SAS Analytics - Lab Report (STATS 5066)

Drought Monitoring: Analysis of meteorological variables based on national weather data using SAS

Submitted to:

Dr Benn Macdonald
Lecturer in Statistics
School of Mathematics & Statistics
University of Glasgow

Report Prepared by:

Student GUID: 2383746W
MSc in Data Analytics
School of Mathematics & Statistics
University of Glasgow

Page count: 22

January 24, 2018

Contents

List of Tables

List of Figures

1 Introduction

The weather forecast has been one of the most important skills for any civilization throughout history. Agriculture remains one of the most important economic sectors in many European countries, thus it is of great importance to maintain high levels of crop production in order to prevent harvest losses caused by extreme weather conditions such as droughts.

This report carries out statistical analysis on a given weather dataset comprising several climatic variables from an unknown European country. One of the primary aims is to notify national authorities about the possibility of a drought event, for which the average temperature level serves as a crucial indicator. Further, it is of interest to investigate whether the mean maximum and mean minimum temperatures differ. Significant temperature variations may lead to swift changes in atmospheric pressure and increase the probability of precipitation. Since droughts can be associated with high evaporation rates, it is important to quantify the influence of wind gust directions on the evaporation. Ultimately, the attempt is to build a probabilistic model that predicts the occurrence of precipitation for the following day.

2 Exploratory Data Analysis

This section explores the nature of the provided data set "*weather.sas7bdat*" and graphically examines the relationship between variables of interest, before attempting to detect further weather characteristics using more formal statistical analysis in section 3.

2.1 The dataset of interest

The given data set consists of 4657 independent weather-related records and contains 24 meteorological variables. Table 1 and Table 2 show the sixteen quantitative as well as eight categorical variables respectively.

Table 1: Quantitative Variables

Variable	**Type**
MinTemp	Continuous
MaxTemp	Continuous
Rainfall	Continuous
Evaporation	Continuous
Sunshine	Continuous
WindGustSpeed	Continuous
WindSpeed9am	Continuous
WindSpeed3pm	Continuous
Pressure9am	Continuous
Pressure3pm	Continuous

Cloud9am	Discrete
Cloud3pm	Discrete
Humidity9am	Continuous
Humidity3pm	Continuous
Temp9am	Continuous
Temp3pm	Continuous

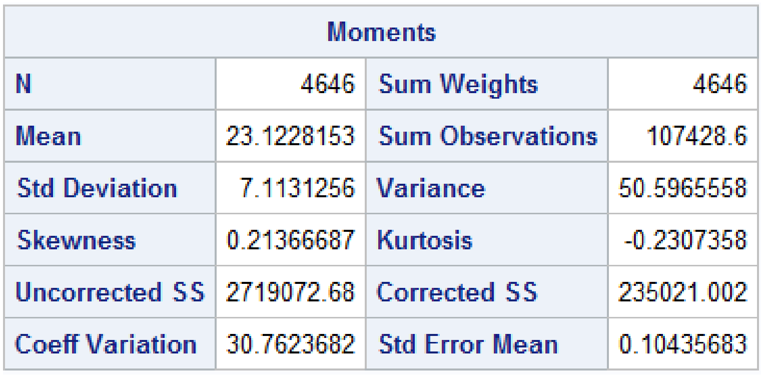
Table 2: Categorical Variables

Variable	**Levels**
NewEquipment	Yes, No
WindGustDir	N, E, S, W, NE, NW, SE, SW, NNE, NNW, SSE, SSW, ENE, ESE, WNW, WSW
WindDir9am	
WindDir3pm	
Location	North, East, South, West, NorthEast, NorthWest, SouthEast, SouthWest
Status	Staff, Student
RainToday	Yes, No
RainTomorrow	Yes, No

2.2 Descriptive Statistics and Graphs

For the country of interest, an average maximum temperature of 25℃ or above is set as a critical threshold, indicative for a drought. In order to decide whether it is required to inform the authorities about a possible drought, we obtain the descriptive statistics (Table 3) as well as histogram and boxplot (Figure 1) for the *MaxTemp* variable and read the mean.

Table 3: Moments of Maximum Temperature



Moments			
N	4646	Sum Weights	4646
Mean	23.1228153	Sum Observations	107428.6
Std Deviation	7.1131256	Variance	50.5965558
Skewness	0.21366687	Kurtosis	-0.2307358
Uncorrected SS	2719072.68	Corrected SS	235021.002
Coeff Variation	30.7623682	Std Error Mean	0.10435683

Maximum temperatures throughout the year have a mean of just over 23℃. With a rather significant standard deviation of 7.11, which means that the mean is different from 25℃.

The logical reason for the standard deviation and variance being relatively high is that the maximum temperature alters during different seasons.

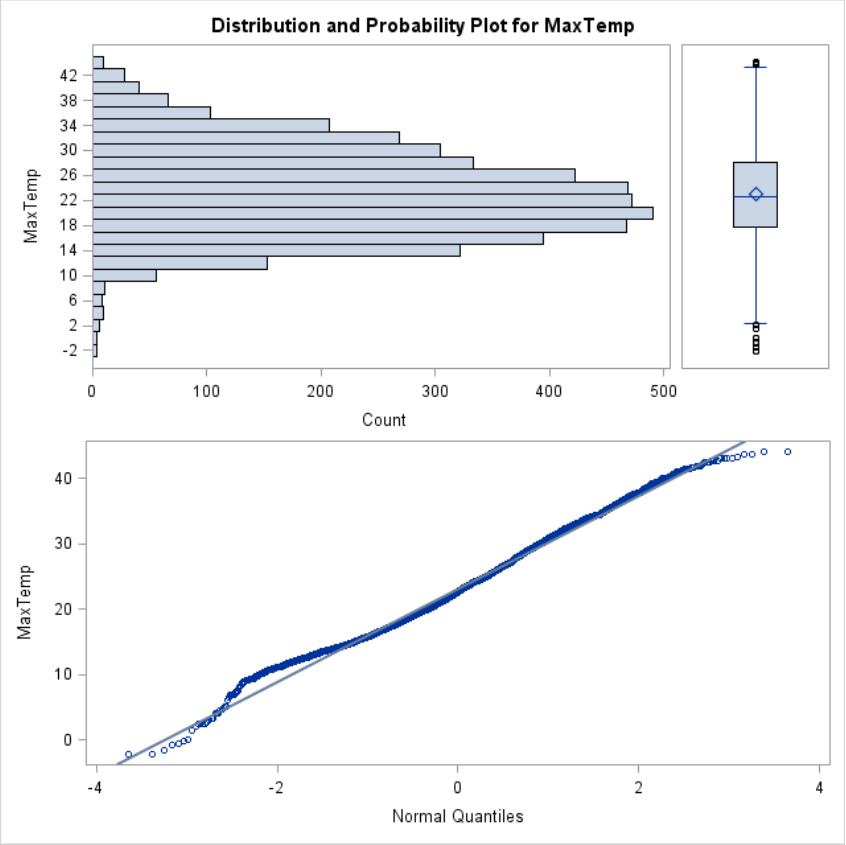


Figure 1: Histogram, Boxplot and Q-Q Plot for variable MaxTemp

From Figure 1, we can recognise that most outliers are located towards the lower part of the temperature scale and that the mean is larger than the median. This indicates that the distribution for the maximum temperature variable is slightly right-skewed. The Q-Q Plot captures this departure from normality in the form of curvature around the main diagonal line. It is easy to see that 75% of the observations approximately fall into maximum temperatures between 18 and 28 degrees Celsius. Since this is a rather high range, this might be an indication for a south European location of the unknown country.

In order to visually detect a difference between the maximum and minimum temperatures, we add a 45-degree reference line to the scatterplot in Figure 2. Along with this line, both temperature measures are equal while departures in either direction suggest there is a difference.

