SMM635 — Data Visualisation

Final course project submission template

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Individual Coursework Submission Form

Specialist Masters Programme

Surname: Pothireddy	**First Name:**Preeti
MSc in: Business Analytics	**Student ID number:** 220036212
Module Code: SMM635	
Module Title: Data Visualisation	
Lecturer: Simone Santoni	**Submission Date:** 07/12/2022

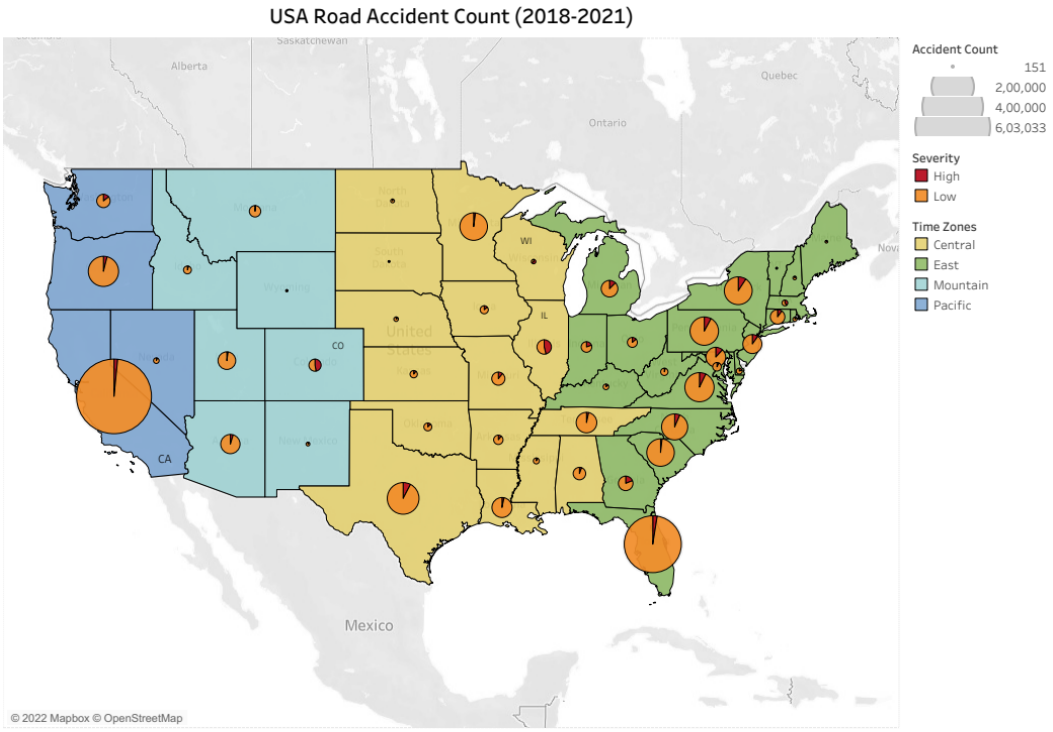
Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.
We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: **Final Mark:**

