Car Accident Severity vs Road Visibility

Group A6

Yixin Ying
Basava Satish Chandra Velagapudi
Sarah Akiki
Eason Tsao
Elene Chikhradze
Sherry Shi
Yongyu Qiu

The problem We Are Seeking to Answer

Historically:

- There tends to be an increase in car accidents, when road visibility is poor
- Studies have explored the impact of weather on road safety

The problem at hand:

- We are investigating the potential correlation between the severity of car accidents and road visibility conditions in the United States

Data Driven approach

- Objective, quantitative evidence that can be used to inform decisions, policies, and interventions
- Uncover patterns, trends, and correlations that might not be visible through theoretical reasoning alone

Possible Outcomes and The Ideal Experiment

Possible outcomes:

- Outcome 1: Fail to reject the null hypothesis
- Outcome 2: reject the null hypothesis

The Ideal Experiment:

- Chi square test of independence

Hypothesis:

- H0: No relationship between the severity of accident and visibility
- Ha: Apparent relationship between the severity of accident and visibility

Data Collection:

- A dataset that collects data consistently over time to minimize seasonal or temporal effects

Confounding Variables:

- Zip code, time zone, humidity, wind speed, and etc.
- Address through regression or stratification

Steps To Perform The Test

Data Preparation:

- Discretize visibility (e.g., Low: 0-1 miles, Excellent: >10 miles).
- Contingency table creation.

Hypothesis Formulation:

- Null: No association.
- Alternative: Significant association.

Chi-Square Test:

- Chi-Square Statistic formula.
- Degrees of freedom: (r-1)(c-1).
- P-value criteria (e.g., $\alpha = 0.05$).

Interpretation:

- Significant (or not) association based on p-value.

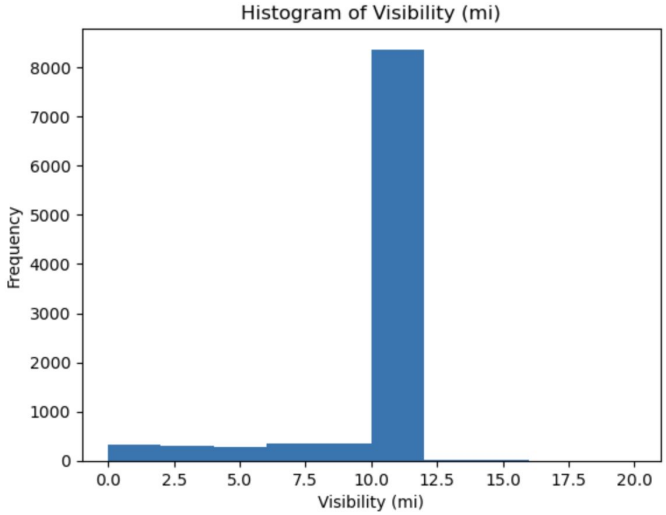
Reporting:

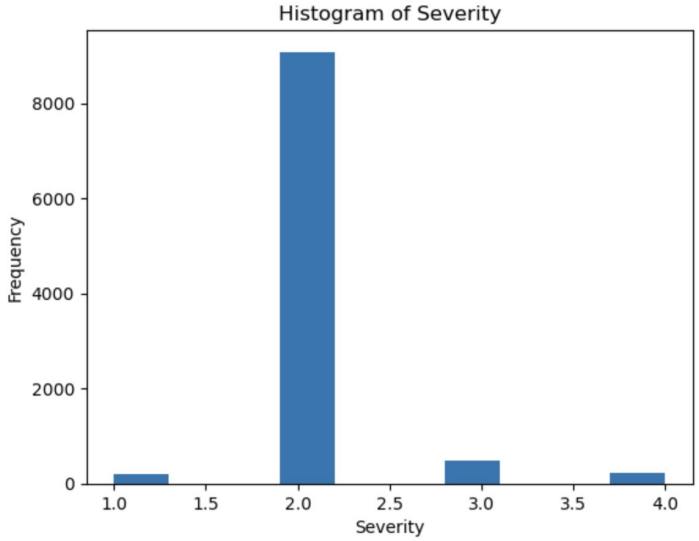
- Test statistic, p-value, conclusion, effect size

