

**Department of DATA ANALYTICAL ENGINEERING**  
**George Mason University**

**A**  
**Project Report on**  
**“National UFO Reporting Center (NUFORC)”**

**Submitted by:**

**Vineel Vishwanth Busi**  
**G#01222602**

**Instructor:**

**Dr. Harry Foxwell, PhD.**

## Table of Contents

Deliverable 1 – .....	3
Deliverable 2 - .....	5
Data Analysis .....	5
Visualizations.....	6
Scatterplot.....	6
Boxplot .....	7
Correlation Analysis .....	9
Linear regression analysis .....	10
hypothesis test .....	11
SQL Schema .....	12
Technical Terms.....	14
Limitations.....	14
Carry outs.....	14
References .....	14

# National UFO Reporting Center (NUFORC)

## Deliverable 1 –

### 1. Who (company, agency, organization) collected the data?

a. Who they are, what do they do?

b. What is their role/purpose?

The National UFO Reporting Center (NUFORC) is an organization in USA founded in 1974 by Robert J. Gribble investigates UFO sightings and/or alien contacts. It has reported 90,000 UFO sightings and which are mostly in the United States. NUFORC also keeps records and provides statistics and graphs to assist others who are researching for information. [1] [2]

I collected the Dataset from Kaggle website where this Dataset consisted over 80,000 records of UFO sightings for the last century. The complete data consists entries where the location of sighting was blank with 0.8146% or have errors or blank time with 8.0237%. As this dataset is of 20th century, the older data might be obscured. [1]

The data consists the information of City, State, Time, Description and duration in seconds and hours of each sighting, comments which says what exactly happened during the sightings.

### About Dataset [1] [3]:

Dataset has 88300 records with 11 columns of different data types below-

1. Datetime (datetime): date and time.
2. City (String): Place where UFO is sighted.
3. State (String): City place where UFO is sighted.
4. Country (String): Country where UFO is sighted.
5. Shape (String): Shape of UFO
6. Duration in seconds (Numeric): The UFO Sighting duration in seconds
7. Duration in Hours/Minutes (Numeric): The UFO Sighting duration in hours of minutes.
8. Comments (String): The comments during UFO sighting
9. Date Posted (datetime): Date of post of Sighting
10. Latitude (Numeric): Latitude of the UFO Sighting
11. Longitude (Numeric): Longitude of the UFO Sighting

## **2. Why did they collect the data?**

Data was collected and tracked the unidentified flying objects along with the duration and what was happened with these objects throughout the century. With this recorded data, national UFO Reporting center threatens the hazards caused by these UFO.

## **3. Describe any privacy, quality, ethical, or other issues with this dataset**

There are no privacy concerns as such. As there is huge data, though there are missing columns, the data even after deletion will not affect the quality of the data. The data is collected by receiving the calls from the Hotline since 1974. Every call is recorded and henceforth we could see few missing data.

## **4. What potential value can be obtained by studying this data?**

- **List some specific questions, and plan to answer them in your analysis**
  1. Understand the areas that has more readings of UFO.
  2. Understand the effects caused by this UFO.
  3. What precautions are taken to overcome these abnormalities.
  4. What is the severiarity between the past and present UFOs over the years.

## **5. Resources: What software and hardware resources will you need to study this data?**

I have a plan to use RStudio and Python on an i7 7th generation with 8GB RAM and 256SSD machine.

## **6. Background & prior studies**

- a. **Identify and briefly discuss one or more other similar studies that were done in the domain of your project**

There are many researches conducted and in the year 2015 there was a reported presented about 10 top states with highest number of UFO Sightings. IN this, California has a recorded UFO sightings of 11202 and the least in North Carolina with 2273 UFO Sightings. [4]

## Deliverable 2 -

### Data Analysis

The taken dataset of NUFORC has so much information and also has many missing values and also the data which are not required for the analysis.

So primarily, data cleaning is required and hence to be done to analyze the data perfectly.

The Duration is removed from the dataset and only the duration in seconds column is taken into consideration as it is of containing same units. Also the comment column was removed as there is nothing to do with that information in doing the analysis through forms.

The following are the 10 columns out of 12 columns remained after the analysis [5]:

1. Datetime: the UFO sighting date and time
2. City: The location name where the UFO sighting happened
3. State: The state of the City where the UFO sighting happened
4. Country: The Country of the State/City where the UFO sighting happened
5. Shape: Shape of the UFO appeared
6. Duration (Seconds): The duration in seconds of UFO sighted
7. Date posted: The date when the UFO sightings posted
8. Latitude: The geographical latitude reading of the UFO Sighted.
9. Longitude: The geographical longitude reading of the UFO Sighted.

