Assessment Task 3: Problem solving task 2: Using aggregation functions for data analysis

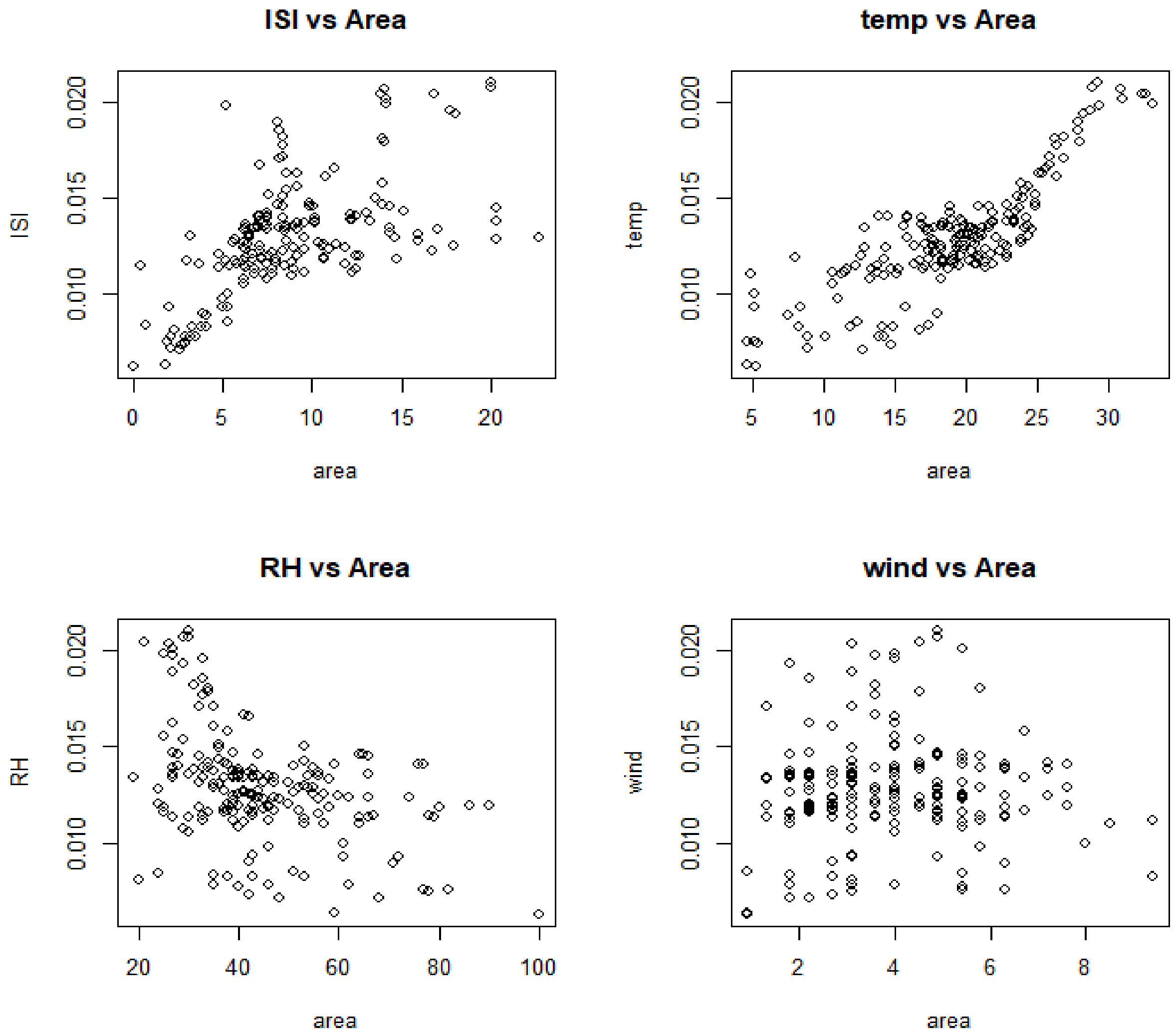
Q1. Understand the data:

(i) And (ii)-This has been done in the R Code.

(iii)Histograms of various variables and the variable of interest Y

For this I have chosen 4 variables as X8: ISI - ISI index from the FWI system, X9: temp - temperature in Celsius degrees, X10: RH - relative humidity in % and X11: wind - wind speed in km/h with the variable of interest X13 i.e. Y: area - the burned area of the forest (in ha).

