Logo, company name

Description automatically generated

**VISUALIZATION PROJECT II**

**(POWER BI)**

**ISM 6361**

**Data Visualization**

Professor Ryan LaBrie

Ngoc Oanh Nguyen

Winter, 2023

Table of Contents

[1. Business reason 1](#_Toc136660667)

[2. Document how/where the data comes from 1](#_Toc136660668)

[3. Document how the tool is used 1](#_Toc136660669)

[4. Explain why the visualizations/charts are chosen 6](#_Toc136660670)

[5. Explanation/analysis of the output 7](#_Toc136660671)

[*— Location Analysis 7*](#_Toc136660672)

[*— Road Condition Analysis 8*](#_Toc136660673)

[*— Accident Description Analysis 9*](#_Toc136660674)

[*— Time Analysis 9*](#_Toc136660675)

[*— Weather Analysis 11*](#_Toc136660676)

[6. Compare/Contrast visualizations from Tableau to PowerBI & Strengths/Weaknesses of each product 11](#_Toc136660677)

[7. The 3W lessons . 12](#_Toc136660678)

[*— What Went Well. 12*](#_Toc136660679)

[*— What Did NOT go Well. 13*](#_Toc136660680)

[*— What Would I do Differently Next Time 13*](#_Toc136660681)

[8. The hyperlink (web address) to Power BI’s Workspace. 13](#_Toc136660682)

# State a business reason for selecting your tools (problem you would like to solve).

Car commuting is considered the most popular method of transportation in the world. The latest 2021 data report 282.4 million vehicles, 228 million licensed drivers, and 3,140 billion miles driven annually. Therefore, the role cars play in daily life is extremely fundamental than ever.

However, whilst technology and new regulations are making vehicles safer, humans cannot remain complacent since car accidents continue to be a leading cause of death in the U.S. In the United States, a car accident occurs every 13 minutes, according to the Bureau of Labor Statistics. Moreover, in 2021, the vehicle death rate increased by 8.5%, the mileage death rate increased by 2.7%, and the population death rate increased by 11% from 2020.

Using statistics to identify the root cause of these car accidents can help prevent more in the future. Thus, in this project, I will use the US Accidents dataset to carefully analysis of the trends that are emerging to improve the quality and safety of traveling on the road.

The US-Accidents dataset with the latest car accident statistics including the leading causes of car accidents, where and when accidents are most likely to occur, and who is most likely to be a victim can be used for numerous applications, such as real-time car accident prediction, studying car accident hotspot locations, casualty analysis, extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence.

# Document how/where you got your data (if it is publicly available, or internal for a work project).

US Accidents is an available dataset on Kaggle ([Dataset Link](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)).

This recent countrywide car accident dataset covers 49 states of the USA and was collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records now. However, when I started this project, this data was just updated to December 2021 with more than 2.5 million records already. Due to that huge amount of data already, I decided to simply utilize the “US\_Accidents\_Dec21\_updated.csv” file until that time without any updating.

Insurance agents and adjusters, vehicle owners and operators, state, regional, and local government agencies, emergency response agencies, auto bodies, mechanic shops, and towing companies could use this dataset for making informed decisions on the cost of insurance, accident likelihood, and safety features; or promulgate a law to make traffic flows and their streets safer; or accurately redetermine pricing for market coverages. Besides, it even enables identifying the root cause of these accidents to help prevent more future research and practice in this field.

# Document how you used the tool. Many tools are super rich in features, and you probably won’t be exploring them all, explain the parts you did use.

Firstly, I have to load the dataset named “US\_Accidents\_Dec21\_updated” which is a csv file into the tool using the “Test/CSV” option inside the “Get Data” widget.

A screenshot of a computer

Description automatically generated ![A screenshot of a computer

Description automatically generated with medium confidence]()

After the dataset is loaded, I need to transform the data to prepare it for my desired visualizations using the Power Query Editor.

