Introduction

The goal of this report is to perform an analysis on a dataset containing information about the weather of a European country. We are interested in alerting officials about the possibility of a drought. We are also interested in whether there is a difference between the mean maximum temperature and the mean minimum temperature. By doing statistical analysis using SAS software, we investigate the answer to these questions. A further of evaporation analysis based on wind gust direction is done. Additionally, we build a model to help with the prediction of rain for that European country.

Exploratory Data Analysis

For this project we will work with 4657 independent records of 24 weather-related variables are contained in the dataset. 16 out of the 24 variables are continuous and the rest are ordinal or categorical. **Table 1** below shows descriptive statistics (minimum, maximum, mean and standard deviation) for the continuous variables of our dataset as well as the number of missing observations for each variable.

Descriptive Statistics for Continuous Variables

Variable	N Miss	N	Minimum	Maximum	Mean	Std Dev
MinTemp	25	4632	-6.80	29.70	12.08	6.42
MaxTemp	11	4646	-2.20	44.10	23.12	7.11
Rainfall	54	4603	0.00	225.00	2.31	8.15
Evaporation	2041	2616	0.00	42.40	5.37	3.90
Sunshine	2275	2382	0.00	14.00	7.59	3.80
WindGustSpeed	333	4324	9.00	106.00	39.68	13.55
WindSpeed9am	44	4613	0.00	65.00	13.87	8.85
WindSpeed3pm	80	4577	0.00	56.00	18.27	8.64
Humidity9am	60	4597	3.00	100.00	69.21	18.98
Humidity3pm	117	4540	3.00	100.00	51.75	20.85
Pressure9am	475	4182	988.50	1039.50	1017.81	6.99
Pressure3pm	481	4176	987.70	1036.90	1015.36	6.90
Cloud9am	1779	2878	0.00	8.00	4.50	2.87
Cloud3pm	1873	2784	0.00	8.00	4.54	2.73
Temp9am	27	4630	-3.40	37.50	16.86	6.52
Temp3pm	84	4573	-4.00	43.50	21.59	6.94

Table 1. Descriptive Statistics for Continuous Variables

A temperature for this country of 25°C or above can alert officials to the possibility of a drought. We are interested on whether we need to alert the officials about this possibility. We look at the statistics of the maximum daily temperature. **Table 1** shows that the mean of

the maximum daily temperature is 23.1°C, the max is 44.1°C and the standard deviation is 7.11. Based on this, we can expect that a mean of 23°C is different than 25°C.

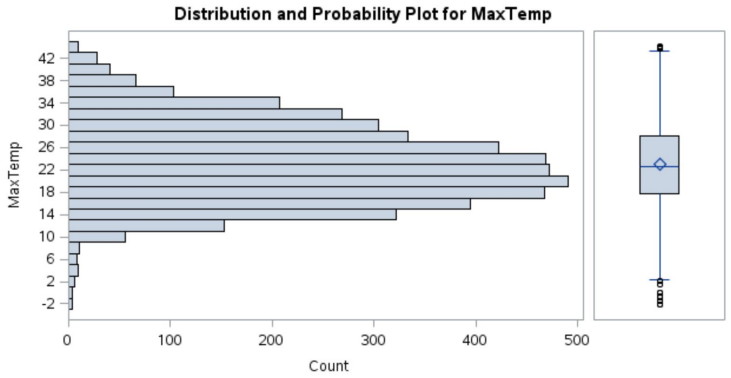

Figure 1: Histogram of maximum daily temperature variable.

Figure 1 below shows the distribution of maximum temperature variable. The boxplot shows that there are multiple outliers towards the low temperatures and few towards the high temperatures.

We are also interested in the minimum temperature and whether there is a difference in the mean between this and the mean of the maximum daily temperature. From **Table 1** we can see that the minimum daily temperature has a mean of 12°C and a standard deviation of 6.42. By having a mean maximum temperature of 23°C and a mean minimum temperature of 12°C we can expect that there is a significant difference between them.

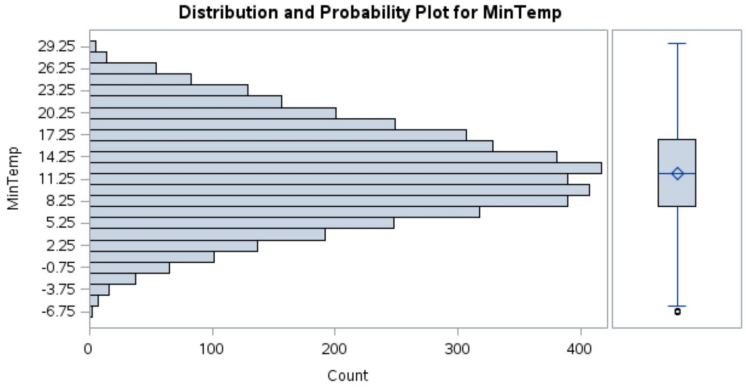

Figure 2: Histogram and boxplot of minimum daily temperature variable.

Figure 2 below shows the distribution of maximum temperature variable. The boxplot shows that there are a few outliers towards the low temperatures.

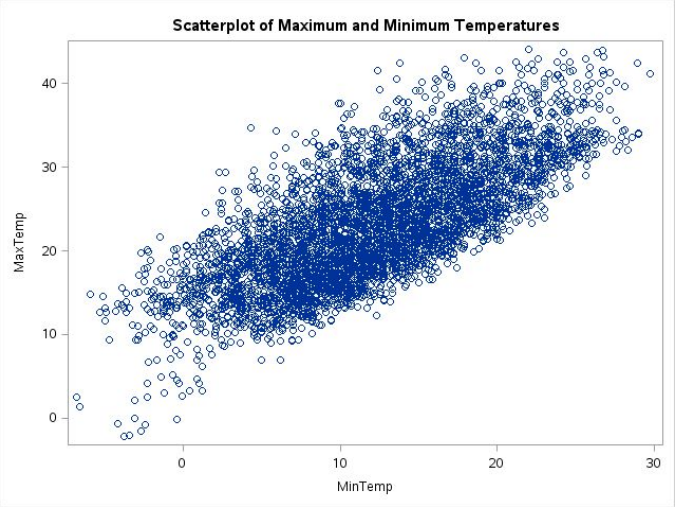

Figure 3: Scatter Plot of Maximum and Minimum Temperature Variables.