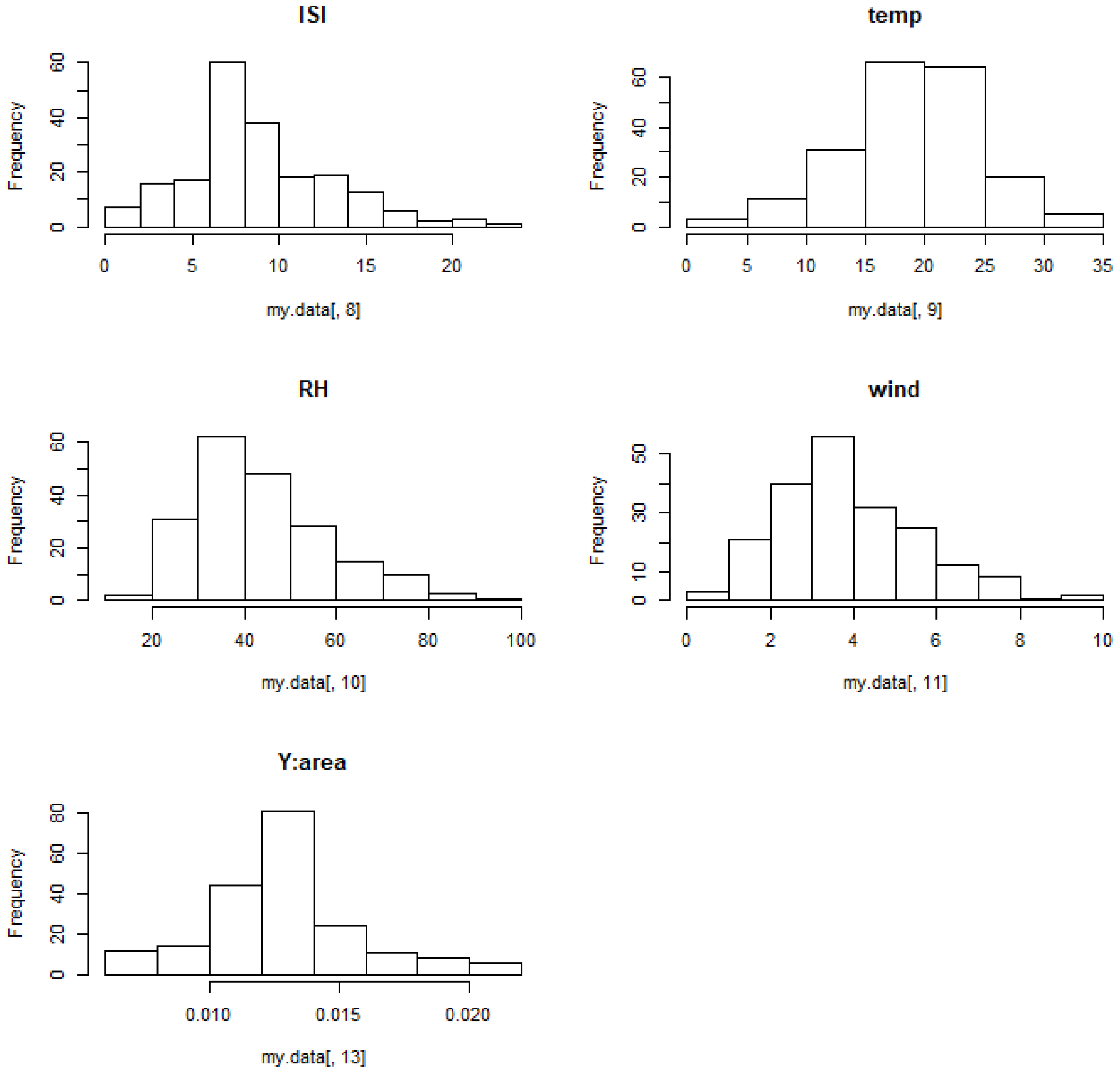
The scatterplots are shown below: -



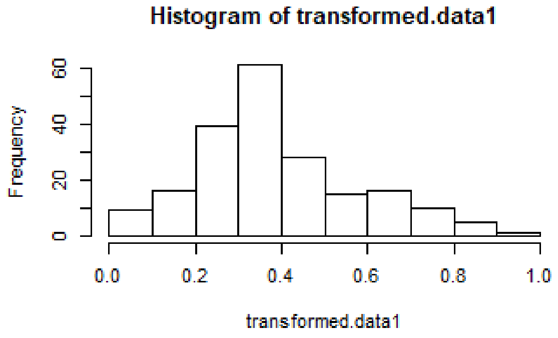
1. The first scatterplots i.e. ISI vs Area shows moderate positive correlation as the burned area increases with increase in ISI and then the scattering of points occur showing moderation in correlation.
2. The Second scatterplots i.e. temp vs Area shows strong positive correlation as all the points align in an increasing diagonal with increase in temp to direct increase in burned area.