A screenshot of a data

Description automatically generated with medium confidence

Since the ID attribute has the text form, I use the “Split Column by Delimiter” to split the “A-” part, remove that column, change the data type to number form to get the identification number only.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence

Similarly, I also use “Split” option with relevant conditions to remove the extra 4-digit zipcode,

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated

By “Replace Values” option, I find similar values to aggregate them into a single value to have a clean and uniform data set, in Weather\_condition for example.

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

To utilize the calculation feature, I create some new measures including the day of the week and the month when the accident happened (DayofWeek & Month), the time the accident occurred (Hour), and the duration of the accident affecting the traffic flow (Duration) using the below formulas.

* DayofWeek=FORMAT(US\_Accidents\_Dec21\_updated[Start\_Time].[Date], “dddd”).
* Month=FORMAT(US\_Accidents\_Dec21\_updated[Start\_Time].[Month], “mmmm”).
* Hour=HOUR(US\_Accidents\_Dec21\_updated[ST]).
* Duration=ABS(DATEDIFF(US\_Accidents\_Dec21\_updated[ST],US\_Accidents\_Dec21\_updated[ET],HOUR)).

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

After finishing transforming data, I close the Power Query Editor, apply changes, and load the dataset into Power BI View for upcoming visualization.

A screenshot of a computer

Description automatically generated with medium confidence

In Power Bi, there is no difference between creating a Worksheet and Dashboard, which means I can put one or many images and related visualizations on just one page. Thus, I just use the “New page” button for getting a new one.

By using the “Data” pane in the right corner, I could choose the measures/data fields to express in the visualizations. The “Visualizations” pane is to select visualization types and features like bar, line, bubble, Q&A, etc. which is appropriate to the characteristics of data in order to express the implication. User can add more visuals from “Power BI Visuals”. I use a lot of add-on visualizations in this project, consist of Radar Chart 2.0.2, Animated Bar Chart Race, DialGauge, Play Axis, Scroller, Timeline 2.4.0, WordCloud 2.0.0, etc.

![A screenshot of a computer

Description automatically generated]() A screenshot of a computer

Description automatically generated with medium confidence

The “Filters” pane is used for filtering the data fields to get wanted values. For example, I get the top 10 states that have the greatest number of accidents in the US from 2016 to 2021 by using this widget.

![A screenshot of a computer

Description automatically generated with medium confidence]()

You can add motion to make the animation of a chart by adding the Animated Bar Chart Race for visualizing and Play Axis for controlling. Hence, by clicking the play button, the bars representing the top 10 states with the greatest number of accidents will move over the time period.

![A screenshot of a computer

Description automatically generated with medium confidence]()

To create a parameter, users could use the Transform Data/Manage Parameters/ New Parameters. However, due to the limitation of this dataset, I cannot set up any appropriate parameter with corresponding formulas.

The slicer is another way of filtering which is displayed on the report page to narrow the portion of the dataset that is show in the other report visualizations. I use it for timeline and state fields.

![A screenshot of a computer

Description automatically generated with medium confidence]()

A map could present a location with measures based on longitude and latitude.

![A map of the world

Description automatically generated with medium confidence]()

I can highlight the same data from different charts by clicking on it.

![A screenshot of a graph

Description automatically generated with low confidence]()

The image and buttons (“Back” button could be embedded to go back to the previous page in the report)

![A close-up of a word cloud

Description automatically generated with low confidence]()

# Explain why you chose which visualizations/charts.

First and foremost, I choose the chart type based on the nature of the data. Every visualization is automatically synced together within each dashboard, so when a specific field is chosen, it and every equivalent statistic will be highlighted as well.

When I want to show how the number of car accidents changes over time in each state, I use the map combined with the proportional pie charts, which are the parameter ones so the measure can be customized, based on the degree of severity, including numbers 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). Besides, the adjacent funnel graph showing the corresponding number of accidents sorted from the largest to the smallest helps for better visualization.

