Exam No. B125405

Assignment 3

SUMMARY

Clustering Summary

KMeans clustering and Hierarchical clustering algorithms are applied. Both algorithms suggest 5 clusters, when number k is either undefined (Hierarchical Clustering) or optimum is found iteratively. The clusters correspond to 2 warm climate zones, of which one has more even weather throughout the year and more rain and other has sharper contrasts between winter and summer weather. 1 cluster is true medium of all the metrics. 2 colder climate zones, one with cool summers and mils winters, other with higher temperature range.

KMeans clustering with K=3 and utilizing variables most correlated with the latitude produce clusters, that correspond quite well to North, Middle, South geographical division.

Ideas for Further Research

Variables preserving temporality of climate could be designed, for example by month. This could lead to discovering different clusters, for example based on the rainiest season.

Classification Summary

Random Forest and Linear Regression algorithms applied to the dataset. Linear Regression with three variables: TempAVG, TempMax and TotalSun far outperforms Random Forest on a test dataset. This can be explained by strong linear relationship between the three variables and the latitude, as well as small number of features and small size of the dataset. Prediction accuracy with Linear Regression algorithm equals 87% on test training dataset and 80% on test dataset.

In conclusion, there can be observed some relationship between higher happiness levels int the UK and weather factors that correlate to Southern latitudes. However, the highest levels of happiness do not follow the same trend. In general, it is not possible to say if there is a relationship between weather and level of happiness. For example, it is possible that Southern part of UK has higher income rate, that could account for higher happiness levels.

Ideas for Further Research

Variables preserving temporality of climate could be designed, for example by month might help to classify more fine-grained regions, for example North West, North East, Middle, South West and South East. Other weather variables (for example, Wind) could be included.

Happiness – Weather Summary

The average rating of happiness per region is divided into three levels “Low”, “Medium” and “High”. Northern Ireland region has significantly higher average happiness rating then the rest of the UK. No significant weather factors were identified to have a relationship with the phenomena. Visually analyzing distribution of “Medium” or “Low” ratings in the rest of the UK reveals concentration of “Medium” ratings in the Southern part. Random Forest algorithm is applied to test if happiness rating levels can be classified using weather variables, because relationship is not linear. Ten fold cross-validation repeated ten times is applied instead of splitting dataset into training and testing parts, because of small size of dataset. The result is 81.5 % accuracy in classifying happiness rating “Low” or “Medium” with TempMax, TempMin, TotalSun and Totalrain variables.

Ideas for Further Research

Other than weather metrics could be explored for the regions that correspond to “High”, “Medium” or “Low” average ratings or percentages of “Very High”, “High”, “Medium” and “Low”. For example, income rate, health.

Automatization Summary

Almost all tasks are automated using R programming language. The “MAIN” is the primary script, where scripts with all the steps performed are included. The “MAIN” script is created to resemble a “table of contents” and illustrate steps performed instead of “one button” solution. It includes 9 scripts from downloading the files to happiness prediction script. The scripts for each step further include functions to perform necessary tasks. All the tasks performed can be replicated from the “MAIN” script. All individual steps and associated functions are described in the report.

Ideas for Further Optimization

Functions could be further generalized, so one function could be used for larger number of tasks. Scripts that include the functions could be refactored into functions that take other functions as an argument. This would reduce amount of code and allow putting parameters in one place, where they could be edited more easily.

Table of Contents

Downloading Weather Files

Files are downloaded using URL list as an input. URL list is obtained from the source of web page https://www.metoffice.gov.uk. Files are written to .txt files in "./MAIN/Data/" folder.

The task is automated with script "/MAIN/Scripts/download_weatherfiles.R"

Weather Data Cleaning

First objective of data cleaning is to create data sets in line with tidy data guidelines:

- one row for one observation;
- one column for one variable;
- consistent data within one table;
- informative variable names;
- only one value in a cell.

The following actions are performed:

- Text in the beginning of the file is removed
- Non-numeric values are removed from the data
- After initial cleaning the code was tested and attempt to convert data to numeric data type and convert to data frame produced warnings: (In as.numeric(current_file$TotalSun) : NAs introduced by coercion(…)In (function (..., deparse.level = 1) ... : number of columns of result is not a multiple of vector length (arg 1)). Debugging code has led to discovery of other non-numeric values in the data and code was adapted accordingly.
- NA inserted to represent missing values
- All data converted to numeric
- Data organized into data frames
- Data frame for every station written into separate .csv files for further analysis and adaptations
- Whitby data file fixed by inserting NA into empty cells, because in Whitby file missing data is not marked by "---".

The tasks are automated with script "/MAIN/Scripts/clean_weatherfiles.R"

Building Weather Data Frame

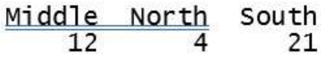
The goal is to build data frame, where each observation corresponds to one weather station. The approach taken is to preserve as much as possible data by taking middle values from all years. For dataset to be representative of seasonal weather specifics, only years where all months are available are included. As a result, created data frame contains no missing values. Median is taken instead of mean to avoid distortions due to possible outliers. In addition to existing variables, average temperature and yearly range are calculated from min and max temperatures.

All tasks are automated with script "MAIN/Script/build_weather_dataframe.R".

For each station following techniques are applied:

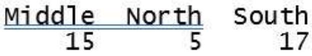
- Mean temperature is derived by adding minimum temperature and maximum temperature for every month and dividing by two. Then yearly mean and median for all years in the data set are calculated. Only the years with full 12 months in the data set are used, to avoid distortion that accounts for example, for only having coldest months available. The variable is representative of average warmth of climate disregarding seasonal fluctuations. The variable is named "TempAVG". Function "MAIN/Script/avgTempMedian_function.R" is used.
- To compensate for information that is lost by deriving average temperature, also temperature range is introduced. Temperature range variable is obtained by finding most extreme Tmax and Tmin of one year and subtracting Tmin from Tmax. Only years where data is available for all 12 months are used, to ensure the coldest and warmest temperatures are present. A median range for all years is found for every station. The variable is named "rangeMedian". Function "MAIN/Script/range_median_function.R" is used.
- Total amount of rain per year is calculated, then median amount of all years is found. Only years with full 12 months data are used. The variable is named "totalRainMedian". Function "MAIN/Script/totalRainMedian_function.R" is used.
- Lowest monthly minimal temperature of each year is found. Then median of all years' minimal temperatures is found. Only years with full 12 months data are used. Representative of winter temperatures. Function "MAIN/Script/tMinMedian_function.R" is used.
- Highest monthly maximum temperature of each year is found. Only years with full 12 months data are used. Then median of all years' maximum temperatures is found. Representative of summer temperatures. Function "MAIN/Script/tMaxMedian_function.R" is used.
- Total amount of sunny hours for each year is calculated. Only years with full 12 months data are used. Then median of all years available is found. Function "MAIN/Script/totalSunMedian_function.R" is used.
- Total count of frost days for each year is calculated. Only years with full 12 months data are used. Then median of all years available is found. Function "MAIN/Script/FrostDaysMedian_function.R" is used.
- Latitudes for each station extractes from .text files. Function "MAIN/Script/get_lat_function.R" is used.
- Longitudes for each station extractes from .text files. Function "MAIN/Script/get_long_function.R" is used.

