FOLLOWING THE FOOTPRINTS OF UFO

ISE 5103 – INTELLIGENT DATA ANALYTICS

STUDY OF THE DATA FROM NATIONAL UFO REPORTING CENTER USING DATA ANALYTICS TECHNIQUES – BY STUDENT X

EXECUTIVE SUMMARY

WHAT ARE WE LOOKING FOR - LOOKING INTO THE UFO DATA-SETS, FEW QUESTIONS THAT NATURALLY ARISE ARE AS FOLLOWS:

- Is there a predictable pattern to the incident occurrence based on time, location, and duration?

  Population of a state has a very high impact on the number of UFO reported.

- Is there a correlation between the color and shape of the object observed?

  Red color was most associated with flashing and blinking lights. Green color was most associated with neon, fireball, falling light with tail.

- Through text mining, find what were the most frequently used words by witnesses to describe an UFO

  Top 25 Words – LIGHT, OVER, OBJECT, BRIGHT, and SKY

- By studying the distribution of time between UFO occurrences, can we predict the chances of next UFO appearance?

  Based on the past occurrences, the following states have highest chances that the UFO will appear in the next 24 hours. CA, TX, FL, AZ has the highest chances. 96% chances!!!

- Is the population has anything to do with the chances of reporting an UFO.

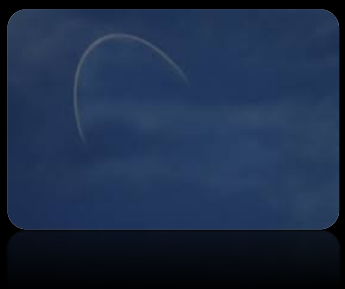
  States WA and NV slightly contradict this claim – It shows low number of reporting.

- Associate the different parameters to bring meaningful insights and patterns.

  Most UFO occurrence was during early morning between midnight and 6:00am, the color was either bright or white light, while the shape was a fireball.

SUMMARY

The UFO data set is available from the National UFO Reporting center website. The data set has a collection of UFO incident reports recorded across the US and Canada. Each incident is recorded with date, time, and place; with a concise summary on the details of the incident. In addition to that, the duration of the incident and the shape of the object witnessed is also recorded. There are over 93,000 incidents reported so far since 1998. Data set is available here: http://www.nuforc.org/webreports.html



RECOMMENDATION

Predictions based on analysis is completed, our recommendations would be to further analyse to include population and other categorical data like industries in the area, education level across population, rate of flight traffic across the country, states/region, Text weight/ sensitivity analysis, etc.

PROBLEM DESCRIPTION

Our project was to isolate study and find a solution for a data-intensive problem using different data analytics techniques learnt in the course over the semester. One of the main requirements of the project was the problem complexity and challenge. After analysis of several data sets, our group finally settled on two. An Oil and Gas data set, and the other was related to Unidentified Flying Object (UFO). Our group decided to choose UFO dataset since Oil and Gas data lacked the necessary complexity in data. This report is the result of the analysis conducted on the UFO data.

UFO DATA:

- Data Set: The UFO data set is extracted from National UFO Reporting center.
  UFO Website: http://www.nuforc.org/webreports.html.
- Existing Data Analytics: Due to the lack of scientific proof related to this set of DATA, any data analytical techniques to better understand and interpret the problem would add to the credibility of the information.
- Volume: The volume of UFO incident data is huge enough to meet the requirement for this project. There are over 93 thousand incidents reported so far since 1998.
- Predictors: The data set has the following predictors of the reported incident
    - Date and time of occurrence
    - Place of the occurrence
    - Shape of the object seen
    - Duration of the incident
    - Summary of the incident with detailed description of the sighting

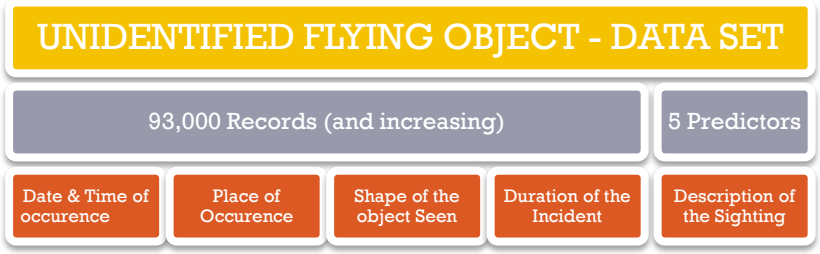
Looking into the UFO data-sets, few questions that naturally arise are as follows:

- Could there be a pattern to the incident occurrence based on time, location, duration?
- Is there a correlation between the color and shape of the object observed?
- From the multiple incidents reported on a specific day across the country, what is the chance that the same object was witnessed by people from different places? In other words, did the UFO follow a certain path?
- Through text mining, can we find out what were the most frequently used words by witnesses to describe an UFO? Find out how much the witnesses were excited or terrified on seeing an UFO.
- By empirically studying the distribution of time between UFO occurrences, can we predict the next UFO appearance?
- Based on the events occurrence across the country, can the population of a state be a factor in affecting the trend of incident reporting.

We would like to draw meaningful insights for the above mentioned questions with the help of various data analysis techniques.