Descriptive Statistics and Visualization

```
sampled_df = df.sample(n=10000, random_state=42)
sampled_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10000 entries, 202953 to 144659
Data columns (total 44 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     10000 non-null  object
 1   Source                 10000 non-null  object
 2   Severity               10000 non-null  int64
 3   Start_Time             10000 non-null  object
 4   End_Time               10000 non-null  object
 5   Start_Lat              10000 non-null  float64
 6   Start_Lng              10000 non-null  float64
 7   Distance(mi)           10000 non-null  float64
 8   Description            10000 non-null  object
 9   Street                 9970 non-null   object
 10  City                   9999 non-null   object
 11  County                 10000 non-null  object
 12  State                  10000 non-null  object
 13  Zipcode                10000 non-null  object
 14  Country                10000 non-null  object
 15  Timezone               10000 non-null  object
 16  Airport_Code           10000 non-null  object
 17  Weather_Timestamp      10000 non-null  object
 18  Temperature(F)         9963 non-null   float64
 19  Wind_Chill(F)          9884 non-null   float64
 20  Humidity(%)            9946 non-null   float64
 21  Pressure(in)           9985 non-null   float64
 22  Visibility(mi)         10000 non-null  float64
 23  Wind_Direction         9918 non-null   object
 24  Wind_Speed(mph)        9918 non-null   float64
 25  Precipitation(in)      9803 non-null   float64
 26  Weather_Condition      9981 non-null   object
 27  Amenity                10000 non-null  bool
 28  Bump                   10000 non-null  bool
 29  Crossing               10000 non-null  bool
 30  Give_Way               10000 non-null  bool
 31  Junction               10000 non-null  bool
 32  No_Exit                10000 non-null  bool
 33  Railway                10000 non-null  bool
 34  Roundabout             10000 non-null  bool
 35  Station                10000 non-null  bool
 36  Stop                   10000 non-null  bool
 37  Traffic_Calming        10000 non-null  bool
 38  Traffic_Signal         10000 non-null  bool
 39  Turning_Loop           10000 non-null  bool
 40  Sunrise_Sunset         9909 non-null   object
 41  Civil_Twilight         9909 non-null   object
 42  Nautical_Twilight      9909 non-null   object
 43  Astronomical_Twilight  9909 non-null   object
dtypes: bool(13), float64(10), int64(1), object(20)
memory usage: 2.6+ MB
```





	Severity	Visibility(mi)
0	3	10.0
1	2	10.0
2	2	10.0
3	3	9.0
4	2	6.0
5	3	7.0
6	2	7.0
7	3	7.0
8	2	5.0
9	3	3.0

	Severity	Visibility(mi)
count	7.551296e+06	7.551296e+06
mean	2.211485e+00	9.090376e+00
std	4.864240e-01	2.688316e+00
min	1.000000e+00	0.000000e+00
25%	2.000000e+00	1.000000e+01
50%	2.000000e+00	1.000000e+01
75%	2.000000e+00	1.000000e+01
max	4.000000e+00	1.400000e+02