To show the top 10 states with the greatest number of accidents in the US over the time frame, the animated bar chart is used with the play button next to it for easy control, so that we can see how this list changes as well as each of these states can be compared. Furthermore, I also incorporated the tree map to emphasize how the accidents of each year contribute to the total number at the end.

To represent the top 10 accident-prone streets in the US, I use the bar graph sorted in descending order with the area chart to show the severity of accidents. The scroller is also placed on the top of the dashboard to emphasize the street having less than one accident in the past 6 years.

For analyzing the road condition when the accidents occurred, I use numerous pie charts to examine the presence of junctions, traffic signals, crossing people, stop signs, bumps, give way signs, no exit signs, and railway with the dial gauge to show the number of accidents. The slicers to choose the specific timeline and state are utilized.

The donut chart indicating the number of accidents by severity is placed next to the bar chart showing the accidents by duration affecting the traffic flows and the emphasized card. Moreover, I also use the Q&A feature to show the top 10 states that have the greatest number of accidents for automatically making a corresponding visualization.

I decide to use the word cloud to delineate the accident description of each accident in order to know what the leading reasons behind those accidents are.

The column chart with absolute values and relative values and the waterfall chart indicates the percentage of accidents by time zone and by year.

To explain in which hour the accidents tend to happen most, noted by gray color, I use the line and column chart colored green for daytime and blue for nighttime. Additionally, the radar chart indicates the overview of accidents percentage for different hours of the day.

To represent the car accidents by day of the week and accidents percentage by month, I use the decomposition tree and funnel chart respectively. It helps me compare the data of each day of the week and each month with each other.

Last but not least, to show the link between weather conditions and the number of accidents, I choose the line chart and “Key Influencers” feature.

# Give an explanation/analysis of the output. What did you learn or uncover?

## Location Analysis

![A map of the world

Description automatically generated with medium confidence]()

This is the map of accident severity by state. As can be seen above, the second degree of severity accounts for the largest number of car accidents, which means most accidents has a medium impact on traffic, resulting in not too short or too long delays. In addition, the pie chart in California state is the biggest one, possibly because it has the most population leading to a large equivalent number of accidents during the study period. Therefore, it is difficult to conclude that commuters encounter a larger possibility of having a car accident when living in here than other states. Other degrees of severity have remained at a low level over the years.

![A screenshot of a computer

Description automatically generated with medium confidence]()

We can see that California always has the greatest number of car crashes and it considerably increases from 2016 to 2021. About 30% of the total accident records of the past 6 years in the US are only from this state. Hence, there should be enacted laws and safety standards for car travel due to this alarming rise in accidents. Moreover, in the first half time period, Oregon ranked second then decreases substantially in the second half, which is a positive signal. Meanwhile, the authorities in Texas and Florida are the second and third highest states for the number of road accidents in the country, should consider the accident situation in their states since they keep surging over time.

![A picture containing text, screenshot, plot, colorfulness

Description automatically generated]()

The column chart indicates that in the last 6 years (2016-2021) Street No. I-95 N is having the highest road accident records at 39.85K. On this street, daily 14 accidents occurred on average. The I-5N is one of the most dangerous ones. Therefore, people should concentrate and extremely be careful when driving on these routes.

## Road Condition Analysis

![A screenshot of a computer

Description automatically generated with medium confidence]()

As can be seen from the report, almost in every case (99.98%), the bumper was absent in the accident spot. In 4.26% of cases, road accidents happened near the crossing. In addition, in 98.62% of cases, there was no stop near the accident area. 13.82% of road accident cases were recorded near the junctions. There are a few accident cases recorded near the railway. 92.59% of road accident cases were recorded far from the traffic signal. As a result, appropriate authorities should take responsibility for car accidents to reduce accidents by increasing the presence of bumper, traffic signals, or stop signs on the street to control the traffic flow.