- Understand the distribution of different data attributes with the help of graphical representations such as histograms and scatter plots.

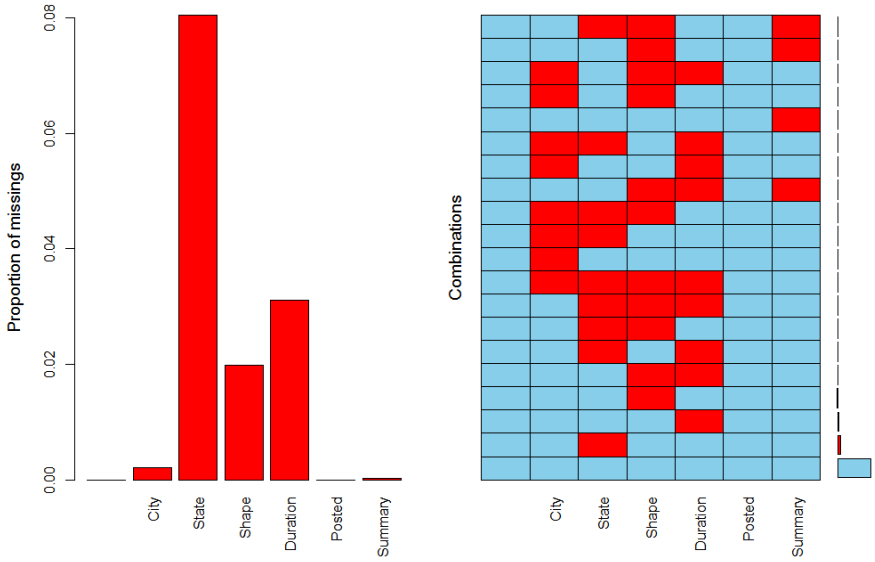
- Apply predictive analysis techniques to find the chances of future incident occurrence.
- Apply classification modeling to categorize the incidents based on incident description.
- Use different text mining techniques to derive information from the unstructured text data.
- Visualize the location based results with the help of Google API and R program.



EXPLORATORY DATA ANALYSIS - THE HIGHLIGHTS

Before we proceeded with the analysis, we had some challenges with the UFO Dataset as listed below.

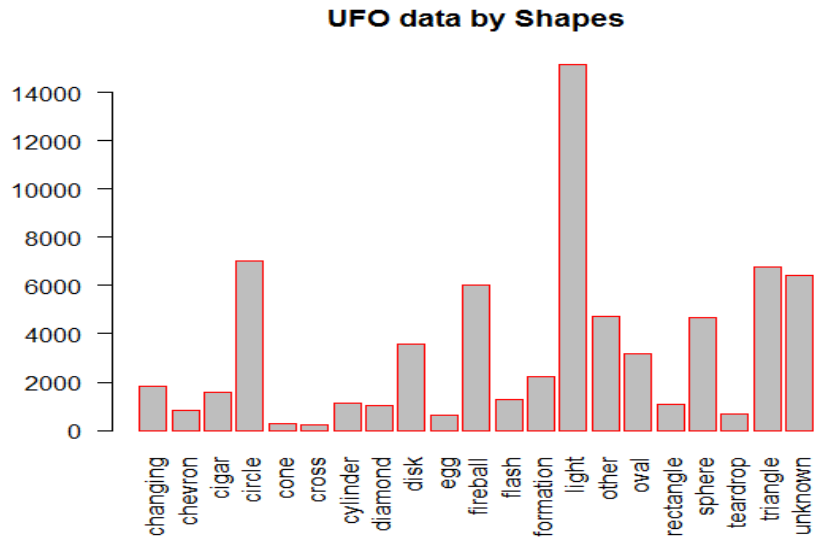
- Inconsistent date and time formats: Time of incident was not recorded for about 2600 incidents. We have done missing value imputation on these cases.
- Inconsistent place information: Around 6500 records had no mention of State. These types of records were included in the analysis where the State parameter was not considered, but excluded in all other cases.
- Missing shape information: Around 1500 records were missing the shape of the UFO. Missing value imputation performed and the value is set as "Unknown" for these cases.
- No Durations: No uniformity followed in recording this value. We categorized and the duration detail is preserved as "In seconds", "In minutes", "In hours".
- Extracting from summary: String search applied on this column to identify the color of the UFO object by matching the details with a list of known colors. A new "color" column created based on the search. Records with no mention of color are set as "Unknown".