Later the date at a glance is seen by using the following R Script:

```
Head(ufo_clear)
```

## Visualizations

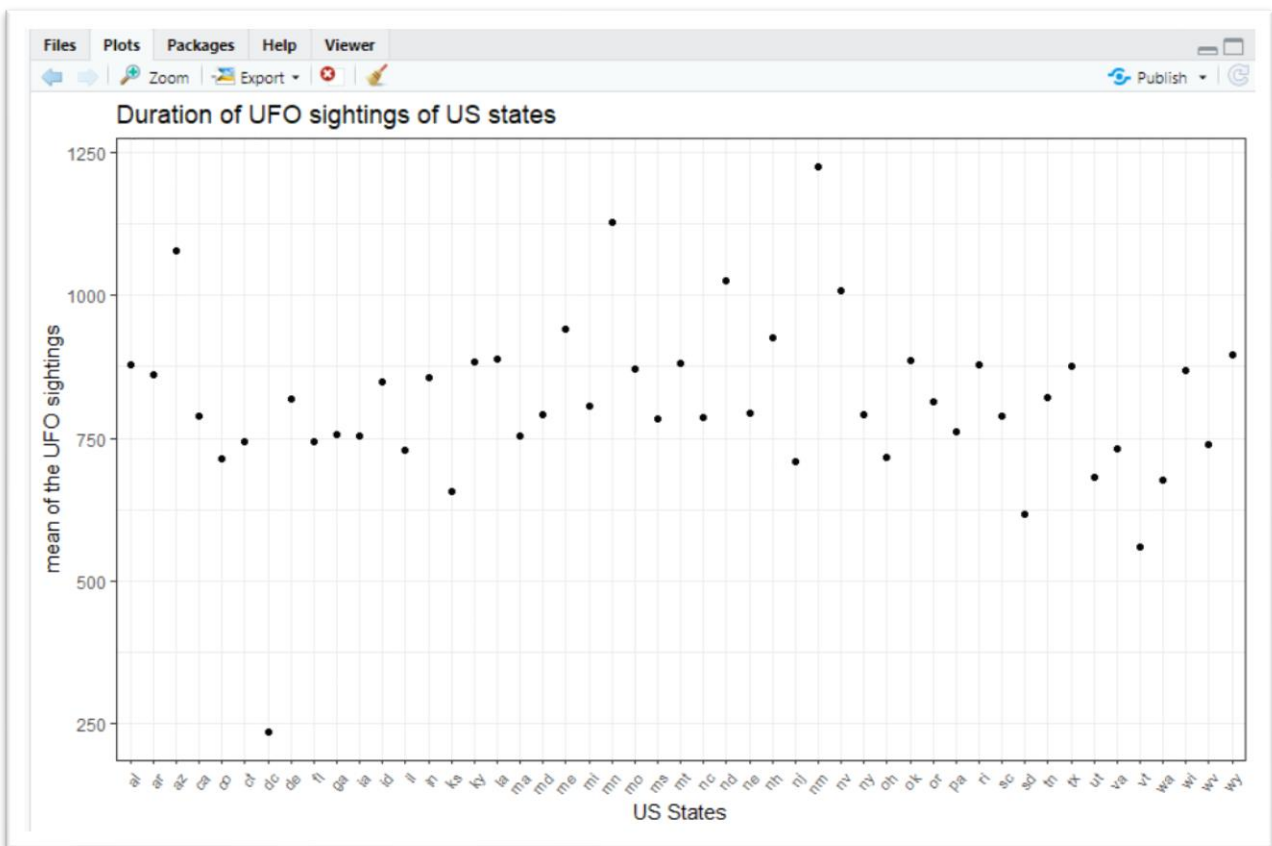
### Scatterplot

To understand the longest sightings of the UFO in USA countrywise, we can use scatter plots this. Before this, the outliers are need to be removed and the duration is calculated only in seconds.

Script:

```
durations_state <- ufo_usa %>%  
  filter(duration < 86400) %>%  
  group_by(state) %>%  
  summarize(mean=mean(duration));  
ggplot(durations_state, aes(x=state, y=mean)) +  
  geom_point() +  
  theme_bw() + theme(axis.text.x = element_text(angle=50, size=8, hjust=1)) +  
  xlab("US States") + ylab("mean of the UFO sightings") + ggtitle("Duration of UFO sightings of US states");
```

Output:



Analysis Conclusion: The graph states that the longest UFO sighting is seen in New Mexico.