- Stations are divided into thirds based on latitudes for classification algorithm. Extreme Northern latitude 60.9 and extreme Southern latitude 49.9 are added to the actual list of stations latitudes, because they are given as reference points in the assignment sheet. Function "MAIN/Script/get_labels.R" is used to assign labels to stations. Variables “Thirds” with coefficient rounded to 3 digits and “Thirds_4” with coefficient rounded to 0 digits are created. Comparing distribution of South, Middle and North for Thirds (picture 1) and Thirds_4 variables (picture 2).



```
Middle  North  South
    12      4     21
```

Picture 1



```
Middle  North  South
    15      5     17
```

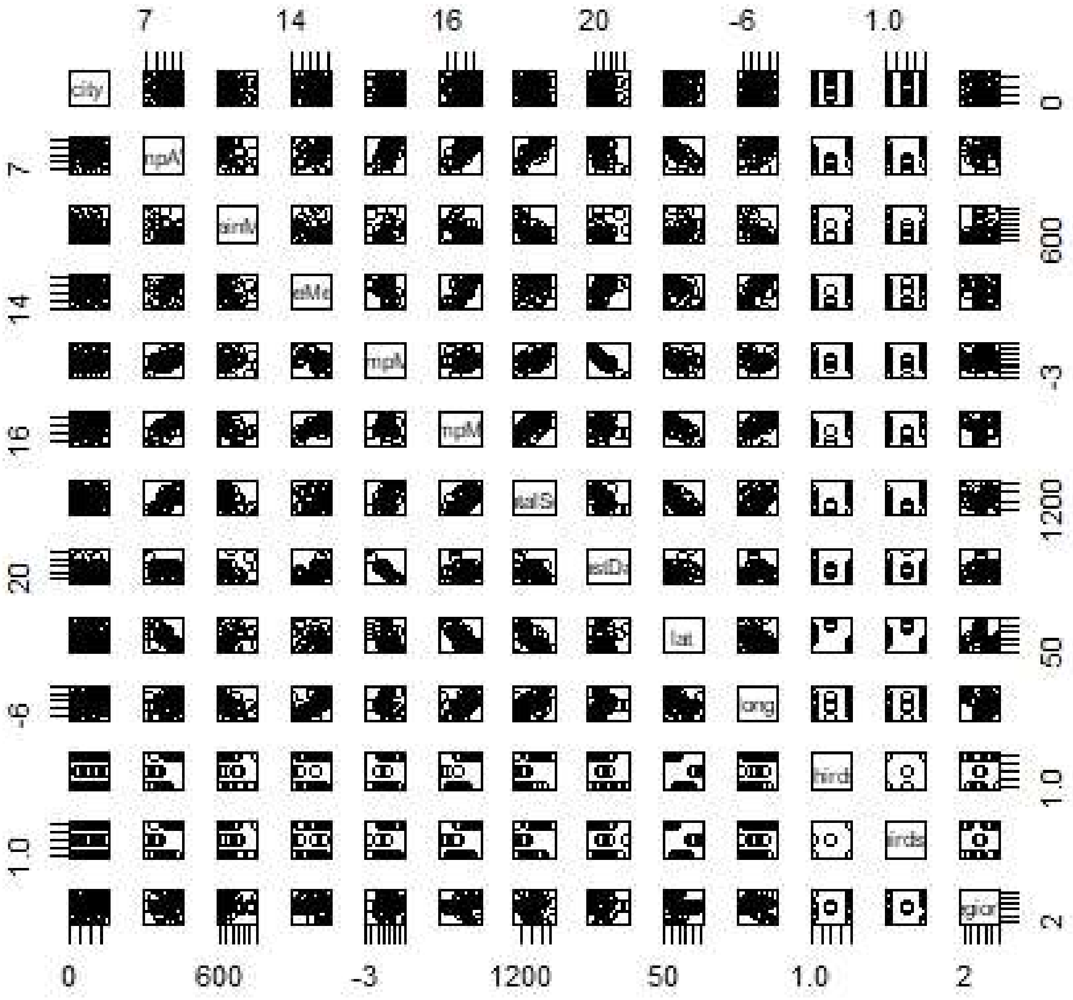
Picture 2

The decision is made to use Thirds_4 variable, because the distribution is more even and better represents geographical distribution of the stations.

- To assign region to each station Kmeans clustering is attempted. Region centers are set as initial centroids for the dataset and iteration is set to 1. However, centroids are moved even after one iteration, so results are not reliable and are dismissed. Code:"MAIN/Script/stations_to_regions_version1.R"
- As an alternative, simple Euclidian distance algorithm is applied: “MAIN/Script/get_distindex_function.R”. The initial results revealed that there was a mistake in the coordinated data for Northern Ireland. The longitude was changed to -5.9 instead of 5.9. However, is still hard to distinguish Northern Ireland from Scotland and Wales using Euclidian distance formula. The list is hand edited using Wikipedia information. 6 stations were reassigned to correct regions:
  ```
  df$regions[df$city == "dunstaffnage"]<-"SCOTLAND"
  df$regions[df$city == "eastbourne"]<-"SOUTH EAST"
  df$regions[df$city == "heathrow"]<-"LONDON"
  df$regions[df$city == "stornoway"]<-"SCOTLAND"
  df$regions[df$city == "tiree"]<-"SCOTLAND"
  df$regions[df$city == "valley"]<-"WALES"
  ```
- File “MAIN/Data/Weather_data.csv” is created.
- Metadata file “MAIN/Data/Meta_Data_Variables.xls” is created.

Exploratory Data Analysis 1

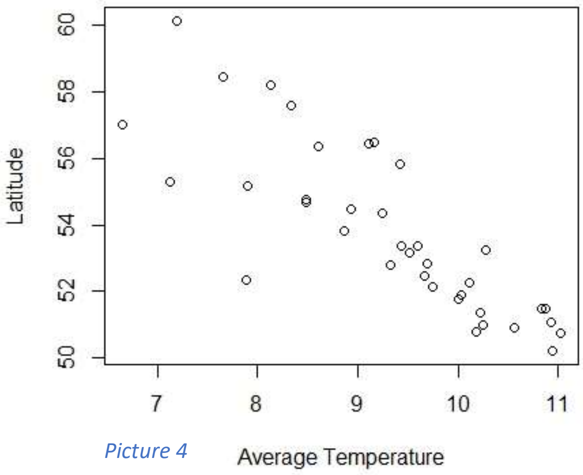
- Plotting entire weather data frame to visually investigate relationship among variables (picture 3):



Picture 3

Following observations are made:

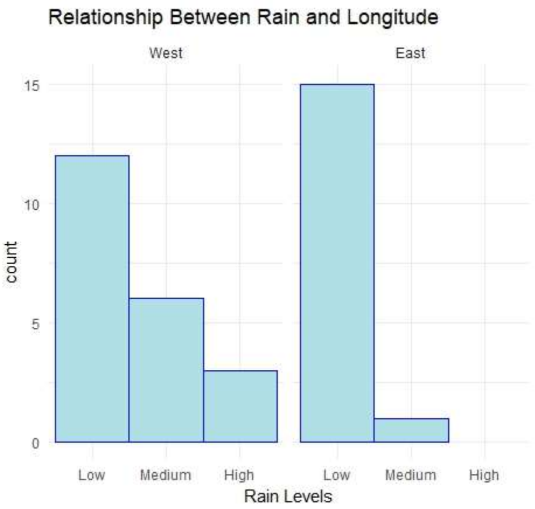
- Some linear relationship between latitudes and TempAVG (picture 4) and TempMax suggests that North – South division can be characterized well by those variables. Average temperature is particularly higher in the South, while North and Middle is more mixed up (picture 8). Frost days are much more scattered in relationship to latitude, probably because of influence of mountain climate and proximity to the coast.
- Plot of longitude and TempMin and Frostdays suggests that temperature gets lowest in the middle and is milder by the coast.
- It is rainier in the Western longitudes. It is also a little bit rainier in the Northern latitudes (pictures 5, 6).
- Well defined clusters are not evident. Some clusters arguably can be identified in graphs between temperature derived variables and TotalRain;
- It is sunnier in the Southern latitudes (picture 7).