Graph 1: Missing Value Information

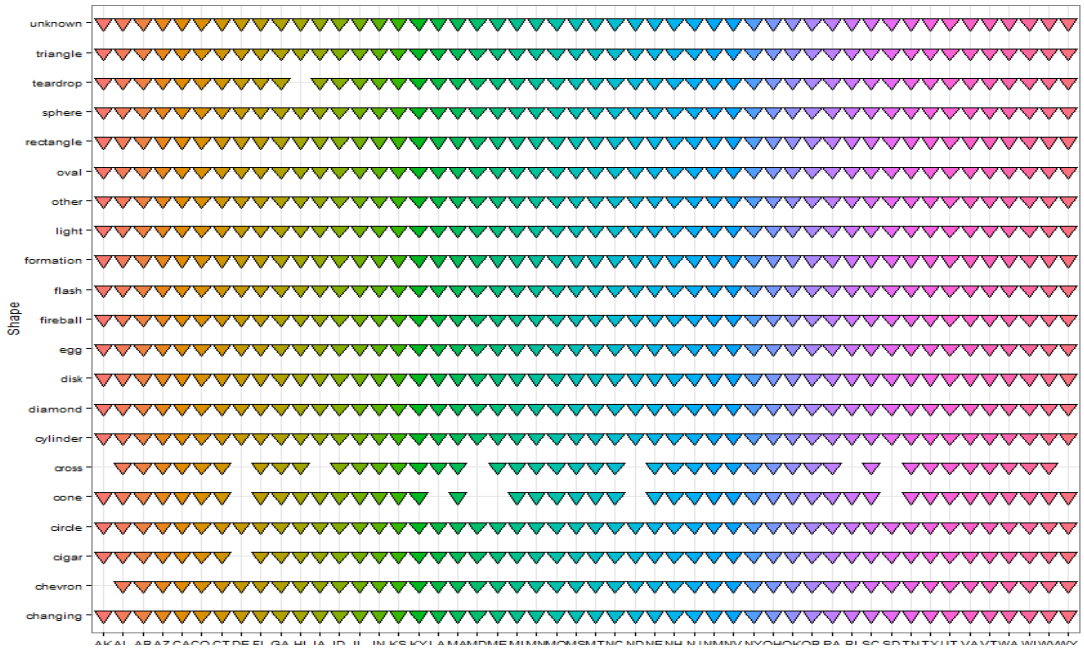
***ANALYSIS:* As mentioned in the analysis above a lot of data did not have 'State' data along with Shapes.**

After our data preparation, we did initial analysis to answer some basic questions like distribution of UFO across states, shapes, etc. We have included our findings in our graphs.



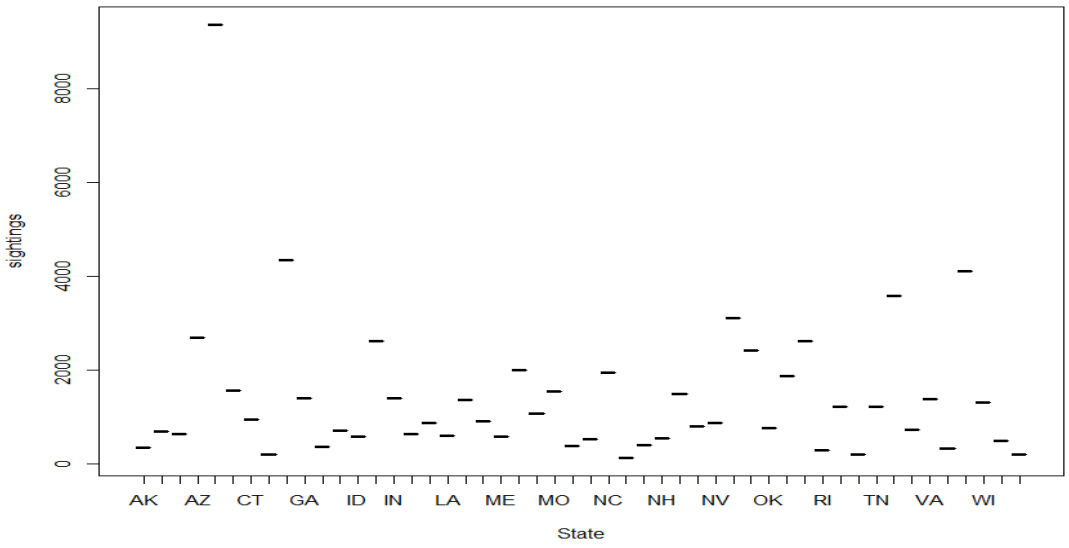
Graph 2: UFO data by shapes

ANALYSIS: ***As can be seen light is the common shape as referred by people across when talking about UFO***



