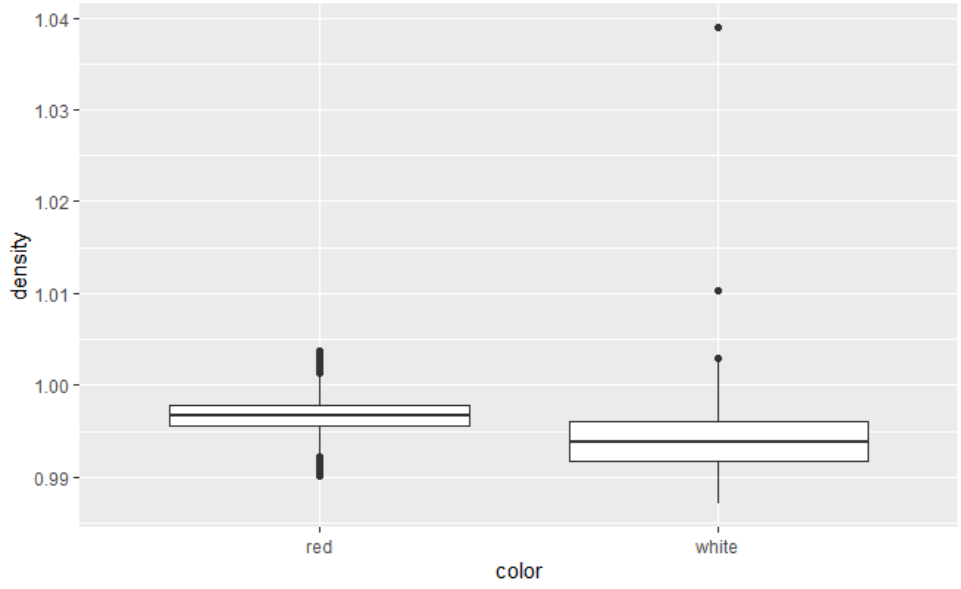
DataAnalytics2021Spring_A7_Diyanko_Bhc

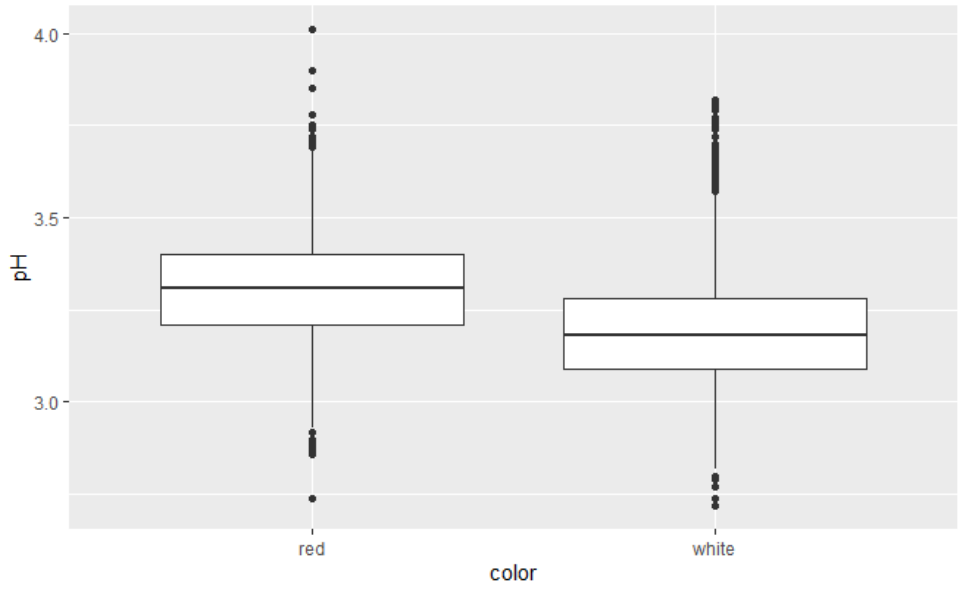
Wine Quality

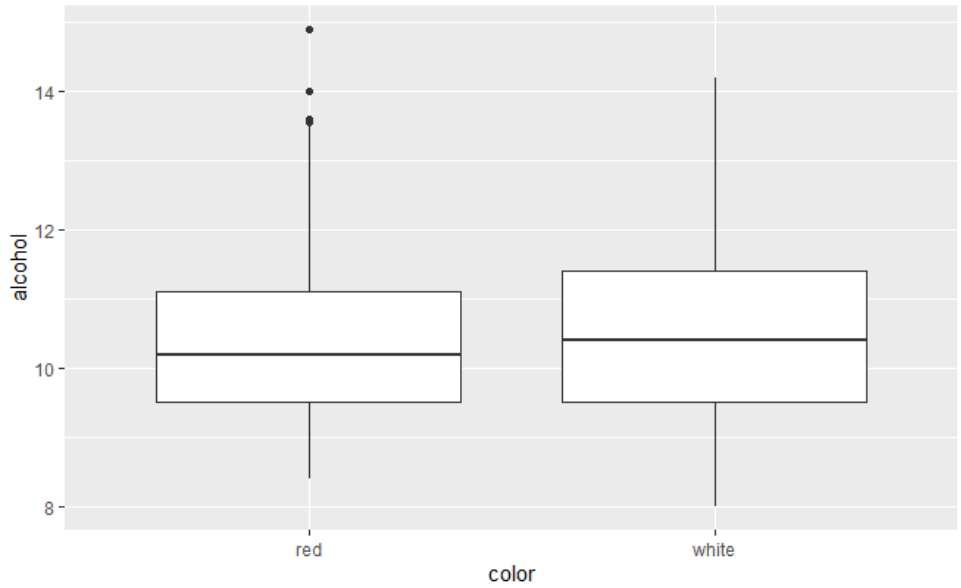
Importing data and EDA

The datasets are loaded separately and are combined wrt color.

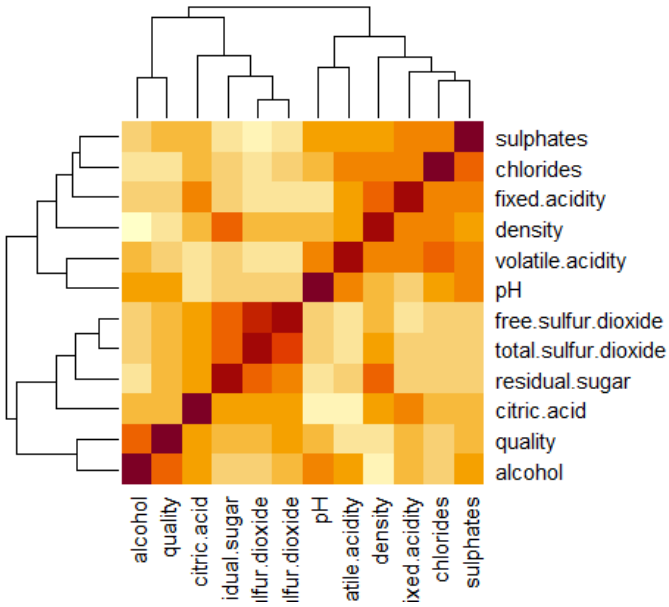
To understand the distribution, boxplots of the features density, pH, and alcohol content are plotted for each of the colors of wine. This shows that there are certainly some differences in pH value and the density.







To understand how each of the factors corelate with all the other factors, we make a heatmap.