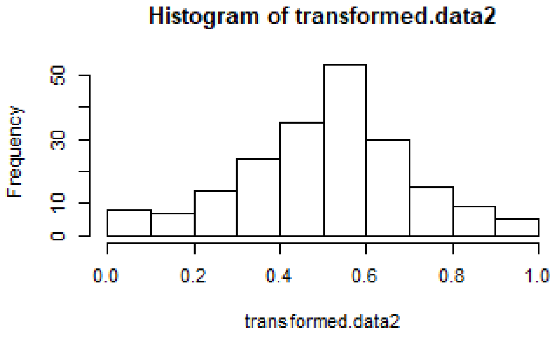
3. The Third scatterplots i.e. RH vs Area shows negative correlation as here with the decrease in Relative humidity there's an increase in the Area Burned.
4. The Fourth scatterplots i.e. wind vs Area shows no correlation since all the points are scattered at random without any close matches.

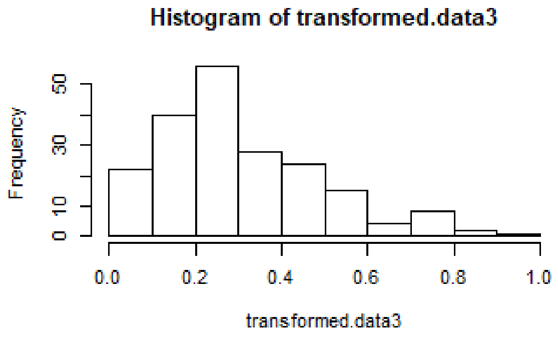
The histograms of all the 5 variables {V8, V9, V10, V11,V13}are shown below(BEFORE SCALING BETWEEN 0 AND 1): -

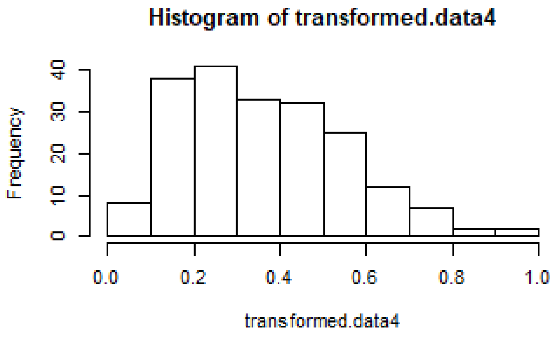


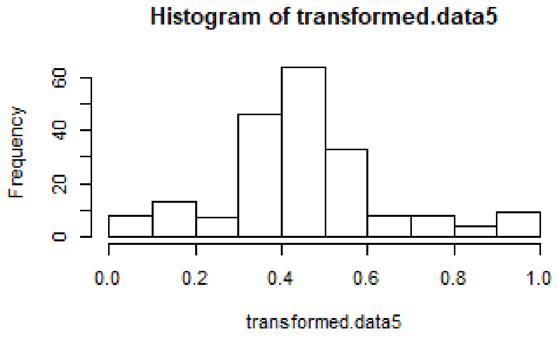
The Histogram after Scaling the graph to the interval [0,1] for normalization purposes is as: -











1. For the first histogram for variable X8 i.e. ISI the graph appears to be skewed right i.e. positively skewed and unimodal with higher values populated near the left end of the graph (between 0.2 and 0.4).
2. For the second histogram for variable X9 i.e. temp the graph appears to be unimodal and somewhat symmetric but acutely skewed on the right at the right end with higher bin widths around 0.5 and 0.6.
3. For the third histogram for variable X10 i.e. RH the graph is skewed on the right i.e. positively skewed and unimodal with the peaks of the graph on the left end between 0.2 and 0.3.

4. For the fourth histogram for variable X11 i.e. wind the graph is skewed right i.e. positively skewed and unimodal with greater accumulation of high values on left.
5. For the fifth histogram for our variable of interest X13 i.e. area the graph appears to be symmetrical and unimodal with highest burned area between 0.4 and 0.5.