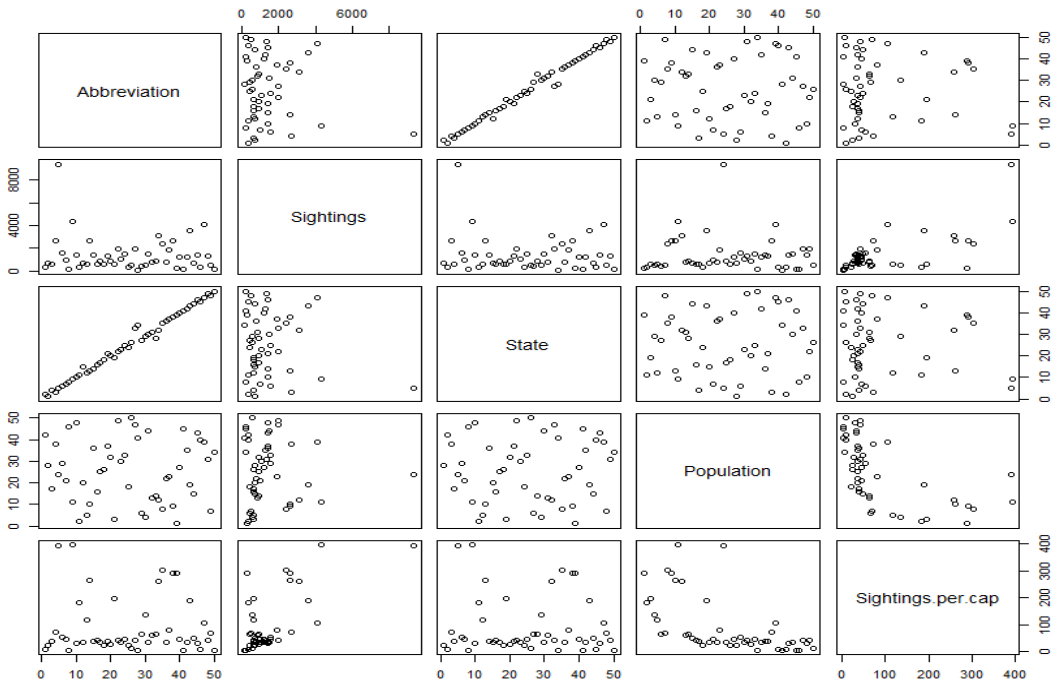
Graph 3: UFO data by Shape and State

ANALYSIS: ***Most of the shapes are there across the states. But IA, AK, DE, LA, SD, ND, WY MD, ME does not have the occurrence of certain shapes***



Graph 4: Initial Sightings by STATE

ANALYSIS: ***Initial analysis reveals that highest state of occurrence is CA that exceeds 800. The lowest is as low as 20's.***



Graph 5: Scatterplot across the predictors

ANALYSIS:
Scatterplot does not reveal a great pattern between STATE and population!!! However look at Sightings per Capital and Population. It looks reverse exponential. Something of interest!

Our basic analysis reveals the following:

> We do not necessarily have a pattern between the predictors except may be population and Sightings per capita. Common shape reference by people for an UFO is "LIGHT" and though most states have these shapes listed some states have certain shapes missing. Some of the common shapes mussing – CONE, CIGAR, CROSS. Considering the closest shape to these would be OVAL, EGG, and FIREBALL etc. which is present in all the states. The lowest frequency of sighting is 20 while the highest frequency is 800. But most of it lie around 100 – 300 range. Based on these initial findings, further analysis was done as in the given in the Analysis Plan.

ANALYSIS PLAN

Based on the above analysis, we decided to proceed with the analysis. First we cleaned up the data for THREE SECTIONS of analysis – Modeling, Visualization and Text Mining. Each had its own problems that were individually addressed. We will first address Modeling.

MODELS:

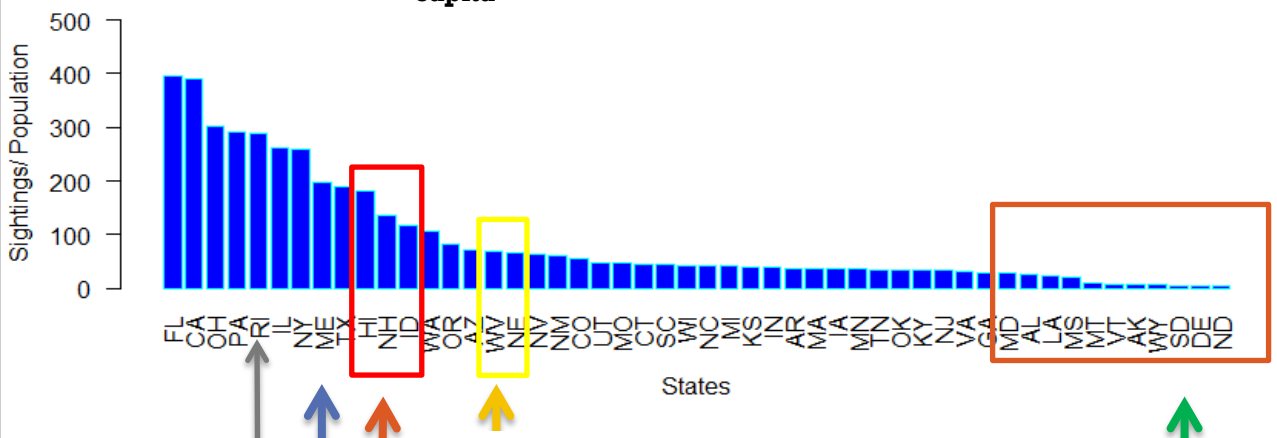
We wanted to model the data based on exploratory data analysis. We tried the following modelling techniques on the data.