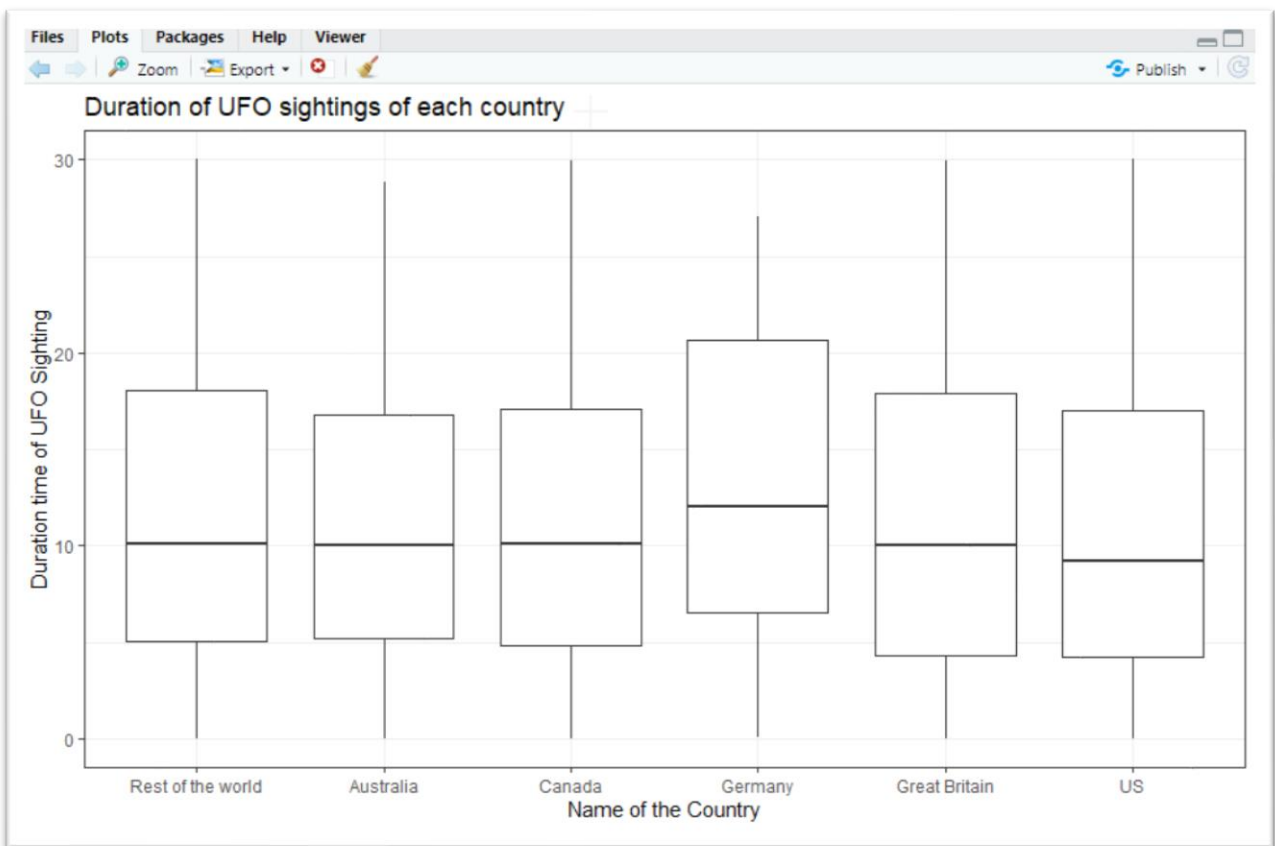
## Boxplot

In order to make a box plot study, the outliers which are existing in the dataset are need to be removed.