## Accident Description Analysis

![A close-up of a word cloud

Description automatically generated with low confidence]()

The word cloud depicts the recorded accident descriptions. The leading reasons for road crashes are accidents, road closed, road blocked, etc., and most on the right side. However, this visualization contains a lot of confusing words which do not contribute meaningfully for readers such as particular names like "Adams", and "2nd St" or general terms like "main", and "lane". It needs to be well-prepared and cleanse the confusing words for better insights.

## Time Analysis

![A screenshot of a graph

Description automatically generated with low confidence]()

The second degree of severity accounts for the greatest number of car accidents at about 90%. Furthermore, of the 2.85M accident cases, the majority of them last from 1 to 2 hours. The provided Q&A shows the corresponding graph according to the search keywords. In this instance, it illustrates the top 10 states that have the greatest number of car accidents by a bar chart, including California, Florida, Texas, etc.

![A picture containing text, diagram, plot, line

Description automatically generated]()

According to the charts, the Eastern time zone region of the US has the highest number of road accident cases (43%) in the past 6 years while the Mountain time zone region of the US has the lowest (6%). Besides, from the right corner figure, it is clear that in the last 6 years (2016-2021), the US accident percentage has increased significantly. More than 75% of the total road accident records of this period happened only within the last 2 years (2020, 2021).

![A picture containing text, screenshot, diagram, plot

Description automatically generated]()

Based on the color displayed in the chart above, we can easily distinguish the car accident tendency during the daytime and the nighttime. Around 18% of road accidents occurred between 6:00 AM to 9:00 AM. In the evening, around 27% of road accidents occurred between 2:00 PM to 6:00 PM. The most-deadliest accident hour is 5:00 PM implies the evening office-returning hour. The second deadliest accident hour is 7:00 AM implying the morning office-going hour. The radar chart also gives the overall picture of the distribution of the number of accidents during the day.

![A picture containing text, screenshot, software, web page

Description automatically generated]()

Based on the funnel chart, around 474K road accidents occurred in December. March is the month with the least number of road accidents in the US (at 158K). Furthermore, about 45% of the road accidents occurred only within the 3 months, October to December (i.e., the transition period from Autumn to Winter). It is probably because the road condition in Winter is much more extreme for driving due to snowy and icy roads, leading to the increased threat of an accident. Furthermore, working days of the week have almost two times higher accident records at about 450K, compared with the weekend days at only around 250K to 300K. In detail, Friday of the week, in which people tend to go out and have fun after an exhausting working week, has the highest percentage of road accidents, while on Sunday, this number is the lowest since people usually spend time at home.

## Weather Analysis

![A screenshot of a graph

Description automatically generated with medium confidence]()

In most road accident cases, the weather was clear and cloudy. What is more, the “Key Influencers” shows that when the wind chill is from 46 to 93 Fahrenheit degrees, it tends to be the key influence in increasing the number of accidents by 2.2 compared to other factors, including temperature, pressure, precipitation, and humidity.

# Compare/Contrast your visualizations from Tableau to PowerBI, what were the strengths/weaknesses of one product’s solutions over the other.

**A picture containing screenshot, text, circle, colorfulness

Description automatically generated![A screenshot of a computer

Description automatically generated with low confidence]()**

I made the visualizations about road accidents using Tableau (the left one) and Power BI (the right one). Both of them show the same insights about the number of road accidents in the US over days of the week, which Friday has the greatest and Sunday is the least.

I mean they both produce a variety of different visualizations, connect to numerous data sources, are code-free and user-friendly, and support interactive interfaces. However, from my limited experience and perspective, I think they are quite the same to some extent when it comes to making visualizations, I like the design language of Tableau rather than Power BI and the transforming data power of Power BI than Tableau. Each of them is superior in their own field.

