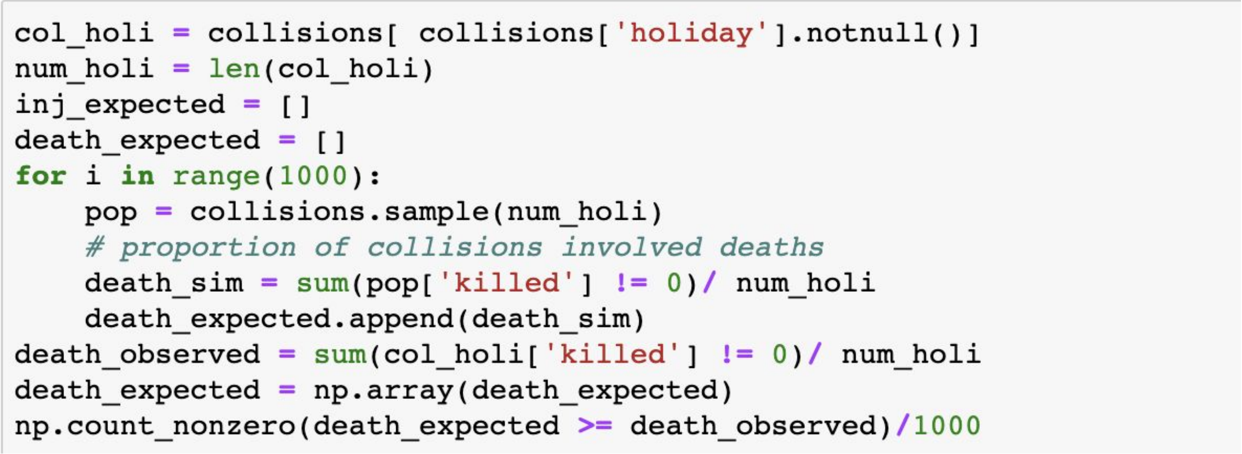
In the first hypothesis testing, we want to perform a test to examine the probability difference between collisions cause deaths on holiday and in a normal day.

In order to find the testing result, we extract a subtable that include all the collisions happened on holiday. We calculate the number of collisions that happened on holiday as `num_holi` and the probability of death occurring in that subtable as our observed statistic (`death_observed`)

Step 1:
Null Hypothesis: There is no difference in the probability of deaths occuring in a traffic collision on a U.S. holiday than a normal day.

Step 2:
Alternative Hypothesis: The probability of deaths occuring in a traffic collision on a U.S. holiday is higher than a normal day.

