SIT718 – REAL WORLD ANALYTICS

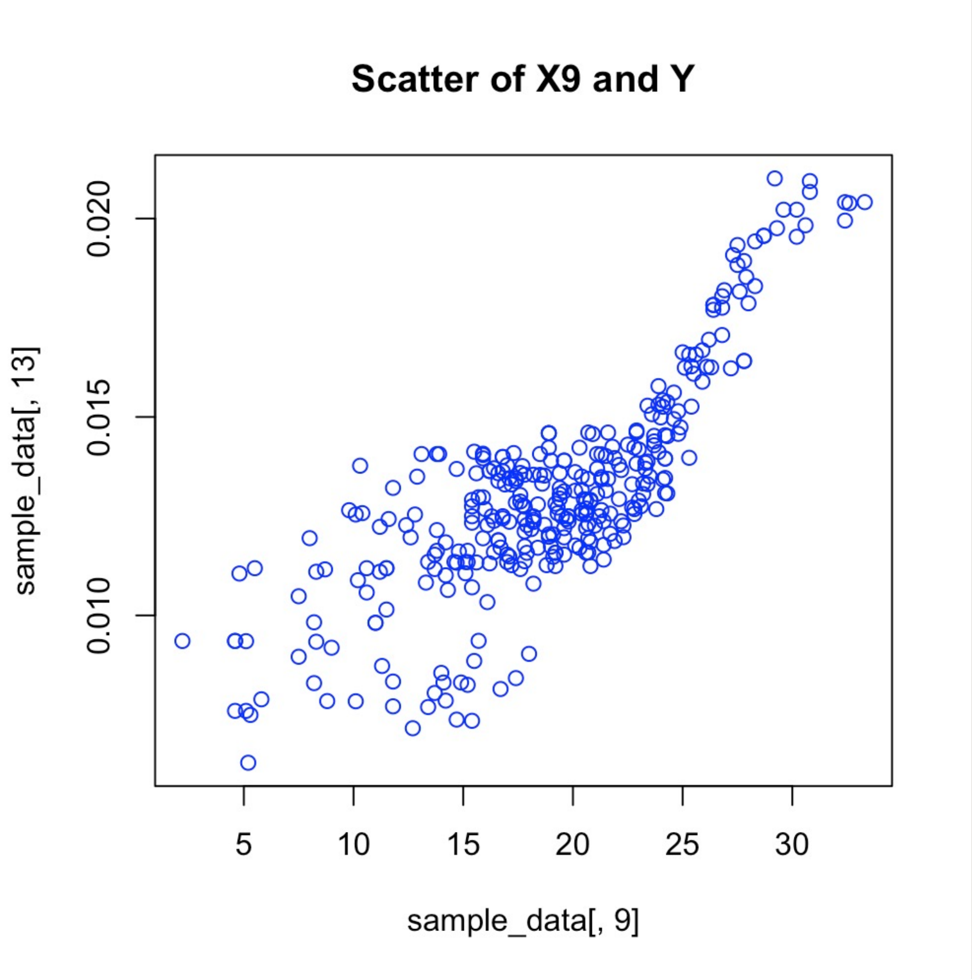
(ASSIGNMENT 2 – TASK 5)

PRAJITH RAJ SURESH

S220156351

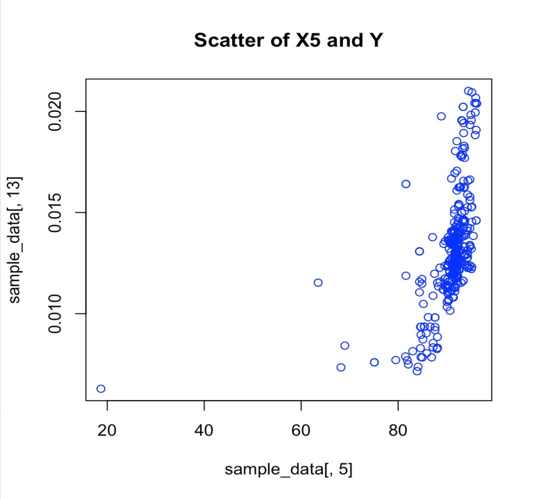
TASK 1: UNDERSTANDING THE DATA

- *Correlation between the variables:-* The correlation between the X variables to Y(area of burned area) provides the relationship and the impact of the variable X on Y. A positive and higher correlation value results in a stronger effect of X on Y and a negative correlation shows that it has a minimum impact on Y. In this case, we see the values are increasing at around 25.

