# Table of Contents

# Introduction

## Consultancy Brief

This report contains an investigation of a data set to aid decision-making. The main objective of this report is to investigate the data set provided to determine which variables relate to Actual Temperature and Apparent Temperature.

The report for this investigation is split into two parts, part one of this report shows an exploration into the data set and the distribution of the data (such as the mean, median, standard deviation, outliers etc.), investigating pairwise relationships among the variables. Additionally, one sample and two sample hypothesis testing on the population mean of some variables in the data set such as the actual and apparent temperature.

Part two of the report aim to show the exact type of relationships (if any) between the variables in the dataset, an investigation of the continuous variables in the dataset using correlation techniques and establishing regression models for variables of interest. Essentially, we will investigate subsets of variables that contributes (or causes) to the actual temperature and apparent temperature variable

## The Data Set

The dataset titled "Weather Data – DATA SET 44" contains weather data in the city of Szeged in Hungary obtained as a subset of a large database published on Kaggle (www.kaggle.com). It contains randomly collected weather data over a 10-year period that spans from the year 2006 to 2016. This amounts to 200 observations, with one row of data per observation. With every observation, it shows the following:

- Date + Time
- Summary
- Precipitation Type (either rain or snow)
- Actual Temperature (measured in °C)
- Apparent Temperature (measured in °C)
- Humidity
- Wind Speed
- Wind Bearing
- Visibility
- Pressure

# Initial Data Analysis

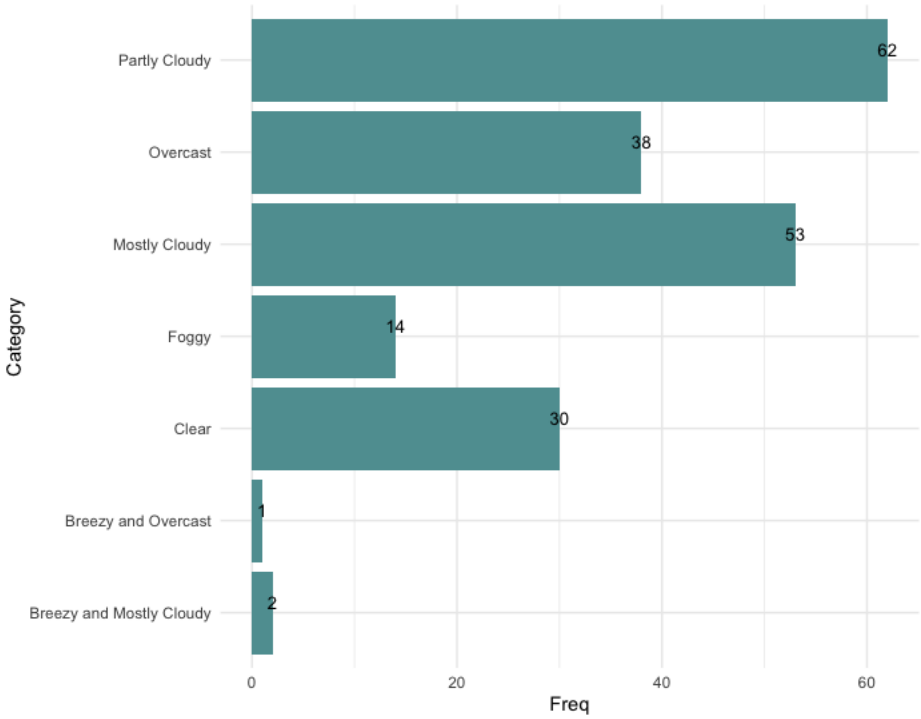
Upon examination of the dataset, the table below shows the variables in the dataset and their respective data types:

| S/N | Variable Name | Data Type |
|---|---|---|
| 1. | Summary | Categorical/Qualitative |
| 2. | Precipitation Type | Categorical/Qualitative |
| 3. | Actual Temperature | Numerical/Quantitative |
| 4. | Apparent Temperature | Numerical/Quantitative |
| 5. | Humidity | Numerical/Quantitative |
| 6. | Wind Speed | Numerical/Quantitative |
| 7. | Wind Bearing | Numerical/Quantitative |
| 8. | Visibility | Numerical/Quantitative |
| 9. | Pressure | Numerical/Quantitative |

The following section shows an exploration and analysis of each of the variable in the dataset.

## Analysis of Summary Variable

The “Summary” variable in the dataset is of a categorical data type. It simply describes the weather condition at a given point in time. The chart below shows a distribution of the various categories in the “summary” variable




The above plot shows there are seven (7) categories in the "summary" variable and their respective distributions. The categories are:

- Breezy and mostly cloudy
- Breezy and overcast
- Clear
- Foggy
- Mostly cloudy
- Overcast and
- Partly cloudy

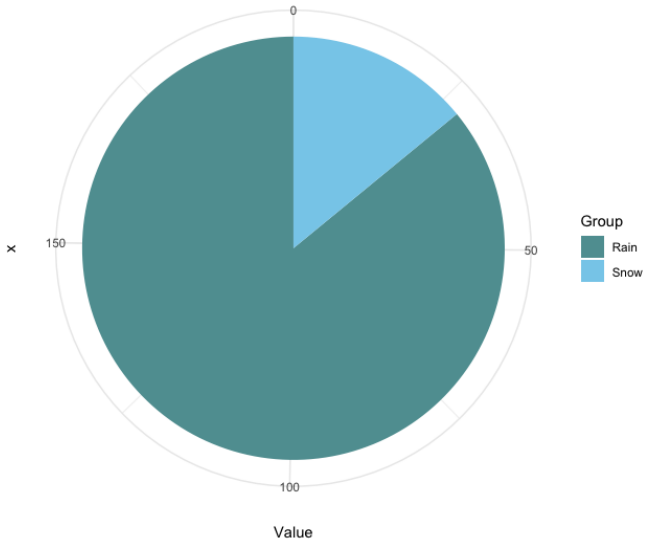
## Analysis of Precipitation Type Variable

The "precipitation type" variable is a categorical data type. This simply describes the type of water falling from the sky. It could be rain, drizzle, snow, hail, sleet, or something even more unusual. Exploring the data in this variable using R, the output is as follows:

```
>
null   rain   snow
   1    171     28
```

The above output indicates that of the 200 observations in the dataset, 171 observations are in the rain category, 28 are in the snow category and there is 1 empty value in the data set as “null” is not a valid precipitation type as such it will be removed from the dataset and replaced with blanks.

The pie chart below shows the occurrence of each group after replacing the null values in the data set with blanks:




## Analysis of Actual Temperature Variable

The “actual temperature” variable is a quantitative (continuous) data type. It expresses the degree of coldness or hotness of the weather as measured by a thermometer. It is measured in degree Celsius (°C).

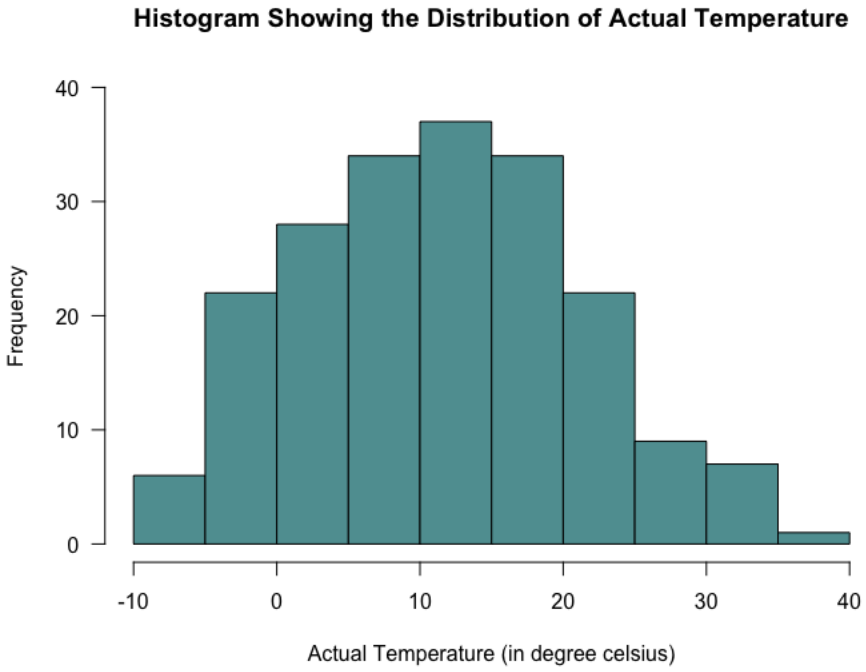
The below R output

>

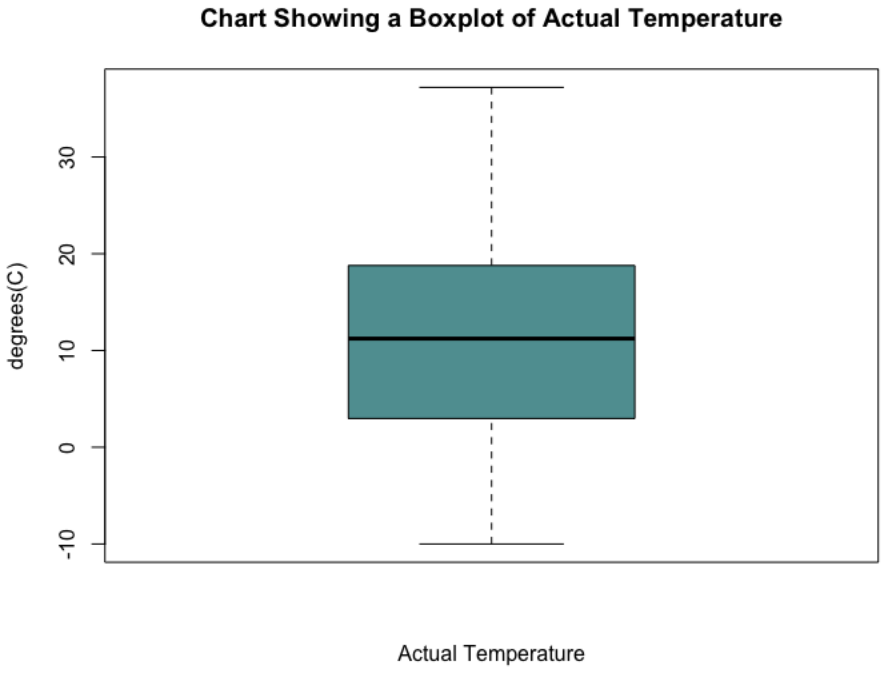
| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| -10.000 | 11.228 | 11.552 | 9.868259 | 37.194 |

The above output shows that the lowest temperature observed in the variable is -10°C while the highest temperature is 37.194°C. The mean and median temperatures observed are 11.552°C and 11.228°C respectively. The standard deviation of the distribution is

9.868259°C which measures the dispersion or variation from the average value of the distribution.