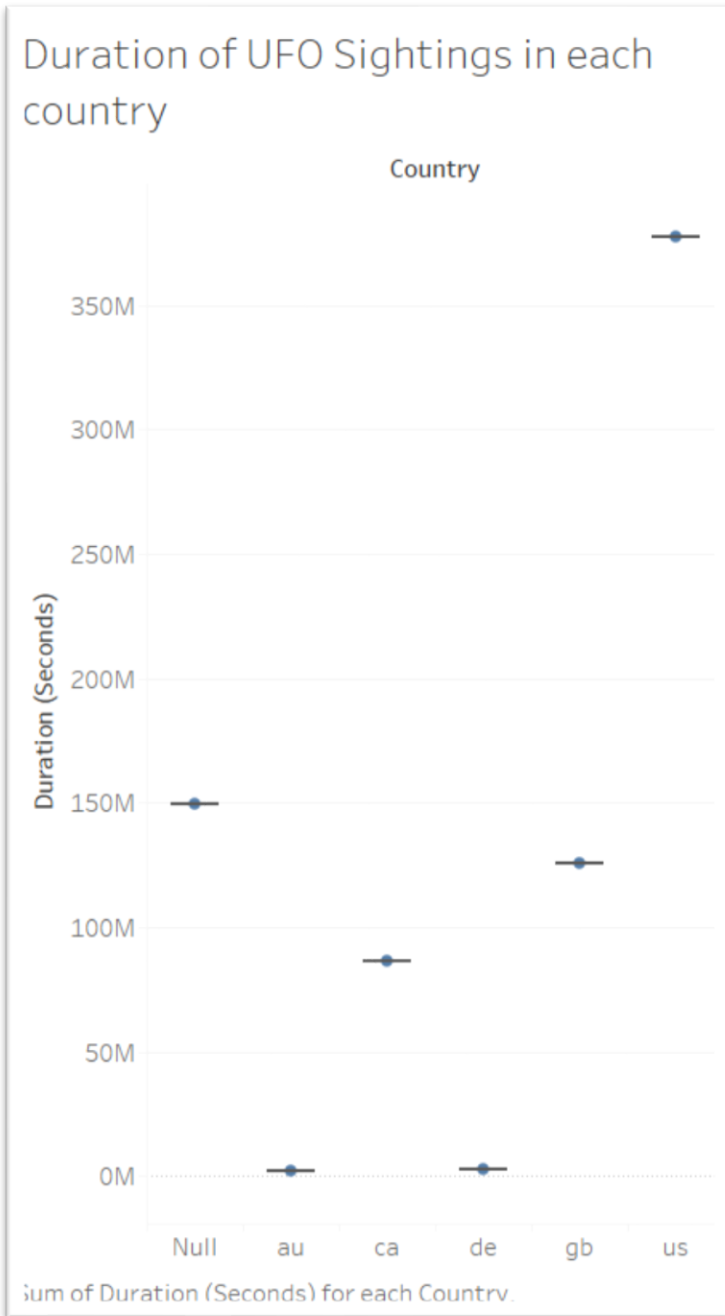
Script:

```
reportedtime_cleared <-  
  report_time %>%  
  filter(duration < 30);  
ggplot(reportedtime_cleared, aes(x=country, y=duration)) +  
  geom_boxplot() +  
  theme_bw() + xlab("Name of the Country") + ylab("Duration time of UFO Sighting") +  
  ggtitle("Duration of UFO sightings of each country");
```

Output:



In Tableau:





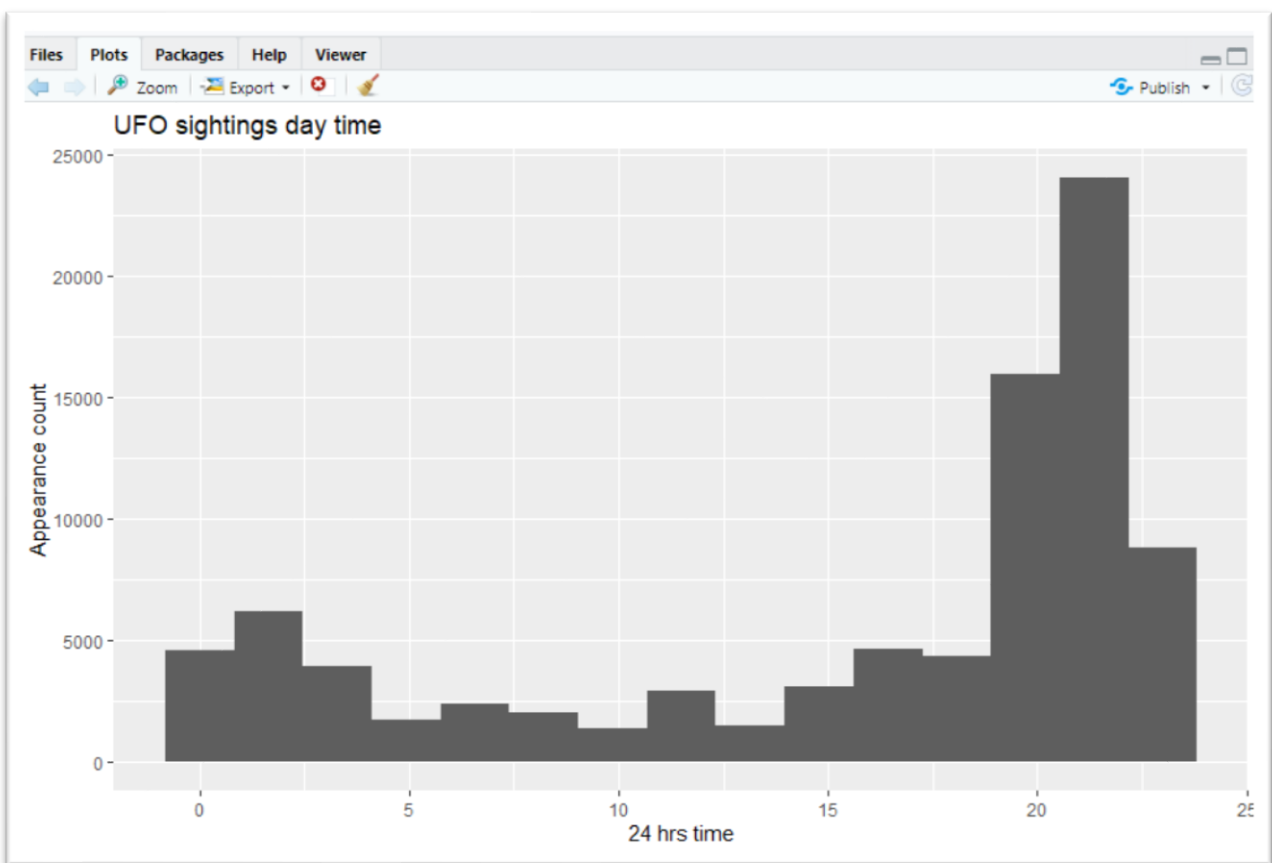
## Correlation Analysis

Inorder to find the correlation, there should be two variables that needs to be compared. Here the comparison is been done between UFO sightings and time of the UFO sighting.

Script:

```
ggplot(ufo_clear, aes(x=hour(datetime))) +  
  geom_histogram(bins=15) +  
  xlab("24 hrs time") + ylab("Appearance count") + ggtitle("UFO sightings day time");
```

Output:



Conclusion: The graph says that the majority of the UFO Sightings are seen mostly at the night. The UFO sightings are also happening in the day light time but are less compared to that of nights.

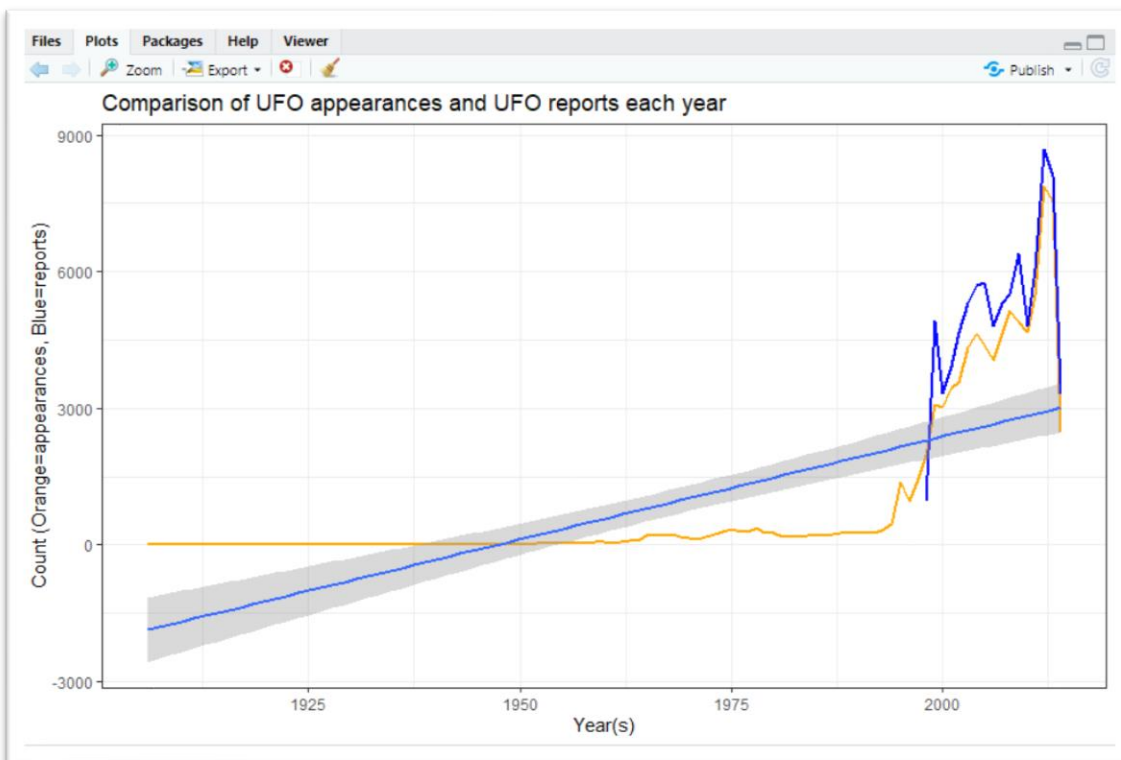
## Linear regression analysis

To find the linear regression, the comparison analysis should happen on – number of UFO Sightings happened against years.