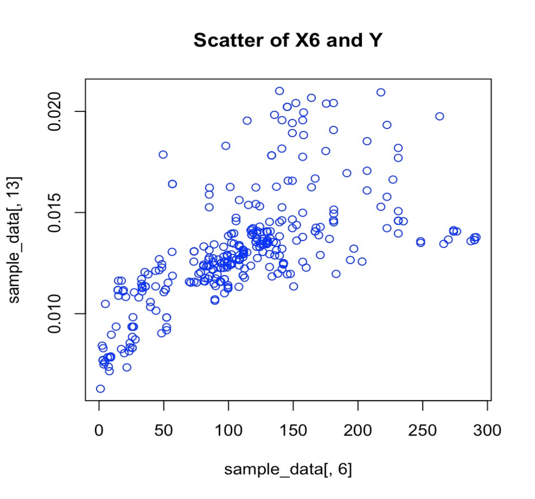


cor(X9,Y)=0.813521

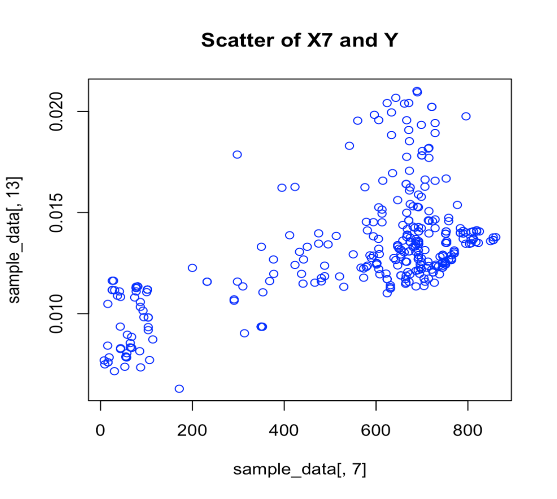
SCATTER PLOTS AND THEIR RELATIVE CORRELATION VALUES



