The impact of the Air Pollution on UFO Sightings

1 Introduction

Premises that have become the basis of the problematic and work of study in this publication, were authors interest in data science, microeconomics and unsupervised learning and also interest in paranormal activities. The paper maintains interdisciplinary market character combining typically economics and data analysis issues with psychology, sociology and behavioral anthropology.

The aim of our work is to present in a theoretical and practical aspect of issues related to air pollution and its reflection on the society kognitive abilities as well as a correlation between the occurence of specific air pollutants and UFO sightings.

Having in mind the purpose of this research, the main hypothesis was formulated, to which the layout and structure of the paper were subordinated. The work has been divided into two parts: theoretical, about the influence of air pollutants to people cognition, and empirical, where data analysis have been conducted with various methods and subordinated conclusion.

Hence, the main idea to conduct such a research, was to check wether it's possible to gather some interesting insight compiling and analysing the two data sets:

- *First about the air pollution with respect to latitude and longtitude of observation*
- *Second is about ufo sights with respect to latitude and longtitude of observation*

Where we were trying to find some correlation between specific air pollutions and UFO sightings, if it true matters in this case? or maybe not?

The first chapter of the work focuses on the review of methods used in the field of data analysis. The considerations will mainly focus on:

- *Spatial data analysis*
- *Unsupervised learning methods*
- *non parametric statistical methods*
- *correlation and linear regression*

The second chapter of the work focuses on summary of the data about UFO sightings, possible explanation of influcence of the air pollutant to people congitive abilities and other scientific theories explaining UFO sightings.

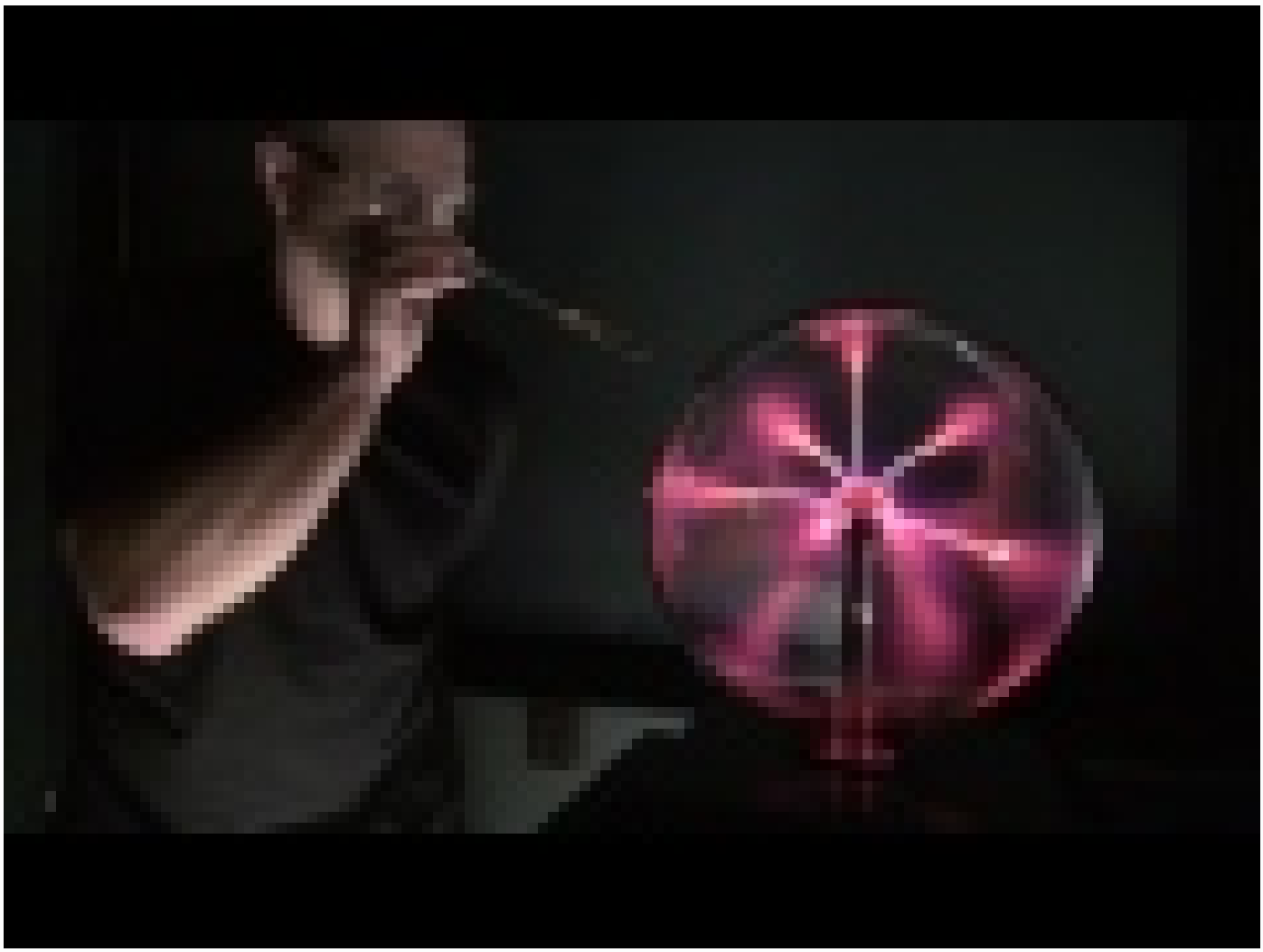
Last chapter is pure data analysis of the obtained data set, to check our assumptions and hypothesis.

This research is based on the idea of Ig Nobel prize, and was conducted with the same ambience.

1.1 Ig Nobel Prize

The Ig Nobel Prizes honor achievements that make people LAUGH, and then THINK. The prizes are intended to celebrate the unusual, honor the imaginative — and spur people's interest in science, medicine, and technology. Every September, in a gala ceremony in Harvard's Sanders Theatre, 1100 splendidly eccentric spectators watch the new winners step forward to accept their Prizes. These are physically handed out by genuinely bemused genuine Nobel Laureates. Thousands more, around the world, watch our live online broadcast

genuine Nobel Laureates. Thousands more, around the world, watch our live online broadcast



Ig Nobel Prize winners in Economics (from last 8 years):

- *Lindie Hanyu Liang, Douglas Brown, Huiwen Lian, Samuel Hanig, D. Lance Ferris, and Lisa Keeping, for investigating whether it is effective for employees to use Voodoo dolls to retaliate against abusive bosses.*
- *Matthew Rockloff and Nancy Greer, for their experiments to see how contact with a live crocodile affects a person's willingness to gamble.*
- *Mark Avis and colleagues, for assessing the perceived personalities of rocks, from a sales and marketing perspective.*
- *The Bangkok Metropolitan Police, for offering to pay policemen extra cash if the policemen refuse to take bribes.*
- *ISTAT — the Italian government's National Institute of Statistics, for including revenue from illegal drug sales, prostitution, smuggling, etc., in GDP reporting, in order to meet an EU regulatory mandate.*
- *The executives and directors of Goldman Sachs, AIG, Lehman Brothers, Bear Stearns, Merrill Lynch, and Magnetar Capital for creating and promoting new ways to invest money—ways that maximize financial gain and minimize financial risk for the world economy, or for a portion thereof.*

2 Characteristics of basic concepts

2.1 Data Analysis Methods

In this chapter we will try to briefly explain the main methods used in this publications. Brace yourself for huge dose of informations!

First of all we will explain briefly about Unsupervised learning methods we will use all along this paper.

The first method used is Cluster Analysis. Cluster analysis is one of the Unsupervised Learning methods where we try to group our observations with respect to similarities among them. The short definition about Cluster analysis looks as follows: ***clustering*** *is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.*

When it comes to understanding these cluster modeling we can divide it into multiple submodels:

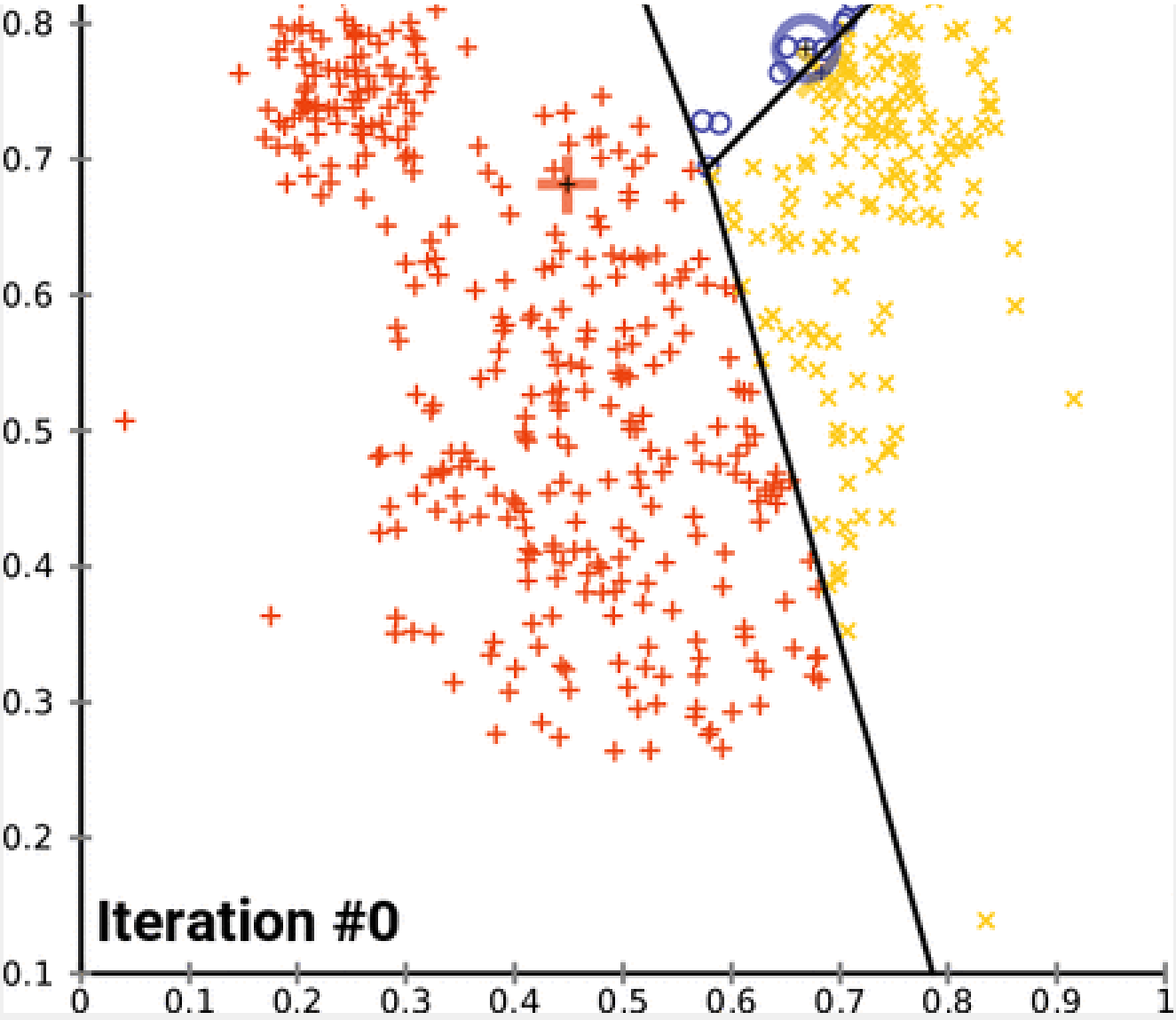
- *Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.*
- *Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.*
- *Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the expectation-maximization algorithm.*
- *Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.*
- *Subspace models: in biclustering (also known as co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.*
- *Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.*
- *Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.*
- *Neural models: the most well known unsupervised neural network is the self-organizing map and these models can usually be characterized as similar to one or more of the above models, and including subspace models when neural networks implement a form of Principal Component Analysis or Independent Component Analysis.*

However in this paper we will focus mainly on Centroid models, especially on CLARA and K-means.

2.1.1 K-Means

K-Means method is one of the most important Clustering methods *where we randomly choose best suited centroid. Then around it, we connect most similiar data points and build our cluster. In Rstudio, to perform K-means clustering we use* `library(facoextra)` *, and function* `eclust(data,"kmeans")` *. If you want to know more about K-means method, you can read it more here. K-means and CLARA analysis is also a part of one of my previous work, if you want to get know about it check out this link.*

Simplified Animated K-Means Clustering:



2.1.2 CLARA

***CLARA (Clustering Large Applications**, is an extension to k-medoids (PAM) methods to deal with data containing a large number of objects (more than several thousand observations) in order to reduce computing time and RAM storage problem. This is achieved using the sampling approach.*

Instead of finding medoids for the entire data set, CLARA considers a small sample of the data with fixed size (sampsize) and applies the PAM algorithm to generate an optimal set of medoids for the sample. The quality of resulting medoids is measured by the average dissimilarity between every object in the entire data set and the medoid of its cluster, defined as the cost function.

CLARA repeats the sampling and clustering processes a pre-specified number of times in order to minimize the sampling bias. The final clustering results correspond to the set of medoids with the minimal cost. The CLARA algorithm is summarized in the next section. More about CLARA here.

Simplified CLARA/PAM animation:

CLARA

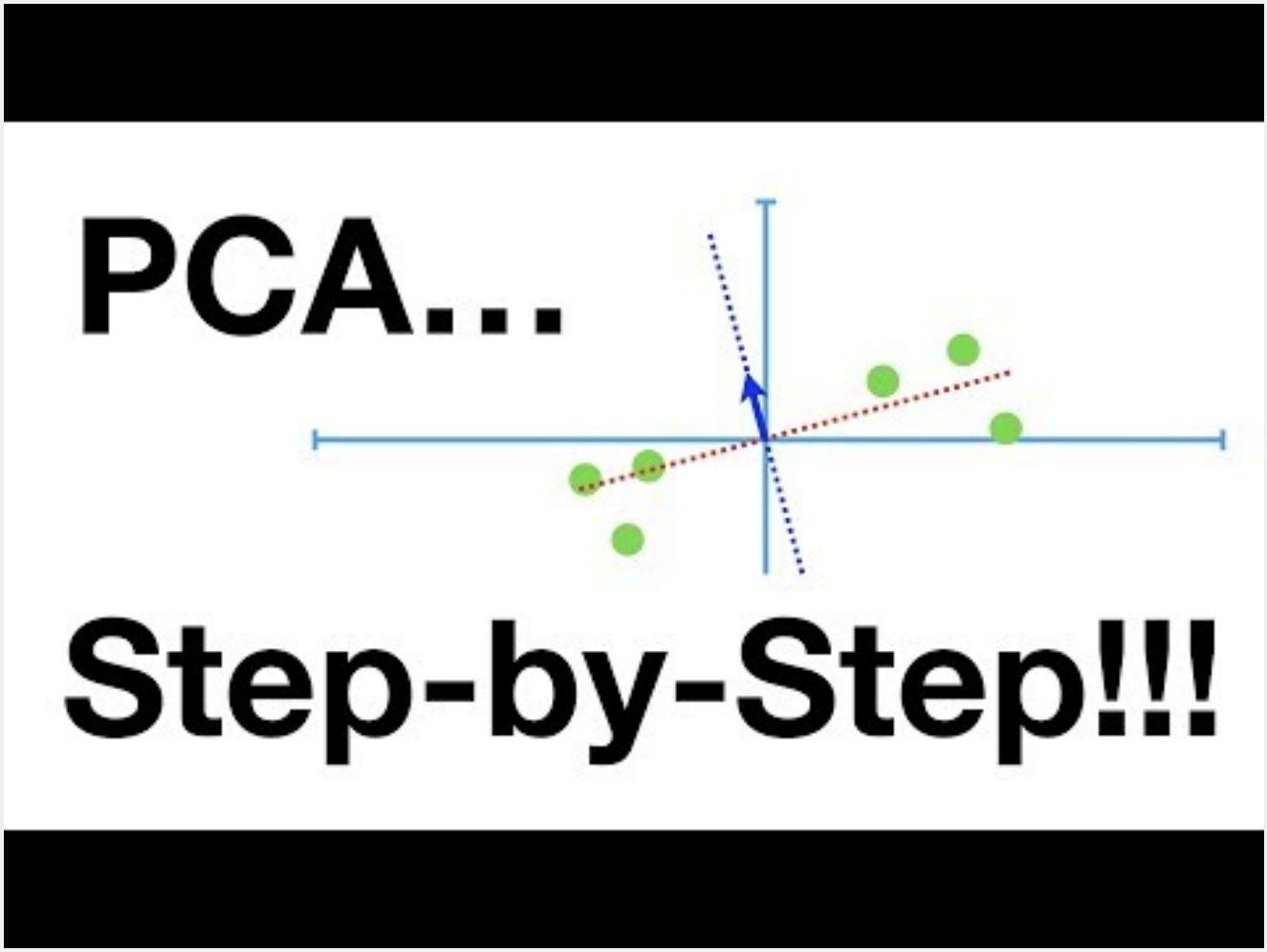
In our case CLARA can be really helpful because we have almost 16 thousand observations, for PAM it would take long to compute.

2.1.3 Principal Component Analysis

2.1.3 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method of factor analysis. We can describe our data set as a cloud of N points in K dimensions, where N is number of observations and K are variables. Goal of PCA is to fit the coordinate system in such a way that first we will maximize the variance of first coordinate, then we maximize it for second etc etc. PCA is widely used to reduce the dimensions of our data set, dropping the least affective ones. When performing the PCA analysis we assignt to each Principal Component some variables that can describe this component best, our goal is to set maximum of 3 Principal Components, because our visual perception can handle only 3D plots.

If you want to know more about PCA, you can read something here, or if you are interested how to perform PCA from scratch, I suggest to watch this Video:

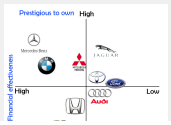


Once again this part is also a part of my previous work, if you want to find more about it check out this link

2.1.4 Multidimensional Scaling

***Multidimensional Scaling (MDS)** is a statistical method to get insights about hidden variables which describes relations, and similiarities among analysed objects. When performing MDS, first of all we have to make a matrix containing distance between values, we can use for this purpose correlation matrix. Scaling drifts get **n** similiar data points which reflets to one dimension in kartesian coordinate system. If n<=3 we can visualise data (up to 3D). If you want to learn more about MDS, check this link.*

Below some simple Multidimensional Scaling graph in the field of Marketing:

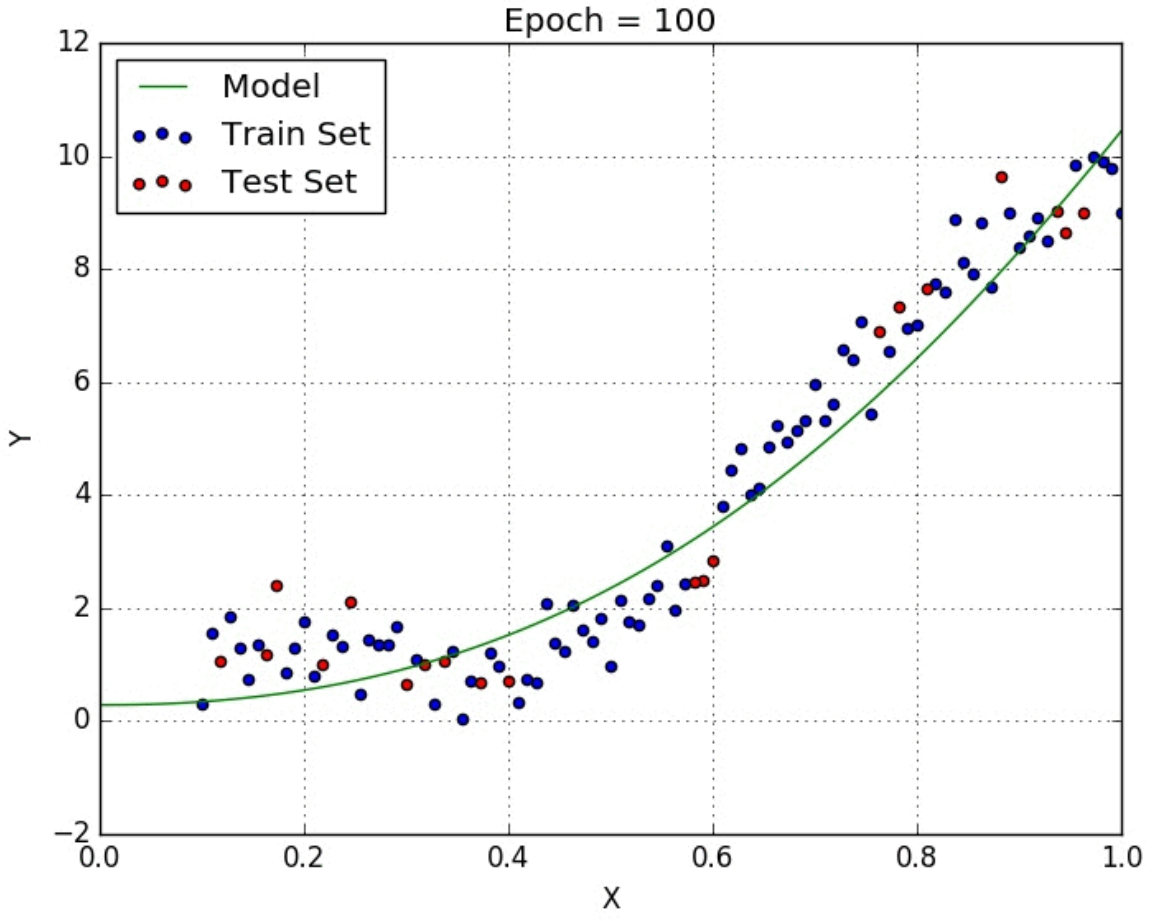


Marketing
MDS

2.1.5 Regression Analysis

I suppose, all of us knows what is Regression, but if not, here we have some definition, In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Simple Regression Animation



2.1.6 Correlation Analysis

In correlation analysis, we estimate a sample correlation coefficient, more specifically the Pearson Product Moment correlation coefficient.