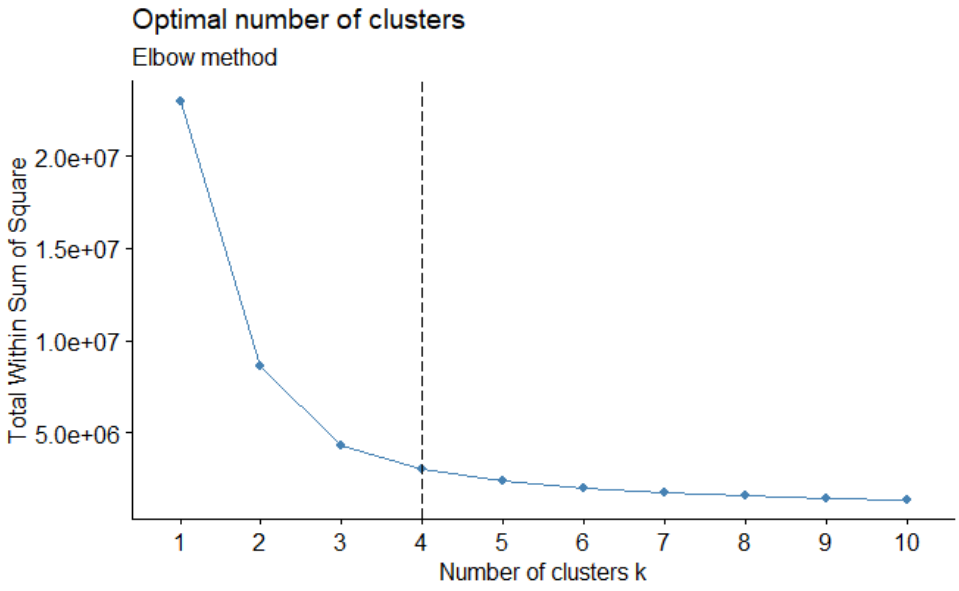


We can see that there seems to be a corellation of the values of free sulphur dioxide and total sulphur dioxide.

Now we start applying 3 models to this dataset.

KMeans

We intend to find the value for k from the elbow plot. The plot is given below.



We calculate the k to be 4. Here is the distribution.

Total Observations in Table: 6497

data1$cluster	data1$quality 3	4	5	6	7	8	9	Row Total
1	3	35	631	854	382	73	3	1981
	4.131	14.460	0.670	0.133	8.539	3.404	1.428	
	0.002	0.018	0.319	0.431	0.193	0.037	0.002	0.305
	0.100	0.162	0.295	0.301	0.354	0.378	0.600	
	0.000	0.005	0.097	0.131	0.059	0.011	0.000	
2	4	77	520	927	431	83	2	2044
	3.133	1.204	34.634	1.355	24.685	8.176	0.116	
	0.002	0.038	0.254	0.454	0.211	0.041	0.001	0.315
	0.133	0.356	0.243	0.327	0.399	0.430	0.400	
	0.001	0.012	0.080	0.143	0.066	0.013	0.000	
3	9	36	496	464	75	20	0	1100
	3.026	0.009	49.618	0.544	63.475	4.918	0.847	
	0.008	0.033	0.451	0.422	0.068	0.018	0.000	0.169
	0.300	0.167	0.232	0.164	0.070	0.104	0.000	
	0.001	0.006	0.076	0.071	0.012	0.003	0.000	
4	14	68	491	591	191	17	0	1372
	9.273	10.987	3.457	0.104	5.962	13.848	1.056	
	0.010	0.050	0.358	0.431	0.139	0.012	0.000	0.211
	0.467	0.315	0.230	0.208	0.177	0.088	0.000	
	0.002	0.010	0.076	0.091	0.029	0.003	0.000	
Column Total	30	216	2138	2836	1079	193	5	6497
	0.005	0.033	0.329	0.437	0.166	0.030	0.001	

We can conclude the following from the above

- Group 1 - most are in the lower to middle category of quality.
- Group 2 - most are in the high category of quality.
- Group 3 - most are in the middle category of quality.
- Group 4 - most are in the middle-high category of quality.

