Documentation - exercise 1
dataset : US Accident

Professor : Dr. Kheradpisheh

Teacher Assistant : MohammadReza Khanmohammadi

By: Katayoun Kobraei

1. Introduction to Dataset

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including multiple APIs that provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about **1.5 million** accident records in this dataset. Check the below descriptions for more detailed information.

US Accident has this features:

1. **ID:** This is a unique identifier of the accident record.
2. **Severity:** Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
3. **Start_Time:** Shows start time of the accident in the local time zone.
4. **End_Time:** Shows end time of the accident in the local time zone. End time here refers to when the impact of an accident on traffic flow was dismissed.
5. **Start_Lat:** Shows latitude in GPS coordinate of the start point.
6. **Start_Lng:** Shows longitude in GPS coordinate of the start point.
7. **End_Lat:** Shows latitude in GPS coordinate of the end point.
8. **End_Lng:** Shows longitude in GPS coordinate of the end point.
9. **Distance(mi):**The length of the road extent affected by the accident.
10. **Distance(mi):** Shows natural language description of the accident.
11. **Number:** Shows the street number in the address field.

12. **Street:** Shows the street name in the address field.
13. **Side:** Shows the relative side of the street (Right/Left) in the address field.
14. **City:** Shows the city in the address field.
15. **County:** Shows the county in the address field.
16. **State:** Shows the state in address field
17. **Zipcode:** Shows the zip code in the address field.
18. **Country:** Shows the country in the address field.
19. **Timezone:** Shows timezone based on the location of the accident (eastern, central, etc.).
20. **Airport_Code:** Denotes an airport-based weather station which is the closest one to location of the
21. **Weather_Timestamp:** Shows the time-stamp of a weather observation record (in local time).
22. **Temperature(F):** Shows the temperature (in Fahrenheit).
23. **Wind_Chill(F):** Shows the wind chill (in Fahrenheit).
24. **Humidity(%):** Shows the humidity (in percentage).
25. **Pressure(in):** Shows the air pressure (in inches).
26. **Visibility(mi):** Shows visibility (in miles).
27. **Wind_Direction:** Shows wind direction.
28. **Wind_Speed(mph):** Shows wind speed (in miles per hour).
29. **Precipitation(in):** Shows precipitation amount in inches, if there is any.
30. **Weather_Condition:** Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
31. **Amenity:** A POI annotation which indicates presence of amenity in a nearby location.
32. **Bump:** A POI annotation which indicates presence of speed bump or hump in a nearby location.
33. **Crossing:** A POI annotation which indicates presence of crossing in a nearby location.
34. **Give_Way:** A POI annotation which indicates presence of give way in a nearby location.
35. **Junction:** A POI annotation which indicates presence of a junction in a nearby location.

36. **No_Exit:** A POI annotation which indicates presence of no exit in a nearby location.
37. **Railway:** A POI annotation which indicates the presence of a railway in a nearby location.
38. **Roundabout:** A POI annotation which indicates the presence of a roundabout in a nearby location.
39. **Station:** A POI annotation which indicates the presence of a station in a nearby location.
40. **Stop:** A POI annotation which indicates presence of stop in a nearby location.
41. **Traffic_Calming:** A POI annotation which indicates presence of traffic calming in a nearby location.
42. **Traffic_Signal:** A POI annotation which indicates presence of traffic signal in a nearby location.
43. **Turning_Loop:** A POI annotation which indicates the presence of a turning loop in a nearby location.
44. **Sunrise_Sunset:** Shows the period of day (i.e. day or night) based on sunrise/sunset.
45. **Civil_Twilight:** Shows the period of day (i.e. day or night) based on civil twilight.
46. **Nautical_Twilight:** Shows the period of day (i.e. day or night) based on nautical twilight.
47. **Nautical_Twilight:** Astronomical_Twilight Shows the period of day (i.e. day or night) based on astronomical twilight.