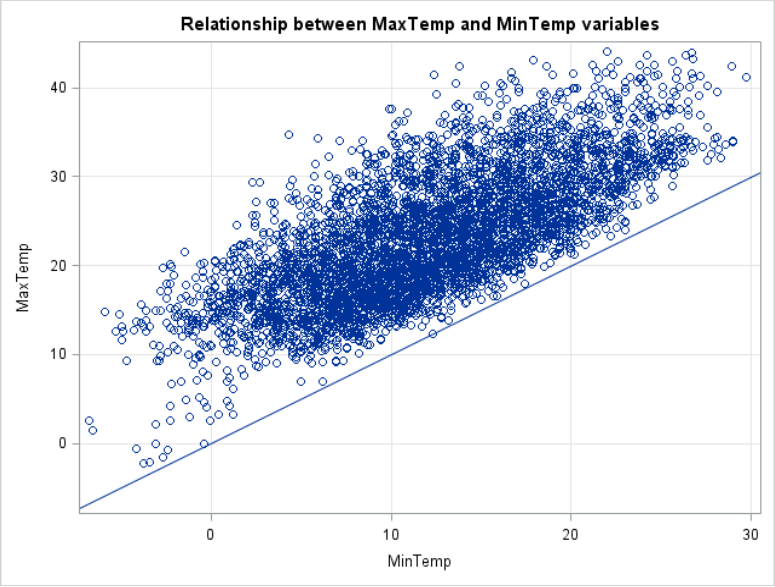


Figure 2: Scatterplot of variables MaxTemp and MinTemp with a reference line

Since all data points lie clearly above the main diagonal, we have a positive difference of roughly 10℃ between the two measures. Thus, it is very unlikely for the means of *MaxTemp* and *MinTemp* to be the same. In Figure 3, we also are interested to find out whether there is a linear association between the minimum and maximum temperature.

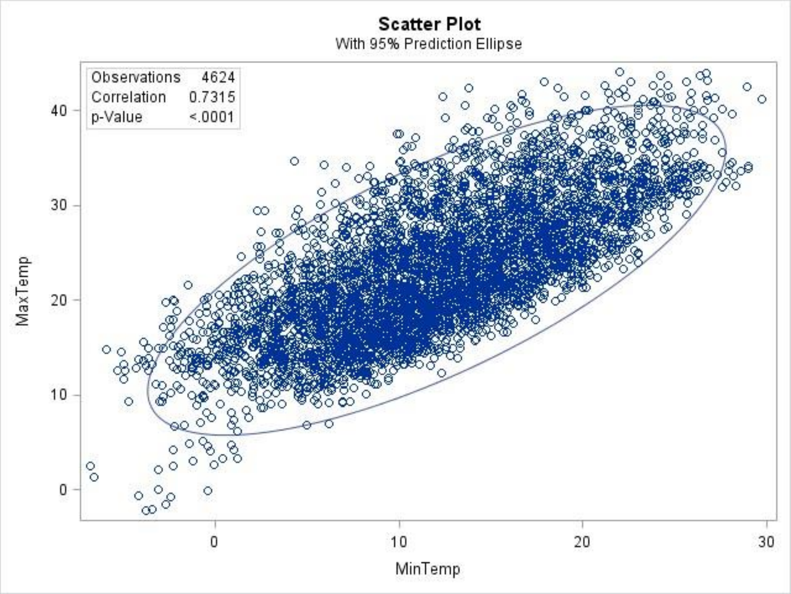


Figure 3: Linear association between variables MaxTemp and MinTemp

For the 4624 observations in total, the correlation coefficient is 0.7315. This means that both temperature measures are highly dependent on one another. This result makes sense since extreme observations for either variable is likely to have a direct effect on the other. For example, a very hot day is usually followed by a milder than average night. The dependency of two variables is an important result for further analysis, as it justifies the choice for a paired t-test rather than a two-sample t-test in chapter 3.2 of this report.

From Figure 4, we note that *Evaporation* has many outliers towards higher values of the scale. This may be attributed to the relatively high temperatures or a geographic location within continental Europe (dryer air leads to higher evaporation).

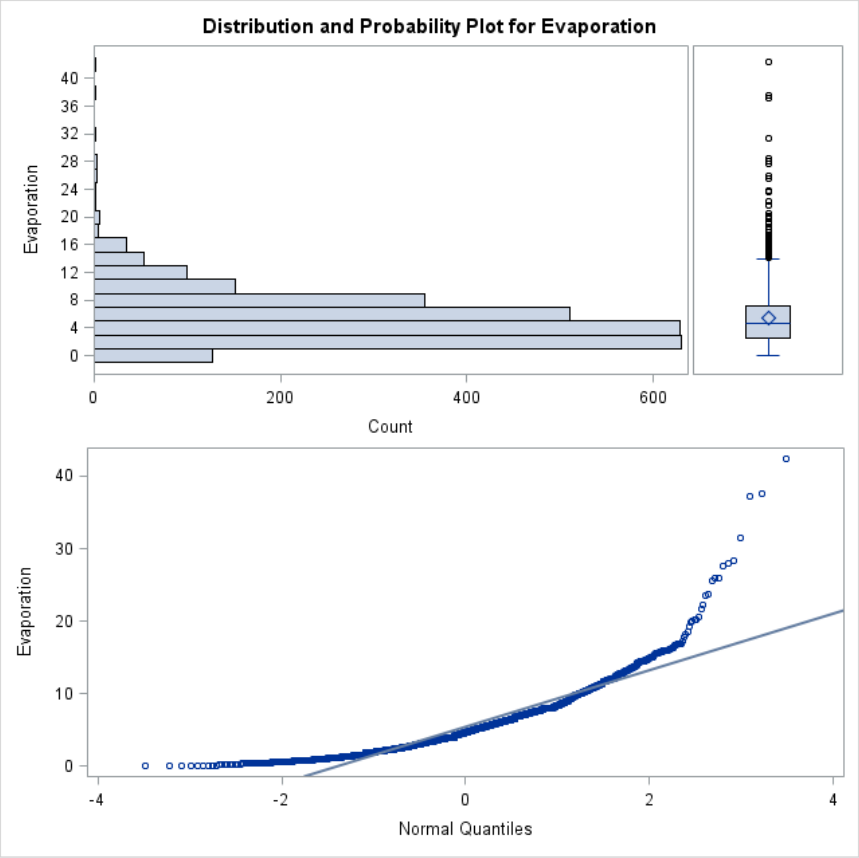
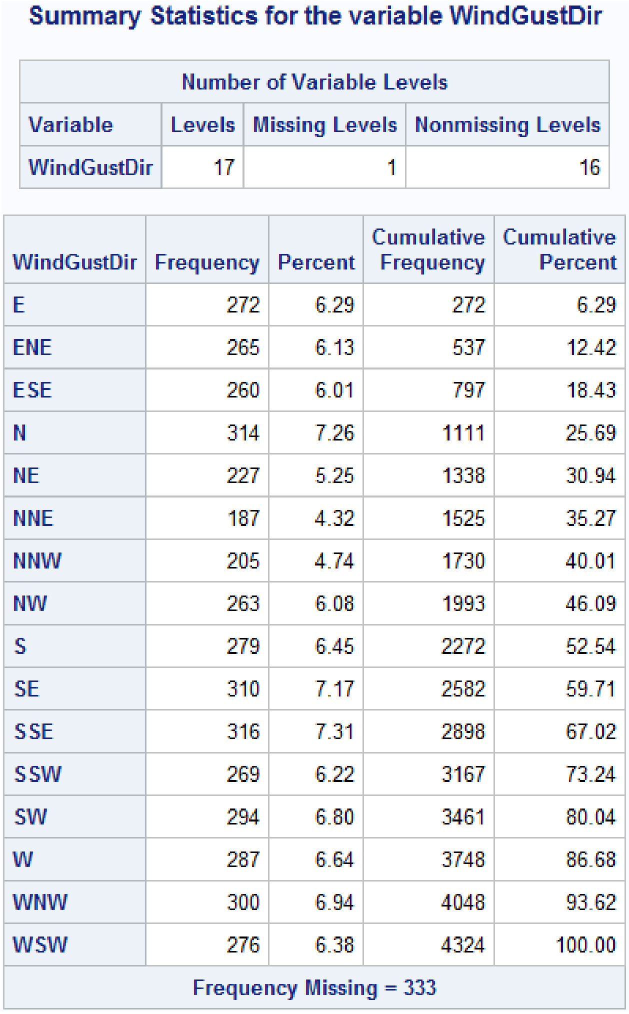


Figure 4: Histogram, Boxplot and Q-Q Plot for the variable Evaporation

There is a considerable right skew in the data and the Q-Q Plot shows a strong curvature. A large number of extreme observations greatly distort the distribution of the evaporation variable. One possible explanation is because there are 2041 missing values for *Evaporation*, which can influence the data distribution and interquartile range.

Finally, we wish to obtain the summary statistics for the *WindGustDir* variable. By comparing the frequencies for different wind gust directions, it can be realised that northerly and southerly wind gusts take place more frequently than those from any other direction do. This is likely to have an impact on evaporation rates as these alter with different wind directions. There are also 333 missing frequencies, which might be contained in the missing level.