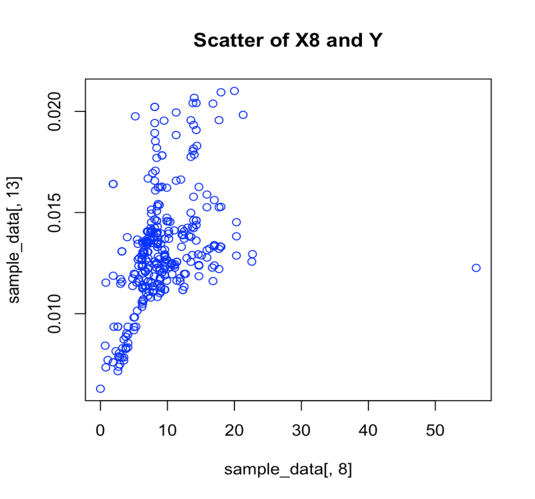
cor(X5,Y)=0.5078001



cor(X6,Y)=0.6471276

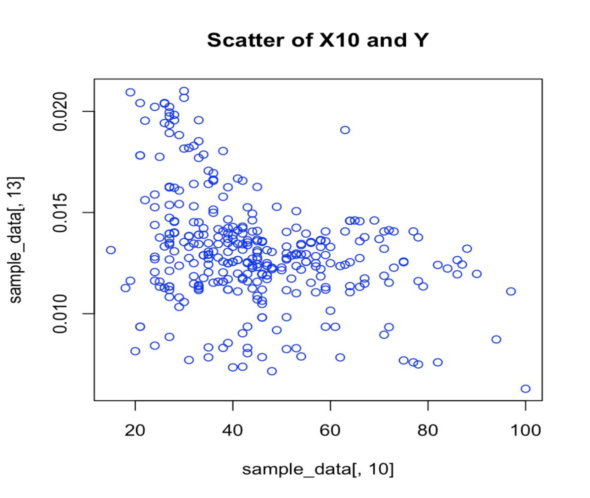


cor(X7,Y)=0.605262

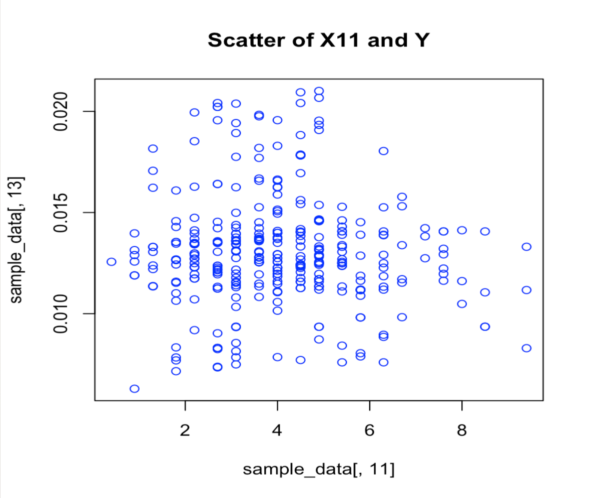


cor(X8,Y)=0.4216162

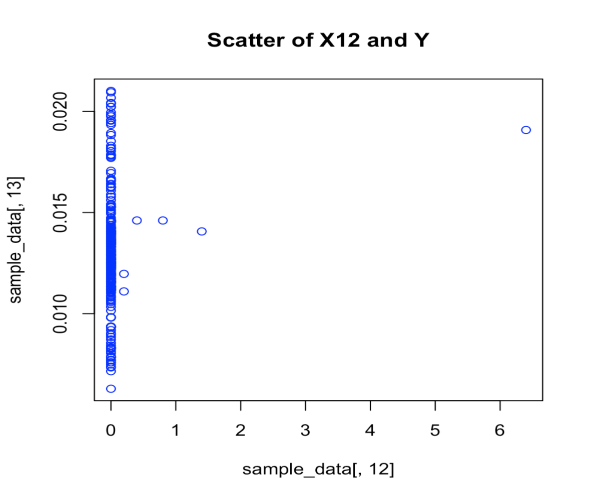
SCATTER PLOTS AND THEIR RELATIVE CORRELATION VALUES