Random Forest

We get an error rate of 29.21% in this model.

```
        OOB estimate of  error rate: 29.21%
Confusion matrix:
  3  4    5    6   7  8 9 class.error
3 0  1   15   10   0  0 0   1.0000000
4 1 25   99   50   2  0 0   0.8587571
5 0  3 1294  410  11  0 0   0.2467986
6 0  4  313 1610  96  1 0   0.2045455
7 0  0   21  284 466  8 0   0.4017972
8 0  0    1   41  47 54 0   0.6223776
9 0  0    0    1   4  0 0   1.0000000
```

Lower error rate is present when the quality is not in the extreme values.

kNN

This was not a good model. The accuracy was only 41%.

Total Observations in Table: 1625

	prediction2					
test_wine_target	4	5	6	7	8	Row Total
3	0	2	2	0	0	4
	0.005	0.581	0.000	0.766	0.030	
	0.000	0.500	0.500	0.000	0.000	0.002
	0.000	0.004	0.002	0.000	0.000	
	0.000	0.001	0.001	0.000	0.000	
4	1	23	15	0	0	39
	18.881	11.657	1.143	7.464	0.288	
	0.026	0.590	0.385	0.000	0.000	0.024
	0.500	0.048	0.018	0.000	0.000	
	0.001	0.014	0.009	0.000	0.000	
5	0	188	203	26	3	420
	0.517	33.969	0.444	36.791	0.003	
	0.000	0.448	0.483	0.062	0.007	0.258
	0.000	0.394	0.247	0.084	0.250	
	0.000	0.116	0.125	0.016	0.002	
6	1	201	443	160	7	812
	0.000	5.854	2.452	0.136	0.168	
	0.001	0.248	0.546	0.197	0.009	0.500
	0.500	0.421	0.538	0.514	0.583	
	0.001	0.124	0.273	0.098	0.004	
7	0	50	143	105	2	300
	0.369	16.451	0.526	39.437	0.021	
	0.000	0.167	0.477	0.350	0.007	0.185
	0.000	0.105	0.174	0.338	0.167	
	0.000	0.031	0.088	0.065	0.001	
8	0	13	17	20	0	50
	0.062	0.192	2.736	11.370	0.369	
	0.000	0.260	0.340	0.400	0.000	0.031
	0.000	0.027	0.021	0.064	0.000	
	0.000	0.008	0.010	0.012	0.000	
Column Total	2	477	823	311	12	1625
	0.001	0.294	0.506	0.191	0.007	

Decisions

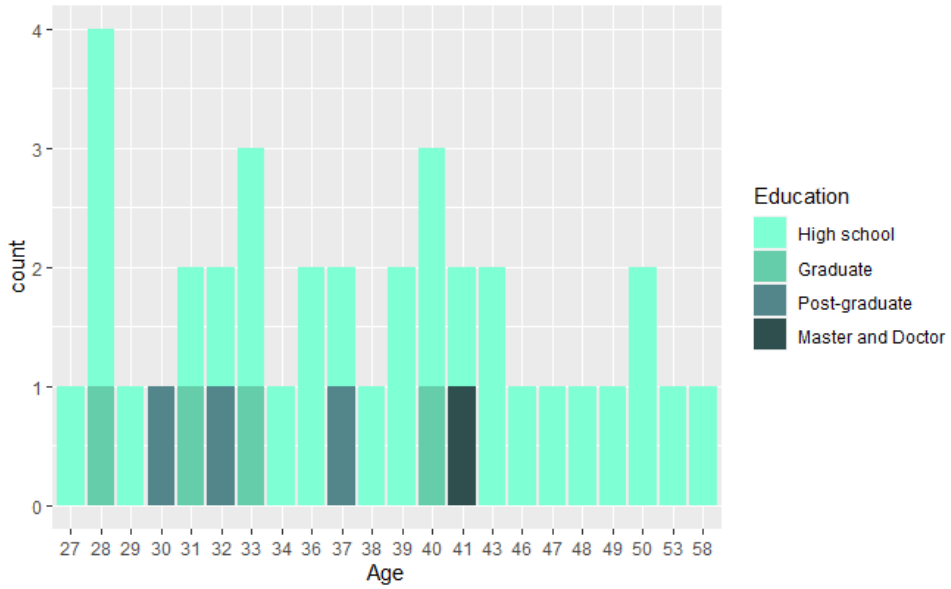
We see that Random Forest is a better model for this data and we can group the wine records on the basis of quality (except extreme values) using KMeans.