The above histogram plot shows a normal distribution of the observations in the dataset as a bell curve and somewhat tails at both ends of the plot.



Furthermore, the boxplot above shows there are no outliers in the variable. Also, no missing values were observed. The zero values and negative values observed are probable due to the nature of the variable as it is possible to have a temperature of 0°C or -10°C which means a very cold weather is observed.

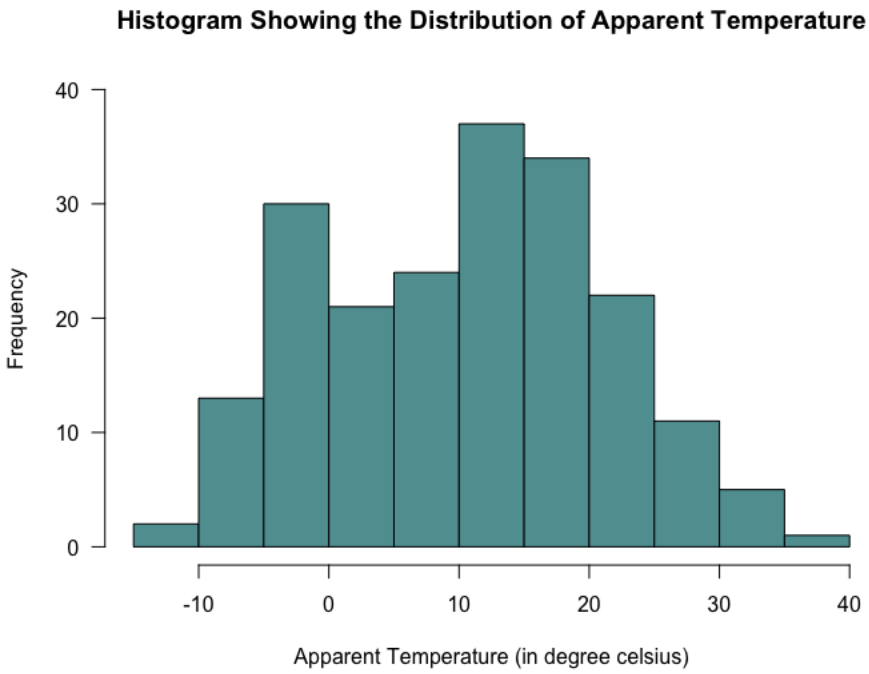
## Analysis of Apparent Temperature Variable

The "apparent temperature" variable is a quantitative (continuous) data type. This is the temperature of what it feels like by humans as opposed to what it is on the thermometer. It is also measured in degree Celsius (°C). The below R output

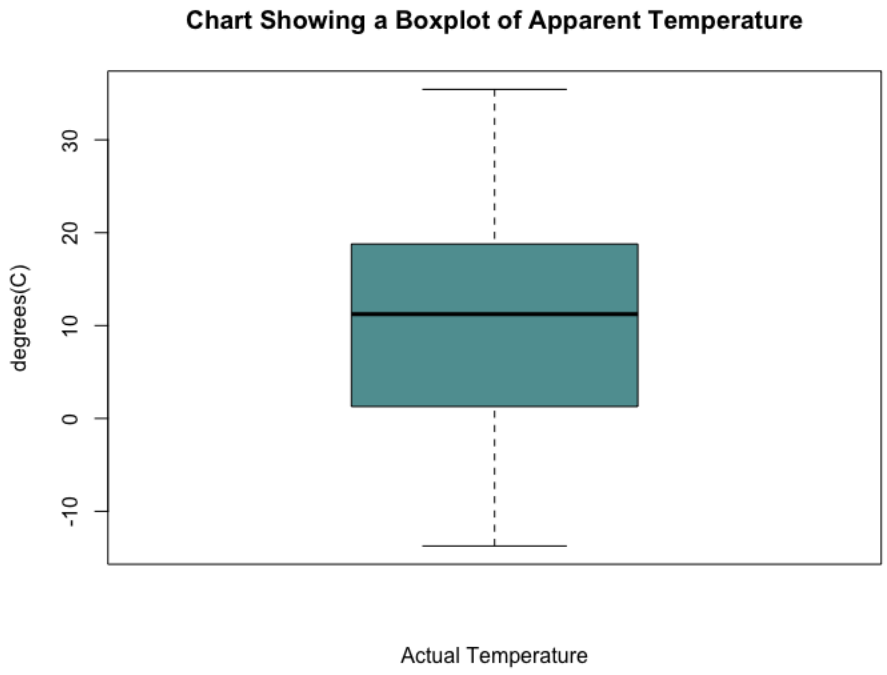
>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| -13.733 | 11.228 | 10.574 | 10.94571 | 35.433 |

The above output shows that the lowest temperature observed in the variable is -13.733°C while the highest temperature is 35.433°C. The mean and median temperatures observed are 10.574°C and 11.228°C respectively while the standard deviation of the distribution from its mean is 10.94571°C.




The above histogram plot slightly appears to be a normal distribution of the observations in the dataset as it tails on both ends and somewhat peaks in the middle.




Furthermore, the boxplot above shows there are no outliers in the variable. Also, no missing values were observed. Similarly, as observed in the actual temperature, it is probable to have zero or negative values due to the nature of the variable.

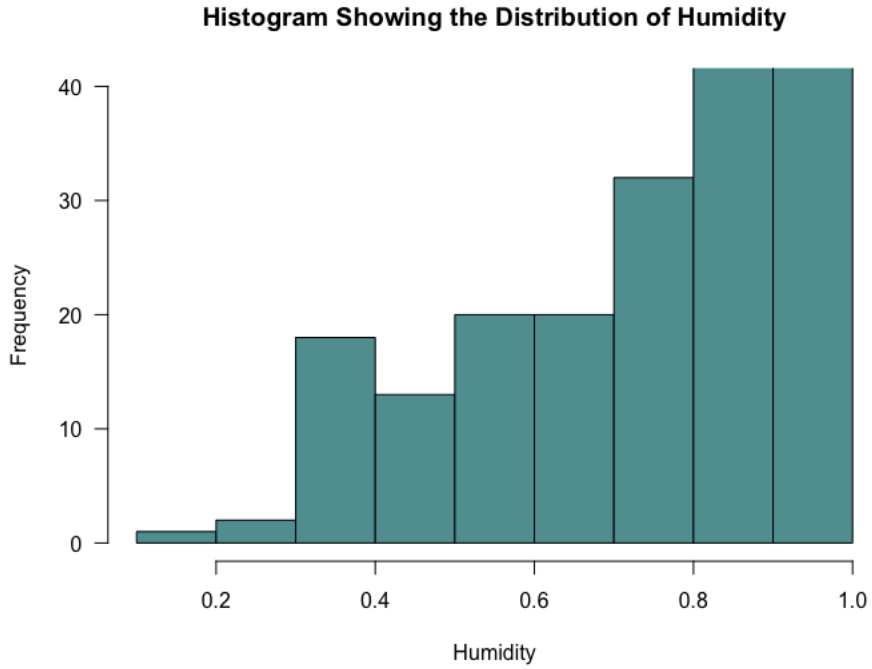
## Analysis of Humidity Variable

The "humidity" variable has a quantitative (continuous) data type. It is a measurement of the amount of water vapour in the air. It is expressed as a proportion. The below R output

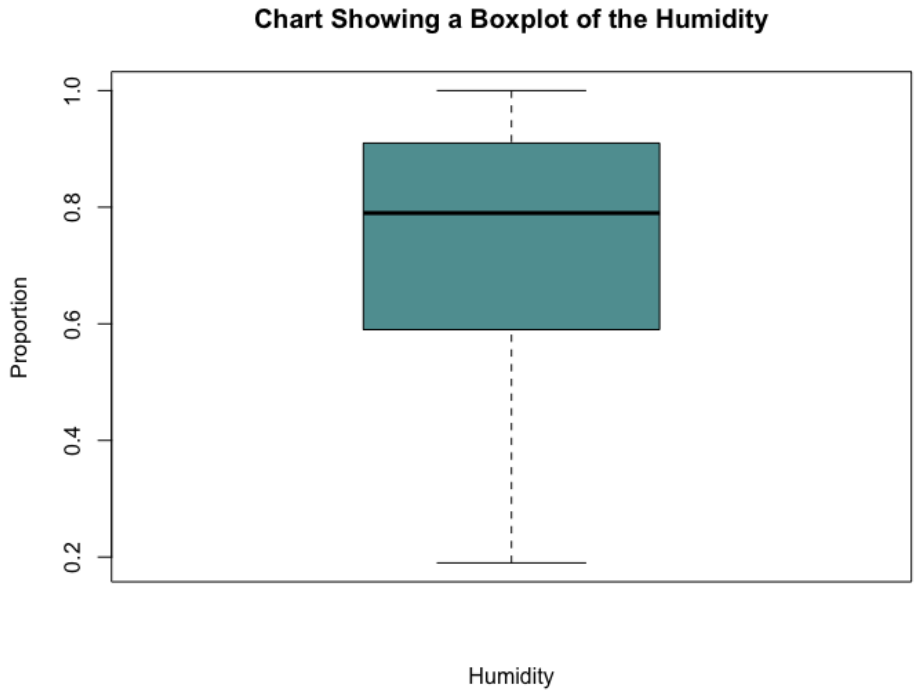
>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| 0.19 | 0.79 | 0.7358 | 0.202699 | 1.0 |

The above output shows that the lowest humidity observed in the variable is 0.19 while the highest is 1.0. The mean and median humidity observed are 0.7358 and 0.79 respectively while the standard deviation from the mean distribution is 0.202699.




The histogram above shows that the distribution is not a normal distribution as the plot appears to be a bit negatively skewed.