cor(X10,Y)= - 0.3367396



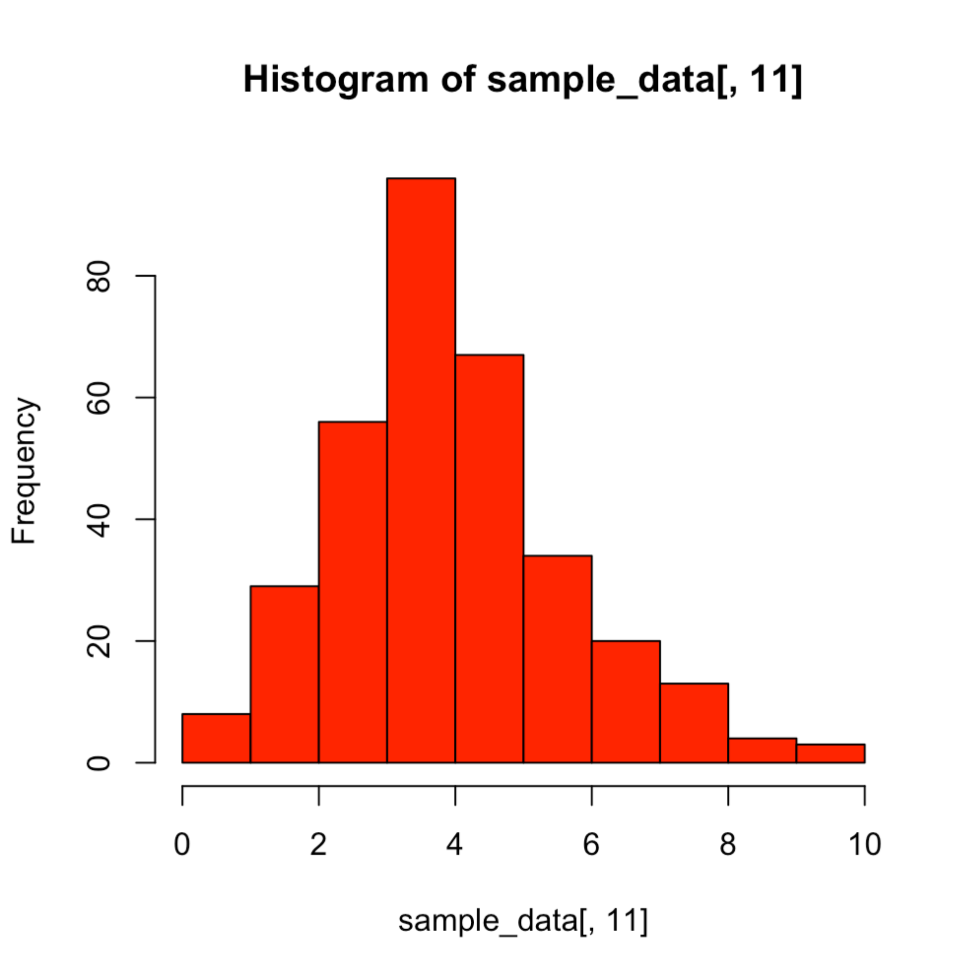
cor(X11,Y)= - 0.06075109



cor(X12,Y)= 0.1244347

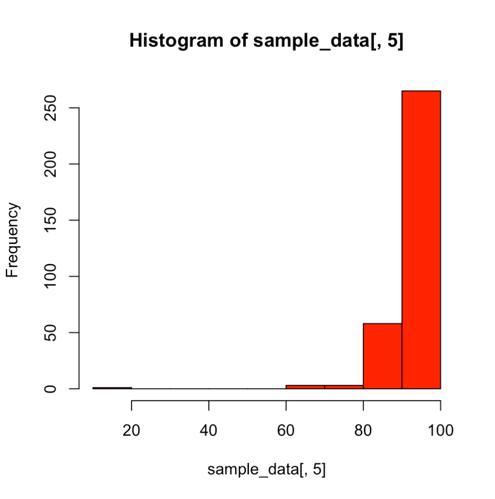
DISTRIBUTION OF DATA WITH HISTOGRAMS

- The distribution of data can be found with the help of histograms. In the figure provided for X11, we can see that that there are more values with wind speed between 3 and 4 kmph. This has a skewness of 0.5433574 and it is a Normal Distribution(close to being Normal).

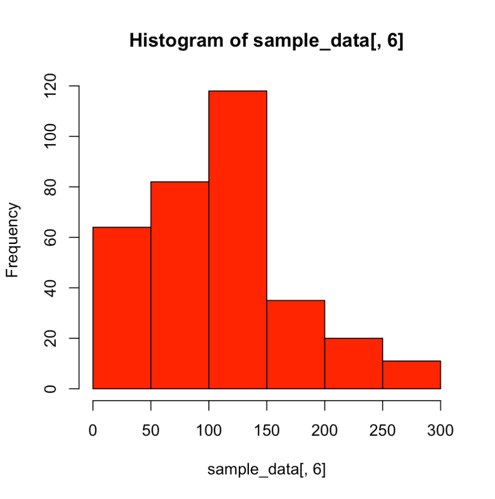


Skewness = 0.5433574

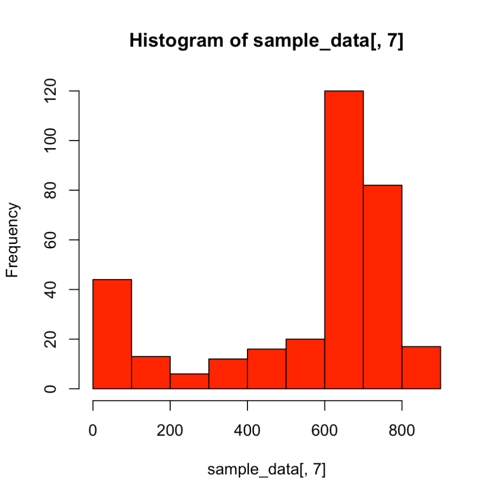
DISTRIBUTIONS AND HISTOGRAMS OF DATA



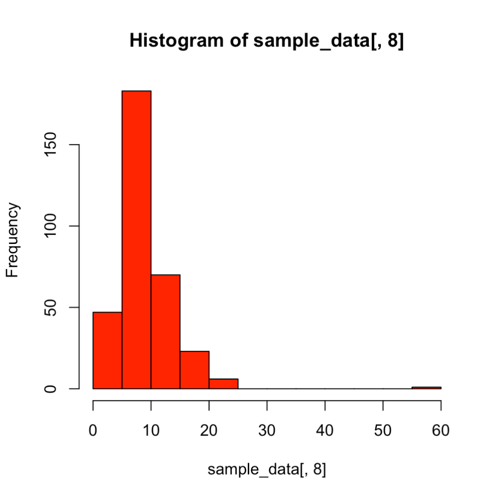
Negatively Skewed Distribution
Skewness= -7.430441