The models were not particularly very accurate but gave us an understanding of the data. Maybe the reason for the bad performance of kNN was that the prediction of the wine quality is not directly dependent upon the factors under consideration.

Absenteeism at work

Importing data and EDA

The data has been imported from the UCI database and the delimiter has been set to ';'. A plot showing the education level with the record of absent hours has been plotted to find out if there is any sort of corelation.



This seems interesting that high-school graduates, specially at a young age have a higher rate than any other category but it may be misleading because of the less amount of data.

Now we start applying 3 models to this dataset.

Random Forest

The dataset has been divided into a 70-30 split. and random forest analysis has been carried out taking into account *Reason.for.absence, Month.of.absence, Day.of.the.week, Distance.from.Residence.to.Work, Age, Disciplinary.failure, Education, Son, Social.drinker, Social.smoker, Pet, Absenteeism.time.in.hours*.

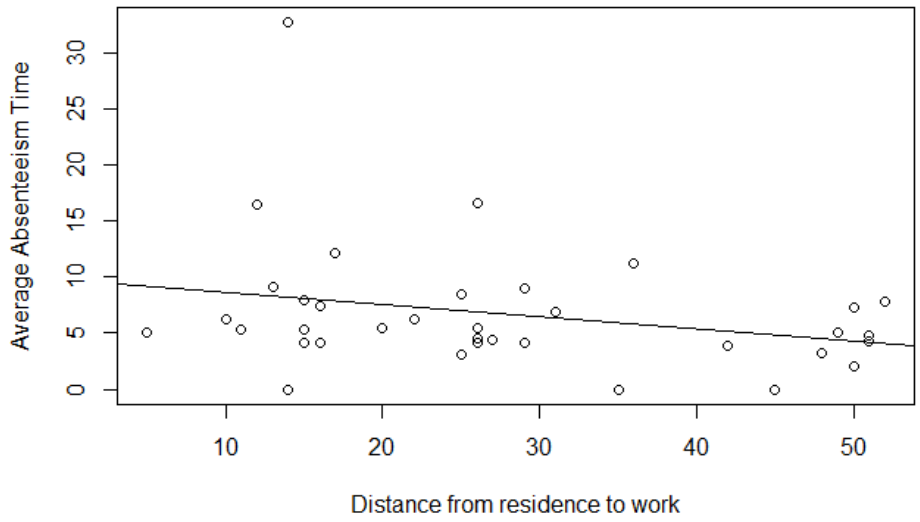
The confusion matrix is as follows. We try to get the test set values on the basis of the actual reason for absence.

	rf.predict															
	1	2	7	8	11	12	13	14	15	19	20	22	23	25	26	27	28
1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	1	0	0	0	2	0	0	0	2	3	0	1	0	1
7	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
8	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1
11	0	0	1	0	0	0	0	2	0	0	0	1	0	0	0	0	1
12	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	1
14	0	0	0	1	2	0	0	2	0	0	1	1	2	0	2	0	0
15	0	0	0	1	1	0	0	0	1	0	1	2	2	0	0	0	0
19	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	1
20	0	0	0	0	1	1	0	5	0	0	1	0	0	0	0	1	1
22	0	0	0	0	2	0	0	1	0	0	1	5	0	0	0	0	0
23	0	0	0	0	1	0	0	2	2	0	0	0	23	1	0	0	8
25	0	0	0	0	0	1	0	1	0	0	0	1	2	0	0	0	0
26	0	0	0	0	1	0	0	1	0	0	2	0	1	1	3	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	7	0
28	0	0	0	0	0	0	0	0	0	0	2	0	11	1	0	1	17

This model has an accuracy of 42.6%.