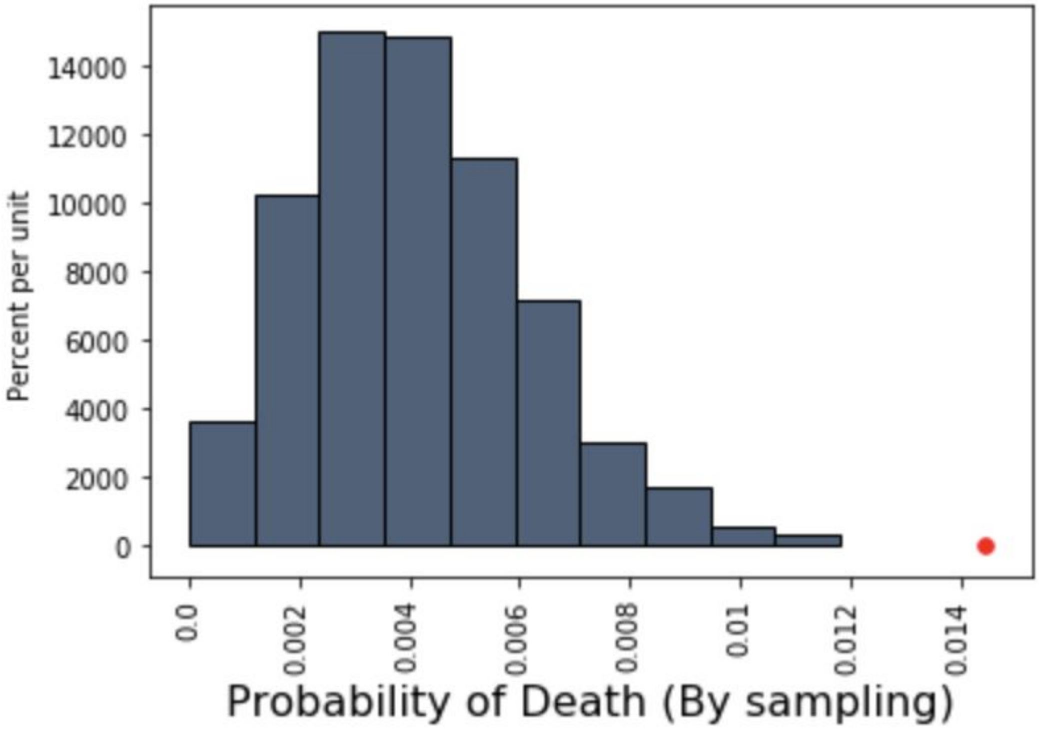
Step 3:
Sampling:
In order to test the hypothesis, we perform sampling to the pd_collisions_datasd.csv data by selecting a subtable which has `num_holi` items. We calculate the probability of death occurring in this sampled subtable as our expected statistic (`death_sim`). Repeat that process 1000 times, append each simulation value to our array (`death_expected`).



```
col_holi = collisions[ collisions['holiday'].notnull()]
num_holi = len(col_holi)
inj_expected = []
death_expected = []
for i in range(1000):
    pop = collisions.sample(num_holi)
    # proportion of collisions involved deaths
    death_sim = sum(pop['killed'] != 0)/ num_holi
    death_expected.append(death_sim)
death_observed = sum(col_holi['killed'] != 0)/ num_holi
death_expected = np.array(death_expected)
np.count_nonzero(death_expected >= death_observed)/1000
```

Step 4:
Level of Significance: 5%

Significant test:
In order to test our null hypothesis, we calculate the probability that the expected value (elements in `death_expected`) is greater than the observed value (`death_observed`). Since the p-value (around 0.00) is less than our significant value, we reject our null hypothesis.



Conclusion:
Based on the testing result, we reject the null hypothesis and conclude that the probability of deaths occuring in a traffic collision on a U.S. holiday is higher than a normal day.

Hypothesis testing 2

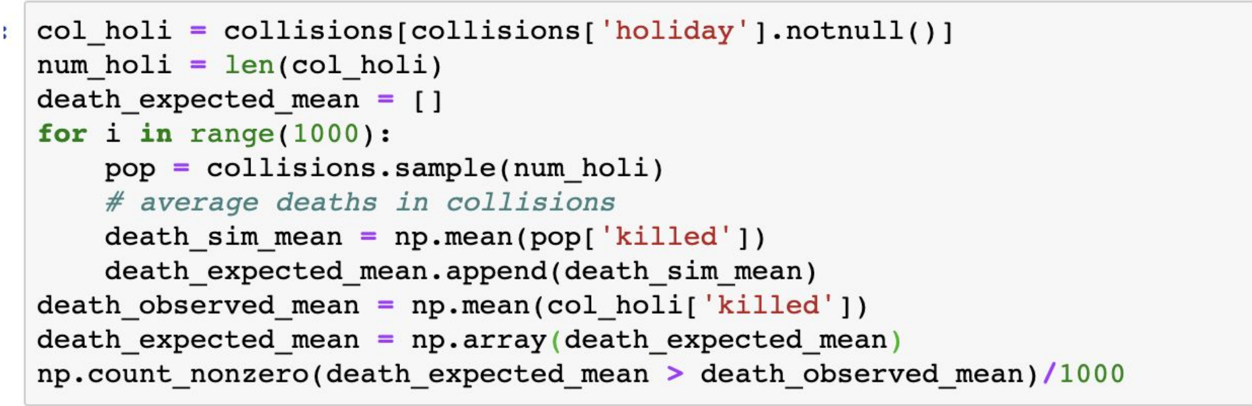
In the second hypothesis testing, we want to perform a test to examine the difference between the average number of deaths on holiday and in a normal day.