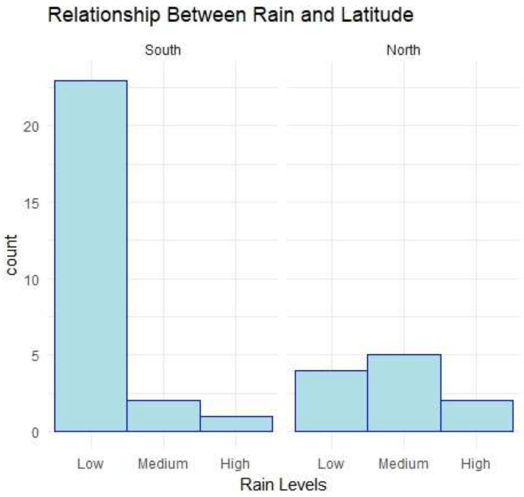
Picture 4

Exploratory data analysis is automated with script:

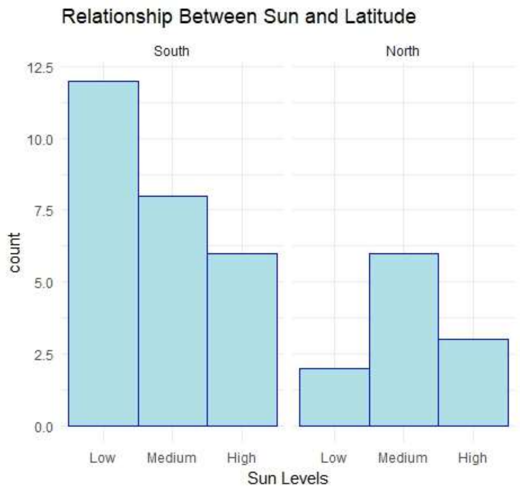
“./MAIN/Scripts/EDA_1.R” .



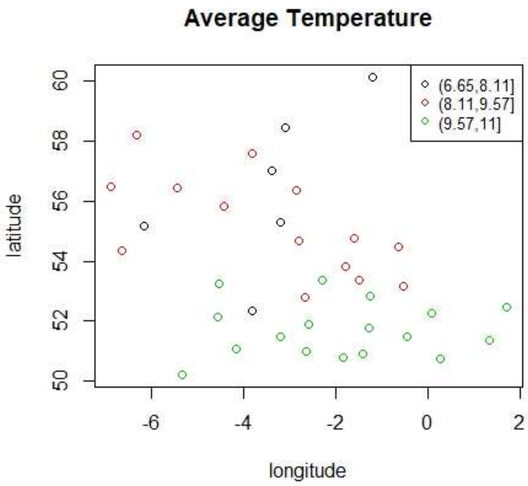
Picture 5



Picture 6

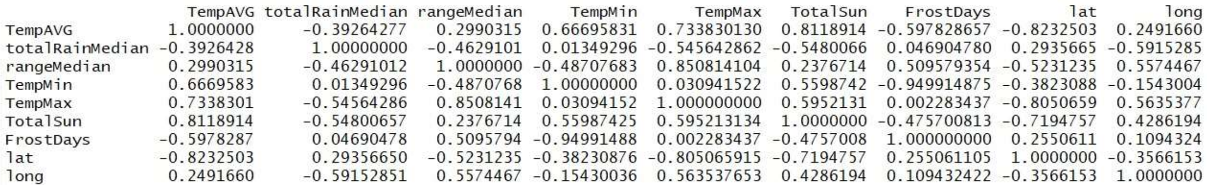


Picture 7



Picture 8

Correlation Matrix



```
                 TempAVG totalRainMedian rangeMedian     TempMin      TempMax   TotalSun    FrostDays        lat       long
TempAVG        1.0000000     -0.39264277   0.2990315  0.66695831  0.733830130  0.8118914 -0.597828657 -0.8232503  0.2491660
totalRainMedian -0.3926428     1.00000000  -0.4629101  0.01349296 -0.545642862 -0.5480066  0.046904780  0.2935665 -0.5915285
rangeMedian    0.2990315     -0.46291012   1.0000000 -0.48707683  0.850814104  0.2376714  0.509579354 -0.5231235  0.5574467
TempMin        0.6669583      0.01349296  -0.4870768  1.00000000  0.030941522  0.5598742 -0.949914875 -0.3823088 -0.1543004
TempMax        0.7338301     -0.54564286   0.8508141  0.03094152  1.000000000  0.5952131  0.002283437 -0.8050659  0.5635377
TotalSun       0.8118914     -0.54800657   0.2376714  0.55987425  0.595213134  1.0000000 -0.475700813 -0.7194757  0.4286194
FrostDays     -0.5978287      0.04690478   0.5095794 -0.94991488  0.002283437 -0.4757008  1.000000000  0.2550611  0.1094324
lat           -0.8232503      0.29356650  -0.5231235 -0.38230876 -0.805065915 -0.7194757  0.255061105  1.0000000 -0.3566153
long           0.2491660     -0.59152851   0.5574467 -0.15430036  0.563537653  0.4286194  0.109432422 -0.3566153  1.0000000
```

Picture 9

- Correlation matrix also show strong relationship between average temperature, maximum temperature, total sun and latitude. It shows some relationship between longitude and rain, range and minimum temperature. It also shows, that some of the variables are significantly correlate, for example FrostDays and TempMin and can be redundant (picture 9).

CLUSTERING: KMEANS

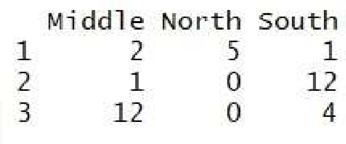
KMeans algorithm is chosen as simple and most widely used clustering algorithm. Euclidian distance metric is used, so data is scaled to avoid distortions due to large numbers. Data is normalized by subtracting mean and dividing by standard deviation.

#1 North, Middle, South clustering

Checking if clusters based on data corresponding to North, Middle, South division by latitude can be created.

- Variables most correlated with latitude are used: TempAVG, TempMax and TotalSun. Connection between TempMax and latitude is also confirmed by the fact that temperature differences are more dominated by latitude in summer (The British Climate). Insolation is also dependent on latitude and Northern regions receive less sunny hours per year.
- K = 3
- The result is 3 clusters of sizes 7, 16, 14. Between_SS / total_SS = 72.5 %
- Clustering using only TempAVG and TempMax variables produces 3 clusters of sizes 7, 18, 12 with between_SS / total_SS = 75.8 %.
- Clustering with only one TempAVG variable produces 3 clusters of sizes 8, 13, 16 with between_SS / total_SS = 85.8 % .