The chart shows a boxplot of the Humidity variable in the dataset, having specified the mean, median and the quartile ranges, there are no outliers in the humidity variable that

can potentially affect any further analysis. Additionally, no missing values were observed in the distribution.

## Analysis of Wind Speed Variable

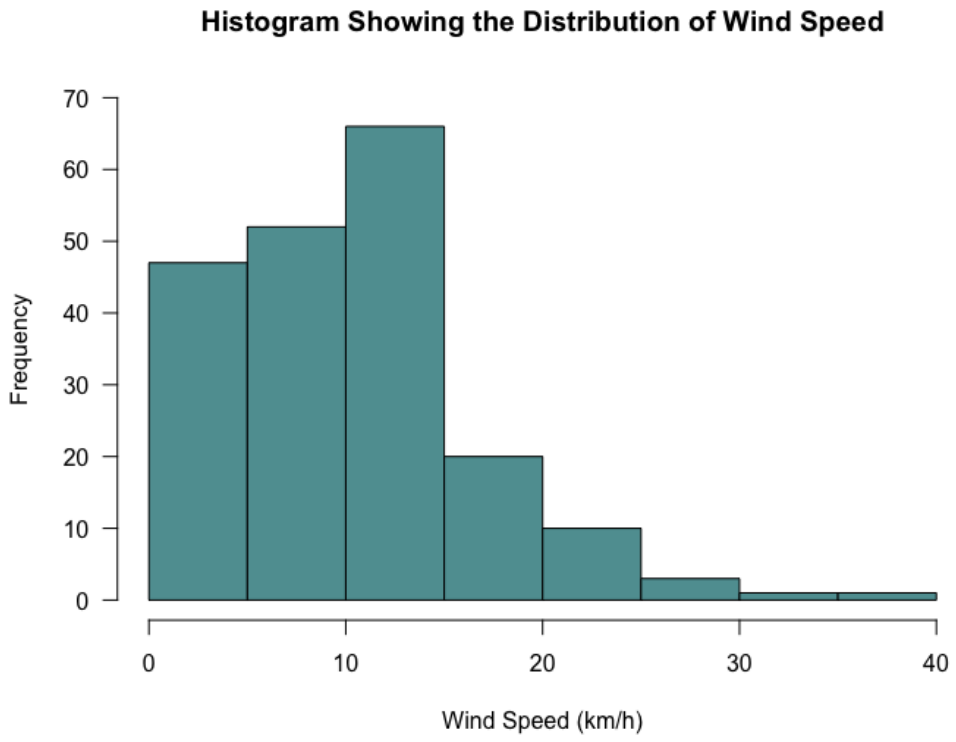
The "wind speed" variable is a quantitative (continuous) data type. This variable describes how fast the wind is moving past a certain point. It is measured in kilometre per hour (km/h). The below R output

>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| 0.00 | 10.26 | 10.36 | 6.136214 | 36.77 |

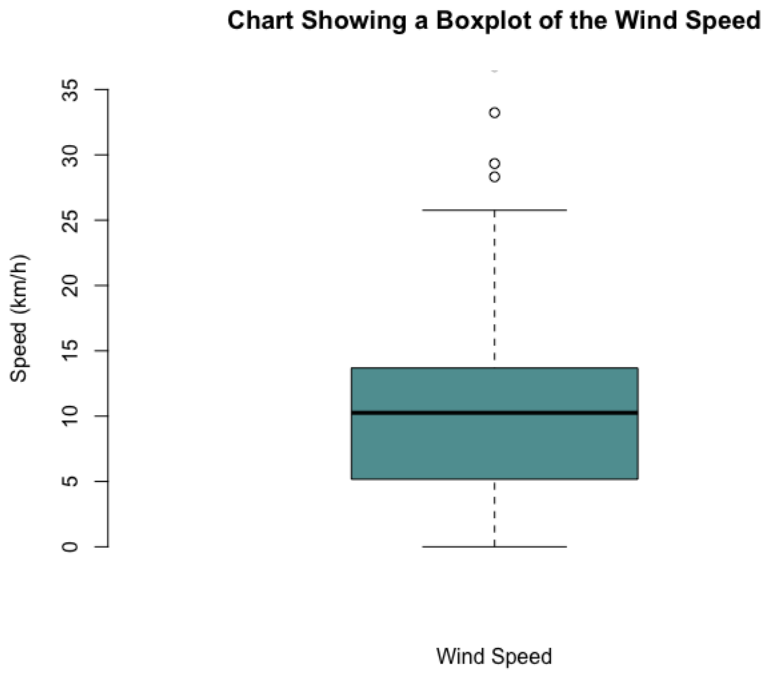
The above output shows that the lowest wind speed observed in the variable is 0.00km/h while the highest is 36.77km/h. The mean and median observed are 10.36km/h and 10.26km/h respectively while the standard deviation from the mean distribution is 6.136214km/h.

The chart below shows the distribution in the observations:




From the histogram above, it indicates that the observations in the wind speed variable are not normally distributed, the distribution appears to be positively skewed. Further

investigation into the variable as seen in the boxplot below indicates that there are outliers however, the values are probable and not extreme, therefore will be left untouched as it shouldn't impact further analysis.



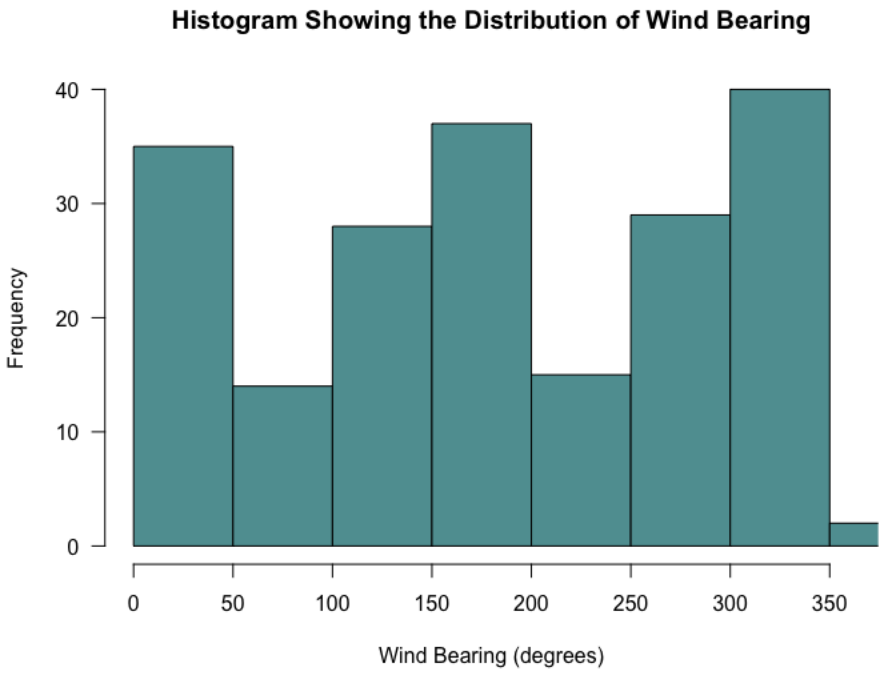

## Analysis of Wind Bearing Variable

The "wind bearing" variable is a quantitative (continuous) data type. This variable indicates the direction from which the wind is blowing. It is measured in degrees (°). The below R output

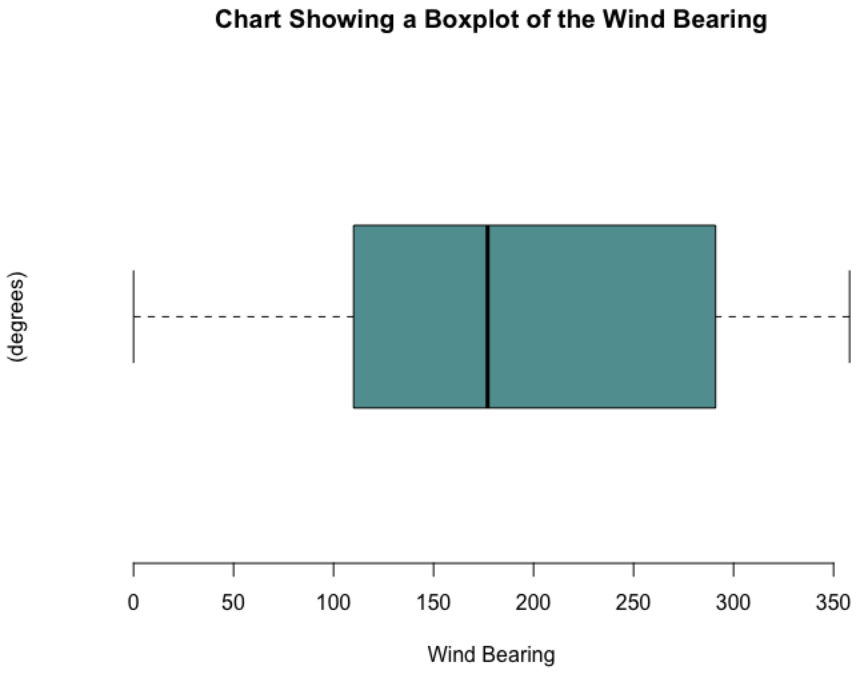
>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| 0.0 | 177.0 | 186.2 | 108.9026 | 358.0 |

The above output shows that the lowest wind bearing observed in the variable is 0.0° while the highest is 358.0°. The mean and median observed are 186.2° and 177.0° respectively while the dispersion from the mean (standard deviation) is 108.9026°.




The histogram shows that the distribution in the wind bearing variable is not a normal distribution and no skew in the distribution is observed from the plot.




The boxplot above shows there are no outliers in the distribution of the wind bearing variable. The zero values observed are probable due to the nature of the variable.