In **Figure 3** we can observe that MaxTemp and MinTemp variables seem linearly correlated but we can see visually that there seems to be a difference between them of approximately 10°C.

We are also interested to see if we can predict evaporation depending on the direction of the wind. We could predict if the wind carries humidity based on its direction. It is of interest to see the summary statistics of our wind gust direction variable, which can be observed in **Table 2** below.

Summary Statistics for WindGustDir variable

Number of Variable Levels			
Variable	Levels	Missing Levels	Nonmissing Levels
WindGustDir	17	1	16

WindGustDir	Frequency	Percent	Cumulative Frequency	Cumulative Percent
E	272	6.29	272	6.29
ENE	265	6.13	537	12.42
ESE	260	6.01	797	18.43
N	314	7.26	1111	25.69
NE	227	5.25	1338	30.94
NNE	187	4.32	1525	35.27
NNW	205	4.74	1730	40.01
NW	263	6.08	1993	46.09
S	279	6.45	2272	52.54
SE	310	7.17	2582	59.71
SSE	316	7.31	2898	67.02
SSW	269	6.22	3167	73.24
SW	294	6.80	3461	80.04
W	287	6.64	3748	86.68
WNW	300	6.94	4048	93.62
WSW	276	6.38	4324	100.00
Frequency Missing = 333				

Table 2: Summary statistics for WindGustDir variable.

We will investigate possible correlation between our continuous variables, which could lead to multicollinearity. Getting inspiration from the internet[1], we use SAS to create the following correlation matrix for our continuous variables.

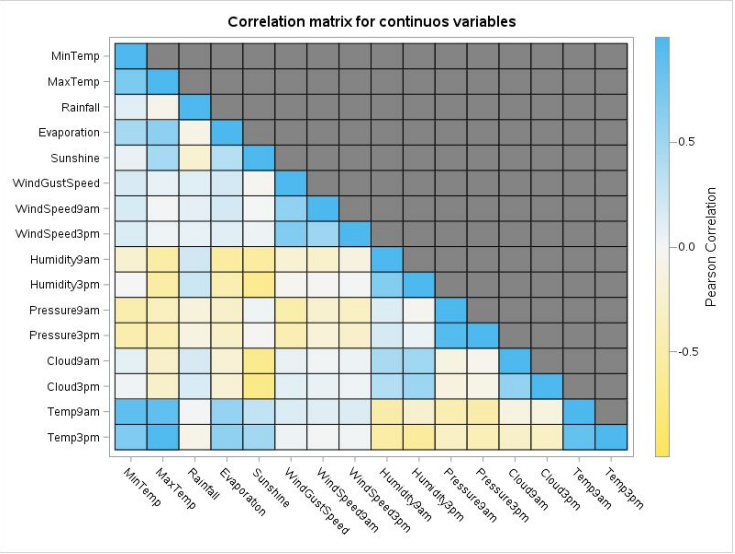

Figure 4: Correlation matrix for continuous variables of our dataset.

We find a high correlation between some of the continuous variables. They are plotted again in **Figure 5** below. We need to keep this in mind in further analysis.

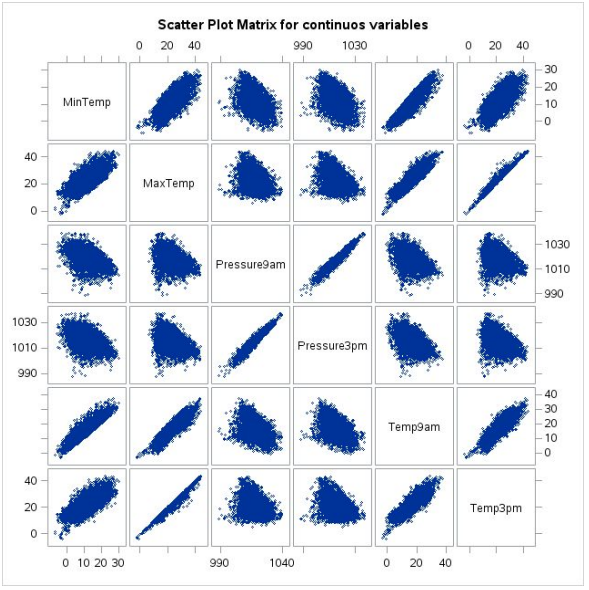

Figure 5: Scatter Plot Matrix for continuous variables.

[1]How to build a correlations matrix heat map with SAS
https://blogs.sas.com/content/sasdummy/2013/06/12/correlations-matrix-heatmap-with-sas/

Formal Analysis

Questions of interest

Assessment of a drought

We are interested in the possibility of a drought in this European country. A temperature for this country of 25°C or above can alert officials to the possibility of a drought. We conduct a one sample t-test to investigate whether the mean maximum temperature is significantly different from 25°C. We obtain a mean of 23.1°C for this variable and a p-value of <0.0001. At a level of significance of 0.05, we reject the null hypothesis and conclude that the mean maximum temperature for this country is different than 25°C.

N	Mean	Std Dev	Std Err	Minimum	Maximum
4646	23.1228	7.1131	0.1044	-2.2000	44.1000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
23.1228	22.9182	23.3274	7.1131	6.9714	7.2608

Table 3. Summary Statistics

DF	t Value	Pr > |t|
4645	-17.99	<.0001

Table 4. P-Value for One Sample T-Test

Table 3 and **Table 4** above show the summary statistics as well as the p-value obtained in the one-sample t-test.

Assumptions

Assumptions for a one-sample test include independent observations and normality in our distribution. Observations are independent as given. However, normality assumption seems dubious by looking at the distribution in **Figure 6** and the Q-Q Plot in **Figure 7** below.