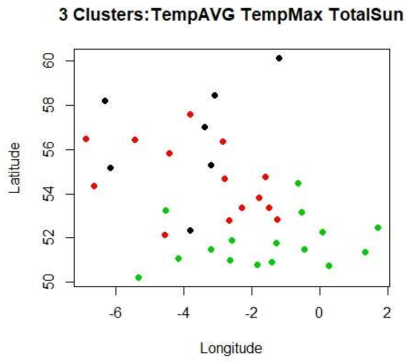
  Confusion matrix shows that clusters produced with TempAVG variable correspond quite well with the North, South, Middle division by latitude (picture 10):



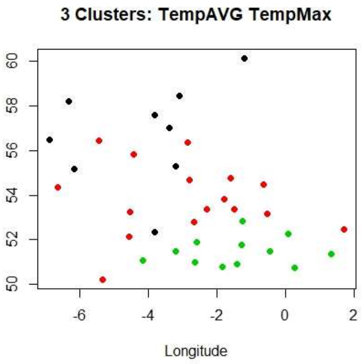
```
  Middle North South
1      2     5     1
2      1     0    12
3     12     0     4
```

Picture 10

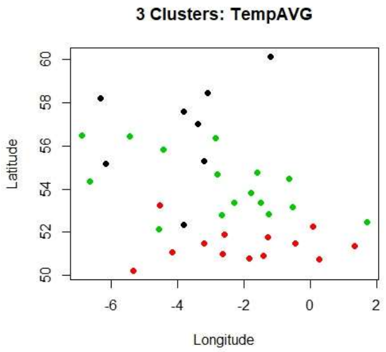
The plots suggest that TempAVG is the best variable to separate stations by latitude (pictures 11, 12, 13). In TempAVG plot South and North are well defined with some more confusion in the Middle (picture 13).



Picture 11



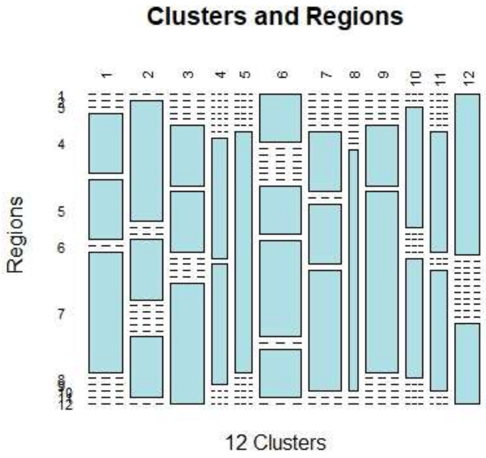
Picture 12



Picture 13

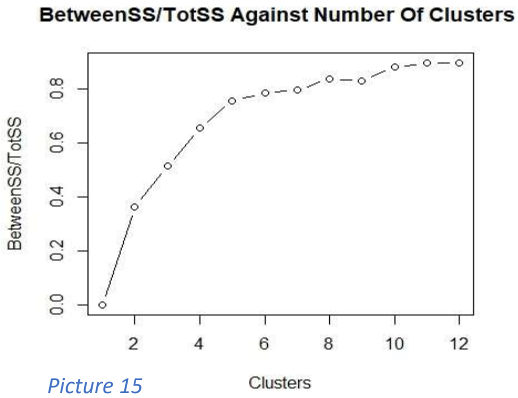
#2 Clustering with all variables

- Looking for natural clusters using all variables.
- KMeans is run algorithm iteratively with number k from 1 to 12. Maximum number 12 is chosen to check if clusters correspond to the regions. Function "MAIN/Script/k_iteration_function.R" that takes dataset ans maximum number of clusters as an argument is used.
- K=12 produces between_SS / total_SS = 89.5 %). Comparing clusters with regions show, that clusters in general do not correspond to regions (picture 14):



Picture 14

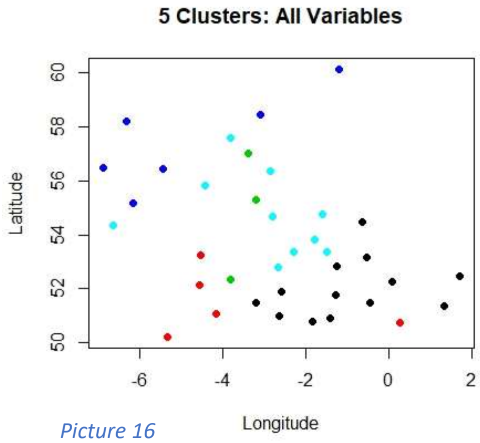
- Elbow test to determine best performing K is implemented using function "MAIN/Script/elbow_test_function.R":



Picture 15

Elbow test suggests that roughly 5 clusters could be the optimal number (picture 15).

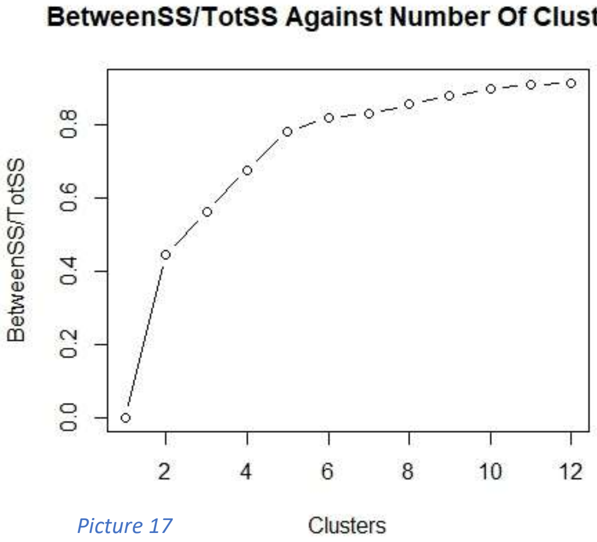
- Geographical representation of 5 clusters with all variables (picture 16):



Picture 16

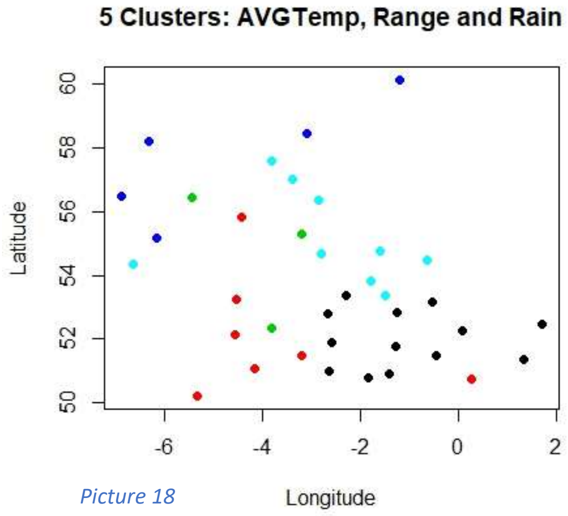
#3 Clustering using temperature and reciprocity variables

- It is possible to come up with different clusters depending on what basis the clustering is done. For weather regionalization temperature and precipitation metrics are most widely used and yield the best results (Yurdanul Unal et al, 2003; Netzel Pawel et al, 2016).
- Some of the temperature measures are highly correlated, so only TempAVG, TempMax and rangeMedian are used.
- Iterative approach is taken for choosing number K. KMEANS is run with K 1:12 and the performance statistics are compared. Functions "MAIN/Script/k_iteration_function.R" and "MAIN/Script/elbow_test_function.R" are used. Elbow graph for temperature and reciprocity variables also suggests 5 to be the optimal number K (picture 17).