In order to find the testing result, we extract a subtable that include all the collisions happened on holiday. We calculate the number of collisions that happened on holiday as `num_holi` and the average number of deaths occurring on holiday as our observed statistic (`death_observed_mean`)

Step 1:
Null Hypothesis: There is no difference in the average number of deaths occuring in a traffic collision on a U.S. holiday than a normal day.

Step 2:
Alternative Hypothesis: The average number of deaths occurring in a traffic collision on a U.S. holiday is higher than a normal day.

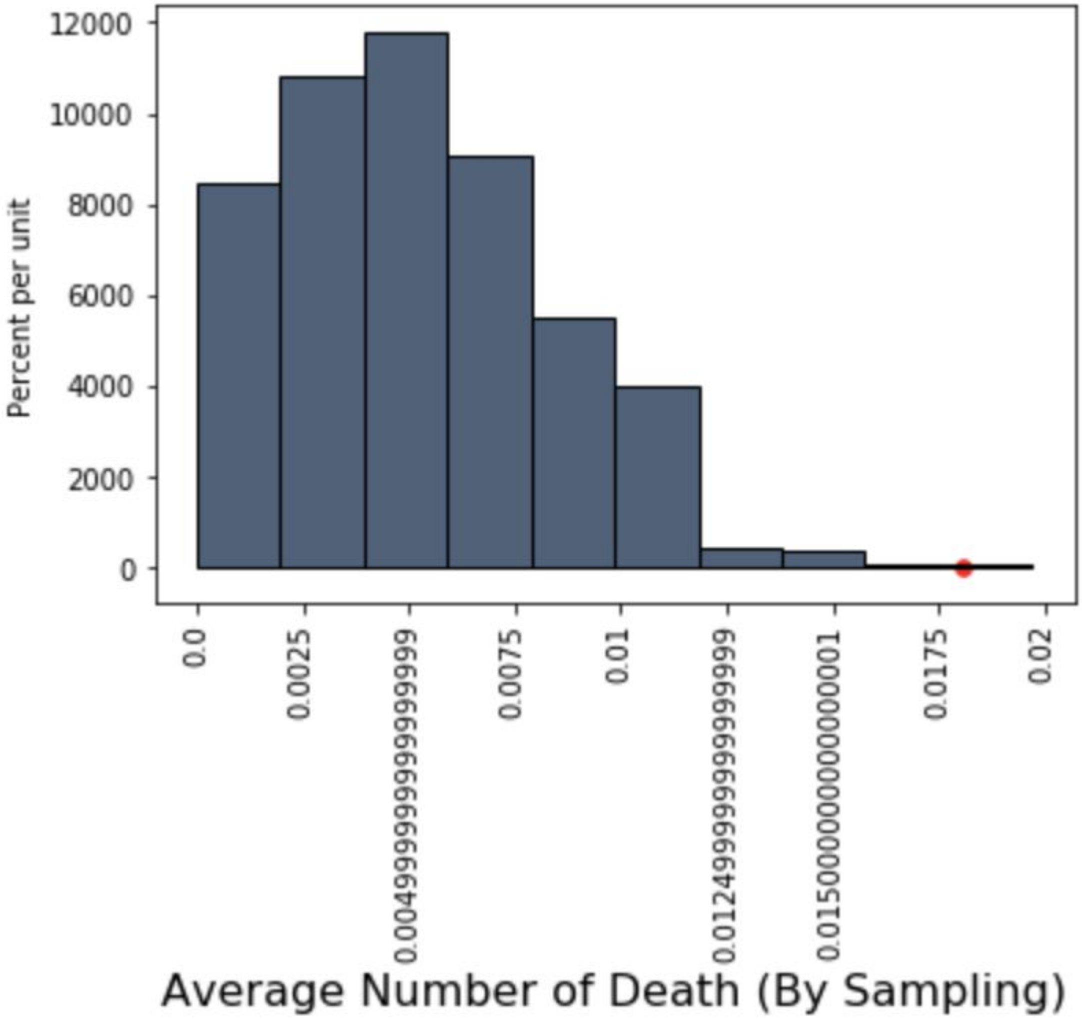
Step 3:
Sampling:
In order to test the hypothesis, we perform sampling to the pd_collisions_datasd.csv data by selecting a subtable which has `num_holi` items. We calculate the average number of death occurring in this sampled subtable as our expected statistic (`death_sim_mean`). Repeat that process 1000 times, append each simulation value to our array (`death_expected_mean`).



```python
col_holi = collisions[collisions['holiday'].notnull()]
num_holi = len(col_holi)
death_expected_mean = []
for i in range(1000):
    pop = collisions.sample(num_holi)
    # average deaths in collisions
    death_sim_mean = np.mean(pop['killed'])
    death_expected_mean.append(death_sim_mean)
death_observed_mean = np.mean(col_holi['killed'])
death_expected_mean = np.array(death_expected_mean)
np.count_nonzero(death_expected_mean > death_observed_mean)/1000
```