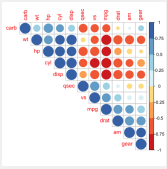
The sample correlation coefficient, denoted r, ranges between -1 and +1 and quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).

The sign of the correlation coefficient indicates the direction of the association. The magnitude of the correlation

coefficient indicates the strength of the association.

For example, a correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

Example of Correlation Analysis in R



Correlation in R

2.1.7 Nonparametric statistics

Nonparametric statistics is the branch of statistics that is not based solely on parametrized families of probability distributions (common examples of parameters are the mean and variance). Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. Nonparametric statistics includes both descriptive statistics and statistical inference

The first meaning of nonparametric covers techniques that do not rely on data belonging to any particular parametric family of probability distributions.

The second meaning of non-parametric covers techniques that do not assume that the structure of a model is fixed. Typically, the model grows in size to accommodate the complexity of the data. In these techniques, individual variables are typically assumed to belong to parametric distributions, and assumptions about the types of connections among variables are also made.

2.1.8 ANOVA

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVA is useful for comparing (testing) three or more group means for statistical significance. It is conceptually similar to multiple two-sample t-tests, but is more conservative, resulting in fewer type I errors, and is therefore suited to a wide range of practical problems.

In our case we have used ANOVA to analyse the table of deviance and decide which variables to exclude from our model.

2.1.9 Logistic Regression

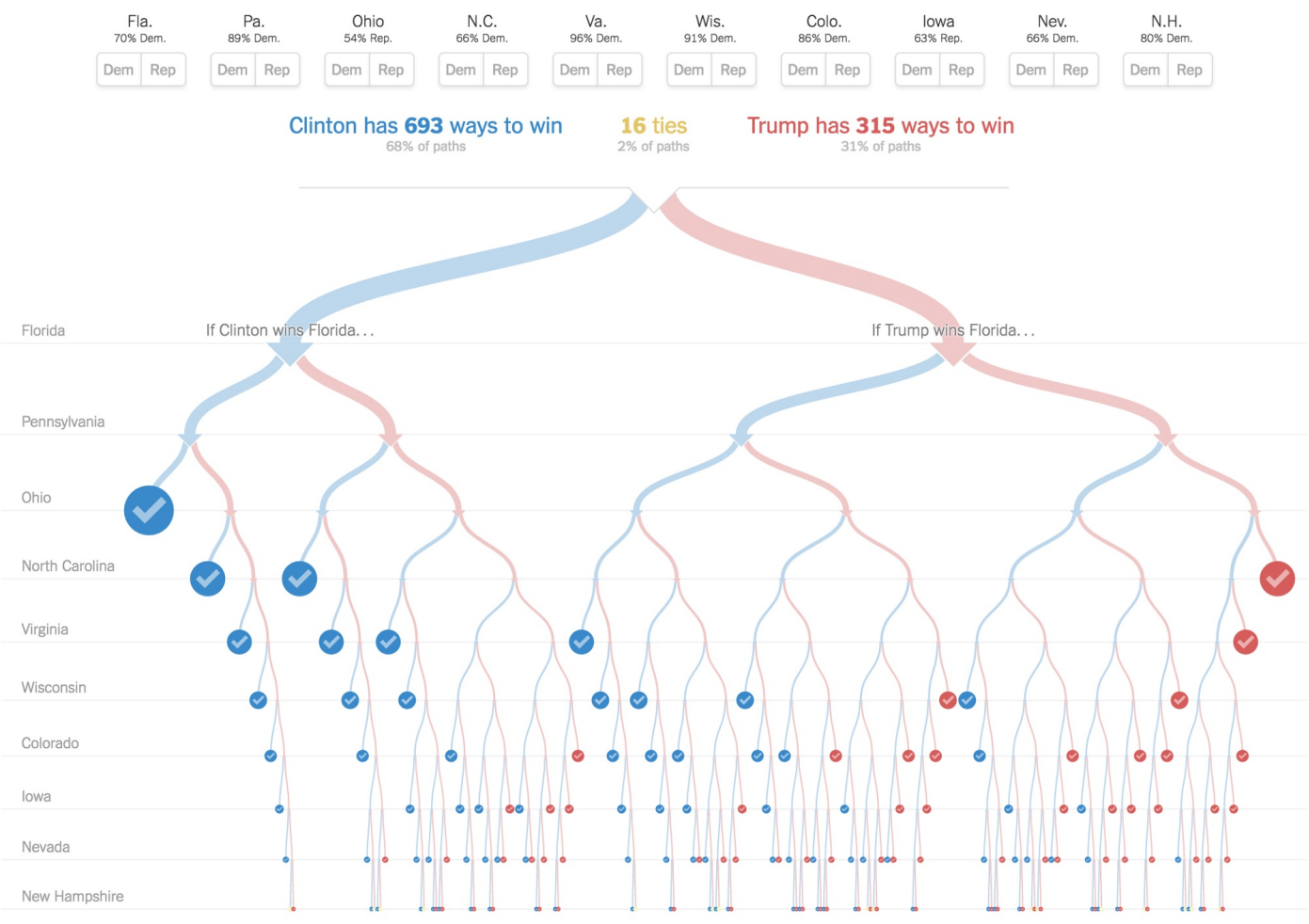
In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

> *In our case we have used Logit regression to check our ET sightings where 0 was no sightings at all and 1 was successful sighting.*

2.1.10 Decision Tree

> *A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.*

> *Example of using Decision Tree to check about prorable outcomes*



> *In our case we have used Decision Tree method to check probable paths that leads to more frequent occurences of UFO.*

3 Analysis of UFO Occurences Analysis and impact of Air Pollution to cognitive performance

3.1 Analysis of historical data of UFO Occurences

3.2 Possible explanations of UFO sightings

> ***This fragment is entirely taken from this site to take closer look on the paranormal activities and possible explanations!!!.***

> *From 1947 to 1970, the United States Air Force conducted investigations into the increasing number of unidentified flying object (UFO) sightings throughout the United States. The purpose of the investigations was to assess the nature of these*

sightings and determine if they posed any potential threat to the U.S.

Three successive projects were created to carry out these investigations: Sign, Grudge, and Blue Book.

Blue Book was the longest and most comprehensive, lasting from 1952 to 1970. A 1966 Air Force publication gave insight into how the program was conducted:

The program is conducted in three phases. The first phase includes receipt of UFO reports and initial investigation of the reports. The Air Force base nearest the location of a reported sighting is charged with the responsibility of investigating the sighting and forwarding the information to the Project Blue Book Office at Wright-Patterson Air Force Base, Ohio

If the initial investigation does not reveal a positive identification or explanation, a second phase of more intensive analysis is conducted by the Project Blue Book Office. Each case is objectively and scientifically analyzed, and, if necessary, all of the scientific facilities available to the Air Force can be used to assist in arriving at an identification or explanation. All personnel associated with the investigation, analysis, and evaluation efforts of the project view each report with a scientific approach and an open mind

The third phase of the program is dissemination of information concerning UFO sightings, evaluations, and statistics. This is accomplished by the Secretary of the Air Force, Office of Information.

After investigating a case, the Air Force placed it into one of three categories: Identified, Insufficient Data, or

Unidentified.

Identified reports are those for which sufficient specific information has been accumulated and evaluated to permit a positive identification or explanation of the object.

Reports categorized as Insufficient Data are those for which one or more elements of information essential for evaluation are missing. Some examples are the omission of the duration of the sighting, date, time, location, position in the sky, weather conditions, and the manner of appearance or disappearance. If an element is missing and there is an indication that the sighting may be of a security, scientific, technical, or public interest value, the Project Blue Book Office conducts an additional investigation and every attempt is made to obtain the information necessary for identification. However, in some instances, essential information cannot be obtained, and no further action can be taken.

The third and by far the smallest group of evaluations is categorized as Unidentified. A sighting is considered unidentified when a report apparently contains all pertinent data necessary to suggest a valid hypothesis concerning the cause or explanation of the report but the description of the object or its motion cannot be correlated with any known object or phenomena.

Sightings resulting from identifiable causes fall into several broad categories:

- *human-created objects or phenomena including aircraft, balloons, satellites, searchlights, and flares;*
- *astronomical phenomena, including meteors and meteorites, comets, and stars;*
- *atmospheric effects, including clouds and assorted light phenomena; and*
- *human psychology, including not only psychological frailty or illness but also fabrication (i.e., hoaxes).*

The conclusions of Project Blue Book were:

1. *no unidentified flying object reported, investigated, and evaluated by the Air Force has ever given any indication of threat to our national security;*
2. *there has been no evidence submitted to or discovered by the Air Force that sightings categorized as unidentified represent technological developments or principles beyond the range of present day scientific knowledge; and*
3. *there has been no evidence indicating that sightings categorized as unidentified are extraterrestrial vehicles.*

In 1967, the Air Force's Foreign Technology Division (FTD), the organization overseeing Blue Book, briefed USAF Gen. William C. Garland on the project. The July 7 report stated that in the 20 years the FTD had reported and examined over 11,000 UFO sightings, they had no evidence that UFOs posed any threat to national security. Furthermore, their evidence "denies the existence of flying saucers from outer space, or any similar phenomenon popularly associated with UFOs."

The FTD reiterated an expanded finding from Project Grudge: "Evaluations of reports of UFOs to date demonstrate that these flying objects constitute no threat to the security of the United States. They also concluded that reports of UFOs were the result of misinterpretations of conventional objects, a mild form of mass hysteria of war nerves and individuals who fabricate such reports to perpetrate a hoax or to seek publicity."

An independent review requested by FTD came to the same conclusion:

Results of

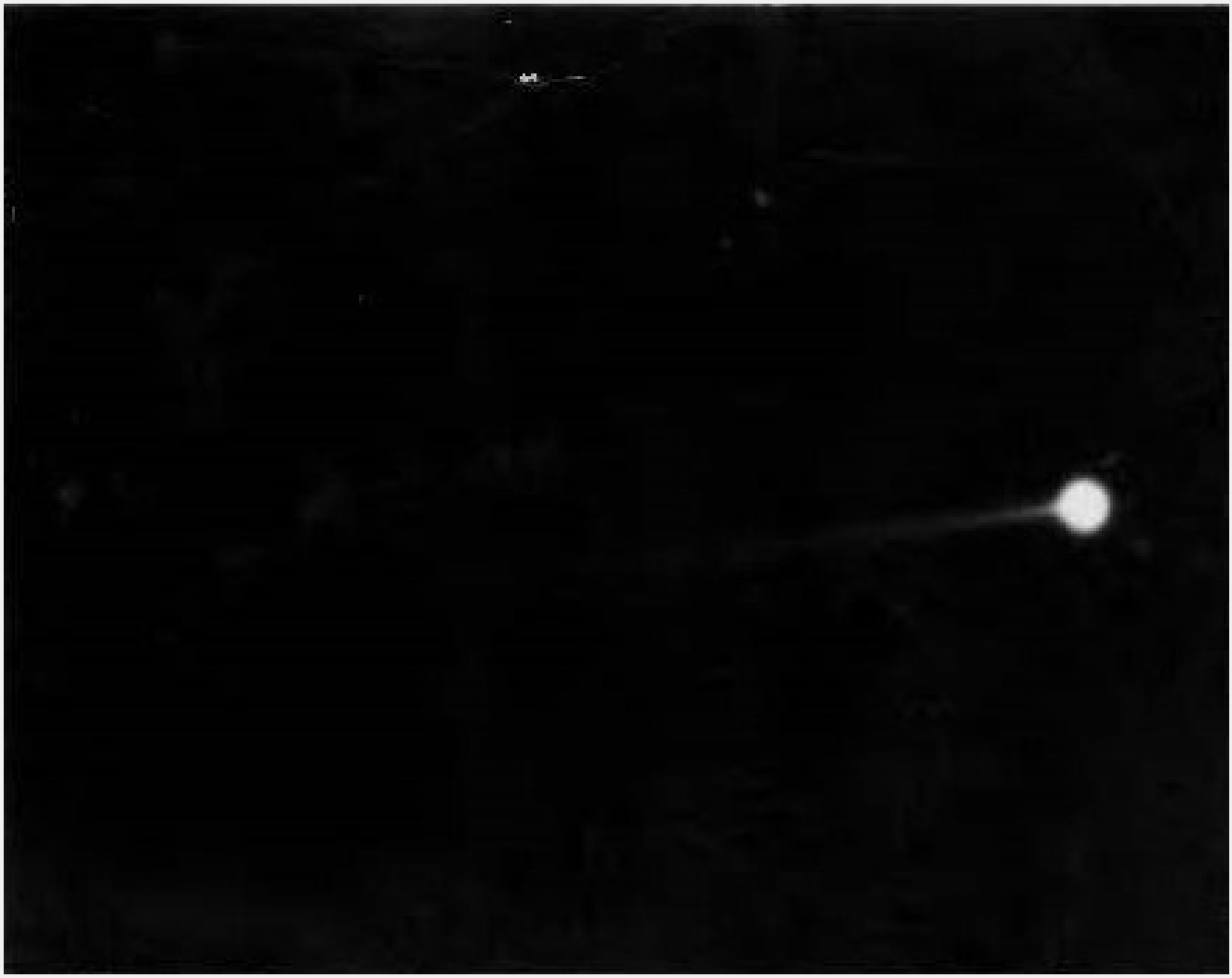
the evaluation of selected cases did not reveal any evidence

of extraterrestrial vehicles nor anything that might be considered beyond the range of present day scientific knowledge. The most probable explanation for the unidentified cases would have to be cast in terms of man made objects, natural phenomena, or psychological causes.

Looking to specific investigation files, we can see what a typical investigation was like, the kinds of documentation and information collected, the investigatory process, and how the Air Force arrived at its conclusions.

Datil, NM, 1950

Cpl. Lertis E. Stanfield, 3024th Air Police Squadron at Holloman Air Force Base in New Mexico, reported seeing a strange object in the sky on the night of February 24/25, 1950. He had a camera with him at the time and took several pictures, including the following:



The details of the sighting were included in an investigation report:

DETAILS:

1. This investigation predicated upon AFCSI Letter No. 85, dated 12 August 1949, to report all sightings of unidentified flying objects.

AT ALAMOGORDO, NEW MEXICO

2. A personal interview has not been conducted with Corporal LERTIS E. STANFIELD, AF-38573711, concerning his sighting of an unusual aerial phenomenon, as he is on detached service at Datil, New Mexico. However, STANFIELD sent the information contained herein in writing to this office.

a. Date of sighting: 24 and 25 February 1950

b. Time of sighting: Object was sighted at 1930 hours on 24 February 1950 and 0200 hours on 25 February 1950. Object remained in view the first time from 1930 to 2200 hours and reappeared at 0200 hours and remained in view approximately 30 minutes.

c. Where sighted: At Datil, New Mexico.
Coordinates: 107° 45' West, 34° 8' North
Object appeared due south of position of observer.

d. Number of Objects: One

e. Observable celestial phenomena or planets that may account for the sighting: None.

f. Distance of object from observer:

(1) Laterally or horizontally: Unable to determine.
(2) Angle of elevation from horizon: Approximately 30°.
(3) Altitude: Not estimated.

g. Time in flight: Approx. 1½ hours - From 2000 hours to 2130 hours.

h. Appearance of object:

(1) Color: White, changing to red and green.
(2) Shape: Object was perfectly round.
(3) Apparent construction: Unable to determine.
(4) Size: Not estimated.

i. Direction of flight: Object appeared to move northwest.

j. Tactics or maneuvers: None

k. Evidence of exhaust: No evidence of exhaust appeared to the naked eye. However, photograph No. 1 indicates some sort of trail from the object.

l. Effect on clouds: No clouds.

m. Lights: Object itself appeared to be a very brilliant light.

n. Supports: None visible.

o. Propulsion: Unable to determine.

p. Control and stability: Movement of object was not erratic, but moved in a straight line.

q. Air ducts: None visible.

r. Speed: 1° per 2 minutes on the azimuth.

s. Sound: None

t. Manner of disappearance: At 2200 hours object disappeared while still high in the sky. Object then re-appeared at 0200 hours 25 February in almost the same position where it disappeared at 2200 hours 24 February.

u. Notes relative to observer:

(1) Name: Cpl. LERTIS E. STANFIELD, AF-38573711
(2) Address: 3024th Air Police Squadron, Holloman AFB.
(3) Occupation: Air Police Range Patrol, Holloman AFB.
(4) Place of business: Datil, New Mexico.
(5) Pertinent hobbies: Hunting and fishing.
(6) Ability to determine: Above average.
(7) Reliability of observer: Very reliable
(8) Former sightings: None
(9) Witnesses: No other witness.

UNCLASSIFIED

[redacted]

CONFIDENTIAL

UNCLASSIFIED

3. Photographs of this object were taken at 1930 hours when it first appeared, by STANFIELD. At 2000 hours the object appeared much brighter than before and had moved west some distance, and appeared much closer. At 2130 hours the object had ceased moving but continued to blink red and green. At 2200 hours object faded from view. Object again appeared at 0200 hours and began a slight movement to the northwest. Object disappeared behind a mountain at approximately 0230 hours, 25 February. Photographs of the object were taken at various times by STANFIELD and are numbered in the order in which they were taken, from 1 to 5.

4. The Electronics and Atmospheric Branch, Holloman AFB, and the Weather Detachment, Holloman AFB, could offer no information which would explain what the object was. Photographs of the object are inclosed herein. A comparison photograph was made by the Photo Lab, Holloman AFB, with the same type of camera, lens, and film, or a round object 3-25/32 inches in diameter. The data regarding this comparison photograph is contained on the back of this photograph, which is No. 6.

Inclosures:

FOR CG, AMC, ATTN: MCIAXO-3

Photographs, Nos. 1, 2, 3, 4, 5 and 6.

FOR HEADQUARTERS OSI

Same as above.