It can be seen that some of these are useless. Therefore, we decided to drop following features:

"**Source**": This column might not be needed for analysis, especially if we are not interested in the source of the data.
"**End_Time**": At the moment we are primarily interested in the start time of accidents and do not need to calculate the duration of accidents.
"**End_Lat**" , **'End_Lng'**: These may not be necessary if we are only concerned with the starting location of accidents.
'**Distance(mi)**': Depending on our analysis goals, this column might not be relevant. For instance, if we focus on the location and conditions of accidents, the distance traveled may not be essential.

'Description': Textual descriptions can be valuable for qualitative analysis but may not be required for quantitative analyses.
'Zipcode', **'Country'**, **'Timezone'**, **'Airport_Code'**: These location-related columns may be redundant. Because we already have 'State' and 'City' information in our features.
'**Weather_Timestamp**': As we are not performing any time-series analysis of weather conditions right now, we do not need this column.
'Turning_Loop': This column has constant values (e.g., always "False"). As a consequence it does not provide useful information.
'**Civil_Twilight**', '**Nautical_Twilight**', '**Astronomical_Twilight**': As they relate to daylight conditions, we do not need them.

2. Abstract

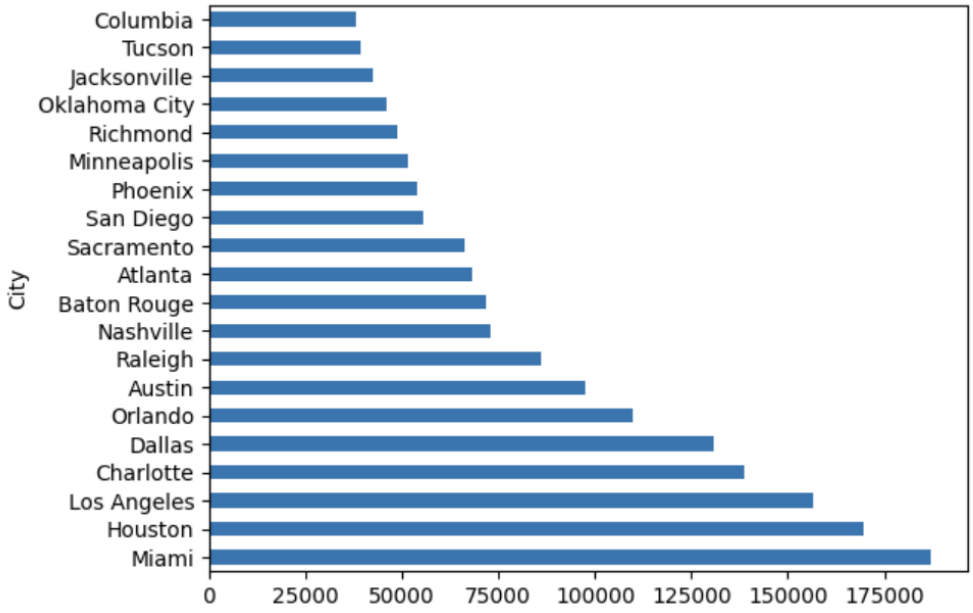
In this dataset we perform preprocessing methods to explore and find hidden relations between features and data. First we define useful and important features to work with. Then we try to find Attributes with null-values to impute them. Next step, we plot data and explore features to find relations. Finally, we claim five hypotheses to see hidden relations between features. We use t-test, u-test, Chi-squared test, regression analysis to get these results.