Visualization #1

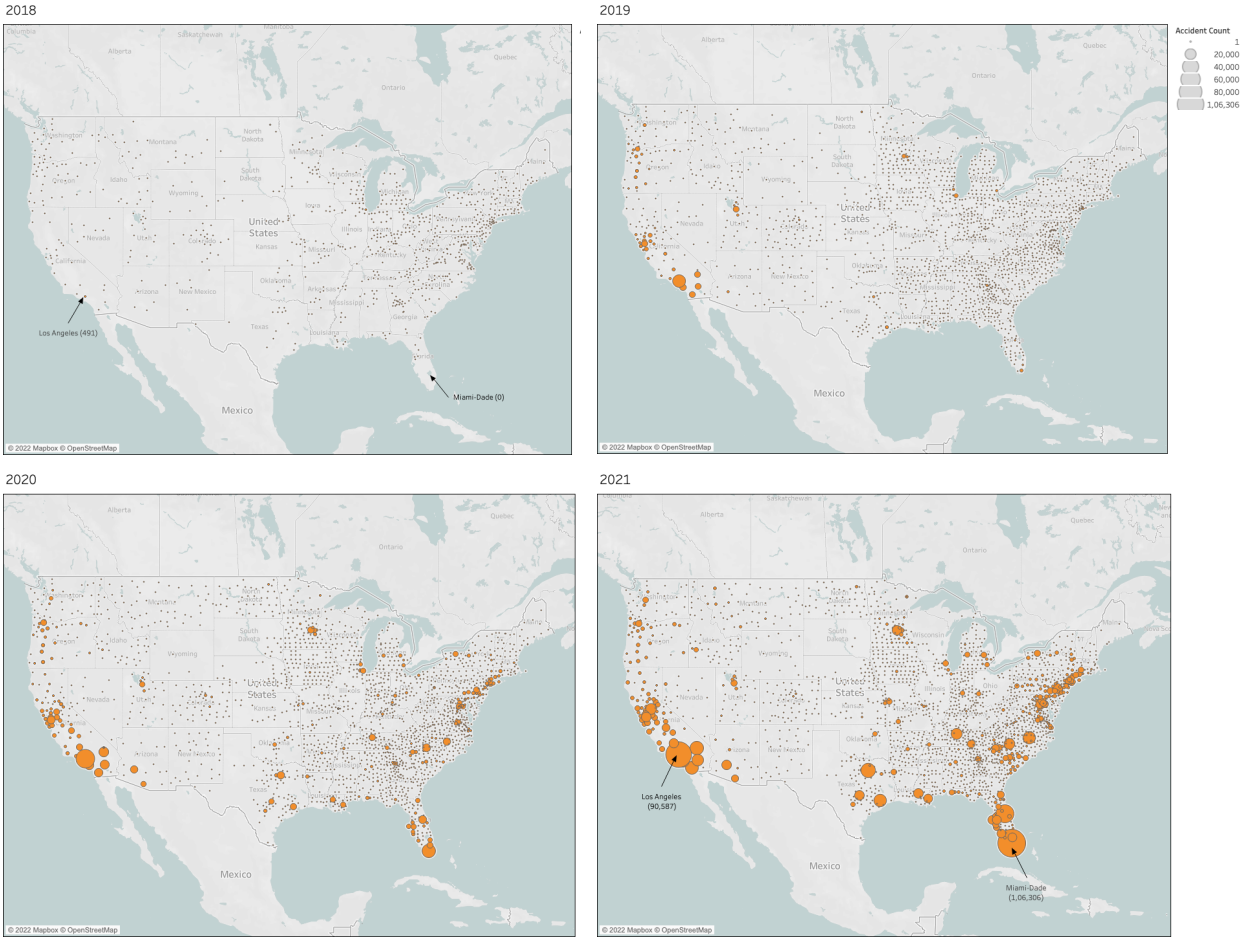
USA Road Accident Count and Accident Severity:



The geospatial map represents the cumulative count of road accidents in the United states over the past four years. The map has been divided into four broad categories based on US time zones: (a) Pacific (dark blue), (b) Mountain (light blue), (c) Central (yellow) and (d) Eastern (green). The size of the pie chart represents accident count. Clearly, California (CA) has the highest count of road accidents. All the states with high road accident count are clustered in Pacific and Eastern time zones. Comparatively, Mountain and Central zones have fewer cases. Further, the pie has been divided into two categories based on the level of severity of car accidents: red representing high severity cases and orange representing low severity cases. Most car accidents across the country are of low severity. Surprisingly, only three states have a significant proportion of high severity cases (i.e Illinois: IL, Colorado: CO, and Wisconsin: WI) and they belong to Mountain and Central zones (i.e zones with less accident count). However, the dataset used had information on only 49 states in the United States and Hawaii was excluded.