- Clustering - K Means
- Time Series Analysis – Exponential Distribution

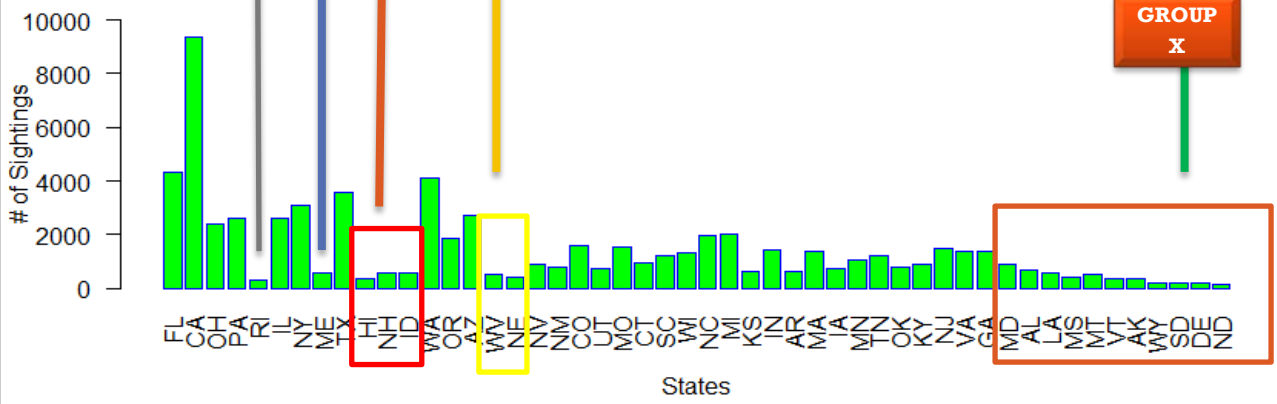
TIME SERIES ANALYSIS – EXPONENTIAL DISTRIBUTION

Our data set had over 90,000 records with increasing value. To predict the data on the occurrence of Flying Object, we decided to take a subset of the data. To choose our test data, we first plotted the frequency of occurrence against the States.

Graph 7: UFO Occurrence across STATES per capita

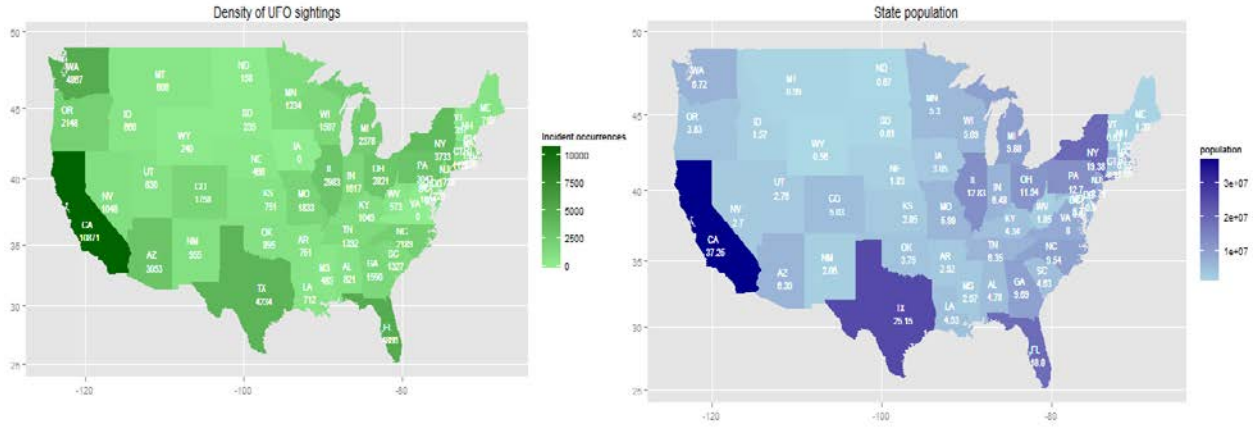


Graph 6: UFO Occurrence across STATES



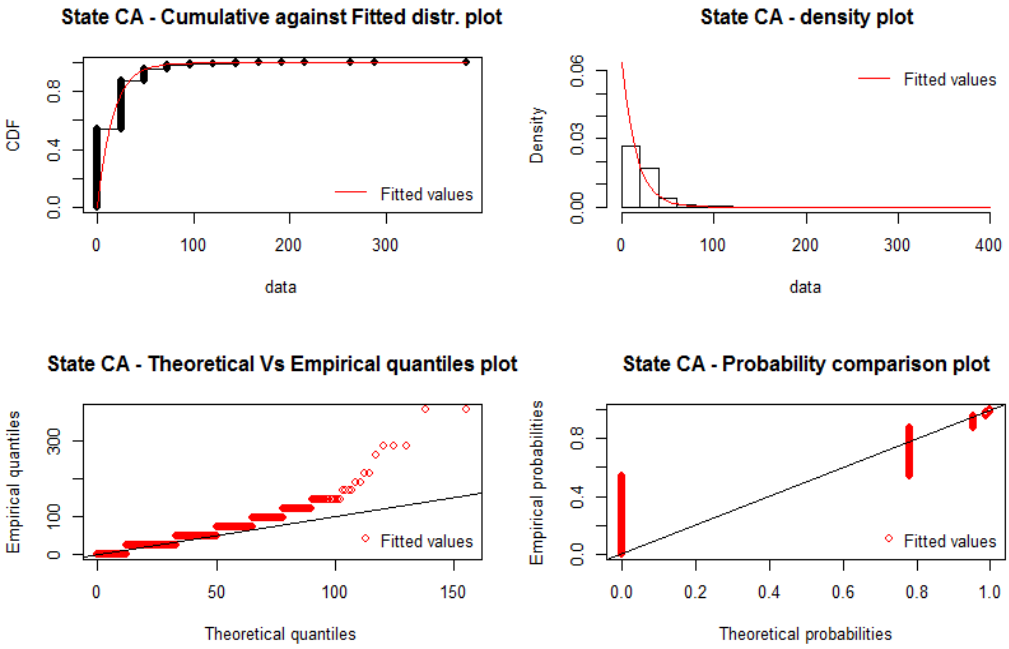
As can be seen from Graph 6, the State of CA has the highest frequency of occurrence of flying object. However, to make the data more palatable, we divided the frequency by the population in the State. This gave us the UFO occurrence per CAPITA. Now the data has changed!!!! Well, FL has the largest occurrence, closely followed by the CA and then the rest. However, group X does not change!!!! The values are still **consistently low**. Surprisingly, RI, ME, HI had sightings/populations. Anyways, we decided to use **CA as our subset TEST data** for our calculations based on the above.