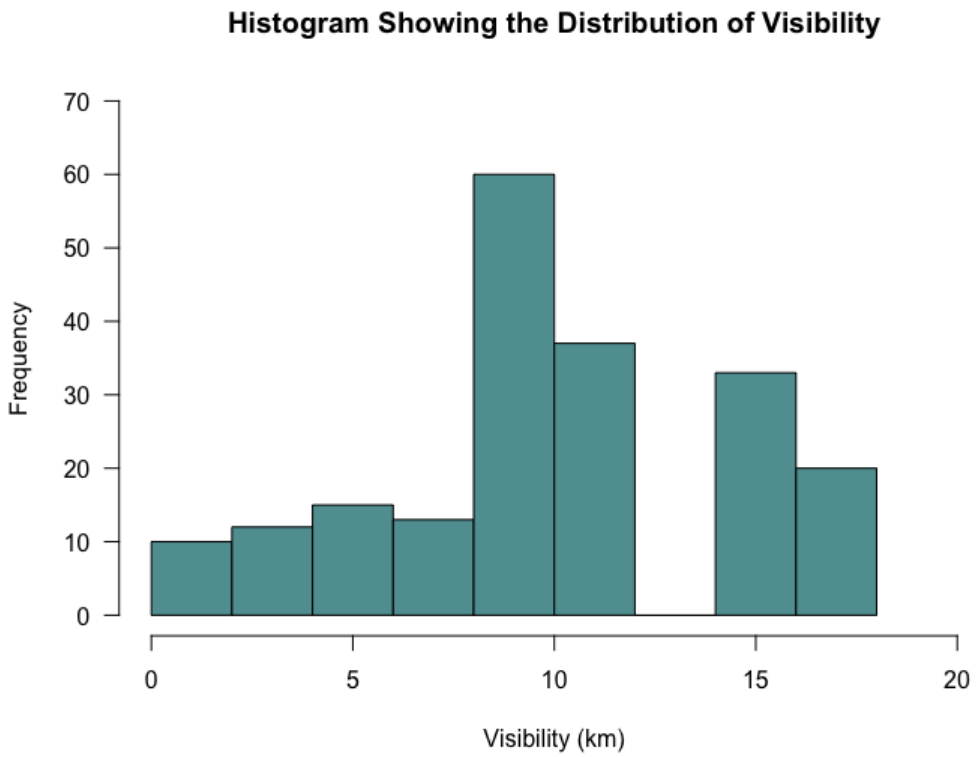
## Analysis of Visibility Variable

The "visibility" variable is a quantitative (continuous) data type. This variable shows a measurement of how far away an object may be seen clearly. It is measured in kilometres (km). The below R output

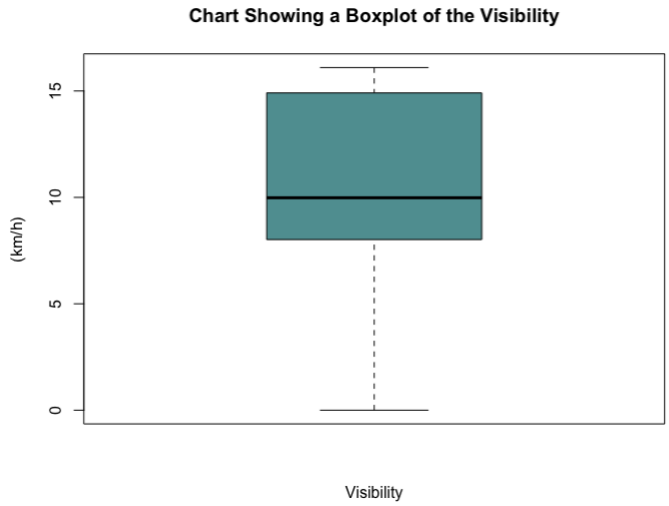
>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| 0.000 | 9.982 | 10.131 | 4.339896 | 16.100 |

The above output shows that the lowest visibility observed in the variable is 0.000km while the highest is 16.1km. The mean and median observed are 10.131km and 9.982km respectively while the standard deviation is 4.339896km.




The histogram shows that the distribution in the visibility variable is not a normal distribution.




Furthermore, the boxplot reveals no outliers, and the zero values recorded are thought to be likely owing to the nature of the variable, as it is conceivable to see nothing when the weather is severely foggy.

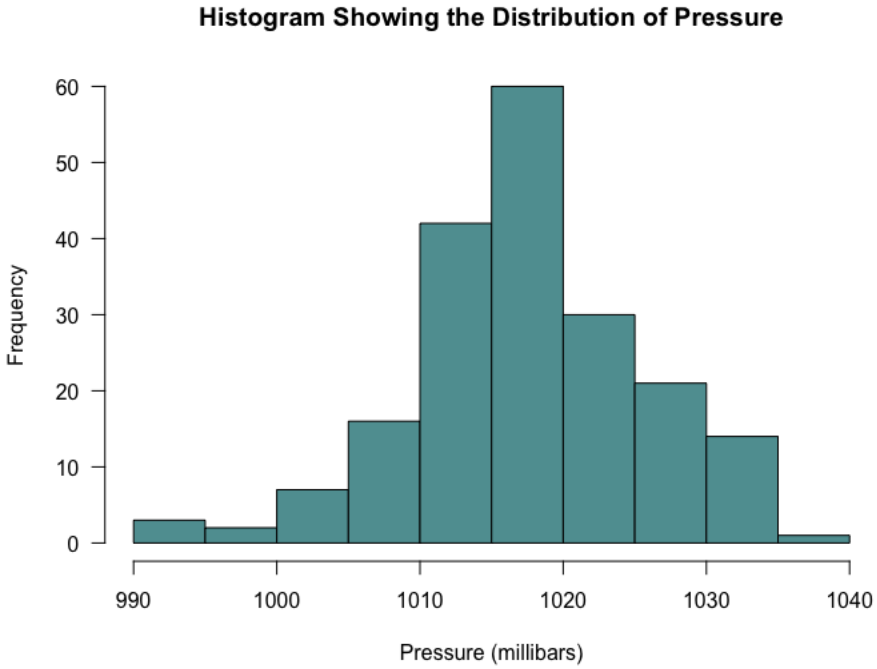
## Analysis of Pressure Variable

The "pressure" variable is a quantitative (continuous) data type. This simply describes the weight of the air, and it is measured in millibars. The below R output

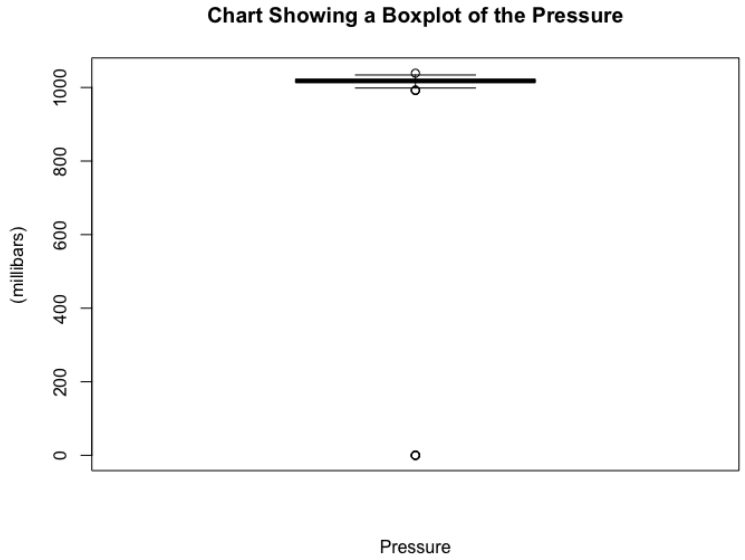
>

| Min. | Median | Mean | SD. | Max. |
|---|---|---|---|---|
| 0.0 | 1017.2 | 997.2 | 143.0436 | 1038.8 |

The above output shows that the lowest pressure observed in the variable is 0.00millibars while the highest is 1038.8millibars. The mean and median observed are 997.2millibars and 1017.2millibars respectively while the standard deviation from the mean is 143.0436millibars. The chart below shows a histogram of the distribution:



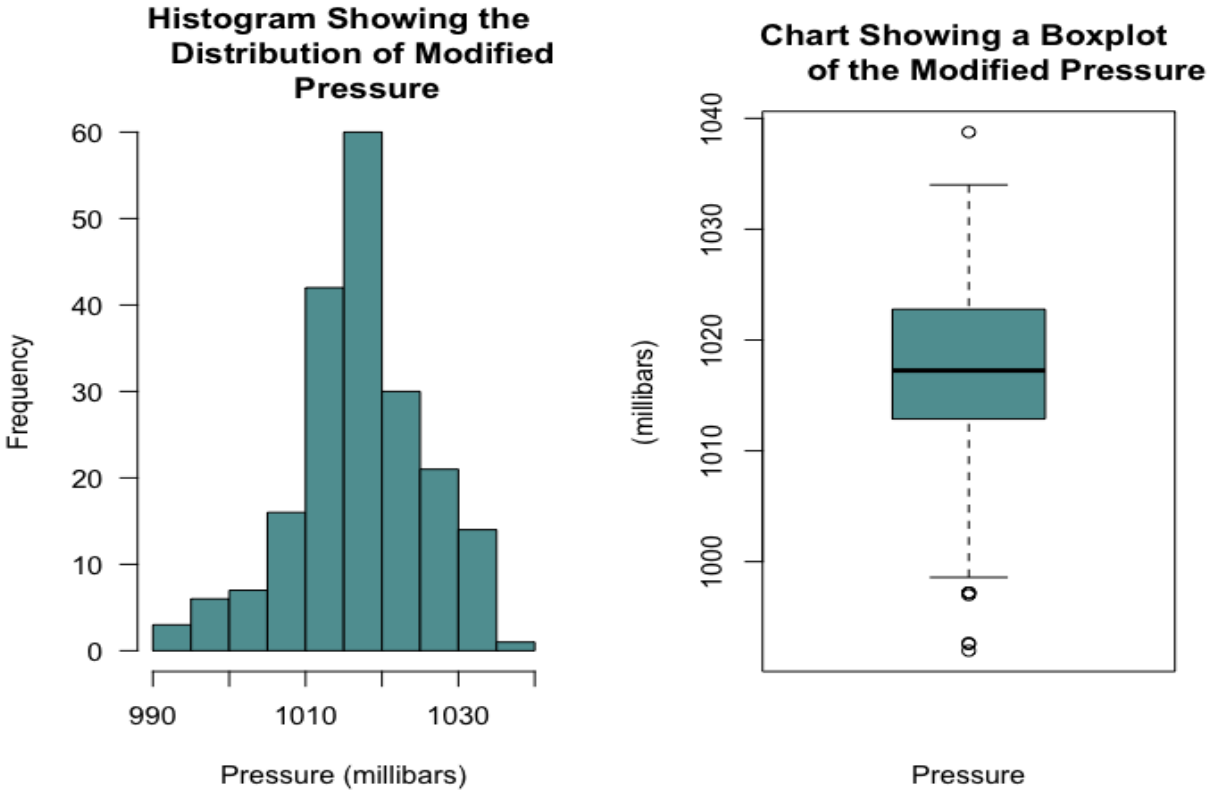

The above chart shows the distribution is a normal distribution as a bell curve can be seen and it somewhat tails on both ends of the plot.