However, feel everything runs more smoothly in Tableau (maybe I used a Macbook which has a stronger configuration for the Tableau project with a smaller dataset, and Windows which is much older for this huge amount of data in this Power Bi one)

Here is the strengths and weaknesses of each product comparison.

|  |  |  |
| --- | --- | --- |
|  | Tableau | Power BI |
| Strengths | * Provide online resources, guides, training, and an online forum - more engaging. * Have high and secure overall performance of this tool. * Create different kinds of visualizations in one shot. * Possible to handle a large amount of data with Tableau. * Incorporate Python or R for performing complex table calculations. | * Affordable tool - free of cost to create datasets, dashboards, and reports. * Easy to use and similar to Excel. * Access data from different data sources from anywhere and anytime multiple times. * Have interactive dashboards - filter, highlight features, etc., in a single click. * The query editor in Power BI - modify data files before loading them into the Power BI * Quick deployment * Integrate with Python and R language for using visualizations |
| Weaknesses | * Not provide an automatic option for refreshing reports and require manual efforts to update data in the back end. * Not provide version control so users cannot get back to the previous level of data once the reports and dashboards are published. * Have a limit of table display - difficulty in displaying data for larger tables. | * The Power BI desktop version is not compatible with iOS. * Have a clunky interface filled with many icons that may block the view of reports and dashboards. * Run slow with a file size greater than 1GB. * Difficult to process larger datasets with complex features. |

# Conclude with the 3 W’s (What Went Well, What Did NOT go Well, What Would you do Differently Next Time).

## What Went Well.

The data preparation did not take so much time for me since the dataset is quite a clean one, so I did not have to cleanse them that much. Everything is relatively in good condition which facilitates me to focus on manipulating the dataset.

Power BI is extremely rich in features and visualizations, thus there are a lot of things to do. I think it is quite similar to Tableau to some extent so I could apply my experience and previous knowledge from the first visualization project to this second one.

The tutorials provided by Professor are very detailed, which made it a lot easier for me to get used to this new tool and to manipulate the data. Moreover, there are a lot of provided documents and related information available on the forum so I could use them as the material.

## What Did NOT go Well.

First, finding a dataset with all the required attributes to be able to create calculations, maps with location, line charts with motion, etc. is not easy at all but extremely time-consuming. In the first place, I chose another dataset but that one is not cleaned with a lot of missing values which are important. Therefore, I had to switch to other datasets before deciding to deal with this US Accidents dataset. However, there is still a lot of missing data and confusingly classified attributes in this dataset, so I have to considerably eliminate the unnecessary ones to get quite a clean one. The dataset is huge for my computer capacity so it ran slow and took a lot of time to load and apply new changes, so I could not add as much as visuals and create new measures or parameters as wanted.

In addition, some attributes are string types such as Description, which can be difficult to illustrate with other metrics. However, I could be able to find a way to utilize that information in the word cloud form, which is not effective as aforementioned.

Besides, this dataset does not have many numeric attributes, which makes it difficult for me to freely choose measures to display on the graph. I only have the US accidents count with their severity including from 1 to 4. This caused me to repeatedly use each of the injured people variables over and over. I tried to create some new measures such as DayofWeek, Month, Hour, and Duration. Because of the limited data, my charts are not quite as beautiful and consistent as I would like. Sometimes the performance is quite forced without really showing the meaning.

## What Would I do Differently Next Time

I want to spend more time finding the data to understand each attribute more comprehensively as well as explore the rich set of available features of Power BI. In particular, I had to carefully consider the amount of numeric data to be able to perform calculations and visualize the data more easily. Moreover, I want to explore the Parameter feature in this tool, which I cannot do in this project yet.

I also want to try a different larger dataset next time with different visualization tools like Metabase which I researched in the KMBI class or Zoho to broaden my knowledge in this field.

# The hyperlink (web address) to Power BI’s Workspace.

[Nguyen\_Ngoc Oanh\_Visualization Project II](https://app.powerbi.com/groups/me/reports/a332f4e7-857f-4731-862b-2cb49e78e76b/ReportSection?experience=power-bi)