Before we did the per capita, we plotted State population Vs UFO Reporting. Population of a state has a very high impact on the number of UFO reported. Population of US states was taken from https://www.census.gov. Population and UFO incidents are mapped spatially in the below maps using ggplot. States WA and NV slightly contradict this claim – It shows low number of reporting.



Graph 8: State population Vs UFO Reporting

We used this subset to fit the time series to a standard probability distribution. We observed that the time series is closely following an Exponential distribution.



Graph 9: Fitting Distribution - CA

- Fitdist (normal distribution) yields a negative likelihood of **-42482**
- Fitdist (Exponential distribution) yields a negative likelihood of **-35004.**

The larger the better therefore, Exponential distribution is the relatively best distribution that fits the data considered.

By estimating the parameter of the Exponential distribution, we computed the chances of UFO appearing in a state, say in the next 24 hours. Based on the past occurrences, the following states have highest chances that the UFO will appear in the next 24 hours. CA, TX, FL, AZ has the highest chances. 96% chances!!! The rest of the states together just have a 4% chance of a UFO appearing in the next 24 hours. This closely follows the Pareto principle – the 80-20 rule; i.e., large number of reporting is from small fraction of states.

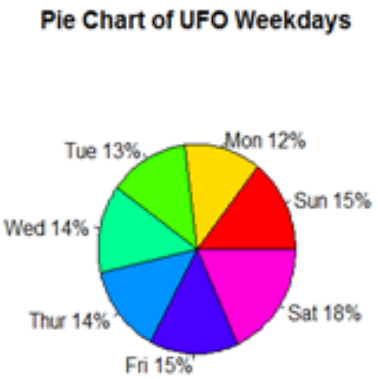
To estimate the probability of event occurring in next 24 hours is given by P(X<= 24) = the cumulative function of exponential distribution = 1-e^(-rate*24) for the State – CA

summary(fitexp)

Fitting of the distribution ' exp ' by maximum likelihood
Parameters :
estimate Std. Error
rate 0.06340195 0.0006567898
Loglikelihood: -35004.44 AIC: 70010.88 BIC: 70018.02
> 1-(exp(-0.06340195*24))

0.781648 (80% probability that an incident can occur in the next 24 hours in state CA)

Since our data contains a lot of **DATE/ DURATION/ TIME factors** which play a vital role in output variable, we wanted to do a time series analysis. We did some exploratory analysis before we did a time analysis. Included below are some of the results of the same.



Graph 10: Highest UFO occurrence based on hours and weekdays

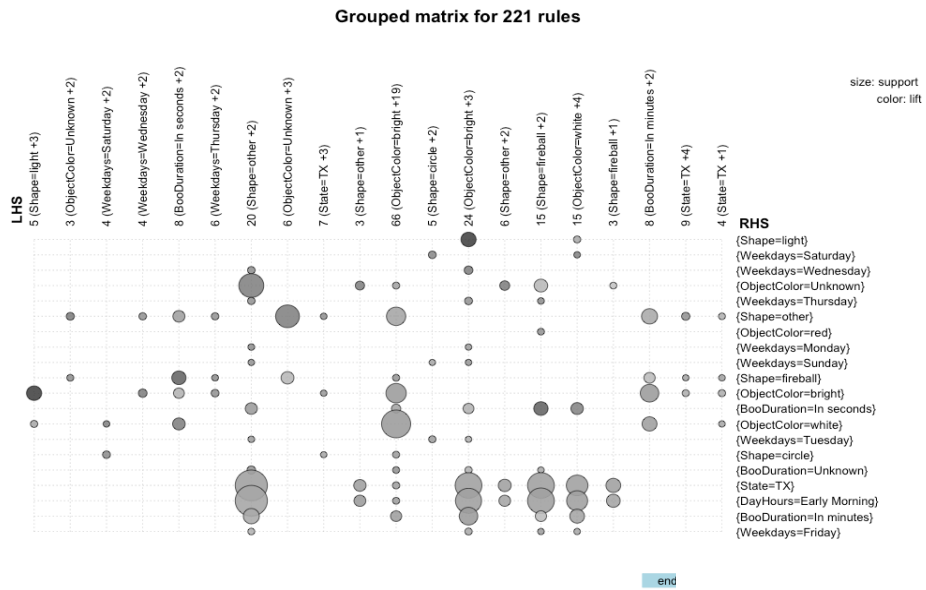
ANALYSIS: ***As can be seen, the hours of the day, makes a difference in the reporting of UFO along with the days of the week. Based on the above, it can be seen that highest reporting hour are the hours 21, 22, 23 (EARLY MORNING with Saturdays as the occurrence day). Though the days of the week does not seem to have as much impact VISUALLY!***