Script:

```
# UFO Sightings every year
# Appearances per year
appearance_year <- ufo_clear %>% group_by(year=year(datetime)) %>%
  summarize(count=n());
# Reports per year
reports_year <- ufo_clear %>% group_by(year=year(date.posted)) %>%
  summarize(count=n());
ggplot(appearance_year, aes(x=year, y=count)) +
  geom_line(size=1, colour="orange") +
  geom_line(data=reports_year, aes(y=count), size=1, colour="blue") +
  geom_smooth(method="lm") +
  theme_bw() + xlab("Year(s)") + ylab("Count (Orange=appearances, Blue=reports)") +
  ggtitle("Comparison of UFO appearances and UFO reports each year")
```

Output:



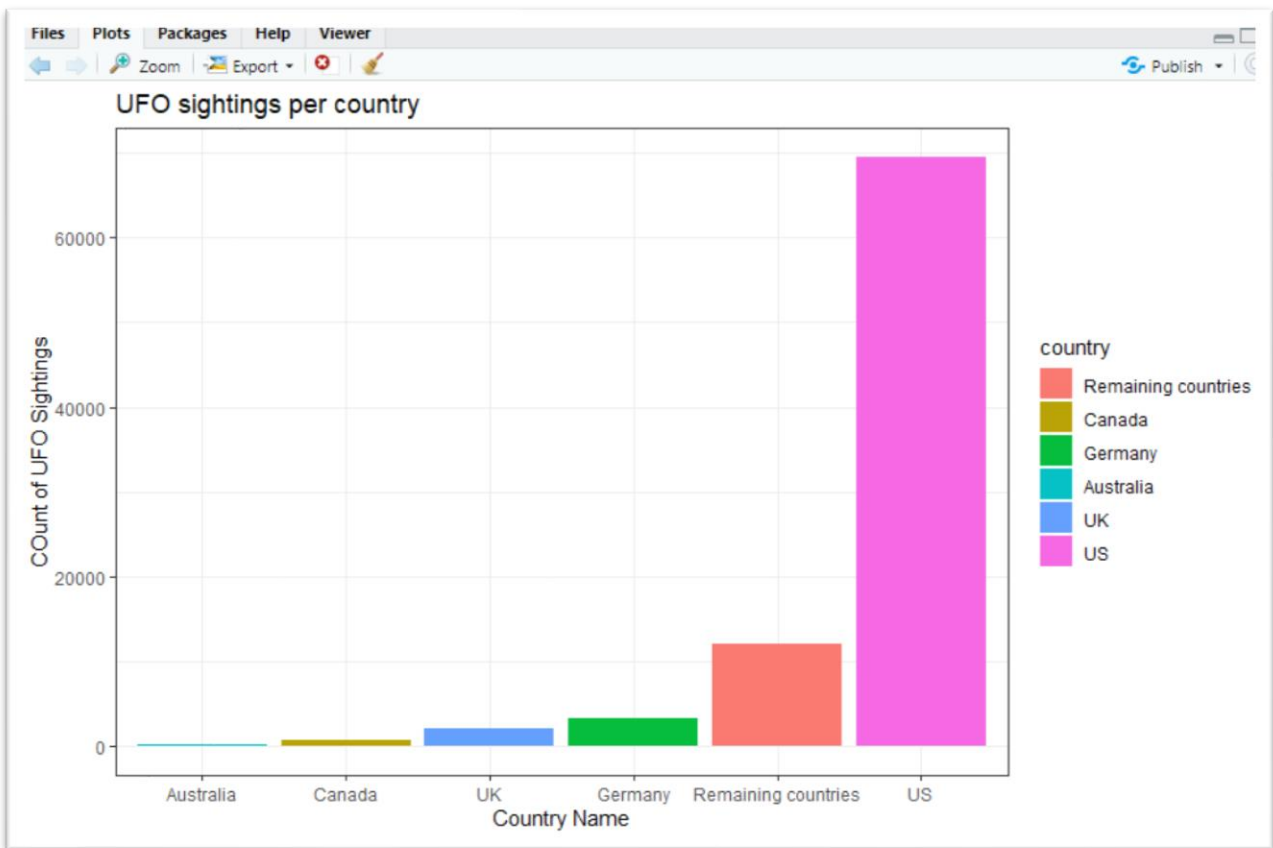
Conclusion Analysis: From the above Linear Regression analysis graph, the UFO sightings are seen everywhere and are increasing year by year.

hypothesis test

Script:

```
levels(ufo_clear$country) <- c("Remaining countries", "Canada", "Germany", "Australia", "UK", "US");  
ggplot(ufo_clear, aes(x=reorder(country, country, FUN=length), fill=country)) +  
  stat_count() +  
  theme_bw() + xlab("Country Name") + ylab("COunt of UFO Sightings") + ggtitle("UFO sightings per  
country");
```

Output:



Analysis Conclusion: From the above Histogram, it is clear that US is the country with huge number of UFO Sightings seen from all these years. There are also other countries which also experienced UFO sightings and the graph is explaining the UFO sightings of other countries too.