FOR COMMANDING OFFICER, AF CAMBRIDGE RESEARCH LABORATORIES

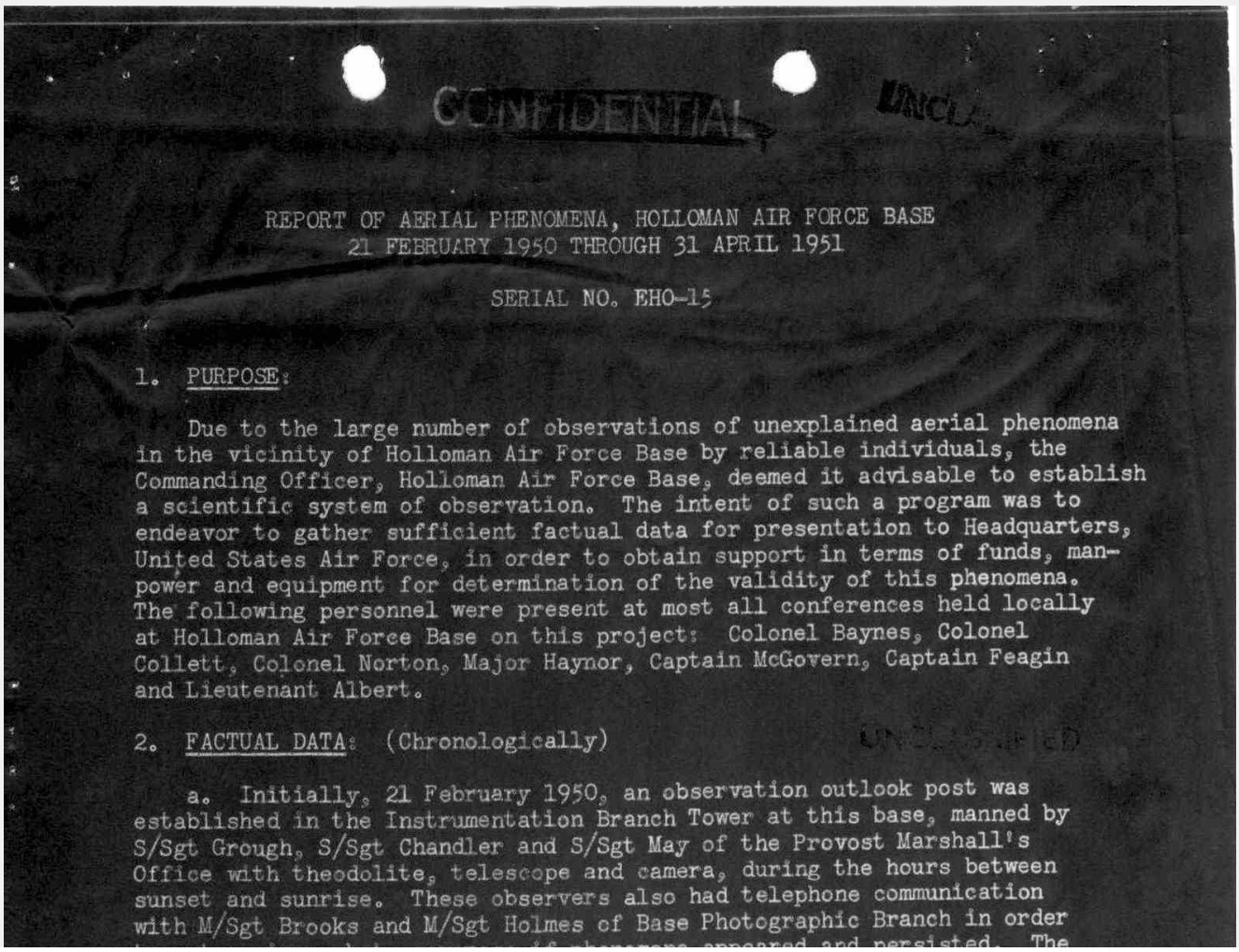
Same as above.

- CLOSED -

 UNCLASSIFIED

CONFIDENTIAL

This was not the first time an unusual sighting had occurred at Holloman. In fact, it was part of a recurring pattern (and one that explains Stansfield's possession of a camera at the time of the sighting).



CONFIDENTIAL UNCL

REPORT OF AERIAL PHENOMENA, HOLLOMAN AIR FORCE BASE
21 FEBRUARY 1950 THROUGH 31 APRIL 1951

SERIAL NO. EHO-15

1. PURPOSE:

Due to the large number of observations of unexplained aerial phenomena in the vicinity of Holloman Air Force Base by reliable individuals, the Commanding Officer, Holloman Air Force Base, deemed it advisable to establish a scientific system of observation. The intent of such a program was to endeavor to gather sufficient factual data for presentation to Headquarters, United States Air Force, in order to obtain support in terms of funds, manpower and equipment for determination of the validity of this phenomena. The following personnel were present at most all conferences held locally at Holloman Air Force Base on this project: Colonel Baynes, Colonel Collett, Colonel Norton, Major Haynor, Captain McGovern, Captain Feagin and Lieutenant Albert.

2. FACTUAL DATA: (Chronologically)

a. Initially, 21 February 1950, an observation outlook post was established in the Instrumentation Branch Tower at this base, manned by S/Sgt Grough, S/Sgt Chandler and S/Sgt May of the Provost Marshall's Office with theodolite, telescope and camera, during the hours between sunset and sunrise. These observers also had telephone communication with M/Sgt Brooks and M/Sgt Holmes of Base Photographic Branch in order

to get maximum photo coverage if phenomena appeared and persisted. The Photographic Branch also supplied Air Police personnel at Datil and Vaughn, New Mexico, with cameras. One of the best pictures of phenomena (Incl "A") was taken from Datil, New Mexico, at 1930 hours 24 February 1950.

At the time, Project Grudge was unable to provide an explanation. However, a decade and a half later, a similar sighting over the Soviet Union provided Blue Book with an answer: a comet.

PROJECT 10073 RECORD

1. DATE - TIME GROUP 24 February 1950 25/0230Z	2. LOCATION Alamogordo, New Mexico
3. SOURCE Military	10. CONCLUSION Astro (Comet) Photo compares with Comet photographs (Evaluation made 1965) (See Photo of Comet SIKI 1962 10 Apr case)
4. NUMBER OF OBJECTS One	
5. LENGTH OF OBSERVATION 1 1/2 hours	11. BRIEF SUMMARY AND ANALYSIS White object changing colors. Appeared to move NW. Observed for 1 hour 30 minutes. Color changes reported as red and green. Tail not visible to naked eye showed on camera. Elevation 30 degrees, azimuth not Reported. Observed in same area 4 hours later. Photos taken. Object appeared only as a brilliant light. Photo taken at 1930 when object first appeared. Disappearance behind mountains.
6. TYPE OF OBSERVATION Gorund Visual	
7. COURSE Appeared to move NW	
8. PHOTOS XX Yes In File ☐ No	
9. PHYSICAL EVIDENCE ☐ Yes XX No	

FORM
FTD SEP 63 0-329 (TDE) Previous editions of this form may be used.

Several sightings of this kind were reported in the desert Southwest around this time. Despite the delay in reaching a conclusion, the similarity of the photographic evidence to known comet sightings led the Air Force to conclude it was dealing with a comet here too.

Redlands, CA, 1958

On December 13, 1958, a man in Redlands, California, snapped a photograph of a strangely shaped object in the sky.

The UFO worksheet described the sighting in detail:

4. TIME AND DATE OF SIGHTING:

a. ZULU TIME-DATE GROUP: 1715Z 13 December 1958

b. LIGHT CONDITIONS (NIGHT, DAY, DAWN, DUSK) Day

5. LOCATION OF OBSERVER. EXACT LATITUDE AND LONGITUDE, OR REFERENCE TO A KNOWN LANDMARK 34° 06' North - 117° 15' West Redlands, California

6. IDENTIFYING INFORMATION OF ALL OBSERVERS:

a. CIVILIAN - MILITARY. NAME, GRADE, ORGANIZATION, DUTY AND ESTIMATE OF RELIABILITY [redacted], Age 26, Address: [redacted], Redlands, California; Occupation: Electrication, Marketeer Co. Redlands, Calif. Very Reliable

7. WEATHER AND WINDS ALOFT - CONDITIONS AT THE TIME AND PLACE OF SIGHTING:

a. OBSERVER'S ACCOUNT OF WEATHER CONDITIONS Clear

b. REPORT FROM NEAREST AWS OR US WEATHER BUREAU: WIND DIRECTION AND VELOCITY IN DEGREES AND KNOTS AT: (IF AVAILABLE)

SURFACE		20,000'	300/26
6,000'	120/06	30,000'	300/39
10,000'	360/07	50,000'	280/44
16,000'	300/21	80,000'	

c. CEILING None

d. VISIBILITY 35 Plus

e. AMOUNT OF CLOUD COVER Light Scattered Cirrus.

f. THUNDERSTORMS IN AREA None

8. ANY OTHER UNUSUAL ACTIVITY OR CONDITION, METEOROLOGICAL, ASTRONOMICAL, ETC.

None

9. INTERCEPTION OR IDENTIFICATION ACTION TAKEN (FOR ADDC OR ADCC)

None

10. LOCATION OF ANY AIR TRAFFIC IN THE AREA AT THE TIME OF SIGHTING

None

11. POSITION, TITLE AND COMMENTS OF THE PREPARING OFFICER, INCLUDING HIS

ANALYSIS Meteoralogical Phenomenn or Jet Aircraft.

12. EXISTENCE OF PHYSICAL EVIDENCE, SUCH AS MATERIALS AND PHOTOGRAPHS Photo available - Camera used - Linhof - wide angle 127mm lens, shutter speed, 1/50 th second, F-16, Photo developed 2215Z 13 December 1958.

Major Loren W. Bruner, USAF
Submitted by

2015Z 16 December 1958
Date/Time Group

UFOB WORK-SHEET

SOURCE

TIME

1. DESCRIPTION OF THE OBJECT:

a. SHAPE Large Jet Aircraft

b. SIZE COMPARED TO A KNOWN OBJECT (USE ONE OF THE FOLLOWING TERMS: HEAD OF A PIN, PEA, DIME, NICKEL, QUARTER, HALF DOLLAR, SILVER DOLLAR, BASEBALL, GRAPEFRUIT, OR BASKETBALL) HELD IN THE HAND AT ABOUT ARMS LENGTH.

c. COLOR Dark

d. NUMBER One

e. FORMATION, IF MORE THAN ONE None

f. ANY DISCERNIBLE FEATURES OR DETAILS 3 dard objects on top of vertical line.

g. TAIL, TRAIL OR EXHAUST, INCLUDING SIZE OF SAME COMPARED TO SIZE OF OBJECT Trail or Exhaust about half size of object.

h. SOUND. IF HEARD, DESCRIBE SOUND None

i. OTHER PERTINENT OR UNUSUAL FEATURES None

2. DESCRIPTION OF COURSE OF OBJECT:

a. WHAT FIRST CALLED THE ATTENTION OF OBSERVER TO THE OBJECT Photo taken for other purpose. First called to attention after development.

b. ANGLE OF ELEVATION AND AZIMUTH (DIRECTION) OF OBJECT WHEN FIRST SIGHTED Easterly to Southeast.

c. ANGLE OF ELEVATION AND AZIMUTH (DIRECTION) OF OBJECT WHEN LAST OBSERVED East - South East

d. DESCRIPTION OF FLIGHT PATH AND/OR MANEUVER OF OBJECT Appeared to travel almost in straight line East - South East.

e. MANNER OF DISAPPEARANCE OF OBJECT To East Southeast

f. LENGTH OF TIME IN SIGHT Approximately 20 seconds.

3. MANNER OF OBSERVATION:

a. USE OF ONE OR ANY NUMBER OF THE FOLLOWING ITEMS: GROUND-VISUAL, GROUND-ELECTRONIC, AIR-ELECTRONIC (IF ELECTRONIC, SPECIFY TYPE OF RADAR) Ground Visual with Camera.

b. STATEMENT AS TO OPTICAL AIDS (TELESCOPES, BINOCULARS, ETC.) None

c. IF THE SIGHTING WAS MADE WHILE AIRBORNE, GIVE TYPE OF ACFT, IDENTIFICATION NO. ALTITUDE, HEADING, SPEED AND HOME STATION None

However, inconsistencies in the reporting led the Air Force to initially determine that the case was impossible to analyze accurately

.

PENNY

5 February 1959

Dear Mr. :

Your report of an unidentified flying object and a negative of the object was received by the Air Technical Intelligence Center on 23 January 1959.

The negative and receipt are returned herewith.

The report contained insufficient data to arrive at a valid conclusion and also contained a serious contradiction which precludes an accurate analysis and/or evaluation of this particular sighting.

You states in the report that you did not see the object until the film had been developed. The exposure for this film was only one fiftieth of a second, yet you say the object was observed for approximately twenty seconds.

This difference in time interval renders this sighting valueless.

Sincerely,

2 Incls

LAWRENCE J. TACKER
Major, USAF
Executive Officer
Public Information Division
Office of Information Services

Mr.
C/O Electric Mfg. Company
Redlands, California

COV. SHEET		SUSPENSE
ORIGIN OF BASIC		DATE
		ASSIGNED BY
DATE	TYPE	NO.

SUBJECT

Unidentified Flying Object

ROUTING

Initial "IN" column to denote review prior to action. Initial "OUT" column to denote review of completed action. (X for action; ✓ for coordination.)

IN	OFFICE	OUT	IN	OFFICE	OUT	IN	OFFICE	OUT	IN	OFFICE	OUT	IN	OFFICE	OUT
	OIN-1			OIN-2			OIN-3			OIN-4			AFOIN	
	OIN-1X			OIN-2X			OIN-3X			OIN-4X			AFOIN-X	
													AFOIN-X	
													AFOIN-X1	
													AFOIN-X2	
													AFOIN-X3	
													AFOIN-X4	
													AFOIN-X5	
													AFOIN-Z	
													CABLES	
												M & R	FILE	
													DISPATCH	

TO: SAFIS-3, Attn: Maj. L. J. Tacker	DATE 30 January 1959
FROM: AFCIN-4E4	COMMENT NO. 1

COMMENTS *(Use reverse, if necessary)*

4E4/Maj. Friend/ac/69216/Bldg 828

1. The attached negative and receipt is to be returned to Mr. .

2. The following is a brief summary of the report:

a. One dark object, shaped like a large jet aircraft, size of a dime at arm length. There were three dark objects on top of the vertical line. Trail or exhaust about half size of object.

b. Observer first saw the object when the film was developed. Object was first and last seen east southeast and traveled in a straight line east southeast. Object was observed for approximately twenty seconds.

c. Sky was clear, with light scattered cirrus clouds. Visibility thirty-five miles. Photographic exposure was one fiftieth of a second.

3. The observer never saw the object until he had the film developed. The exposure was only one fiftieth of a second. Yet he says the object was observed for approximately twenty seconds.

4. The observer determined from one photograph that the object was traveling to the east southeast. To determine the direction of movement with such a short exposure is impossible, even for the most experienced photo interpreter.

5. An analysis of the photograph was made and the object was determined to be a lenticuler cloud.

H/K Gilbert

2 Incls:
1. Negative
2. Receipt for photo negative

H. K. GILBERT
Colonel, USAF
AFCIN-4E

DISPATCHED
FEB 5 12 03 28
ATIC
2

All of these sighted were explained as initially misinterpreted natural occurrences. In the next post of the series, we'll turn our attention to sightings ultimately identified as human-created objects and one sighting truly classified as a UFO.

As we could read above, UFO sightings mostly are misinterpreted weahterlike or other influences, hence we will test the impact of air pollutants too!

**If you want you can read even more, here, or here, or just watch this*

3.3 Impact of Air Pollution on cognitive abilities and visibility

> *In this part we will take closer look on how Air Pollution affets visibility, visual performance and overall cognitive abilities of human beeing. We set this as our basis to explain the UFO Sightings. We assume that air pollutants have an impact on brain cognitive abilities and overall visibility, thus some can claim that they have seen UFO.*

3.3.1 Impact on cognitive abilities

> *First of all we will to make a brief summary to how air pollutants affect brain cognitive abilities, quoting some Chinesee research article about it. Full article you can find here.*

3.3.1.1 Data Set

> *Data set of this article research is based on several sources. The cognitive test scores come from the CFPS, a nationally representative survey of Chinese families and individuals. The waves 2010 and 2014 contain the same cognitive ability module, that is, 24 standardized mathematics questions and 34 word-recognition questions. All of these questions are sorted in ascending order of difficulty, and the final test score is defined as the rank of the hardest question that a respondent is able to answer correctly. The survey also provides exact information about the geographic locations and dates of interviews for all respondents, which enables us to match test scores with local air quality data more precisely. Air quality is measured using the air pollution index (API), which is calculated based on daily readings of three air pollutants, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), and particulate matter smaller than 10 μm (PM10). The API ranges from 0 to 500, with larger values indicating worse air quality. Daily API observations are obtained from the city-level air quality report published by the Chinese Ministry of Environmental Protection. The report includes 86 major cities in 2000 and covers most of the cities in China in 2014.*

3.3.1.2 Econometric Model

> *Econometric Model of this research looks as follows: $$Score_{ijt} = \alpha_{1}P_{jt}+\alpha_{2} \cdot \frac{1}{k}\sum_{n=0}^{k-1} P_{j,t-n}+X_{ijt}^{'}\beta+W_{jt}^{'}\phi+T_{jt}^{'}\gamma+\lambda_{i}+\delta_{j}+\eta_{t} + f (t) +\epsilon_{ijt}$$*

Description down below:

The dependent variable Scoreijt is the cognition test scores of respondent i in county j at date t. Pjt is the contemporaneous air quality measure at date t. The key variable ð1=kÞPk−1 n=0Pj,t−n is the mean API reading in the past k days, which measures cumulative exposure. Xijt is a set of the observable demographic correlates of the respondents. We also control for a vector of contemporaneous weather conditions Wjt and a vector of county-level characteristics Tjt to account for factors that are correlated with both test scores and air quality. λi denotes individual fixed effects. δj represents county fixed effects, which cannot be wiped out by individual fixed effects since some respondents do not live in the same counties across the two waves. ηt indicates month, day of week, and postmeridiem hour fixed effects. f(t) is the quadratic monthly time trend that ranges from 1 (January 2010) to 60 (December 2014). «ijt is the error term. SEs are clustered at the county level.

3.3.1.3 Conclusion

As we dont want to copy paste all of the article here we will just skip to the conclusion, whole article is available here.

This paper estimates the contemporaneous and cumulative impacts of air pollution on cognition by matching the scores of verbal and math tests given to people age 10 and above in a nationally representative survey with local air quality data according to the exact dates and locations of the interviews. We find that accumulative exposure to air pollution impedes verbal test scores. As people age, the negative effect becomes more pronounced, especially for men. The gender gap is particularly large for the less educated. Our findings about the damaging effect of air pollution on cognition, particularly on the aging brain, imply that the indirect effect on social welfare could be much larger than previously thought. A narrow focus on the negative effect on health may underestimate the total cost of air pollution.