3. EDA (Exploratory data analysis

1. Cities with highest accident rates

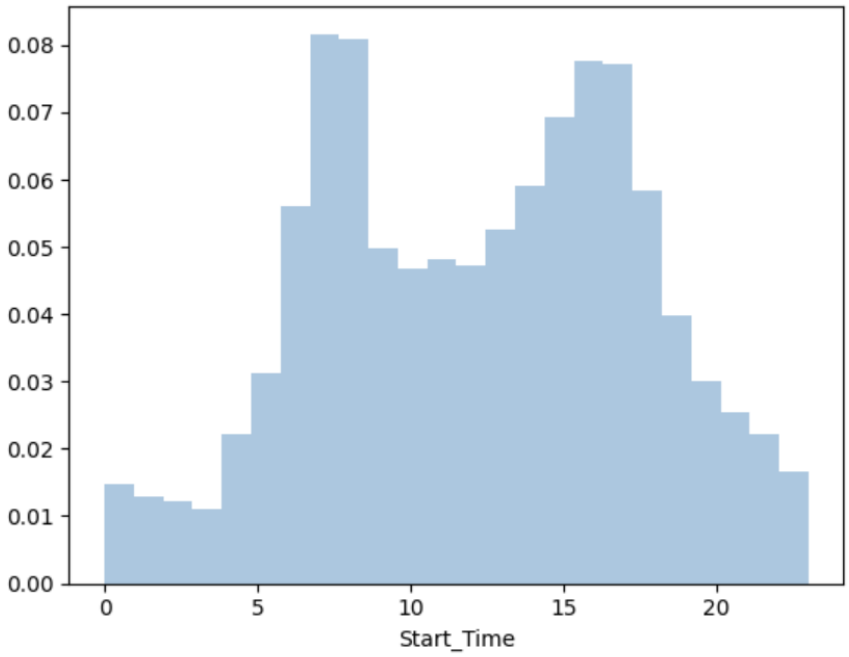
These 20 cities have the highest accident rate in US:

1. Miami 187170
2. Houston 169609
3. Los Angeles 156491
4. Charlotte 138652
5. Dallas 130939
6. Orlando 109733
7. Austin 97359
8. Raleigh 86079
9. Nashville 72930
10. Baton Rouge 71588
11. Atlanta 68186
12. Sacramento 66264
13. San Diego 55504
14. Phoenix 53974
15. Minneapolis 51488
16. Richmond 48845
17. Oklahoma City 46092
18. Jacksonville 42447
19. Tucson 39304
20. Columbia 38178

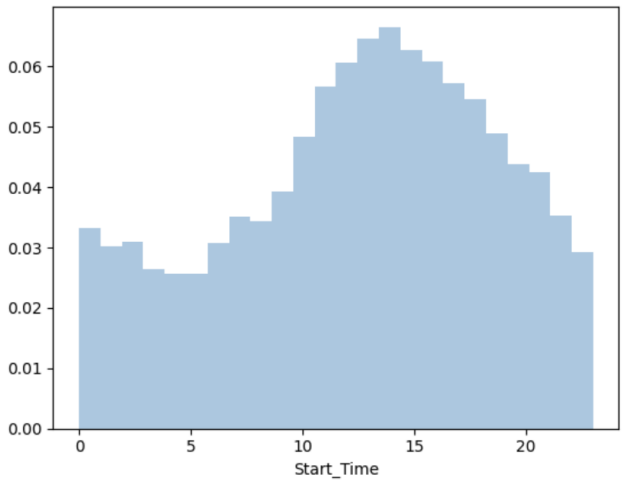


2. **What time of the day do the most accidents happen?**

   First we convert our data to dat-time to find these pick hours:

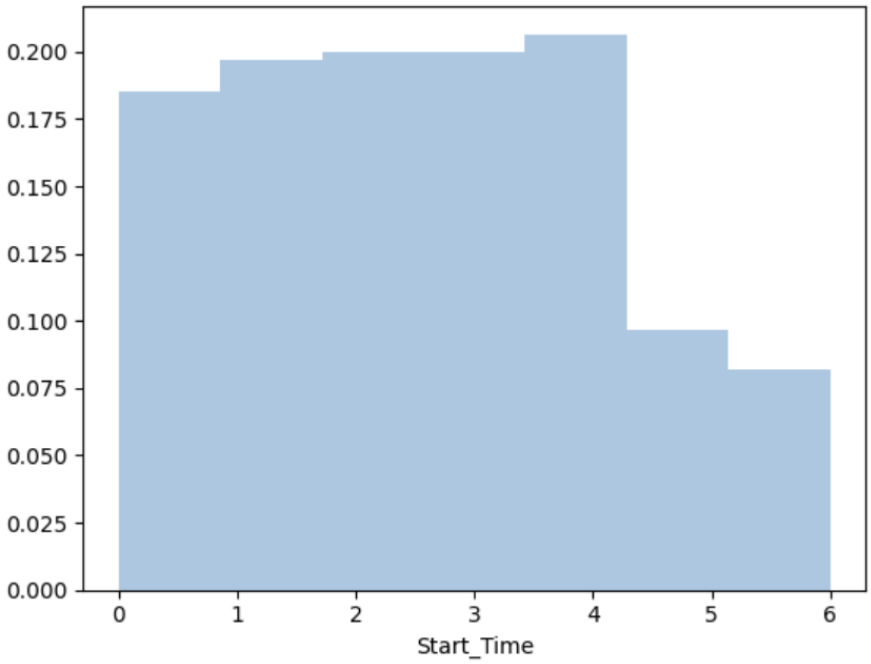


As we can see a high percentage of accidents occur between 6 am to almost 9 am (probably people in a hurry to get to work) Next highest percentage is 1 pm to 6 pm.
Now we check if weekend has the same peak hour as weekdays.



As we can see on Sundays, the peak occurs between 11 am and 5 pm, unlike weekdays

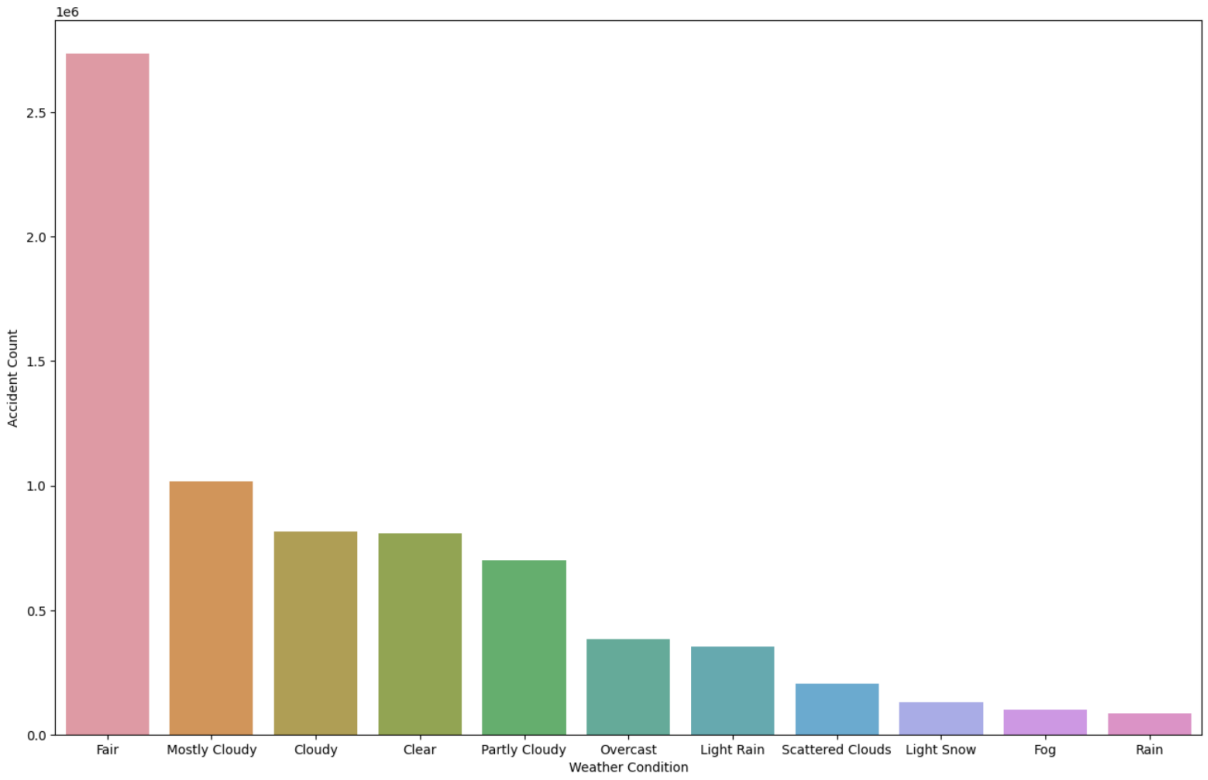
3. **Which day of the week do the accidents happen?**



Surprisingly, weekdays have a higher accident rate compared to weekends.