Visualization #2

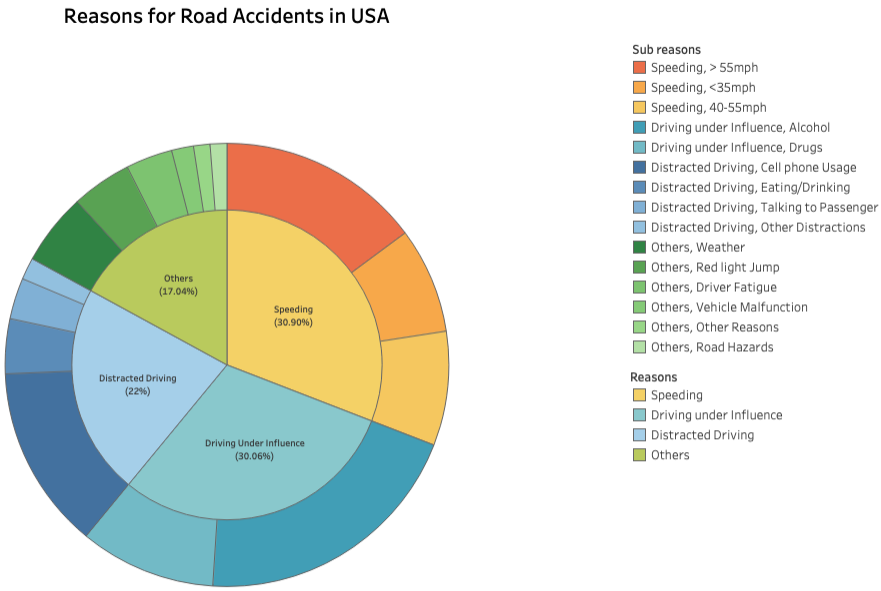
Change in road accident count over the past four years, by county:



The geospatial maps highlight the changes in US road accident count over the past four years, by county. The size of the circles represent the count of road accidents county-wise. Annotations have been used where required to aid comprehension. From 2018 to 2021, we observe a dramatic jump in the count of cases. In 2018, there was no particular hotspot for road accidents, but over the span of the next three years, we can see the emergence of certain counties (such as Los Angeles and Miami-Dade) becoming breeding grounds for road accidents in the US. In fact, in 2018, Miami Dade had 0 road accidents and by the year 2021, this number rose to 106,306. On the other hand, counties in South Dakota, Wyoming and Nevada have consistently been safe from road accidents. Interestingly, even during the pandemic (2020-2021), with travel restrictions, we would expect the road accident count to drop, but surprisingly this did not hold true in the United States.

Visualization #3

Reasons for Road Accidents:



The nested pie chart represents the reasons for road accidents in the US. The pie chart has been divided into four broad categories: Speeding (represented by yellow), Driving under Influence (represented by turquoise), Distracted Driving (represented by blue) and other reasons (represented by green). Further, these categories have been divided into subcategories, categorized by varying intensities of their respective colors. The pies have been arranged in descending order to aid comprehension. Borders have been utilized for easier differentiation. Some key insights from analyzing the pie chart are:

a) Speeding (30.90%) and Driving under Influence (30.06%) are the two most frequent factors for road accidents in the US.