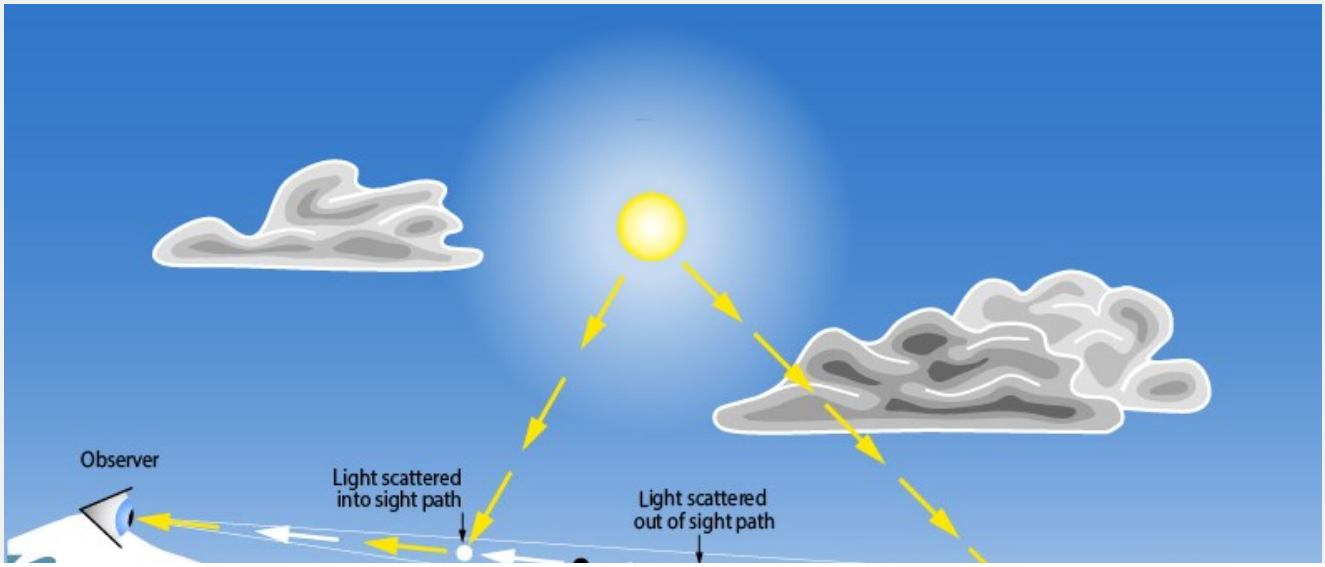
3.3.2 Impact on weather and clarity

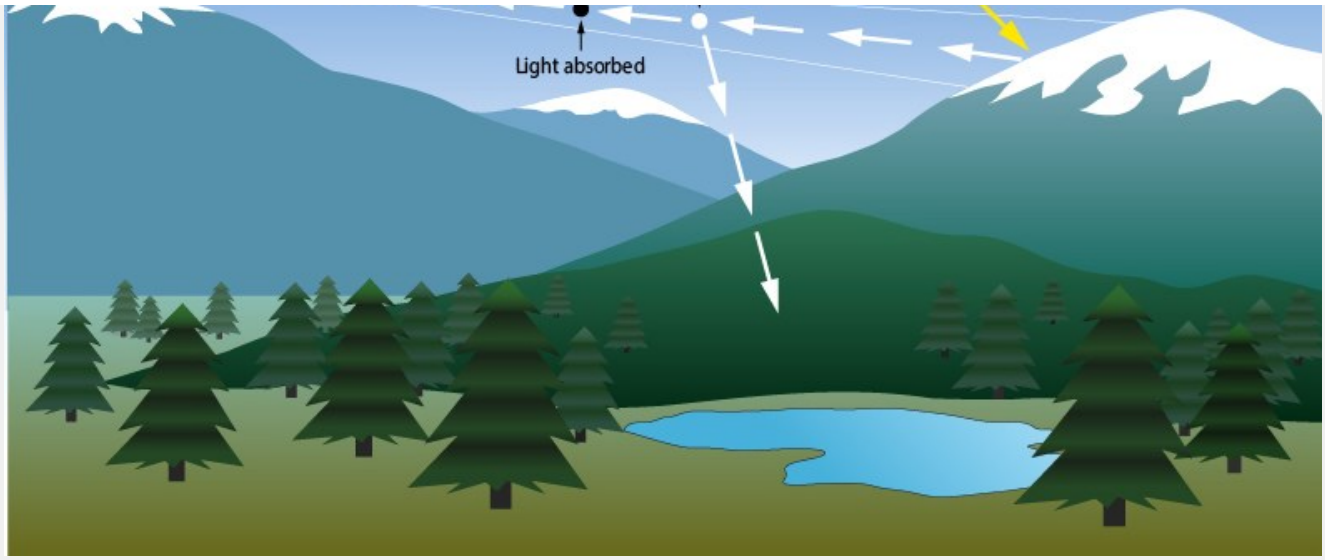
3.3.2.1 Basic Informations

This part has been taken from here.

Air pollution can create a white or brown haze that affects how far we can see. It also affects how well we are able to see the colors, forms, and textures of natural and historic vistas.

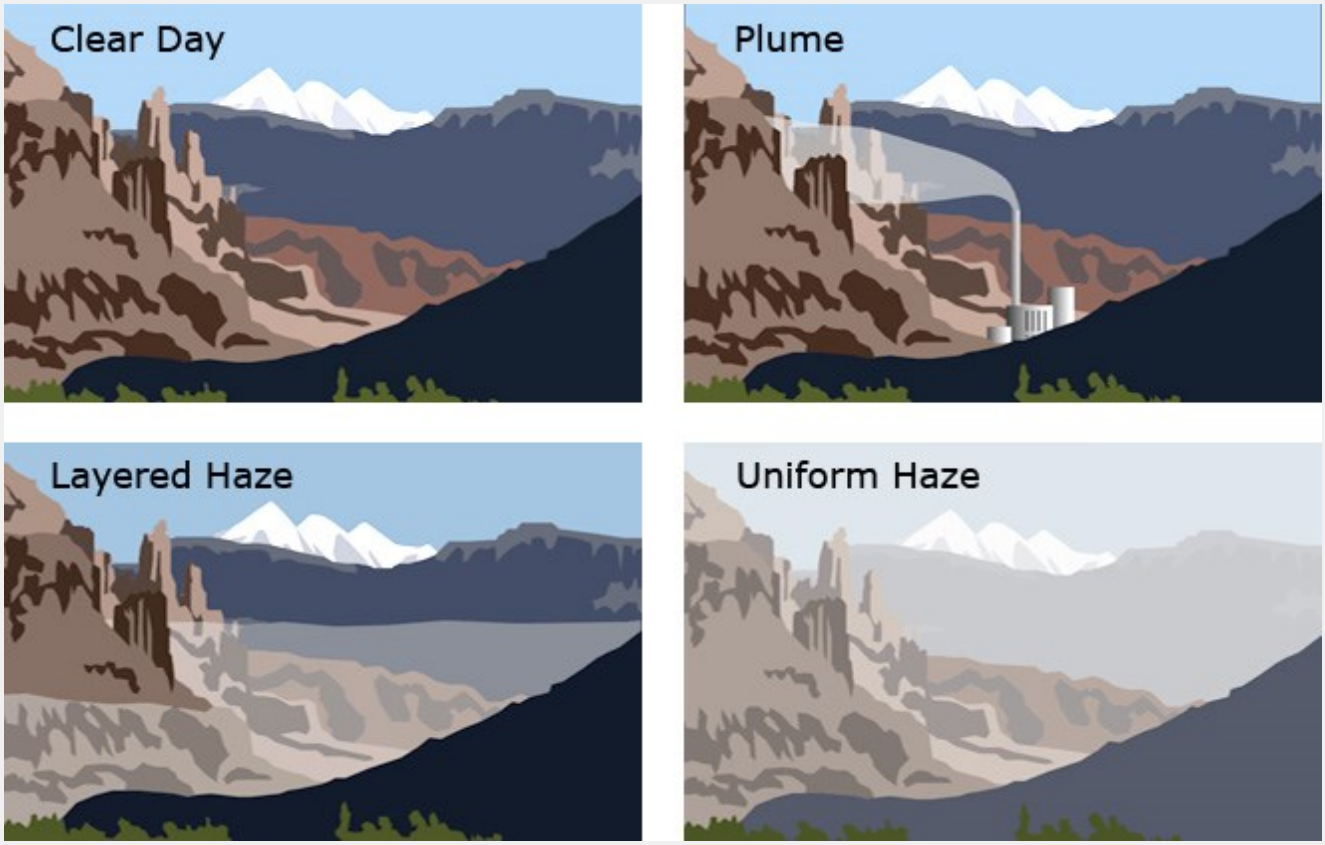
Haze is caused when sunlight encounters tiny particles in the air. The particles scatter light into and out of the sight path and absorb some light before it reaches your eyes. The more particles in the air, the more scattering and absorption of light to reduce the clarity and colors of what you see. Some types of particles scatter more light, especially when it is humid. Haze is mostly caused by air pollution from human activity including industry, power generation, transportation, and agriculture. Natural haze from dust, wildfires, and more also occurs in many parks.





Air pollution does not impact views on clear days but can be seen as a plume, layered haze, or uniform haze when air pollution is present. A plume of air pollution is a tight, vertically constrained layer of air pollution coming from a point source (such as a smoke stack). Layered haze is any confined layer of pollutants that creates as a visible contrast between that layer and the sky or landscape behind it. In an unstable atmosphere, plumes and layers mix with the surrounding atmosphere creating a uniform haze or overall reduction in air clarity.

Plumes and layered haze are more common during cold winter months when the atmosphere is more stagnant. Uniform haze occurs when warm turbulent air causes atmospheric pollutants to become well mixed.



3.3.2.2 Conclusion

> *As we can read from above, air pollution can have impact on forming haze, and the proces how light is absorbed by air particles, thus we can conclude that it can be main problem when it comes to truly distinguish if someone have seen ufo or not.*

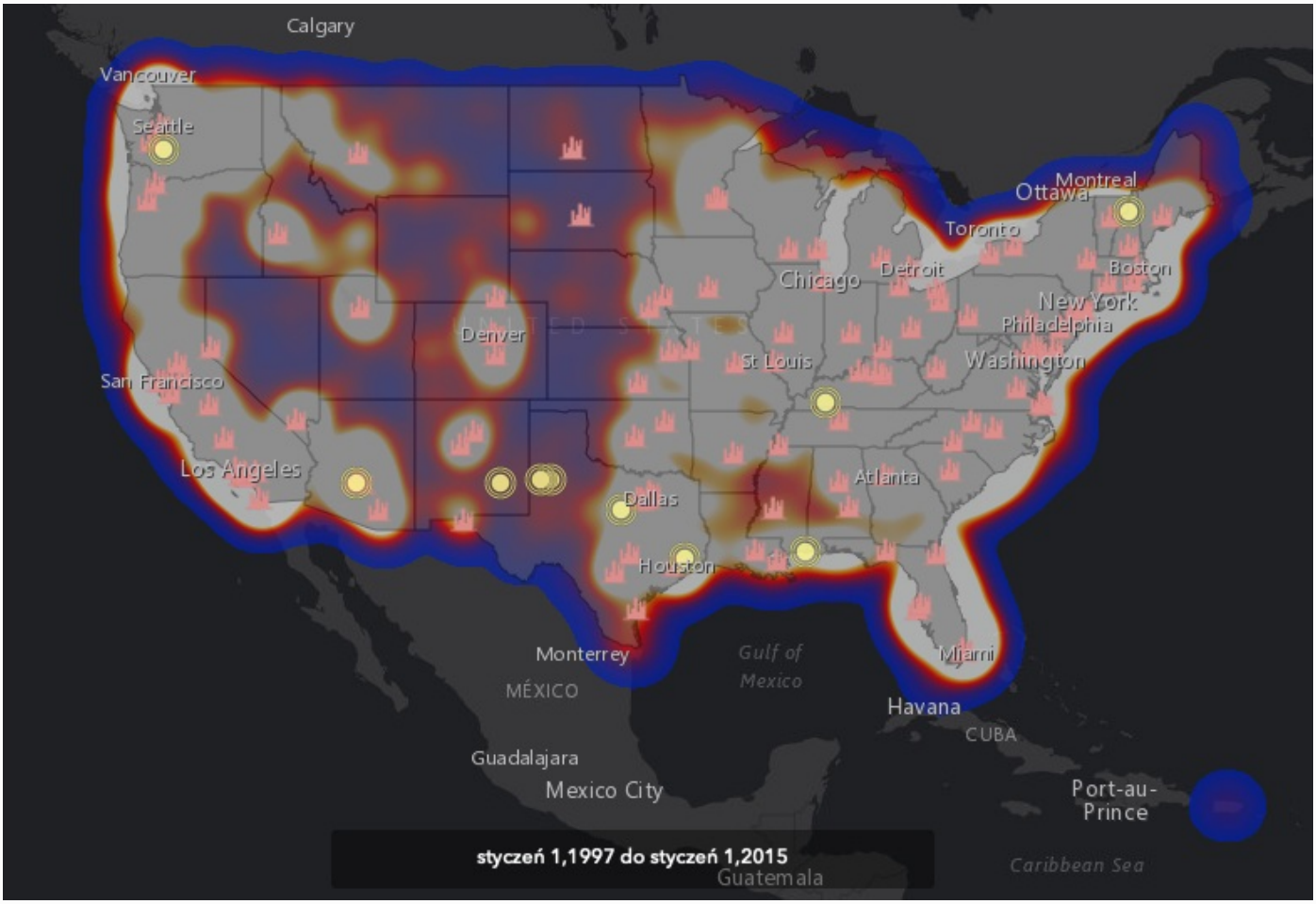
3.3.3 UFO Sightings MAP with comparision to Pollution MAP in USA

3.3.3.1 Comparision

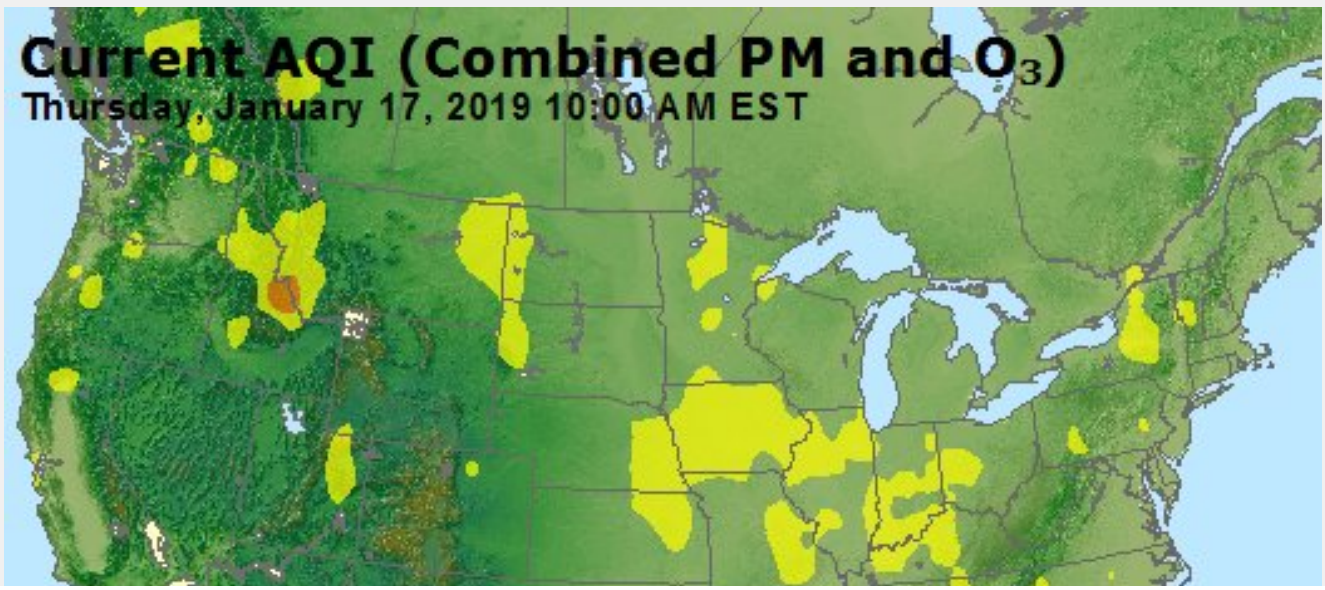
3.3.3.1 Comparision

> *Below we have rendered map of UFO sightings in the usa from the year 1997 to 2015. Red area represent general ufo sightings area, the more intensive, the most UFO sightings occured there. Yellow dots are historical documented UFO sightings, with photos etc.*

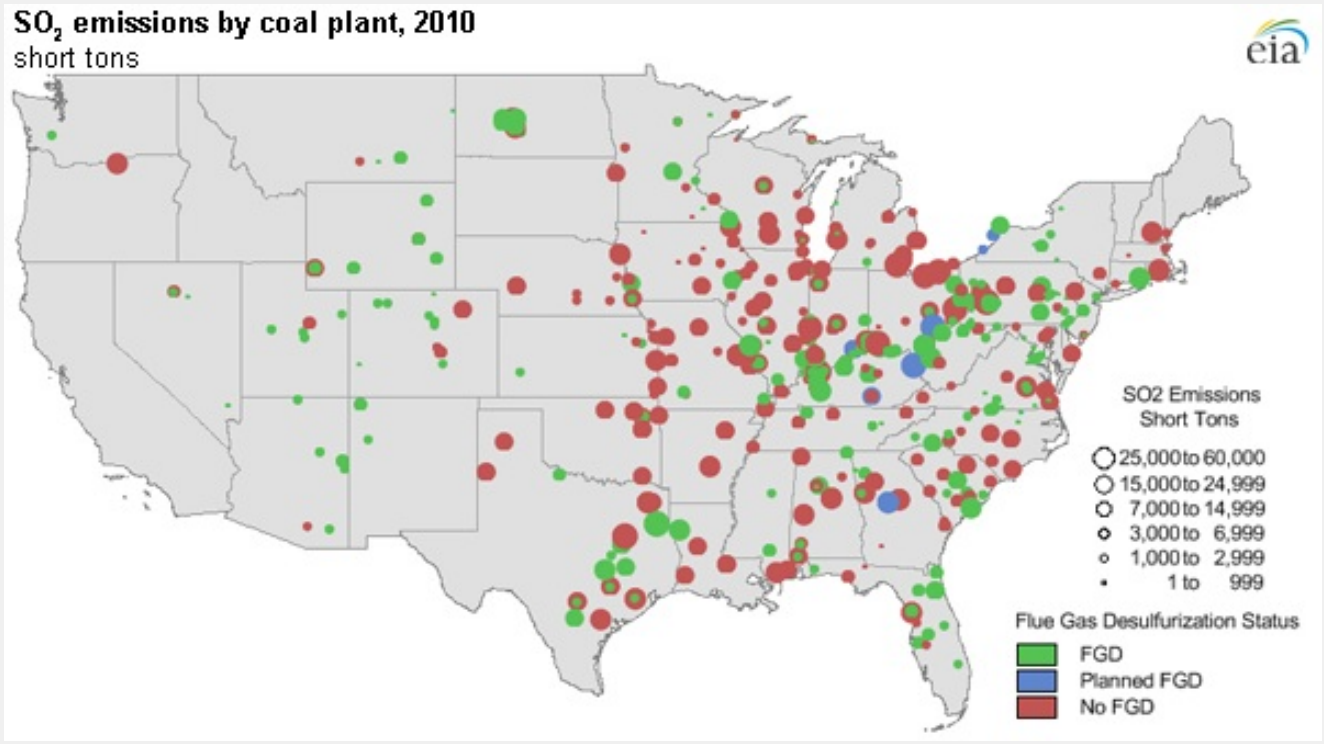
> *Now we will compare it to randomly chosen periods when air was polluted by the pollutants gathered in our data set, just to see in which regions, the pollution in average occurs most often. Here Hovewer we have Animated NO2 pollution map from 2005 to 2011 just to see in average, where the pollutants occurs the most.*



> *Below we have combined O3 and PM pollution Map in 2019*



Below Sulfur Dioxide pollution map in 2010



At the first glance we can see that occurences of UFO is higher in areas highly polluted.(We provide this example maps to just check at the first glance is there any relation between air pollution and ufo sightings based on those heatmaps, this is not any kind of statistical or computational based observation)

4 The impact of the Air Pollution on UFO Sightings: Analysis

4.1 Research Goals, Problems and Hypothesis

The aim *of this analysis is to verify if there really exist any air pollution influence on the UFO sightings. The derivative goal is also to check how it can be connected to other issues about UFO sigthtings.*

The research problems undertaken in analysis are:

- *Determine what factors affect the UFO sightings the most*
- *Identification the nature of factors and air pollutants on UFO sigtings*
- *Identifying in which states the UFO occurences with respect to air pollutants were the highest*

- *Indication the reasons of the sigthtings (influence of the air pollutants to cognitive abilities)*

*The main hypothesis formulated in the work is: **"There is existing connection between air pollution level and UFO Sightings".** Hence, our main goal of course is to test the data to verify the hypothesis. This process will take place based on various data analysis method mentioned in chapter two. We will also compare the result to the potential explanation of UFO sightings mentioned in chapter three, to make sure if our analysis is different from others (ex.impact of weather on UFO sightings). However, the choice of analysing this exact data set is supported by Ig Nobel work intention.*

Our data set is combined UFO sigthtings with respect to longtitute, latitude and air pollution with the same longtitude and latitude parameters. Belowe you can find brief view of data:

Before the analysis we will plot missing values in our dataset to decide whether any variable should be removed.