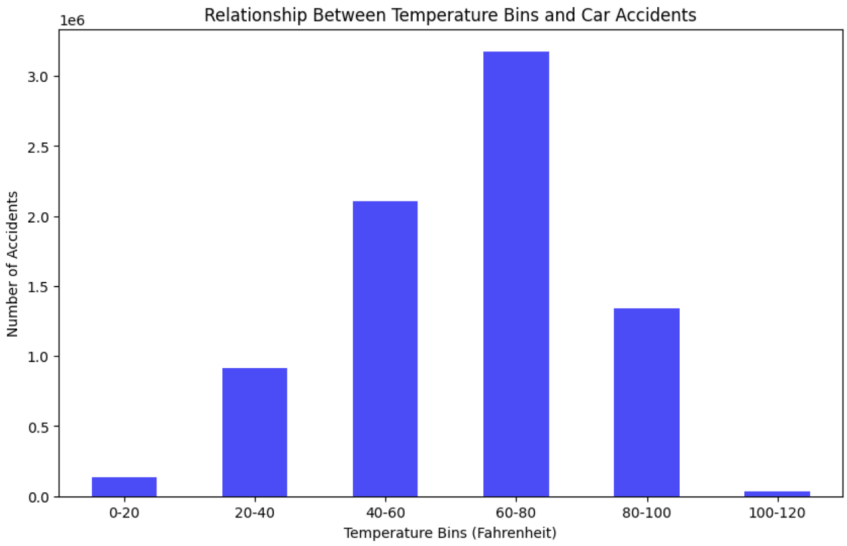
4. Which weather has the highest accident rate?

4.1. Weather condition



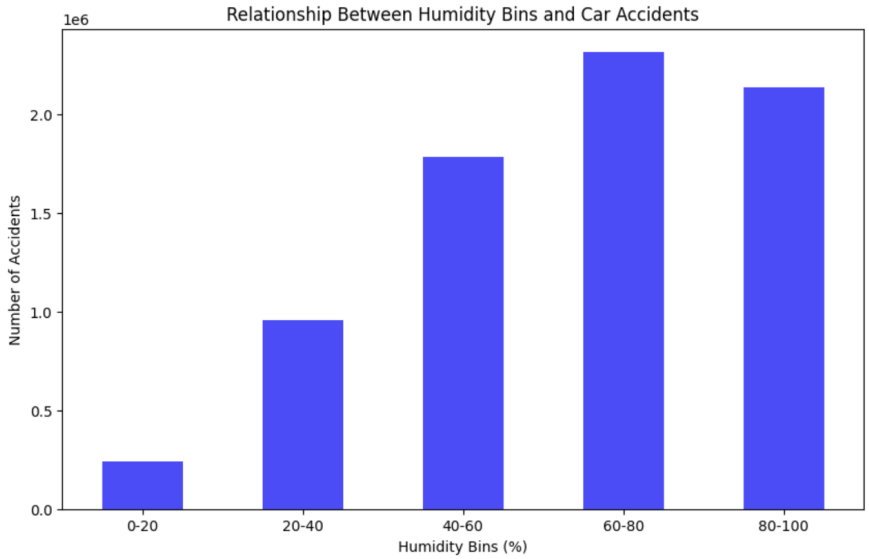
There is a high chance of accidents in fair weather. Surprisingly we have much less accidents in rainy weather. Maybe people become a careful driver in rainy days and this prevent accidents.