CLUSTERING – KMEANS

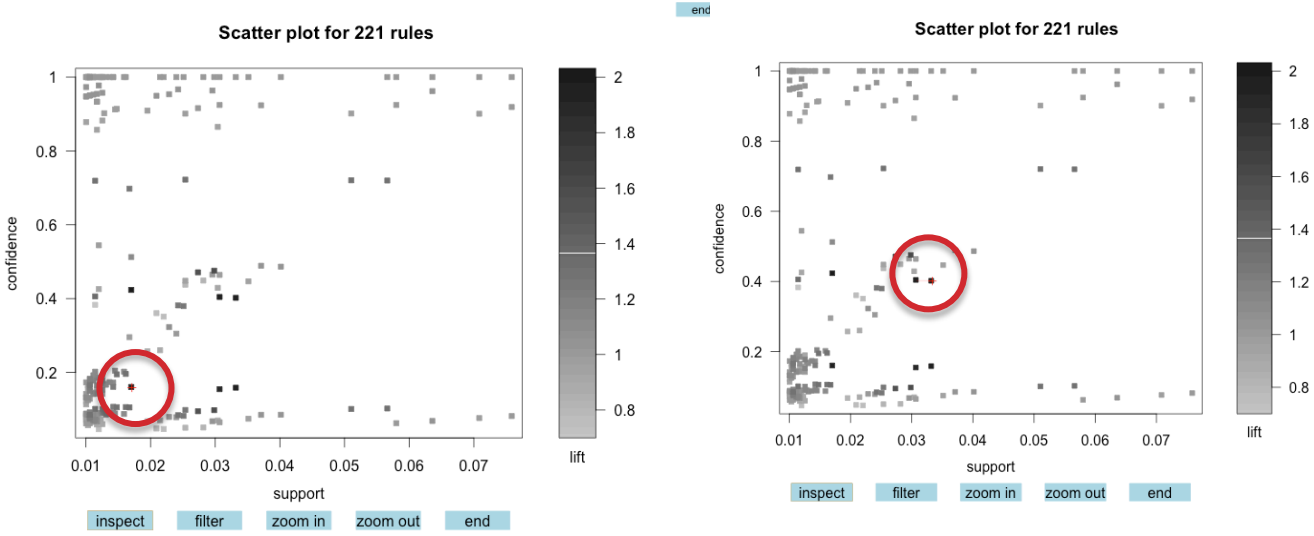
We also did clustering model for the data. However due to the variability in the data, the clusters did not yield a very intuitive result and so we decided to choose the exponential distribution.

Another key unsupervised modelling technique used was Association Mining. We got some very interesting results as shown in GRAPH 11.

Graph 11: Association Mining – Grouped Matrix



- Most UFO occurrence was during early morning between midnight and 6:00am, the color was either bright or white light, while the shape was a fireball.
- When the color was white or bright light, it lasted only for seconds.
- Most occurrences were on weekend, and the least was on Monday.



Graph 12: Association Mining – Interactive Plot

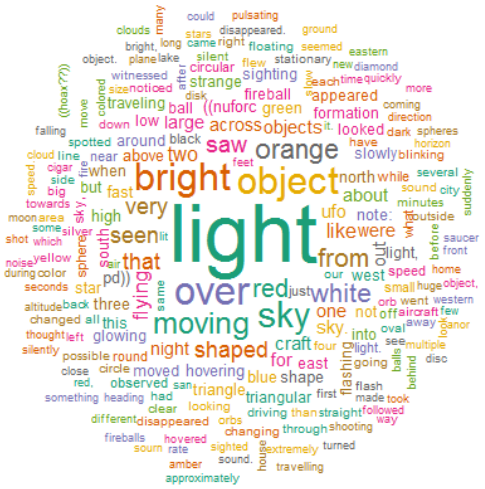
As can been, the results are very similar to the grouped matrix.

lhs	rhs	support	confidence	lift
{Shape=light, BooDuration=In minutes}	{ObjectColor=bright}	0.0170106	0.1605263	1.944755
{Shape=fireball, DayHours=Early Morning	{BooDuration=In seconds}	0.0273285	0.4711538	1.545798

The Final ANALYSIS that we did was Text Mining using the r package on the predictor variable Summary. We found the TOP 25 words that people used as listed below along with the Word Cloud.