## SQL Schema

SQL schema for the data, and demonstrate several basic SQL-based queries of the dataset

### Step 1: Creation of Table:

```
Create TABLE ufodata
(
  datetime varchar(255),
  city varchar(255),
  state varchar(255),
  country varchar(255),
  shape varchar(255),
  duration_seconds int,
  duration_hours_min varchar(255),
  comments varchar(255),
  dateposted varchar(255),
  latitude int,
  longitude int
);
```

### Outcome:

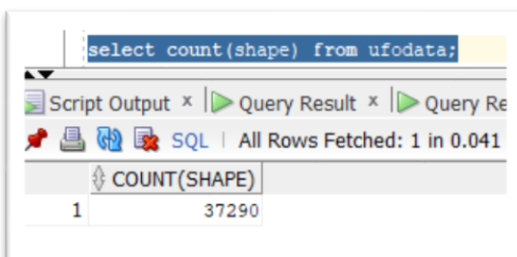
Table ufodata crated.

Step 2: Imported the National UFO Reporting Center dataset into the Table.

Step 3: Some basic SQL queries for the National UFO Reporting Center are as below:

Query: select count(shape) from ufodata;

Output:

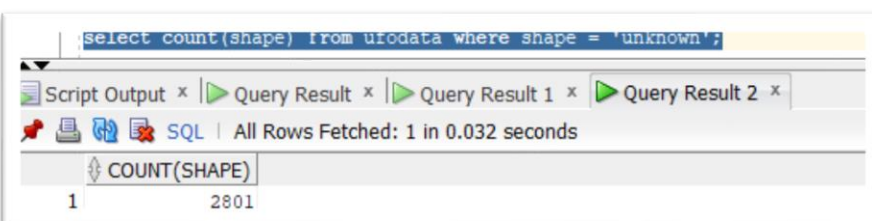


The screenshot shows a SQL query window with the query 'select count(shape) from ufodata;'. Below the query, there are tabs for 'Script Output', 'Query Result', and 'Query Re'. The 'Query Result' tab is active, showing a table with one row and one column labeled 'COUNT(SHAPE)'. The value in the row is 37290.

COUNT(SHAPE)
37290

Query: select count(shape) from ufodata where shape = 'unknown';

Output:

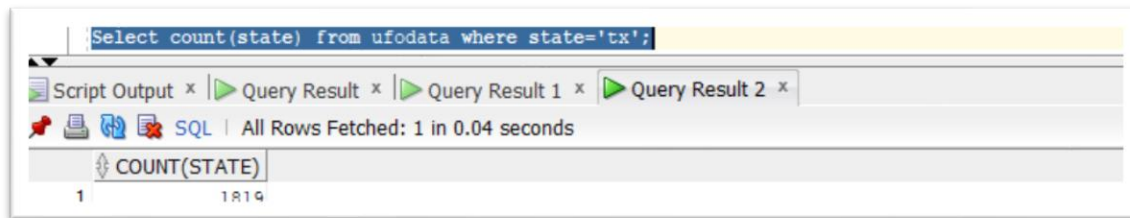


The screenshot shows a SQL query window with the query 'select count(shape) from ufodata where shape = 'unknown';'. Below the query, there are tabs for 'Script Output', 'Query Result', 'Query Result 1', and 'Query Result 2'. The 'Query Result' tab is active, showing a table with one row and one column labeled 'COUNT(SHAPE)'. The value in the row is 2801.

COUNT(SHAPE)
2801

Query: Select count(state) from ufodata where state='tx';

Output:

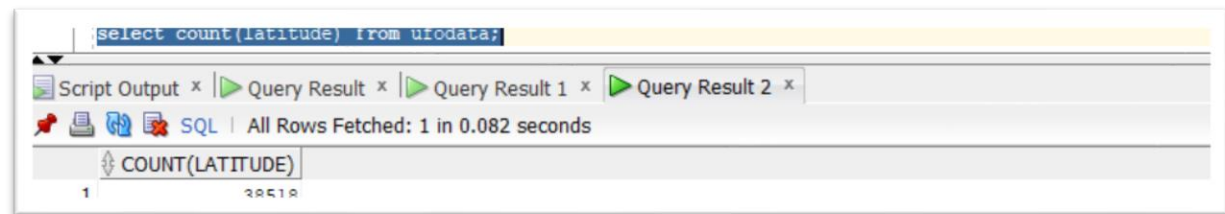


The screenshot shows a SQL query window with the query: `Select count(state) from ufodata where state='tx';`. The results pane shows a single row with the column header `COUNT(STATE)` and the value `1819`. The status bar indicates "All Rows Fetched: 1 in 0.04 seconds".