Step 4:
Level of Significance: 5%

Significant test:
In order to test our null hypothesis, we calculate the probability that the expected value (elements in `death_expected_mean)` is greater than the observed value (`death_observed_mean`). Since the p-value (around 0.001) is less than our significant value, we reject our null hypothesis.



Conclusion:
Based on the testing result, we reject the null hypothesis and conclude that the average number of deaths occuring in a traffic collision on a U.S. holiday is higher than a normal day.

Hypothesis testing 3

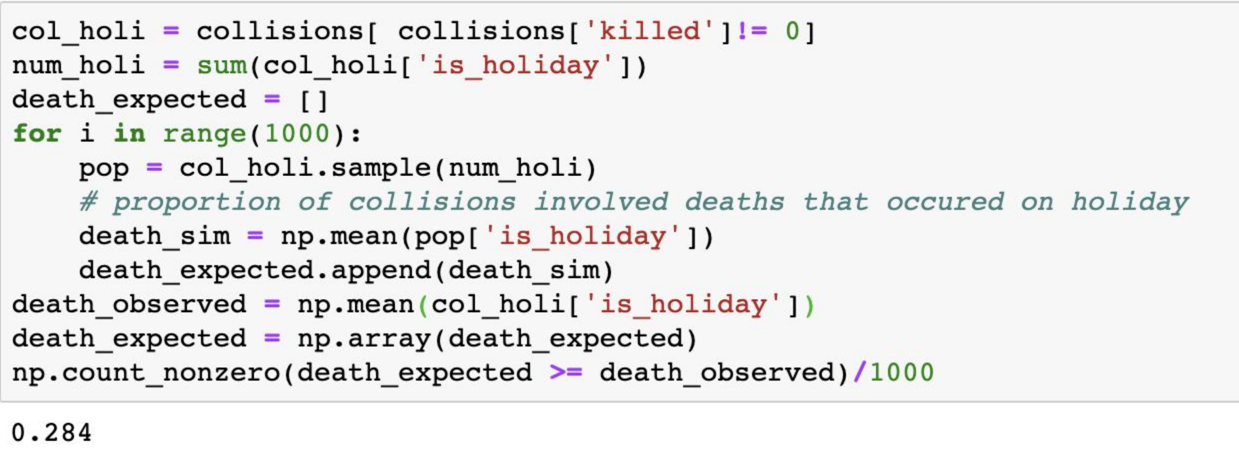
In the third hypothesis testing, we want to pursuit the test from a different perspective. Given a collision that causes deaths, we want to compare the probability of the collision happening on holiday.

In order to find the testing result, we extract a subtable that include all the collisions that have caused deaths. We calculate the number of collisions that happened on holiday as `num_holi` and the proportion that the collision happened on holiday as our observed statistic (`death_observed`)

Step 1:
Null Hypothesis: Given the collision causes deaths, there is no difference between the probability of happening on holiday or in a normal day.

Step 2:
Alternative Hypothesis: Given the collision causes deaths, the probability that the collision happens on holiday is higher than that in a normal day.

Step 3:
Sampling:
In order to test the hypothesis, we perform sampling to the `col_holi` table (collisions that cause deaths) by selecting a subtable which has `num_holi` items. We calculate the probability of the collision happening on holiday as our expected statistic (`death_sim`). Repeat that process 1000 times, append each simulation value to our array (`death_expected`).