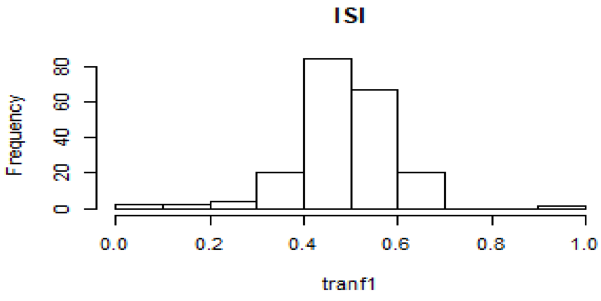
Q2. Transform the data

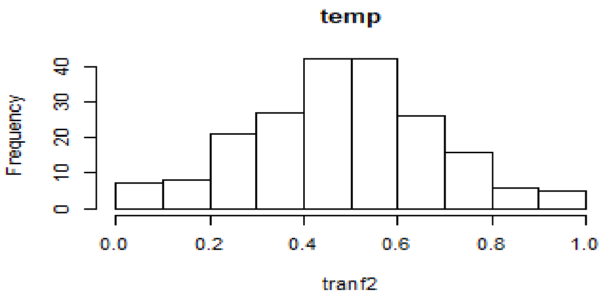
(i) The transformation I applied was Polynomial Transformations where I did t^p(where t is the data) and I chose the value of P to be between 0<*p*<1 as most of the graphs I encountered were positively skewed and had fewer very high values.

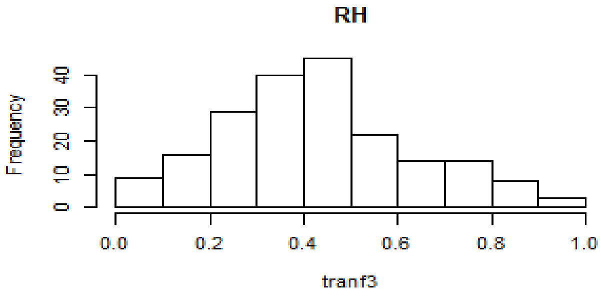
So, the transformation goes like:

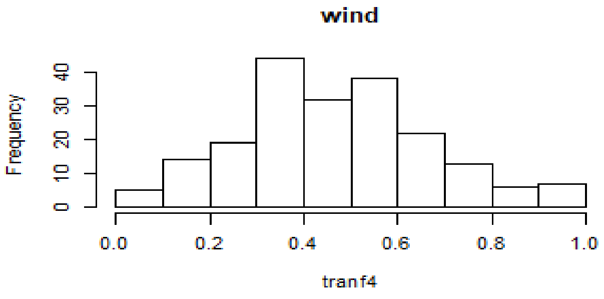
- `tranf1=tranformed.data1^0.37`
- `tranf2=tranformed.data2^1.1`
- `tranf3=transformed.data3^0.7`
- `tranf4=transformed.data4^0.75`
- `tranf5=transformed.data5^0.9`

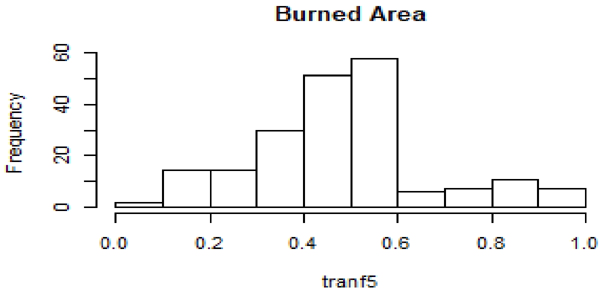
(ii)











1. From the ISI it is evident that it shows positive relationship with the Burned Area i.e. at the middle when the ISI index is higher the burned area is more or increased.
2. From the temp also shows positive relationship with the Burned Area which is with rise in temperature the burned area is more and gradually decreasing with fall in temp.
3. For the Relative humidity(RH) it shows negative relationship as with gradual decrease in humidity there's increase in burned area.
4. With the wind there is no specific relationship as all the points (in scatterplot) or the values(histogram) don't follow a common pattern for evaluation which means wind has no/random effect on the area burned.

Q3. Build models and investigate the importance of each variable

(i) Done in R Code.

(ii) Done in R Code.