The boxplot above shows extreme outliers in the dataset, upon further investigation of the observations, there appears to be a couple of zero values in the variables which are highly improbable as zero pressure only exists in a perfect vacuum as such the values are deemed as missing values. Consequently, such values have been replaced with the

mean of the distribution (997.2) so as not to inflate the significance of the distribution or the outliers (if any) in the variable.

Below is the histogram and boxplot of the modified variable in the modified dataset. The outliers in the boxplot are probable and acceptable therefore won't impact any further analysis. Additionally, the histogram still shows a normal distribution after replacing the zero values with the mean of the distribution however, a slight change in the tails of the plot was observed.



## Modification of the “Summary” Variable

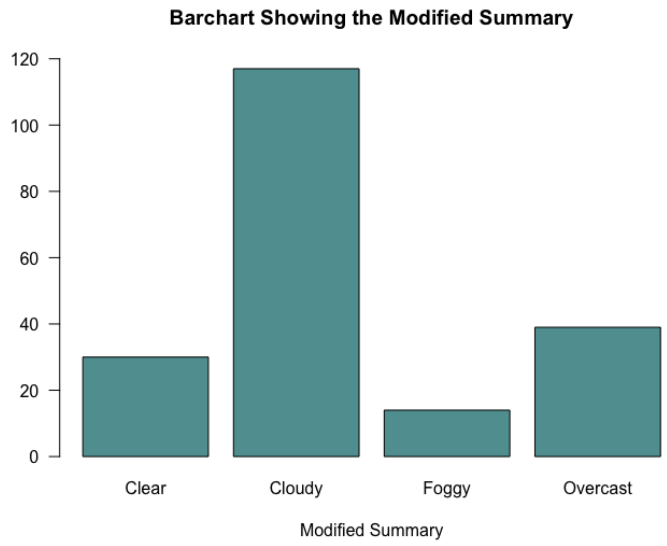
The summary variable's initial data analysis reveals multiple categories that overlap (such as breezy and mostly cloudy, breezy and overcast, mostly cloudy, overcast etc.) The "summary variable" currently has seven (7) categories which needs to be reduced for the purpose of simplification during subsequent analysis.

The R output shows a summary statistic of the modified “summary” variable:

>

| Clear | Cloudy | Foggy | Overcast |
|---|---|---|---|
| 30 | 117 | 14 | 39 |

Below is the graphical representation of the modified summary variable in the dataset:




The categories in the “summary variable” have been reduced from seven categories to four categories. They are:

- Clear
- Cloudy
- Foggy
- Overcast

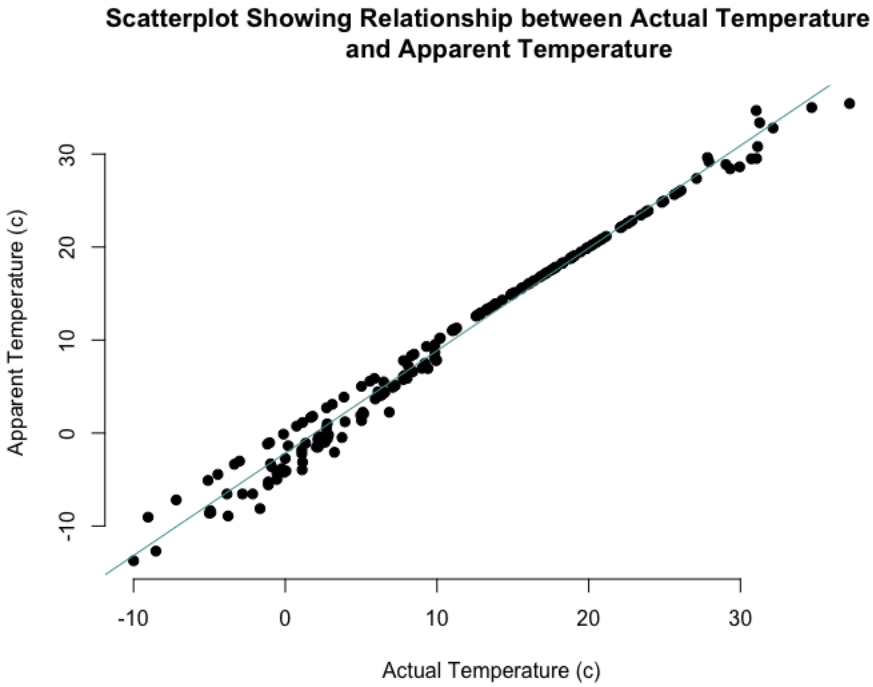
Subsequently, this modified summary variable will be used for any further analysis of the dataset.

# Pairwise Relationships Among Variables

In the dataset, there are two (2) main variables of interest, actual temperature, and apparent temperature. The section below aims to investigate relationships between the variables of interest and other variables in the data set. To investigate the relationships among variables we will use scatterplots.

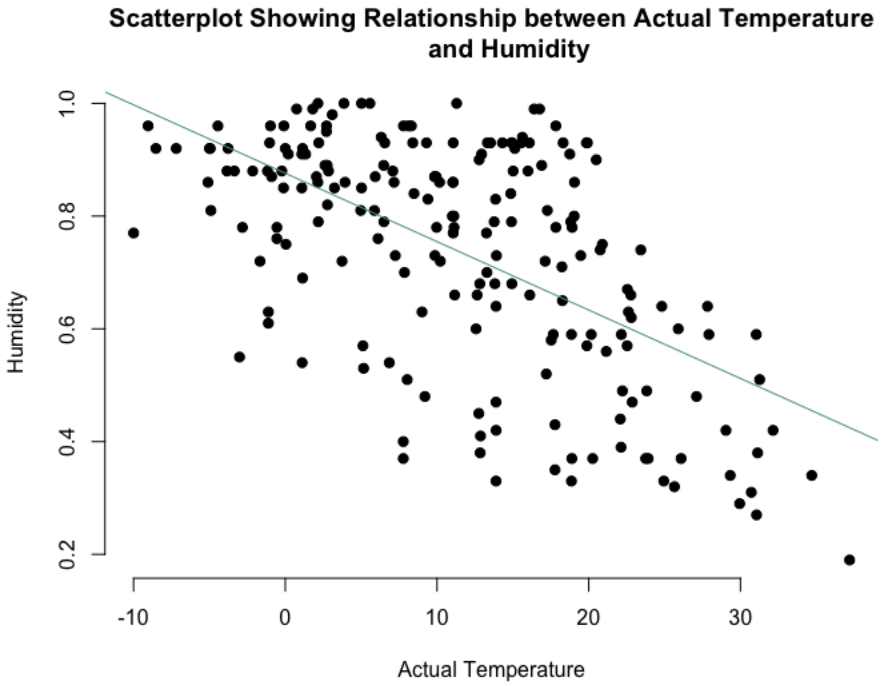
## Investigating Variables Related to Actual Temperature

### Relationship Between Actual Temperature and Apparent Temperature



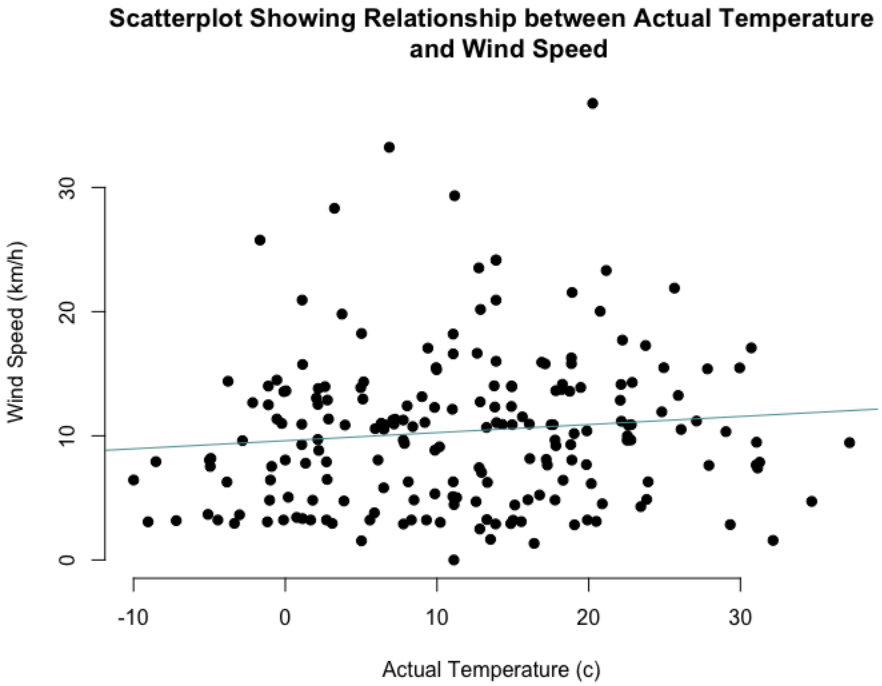

The scatter plot indicates a strong positive relationship between the actual temperature and the apparent temperature.