Regression

This is to find out if the factor, Distance.from.Residence.to.Work was related to the absenteeism time of the employee.



We notice no significant relation between these variables.

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     9.66363    2.00147   4.828 2.69e-05 ***
Distance.from.Residence.to.Work -0.10750    0.06538  -1.644    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.617 on 35 degrees of freedom
Multiple R-squared:  0.07172,   Adjusted R-squared:  0.04519
F-statistic: 2.704 on 1 and 35 DF,  p-value: 0.1091
```

The R-squared value is 0.4519 and the p-value is 0.1091.

SVM

Now we try to use the SVM model to understand if the absence is due to any of the factors of *ID,Reason.for.absence, Month.of.absence,Day.of.the.week,Age,Education,Social.drinker.*

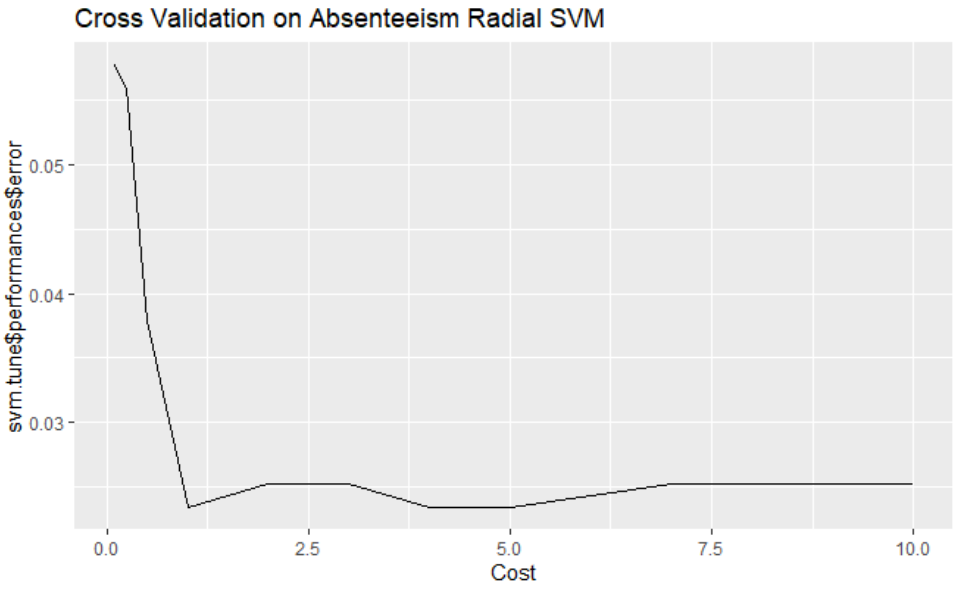
We make a 75-25 split of the data and perform SVM with a radial kernel.

The 10-fold cross-validation has been carried out and the tune function was used to find the best cost value. We find it to be 4. Then, we try to predict the test set-values.

Description: df[,3] [10 x 3]

cost <dbl>	error <dbl>	dispersion <dbl>
0.10	0.057759...	0.031651...
0.25	0.055941...	0.033486...
0.50	0.037889...	0.021647...
1.00	0.023441...	0.017014...
2.00	0.025227...	0.017298...
3.00	0.025227...	0.017298...
4.00	0.023409...	0.016949...
5.00	0.023409...	0.016949...
7.00	0.025259...	0.019359...
10...	0.025259...	0.019359...

1-10 of 10 rows




```
   predicted_svm
      0   1
  0 172   5
  1   2   6
[1] 0.9621622
```

We find the accuracy to be 96.21%. We can also see the predicted value up above.

Decisions

We can say that the SVM performed very well with a high accuracy rate. The number of items with a disciplinary failure was very less and that could have been a factor.

The regression model showed us that there was effectively no correlation with the distance from work with the absence data. It was not very successful.

The random forest worked quite well in this scenario. The dataset is not large enough to make good predictions on the reason for absence. The model did not perform well.