Table 4: Summary Statistics for the WindGustDir variable



Summary Statistics for the variable WindGustDir

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				

3 Formal Analysis

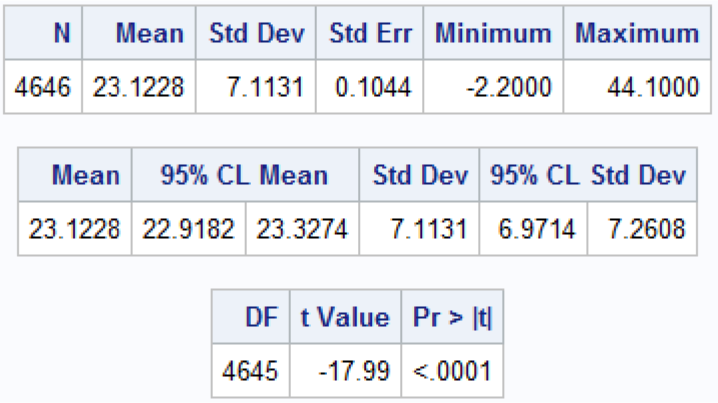
This part will more thoroughly investigate the findings already obtained in the previous Explanatory Analysis and move from the initial purpose of weather monitoring to weather forecasting at the end. The beginning sections cover mainly hypothesis testing, while the latter parts focus more on modelling and prediction. We start the analysis motivated by our main questions of interest.

3.1 Drought Detection

In order to examine the hypothesis that the mean maximum temperature is similar to the drought temperature of 25℃, it is necessary to perform a one-sided t-test with the null hypothesis: $H_0 \colon \mu_{MaxTemp} = 25$.

The summary statistics of this test are shown in Table 5.

Table 5: Summary Statistics and p-value for a one-sample t-test



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

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

The mean of the variable *MaxTemp* for this sample is 23.1228, which is significant (t = -17.99, p-value < 0.0001) at the 5% level. Therefore, it is likely for the population to mean to differ from the null value of 25. Additionally, Figure 5 below shows a shift of the 95% confidence interval (22.9182, 23.3274) to the left of the null value, thus it is not included in the aforementioned interval.

We reject the null hypothesis and conclude that the mean maximum temperature is highly likely to be lower than 25°C. Based on this result it is unlikely for a drought to take place and consequently, there is no need to inform the national authorities about such.

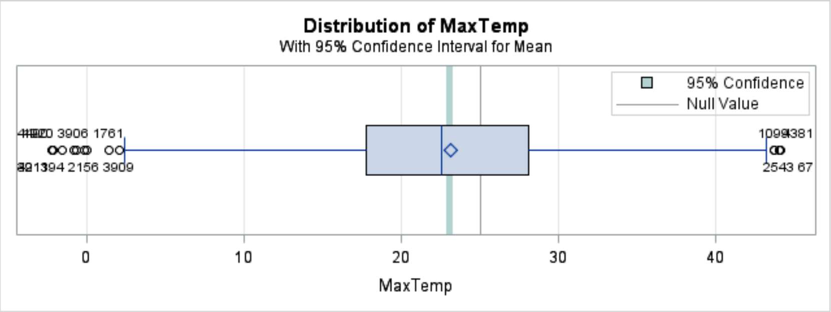


Figure 5: Boxplot with 95% CI for the mean of variable MaxTemp

In order to check the validity of the one-sample t-test, it is of importance to investigate whether the assumptions of independent observations and normality are satisfied. According to the study design, all 24-recorded meteorological variables are independent observations. Figure 6, suggest a right-skewed distribution for the *MaxTemp* variable as the kernel is shifted towards lower values of the maximum temperature. The notable departures of the residuals in the tails of the Q-Q Plot make normality seem questionable.

Since not all assumptions for the t-test are satisfied, it would make sense to proceed with a non-parametric test such as the Wilcoxon Signed-rank test, which does not require strict distributional requirements. Interestingly, the outcome of this test would be identical to the one-sample t-test, since by definition the median (η) in a right-skewed distribution is smaller than the mean ($\eta_{MaxTemp} = 22.6$). Thus, we reject the null hypothesis (H_0: $\eta_{MaxTemp} = 25$) and conclude that with a confidence level of 95%, the median maximum temperature is likely to be lower than 25℃.

As a result, both tests reject the possibility of a drought in the given European country.

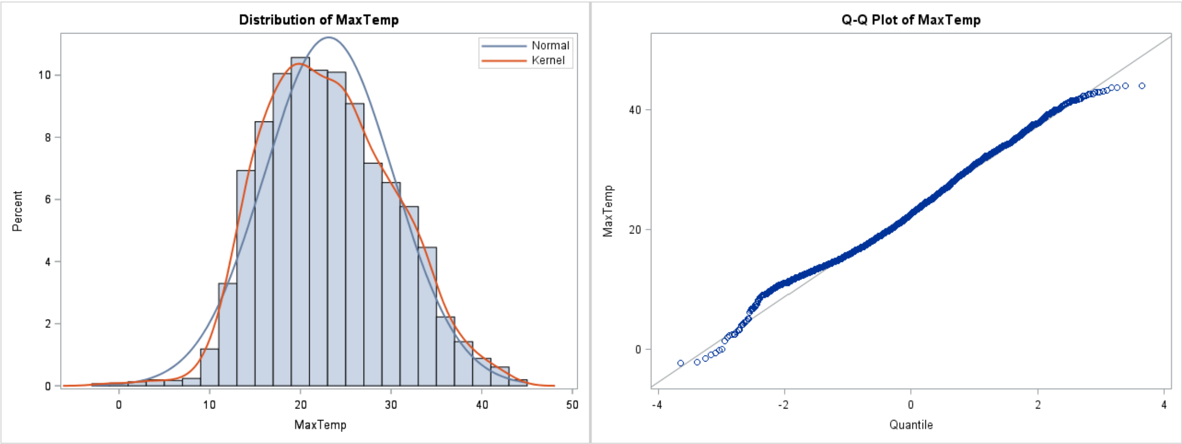


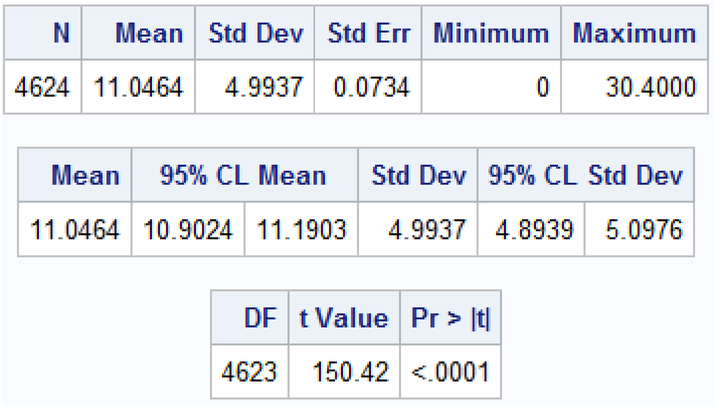
Figure 6: Histogram and Q-Q Plot for the variable MaxTemp

3.2 Mean Temperature Differences

Since national authorities want to stay informed about the risk of extreme weather conditions, the second question of interest is whether the mean maximum temperatures and mean minimum temperatures are different. Because the temperature measures come from two linearly related samples (see correlation in Figure 3 of chapter 2.2), we run a paired t-test.

Formally, the hypothesis is H0: μ_D =0.

Table 5: Summary Statistics of the paired t-test



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

The two temperature measures are paired variables (*MaxTemp – MinTemp)* with a sample size of 4624. The summary statistics of the difference is displayed (mean, standard deviation, and standard error) along with their confidence limits. The paired t-test is significant (t = 150.42, p < 0.0001) at the 5% level, indicating that there's a difference between the mean maximum temperature and the mean minimum temperature. At the same time, the 95% CL for the mean difference is located between 10.9024 and 11.1903. Figure 7, shows the positive deviation from the null value.

Consequently, we reject the null hypothesis and conclude that there is a statistically significant difference between the two temperature measures. On average, the maximum temperature is roughly 11℃ higher than the minimum temperature, possibly a risk factor for sudden weather changes, strong winds or precipitation.