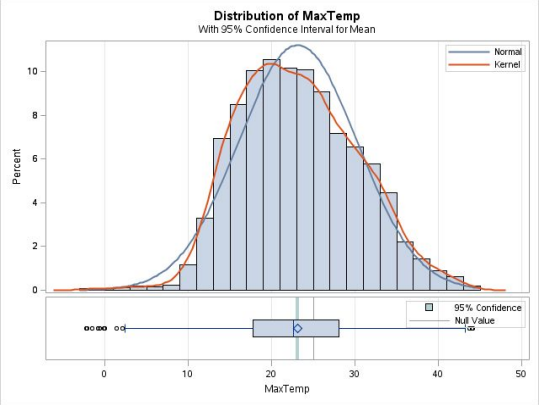

Figure 6. Distribution of MaxTemp

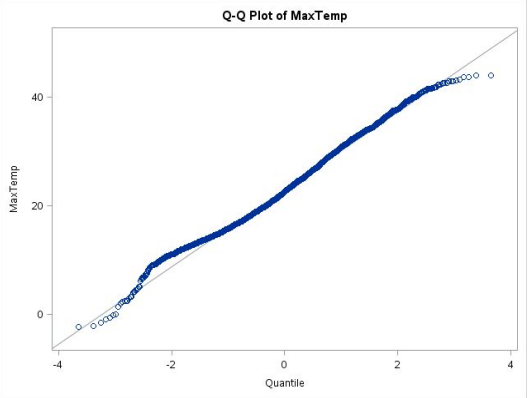

Figure 7. Q-Q Plot of Max Temp

Figure 8 and **Figure 9** below show that the value of 25°C lies outside the confidence interval for the mean of maximum temperature.

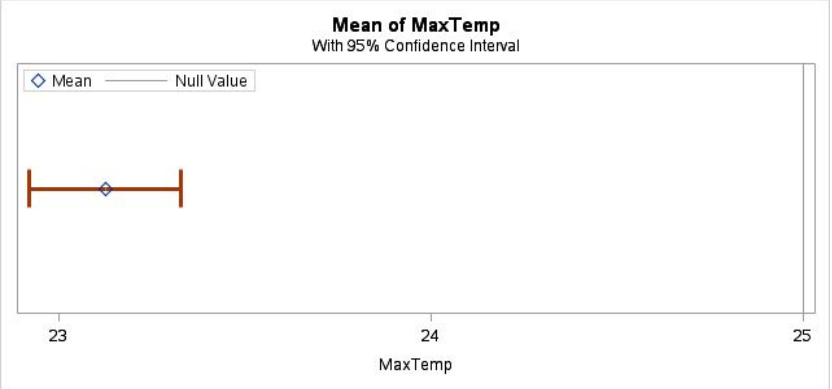

Figure 8. Confidence Interval for the mean of MaxTemp including the null hypothesis

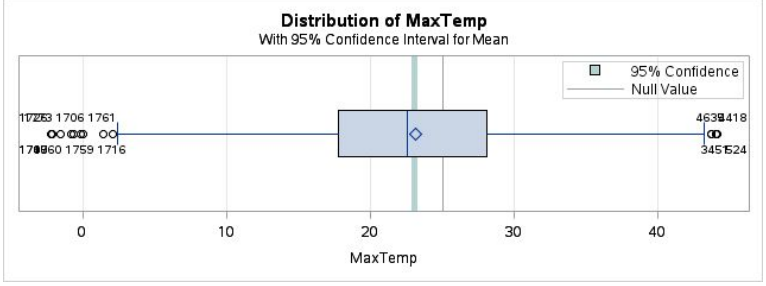

Figure 9. Boxplot of the distribution of MaxTemp

Analysis of the difference between mean maximum and mean minimum temperature

To test whether there's a difference between the mean maximum temperature and the mean minimum temperature we run a paired t-test. The variables MaxTemp and MinTemp are the paired variables with a sample size of 4624. With a p-value of <0.0001, at a level of significance of 0.05, we reject the null hypothesis and conclude that the mean of the maximum and the mean of the minimum daily temperatures are significantly different. The confidence intervals shown in **Table 5** that it is highly likely that difference in means lies between 10.9°C and 11.1°C.

N	Mean	Std Dev	Std Err	Minimum	Maximum
4624	11.0464	4.9937	0.0734	0	30.4000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
11.0464	10.9024	11.1903	4.9937	4.8939	5.0976

DF	t Value	Pr > |t|
4623	150.42	<.0001

Table 5. Summary statistic tables of paired t-test.

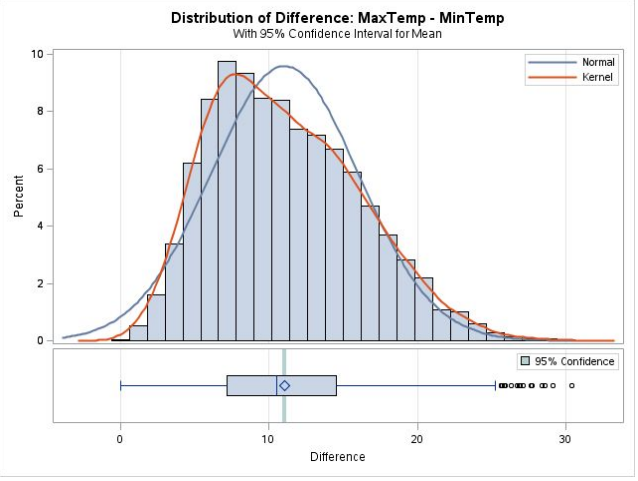

Figure 10. Distribution of difference of MaxTemp and MinTemp.

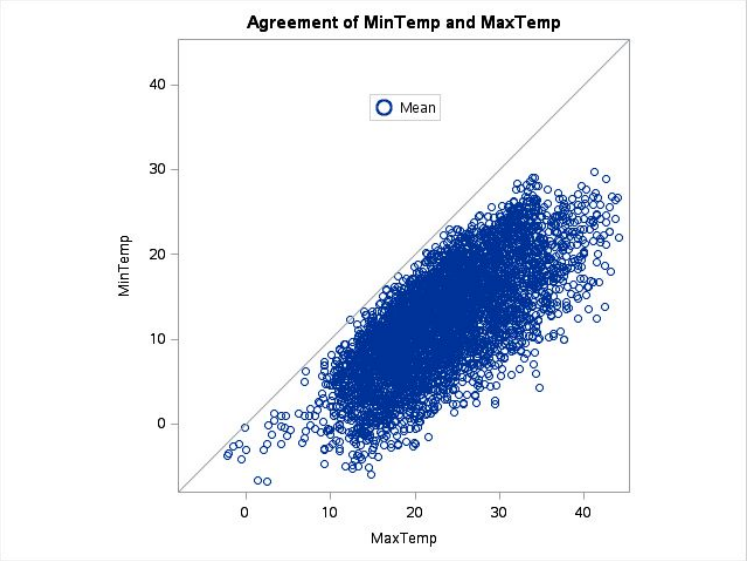

Figure 11. Agreement of MinTemp and MaxTemp.