## Relationship Between Actual Temperature and Humidity



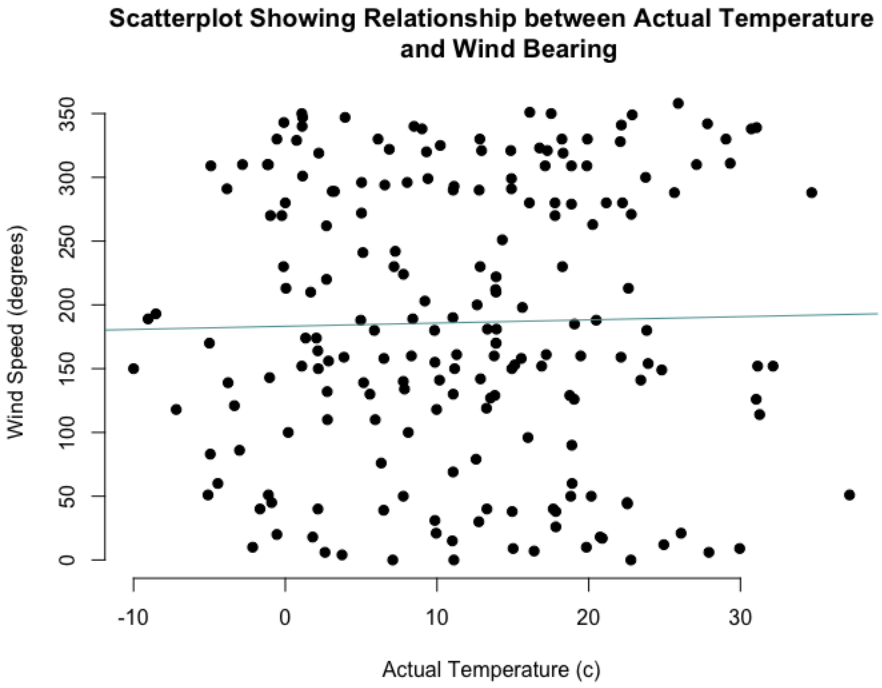

The scatter plot indicates a somewhat weak negative relationship between the actual temperature and the humidity.

## Relationship Between Actual Temperature and Wind Speed



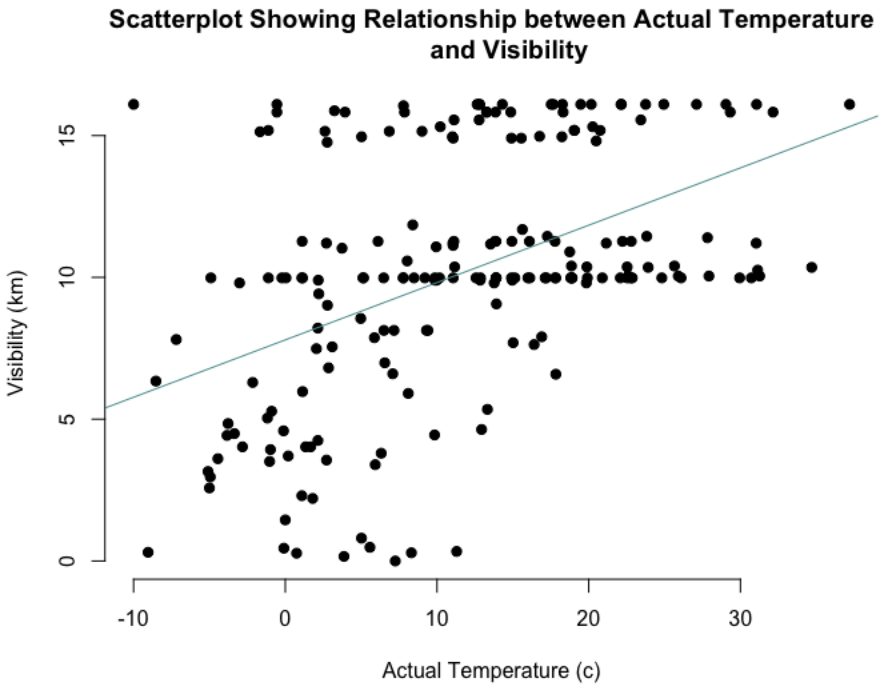

The scatter plot indicates that there is no relationship between the actual temperature and the wind speed.

## Relationship Between Actual Temperature and Wind Bearing



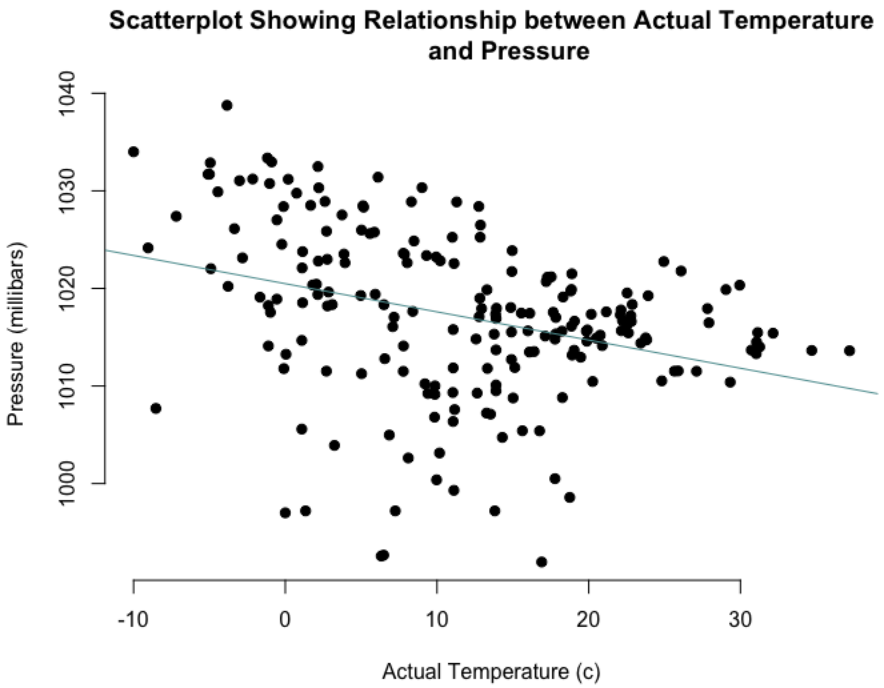

The scatter plot indicates that there is no relationship between the actual temperature and the wind bearing.

## Relationship Between Actual Temperature and Visibility



The scatter plot indicates that there may be a weak positive relationship between the actual temperature and the visibility.

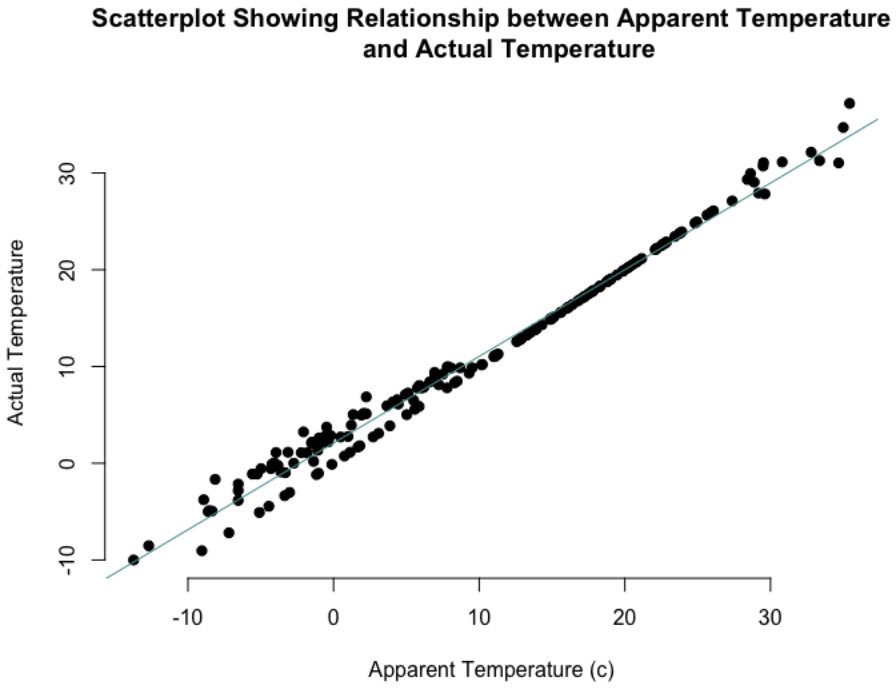
### Relationship Between Actual Temperature and Pressure




The scatter plot indicates that there is a moderately weak negative relationship between the actual temperature and the pressure (modified pressure as outlined in the Initial Data Analysis section).

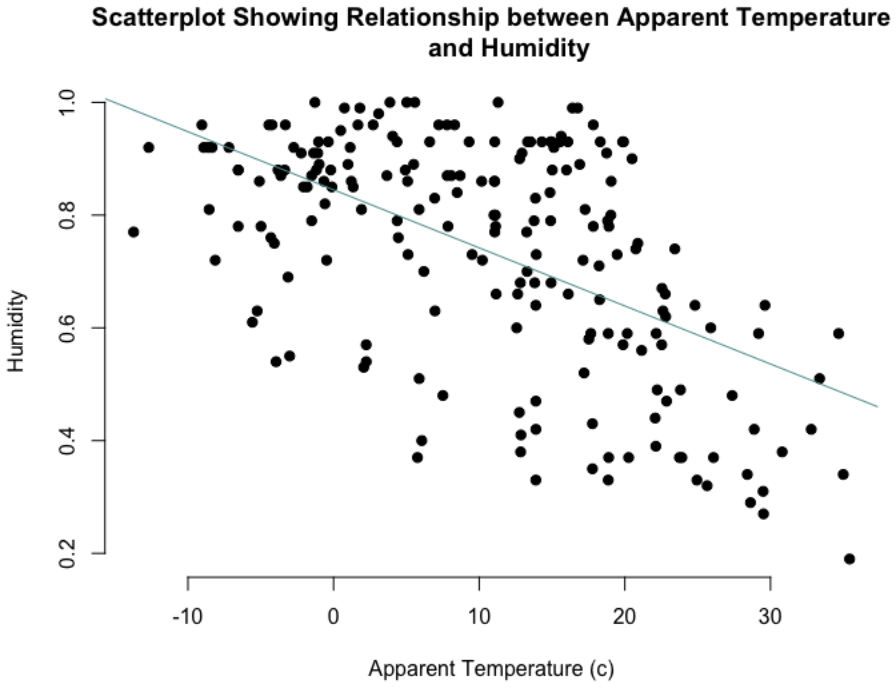
### Investigating Variables Related to Apparent Temperature