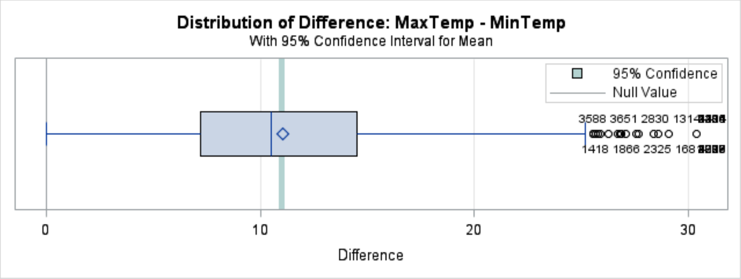


Figure 7: Boxplot with 95% CI for the mean difference MaxTemp – MinTemp

We proceed by analysing the validity of the paired t-test assumptions, which are similar to the ones of the one-sample t-test. Since by study design all observations are independent of one another, differences between *MaxTemp* and *MinTemp* must also be independent. Again, Figure 8 depicts right-skewed data and it is possible to recognise the curvature in the lower tail of the residuals in the Q-Q Plot. Consequently, the normality assumption appears dubious.

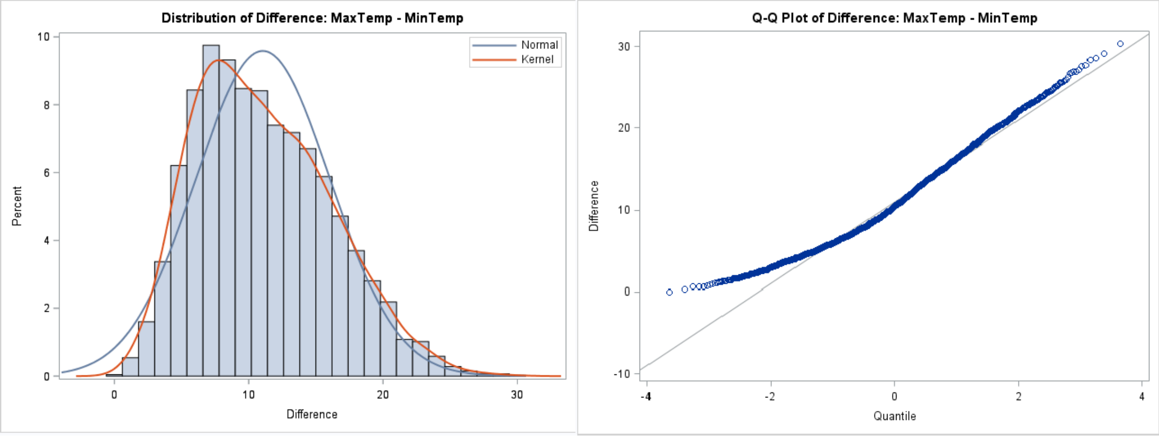


Figure 8: Histogram and Q-Q Plot for the difference MaxTemp – MinTemp

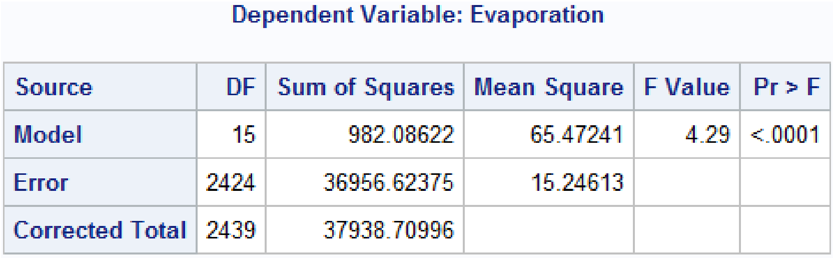
A possible solution to justify the conclusions obtained in the paired t-test is to run a paired Wilcoxon Signed-rank test and check whether it leads to the same result. If this is the case, then the paired t-test is robust towards normality violations.

3.3 Association between Evaporation and Wind gust directions

Depending on the country where the recordings took place, different directions of the strongest wind gusts could have varying influences on the evaporation rate. This can be explained by the fact that air masses over continental regions tend to be dryer (high evaporation), whereas winds coming from the sea are usually more humid (low evaporation).

In order to check if there is a difference in the mean evaporation between the strongest wind gust directions, we use the PROC GLM Procedure and perform a one-way ANOVA model for multiple comparisons with *Evaporation* and *WindGustDir* as dependent and explanatory variables respectively.

Table 6: One-Way ANOVA with *WindGustDir* as an explanatory variable



Dependent Variable: Evaporation

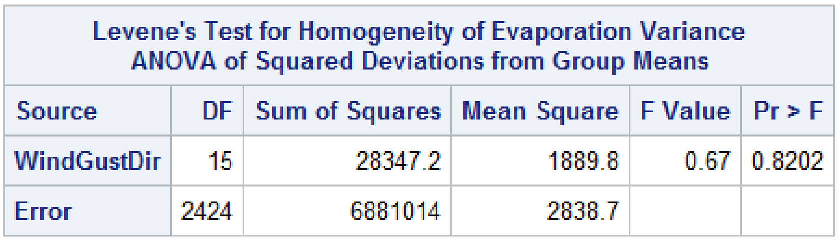
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	982.08622	65.47241	4.29	<.0001
Error	2424	36956.62375	15.24613		
Corrected Total	2439	37938.70996			

Table 6, demonstrates that the effect of the wind gust direction is significant (F=4.29, p-value < 0.0001), so there are differences between the various wind directions. It is necessary to mention that the ANOVA only concerns full data, where both *Evaporation* and *WindGustDir* was recorded and indicates chunks of days, where these values are missing. We go on to verify the ANOVA assumptions of:

(1) independent responses for a given group,
(2) normally distributed residuals,
(3) equality in population variances.

Since the evaporation rate differs depending on individual observations within each level of *WindGustDir* the first assumption seems satisfied. The Fit Diagnostics produced in SAS shows that normality does not hold (strong curvature in the Q-Q Plot), but does not reveal whether the population variances are homogenous. We first can obtain a clearer result by running Laverne's test.

Table 7: Result from the Levene's Test



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		

The result from Table 7 indicates that the Lavene's test is not statistically significant (F = 0.67, p = 0.8202) at a 5% level. Thus, we do not reject the null hypothesis of homogenous population variances. Therefore, the third assumption of the ANOVA model is satisfied.

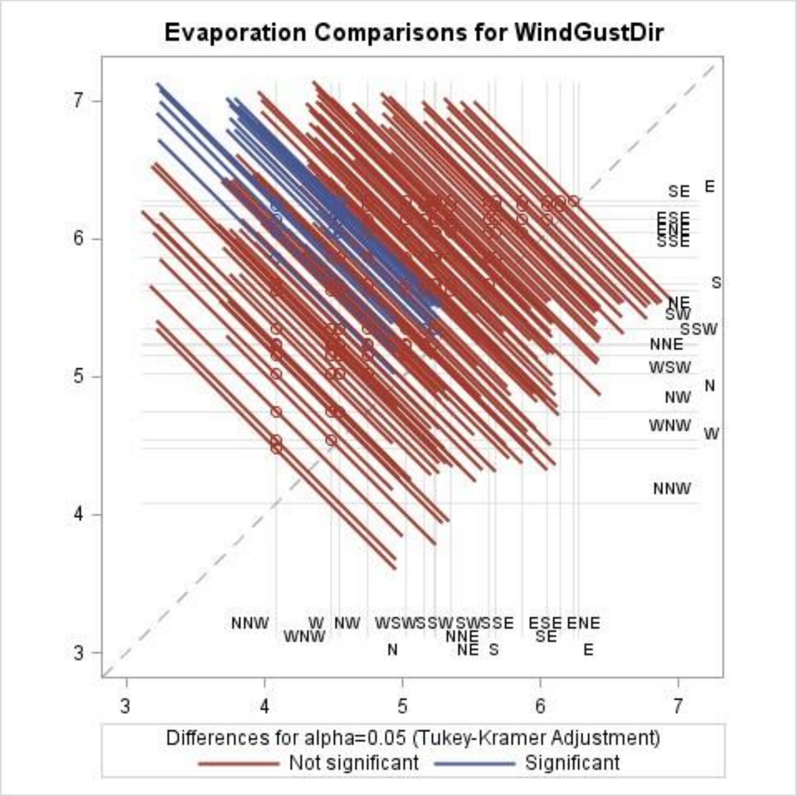


Figure 9: Diffogram of the Evaporation comparisons for variable WindGustDir

Since there are differences between wind gust directions, we perform a post hoc analysis using Tukey's honestly significant difference test as multiple comparisons adjustment model.

The Diffogram in Figure 9 shows that with regard to evaporation rates, the East direction is significantly different from the West, West-North-West and North-North-West directions. Similarly, the South-East, the East-North-East and the East-South-East winds also differ significantly from the same set of directions. Finally, the South-East gust direction indicates a significant difference for evaporation compared to North-North-West direction. The table of p-values for the mean wind gust effect confirms this result and is in each case p<0.05. Therefore, we conclude that the means for the above-listed groups are not the same.