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7551296 entries, 0 to 7728393
Data columns (total 2 columns):
 #   Column          Dtype
---  ------          -----
 0   Severity        int64
 1   Visibility(mi)  float64
dtypes: float64(1), int64(1)
memory usage: 172.8 MB
```

Data Preparation

Discretize visibility i.e, Categorize ‘visibility’ into ‘visibility category’ based on the following rules:

Low Visibility: Less than 1 mile (0-1 miles)
Moderate Visibility: 1 to 3 miles (1-3 miles)
Good Visibility: 3 to 10 miles (3-10 miles)
Excellent Visibility: More than 10 miles (> 10 miles)

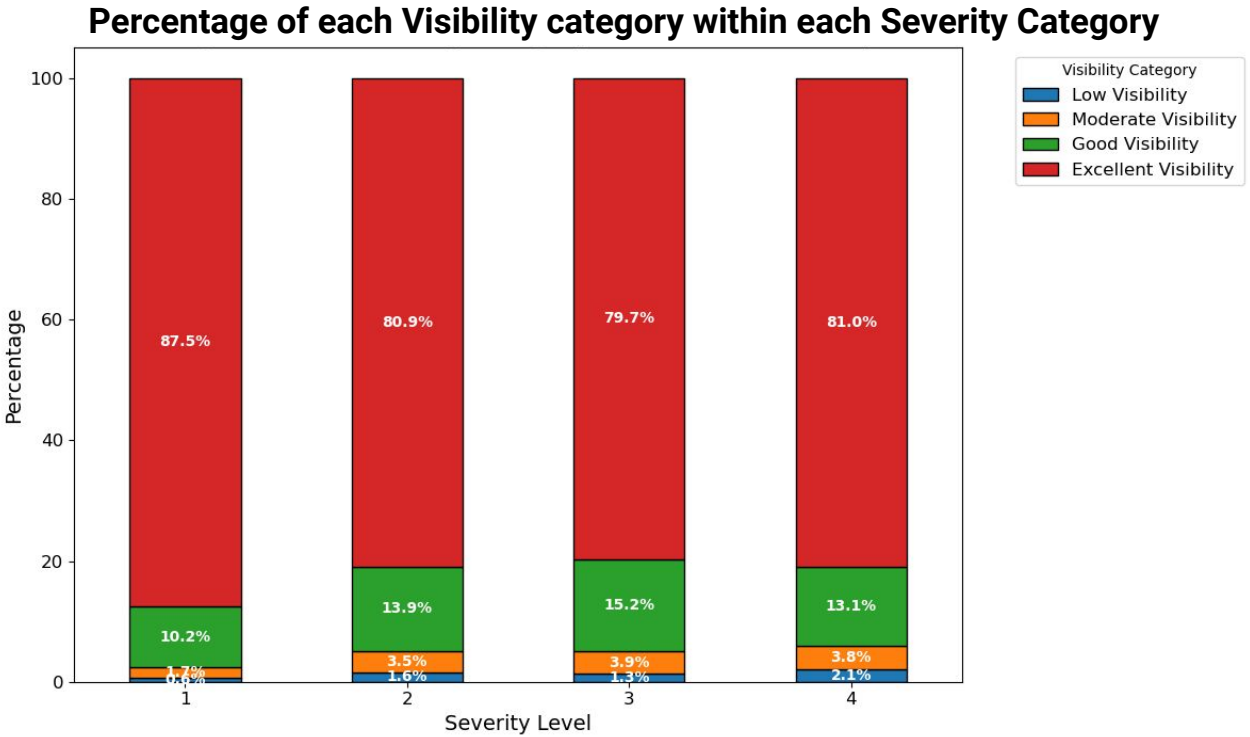
	Severity	Visibility(mi)	Visibility Category
0	3	10.0	Excellent Visibility
1	2	10.0	Excellent Visibility
2	2	10.0	Excellent Visibility
3	3	9.0	Good Visibility
4	2	6.0	Good Visibility

Contingency Table:

Severity	Excellent Visibility	Good Visibility	Moderate Visibility	Low Visibility	Total by Severity
1	58113	6755	1145	418	66431.0
2	4871883	838746	211998	96117	6018744.0
3	1011275	192280	48877	16394	1268826.0
4	159805	25821	7467	4202	197295.0
Total by Visibility	6101076	1063602	269487	117131	NaN

Null Hypothesis H0: No significant association between Road Visibility and Severity of an Accident
Alternative Hypothesis H1: Significant association between Road Visibility and Severity of an Accident

Visualizations:



Chi-Square Test:

```
from scipy.stats import chi2_contingency
# Perform the chi-square test of independence
chi2, p, dof, expected = chi2_contingency(contingency_table)

chi2, p, dof, expected
```

```
(4962.777704613812,
 0.0,
 9,
 array([[1.03043629e+03, 2.37075740e+03, 9.35682358e+03, 5.36729827e+04],
        [9.33590080e+04, 2.14794025e+05, 8.47741653e+05, 4.86284931e+06],
        [1.96812386e+04, 4.52812487e+04, 1.78714471e+05, 1.02514904e+06],
        [3.06031715e+03, 7.04096855e+03, 2.77890519e+04, 1.59404662e+05]]))
```

Chi-Square statistic: 4962.78; p-value = 0; degrees of freedom: 9

The very small p-value (0.0) suggests that there is a statistically significant association between the "Severity of Accident" and "Visibility of Road" categories.

This means that the distribution of accident severities is dependent of the visibility conditions

Limitations

- Variable is categorical - hard to measure
- Chi square tests association, excludes direction of association
- Further analysis - regression model
- Confounding variables
  - Road Conditions
  - Time of Day
  - Weather Conditions
  - Environmental Features
  - Driver Behavior

Future Expectations

- Logistic regression to address ordinal data and confounding variables
- Improved understanding for risk assessment
- May alter road signs or speed control if finding is significant
- Greater awareness about road conditions
- Reduction in accident severity


Thank You!