## Relationship Between Apparent Temperature and Actual Temperature



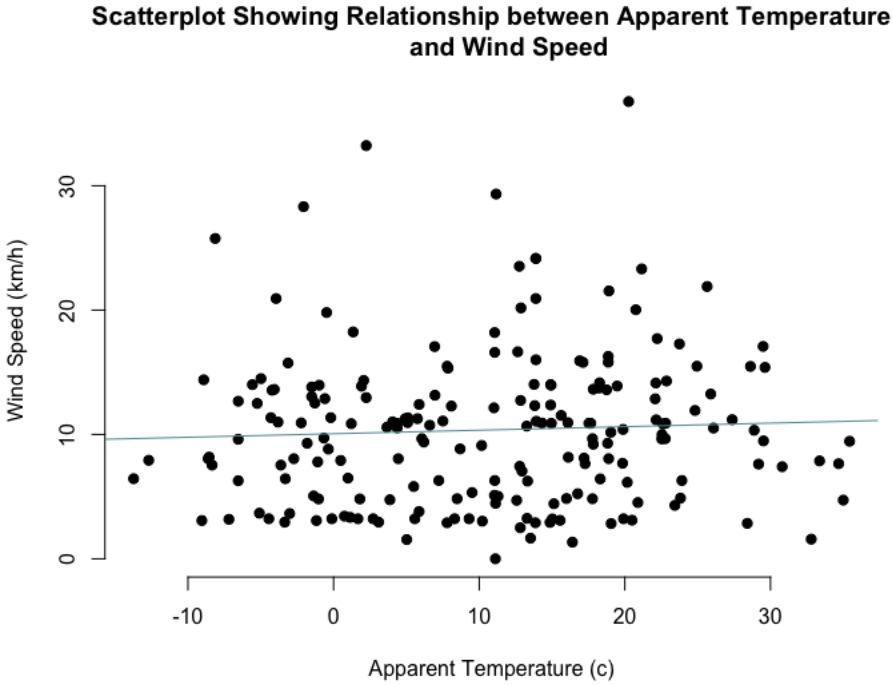

The above scatter plots indicate a strong positive relationship between apparent temperature and actual temperature.

## Relationship Between Apparent Temperature and Humidity



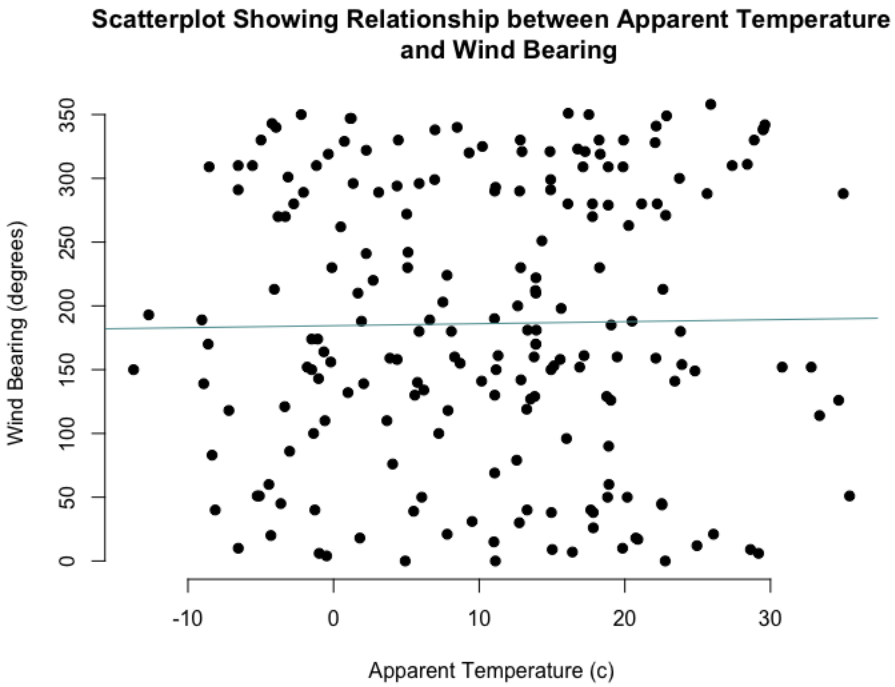
The scatter plot indicates that there is a moderate negative relationship between the apparent temperature and the humidity.

## Relationship Between Apparent Temperature and Wind Speed



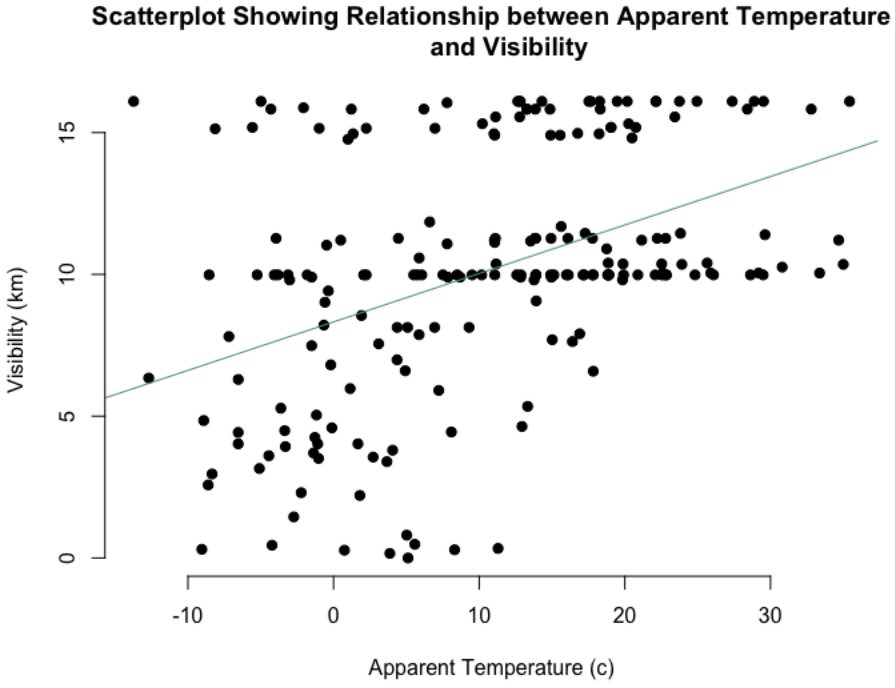

The scatter plot indicates that there is no relationship between the apparent temperature and the wind speed.

## Relationship Between Apparent Temperature and Wind Bearing



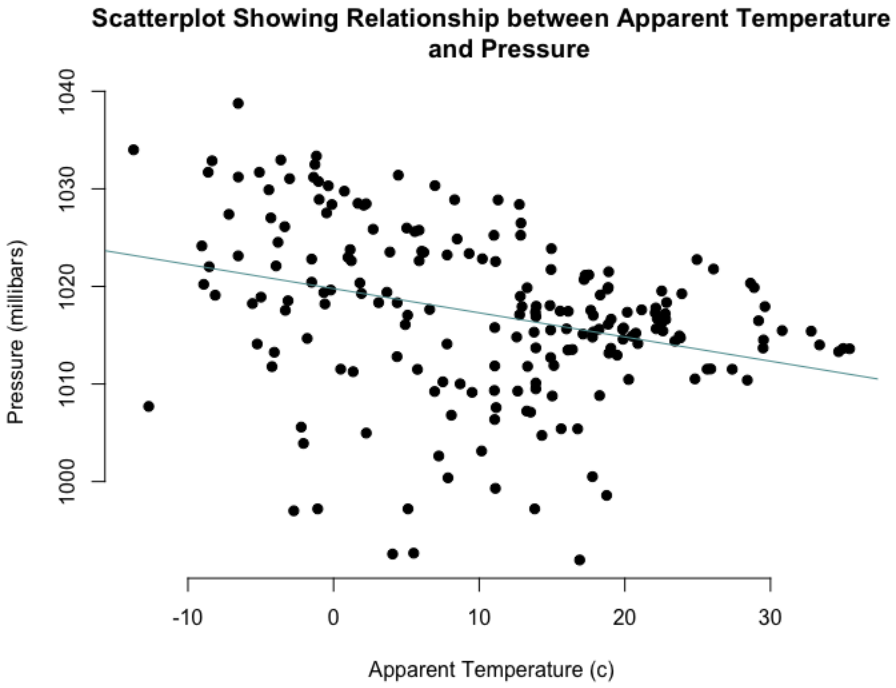

The scatter plot indicates that there is no relationship between the apparent temperature and the wind bearing.

## Relationship Between Apparent Temperature and Visibility




The scatter plot indicates a slightly weak positive relationship between the apparent temperature and the visibility.

### Relationship Between Apparent Temperature and Pressure




The scatter plot indicates that there is a weak negative relationship between the apparent temperature and the pressure variable (modified pressure as outlined in the Initial Data Analysis section).

## Investigating Other Variables

### Investigating One Categorical Variable and One Continuous Variable

Examining the modified dataset, we will consider the mean of the variables of interest (the actual temperature and the apparent temperature) which are continuous variables against the modified summary variable (which is a categorical variable).

>

| Clear | Cloudy | Foggy | Overcast |
|---|---|---|---|
| 10.1225926 | 14.5363723 | 0.9666667 | 7.4967236 |

The above shows the mean actual temperature for each of the categories in the modified summary variable. The mean actual temperature for a “Clear” weather is 10.1225926°C,

for a “Cloudy” weather is 14.5363723°C, 0.9666667°C for a “Foggy” weather and 7.4967236°C for an “Overcast” weather.

>

| Clear | Cloudy | Foggy | Overcast |
|---|---|---|---|
| 9.1090741 | 13.7823837 | -0.2611111 | 5.9628205 |

The above shows the mean apparent temperature for each of the categories in the modified summary variable. The mean apparent temperature for a “Clear” weather is 9.1090741°C, for a “Cloudy” weather is 13.7823837°C, -0.2611111°C for a “Foggy” weather and 5.9628205°C for an “Overcast” weather.

# Hypothesis Testing

In this section, inferences from the dataset primarily on the actual temperature and apparent temperature will be made.

## One Sample Hypothesis Test

To test the difference between matched pairs. In this case, the actual temperature variable as it is a variable of interest it would be interesting to find out if the population mean is above or below the sample mean. In the Initial Data Analysis section, we established that the sample mean of the actual temperature is 11.552°C. Therefore, we are testing a value that is above the sample mean (14°C).