Overall, the results indicate significant differences between westerly and easterly winds in terms of effects on evaporation. One possible explanation for the high evaporation associated with wind gusts from the east is because winds blow from continental (dryer) air masses in mainland Europe or Russia, whereas west winds originate from the Atlantic Ocean (humid) and reduce evaporation. More surprisingly, northern and southern winds do not show significant differences. This could be an indication that the country of interest may be located relatively central within Europe, perhaps halfway between the Baltic Sea in the North and the Mediterranean Ocean in the South.

3.4 Model Selection

In order to investigate the relationship between *Evaporation* (dependent variable) and all other variables (explanatory), it is necessary to choose the most appropriate regression model. In SAS, we use PROC GLMSELECT and specify two different directions, namely Stepwise Selection and Forward Selection. By further selecting the Schwarz Bayesian Information Criterion, it is possible to compare the models. For the model selection procedure, we ignore the variable *RainTomorrow* and limit our scope to the multiple linear regression approach. For simplicity reasons, our aim is to choose the simplest model with the lowest number of explanatory variables.

We start with the stepwise selection procedure and read the six variables contained in the final model from Table 8. Including any other variables means the SBC is no further improvable.

Table 8: Stepwise Selection Summary using SBC local minimum criteria

Step	**Effect Entered**	**Number Effects In**	**Number Parms In**	**Model R-Square**	**Adjusted R-Square**	**F Value**	**Pr > F**
0	Intercept	1	1	0.0000	0.0000	0.00	1.0000
1	MaxTemp	2	2	0.4181	0.4178	1292.78	<.0001
2	Humidity9am	3	3	0.5001	0.4995	294.71	<.0001
3	WindSpeed9am	4	4	0.5207	0.5199	77.39	<.0001
4	MinTemp	5	5	0.5268	0.5257	23.00	<.0001
5	Humidity3pm	6	6	0.5304	0.5291	14.02	0.0002
6	RainToday	7	7	0.5325	0.5309*	7.80	0.0053
* Optimal Value of Criterion							

We continue with the forward selection procedure (Table 9) and note that this gives the same result as the previous stepwise solution.

Table 9: Forward Selection Summary using SBC local minimum criteria

Step	**Effect Entered**	**Number Effects In**	**Number Parms In**	**Model R-Square**	**Adjusted R-Square**	**F Value**	**Pr > F**
0	Intercept	1	1	0.0000	0.0000	0.00	1.0000
1	MaxTemp	2	2	0.4181	0.4178	1292.78	<.0001
2	Humidity9am	3	3	0.5001	0.4995	294.71	<.0001
3	WindSpeed9am	4	4	0.5207	0.5199	77.39	<.0001
4	MinTemp	5	5	0.5268	0.5257	23.00	<.0001
5	Humidity3pm	6	6	0.5304	0.5291	14.02	0.0002
6	RainToday	7	7	0.5325	0.5309*	7.80	0.0053
* Optimal Value of Criterion							

Since both selection procedures are identical, it does not matter which one we chose. Either of the models explains approximately 53% of the variability in *Evaporation*.

The final model specification with six independent variables is:

Evaporation = Intercept, MaxTemp, Humidity9am, WindSpeed9am, MinTemp,
Humidity3pm, RainToday

3.5 Clustering Categorical Data

Combining categorical inputs into clusters can be beneficial since it allows for simpler calculations especially when confronted with a large number of levels. Clustering can reduce the dimensionality of our data and prevent overfitting when carrying out model predictions. We can cluster categorical variables using PROC CLUSTER in SAS, which relies on Greencare's method.

Strategically our goal is to cluster only those categorical variables with many levels. In light of our dataset, the variables, *Wind Gust Direction*, *Wind Direction at 9am* and *Wind Direction* at 3pm have all 16 levels. Within SAS, we obtain the Cluster History for each variable and chose 0.93 as a satisfactory R-Squared cut-off level. Based on this subjective cut-off, it is now possible to group common levels into different clusters. The number of clusters within each categorical variable depends on how informative common group levels are. Generally, it is advisable to find the lowest number of clusters in order to guarantee modelling simplicity.

The first cluster analysis (Figure 10) involves the *WindGustDir* variable.

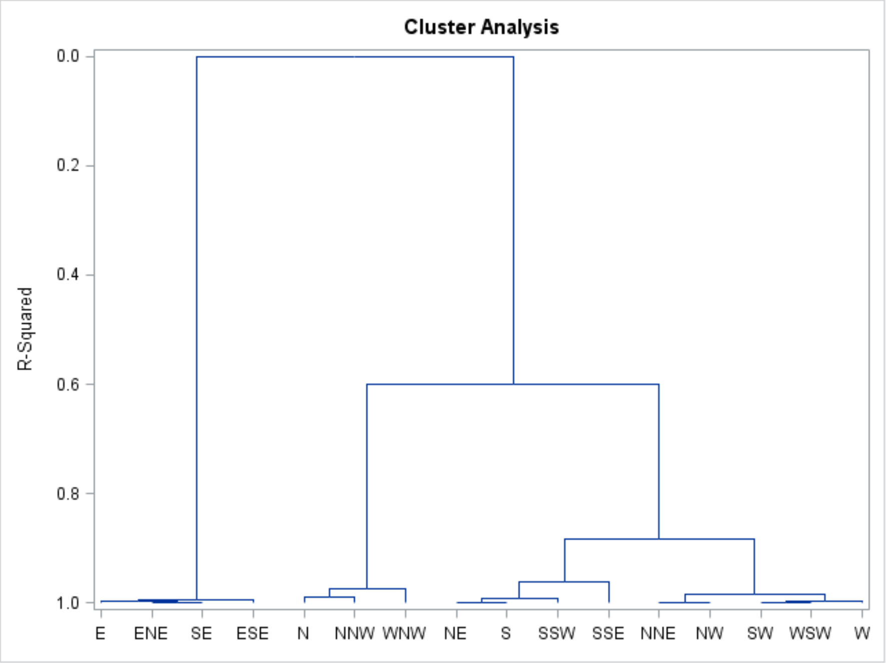


Figure 10: Dendrogram and R-Squared values for variable *WindGustDir*

This allows to create four different clusters, these are:

Cluster 1	E NNE ESSE SW NE ENE

Cluster 2	S SE SSE
Cluster 3	N W NW SSW WSW
Cluster 4	NNW WNW

The second cluster analysis (Figure 11) is about the variable *WindDir9am* and we get.

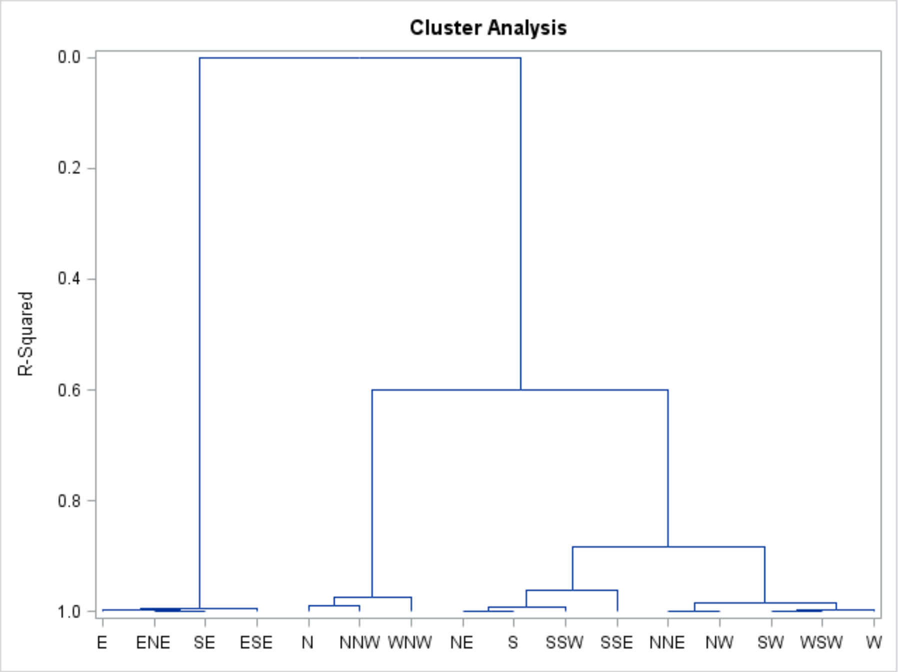


Figure 11: Dendrogram and R-Squared values for variable *WindDir9am*

This enables us to create three different clusters, these are:

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

At last, the third cluster analysis (Figure 12) is about the *WindDir3pm* variable and we obtain the following clusters.