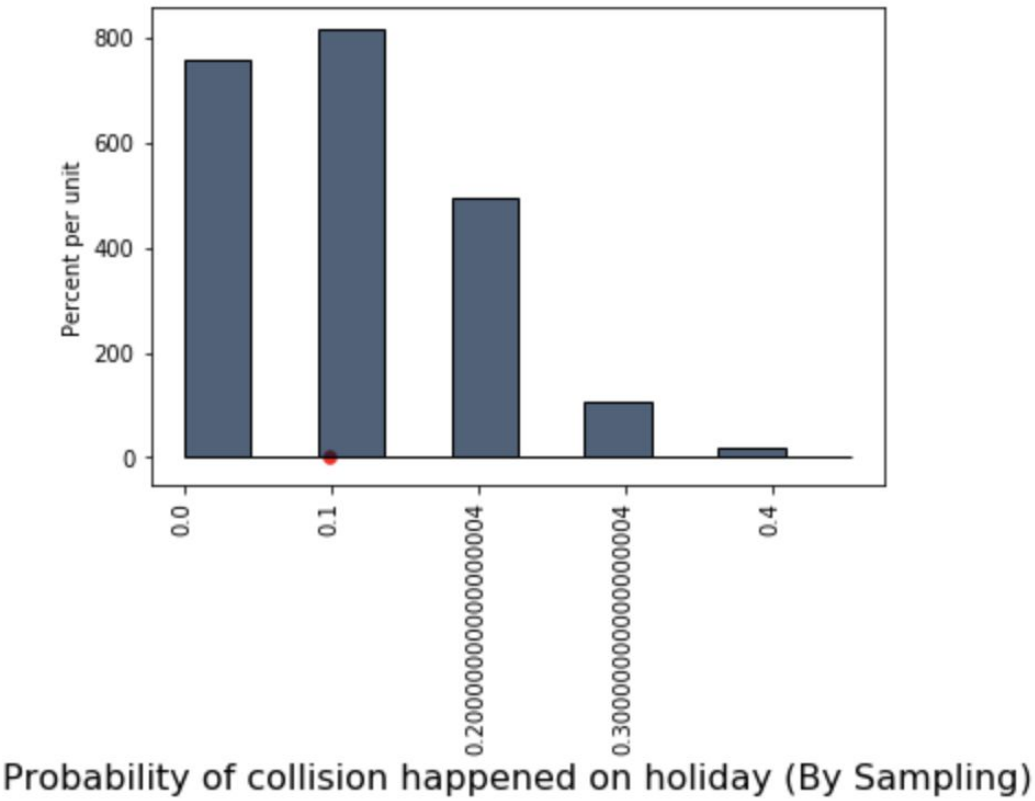
```
col_holi = collisions[ collisions['killed']!= 0]
num_holi = sum(col_holi['is_holiday'])
death_expected = []
for i in range(1000):
    pop = col_holi.sample(num_holi)
    # proportion of collisions involved deaths that occured on holiday
    death_sim = np.mean(pop['is_holiday'])
    death_expected.append(death_sim)
death_observed = np.mean(col_holi['is_holiday'])
death_expected = np.array(death_expected)
np.count_nonzero(death_expected >= death_observed)/1000
```

```
0.284
```

Step 4:
Level of Significance: 5%

Significant test:
In order to test our null hypothesis, we calculate the probability that the expected value (elements in `death_expected`) is greater than the observed value (`death_observed`). Since the p-value (around 0.25-0.30) is greater than our significant value, we cannot reject our null

hypothesis.