Figure 11 shows that most of the difference in means are different than zero.

Assumptions

While the assumption of independence observations holds, the paired t–test also assumes that the differences between pairs are normally distributed. This assumption seems dubious.

Assessment of correlation between evaporation with strongest wind gust

It is quite possible that evaporation is associated with the strongest wind gust direction for a day. By using PROC GLM in SAS, we perform a one-way ANOVA test with Evaporation as response and WindGustDir as predictor. Our goals is to check whether there is a difference in the mean evaporation between the different strongest wind gust directions.

After performing one-way ANOVA test, we find the model to be significant by looking at the p-value (<0.001) in **Table 6**. We will now proceed to check the assumptions of normality and equal variances.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	982.08622	65.47241	4.29	<.0001
Error	2424	36956.62375	15.24613		
Corrected Total	2439	37938.70996			

Table 6. ANOVA table of model with Evaporation as response, WindGustDir as predictor

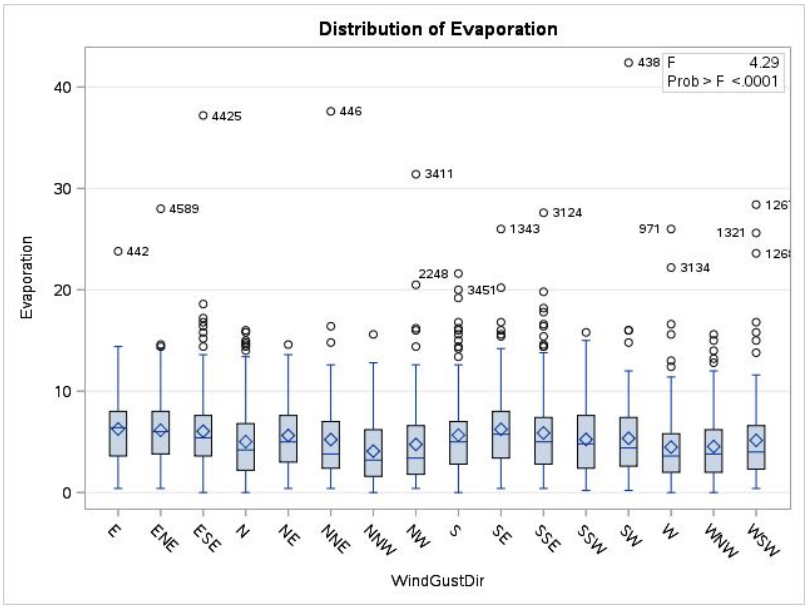

Figure 12. Distribution of Evaporation

We perform the normality test using `PROC UNIVARIATE` to see the scores of Kolmogorov-Smirnov. **Table 7** shows the p-value of 0.01 for Kolmogorov test. We reject the null hypothesis of normality. This means that the assumption of normality does not hold.

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.09384	Pr > D	<0.0100
Cramer-von Mises	W-Sq	8.768635	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	56.58408	Pr > A-Sq	<0.0050

Table 7: Tests for Normality

We perform Levene's test to check the assumption of equal variances using the same `PROC GLM` in `SAS` but setting the option `hovtest=levene`. **Table 8** shows the results for the test. Since p-value 0.82 > 0.05, we do not reject the hypothesis of equal variances and conclude that the assumption holds. Since the assumption holds, we can proceed to do a post hoc test using Tukey's method to see which specific groups differ.

Levene's Test for Homogeneity of Evaporation Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
WindGustDir	15	28347.2	1889.8	0.67	0.8202
Error	2424	6881014	2838.7		

Table 8: Levene's Test for Homogeneity of Evaporation Variance

We now run Tukey's honestly significant difference (HSD) post hoc test using `PROC MIXED` in SAS. **Figure 13** below shows the results of our test. It is quite difficult from this graphic to detect which groups differ. By using `PROC ANOVA` and lines options, we create a Lines Plot that helps to detect visually which groups differ. For example, we can conclude that groups from level *E* and level *NNW* have different means since none of the three bars is the same for those 3 groups.

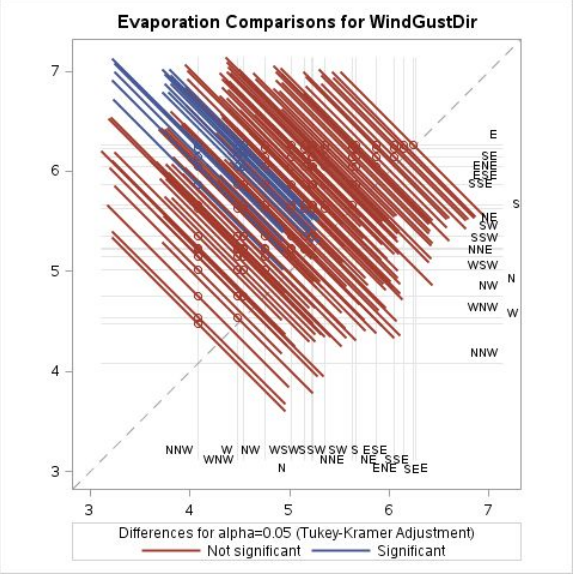

Figure 13. Diffogram of the Evaporation Comparisons for WindGustDir using Tukey Grouping