No.	Shape	Count	No.	Shape	Count
1	light	31633	14	that	5504
2	over	13584	15	shaped	5394
3	object	13081	16	flying	4918
4	bright	13058	17	like	4716
5	sky	11770	18	two	4407
6	moving	9029	19	ufo	4186
7	orange	7733	20	for	4152
8	white	7173	21	craft	4113
9	from	7153	22	sky.	4018
10	red	6538	23	objects	3887
11	saw	6441	24	across	3594
12	very	5744	25	out	3576
13	seen	5650			

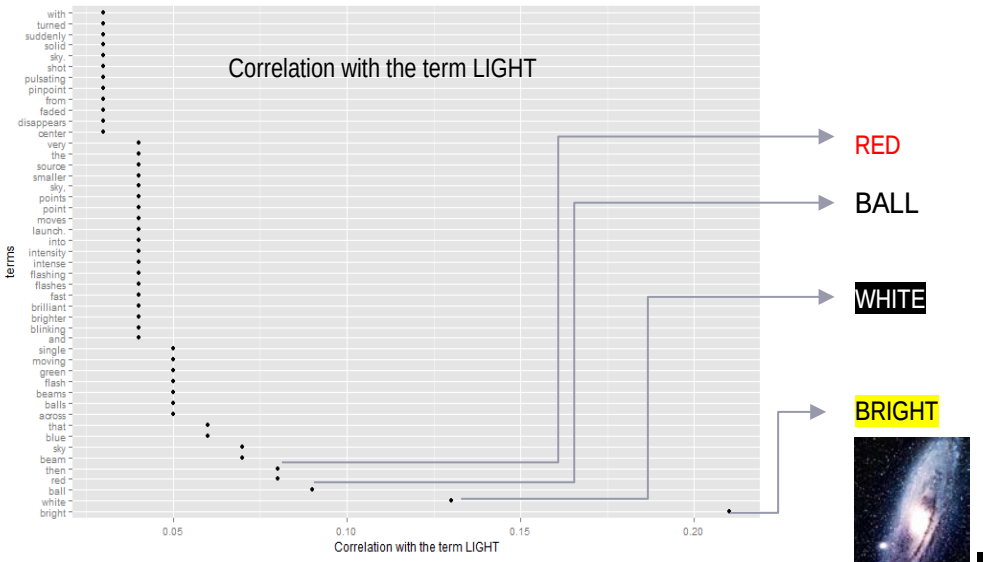
Graph 14: TOP 25 Words USED



Graph 15: Word CLOUD – Corpus – CA

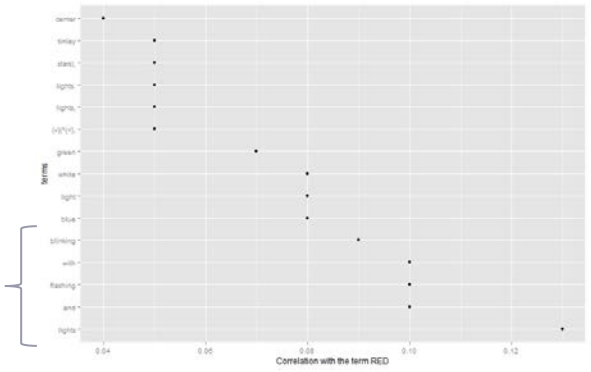
We wanted to find out the association between words to figure would if it would add meaning to other predictors. We have included the results below:

Association matrix

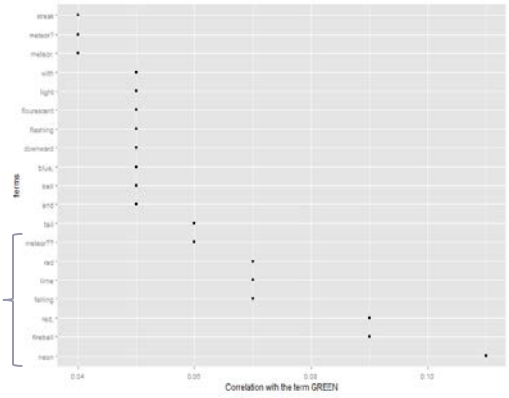


Graph 16: Association with Text Mining

created based on the most frequent terms. Correlation limit varies for each of these terms. For ex: consider the term "light", which tops the list of the most frequently used words. Below is the correlation plot with limit set as 0.03.



Red color was most associated with flashing and blinking lights.



Green color was most associated with neon, fireball, falling light with tail.

Graph 17: Association with Text Mining – RED and Green COLOR

As expected Red Color is associated with Flashing as in Graph 17, it is surprising that Green is associated with NEON, fireball and failing light with tail. This could mean a number of things including airship!!

CONCLUSION

Population can have an impact on the UFO Frequency. Color does have a significant association with lights and sometimes with FLASHIN and BLINKIING. When we are trying to denote a UFO, the most common English terminology used is LIGHT. CA and FL have the largest occurrence of UFO sightings and the chances of it happening again are high. Association between UFO occurrences and time from midnight to 6:00am is very high. Usually during that time, people report seeing UFO's of color was either bright or white light, while the shape was a fireball. Data analysis on the UFO data was using the following data analysis techniques:

- Missing data analysis and imputation
- Data transformation
- Text mining
- Association mining
- Time series distribution
- Spatial and Word cloud visualization
- Other basic plots and histograms

REFERENCESES

Data set - http://www.nuforc.org/webreports.html; Image courtesy - http://www.clipartlord.com/; http://www.openminds.tv/ufo-hotspots-805/12564; http://www2.shutterstock.com/similar-47343178/stock-vector-ufo-ships-flying-in-galaxy.html

FOLLOWING THE FOOTPRINTS OF UFO

ISE 5103 – INTELLIGENT DATA ANALYTICS

BY STUDENT X