4.2. Weather temperature



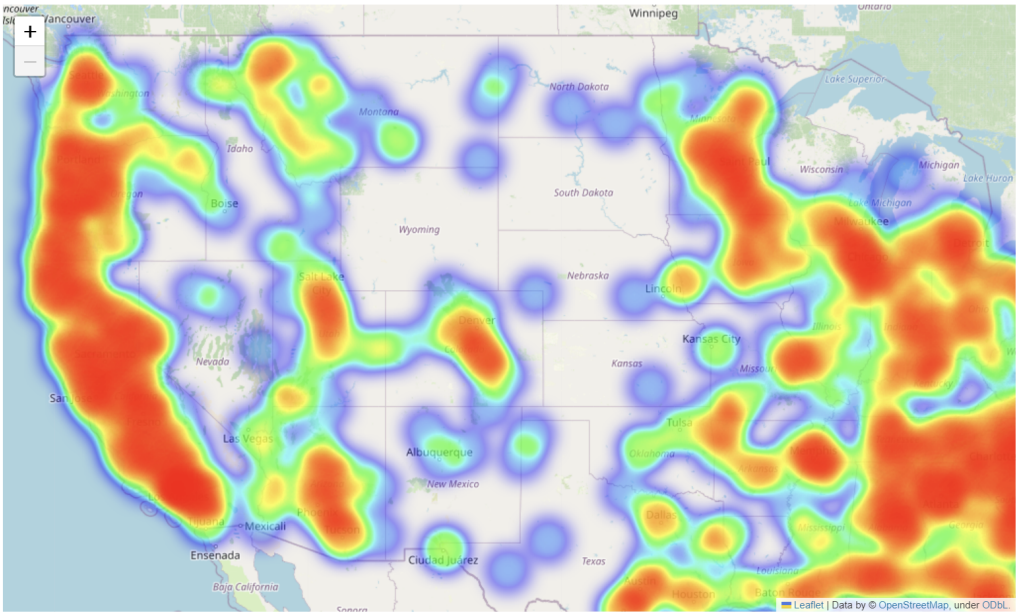
We see that most accidents happen in 60-80 Fahrenheit temperatures. It is a regular temperature so there are more significant features affect rate of accidents.

4.3. Weather humidity



We can see we have more accidents when there is more humidity in weather.

5. Location analysis



Using folium library we can see exact location on map has a high chance to have accident.

5. Hypothesis tests

5.1. Temperature and Accident Severity

Null Hypothesis (H0): There is no significant difference in accident severity between hot (e.g., temperature > 90°F) and cold (e.g., temperature < 32°F) weather.

Alternative Hypothesis (H1): There is a significant difference in accident severity between hot and cold weather.

```
Independent Samples t-test: t-statistic = -51.44560494146428, p-value = 0.0
Mann-Whitney U Test: U-statistic = 68678613960.5, p-value = 0.0
Reject the null hypothesis. There is a significant difference in accident severity between hot and cold weather.
```

The results of the t-test and Mann-Whitney U test indicate a significant difference in accident severity between hot and cold weather conditions:

Independent Samples t-test:

t-statistic: -51.44560494146428
p-value: 0.0
The t-statistic essentially quantifies how far off the average accident severity in our sample is from the population's average. When it's negative, like in this instance, it indicates that the average accident severity in hot weather is significantly lower than in cold weather. Now, the incredibly low p-value, which is practically zero, is crucial. It tells us that we've got substantial evidence against the null hypothesis. In simpler terms, it means there's a significant contrast in crash severity between hot and cold weather conditions.

Mann-Whitney U Test:

U-statistic: 68678613960.5
p-value: 0.0
The Mann-Whitney U test is a powerful tool for comparing two sets of data that are not necessarily normally distributed. The U-statistic it generates represents the sum of ranks for one sample relative to the other. In this case, the remarkably low p-value of 0.0 from the Mann-Whitney U test is incredibly significant. It tells us there's compelling evidence against the null hypothesis. This means there's a substantial difference in accident severity between hot and cold weather conditions, and it holds true no matter the specific distribution of the data.

Both of these tests offer compelling evidence to reject the null hypothesis, conclusively demonstrating a statistically significant contrast in accident severity between hot and cold weather conditions. The negative t-statistic implies that, on average, accident

severity is lower in hot weather. The exceptionally low p-values further underscore the profound significance of this discrepancy.

5.2. Chi-squared test hypothesis test on the 'Weather_Condition' feature and accident severity

Null Hypothesis (H0): There is no significant association between weather conditions and accident severity. In other words, the distribution of accident severity is the same across different weather conditions.