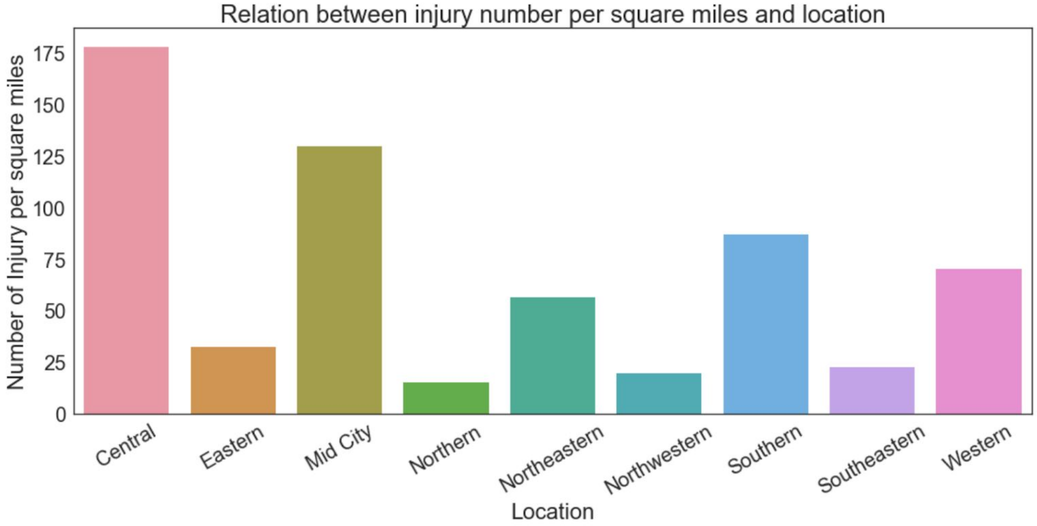


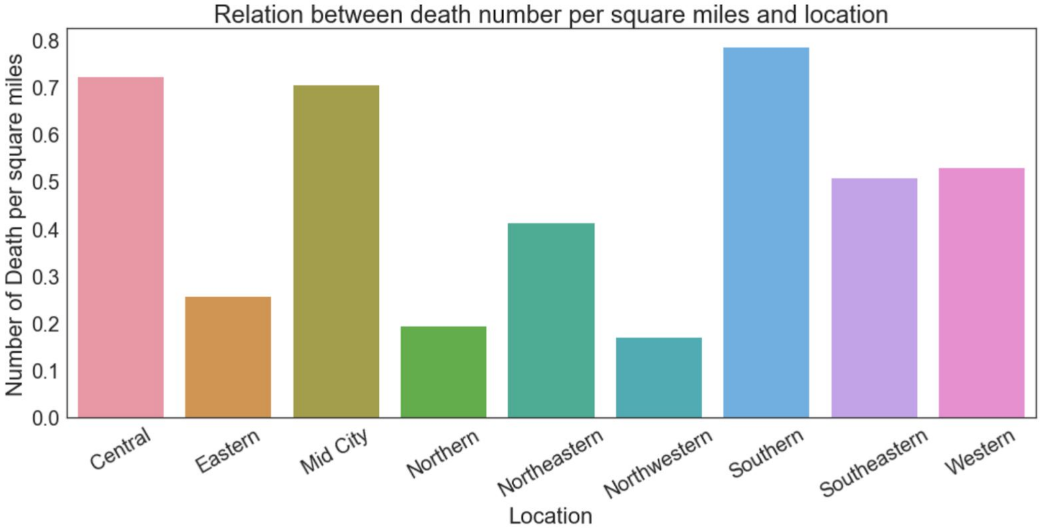
Conclusion:
Based on the testing result, we cannot reject the null hypothesis and conclude that given the collision causes deaths, the probability of that collision happened on holiday and the probability of that collision happened in a normal day does not have a significant difference.

Relationships between the injured/death number per square miles and the location

We set out to explore the relationships between the injured/death number per square miles and the location. We mainly investigated the data for the number of injury/death and the police beat. After searching on the official website, we decided to divide the police beat into 9 different divisions based on their locations according to San Diego Police Department. By taking the first digit of each police beat, we were able to categorize each police beat into different divisions. After looking up the area of each division, we were able to normalize the total injury number. We decided to present our findings on a bar chart to show how the average injured/death number correlates to the location.

From the bar chart, we can see that Central, Mid City, and Southern has the highest average injury/death number, where Central division has the highest injury number, and Southern division has the highest death number. This finding is reasonable since these three divisions are the closest to the downtown area of San Diego, where traffic is busy.





WORKS CITED

Collisions: https://data.sandiego.gov/datasets/police-collisions/

Traffic Volumes: https://data.sandiego.gov/datasets/traffic-volumes/

San Diego Police Department data: https://www.sandiego.gov/police/services/divisions