COUNT(STATE)
1819

Query: select count(latitude) from ufodata;

Output:

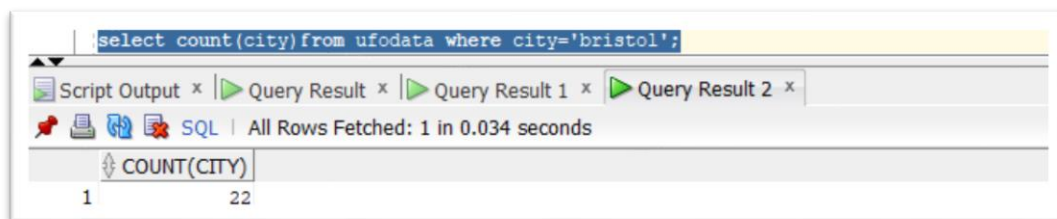


The screenshot shows a SQL query window with the query: `select count(latitude) from ufodata;`. The results pane shows a single row with the column header `COUNT(LATITUDE)` and the value `38518`. The status bar indicates "All Rows Fetched: 1 in 0.082 seconds".

COUNT(LATITUDE)
38518

Query: select count(city)from ufodata where city='bristol';

Output:

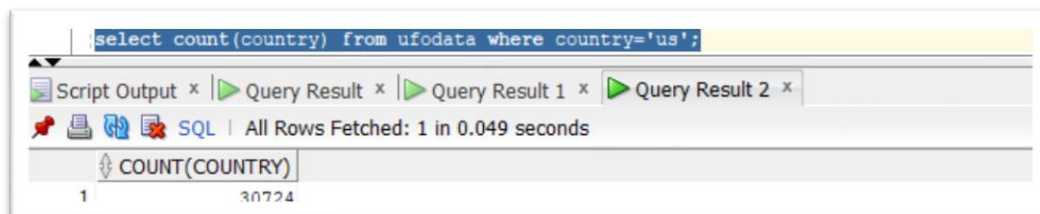


The screenshot shows a SQL query window with the query: `select count(city)from ufodata where city='bristol';`. The results pane shows a single row with the column header `COUNT(CITY)` and the value `22`. The status bar indicates "All Rows Fetched: 1 in 0.034 seconds".

COUNT(CITY)
22

Query: select count(country) from ufodata where country='us';

Output:



The screenshot shows a SQL query window with the query: `select count(country) from ufodata where country='us';`. The results pane shows a single row with the column header `COUNT(COUNTRY)` and the value `30724`. The status bar indicates "All Rows Fetched: 1 in 0.049 seconds".

COUNT(COUNTRY)
30724

## Technical Terms

**Ggplot:** This is data visualization package used for statistical programs in the Rstudio tool.

**Regression:** This is the relation between two variables which will give the statistical graphical representation with linear, Multiple and Logistic relationship.

**SQL Schema:** A schema is the collection of DB objects associated with one DB.

**Correlation:** This is the mutual relationship between two variables

**Scatter Plot:** This is used to present any correlation between two variables plotted.

**Box Plot:** This is the graphical representation of the data with Minimum, maximum, first quarter, medium and third quarter.

## Limitations

The dataset which has been taken is very huge and the data has only one column of Numeric data which made only fewer points on some graphs. Also the data is not completely full as it looks like and needs very much time to clean the data. But yes, this dataset project helped me to learn many things.

## Carry outs

By doing this project exercise, I have found that we can work so much on any given dataset and thus extract much more information and also can be able to predict which is to be good grade. As I have worked on tools like RStudio, Tableau, Python and SQL, I learnt so much and still found that there are many concepts where I can practice them so that I can get the wanted information in whatever the way the requirement is.

## References

- [1] NUFORC, Kaggle, 16 11 2016. [Online]. Available: <https://www.kaggle.com/NUFORC/ufo-sightings>. [Accessed 9 11 2019].
- [2] "National UFO Reporting Center," Wikipedia, [Online]. Available: [https://en.wikipedia.org/wiki/National\\_UFO\\_Reporting\\_Center](https://en.wikipedia.org/wiki/National_UFO_Reporting_Center). [Accessed 9 11 2019].
- [3] N. Donges, "Towards Datascience," 18 March 2018. [Online]. Available: <https://towardsdatascience.com/data-types-in-statistics-347e152e8bee>. [Accessed 9 11 2019].
- [4] ABC News, 2 Apr 2015. [Online]. Available: <https://abcnews.go.com/US/ufo-sightings-top-10-states-highest-number-reports/story?id=30061767>. [Accessed 9 11 2019].
- [5] G. R. K. Hugo Frezat, "Analysis of UFO Sightings Across the World," rpubs, 19 1 2018. [Online]. Available: [https://rpubs.com/ganapathy\\_ram/ufo\\_analysis](https://rpubs.com/ganapathy_ram/ufo_analysis). [Accessed 25 11 2019].

[6] THE NATIONAL UFO REPORTING CENTER , 4 Oct 2019. [Online]. Available: <http://www.nuforc.org/>.  
[Accessed 9 11 2019].