Slightly Positive Distribution
Skewness=0.5358228

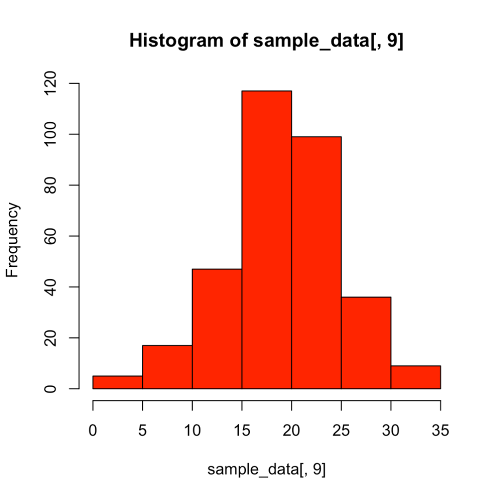


Negatively Skewed Distribution
Skewness= -1.136233

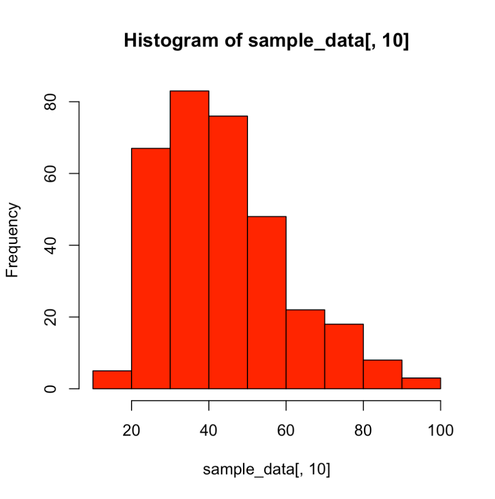


Positively Skewed Distribution
Skewness=2.99424

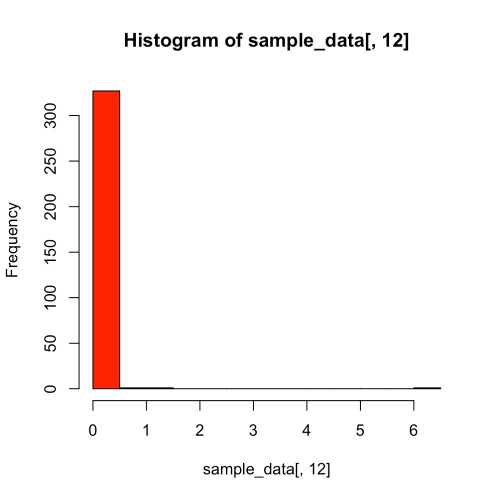
DISTRIBUTIONS AND HISTOGRAMS OF DATA



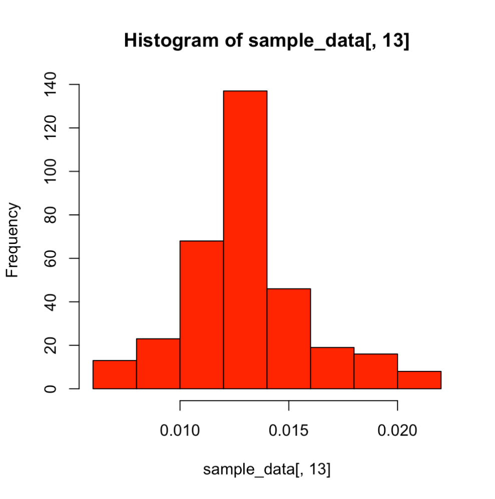
Slightly Negatively Skewed Distribution
Skewness= -0.243186



Slightly Positive Distribution
Skewness=0.8749845



Positively Skewed Distribution
Skewness= 16.47988



Slightly Positively Skewed Distribution
Skewness=0.5349772

TASK 2: - TRANSFORMING THE DATA

- The four variables chosen for the transformation are **X6, X9, X10, X11**.
- The transformation applied on each variable help to scale down and standardize with the same units.
- X6: **Polynomial and Linear Scaling.**
- X9: **Linear Scaling.**
- X10: **Negation, Polynomial and Linear Scaling.**
- X11: **Negation, Polynomial and Linear Scaling.**
- X13: **Polynomial and Linear Scaling.**

ERROR MEASURES, CORRELATION COEFFICIENTS AND SUMMARY OF DATA (CHOQUET MODEL)

ORNESS : 0. 506050111885507

SHAPEY VALUES:
0.461740045149982,0.302985694058228, 0.157736441077396,0.0775378197143942

RMSE	0.0904894357022566
Average Absolute Error	0.0703562952074658
Pearson Correlation	0.873278131628413
Spearman Correlation	0.879202963861407