Picture 17

- Geographical representation of clusters shows some correspondence with North, South, East, West and Middle division (picture 18):



Picture 18

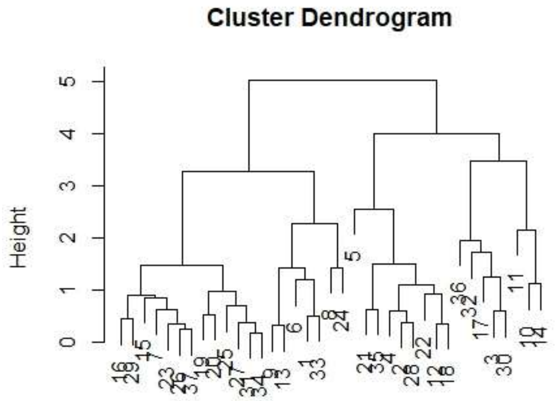
- With AVGTemp, totalRainMedian and range Median variables between_SS / total_SS = 78.1 % as compared to 75.6 % with all variables.

All the steps can be reproduced using script "/MAIN/Script/CLUSTERING.R".

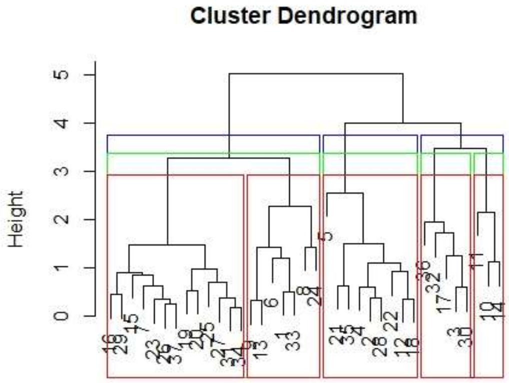
CLUSTERING: Hierarchical clustering

Hierarchical clustering is applied to investigate relationship between the clusters and look what results can be obtained without predefined number of clusters. Method: default, complete linkage.

- "TempAVG", "totalRainMedian" and “rangeMedian" variables are selected, because there were no significant improvement, when using all variables with KMeans aglgorithm. Moreover, resulting clusters seem to be better defined with only these three variables.
- Distance matrix is created.
- Hierarchical model (picture 19):

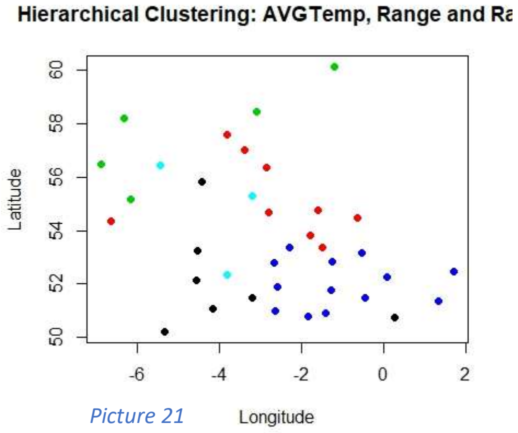


Picture 19



Picture 20

- Cutting tree to obtain 3, 4 or 5 clusters. Number 5 seems to produce best defined clusters (picture 20).
- Geographical representation of the 5 clusters (picture 21):



Picture 21

- Summary statistics for the clusters show (statistics can be found in “MAIN/Graphs/CLUSTERING/summary_statistics_clusters.csv”.) following characteristics for each cluster:
  1 – high average temperature (mean 10.46), medium amount of rain (mean 978.4), sunny (mean 1570), annual temperature fluctuation is small (mean range 17.88).

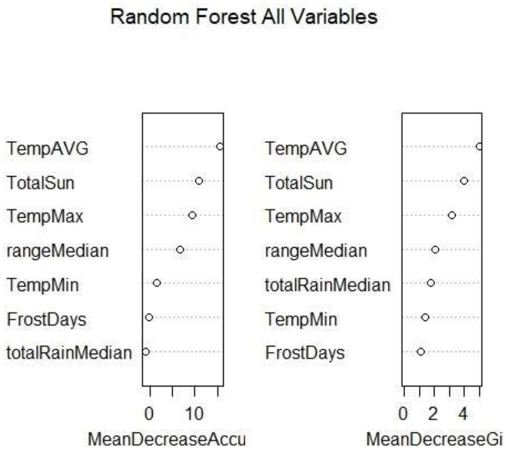
4 – average temperature is also high (mean 10.00), dry (rain mean 674.6), but yearly range is higher (mean 21.2), colder winters and hotter summers.
2 – average temperature is medium (mean 8.56), medium amount of rain (mean 736.6), range is quite high (mean 19.74) due to cold winters (frost days mean : 53.56). Medium number of sunny hours (mean 1342).
3 – cold average temperature (mean 8.0), rainy (mean: 1130.6), annual range is small (mean 15.6) cool summers and mild winters.
5 – cold average temperature (mean 8.0), heavy rain (mean 1676), little sun (mean 1171), annual range is higher then in cluster number 3 (mean 18.77) resulting in warmer summers and colder winters.

- All tasks are automated with script "MAIN/CLUSTERING.R".

CLASSIFICATION: Random Forest

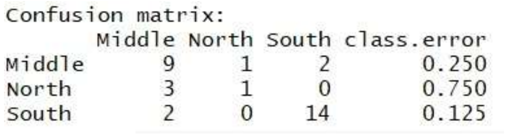
Random Forest algorithm is selected because it is a general-purpose algorithm, suitable for diverse types, it can be used both with numeric and categorical variables or with correlated variables. In addition, Random Forest has in-built feature selection to determine variables with best predictive power and accuracy testing in training data set.

- Divide data set into training and testing (last five stations).
- Set seed for reproducibility and comparability.
- Set importance = true to get insights about predictive power of individual features.
- First Random Forest with all weather features (not including longitude, latitude, and regions).
- Out Of Bag estimate of error rate is 28.12%. This can be viewed as 71.88% percent accuracy.
- Variable importance plot identifies TempAVG, TotalSun and TempMax as three most predictive variables. It confirms correlations between those variables and latitude (picture 22).



Picture 22

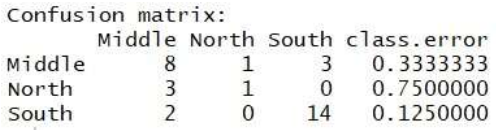
- However, running RandomForest with only TempAVG and TempMax produces better results, than using all three: 75% accuracy with TempAVG and TempMax as compared to 68.75 % accuracy with TempAVG, TotalSun and TempMax.
- Confusion matrix with TempAVG and TempMax variables (picture 23):



```
Confusion matrix:
       Middle North South class.error
Middle      9     1     2       0.250
North       3     1     0       0.750
South       2     0    14       0.125
```