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

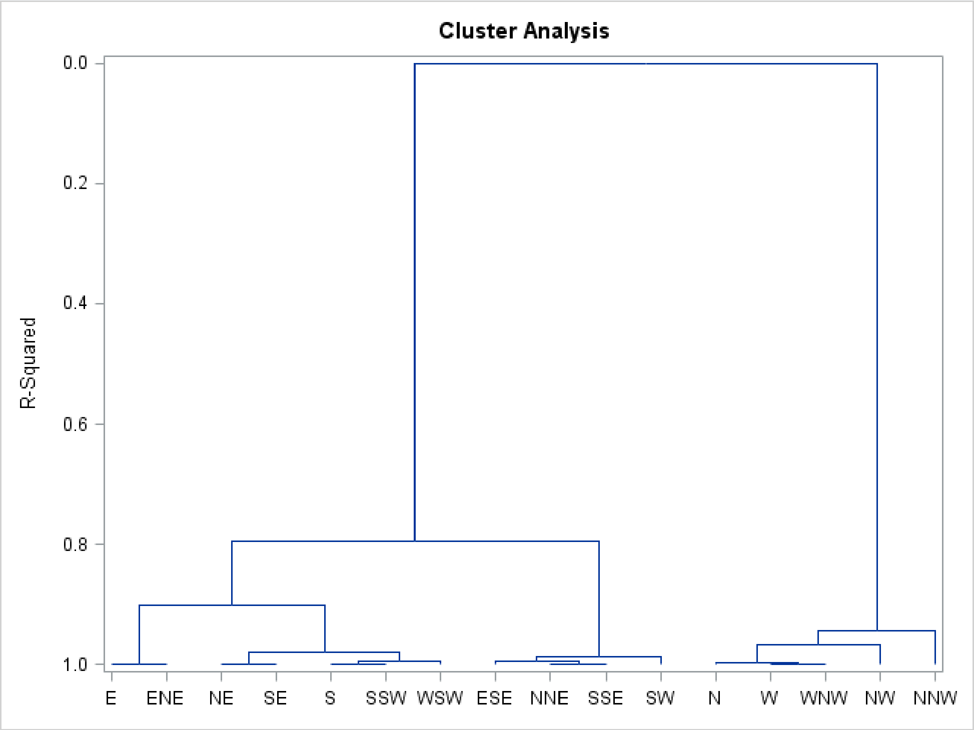


Figure 12: Dendrogram and R-Squared values for variable *WindDir3pm*

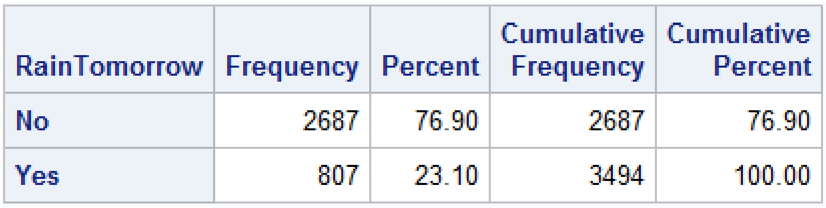
3.6 Splitting the Data

In order to build powerful predictive models, we rely on the idea of splitting data into a training set and a test set. This concept stems from Machine Learning and reduces the risk of overfitting the model. Model fit is a good criterion on the basis of inference, however, when the focus is at learning we wish to be able to generalize the extracted knowledge to untrained or new data. Thus, we need to make room for untreated test data to be able to validate the predictive performance of our fitted model from the training data.

A common way to guarantee a tolerable trade-off between model fit and learning rate is to split the data into approximate proportions of 75% (training set) and 25% (test set). In SAS, we use PROC SURVEYSELECT together with stratified random sampling. The target for the predictive model is the binary variable *RainTomorrow*. The two possible levels are “**Yes**” and “**No**”.

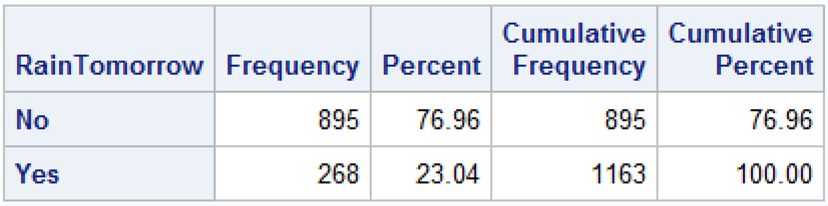
Consulting the frequencies within the training and test set should report roughly identical proportions of “**Yes**” and “**No**” respectively. We examine whether this is the case and run PROC FREQ on both datasets (Table 10 and Table 11).

Table 10: Percentage of Yes and No in the training set



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

Table 11: Percentage of Yes and No in the test set



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

Both tables suggest that there is an equal percentage of **Yes** and **No** within the training data and test data. The code regarding the data split is listed below.

SAS-Code:

```
data weather1;
        set work.weather_combined;
run;

proc sort data=weather1 out=work.weather_sort;
        by RainTomorrow;
run;

proc surveyselect noprint data=weather_sort samprate=.75 outall out=weather_sampling;
        strata RainTomorrow;
run;

data  weather_training(drop=selected  SelectionProb  SamplingWeight)
        weather_test(drop=selected   SelectionProb   SamplingWeight);
        set work.weather_sampling;
        if selected then output weather_trainingset;
        else output weather_testset;
run;

proc freq data=work.weather_trainingset;
  tables RainTomorrow;
run;

proc freq data=work.weather_testset;
  tables RainTomorrow;
run;
```

3.7 Variable Screening

In the logistic regression framework, we have the interest to find out whether there are irrelevant variables or evidence of non-linear relationships for the continuous variables in the training data. We do this by implementing the Spearman Correlation Coefficient and Hoeffding's Dependance Coefficient.

While the Spearman's correlation is a measure of a monotonic relationship between two variables, it does not assume linearly related variables or normality (such as Pearson's correlation does) and can be used for ordinal variables. Further, Spearman's correlation is more robust to outliers and a value close to zero represents an absence of a monotonic relationship between variables, whereas values close to the ends of the scale (-1 to 1) indicate a strong negative/positive monotonic association.

Hoeffding's D correlation (with support -0.5 to 1) is a measure of deviation from independence; it can account for monotonic, non-monotonic and linear associations as well as non-functional relationships. The signs do not have an interpretation.

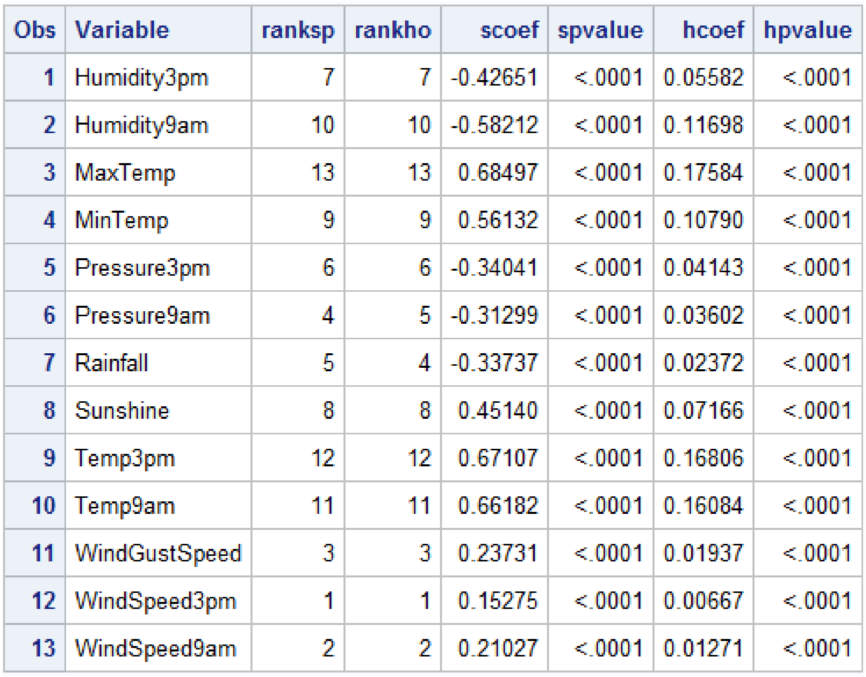
For analytical purposes, we look at the Spearman Rank (SK) and Hoeffding Rank (HK), they relate to each other as follows.