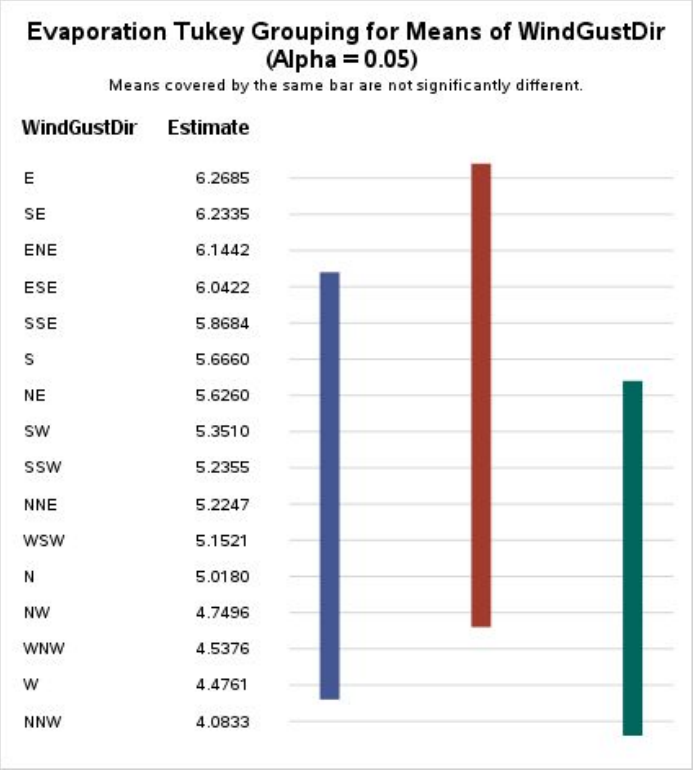

Figure 14. Mean Lines Plot of Evaporation Tukey Grouping

Fitting a Model for Evaporation

Using `PROC GLMSELECT` in SAS with Schwarz Bayesian Information Criterion (SBC) we will proceed to compare models using backward and stepwise directions. We use evaporation as the response variable and all other variables as potential predictors. We will not include the variable RainTomorrow in the modelling process, neither we will consider any interaction terms, transformation or higher order terms.

We get the following models:

Backward Direction Model (10 independent variables): `Intercept MinTemp MaxTemp WindGustSpeed WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm RainToday`

Stepwise Direction Model (6 independent variables): `Intercept MinTemp MaxTemp WindSpeed9am Humidity9am Humidity3pm RainToday`

By using different directions for the same criterion (SBC), we get different final models. The model built using Stepwise Direction has only has 6 variables while the one built using

Backward direction has 10 predictors. Having a model with less variables reduces in complexity. Having that in mind, we prefer the model with less variables (stepwise direction). For making our decision, we also compared the Adjusted R-Squared statistic shown in **Table 9** and **Table 10**, ensuring that the numbers do not differ too much. Model found using stepwise direction explains 53% of the variance while the one using backward direction explains 54%.

Root MSE	2.54126
Dependent Mean	5.43648
R-Square	0.5472
Adj R-Sq	0.5447
AIC	5173.41169
AICC	5173.58619
BIC	3371.59266
C(p)	91.39835
PRESS	11747
SBC	3430.86876
ASE	6.41857

Table 9. Fit Statistics for Backward direction model

Root MSE	2.57948
Dependent Mean	5.43648
R-Square	0.5325
Adj R-Sq	0.5309
AIC	5223.19080
AICC	5223.27115
BIC	3421.23662
C(p)	144.37679
PRESS	12058
SBC	3458.66348
ASE	6.62783

Table 10. Fit Statistics for Stepwise direction model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	13972	1397.20032	216.35	<.0001
Error	1790	11560	6.45802		
Corrected Total	1800	25532			

Table 11. ANOVA for Backward direction model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	13595	2265.85512	340.54	<.0001
Error	1794	11937	6.65369		
Corrected Total	1800	25532			

Table 12. ANOVA for Stepwise direction model

Data Preparation

Categorical Inputs

When doing predictive modeling, problems may arise when dealing with too many levels in categorical predictors. Having predictive variables with too many levels can increase dimension in the input space, leading to potential issues such as overfitting. One solution for dealing with categorical inputs with too many levels is clustering categorical inputs.

Clustering Categorical Input Levels

There are different solutions for combining categorical nominal inputs, one of them is Greenacre's method (1988)[2]. We will use this method in SAS using `PROC CLUSTER` procedure combined with `METHOD=WARD` for variables in our dataset with too many levels. First we find the number of levels of all our categorical variables, which can be observed in **Table 13**.

Categorical Variables

Number of Variable Levels			
Variable	Levels	Missing Levels	Nonmissing Levels
Location	8	0	8
NewEquipment	2	0	2
WindGustDir	17	1	16
WindDir9am	17	1	16
WindDir3pm	17	1	16
Status	2	0	2
RainToday	3	1	2
Cloud3pm	10	1	9
Cloud9am	10	1	9

Table 13: Number of levels for Categorical Variables.

We notice that "WindGustDir", "WindDir9am" and "WindDir3pm" have too many levels. We will proceed to cluster them using Greenacre's method.

We will start analysing WindGustDir variable using RainTomorrow. PROC CLUSTER provides a Cluster History table as well as a dendrogram that helps for interpretation. There are ways to calculate the optimal number of clusters through the chi-square statistic and the associated p-value computation. However, we will select the final number of significant clusters through a "cut-off" of .95 in the R-Square statistic. Using this procedure, we end up with a total of 4 significant clusters for WindGustDir.

Table 14 below shows all the clusters that are combined keeping an R-Square of at least 0.95. The dendrogram in **Figure 15** helps visually to spot which levels need to be combined. After selecting the clusters, we proceed to combine them into a new variable using a data step in SAS with `IF/ELSE` conditionals. **Table 15** shows the final set of clusters for WindGustDir.

[2] Greenacre, M.J. Journal of Classification (1988) 5: 39. https://doi.org/10.1007/BF01901670

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
15	E	NNE	459	0.0000	1.00	
14	N	W	601	0.0000	1.00	
13	ESE	SW	554	0.0002	1.00	
12	CL14	NW	864	0.0004	.999	
11	SE	SSE	626	0.0008	.999	
10	NNW	WNW	505	0.0013	.997	
9	CL13	NE	781	0.0023	.995	
8	S	CL11	905	0.0029	.992	
7	SSW	WSW	545	0.0049	.987	
6	CL15	CL9	1240	0.0070	.980	
5	CL12	CL7	1409	0.0209	.959	
4	CL6	ENE	1505	0.0301	.929	
3	CL5	CL10	1914	0.0510	.878	
2	CL4	CL8	2410	0.0838	.794	
1	CL2	CL3	4324	0.7943	.000	