Picture 23

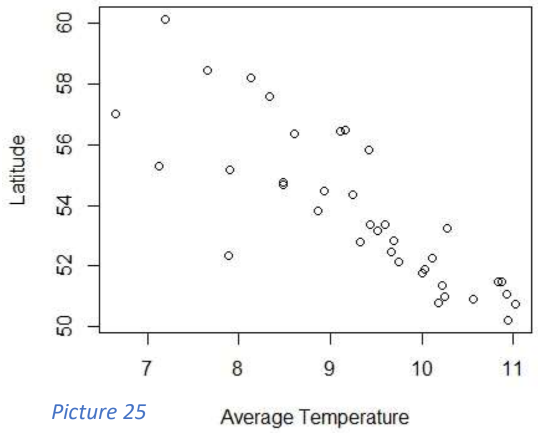
- Confusion matrix all variables (picture 24):



```
Confusion matrix:
       Middle North South class.error
Middle      8     1     3   0.3333333
North       3     1     0   0.7500000
South       2     0    14   0.1250000
```

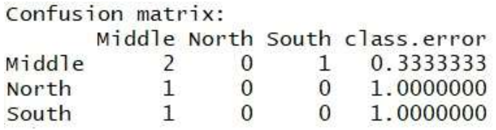
Picture 24

- Comparing confusion matrices for all variables and 2 variables shows, that using only TempAVG and TempMax variables allow to decrease error rate for classifying Middle region. North region is still classified poorly, probably because data is skewed. Moreover, plot shows that relationship between latitude and temperature is more spread at higher latitudes (picture 25):



Picture 25

- Applying Random Forest with best performing TempAVG and TempMax variables to the testing data set (last five stations alphabetically) produces only 40% accuracy (picture 26).

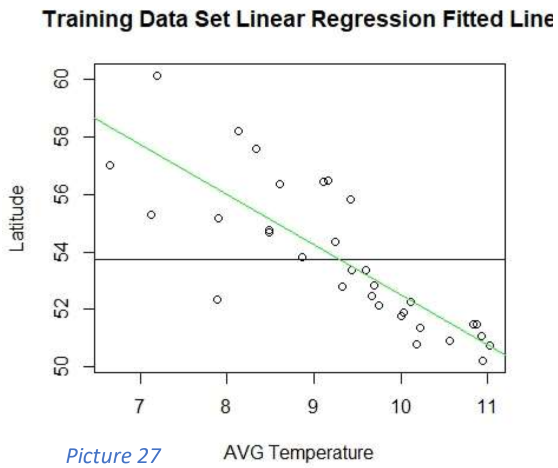


```
Confusion matrix:
       Middle North South class.error
Middle      2     0     1   0.3333333
North       1     0     0   1.0000000
South       1     0     0   1.0000000
```

Picture 26

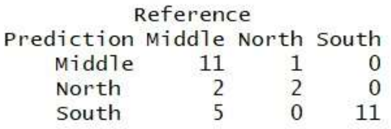
CLASSIFICATION: Linear Regression

- The results with Random Forest are not satisfying thus the alternative algorithm is tested. Exploratory data analysis has shown linear relationship between latitude and TempAVG variables. In addition, clustering with KMeans algorithm revealed best defined clusters with TempAVG variables. Based on the above, Linear Regression algorithm is tested.
- The data scaled because rain and sunny hours measurements result in much higher numbers than temperature variables.
- Linear Regression algorithm is applied with TempAVG variable to training data set (picture 27):



Picture 27

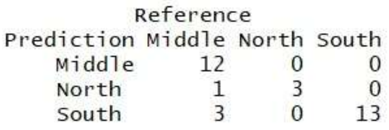
- Labels North, South or Middle are obtained for each predicted latitude using function "Script/get_predicted_labels_function.R".
- Labels are predicted with 75% accuracy (picture 28).



```
          Reference
Prediction Middle North South
    Middle     11     1     0
    North       2     2     0
    South       5     0    11
```

Picture 28

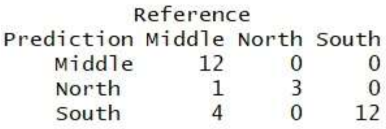
- Linear Regression model with three variables, identified as most predictive by Random Forest algorithm (TempAVG, TempMax and TotalSun) is fitted to training data set.
- The accuracy has increased to 87 % (picture 29):



```
          Reference
Prediction Middle North South
    Middle     12     0     0
    North       1     3     0
    South       3     0    13
```

Picture 29

- Model with totalRainMedian variable is applied to the training dataset. The accuracy has decreased to 84% with rain variable included (picture 30).



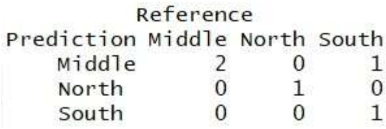
```
            Reference
Prediction Middle North South
    Middle     12     0     0
    North       1     3     0
    South       4     0    12
```

Picture 30

The confusion matrix shows that the algorithm performed worse at classifying South. The rain variable has stronger relationship with the longitude, therefore can interfere with predicting the latitudes.

- Linear model to the test data set is applied with TempAVG, TempMax and TotalSun variables.

- Labels North, South or Middle are obtained for each predicted latitude using function "Script/get_predicted_labels_function.R".
- Confusion matrix comparing actual and predicted labels for test data set (picture 31):



```
            Reference
Prediction Middle North South
    Middle      2     0     1
    North       0     1     0
    South       0     0     1
```

Picture 31

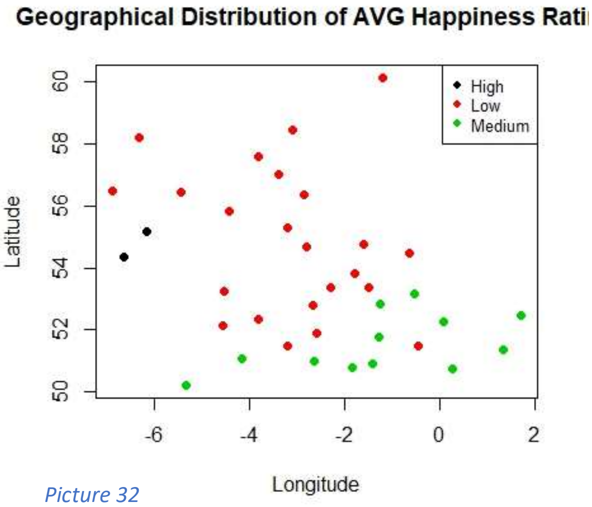
- The predictions with Linear Regression model with one TempAVG variable are made with 80 % accuracy.
- All tasks are automated with script "MAIN/Script/CLASSIFICATION.R".

Happiness Data Preparation and Building Happiness-Weather Data Set

- Happiness data consists of only 4 files, so it is downloaded manually. Sheets containing other than happiness data are removed and files are saved as .csv for importing into R.
- Data sets are imported to R. First five lines with text are removed. Following variables with following names are preserved:
  "AreaCode" – area code;
  "regions" – area name;
  "Low" – percentage of records corresponding to low happiness levels per area;
  "Medium" - percentage of records corresponding to medium happiness levels per area;
  "High" - percentage of records corresponding to high happiness levels per area;
  "VeryHigh" - percentage of records corresponding to very high happiness levels per area;
  "AVGRating" – average rating of happiness per area.
- Data on regional level is used. Above mentioned variables, corresponding to regions in "/MAIN/Utilities/regions.txt" file are selected.
- Function "/Script/Clean_Hap_function.R" is used for above mentioned tasks.
- CSV files "/Data/Happ/hap_1415.csv", "/Data/Happ/hap_1314.csv", "/Data/Happ/hap_1213.csv" and "/Data/Happ/hap_1112.csv" are created.
- 4 data sets for merged into one file "MAIN/Data/Happ/hap_by_region_allyears.csv"
- Comparing happiness AVGRating between years shows slight differences. Therefore, average for 4 years is calculated.
- Labels for AVGRating "High", "Medium" and "Low" are added using function "MAIN/Script/get_labels.R". Northern Ireland region has much higher average rating than all the rest of the UK, so effectively ratings were divided into "High" for Northern Ireland and "Medium" or "Low" for all the rest regions.
- File with averages of all years is created: "MAIN/Data/Happ/hap_means_allyears.csv".
- Weather and Happiness datasets are merged on "regions" basis and file "MAIN/Data/Happ/weather_happiness.csv" is created.
- All the tasks are automated with script "MAIN/Script/build_happiness_dataframe.R".