- High SK and High HK: indicates a monotonic relationship
- Low SK and High HK: indicates a non-monotonic relationship
- High SK and Low HK: indicates a monotonic relationship
- Low SK and Low HK: indicates a weak relationship

We can now go on to check whether some of the variables in our training data are irrelevant or have a non-monotonic relationship. One possible way to determine this is to use the hypothesis test H_0: "*There is no monotonic association between variables X and Y*". We obtain the p-values for each variable in Table 12.

Since all p-values are less than 0.001, we conclude that none of the variables appears to be irrelevant or non-monotonic.

Table 12: Rank of Spearman Correlations and Hoeffding's D Correlations



Obs	Variable	ranksp	rankho	scoef	spvalue	hcoef	hpvalue
1	Humidity3pm	7	7	-0.42651	<.0001	0.05582	<.0001
2	Humidity9am	10	10	-0.58212	<.0001	0.11698	<.0001
3	MaxTemp	13	13	0.68497	<.0001	0.17584	<.0001
4	MinTemp	9	9	0.56132	<.0001	0.10790	<.0001
5	Pressure3pm	6	6	-0.34041	<.0001	0.04143	<.0001
6	Pressure9am	4	5	-0.31299	<.0001	0.03602	<.0001
7	Rainfall	5	4	-0.33737	<.0001	0.02372	<.0001
8	Sunshine	8	8	0.45140	<.0001	0.07166	<.0001
9	Temp3pm	12	12	0.67107	<.0001	0.16806	<.0001
10	Temp9am	11	11	0.66182	<.0001	0.16084	<.0001
11	WindGustSpeed	3	3	0.23731	<.0001	0.01937	<.0001
12	WindSpeed3pm	1	1	0.15275	<.0001	0.00667	<.0001
13	WindSpeed9am	2	2	0.21027	<.0001	0.01271	<.0001

Another, approach to conduct variable screening is to examine the rank based relationship in the Scatterplot of Figure 13.

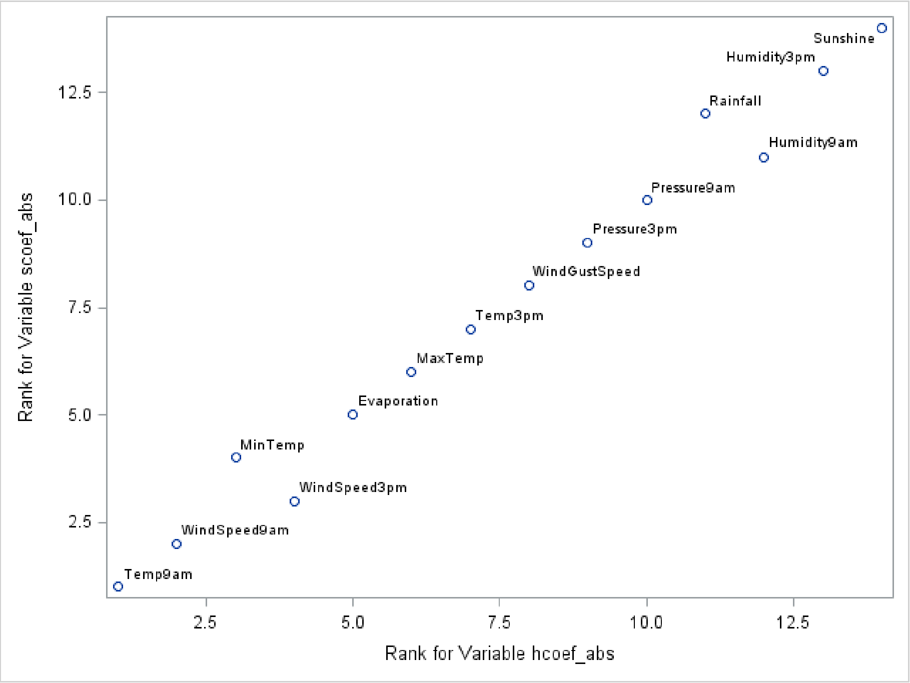


Figure 13: Scatterplot of the Spearman Ranks vs the Hoeffding's D Ranks

Since the bottom right corner (Low SK and High HK) indicates potential non-linearity but there is no variable in this area, we can confirm the result from the previous hypothesis test and conclude that there is not enough evidence of variables with the non-linear association.

3.8 Variable Selection

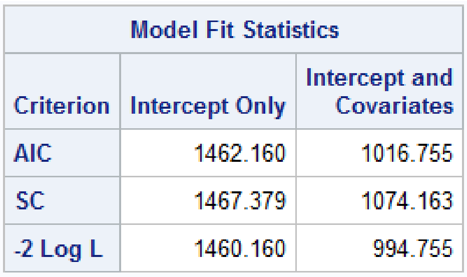
We are interested in building a model that can most accurately predict rain for the next day and since the response, *RainTomorrow* is a binary variable we run a logistic regression on our training data. In order to simplify the process of finding a set of optimal predictors, it is better to assume the absence of higher order terms or any other non-linear transformation. There are three possible automatic model selection directions to reduce the full model to an optimal final model with fewer predictors namely, forward, stepwise or backward.

Using automated forward selection, we get the "optimal" final model:

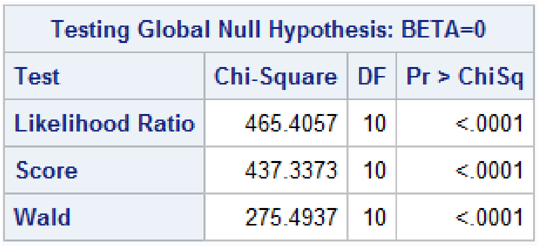
RainTomorrow = Intercept RainToday MaxTemp Sunshine Humidity3pm Pressure3pm
WindGustSpeed WindDir3pmCombinded WindSpeed3pm WindSpeed9am

As shown in Table 13 the model is valid for the 3 tests and has an AIC= 1016.755

Table 13: Model Fit Statistics and Global Null Hypothesis Table



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1462.160	1016.755
SC	1467.379	1074.163
-2 Log L	1460.160	994.755



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	465.4057	10	<.0001
Score	437.3373	10	<.0001
Wald	275.4937	10	<.0001

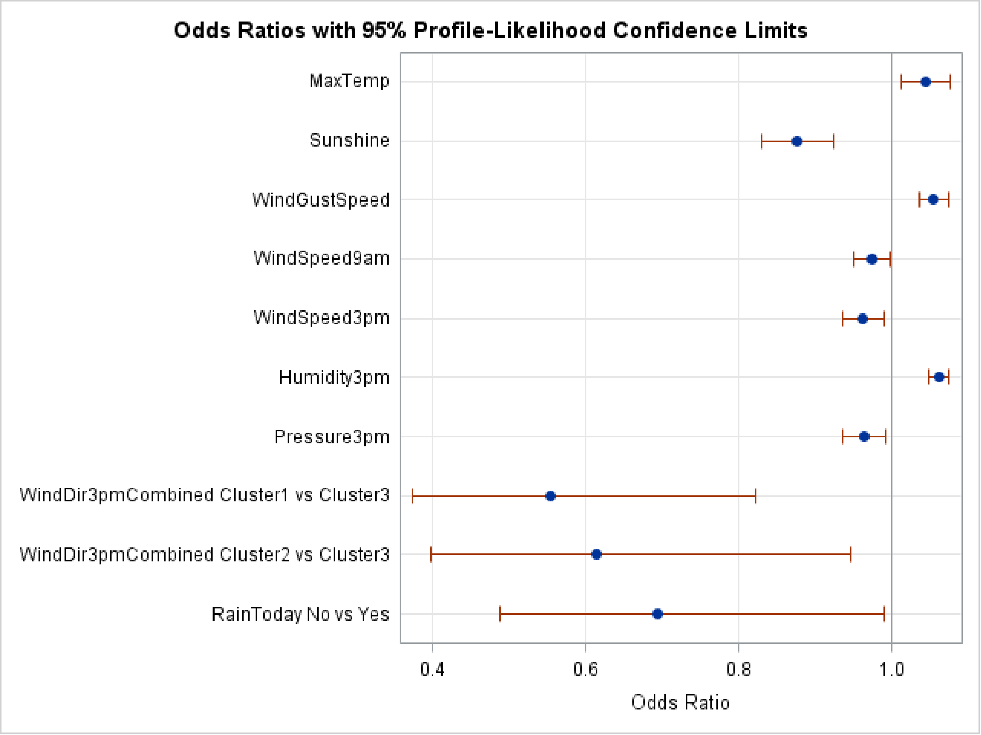


Figure 14: Odds Ratios with 95% Confidence Limits

Since none of our predictors crosses the 1.0 horizontal, reference line they are all significant and should be kept in the model specification.