Table 14: Cluster History Table for *WindGustDir* variable

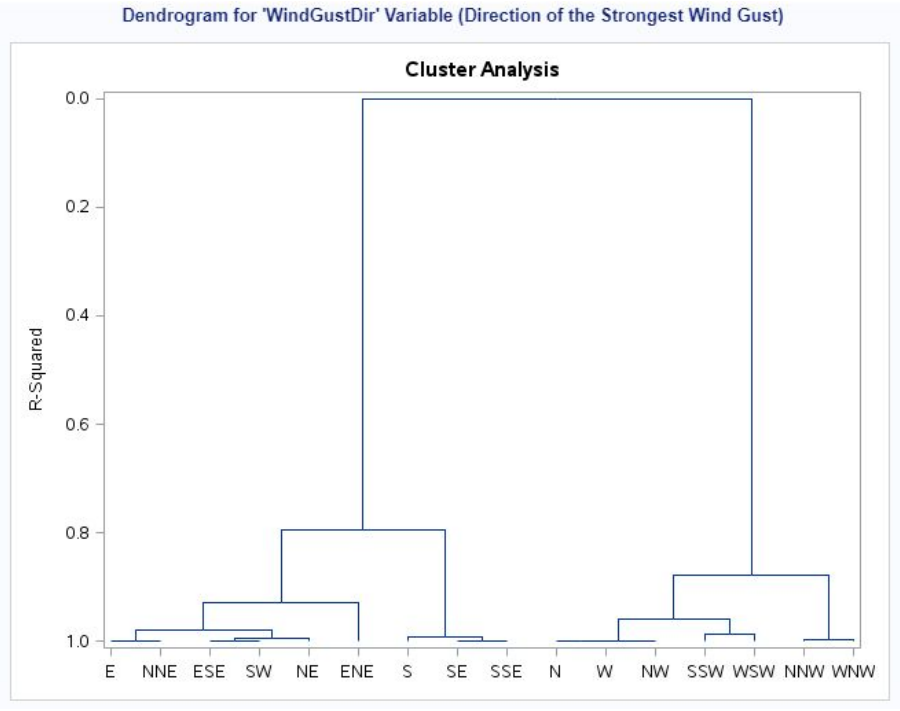

Figure 15: Dendrogram of Clusters versus R-Square Values for *WindGustDir* variable.

WindGustDir Clusters	
Cluster 1:	E NNE ESE SW NE ENE
Cluster 2:	S SE SSE

Cluster 3:	N W NW SSW WSW
Cluster 4	NNW WNW

Table 15: Cluster History Table for *WindGustDir* variable

This procedure is repeated for both WindDir9am and WindDir3m variables (plots not included). The final number of clusters for both variables is 3 and 4 respectively. They are shown in **Table 16** and **Table 17** below.

WindDir9am Clusters	
Cluster 1:	E ENE SE ESE
Cluster 2:	N NNW WNW
Cluster 3:	NE S SSW SSE NNE NW SW WSW W

Table 16: Cluster History Table for *WindDir9am* variable

WindDir3pm Clusters	
Cluster 1:	E ENE
Cluster 2:	NE SE S SSW WSWS
Cluster 3:	ESE NNE SSE SW
Cluster 4	N W WNW NW NNW

Table 17: Cluster History Table for *WindDir3pm* variable

Data Splitting

When predictive models are assessed using the same data that was used to fit the model, better assessment statistics are found, which is called “optimism bias”. In order to prevent optimism bias, assessment needs to be done in a different set of data, which is called “honest assessment”.

We will split our dataset for our predictive model using stratified random sampling. This can be achieved in SAS by using PROC SURVEYSELECT command. Data will be splitted into training data (75% of original dataset), and test data (25% of original dataset) using RainTomorrow as our target variable. After splitting the data, we use PROC FREQ to ensure that we have same proportion of RainTomorrow events in both datasets, which we can

confirm in **Table 18** and **Table 19** below. There is 76.9% of "No" in training and 76.96% in test dataset.

Frequencies in training dataset

RainTomorrow	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	2687	76.90	2687	76.90
Yes	807	23.10	3494	100.00

Figure 16. Frequencies in training dataset.

Frequencies in test dataset

RainTomorrow	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	895	76.96	895	76.96
Yes	268	23.04	1163	100.00

Figure 17. Frequencies in test dataset.

Code:

```
data weather;
       set input.weather;
run;

proc sort data = weather out = sorted;
       by RainTomorrow;
run;

proc surveyselect noprint data = sorted samprate=.75 outall out = sampling;
       strata RainTomorrow;
run;

data training(drop=selected SelectionProb SamplingWeight)
        test(drop=selected SelectionProb SamplingWeight);
       set work.sampling;
       if selected then output training;
       else output test;
run;

*Run proc freq on both datasets to ensure equal proportion of target variable.

*76.90 percentage of NO;
proc freq data = training;
  tables RainTomorrow;
run;

*76.96 percentage of NO;
proc freq data = test;
```

```
  tables RainTomorrow;
run;
```

Variable Screening

Before proceeding to the independent variable selection, we will proceed to verify the linearity assumption in logistic regression. Even though logistic regression does not require a linear relationship between the response variable and predictors, it does assume a linear relationship between the independent variables and their respective log odds[3].

Plotting empirical logit plots is one way for detecting nonlinearity. A different approach is by finding non-monotonic relationships by comparing Hoeffding's D and Spearman's correlation ranks. If a non-monotonic relationship is found, further analysis would need to be performed with empirical logit plots to detect potential nonlinear relationships. By using this analysis method, we will also be able to find weak associations in variables. Variables with weak associations are irrelevant for our predictive model, and they could be excluded in the analysis. **Table 20** shows the method to find either non-monotonic or weak associations in our variables.

Spearman's Rank	**Hoeffding's D Rank**	**Association**
Low	High	Non-monotonic
Low	Low	Weak

Table 20: Associations between Spearman's Rank & Hoeffding's D Rank

We will perform this analysis and try to find variables with Spearman's low rank and detect if they are either non-monotonic or irrelevant. It is worth noting that if the Spearman's rank is high, the association will be monotonic no matter what Hoeffding's D rank is. Monotonic associations are not a problem in logistic regression.