(iii)

- Error Measures of my data are: -

	WAM	PM05	PM2	OWA	Choquet Integral
RMSE	0.204667057386343	0.205038998975752	0.204527833289059	0.227225380347858	0.204205284105195
Av. abs error	0.152652588851742	0.15229388216696	0.152133042970558	0.170009313096586	0.150107547738476
Pearson Correlation	0.0852215382759432	0.0905006407772445	0.089136643567343	-0.12688979908082	0.0661316093432207
Spearman Correlation	0.0641964956431061	0.0677471336743995	0.0654910459688375	-0.109634638716523	0.0609713016715941
Orness				0.476092473235156	0.462963730677258

- Weight/parameters from the data are: -

	WAM	PM05	PM2	OWA	Choquet Integral (shapley measures)
X8: ISI	0.856629524964571	0.831042834039886	0.899970930578058	0.340314517985807	0.777031703737445
X9: temp	0	0	0	0	0.0359509484869069
X10: RH	0.00531239879796766	0.00616316057063664	0.00810111035149304	0.550779026337116	0.0656290519733372
X11: wind	0.138058076237461	0.162794005389477	0.0919279590704485	0.108906455677078	0.121388295808774

- Weights of Choquet integral: -

Index	binary number fm.weights	Binary numbers	Fuzzy measures
1	0	0001	V({1})
2	0.564591520748351	0010	V({2})
3	0.611828272513259	0011	V({1,2})
4	0	0100	V({3})
5	0	0101	V({1,3})
6	0.564591520748351	0110	V({2,3})
7	0.611828272513259	0111	V({1,2,3})
8	0	1000	V({4})
9	0	1001	V({1,4})
10	0.564591520748351	1010	V({2,4})
11	0.747777811857352	1011	V({1,2,4})
12	0	1100	V({3,4})
13	0.164819948552539	1101	V({1,3,4})
14	0.564591520748351	1110	V({2,3,4})
15	1	1111	V({1,2,3,4})

(iv)

a. The model I used is Choquet integral and it is good or better than others as it has lowest RMSE and average absoluter error and moreover it has values of Pearson and spearman correlation closer to 1 w.r.t. other models. So, it shows and provides the best fitting model w.r.t. other models.

b. The importance is evident from the weights each of the variable carry and the most important out of these is variable X8 i.e. ISI index from FWI system as it has the highest weight of 0.777031703737445 and then comes X11: wind after which comes X10: RH and at last the least important according to weight is X9: temp which is 0.0359509484869069.

c. Interactions:
   - V[1]+V[2]= 0.564591520748351<V[1,2]= 0.611828272513259→So, this is redundant.
   - V[1]+V[3]=0=V[1,3]→So, this is additive.
   - V[2]+V[3]= 0.564591520748351=V[2,3]→So, this is additive.
   - V[2]+V[4]= 0.564591520748351=V[2,4]→So, this is additive.
   - V[3]+V[4]=0=V[3,4]→So, this is additive.
   - V[1,2]+V[3]= 0.611828272513259=V[1,2,3]→So, this is additive.
   - V[1,2]+V[4]= 0.611828272513259<V[1,2,4] =0.747777811857352→So, this is redundant.
   - V[1,3]+V[4]=0<V[1,3,4]= 0.164819948552539→So, this is redundant.
   - V[2,3]+V[4]= 0.564591520748351=V[2,3,4]→So, this is additive.
   - V[1,2]+V[3,4]= 0.611828272513259<V[1,2,3,4]=1→So, this is redundant.

   There seems to be no complimentary interactions just redundant and additive.

d. Better models may favour the inputs that are important along with their relationships, coefficient signs, and effective magnitude. So, I would say that better models do favour higher inputs as they create a significant contribution to the aggregation functions and may play a major part in signifying the value of the output due to their relative higher magnitudes.
Moreover, if the higher inputs contain higher weights the aggregation performed may be biased on one-side and the graph may be skewed towards the lower inputs.
In a case where lower inputs have greater weights; in that case there may be a tendency that the higher magnitude of higher inputs may balance out the product (weight and lower input) of the lower input. But, to sum up at last a good model would favour higher inputs.

Q4. Use your model for prediction

(i) **Done in R Code.**

(ii) The output I obtained is 0.01435504 and I think it is reasonable as evident from the Forest718.txt database the values of Burned area where X8 i.e. ISI =7.6 is 0.013589446997002 and the output obtained is around the original data value.

(iii) The ideal conditions under which a lower burned area will result is, if the variables X8 i.e. ISI and X9 i.e. temp decrease (as positive correlation) at a rate and the variable X10 i.e. Relative humidity increase (as negative correlation) at the same rate irrespective of the X11 i.e. the wind(as it has no specific relation to Y:area)→in order to decrease/lower the total Burnt area of Forest Fires.
So, ISI and temperature would affect the burned area most and could make it maximal whereas relative humidity affects the burned area inversely and could make it minimal.