Alternative Hypothesis (H1): There is a significant association between weather conditions and accident severity. It means that the distribution of accident severity is not the same across different weather conditions, indicating that weather conditions have an impact on accident severity.

```
Chi-squared statistic = 357164.5565118593
P-value = 0.0
Reject the null hypothesis. Weather condition has a significant effect on accident severity.
```

Chi-squared statistic: 357164.5565118593

The Chi-squared statistic serves as a metric for assessing the level of association between the 'Weather_Condition' and 'Severity' variables. In this specific scenario, the substantial value of the Chi-squared statistic signifies a noteworthy disparity between the observed and expected frequencies within the contingency table.

P-value: 0.0

When the p-value is exceedingly low, as observed here (exactly 0.0), it signifies robust evidence against the null hypothesis. This essentially implies that the distribution of accident severity, as observed, significantly deviates from what would be anticipated if there were no connection between weather conditions and accident severity.

Consequently, the findings lead us to reject the null hypothesis (H0). it means that there is a noteworthy association between weather conditions and accident severity. The remarkably low p-value reinforces the notion that the observed differences in accident severity concerning various weather conditions are highly improbable to be the result of random chance. Hence, it can be inferred that weather conditions indeed exert a substantial influence on accident severity, implying that specific weather conditions may result in varying levels of accident severity.

5.3. Wind Speed and Accident Severity

Null Hypothesis (H0): There is no significant difference in accident severity between high wind speed conditions (e.g., wind speed > 20 mph) and low wind speed conditions.

Alternative Hypothesis (H1): There is a significant difference in accident severity between high wind speed and low wind speed conditions.

```
Independent Samples t-test:
t-statistic = 11.414913077497253
P-value = 3.6197489021470266e-30
Mann-Whitney U Test:
U-statistic = 582873494692.5
P-value = 6.724684713723917e-11
Reject the null hypothesis. There is a significant difference in accident severity between high wind speed and low wind speed conditions.
```

The results presented include the outcomes of two distinct statistical tests: the Independent Samples t-test and the Mann-Whitney U Test. Both tests were carried out with the aim of determining whether a noteworthy disparity exists in accident severity between conditions characterized by high wind speed and those with low wind speed.

Independent Samples t-test:

t-statistic = 11.414913077497253*: The t-statistic measures the extent of differentiation between the means of two datasets. A larger t-statistic implies a more substantial dissimilarity between the means.

P-value = 3.6197489021470266e-30*: The p-value signifies the likelihood of the observed distinction arising purely by random chance. A lower p-value signifies stronger evidence against the null hypothesis.

Interpretation: The exceedingly minuscule p-value (close to zero) strongly indicates the presence of compelling evidence against the null hypothesis. In this context, the null hypothesis posits that there is no significant variance in accident severity between high wind speed and low wind speed conditions. Given that the p-value is less than a pre-defined significance level (commonly 0.05), we opt to reject the null hypothesis.

Mann-Whitney U Test:

U-statistic = 582873494692.5*: The Mann-Whitney U statistic gauges the disparities between the distributions of two datasets. A higher U-statistic hints at more pronounced distinctions between these distributions.

P-value = 6.724684713723917e-11*: Similar to the t-test, the p-value in the Mann-Whitney U test gauges the likelihood of the observed variations between the two datasets occurring by mere chance.

Interpretation: The modest p-value (close to zero) derived from the Mann-Whitney U test also lends robust support against the null hypothesis. This outcome underlines the assertion that a meaningful divergence in accident severity does indeed exist between circumstances characterized by high and low wind speeds.
In both these tests, the evidence cohesively points to a conclusive verdict: a significant contrast in accident severity prevails between instances of high wind speed and those of low wind speed. In essence, this implies that accident severity varies appreciably depending on whether high or low wind speed conditions are in effect.

5.4. Temperature and Humidity Interaction

Null Hypothesis (H0): There is no significant interaction effect between temperature and humidity on accident frequency.