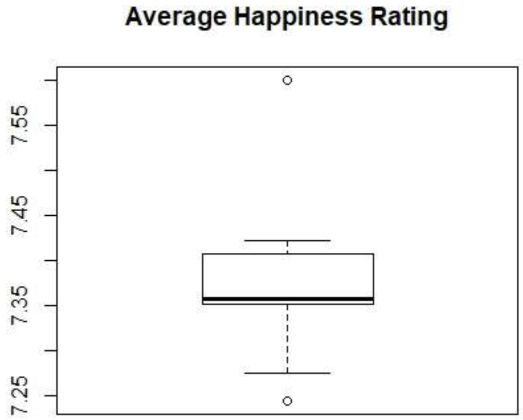
Exploratory Data Analysis 2

Plotting AVGRating geographically reveals that apart from Northern Ireland and London, datapoints corresponding to “Medium” tend to cluster in the Southern part of the UK (picture 32).



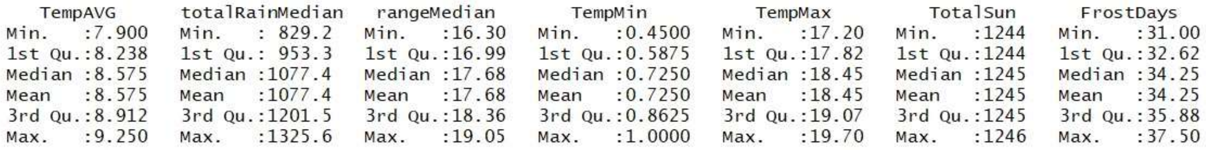
Picture 32

- Northern Ireland average happiness rating is much higher than the rest of the UK, so it is checked with the box plot (picture 33).



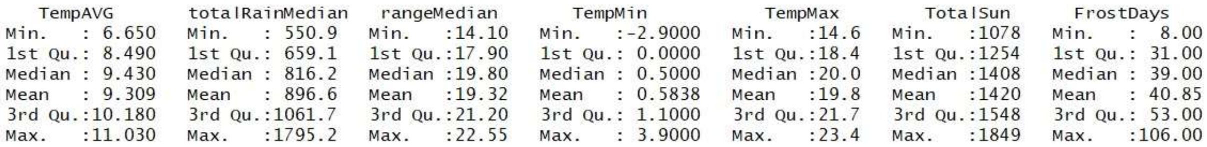
Picture 33

- The result suggests that Northern Ireland average happiness rating can be considered an outlier.
- Comparing summary statistics for Northern Ireland (picture 34) to all regions statistics does not show any obvious deviations from the all regions averages (picture 35).



```
    TempAVG         totalRainMedian    rangeMedian        TempMin          TempMax           TotalSun        FrostDays    
 Min.   :7.900   Min.   : 829.2   Min.   :16.30   Min.   :0.4500   Min.   :17.20   Min.   :1244   Min.   :31.00  
 1st Qu.:8.238   1st Qu.: 953.3   1st Qu.:16.99   1st Qu.:0.5875   1st Qu.:17.82   1st Qu.:1244   1st Qu.:32.62  
 Median :8.575   Median :1077.4   Median :17.68   Median :0.7250   Median :18.45   Median :1245   Median :34.25  
 Mean   :8.575   Mean   :1077.4   Mean   :17.68   Mean   :0.7250   Mean   :18.45   Mean   :1245   Mean   :34.25  
 3rd Qu.:8.912   3rd Qu.:1201.5   3rd Qu.:18.36   3rd Qu.:0.8625   3rd Qu.:19.07   3rd Qu.:1245   3rd Qu.:35.88  
 Max.   :9.250   Max.   :1325.6   Max.   :19.05   Max.   :1.0000   Max.   :19.70   Max.   :1246   Max.   :37.50  
```

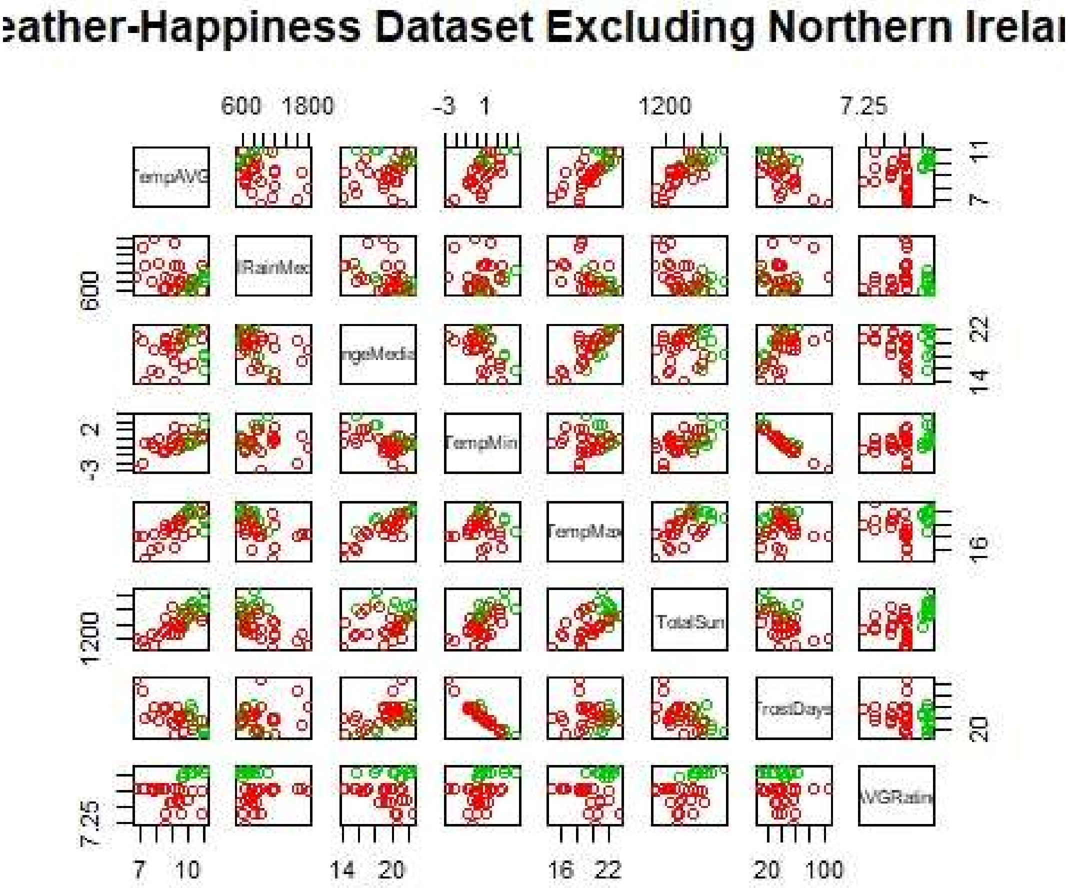
Picture 34



```
    TempAVG         totalRainMedian    rangeMedian        TempMin            TempMax        TotalSun         FrostDays
 Min.   : 6.650   Min.   : 550.9   Min.   :14.10   Min.   :-2.9000   Min.   :14.6   Min.   :1078   Min.   :  8.00
 1st Qu.: 8.490   1st Qu.: 659.1   1st Qu.:17.90   1st Qu.: 0.0000   1st Qu.:18.4   1st Qu.:1254   1st Qu.: 31.00
 Median : 9.430   Median : 816.2   Median :19.80   Median : 0.5000   Median :20.0   Median :1408   Median : 39.00
 Mean   : 9.309   Mean   : 896.6   Mean   :19.32   Mean   : 0.5838   Mean   :19.8   Mean   :1420   Mean   : 40.85
 3rd Qu.:10.180   3rd Qu.:1061.7   3rd Qu.:21.20   3rd Qu.: 1.1000   3rd Qu.:21.7   3rd Qu.:1548   3rd Qu.: 53.00
 Max.   :11.030   Max.   :1795.2   Max.   :22.55   Max.   : 3.9000   Max.   :23.4   Max.   :1849   Max.   :106.00
```