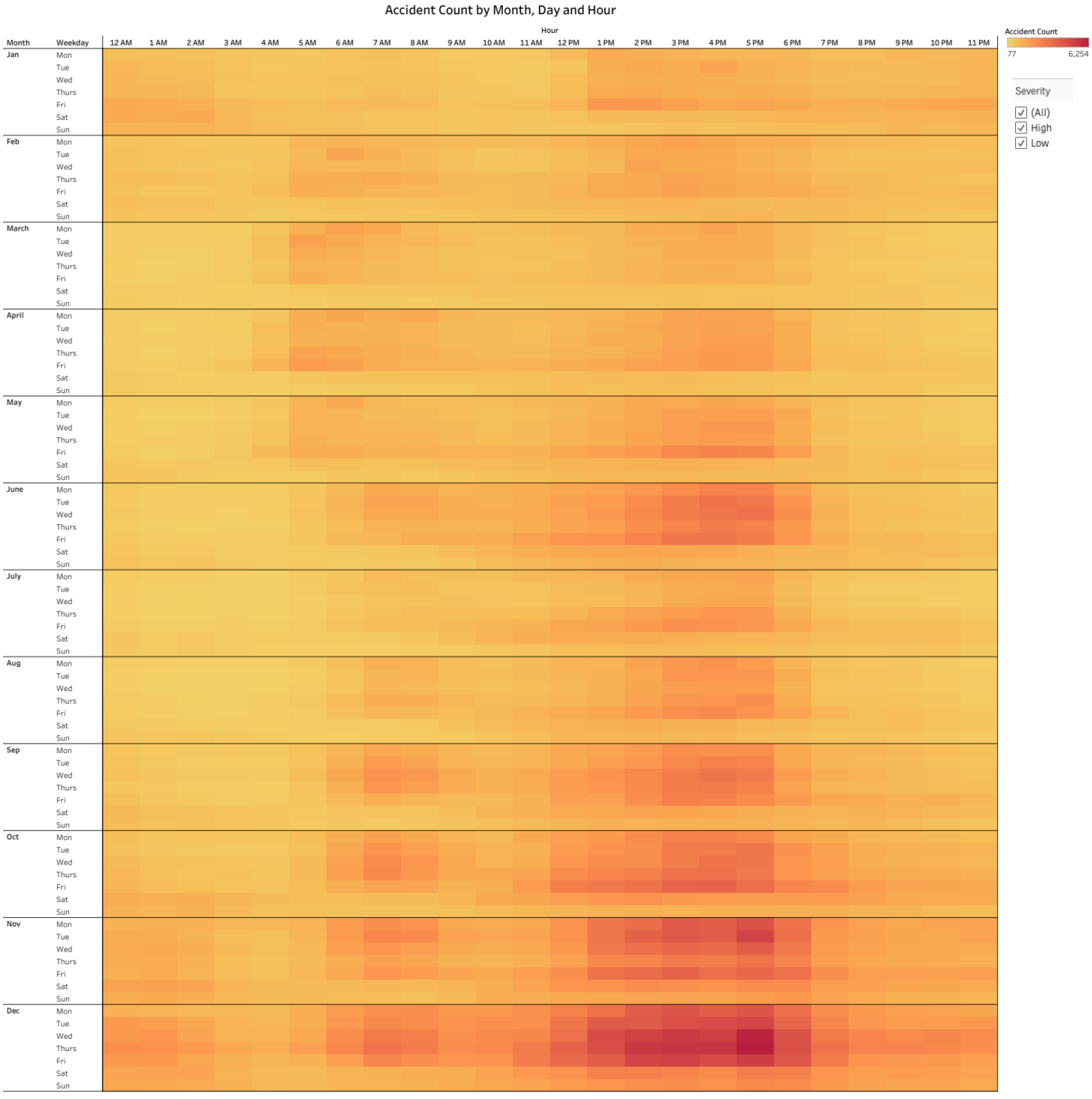
b) The majority of accidents due to speeding frequently occur on roads with speed limits greater than 55 miles per hour.

c) Drunk Driving is a major contributor in ‘driving under influence’ accident count