Alternative Hypothesis (H1): There is a significant interaction effect between temperature and humidity on accident frequency.

```
Regression Analysis for Temperature as the Dependent Variable:
P-value = 2.0013882436322942e-43
Reject the null hypothesis. There is a significant relationship between Humidity and Temperature.
Regression Analysis for Humidity as the Dependent Variable:
P-value = 2.756033186447389e-44
Reject the null hypothesis. There is a significant relationship between Temperature and Humidity.
```

P-value = 2.0013882436322942e-43
"Dependent Variable" refers to the variable you are trying to predict or explain, which, in this case, is 'Temperature(F)'.
"Independent Variable" refers to the variable(s) used to predict or explain the dependent variable. In this analysis, 'Humidity(%)' is the independent variable.
P-value is a measure of the evidence against a null hypothesis. In this case, the null hypothesis is that there is no relationship between 'Temperature' and 'Humidity'.
The extremely low p-value (close to zero) indicates strong evidence against the null hypothesis.

Since the p-value is less than the chosen significance level (alpha), which is typically 0.05, we reject the null hypothesis.
The conclusion is that there is a significant relationship between 'Humidity' and 'Temperature'. In other words, 'Humidity' has a significant effect on 'Temperature', and the two variables are correlated.
Regression Analysis for Humidity as the Dependent Variable:

P-value = 2.756033186447389e-44
In this analysis, 'Humidity(%)' is the dependent variable, and 'Temperature(F)' is the independent variable.
The p-value is again extremely low, indicating strong evidence against the null hypothesis.
Since the p-value is less than the chosen significance level, we reject the null hypothesis.

Therefore there is a significant relationship between 'Temperature' and 'Humidity' when 'Humidity' is the variable being explained. Both regression analyses result in extremely low p-values, leading to the rejection of the null hypothesis in both cases. This means that there is a strong and statistically significant relationship between 'Temperature' and 'Humidity'. The exact nature of this relationship (positive or negative correlation) would require further analysis and interpretation of the coefficients from the regression models.

5.5. Temperature and Humidity Interaction

Null Hypothesis (H0): There is no significant difference in accident frequency during rainy conditions (e.g., precipitation > 0.1 inches) compared to non-rainy conditions.

Alternative Hypothesis (H1): There is a significant difference in accident frequency during rainy and non-rainy conditions.

```
Independent Samples t-test:
t-statistic = 39.99861476815193
P-value = 0.0
Mann-Whitney U Test:
U-statistic = 429216534009.0
P-value = 1.4382853799752817e-275
Reject the null hypothesis. There is a significant difference in accident frequency during rainy and non-rainy conditions.
```

Our objective was to ascertain whether there exists a noteworthy dissimilarity in accident severity across a range of distinct road conditions. Two testing methodologies, the Analysis of Variance (ANOVA) and the Kruskal-Wallis test, were employed. The ensuing insights offer an interpretation of the outcomes:

Analysis of Variance (ANOVA):

F-statistic = 16216.085684169951*: The F-statistic represents the ratio of the variance among sample means to the variance within the samples. A larger F-statistic indicates more substantial differences between the sample means.

P-value = 0.0: The null hypothesis (H0) for this scenario postulates the absence of any meaningful discrepancy in accident severity among different road conditions. Conversely, the alternative hypothesis (H1) posits that there is indeed a substantial variation in accident severity across these road conditions.

Kruskal-Wallis Test:

P-value = 0.0*: The Kruskal-Wallis test, a non-parametric counterpart to ANOVA, is employed to compare three or more groups when ANOVA's underlying assumptions are not satisfied.

Interpretation:
Both the ANOVA and Kruskal-Wallis tests have yielded p-values exceedingly close to zero, indicative of compelling evidence against the null hypothesis. Consequently, in both instances, the null hypothesis is firmly rejected.

Conclusion:
Drawing from the outcomes of both tests, we can confidently conclude that there is an appreciable variation in accident severity among different road conditions. In simpler terms, the nature of the road condition does indeed exert a discernible influence on the severity of accidents. This revelation implies that specific road conditions are associated with a greater likelihood of more severe accidents compared to others.

6. Conclusion

We find all this valuable information to understand the factors that contribute to accident severity, which can be important for road safety measures and policy decisions.
There are too many factors that we can prevent from happening to decrease the rate of accidents in the US.