Picture 35

- Without having obvious climatic differences that could be connected to much higher happiness rating, assumption is made, that happiness in Northern Ireland depends on other than weather factors.
- Checking relationship between weather and happiness levels for the rest of the UK, Northern Ireland excluded (picture 36). The “Medium” happiness levels in the plot are colored green, “Low” – red.

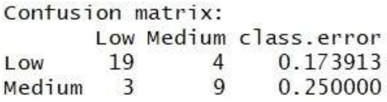


Picture 36

- The plot shows positive relationship between higher levels of average happiness rating with and number of sunny hours per year (TotalSun). There is some relationship with other weather variables, mostly correlated with the latitude, for example maximum temperature (TempMax).
- All tasks are automated with script “MAIN/Script/EDA_2.R”.

Classification Happiness Levels: Random Forest

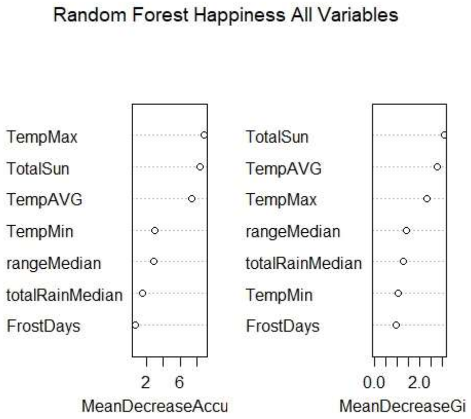
- Random Forest algorithm is applied to check if it is possible to use weather variables to predict “Low” or “Medium” level of average happiness rating per region.
- Random Forest is chosen as a general-purpose algorithm, which does not assume linear relationship and can be applied with correlated variables.
- Random Forest with all weather variables and default parameters is applied. The result is 80% accuracy (picture 37).



```
Confusion matrix:
       Low Medium class.error
Low     19      4    0.173913
Medium   3      9    0.250000
```

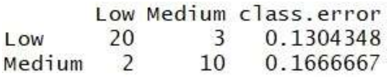
Picture 37

- Variables importance plot for all variables suggests TempMax, TotalSun, and TempAVG to be the most predictive variables (picture 38).



Picture 38

- However, experimentation with different combinations of variables shows that using TempMax, TempMin, TotalSun and Totalrain gives the best accuracy: 85.71%. Error rate for each variable has also decreased (picture 39).



```
       Low Medium class.error
Low     20      3   0.1304348
Medium   2     10   0.1666667
```

Picture 39

- 10-fold cross-validation repeated 10 times is applied, because the data set and the number of features is small.
- The proportion of “Low” and “Medium” average rating levels in the dataset is uneven: Low:23; Medium:12. Thus it is important to preserve the same proportion then sampling for cross-validation.
- Parameter tuning is set to be adjusted automatically with 3 tries.
- The best performing result is 81.5 % accuracy, Kappa = 0.59, with mtry = 2 .
- There can be observed some relationship between higher happiness levels int the UK and weather factors that correspond to Southern latitudes. It can be used for prediction. However, the highest levels of happiness do not follow the same trend. In general, it is not possible to say if there is a relationship between weather and level of happiness. For example, it is possible that Southern part of UK has higher income rate, that could account for higher happiness levels.

Moreover, some studies suggest that the relationship between feeling of happiness and the weather (Mark Easton, 2012).

- All tasks are automated with script “MAIN/Script/Happiness_Prediction.R”.

References:

1. Pawel Netzel and Tomasz Stepinski (2016) On Using a Clustering Approach for Global Climate Classification. *Journal of Climate. American Meteorological Society.* *https://journals.ametsoc.org/doi/10.1175/JCLI-D-15-0640.1. Accessed 5/30/2018*
2. Yurdanur Unal et al (2003) REDEFINING THE CLIMATE ZONES OF TURKEY USING CLUSTER ANALYSIS. *INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol. 23: 1045 – 1055. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.7795&rep=rep1&type=pdf . Accessed 5/30/2018*
3. The British Climate. *http://www.lordgrey.org.uk/~f014/usefulresources/aric/Resources/Teaching_Packs/Key_Stage_4/Weather_Climate/11.html*
4. Hands-on Tutorials in R with Hefin Rhys. Published on 6 Aug 2017. *https://www.youtube.com/watch?v=PX5nSBGB5Tw&t=1699s . Accessed 6/1/2018*
5. UK Regional Weather Differences*. http://www.foreignstudents.com/guide-to-britain/british-culture/weather/location . Accessed 6/4/2018.*
6. David Langer (2014) Introduction to Data Science with R (2). *https://www.youtube.com/watch?v=u6sahb7Hmog&list=PLTJTBoU5HOCRrTs3cJK-PbHM39cwCU0PF&index=2*
7. David Langer (2014) Introduction to Data Science with R - Cross Validation. https://www.youtube.com/watch?v=84JSk36og34&t=2401s
8. Christoph Scherber (2013) Statistics with R (1) - Linear regression. *https://www.youtube.com/watch?v=Xh6Rex3ARjc*
9. Mark Easton (2012) Does sunshine make us happier? https://www.bbc.com/news/uk-18986041

List of files:

- MAIN/ MAIN.R – primary script for replicating data preparation and analysis
- MAIN/Scrip/ - all script files and functions that are run from the MAIN.R script
- MAIN/.Rhistory
- MAIN/Utilities/ - regions.txt, stations.txt, url_list.txt. Files necessary to run the code.
- MAIN/Data/Happ/ - hap_1112.csv, hap_1213.csv, hap_1314.csv, hap_1415.csv. Files prepared to be loaded in R.
- /Graphs/ – includes .jpeg graphs illustrating the analysis and summary_statistics_clusters.xls file.
- Data/Datasets/ - includes datasets created in the analysis process and metadata file.
- Data/Raw_Weather_files/ - .txt raw weather data files.
- Data/Raw_Happiness_files/ - .xls raw happiness data files
- Data/Edited_Weather_files/ - edited and cleaned weather .csv files
- Data/Edited_Happiness_files – edited .csv happiness files