Number of Digits	Binary number fm.weights
1	0.289099857253279
2	0.417931082479919
3	0.83882269368044
4	0.173473206702576
5	0.378346190488353
6	0.449375478439664
7	1
8	0
9	0.597899884314362
10	0.417931082479919
11	0.83882269368044
12	0.173473206702576
13	1
14	0.449375478439664
15	1

IMPORTANCE OF EACH CHOSEN VARIABLE

- X6: The **DMC(Duff Moisture Code)** value in the soil represents moisture conditions for the equivalent of 15-days and provide insight to live fuel moisture stress.
- X9: **Temperature** plays a vital role in the fast spread of forest fire, wherein, extreme temperatures can further fasten and widen the spread of the forest fire.
- X10: **Relative Humidity** can be seen as a subset of high temperatures. Lower the relative temperature, higher is the temperature in the air which causes dryness in the air, thereby, helping the spread of the fire.
- X11: **Wind** plays a vital role in spreading the fires from one part to the other. The more the wind, the easier the fire spreads.

TASK 3: - BUILDING MODELS AND FINDING THE BEST FIT MODEL

- This task involves building models with aggregate functions keeping our transformed data in hand.
- On applying with different aggregate functions such as **Weighted Arithmetic Mean, Weighted Power Mean with p-value: {0.5, 2}, Ordered Weighted Average and Choquet Integral, the Choquet Integral** value came out with the best results.
- **The Choquet Integral** was chosen over with evidences from having the least **RMSE, Avg. Abs. Error** and the best value for **Pearson and Spearman Correlations.**

TASK 4: PREDICTION FOR THE DATA MODEL

- The **Predicted value** for the variable of interest **Y(Area of forest burnt)** by using the best fit model of Choquet Integral is **0.03486013**.
- The Actual value of Y which was provided as a value measure was **0.0146** and in comparison, to the value of the predicted value, it only differs by a mere **0.02026013.**
- This is a highly reasonable value of prediction given it differs from the actual value by minute decimals.

BEST CONDITIONS FOR THE CHOSEN VARIABLES

- The best conditions for each of the chosen variable in terms of having a high effect on a large burned area are:
- X6 - **DMC(Duff Moisture Code):** A high value of DMC value ranging from (41-60 and 61+) will make for a high hazard and create a huge forest fire.
- X9 - **Temperature:** A Higher value of temperature will result in a more severe spread of forest fire by assisting with more heat.
- X10 - **Relative Humidity:** A lower value of Relative Humidity will result in making the air dry, thereby, raising the air temperature resulting in huge forest fires.
- X11 - **Wind:** The higher the value of wind speed, the greater is going to be the spread of the forest fire to different regions of the forest, resulting in large burned areas.

IMPLICATIONS AND LIMITATIONS OF THE CHOQUET INTEGRAL MODEL

- ***Choquet Integral*** is an aggregation function defined with respect to the **fuzzy measure**. A fuzzy measure is a set function, acting on the domain of all possible combinations of a set of criteria. The complexity is therefore exponential of $\mathbf{2}^{n}$ subsets, where *n* is the number of criteria.
- The **advantages** of the *Choquet Integral* is based on the use of fuzzy measure in its computation, which allows it to consider the i**nteraction between all possible combinations of criteria.**
- The **limitation** of a Choquet integral is, if a fuzzy measure is additive, the criteria do not interact with each other, and the interaction indices of these criteria are equal to zero. Therefore, if we think the criteria is mutually preferably **independent**, the corresponding interaction indices are equal to zero.
- If the expert suggests that the criteria are preferably **dependent**, then it is possible to formalize this only by means of partial weak order on the set of criteria realizations (training set). **No other method of formalization of the criteria preferred dependence and independence has not been proposed**.

REFERENCES

- "'UCI Machine Learning Repository: Forest Fires Data Set'. Archive.ics.uci.edu. N.p., 2017,http://archive.ics.uci.edu/ml/datasets/forest+fires."
- "Simon James(2016) An Introduction to Data Analysis using Aggregation Functions in R,Springer, Deakin University Library, Melbourne"
- https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system
- https://www.science.org.au/curious/earth-environment/things-you-need-know-about-bushfire-behaviour
- https://wildfire.alberta.ca/wildfire-status/fire-weather/understanding-fire-weather.aspx
- https://www.researchgate.net/publication/266687131_Problems_of_Choquet_Integral_Practical_Applications

THANK YOU!