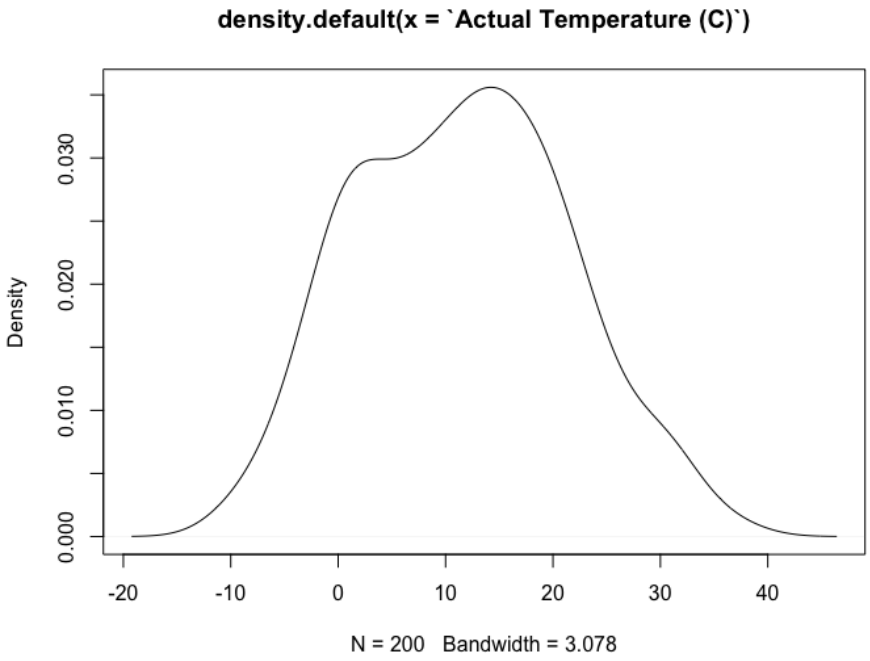
The two competing hypotheses are the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$).

($H_0$): The average actual temperature from 2006 to 2016 is below 14°C

($H_1$): The average actual temperature from 2006 to 2016 is above 14°C

To ascertain whether the mean actual temperature from 2006 to 2016 is 14°C at a 95% confidence level (5% significance level) we check for normality in the distribution. If the distribution is normal, we carry out a parametric test otherwise, a non-parametric test will be used.

From the Initial Data Analysis section, exploration of the actual temperature, the histogram shows a normal distribution however, further test for normality is required to determine whether a parametric or non-parametric test will be used for the hypothesis testing.




The chart above shows a "bell curve" of the distribution in the actual temperature which indicates normality. Additionally, to further test for normality, the Shapiro test could be deployed.

Shapiro Wilk Normality Test:

($H_0$): There is normality in the distribution

($H_1$): There is no normality in the distribution

The R output below shows the results for the normality distribution

>

Shapiro-Wilk normality test

data: Weather_Data_DATA_SET_44_Modified$`Actual Temperature (C)`

W = 0.98889, p-value = 0.1223

The above results indicate a p-value of 0.1223 (at a 5% significance level) therefore we accept $H_0$ since the p-value is greater than the level of significance.

The findings from normality test reinforce the findings from the initial data analysis and consequently, a parametric test will be used for the hypothesis testing.

## Parametric Test

As the findings from the normality test indicates a normal distribution, a parametric test for a mean on the actual temperature will be tested.

To determine whether there is sufficient evidence available from the data set to conclude that the average actual temperature from 2006 to 2016 is above 14°C at a 95% confidence interval.

($H_0$): The mean actual temperature from 2006 to 2016 is below 14°C

($H_1$): The mean actual temperature from 2006 to 2016 is above 14°C

$H_0: \mu \leq 14$ $H_1: \mu > 14$ (one-tailed)

Below is the R output of the test

>

One Sample t-test

data: Weather_Data_DATA_SET_44_Modified$`Actual Temperature (C)`
t = -3.5087, df = 199, p-value = 0.9997
alternative hypothesis: true mean is greater than 14
95 percent confidence interval:
10.39856 Inf
sample estimates:
mean of x
11.55169

The above output shows that the value of the test statistic t is -3.5087, the degrees of freedom df is 199 and a p-value of 0.9997.

## Conclusion of One Sample Hypothesis Test

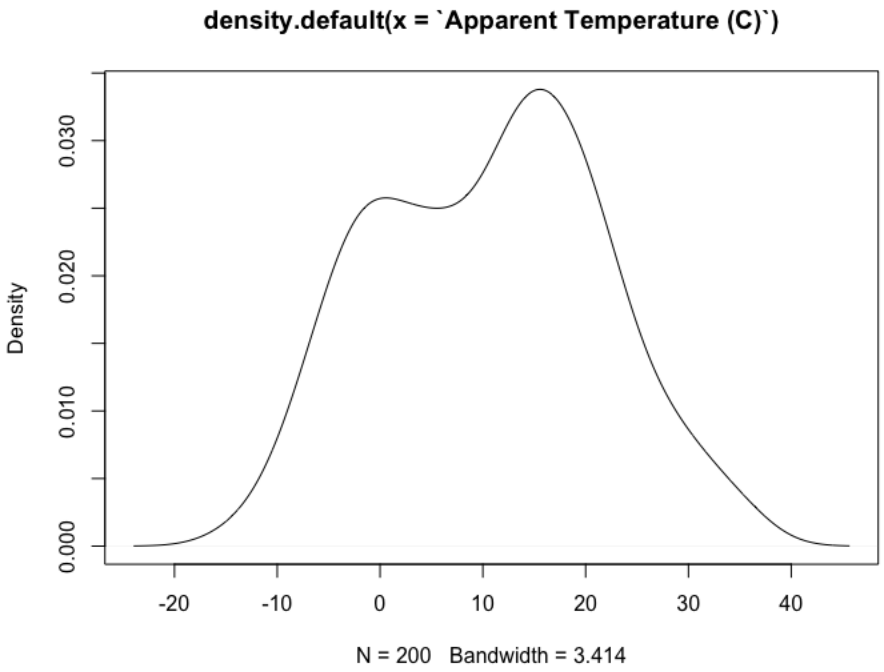
Based on the results of the one sample t-test, a p-value of 0.9997 which is greater than the significance level 5%, we can conclude that there isn't sufficient evidence to reject the null hypothesis at a 5% level of significance, in fact, there is strong evidence to accept the null hypothesis as the p-value is very close to 1.

Therefore, the null hypothesis $H_0$ is accepted and the alternative hypothesis $H_1$ is rejected which means that based on sample data, the population mean actual temperature from 2006 to 2016 is below 14°C.

## Two Sample Hypothesis Test

To test the difference between two population means. In this case, the apparent temperature of the "cloudy" weather and "overcast" weather. Interestingly, from the Modification of the Summary Variable section, these two groups (amongst cloudy, clear, foggy, and overcast) are the groups with new observations as an outcome of adjusting the summary variables hence, the test to ascertain if there is a difference in the mean between the two population groups.

The two competing hypotheses are the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$). We test to check whether the average apparent temperature of when it is cloudy is equal to the average apparent temperature of when it is overcast.




The above plot shows the distribution of the apparent temperature, and it appears to be a normal distribution as it tails on both ends and peak somewhat in the middle.

### Shapiro Wilk Normality Test:

Further test for normality is required to determine whether a parametric or non-parametric test will be used for the hypothesis testing.

($H_0$): There is normality in the distribution

($H_1$): There is no normality in the distribution

The R output below shows the results for the normality distribution of the apparent temperature for the “cloudy” group

>

```
Shapiro-Wilk normality test

data:  cloudy_df$`Apparent Temperature (C)`
W = 0.98181, p-value = 0.1141
```

The above results indicate a p-value of 0.1141 (at a 5% significance level) therefore we accept $H_0$ since the p-value is greater than the level of significance. Therefore, there is normality in the cloudy group observations.

The R output below shows result for the normality distribution of the apparent temperature for the “overcast” category

>

```
Shapiro-Wilk normality test
data:  overcast_df$`Apparent Temperature (C)`
W = 0.96472, p-value = 0.2553
```

The above results indicate a p-value of 0.2553 (at a 5% significance level) therefore we accept $H_0$ since the p-value is greater than the level of significance. Therefore, there is normality in the overcast group observations.

The findings from Shapiro tests indicates there is normality in the distribution of the two groups (cloudy and overcast) therefore, a parametric test for if there is a difference in the mean apparent temperature between “cloudy” and “overcast”.

### Parametric Test

To determine whether the mean apparent temperature of a "cloudy" weather is different from the mean apparent temperature of an "overcast" weather at a 95% confidence interval.

($H_0$): There is no difference between the mean apparent temperature of a cloudy weather and an overcast weather

($H_1$): There is a difference between the mean apparent temperature of a cloudy weather and an overcast weather

$H_0$: $\mu_{cloudy} = \mu_{overcast}$

$H_1$: $\mu_{cloudy} \neq \mu_{overcast}$ (two-tailed)

### *Bartlett Test of Homogeneity of Variances*

To ascertain whether we carry out the test under the assumption of equal variance we perform a Bartlett Test of Homogeneity of Variances.

$H_0$: $\sigma^2_{cloudy} = \sigma^2_{overcast}$

$H_1$: $\sigma^2_{cloudy} \neq \sigma^2_{overcast}$

Below is the R output of the test of homogeneity of variance

>

Bartlett test of homogeneity of variances

data: Weather_Data_DATA_SET_44_Modified$`Apparent Temperature (C)` by Weather_Data_DATA_SET_44_Modified$`Modified Summary`

Bartlett's K-squared = 6.1788, df = 3, p-value = 0.1032

The p-value is 0.1032, therefore the null hypothesis is accepted at a 5% significance level and can use the "equal variances assumed" version of the two-sample t-test.

Below is the R output of the two-sample t-test (equal variance assumed)

>

Two Sample t-test

data: cloudy_df$`Apparent Temperature (C)` and overcast_df$`Apparent Temperature (C)`

```
t = 4.1881, df = 154, p-value = 4.719e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.131124 11.508002
sample estimates:
mean of x mean of y
13.782384 5.962821
```

The above output shows a p-value of 4.710e-05 (0.00004719) with a test statistic t of 4.1881 and degrees of freedom df is 154

## Conclusion of Two Sample Hypothesis Test

Based on the results of the two-sample t-test, a p-value of 0.00004719 which is way less than the significance level 5%, we can conclude that there is sufficient evidence to reject the null hypothesis at a 5% level of significance.

Therefore, the null hypothesis $H_0$ is rejected and the alternative hypothesis $H_1$ is accepted which means that based on sample data, there is a difference in the mean apparent temperature of “cloudy” and “overcast” weather groups.