Hide

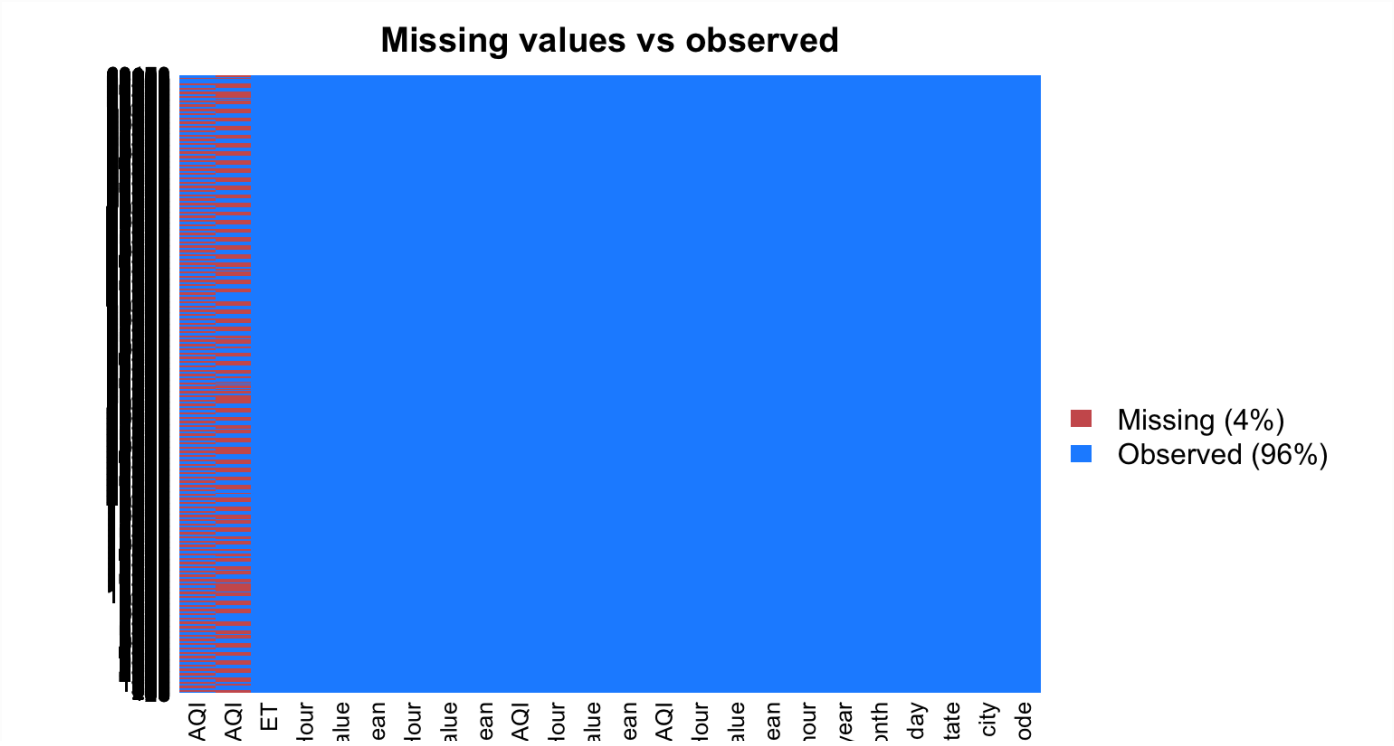
```
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

Hide

```
missmap(data, main = "Missing values vs observed")
```



CO..
SO2..
O.1st.Max.H
O.1st.Max.Va
CO.M
O2.1st.Max.H
2.1st.Max.Va
SO2.M
O3..
O3.1st.Max.H
3.1st.Max.Va
O3.M
NO2..
O2.1st.Max.H
2.1st.Max.Va
NO2.M
mo
s
State.C

> *As we can see, variable CO.AQI and SO2.AQI contain many missing values. These variables are both air quality indices throughout a day. Involving them would mean cutting out many observations from the dataset, thus we decided not to include them in our analysis. We also cut out the variables that are qualitative and not quantitative (state, city and State.Code).*

> *After a short analysis, we also decided to drop variable 'year'. The database was constructed in such way that every observation after year 2000 has variable ET equal 1. In such case including year would end up in false conclusions. We chose to treat month as a discrete variable, while other variables will be treated as continous ones.*

Hide

```
clean.data <- subset(data, select = -c(CO.AQI, SO2.AQI, State.Code, city, state, year))
clean.data$month <- as.factor(clean.data$month)
```

4.2 Analysis

4.2.1 Linear regression

> *First of all we will start with simple Linear Regression. We will split the data into in-sample and out-of-sample datasets in order to be able to test our model. We will call them the training set and the testing set respectively. Then we will perform linear regression on the training sample.*

Hide

```
set.seed(123)
test_ind <- sample(seq_len(nrow(clean.data)), size = 5000)

train <- clean.data[-test_ind, ]
test <- clean.data[test_ind, ]

linearMod <- lm(ET ~ ., data=train)
summary(linearMod)
```

```
##
## Call:
## lm(formula = ET ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05185 -0.01900 -0.00962  0.00001  1.01091
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.999e-02  2.888e-03  13.846  < 2e-16 ***
## day          -1.123e-03  4.612e-05 -24.341  < 2e-16 ***
## month2       -5.273e-03  2.511e-03  -2.100 0.035731 *
## month3       -1.148e-02  2.418e-03  -4.746 2.08e-06 ***
## month4       -8.228e-03  2.294e-03  -3.587 0.000334 ***
## month5       -1.160e-02  2.313e-03  -5.015 5.33e-07 ***
## month6       -1.459e-02  2.311e-03  -6.312 2.77e-10 ***
## month7       -1.279e-02  2.329e-03  -5.493 3.97e-08 ***
## month8       -1.528e-02  2.290e-03  -6.672 2.55e-11 ***
## month9       -1.381e-02  2.242e-03  -6.160 7.33e-10 ***
## month10      -1.154e-02  2.194e-03  -5.261 1.44e-07 ***
```

```
## month11           -4.444e-03  2.489e-03  -1.785 0.074217 .
## month12           -8.721e-03  2.392e-03  -3.645 0.000267 ***
## hour               8.564e-04  5.840e-05  14.666  < 2e-16 ***
## NO2.Mean           1.075e-05  1.081e-04   0.099 0.920810
## NO2.1st.Max.Value -7.168e-05  1.170e-04  -0.613 0.540188
## NO2.1st.Max.Hour  -1.098e-05  5.783e-05  -0.190 0.849380
## NO2.AQI           -6.114e-05  1.329e-04  -0.460 0.645600
## O3.Mean           -4.086e-01  9.391e-02  -4.351 1.36e-05 ***
## O3.1st.Max.Value   2.544e-01  9.365e-02   2.716 0.006612 **
## O3.1st.Max.Hour   -5.755e-05  1.004e-04  -0.573 0.566517
## O3.AQI            -4.385e-05  5.491e-05  -0.799 0.424563
## SO2.Mean          -8.306e-04  1.683e-04  -4.936 8.01e-07 ***
## SO2.1st.Max.Value -8.936e-06  6.361e-05  -0.140 0.888281
## SO2.1st.Max.Hour   6.246e-05  6.353e-05   0.983 0.325538
## CO.Mean           -6.481e-04  2.158e-03  -0.300 0.763938
## CO.1st.Max.Value  -2.820e-03  1.016e-03  -2.775 0.005528 **
## CO.1st.Max.Hour    1.233e-05  5.603e-05   0.220 0.825807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09782 on 58145 degrees of freedom
## Multiple R-squared:  0.01768,    Adjusted R-squared:  0.01723
## F-statistic: 38.77 on 27 and 58145 DF,  p-value: < 2.2e-16
```

> *In the first place we see that p-value for some, but no all variables are smaller than 0.05. Small p-value indicates strong evidence against the null hypothesis, therefore we can reject the null hypothesis that variables are not statistically significant. We will run the anova() function on the model to analyse the table of deviance and decide which variables to exclude from our model.*

Hide

```
anova(linearMod, test="Chisq")
```

> *A large p-value implies that the model without the variable explains about the same amount of variation. We can see that some variables seem to improve the model even though they didn't seem statistically significant. Thus, we will include in our model these variables which were either statistically significant or had low p-value in ANOVA test.*

Hide

```
linearMod2 <- lm(ET ~ day+month+hour+NO2.Mean+O3.Mean+O3.1st.Max.Hour+SO2.Mean+CO.Mean+CO.1st.Max.Value, data
summary(linearMod2)
```

```
##
## Call:
## lm(formula = ET ~ day + month + hour + NO2.Mean + O3.Mean + O3.1st.Max.Hour +
##     SO2.Mean + CO.Mean + CO.1st.Max.Value, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05143 -0.01896 -0.00963 -0.00002  1.00704
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.038e-02  2.606e-03  15.494  < 2e-16 ***
## day              -1.122e-03  4.609e-05 -24.340  < 2e-16 ***
## month2           -5.087e-03  2.500e-03  -2.035 0.041838 *
## month3           -1.098e-02  2.383e-03  -4.607 4.09e-06 ***
## month4           -7.686e-03  2.248e-03  -3.419 0.000630 ***
## month5           -1.062e-02  2.263e-03  -4.695 2.68e-06 ***
## month6           -1.348e-02  2.266e-03  -5.951 2.68e-09 ***
```

```
## month7           -1.125e-02  2.263e-03  -4.970 6.72e-07 ***
## month8           -1.375e-02  2.219e-03  -6.193 5.92e-10 ***
## month9           -1.245e-02  2.185e-03  -5.697 1.22e-08 ***
## month10          -1.053e-02  2.154e-03  -4.891 1.00e-06 ***
## month11          -4.027e-03  2.475e-03  -1.627 0.103807
## month12          -8.601e-03  2.378e-03  -3.616 0.000299 ***
## hour              8.553e-04  5.840e-05  14.647  < 2e-16 ***
## NO2.Mean         -1.420e-04  5.293e-05  -2.683 0.007306 **
## O3.Mean          -1.688e-01  4.148e-02  -4.068 4.74e-05 ***
## O3.1st.Max.Hour  -4.905e-05  9.941e-05  -0.493 0.621724
## SO2.Mean         -8.095e-04  9.927e-05  -8.154 3.58e-16 ***
## CO.Mean          -7.418e-04  2.100e-03  -0.353 0.723946
## CO.1st.Max.Value -2.678e-03  9.780e-04  -2.738 0.006182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09782 on 58153 degrees of freedom
## Multiple R-squared:  0.01745,    Adjusted R-squared:  0.01713
## F-statistic: 54.36 on 19 and 58153 DF,  p-value: < 2.2e-16
```

> *Negative coefficients suggest that all other variables being equal, the higher value of the independent variable, the smaller value of the dependent variable. Unfortunately, ET is a dummy variable. It is equal 1 when seeing of UFO occurred and 0 otherwise. Therefore the interpretation cannot be that straightforward. Instead, we should concentrate on probability of an event occuring. The probabilty of y=1 is as follows:*

\[E(y) = p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \] > Thus the interpretation of coefficients should be that a negative coefficient ceteris paribus suggests a decrease in probability of having seen UFO. The huge downside of the linear model is that theoretical values, i.e. in-sample forecasts, may be outside the range [0,1]. What's more, we cannot interpret the measurements of the goodness of fit, e.g. \(R^2\). Additionally, there are also issues concerning heteroscedasticity and lack of normality of residuals. Despite that, we will test the predictive ability of the model. In the testing sample, there are 49 out of 5000 observations with value of ET equals 1. This means, that our model, if well specified, should have accuracy higher than 0.9902

Hide

```
fitted.results <- predict(linearMod2,newdata=subset(test,select=-c(ET)),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$ET)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.9902"
```

> *The 0.9902 Accuracy means that our model is as good in terms of prediction as always assuming 0, meaning it might be not the best predictive model.*

4.2.2 Binomial logistic regression

> *Logistic regression is a method for fitting a regression when the dependent variable is a categorical (i.e. dummy) variable. Theoretically, it should fit our dataset better. We will fit a logistic regression model on our training dataset in order to asses that.*

Hide

```
logitMod <- glm(ET ~ ., data=train, family=binomial(link='logit'))
summary(logitMod)
```

```
##
## Call:
## glm(formula = ET ~ ., family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8544  -0.1386  -0.0717  -0.0370   4.1817
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.947e+00  2.838e-01  -6.862 6.80e-12 ***
## day               -1.588e-01  7.492e-03 -21.196  < 2e-16 ***
## month2            -2.924e-01  2.076e-01  -1.409 0.158894
## month3            -9.337e-01  2.320e-01  -4.025 5.70e-05 ***
## month4            -5.305e-01  1.944e-01  -2.728 0.006366 **
## month5            -8.402e-01  2.132e-01  -3.941 8.12e-05 ***
## month6            -1.257e+00  2.222e-01  -5.660 1.52e-08 ***
## month7            -1.031e+00  2.109e-01  -4.889 1.01e-06 ***
## month8            -1.456e+00  2.289e-01  -6.361 2.01e-10 ***
## month9            -1.213e+00  2.072e-01  -5.854 4.79e-09 ***
## month10           -1.085e+00  2.023e-01  -5.363 8.20e-08 ***
## month11           -3.662e-01  2.032e-01  -1.802 0.071565 .
## month12           -7.426e-01  2.207e-01  -3.365 0.000765 ***
## hour               9.735e-02  6.912e-03  14.084  < 2e-16 ***
## NO2.Mean           1.346e-02  1.329e-02   1.013 0.311182
## NO2.1st.Max.Value -2.636e-02  3.781e-02  -0.697 0.485670
## NO2.1st.Max.Hour  -2.039e-03  6.150e-03  -0.332 0.740263
## NO2.AQI            7.487e-03  4.029e-02   0.186 0.852564
## O3.Mean           -5.171e+01  1.011e+01  -5.115 3.14e-07 ***
## O3.1st.Max.Value   2.980e+01  1.114e+01   2.675 0.007465 **
## O3.1st.Max.Hour   -8.021e-03  1.034e-02  -0.776 0.437755
## O3.AQI            -3.182e-03  7.157e-03  -0.445 0.656587
## SO2.Mean          -1.676e-01  3.023e-02  -5.543 2.97e-08 ***
## SO2.1st.Max.Value -1.916e-04  1.054e-02  -0.018 0.985490
## SO2.1st.Max.Hour   8.076e-03  6.572e-03   1.229 0.219157
## CO.Mean           -2.004e-01  2.723e-01  -0.736 0.461848
## CO.1st.Max.Value  -4.283e-01  1.513e-01  -2.831 0.004646 **
## CO.1st.Max.Hour    4.165e-03  5.957e-03   0.699 0.484426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6426.0  on 58172  degrees of freedom
## Residual deviance: 5245.8  on 58145  degrees of freedom
## AIC: 5301.8
##
## Number of Fisher Scoring iterations: 9
```

> *Then we will run anova() function again, to decide which variables improve our model.*

Hide

```
anova(logitMod, test="Chisq")
```

> *The procedure is the same as the one regarding linear model. We choose the variables that seem to be statistically significant or have low p-value in ANOVA.*

Hide

```
logitMod2 <- glm(ET ~ day+hour+NO2.Mean+O3.Mean+O3.1st.Max.Hour+SO2.Mean+CO.Mean+CO.1st.Max.Value , data=trai
summary(logitMod2)
```

```
##
## Call:
## glm(formula = ET ~ day + hour + NO2.Mean + O3.Mean + O3.1st.Max.Hour +
##     SO2.Mean + CO.Mean + CO.1st.Max.Value, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7730  -0.1429  -0.0746  -0.0396   4.1472
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.636271   0.216924 -12.153  < 2e-16 ***
## day               -0.158210   0.007445 -21.252  < 2e-16 ***
## hour               0.096437   0.006900  13.977  < 2e-16 ***
## NO2.Mean          -0.011014   0.006519  -1.690   0.0911 .
## O3.Mean          -26.971608   4.292391  -6.284 3.31e-10 ***
## O3.1st.Max.Hour   -0.004297   0.010296  -0.417   0.6764
## SO2.Mean          -0.162148   0.018821  -8.615  < 2e-16 ***
## CO.Mean           -0.076276   0.262124  -0.291   0.7711
## CO.1st.Max.Value  -0.320398   0.140951  -2.273   0.0230 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6426.0  on 58172  degrees of freedom
## Residual deviance: 5330.4  on 58164  degrees of freedom
## AIC: 5348.4
##
## Number of Fisher Scoring iterations: 9
```

> *We can see that dropping variables improved the AIC statistic. While there is no perfect substitute of\(R^2\) for logistic regression, we will use McFadden pseudo R squared index, which can be used to assess the model fit.*

Hide

```
library(pscl)
pR2(logitMod)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -2.622896e+03 -3.212984e+03  1.180177e+03  1.836575e-01  2.008296e-02
##          r2CU
##  1.920334e-01
```

Hide

```
pR2(logitMod2)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -2.665181e+03 -3.212984e+03  1.095605e+03  1.704966e-01  1.865733e-02
##          r2CU
##  1.784014e-01
```

> *The pseudo R squared is higher for the first model, which should comes as no surprise as it contains more variables.*

Nevertheless, we will be using the second model for the further analysis. Now we will assess the predictive power of the logit model.

Hide

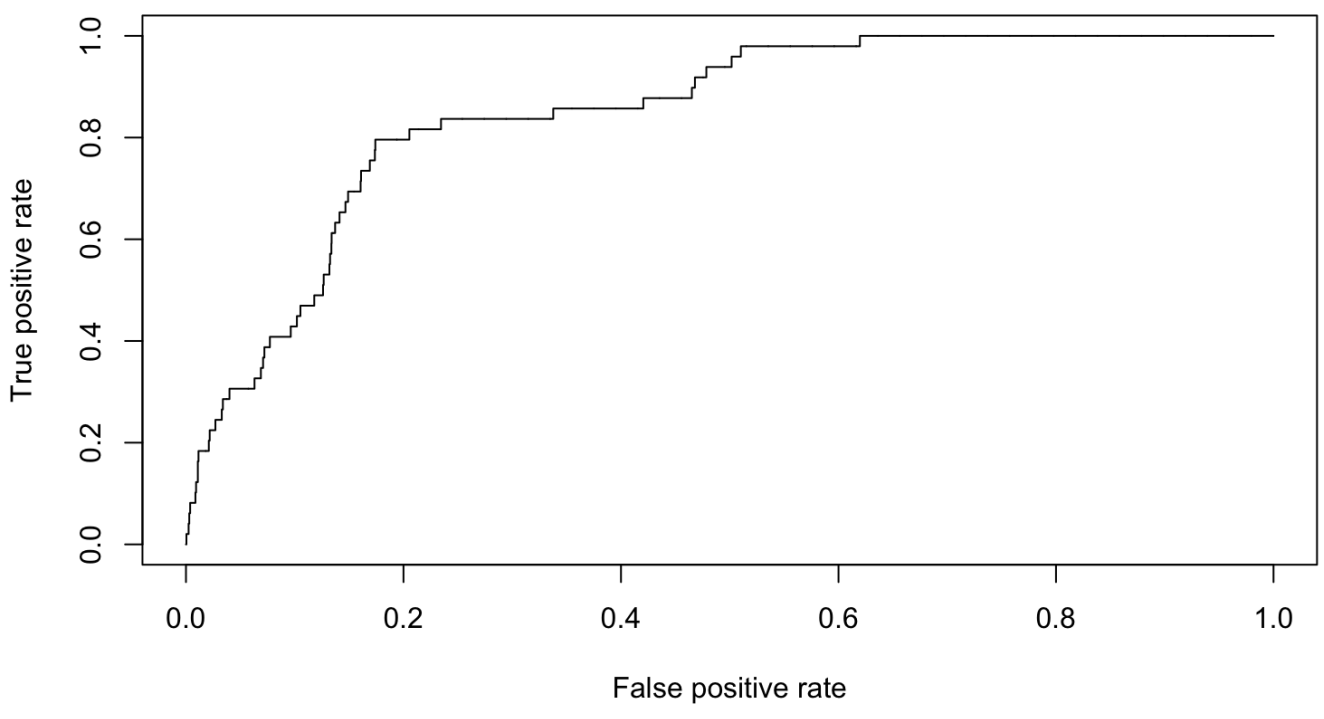
```
fitted.results <- predict(logitMod2,newdata=subset(test,select=-c(ET)),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$ET)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.9902"
```

The Accuracy is exactly the same as for the linear model. Logit model did not improve predictive accuracy. Finally, we will plot the ROC curve and substract the Area Under the Curve (AUC). ROC is a plot of true positive rate versus the false positive rate. Generally, a good predictive model should have an AUC closer to 1 than 0.5

Hide