For performing this test in SAS, different procedures are needed, such as `PROC CORR` and `PROC RANK`. After performing the test, we get the following output:

[3] Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011;18:1099–104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Rank of Spearman Correlations and Hoeffding's D Correlations

Obs	Variable	Spearman's rank of variables	Hoeffding's D rank of variables	Spearman Correlation	Spearman p-value	Hoeffding's D Correlation	Hoeffding's D p-value
1	WindSpeed3pm	1	1	0.15902	<.0001	0.00744	<.0001
2	WindSpeed9am	2	2	0.20295	<.0001	0.01155	<.0001
3	WindGustSpeed	3	3	0.22990	<.0001	0.01800	<.0001
4	Pressure9am	4	5	-0.32325	<.0001	0.03803	<.0001
5	Pressure3pm	5	6	-0.35161	<.0001	0.04401	<.0001
6	Rainfall	6	4	-0.35643	<.0001	0.02719	<.0001
7	Humidity3pm	7	7	-0.43737	<.0001	0.05787	<.0001
8	Sunshine	8	8	0.46747	<.0001	0.07533	<.0001
9	MinTemp	9	9	0.54508	<.0001	0.10003	<.0001
10	Humidity9am	10	10	-0.57688	<.0001	0.11428	<.0001
11	Temp9am	11	11	0.64644	<.0001	0.15118	<.0001
12	Temp3pm	12	12	0.65899	<.0001	0.15866	<.0001
13	MaxTemp	13	13	0.67640	<.0001	0.16983	<.0001

Table 21: Rank of Spearman Correlations and Hoeffding's D Correlations

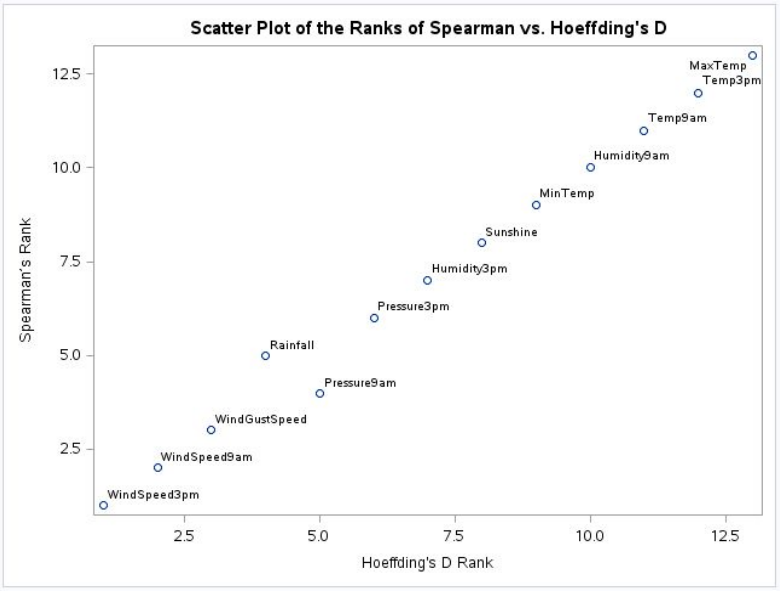

Figure 18: Scatter Plot of the Ranks of Spearman vs Hoeffding's D

For detecting non-monotonic, we look for low ranks in "*Spearman's rank of variables*" column and high rank in "*Hoeffding's D rank of variables*" column, which we can spot in **Figure 19** in the bottom right corner. There are no variables in this corner. We conclude that there is not sufficient evidence of variables with non-monotonic relationship.

Generally, irrelevant variables are found in the bottom left corner due to their poor rank on both metrics. Criterion for deciding which variables to eliminate is subjective. In our case, we will use the p-values from **Table 21** for both metrics. By looking at **Table 21** we can confirm there is no variable with p-values higher than 0.05. and conclude that we do not need to eliminate any variable at this stage.

Subset Selection

Now that we have concluded preparing our data, we will proceed to select the most relevant variables for our prediction model. We will not use any interaction terms, transformation or higher order terms. The goal for our prediction model is to use RainTomorrow as the response variable and all the other variables as potential predictors. We will use `PROC LOGISTIC` procedure in SAS to accomplish this task. This task will be used with our training data. By using *stepwise automatic model selection* we get a final model with 8 predictors plus intercept. Our final model is:

`Model:` `Intercept WindGustSpeed WindSpeed9am WindSpeed3pm Humidity3pm Pressure9am Pressure3pm RainToday Cloud3pm`

By looking at **Table 22** we ensure that we have reached convergence. As seen in the **Table 23** and **Table 24** we can also appreciate that the model is valid for Likelihood Ratio, Score and Ratio tests and has an AIC = 996.093.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Table 22: Model Conversion Table

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1459.777	996.093
SC	1464.985	1079.431
-2 Log L	1457.777	964.093

Table 23. Model Fit Statistics

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	492.4958	15	<.0001
Score	458.1894	15	<.0001
Wald	273.4012	15	<.0001

Table 24. Results for Likelihood Ratio, Score and Wald Tests

Figure 19 below shows how the predictors of our model are relevant. If they cross 1.00 it means that they are not relevant. Even though some levels of Cloud3pm are not relevant, predictor is kept in the model since other levels are.

