IS 590 DW SP20: Data Warehousing & BI
Final Project
Akshay Bafna

I'm sure that in our childhood we all get fascinated after hearing & seeing movies related to aliens & UFOs. Similarly, I have always been interested to know more about the outer world. By analyzing this dataset, it helped me in a way to explore & learn about it. The Target of my final project is to generate **a Descriptive Report** using **the Power BI** platform for the dataset of **UFO Sightings**. The dataset is available on *Kaggle* from NUFORC.
The link for the dataset on the Kaggle is https://www.kaggle.com/NUFORC/ufo-sightings .
The dataset consists of 88k rows of complaints with 11 columns, but for the report generation, I have confined the analysis for the US region. The dataset used for my analysis is 70k rows. Data contains the city, state, time, description, and duration of each sighting. The UFO Sights are from the year 1910 to 2014 in different parts of the US.

The main objective of my analysis is to understand the nature & the pattern of seeing the UFOs in the countrywide. My focus for analysis is on the location of the sight, duration, shape of the UFOs. Further, I would like to understand the period of sight based on daytime, days in a month, month & year. Lastly, I will try to understand the comments based on the word cloud. The other part is to come up with the findings from the report and the inference I learned from the analysis of the dataset.

There were a few challenges that I faced while working on the overall project. This encounter helped me to learn new skills and improve my thought process for the analysis.

Visualizing the incidents based on the geo map brought some issues, as the states in the dataset used the abbreviation. So, while generating the visualization in Power BI, the information is seen at other locations of the world. However, I resolved it by making relevant changes in the dataset by changing the abbreviation with the name of states in the US & further, using the filter for the country to confine the occurrence of the incidents.

To analyze the most prominent shape, I saw a lot of blanks in the dataset which was quite confusing. So, I assumed it to be a good test to define a false claim of viewing the UFO. Hence, I concluded that if there is no shape seen it will be a false claim and removed the rows which prompted with no shape. But, on my further analysis by generating a word cloud most of the comments mentioned light as the keyword. Further, the 24 hours day duration chart (9th chart) saw incidents of viewing UFOs in the daytime. This completely changed my notion of thinking about the shape which I assumed to determine the valid incidents. Finally, I included 'blank' incidents of shape, as it is possible that in some incidents light may be too intense & sharp which can be seen for a short duration that had caused difficulty in commenting on the shape.

It was quite interesting & new experience of generating the word cloud using power BI. The word cloud was generated using the comments of the incidents. The power BI has a feature of removing

some common stop words automatically. To remove specific words that are not relevant to the dataset, the developer must add the stop words in the dialogue box. It took some amount of time to learn the process of removing the stop word & understand the relevance of the words for the dataset. It was challenging for me to decide whether a word should be kept or removed from the word cloud. For example, The interesting word for me was '((#HOAX??))'. I removed the word from the word cloud, as it was not properly described in the comments and was added at the start of the comment with similar punctuation marks.
Some other list of stop words is: "HOAX 44 33 03 2 5 1510 13 12 15 23 11 09 39 20 6 05 21 10 1 01 02 2055 3 25 8 07 30 4 7 s V 45 39s 4 9 06 PD 08 00 very one each two away saw seen minutes quot looked Note" (the numbers & words which I removed was based on research on stop words & thinking to bring effectiveness). the range of the word count for the word cloud was from 50 to 200.

I also learned to write the code in power BI using Data Analysis Expressions (**DAX**). It is a library of functions and operators that can be combined to build formulas and expressions. I used it to generate duration bins for the duration UFOs were visible to the people. The major challenge I faced while generating the visualization was that most of the incident's duration were confined in 300 seconds and very few incidents were of long duration such as 16000 sec (4 hrs. approx.), hence, the histogram was not very uniform across the axis with a lot of gaps in the chart. Hence, I further decided to enclose a histogram (8th chart) for the duration to 350 sec and represent a maximum incident-specific line chart (6th chart) to represent the duration of maximum incidents.

Result or product of my project

The outcome from my analysis of the UFO's dataset for the US region is that most of the states which saw such incidents are located close to western coast such as Seattle, Los Angeles, Houston. The prominent reason I see for such occurrence is that these cities have clear skies at night and pleasant weather for most of the year whereas the east coast is always prone to natural calamities such hurricane, heavy snowfall. Some cities act as an exception, in this case, they are Chicago, Miami.
With the help of the day duration bar chart (9th chart), most of the incidents occurred during the time frame of 6 pm to 11 pm. It was quite interesting to know that some incidents also took place in daylight time such as in the morning or noon. It can be a good focus point to further analyze the incidents and find the similarity among them.
The dataset includes incidents from 1910 to 2014, but there is a rise in cases after 2000 which can be due to the rise in the technology. There has been an exponential growth & advancement of technology which has made people tech-savvy & accessible to smartphones. Making videos, capturing images & sharing is quite easy in today's world. On the other hand, we can also say that governments have shown recent interest in topics such as space exploration, finding earth twin, rise in the private sector making space travel a cost-effective venture. Hence, most of the government & agencies are conducting covert projects which are not disclosed to common people. The trial act may have been noted by the people can be a possible UFO event.
Most of the incidents are reported in the mid-year i.e. July, August & on the 15th-16th of the month which is the start of the fall season in the US. It was quite astonishing to learn that there have been some incidents where UFOs were visible for more than an hour but still there is very limited

information which is shared in public about such incidents. We can conclude such incidents as a hoax because most of the incidents are confined for less than 4-5 mins only.
Some interesting words which I see in the word cloud are related to the shape, color, direction, and related to the formation. The NUFORC word is occurring in the word cloud because the organization has mentioned some conclusion along with the comments to bring clarity of the sight. For example: on some incidents, people viewed mars and concluded it to be a UFO.
The prominent shape sighted of UFO is a triangle, circle, disk & light, where light is the most reported. The possibility of light as a main shape of the UFO is that the observer has limited exposure of event or the event would be too illuminating that they reported light as the shape. This creates some suspicion of being such incidents as Hoax.