```
library(ROCR)
p <- predict(logitMod2,newdata=subset(test,select=-c(ET)),type='response')
pr <- prediction(p, test$ET)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```



Hide

```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8479631
```

> *AUC is equal 0.85 which suggests that our model has a good predictive power. Even though, it doesn't beat the benchmark which is a model that assigns always 0, which might be pretty disappointing.*

4.2.3 Decision tree

> *In order to plot a decistion tree, firstly we will create variable ET2 and assign to it value 'yes' if ET is equal 1 and 'no' otherwise. Then we will use the ctree() function to create a decision tree and plot it.*

Hide

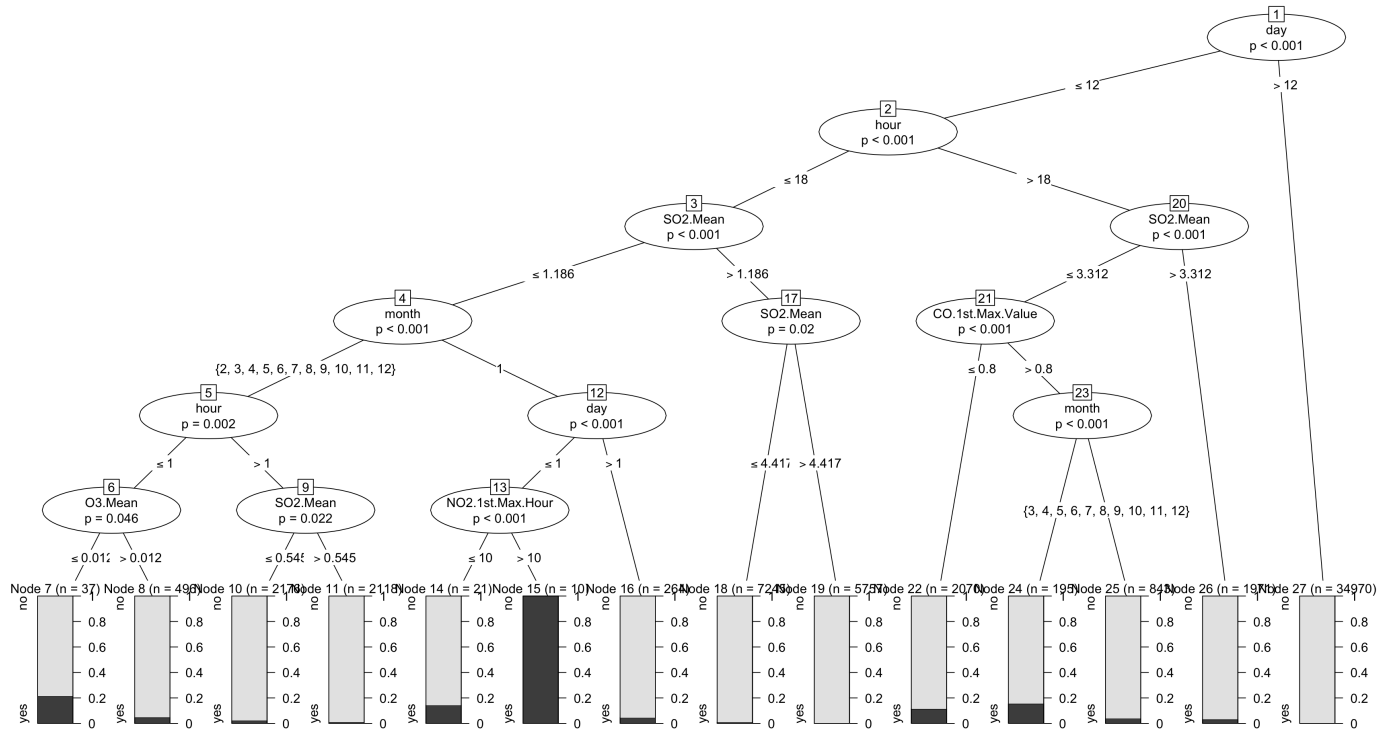
```
library(party)
train$ET2 <- ifelse(train$ET==1, 'yes', 'no')
train$ET2 <- as.factor(train$ET2)
output.tree <- ctree(ET2 ~ ., data = subset(train, select=-c(ET)))

output.tree
```

```
##
##   Conditional inference tree with 14 terminal nodes
##
## Response:  ET2
## Inputs:  day, month, hour, NO2.Mean, NO2.1st.Max.Value, NO2.1st.Max.Hour, NO2.AQI, O3.Mean, O3.1st.Max.Valu
## Number of observations:  58173
##
## 1) day <= 12; criterion = 1, statistic = 585.34
##   2) hour <= 18; criterion = 1, statistic = 208.587
##     3) SO2.Mean <= 1.185714; criterion = 1, statistic = 50.235
##       4) month == {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}; criterion = 1, statistic = 61.877
##         5) hour <= 1; criterion = 0.998, statistic = 15.324
##           6) O3.Mean <= 0.011792; criterion = 0.954, statistic = 14.26
##             7)*  weights = 37
##           6) O3.Mean > 0.011792
##             8)*  weights = 496
##         5) hour > 1
##           9) SO2.Mean <= 0.545455; criterion = 0.978, statistic = 10.354
##             10)*  weights = 2176
##           9) SO2.Mean > 0.545455
##             11)*  weights = 2118
##       4) month == {1}
##         12) day <= 1; criterion = 0.999, statistic = 16.171
##           13) NO2.1st.Max.Hour <= 10; criterion = 1, statistic = 18.215
##             14)*  weights = 21
##           13) NO2.1st.Max.Hour > 10
##             15)*  weights = 10
##         12) day > 1
##           16)*  weights = 264
##     3) SO2.Mean > 1.185714
##       17) SO2.Mean <= 4.416667; criterion = 0.98, statistic = 20.824
##         18)*  weights = 7245
##       17) SO2.Mean > 4.416667
##         19)*  weights = 5757
##   2) hour > 18
##     20) SO2.Mean <= 3.3125; criterion = 1, statistic = 49.785
##       21) CO.1st.Max.Value <= 0.8; criterion = 0.999, statistic = 29.073
##         22)*  weights = 2070
##       21) CO.1st.Max.Value > 0.8
##         23) month == {1, 2}; criterion = 1, statistic = 51.972
##           24)*  weights = 195
##         23) month == {3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
```

```
##           25)*  weights = 843
##     20) SO2.Mean > 3.3125
##       26)*  weights = 1971
## 1) day > 12
##   27)*  weights = 34970
```

Hide

```
plot(output.tree)
```



> UFO sightings seem always to happen when the day of month is less than 13. Interestingly, there is one node, for which the probability of seeing a UFO equals 100%. We will now test the predictive power of our decision tree.

Hide

```
test$ET2 <- ifelse(test$ET==1, 'yes', 'no')
test$ET2 <- as.factor(test$ET2)
fitted.results <- Predict(output.tree, newdata = subset(test, select=-c(ET)))
misClasificError <- mean(fitted.results != test$ET2)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.9902"
```

> *The decision tree achieves exactly the same accuracy as both linear and logistic models. Unfortunately for our attempts to test whether air quality influences the UFO sightings, the time and date seem to be more crucial than air pollution in regards to experience of an UFO sighting. Therefore we will also render a decision tree based solely on day, month and hour and test its accuracy.*

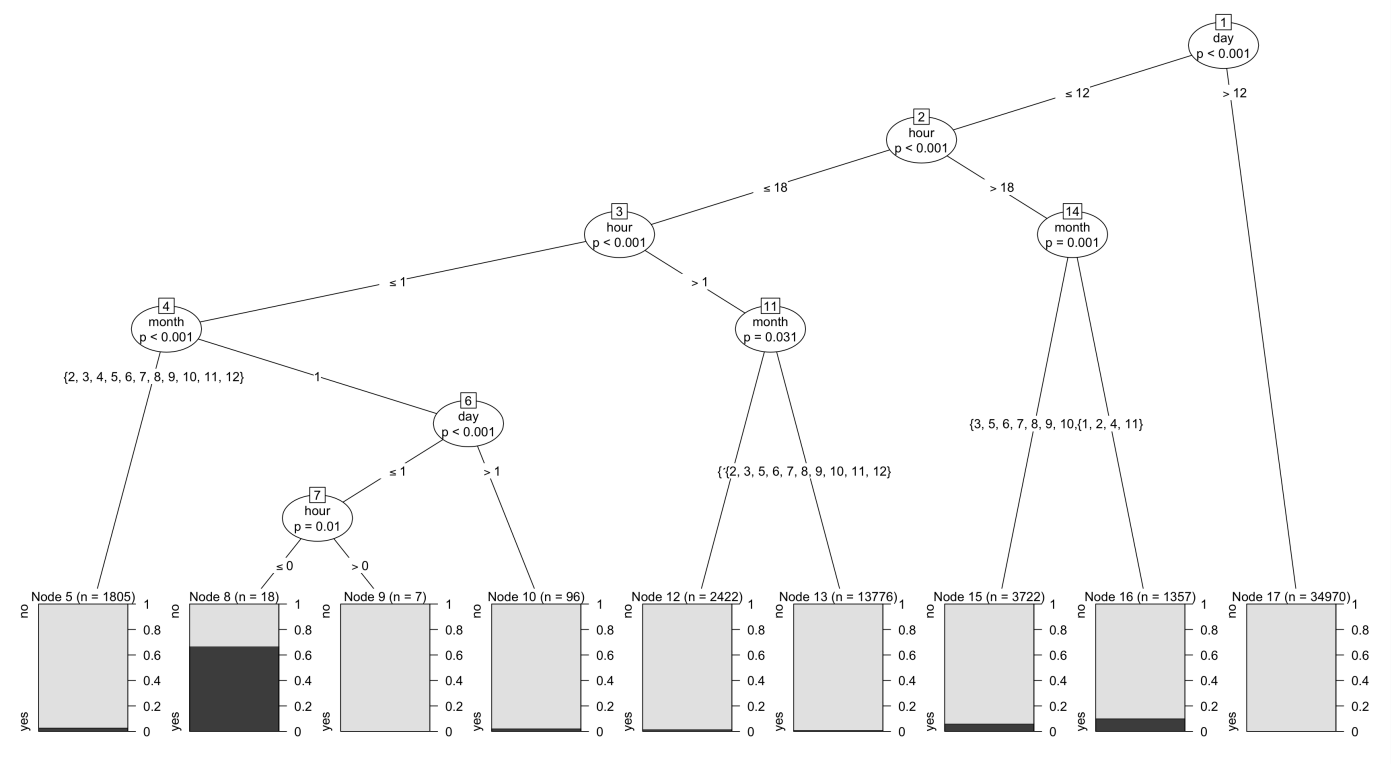
Hide

```
output.tree2 <- ctree(ET2 ~ ., data = subset(train, select=c(ET2, hour, day, month)))
output.tree2
```

```
##
##   Conditional inference tree with 9 terminal nodes
##
## Response:  ET2
## Inputs:  hour, day, month
## Number of observations:  58173
##
## 1) day <= 12; criterion = 1, statistic = 585.34
##   2) hour <= 18; criterion = 1, statistic = 208.587
##     3) hour <= 1; criterion = 1, statistic = 50.235
##       4) month == {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}; criterion = 1, statistic = 40.453
##         5)*  weights = 1805
##       4) month == {1}
##         6) day <= 1; criterion = 1, statistic = 21.727
##           7) hour <= 0; criterion = 0.99, statistic = 8.615
##             8)*  weights = 18
##           7) hour > 0
##             9)*  weights = 7
##         6) day > 1
##           10)*  weights = 96
##     3) hour > 1
##       11) month == {1, 4}; criterion = 0.969, statistic = 24.588
##         12)*  weights = 2422
##       11) month == {2, 3, 5, 6, 7, 8, 9, 10, 11, 12}
##         13)*  weights = 13776
##   2) hour > 18
##     14) month == {3, 5, 6, 7, 8, 9, 10, 12}; criterion = 0.999, statistic = 33.324
##       15)*  weights = 3722
##     14) month == {1, 2, 4, 11}
##       16)*  weights = 1357
## 1) day > 12
##   17)*  weights = 34970
```

Hide

```
plot(output.tree2)
```



Hide

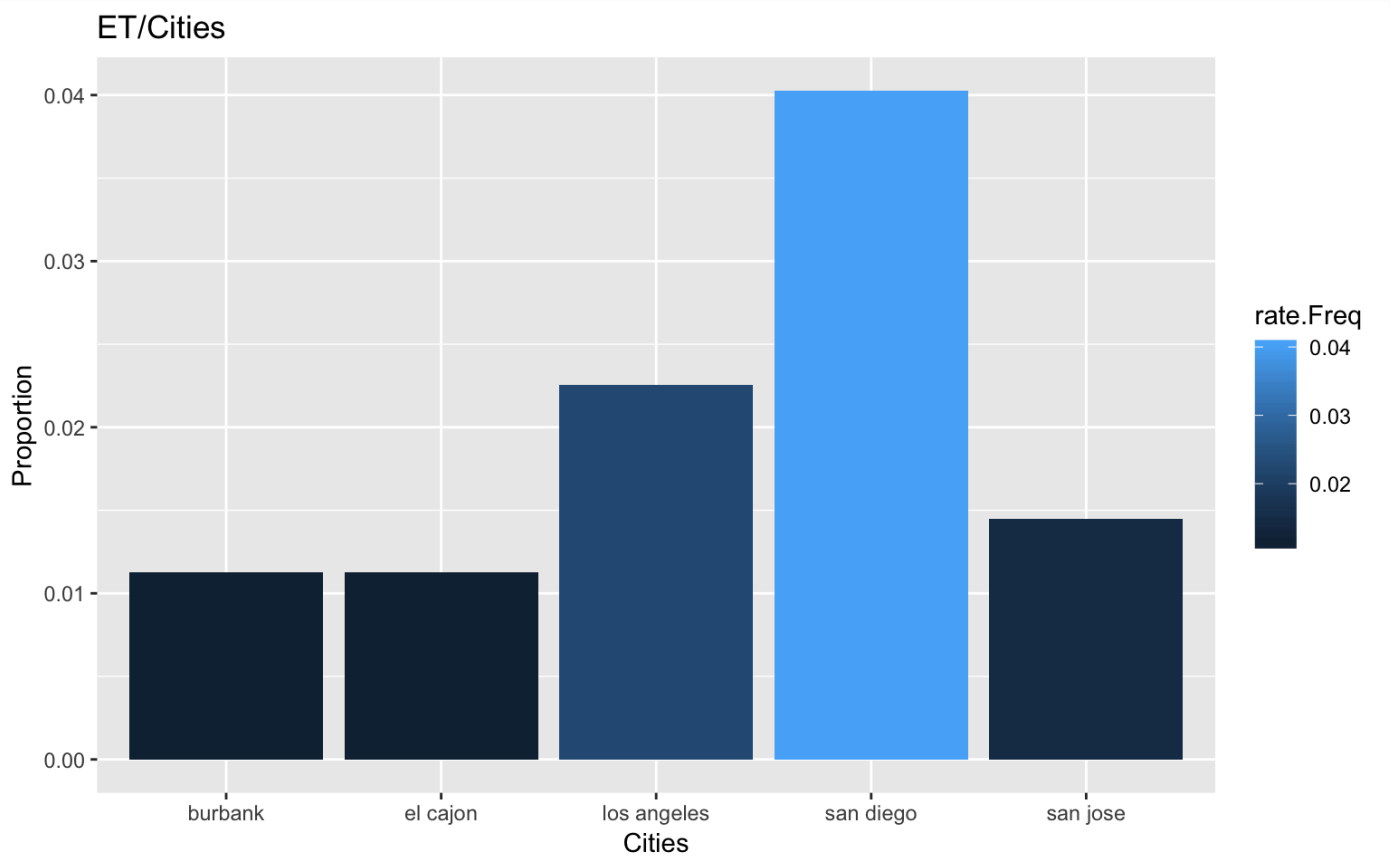
```
fitted.results <- Predict(output.tree2, newdata = subset(test, select=-c(ET)))
misClasificError <- mean(fitted.results != test$ET2)
print(paste('Accuracy',1-misClasificError))
```