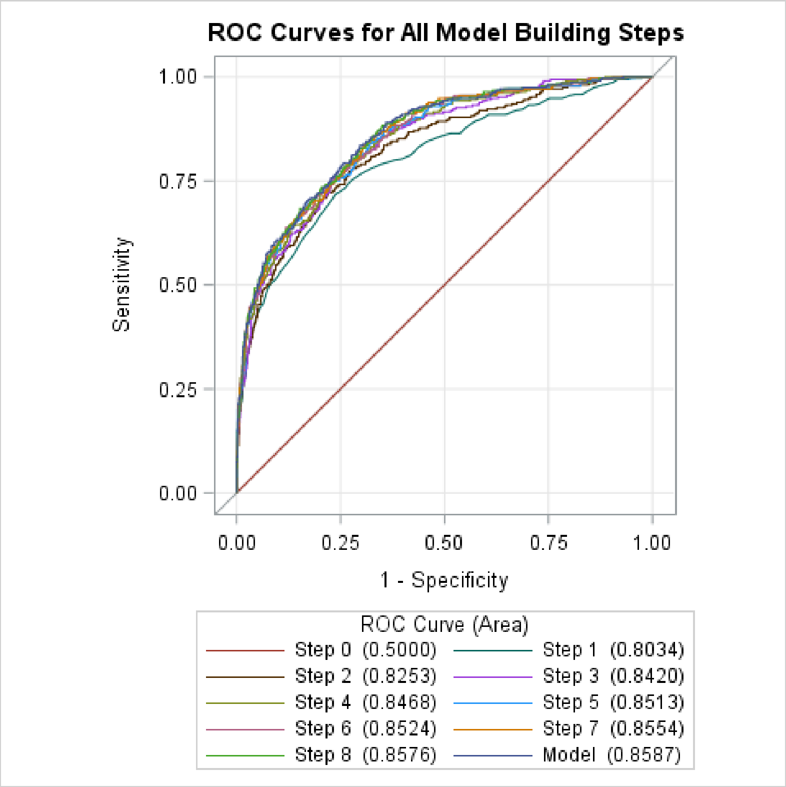


Figure 15: ROC curves for all model-building steps

In Figure 15, we can observe each step in building the optimal model using the forward selection procedure at a p-value of 0.05. Each newly added predictor improves the area under the curve further and the optimal model after step 8 has the maximum AUC=0.8587.

3.9 Model Comparison and Performance Assessment

In practice we wish to simplify predictions as much as possible, thus we only select the predictors *RainToday* and *MaxTemp* to be included in the logistic regression and compare the predictive performance of this “simple model” with the previously obtained “complex model”.

Using only two predictors, we get the “simple” model:

RainTomorrow = Intercept RainToday MaxTemp

The SAS-Code below carries out the comparison between the simple and the optimal model along with a graph for the ROC curves.

SAS Code:

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

data work.weather_test;
```

```
        set work.weather_test;
        if RainTomorrow = "No" then RainTomorrowBin = 0;
        else RainTomorrowBin = 1;
run;

proc logistic data=work.weather_training;
        class RainToday ;
        model RainTomorrowBin(event='1')=Sunshine WindGustSpeed WindSpeed3pm
               Humidity3pm Pressure9am Pressure3pm;
        score data=work.weather_test out=testAssess(rename=(p_1=p_complex)) outroc=work.roc;
run;

proc logistic data=work.weather_training;
        class RainToday;
        model RainTomorrowBin(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 RainTomorrowBin(event='1')=p_complex p_simple/nofit;
        roc "Optimal Model (8 pred.)" p_complex;
        roc "Simple Model (2 ped.)" p_simple;
        roccontrast "Model Comparison";
run;
```

As seen in Figure 16 below, the AUC of the optimal model is higher than the simple model (0.886 > 0.688). On top of that, the ROC Association Statistics (Table 14) is also in favour of the optimal model.

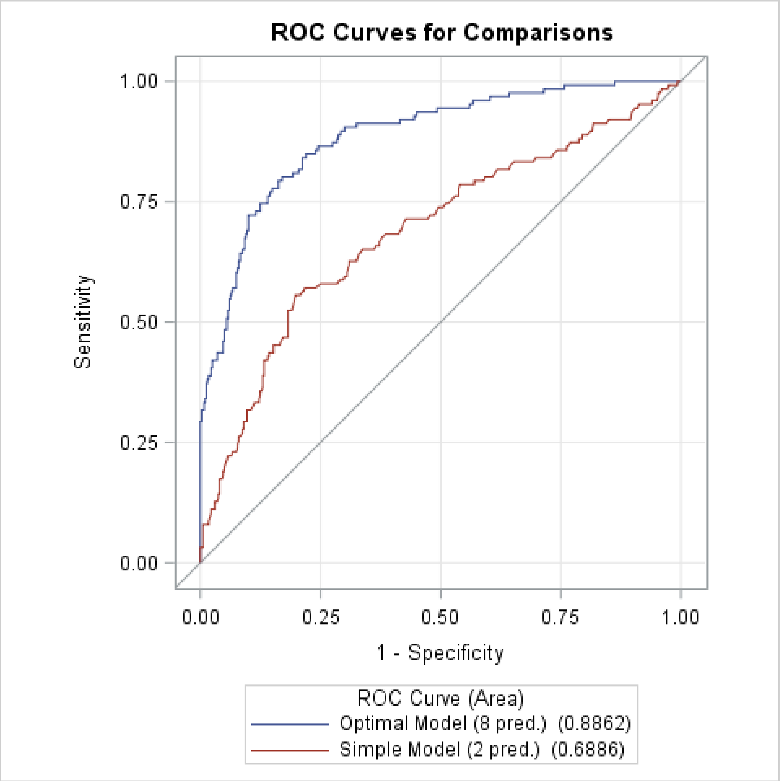
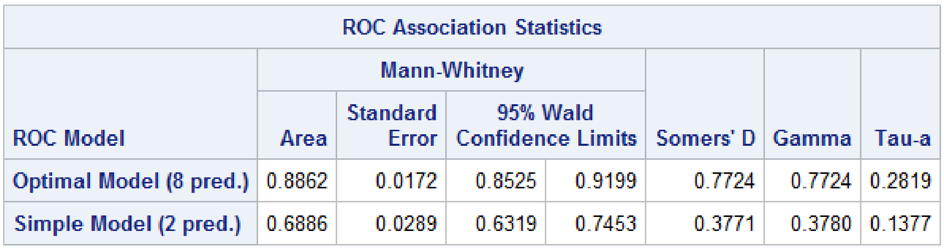


Figure 16: ROC curve comparison between optimal and simple models

Table 14: ROC Association Statistics



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Optimal Model (8 pred.)	0.8862	0.0172	0.8525	0.9199	0.7724	0.7724	0.2819
Simple Model (2 pred.)	0.6886	0.0289	0.6319	0.7453	0.3771	0.3780	0.1377

Therefore, we conclude that the optimal model with eight predictors has better performance and should be preferred over the simple model with just two predictors.

4 Conclusion

This report served to understand the ways in which it is possible to use the provided dataset in order to predict the weather and inform national authorities about weather extremes. We have shown that based on the mean maximum temperature, there is a very low risk of a drought at the unknown European county. The difference between the mean maximum and mean minimum temperature is more than 10℃ in this country, which is quite a large deviation. Since marked temperature differences can cause abrupt changes in air pressure, it is a good idea to inform authorities about the possibility of extreme weather events. It was also possible to determine that the highest evaporation rates are associated with wind gusts from easterly directions (dry continental air masses), while the lowest rates coincide with winds from the west (humidity from the ocean).

Through the GLM approach, it was possible to find a reduced model with only six, instead of all variables in order to describe changes in evaporation rates. Interestingly, varying the model direction (stepwise and backward) under the Schwarz Bayesian Information Criterion did not change the outcome and gave us identical final models. Since the variables related to wind direction have many levels, we combined them into similar clusters in order to reduce the modelling complexity. The chosen 0.93 R-Squared cut off level guarantees the minimal loss of information while maintaining the fit of the data high. Analysis of Spearman's correlation ranks and Hoeffding's Dependence ranks indicated that there was a lack of evidence for non-linearity or irrelevancy among the included continuous variables.

Finally, we found a probabilistic model, which can make predictions about the likelihood of rain for the next day. We have built and compared two distinct models (simple and optimal), using logistic regression with forward selection. It was discovered that the predictors *RainToday* and *MaxTemp* have significant effects on determining whether it will rain tomorrow in the simple model. However, the best predictive performance was achieved with the optimal model consisting of eight significant predictors.