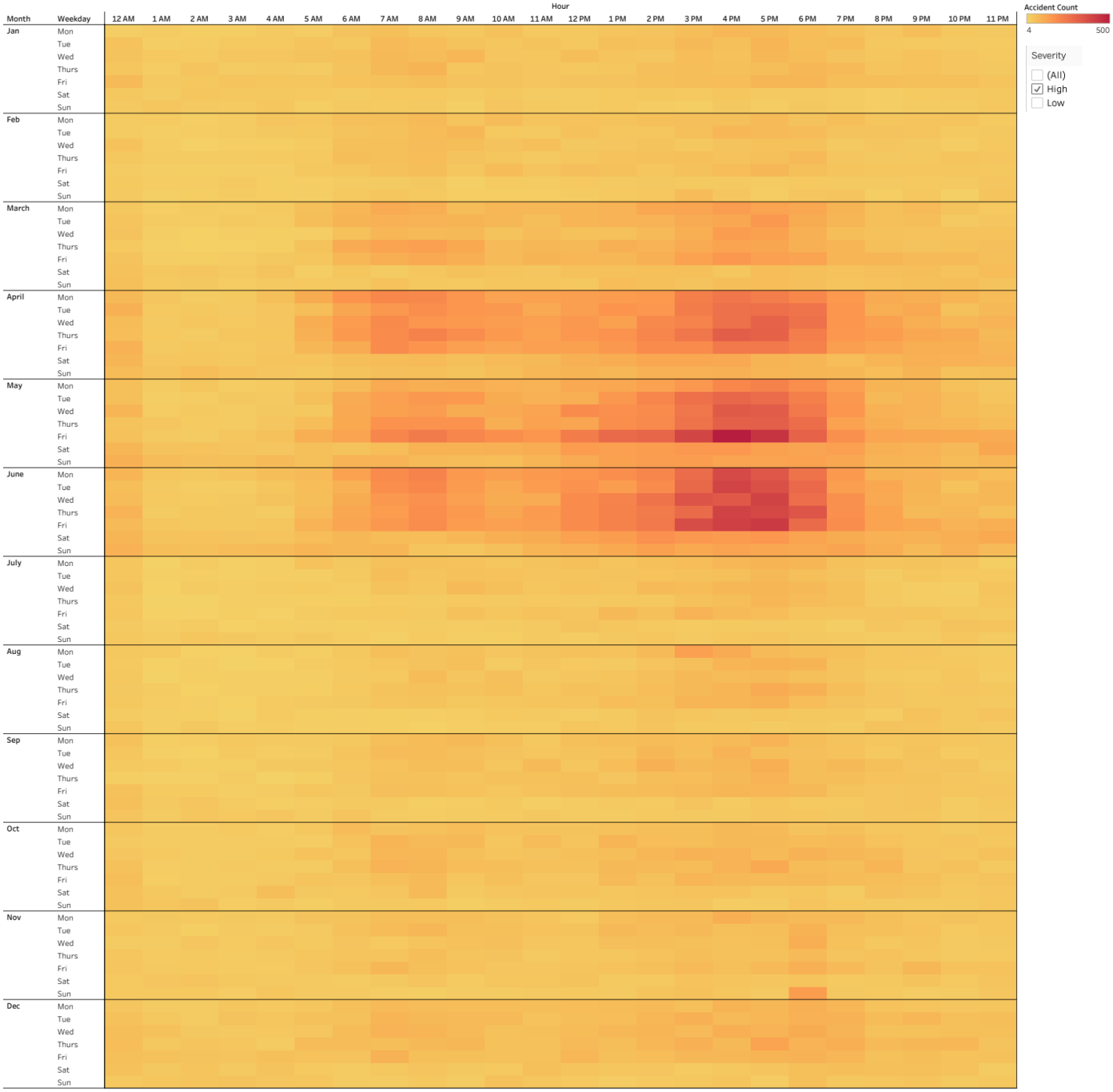
d) Cell phone usage while driving plays a major role in ‘distracted driving’ accident count.

Visualization #4

Accident Count by Month, Day and Hour:



Accident Count by Month, Day and Hour

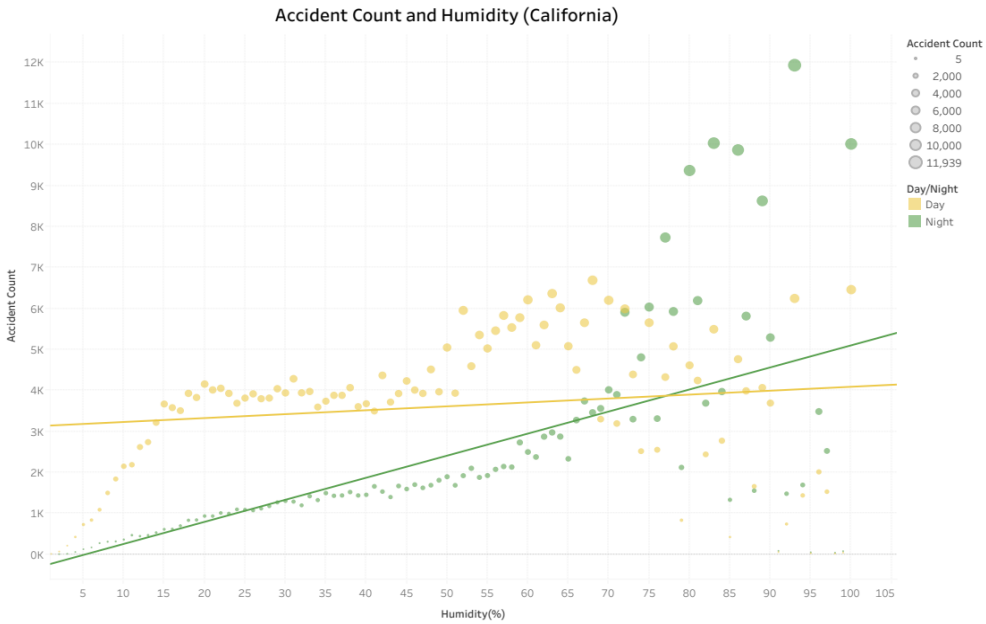


The heat map represents the road accident count in the United States across three categories: month, day and hour. The intensity of color represents the count of road accidents. From figure one, it is evident that October, November and December (being holiday seasons) experience the highest number of road accidents in the United States. Surprisingly, weekdays have a relatively higher number of cases as compared to weekends. A clear pattern can also be observed when it comes to time: most accidents tend to occur between 3pm to 5pm. In figure two, on filtering only high severity cases, a significant change is evident. High severity accidents tend to be more prominent during the first six months of the year. Clearly, high severity cases occur significantly in months of April, May and June. However, in terms of days and hours, the pattern remains similar to that of figure one. During EDA, on analyzing the distribution of severity (originally given on a scale of one to four), it was evident that that data was not normally distributed. To avoid skewness, one and two were combined as low severity cases, while three and four were combined as high severity cases.

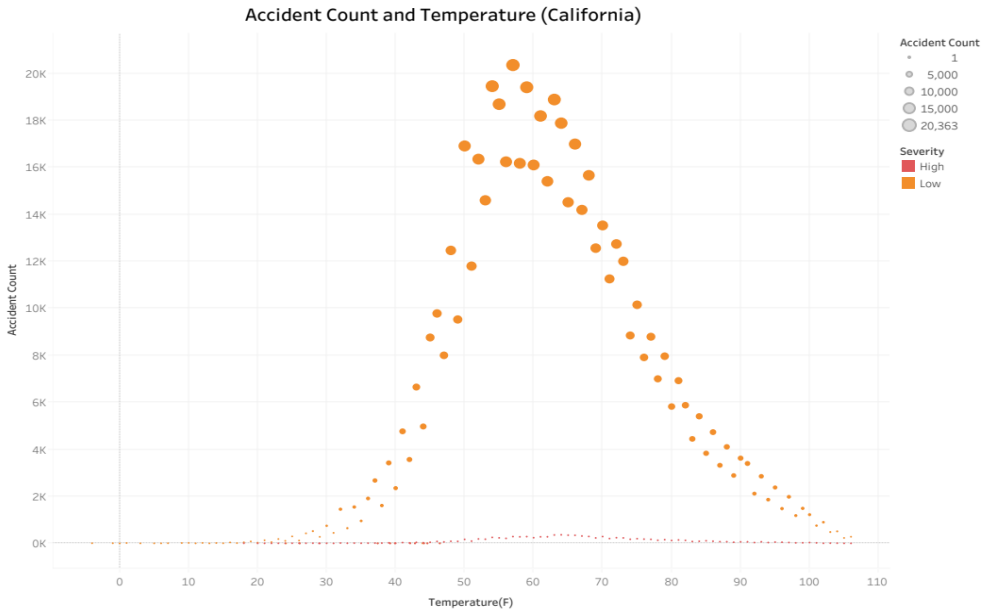
Visualization #5

***The interesting insights**:*

a) Does humidity have an affect on the accident count in California?



b) Does temperature have an affect on the accident count in California?



For the next two visualizations, I focused primarily on California, since it consistently recorded the highest cases of car accidents in the US and I analyzed if weather (humidity & temperature) had an impact on road accidents in CA. First, I analyzed the relation between humidity percent and accident count using a scatter plot. The yellow dots represent day time cases while the green dots represent night cases. The size of the circles represent the accident count. Trend lines and grid lines have been utilized to aid comprehension. Regardless of the humidity percent, the count of road accidents during day time in California remains more/less consistent. Surprisingly, at night, as the humidity percent increases, the count of road accidents also observes a steep rise. Next, I tried to analyze if temperature had any impact on accident count in California, using a scatter plot. The orange dots represent low severity cases and the red dots represent high severity cases. High severity cases remain unaffected by temperature but when it comes to low severity cases we can see a discernible pattern (i.e low severity accidents seem to rise with increasing temperature, reach a peak at 60^{O} F and then observe a dip).

References:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
- National Highway Traffic Safety Administration (https://www.nhtsa.gov/)