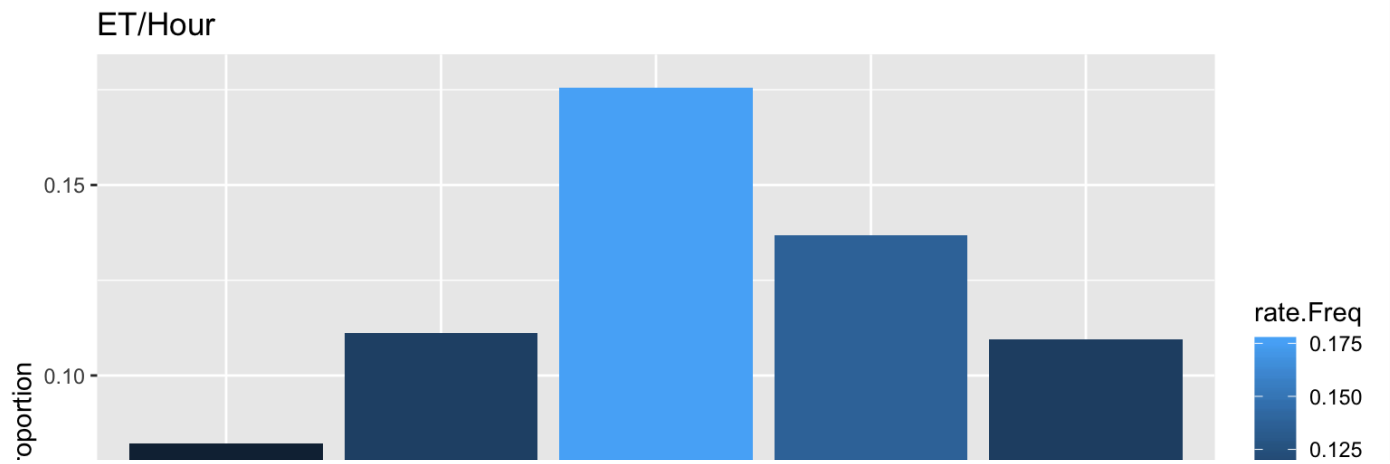
```
## [1] "Accuracy 0.9906"
```

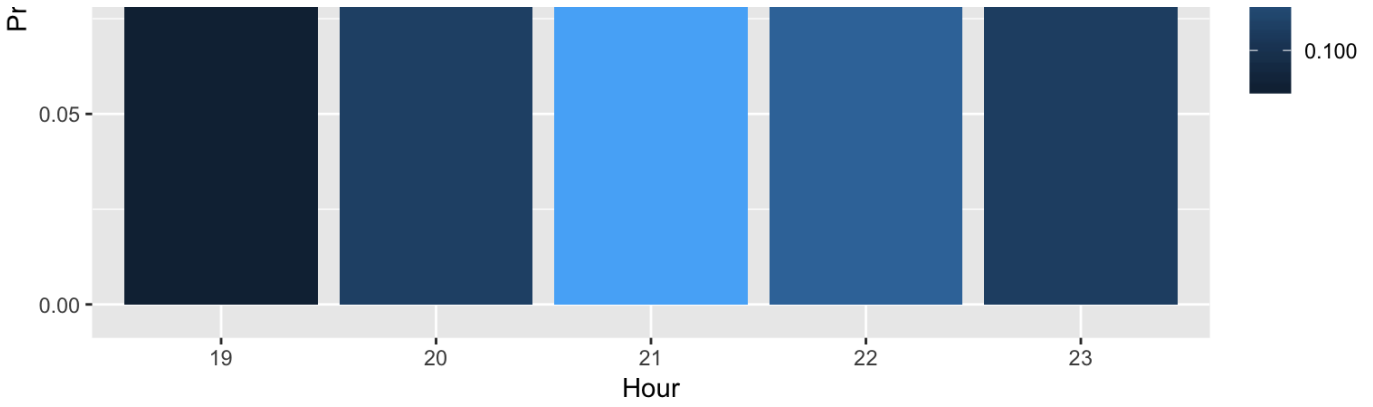
> *Again, there is one node with probability of seeing a UFO more than 50%. The odds for seeing a UFO seem to be the highest on 1st day of January exactly at midnight. It is worth noting that day preceding this date is 31st of December, which is New Year's Eve. We should keep in mind that this holiday usually involves drinking alcohol, which might result in short-term vision-altering effects. The accuracy of our predictive model slightly increased Sadly, this model does not tell anything about the influence of air quality on probability of seeing a UFO.*

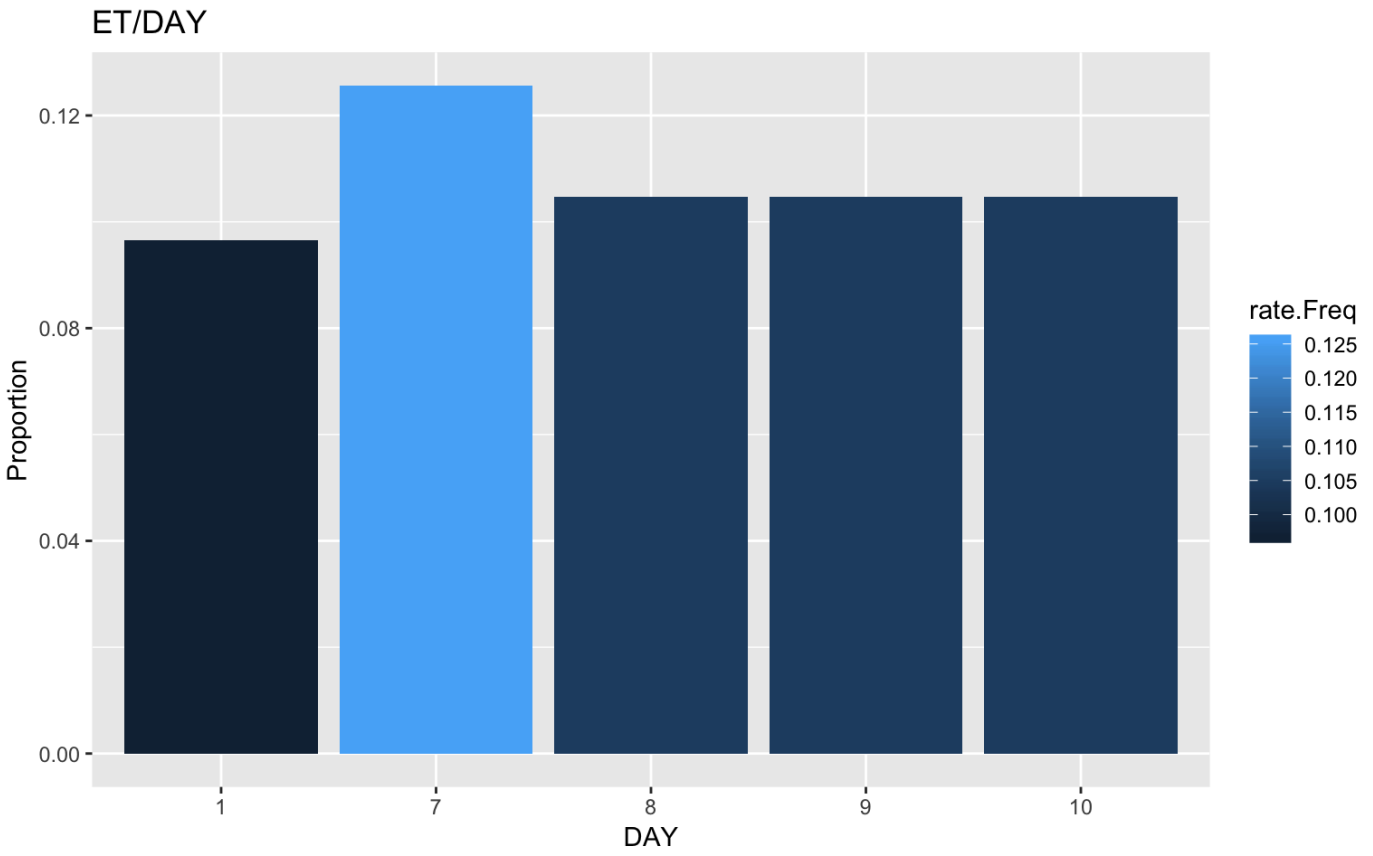
> *At the end we Will plot some basic graphs to check the rate of occurences by country and state in our general data set, and next we will do the same with clustered one, to check wether Clustering will be usefull for grouping our variables with respect to ET.*

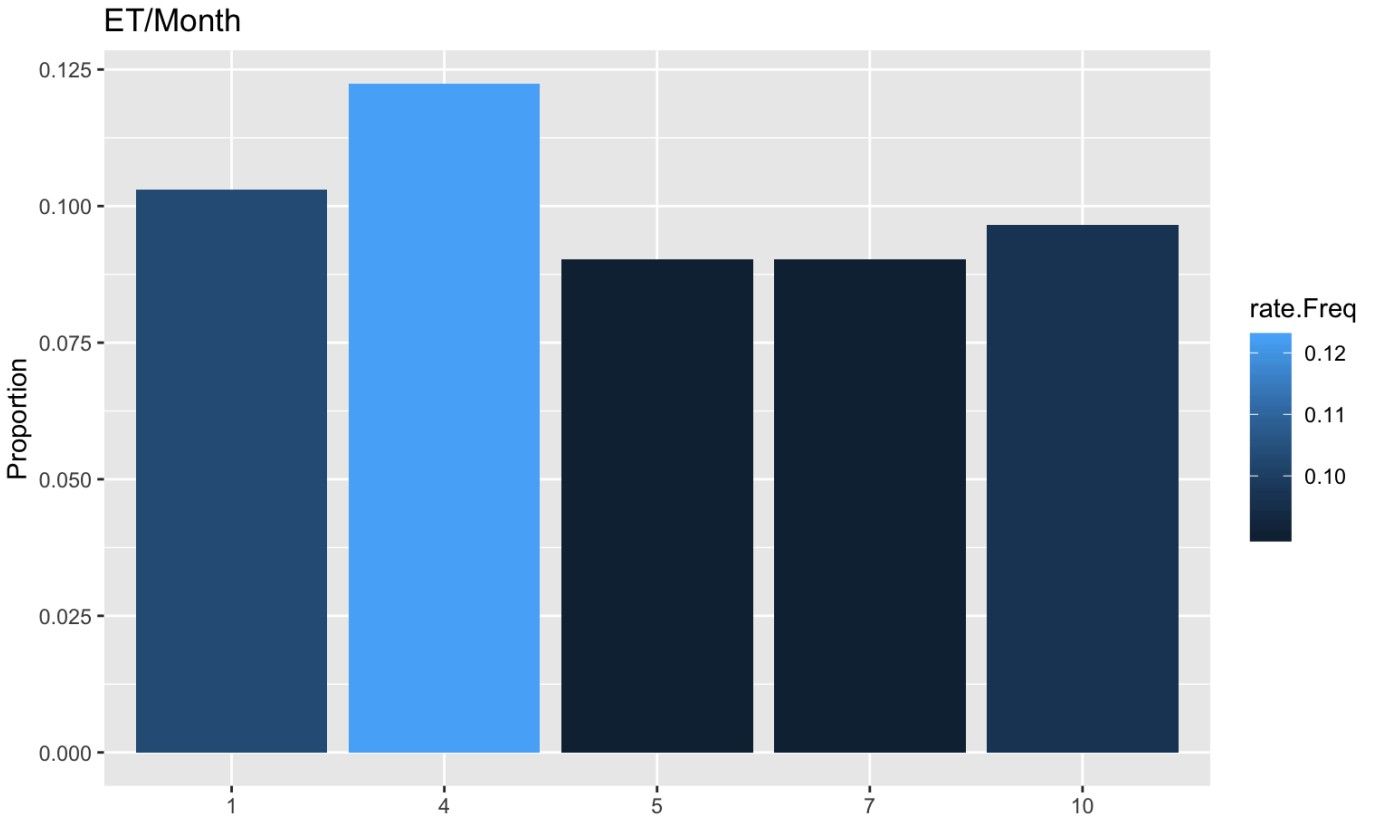
> *Now Let's plot it and see the results*






Pr
0.05
0.00
19
20
21
22
23
Hour
0.100

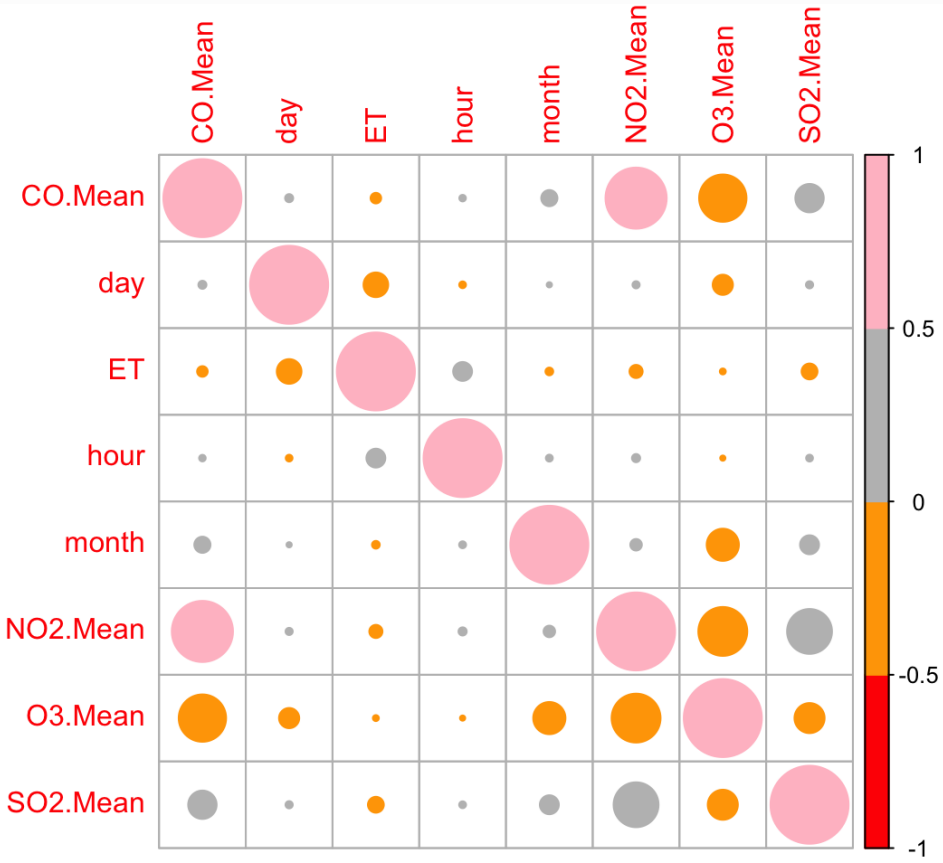

ET/DAY
0.12
0.08
0.04
0.00
Proportion
1
7
8
9
10
DAY
rate.Freq
0.125
0.120
0.115
0.110
0.105
0.100


ET/Month
0.125
0.100
0.075
0.050
0.025
0.000
Proportion
1
4
5
7
10
rate.Freq
0.12
0.11
0.10

Month

> *And check simple Correlation Plot*

```
## corrplot 0.84 loaded
```



> *We can derive from the correlation plot that NO2 and CO are highly positively correlated, SO2 and NO2 also, where O3 and CO are negatively correlated.*

> *So as we can see, if we go to San Diego on 7th of April aroud 21 local time, we have the highest probability to see UFO… :)*

4.2.4 Clustering

> *First of all let's start with simple Cluster Analysis, thus we will need some libraries*

Hide

```
library(cluster)
library(factoextra)
library(flexclust)
library(fpc)
library(clustertend)
library(ClusterR)
library(psych)
library(fitdistrplus)
library(logspline)
library(NbClust)
library(ggplot2)
library(reshape2)
library(gridExtra)
library(tadaatoolbox)
```

```
library(sjPlot)
```

> *Let's standarize the data*

Hide

```
data <- na.omit(data)
data.pollutants = as.data.frame(lapply(data[,8:24], as.numeric))
data.pollutants.s = scale(data.pollutants)
```

> *Now we will check optimal number of clusters for K-means and CLARA*

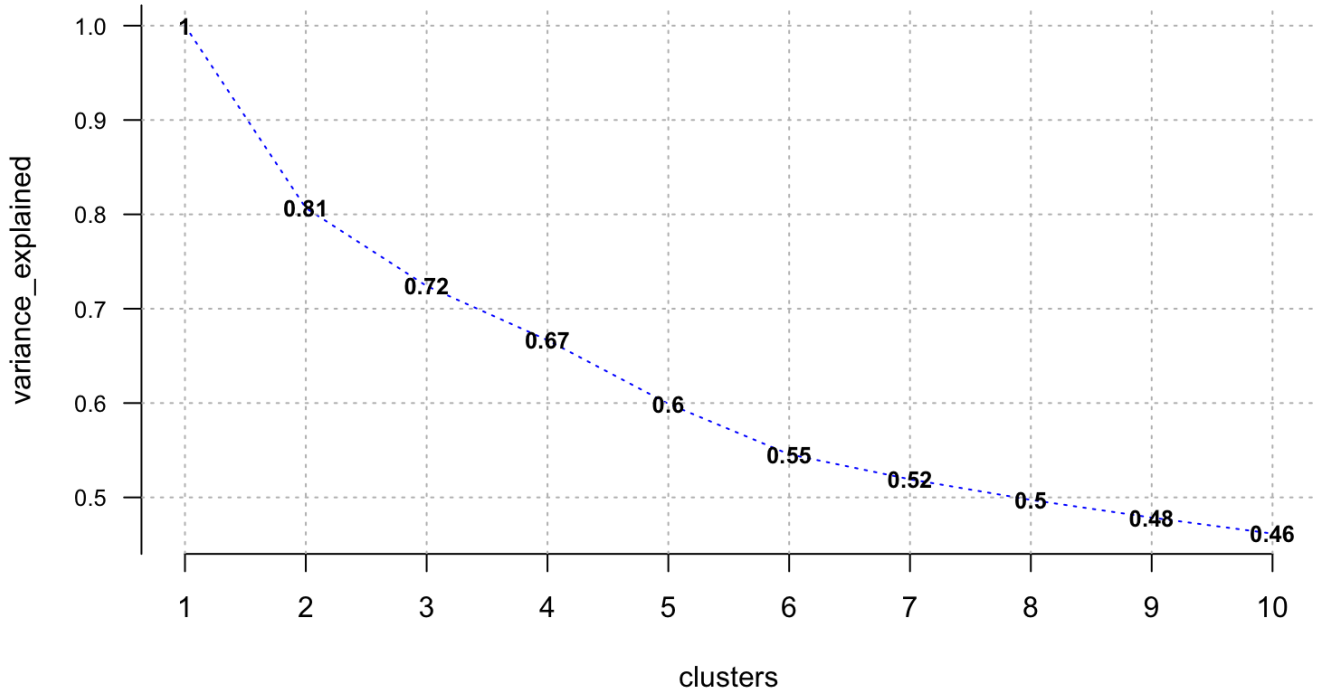
Hide

```
#optimal number of clusters using CLARA
clara.best = pamk(data.pollutants.s, krange=2:10,criterion="multiasw", usepam=FALSE, alpha=0.001, diss=inheri
print(clara.best)
```

```
## $pamobject
## Call:     clara(x = sdata, k = k)
## Medoids:
##        NO2.Mean NO2.1st.Max.Value NO2.1st.Max.Hour    NO2.AQI    O3.Mean
## [1,] -0.1278399        -0.2071605       -0.5588933 -0.2407692 -0.3175290
## [2,]  0.7944917         0.4841700        0.8369525  0.5395360 -0.9910459
## [3,]  0.1065860         0.1119151        1.2176377  0.1193717  0.8232522
##      O3.1st.Max.Value O3.1st.Max.Hour     O3.AQI   SO2.Mean
## [1,]       -0.2539366       0.1948620 -0.2807528 -0.4895510
## [2,]       -0.7676321      -0.0482277 -0.6367021 -0.1277774
## [3,]        1.3442274      -0.2913174  1.0095637  0.5386477
##      SO2.1st.Max.Value SO2.1st.Max.Hour    SO2.AQI    CO.Mean
## [1,]        -0.6148565       -0.5446387 -0.6632308 -0.3375608
## [2,]        -0.1548183        1.2354525 -0.1428479  0.6588803
## [3,]         1.3786425       -0.3962977  1.4761214 -0.6516584
##      CO.1st.Max.Value CO.1st.Max.Hour     CO.AQI         ET
## [1,]       -0.2418226      0.29316740 -0.2209789 -0.1057523
## [2,]        1.0080317      1.73324643  1.0138423 -0.1057523
## [3,]       -0.5542861      0.05315423 -0.4953836 -0.1057523
## Objective function:   3.471163
## Clustering vector:    int [1:15822] 1 1 2 1 1 2 1 2 2 2 2 2 1 1 1 1 1 1 ...
## Cluster sizes:            9360 3470 2992
## Best sample:
##  [1]   179   638  1595  2070  2111  2345  2673  2764  2823  3060  3302
## [12]  3468  3485  5043  5355  5564  5715  5725  6338  6908  7859  7862
## [23]  8205  8348  8679  9366  9612 10025 10089 10648 11400 11429 11439
## [34] 11656 11702 11828 11981 12116 12978 13341 13411 13777 14149 14690
## [45] 14711 15470
##
## Available components:
##  [1] "sample"     "medoids"    "i.med"      "clustering" "objective"
##  [6] "clusinfo"   "diss"       "call"       "silinfo"    "data"
##
## $nc
## [1] 3
##
## $crit
##  [1] 0.00000000 0.17773545 0.18318481 0.13060202 0.09119887 0.08513966
##  [7] 0.12752930 0.12126435 0.08269518 0.11749882
```