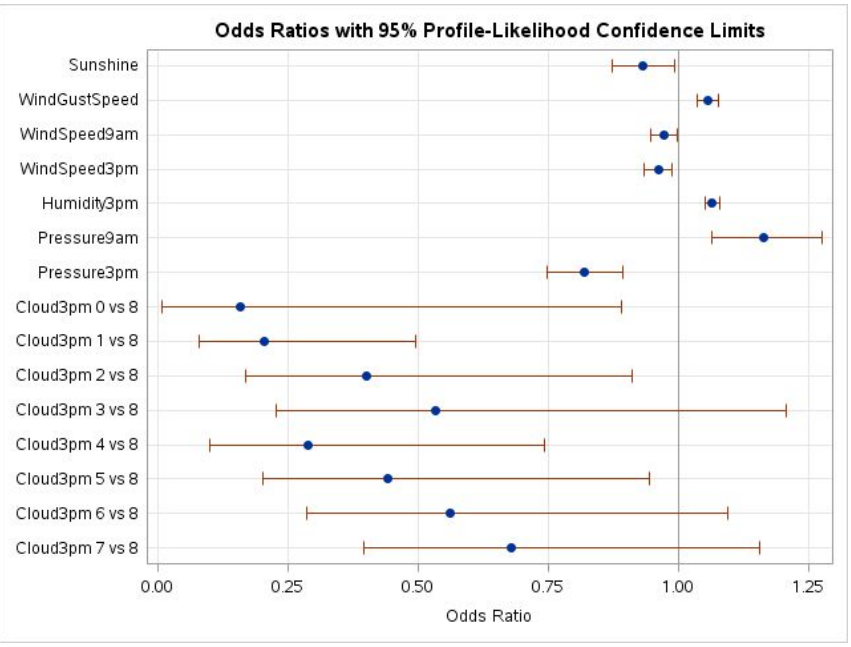

Figure 19: Odds Ratios with 95% Confidence Limits

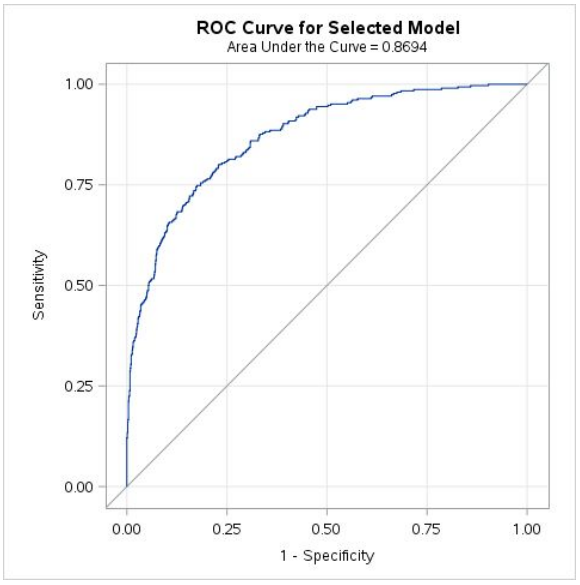

Figure 20: ROC Curve for Selected Model

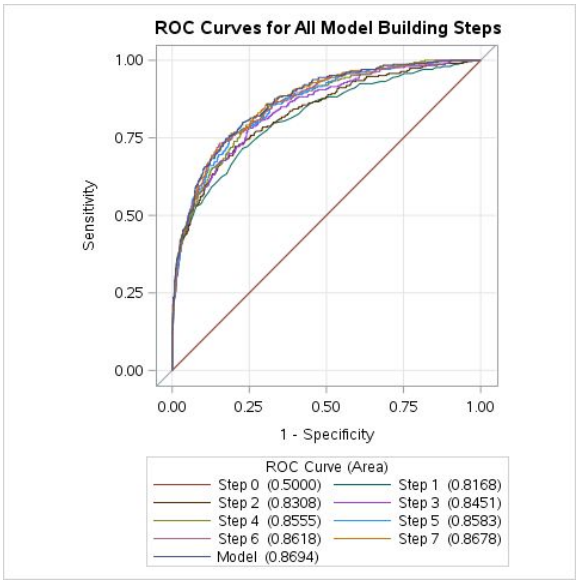

Figure 21: ROC Curves for All Model Building Steps

Figure 20 shows above the improvement in ROC curve after each step. **Figure 21** shows that our final model fits 0.86 of the data. We conclude that our model has a relatively good performance.

Measuring Model Performance

It is of interest to see whether a specific subset of the variables can be used to make accurate predictions about whether it will rain tomorrow. We will consider a model that only uses the variables "RainToday" and "MaxTemp" as predictors. By using PROC LOGISTIC in SAS, we will compare the model predictive performance with the "best" model obtained in the previous section using ROC curves. We get the following information:

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Best Model	0.8630	0.0164	0.8309	0.8951	0.7260	0.7260	0.2638
Simple Model	0.6759	0.0275	0.6220	0.7299	0.3519	0.3525	0.1279

Table 25: ROC Association Statistics

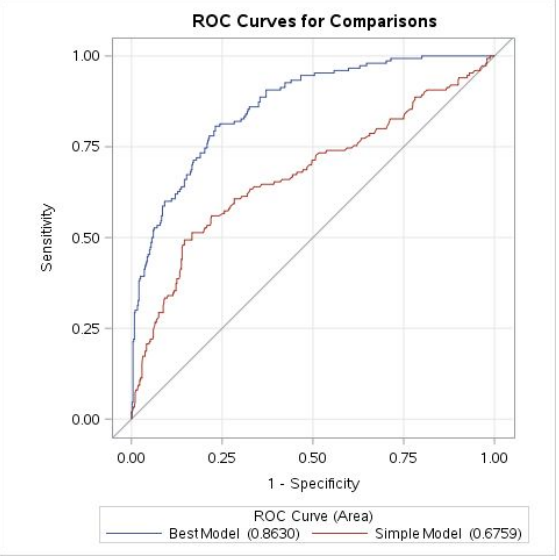

Figure 22: ROC Curves Comparisons

As seen in **Table 25** and **Figure 22** we can appreciate that the Area covered by the "Best" model is 0.8630 vs only 0.6759 from the simple model. We can conclude that the "Best" model found in previous sections has a better performance.

Code:

```
data work.training;
       set work.training;
       if RainTomorrow = "No" then RainTomorrowBinary = 0;
       else RainTomorrowBinary = 1;
run;
```

```
data work.test;
       set work.test;
       if RainTomorrow = "No" then RainTomorrowBinary = 0;
       else RainTomorrowBinary = 1;
run;

proc logistic data=work.training;
       class RainToday;
       model RainTomorrowBinary(event='1')=WindGustSpeed WindSpeed9am
WindSpeed3pm Humidity3pm Pressure9am Pressure3pm RainToday Cloud3pm;
       title "ROC Curve for complex model";
       score data=work.test out=testAssess(rename=(p_1=p_complex))
outroc=work.roc;
run;

proc logistic data=work.training;
       class RainToday;
       model RainTomorrowBinary(event='1')=RainToday MaxTemp;
       score data=work.testAssess out=testAssess(rename=(p_1=p_simple))
outroc=work.roc;
run;

proc logistic data=work.testAssess;
       model RainTomorrowBinary(event='1')=p_complex p_simple/nofit;
       roc "Best Model" p_complex;
       roc "Simple Model" p_simple;
       roccontrast "Comparing Models";
run;
```

Conclusion

After performing our we have found a relatively good model to make accurate predictions about whether it will rain tomorrow in case that we are worried of a drought. Even though the model seems to be good, improvements could be done to the final model in order to improve its performance. We could do further analysis dealing with missing values and perform imputation or variable clustering. It is also worth noting that our final model has variables that are linearly correlated, which could cause multicollinearity. Further analysis and model tuning is advised.