Hide

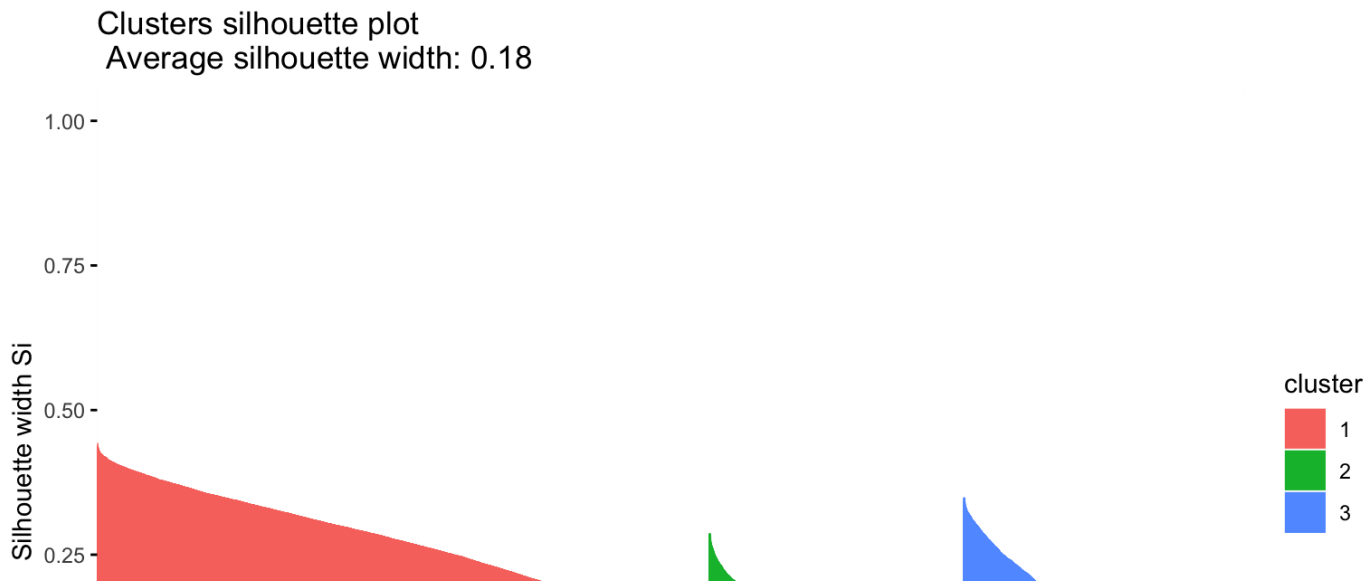
```
#optimal numer of clusters using K-means
optimal_set = Optimal_Clusters_KMeans(data.pollutants.s, max_clusters=10, plot_clusters = TRUE)
```

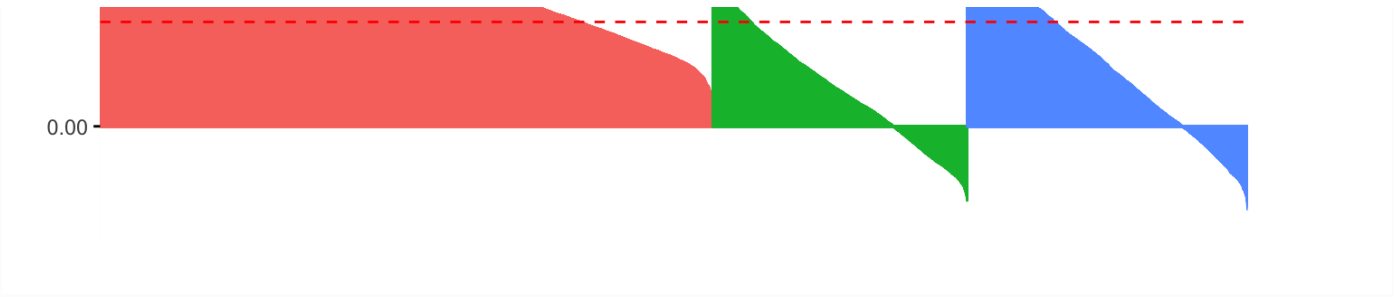


> *As we can see optimal number of clusters for CLARA and K-means is respectively 3 and 2. However, for this big data set we will make it equal, setting K-means number of clusters also to 5. Now let's take a look on the silhouette of Kmeans and CLARA*

> ***K-means silhouette***

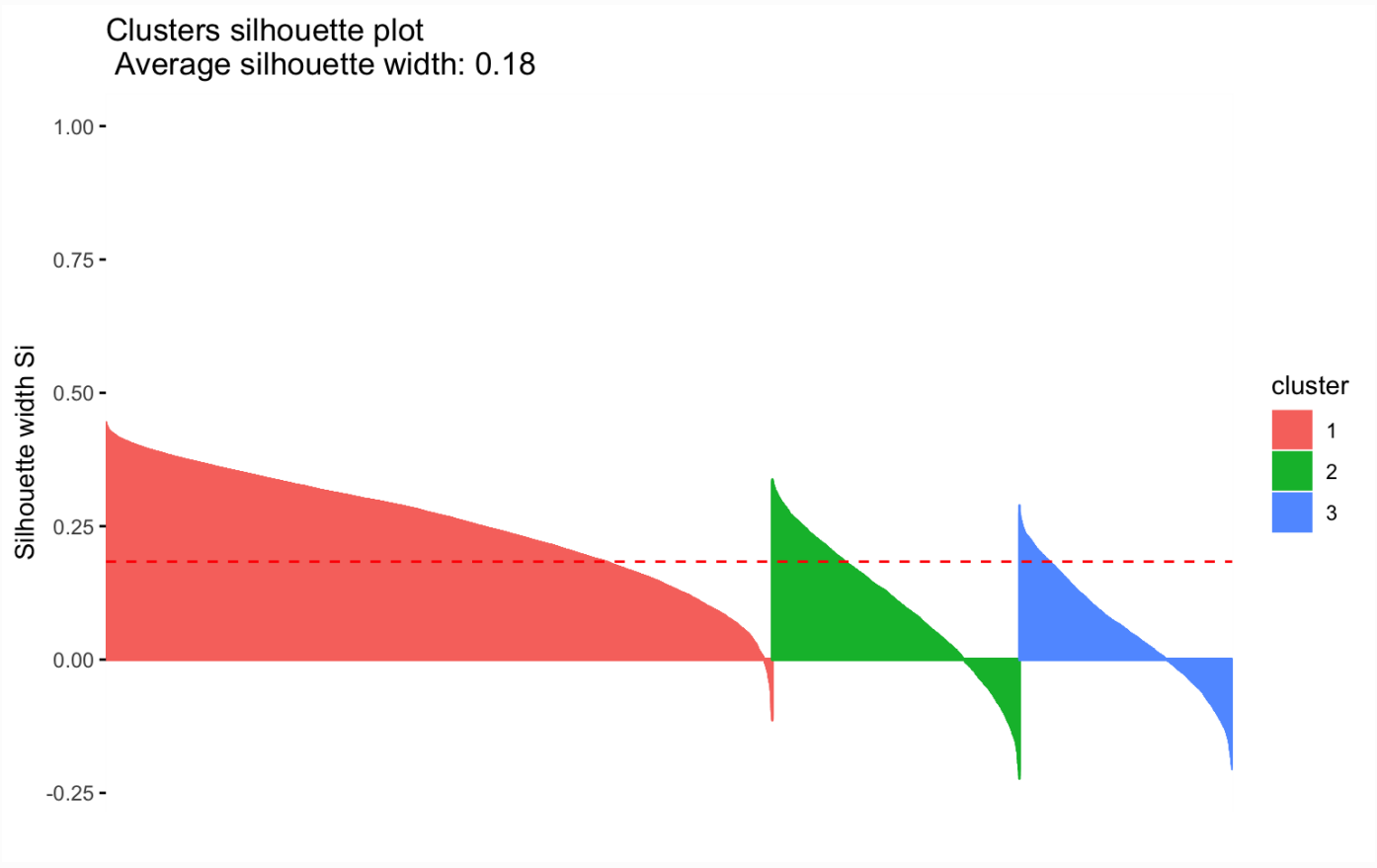
```
##   cluster size ave.sil.width
## 1       1 8446          0.26
## 2       2 3508          0.07
## 3       3 3868          0.11
```





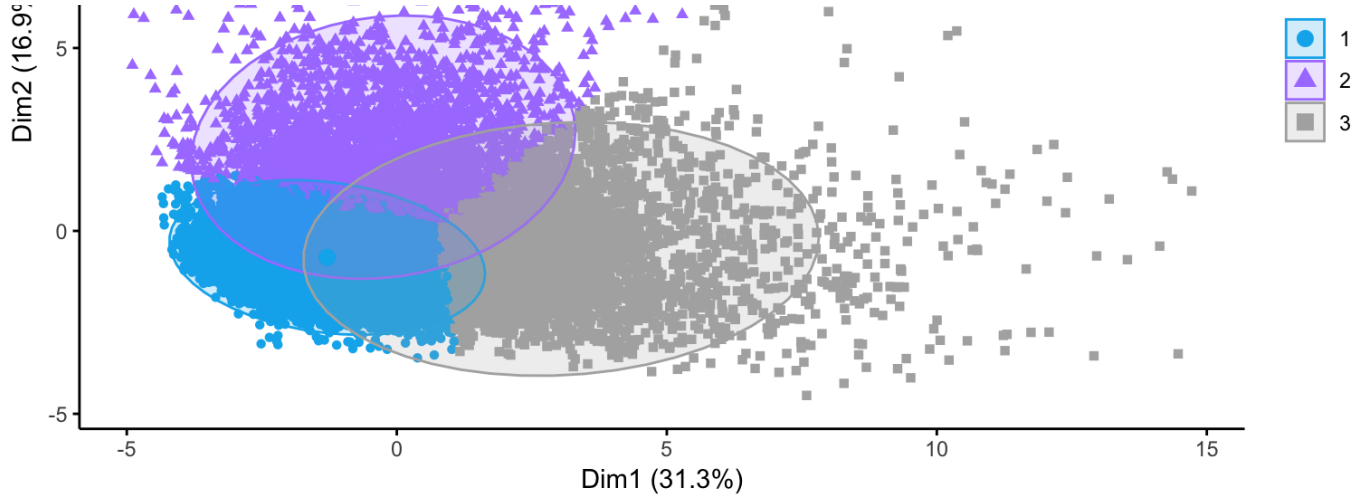
CLARA silhouette

```
##   cluster size ave.sil.width
## 1       1 9360          0.25
## 2       2 3470          0.10
## 3       3 2992          0.06
```



Analysing the silhouettes above, we will opt for K-means method as it provides better distribution with less minus silhouettes.

Therefore we will plot only K-means solution



Concluding from the plot, we can see a large number of outliers, in next step we will assign number of cluster to each row of the data set

Here we divide our data sets to 3 smaller data sets with respect to cluster number, to check statistics, and occurences of pollutants to check which ones are dominating

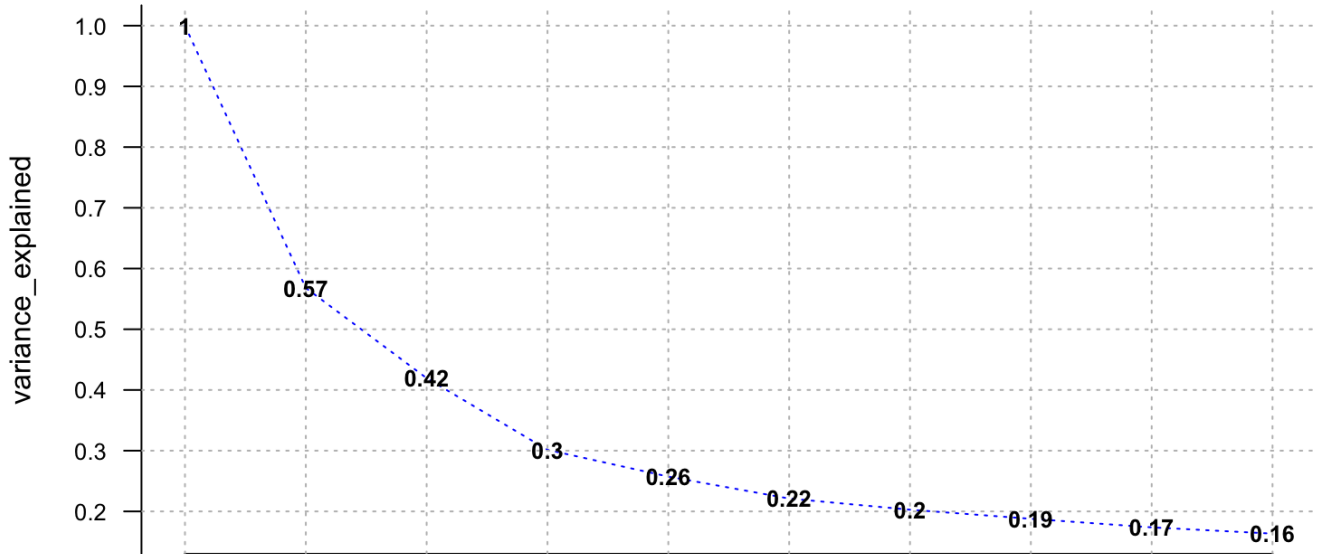
Let's describe each

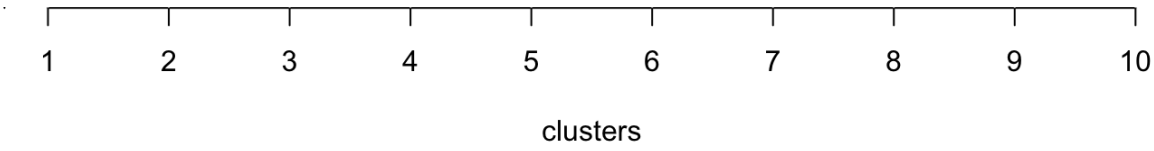
Cluster no.1

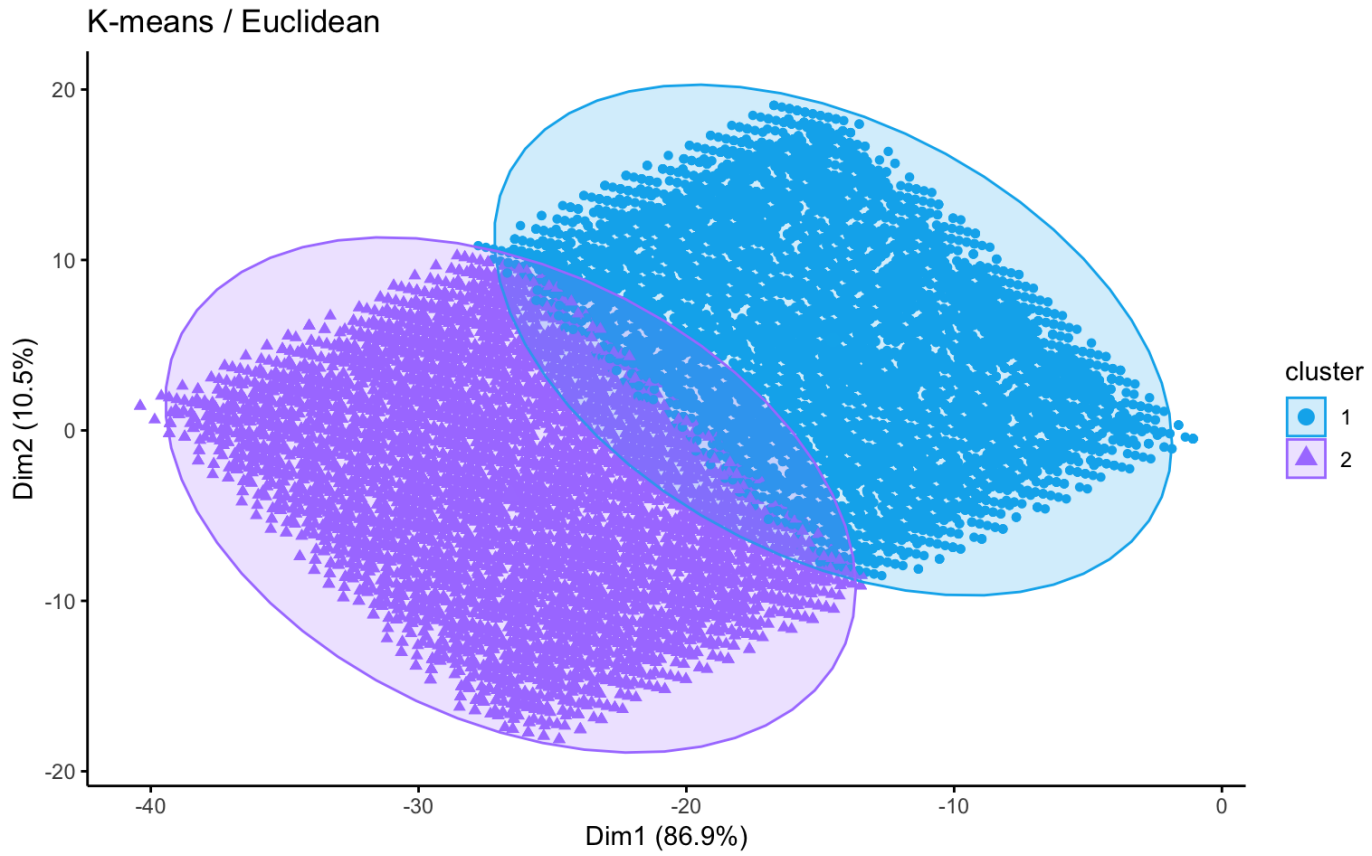
Cluster no.2

Cluster no.3

As we can see Cluster 2 and 3 are similiar to each other in terms of data. However as we conducted in the previous analysis, its pointless to check for Air pollutants in terms of UFO Sightings, because we concluded that there is no significant relation between those variables. Thus, we will run clustering in terms of hour,date, month and add city to compare with previous results.







Cluster no.1

Cluster no.2

We can see that basicaly these clusters are pretty much the same, Thus Cluster analysis in this case is useless.

5 Conclusion

Summarizing, we have proven that in our dataset, generaly speaking, Air Pollutants are not the base of UFO Sightings. Thus, we reject our Thesis that ***"There is existing connection between air pollution level and UFO Sightings"****.. However, we proven that there is relation between UFO Sightings and variables like month,date and hour. Unlikely as we could see in the first analysis, it is kinda misleading beacause, it showed that the night of 31th December - 1st of January is the most probable date to see UFO (100%), but we can explain it due to the New Year party and excessive alcohol intake.*

Rejecting Hypothesis that air pollutants doesn't affect UFO Sightings, we also rejected our derivetive goals such as: Indication the reasons of the sigthtings (influence of the air pollutants to cognitive abilities. As we cannot say that air pollutants are the main factor to cause fake-UFO sightings, we cannot derive from it any other conclusions.

Happily we defined that In positive correlation with UFO Sightings are the hour, where of course if its darker, the more UFO we see.

> *We also have shown frequency ratio histograms to get more insights by the chosen factors (date, month,hour and city).*

> *Despite the fact that we could not prove any relation between air pollutants and UFO Sightings, in some aspect because of data set, we made full analysis and found out some basic, although funny insights.*

6 References

https://www.arcgis.com/apps/webappviewer/index.html?id=ddda71d5211f47e782b12f3f8d06246e

https://en.wikipedia.org/wiki/Ig_Nobel_Prize

https://en.wikipedia.org/wiki/Ig_Nobel_Prize

https://www.datanovia.com/en/lessons/clara-in-r-clustering-large-applications/

https://en.wikipedia.org/wiki/K-medoids

https://en.wikipedia.org/wiki/Regression_analysis

https://en.wikipedia.org/wiki/Correlation_and_dependence

https://en.wikipedia.org/wiki/Analysis_of_variance

https://en.wikipedia.org/wiki/Logistic_regression

https://en.wikipedia.org/wiki/Decision_tree

https://prologue.blogs.archives.gov/2018/04/16/ufos-natural-explanations/

https://www.pnas.org/content/115/37/9193

https://www.nps.gov/subjects/air/visibility.htm