Project 2

Shauna Le, Aaron Chien, Yoojin Min, Doris Wu

2022-11-04

Introduction

Hello, in this project by Shauna Le, Aaron Chien, Yoojin Min, and Doris Wu, we will be exploring the Climate Change in Delhi. We would be trying to model the temperature in Delhi and trying to predict climate change. Weather data was collected in the city of Delhi from the period of 4 years (from 2013 to 2017). Our data can be found here: **Daily Climate time series data (https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data)**. Our variables are *date* in the MM/DD/YYYY format, *meantemp* for the average temperature of that day from three hour intervals in Celsius, *humidity* is the humidity levels (units are grams of water vapor per cubic meter volume of air), *wind_speed* is the wind speed measured in kmph, and *meanpressure* is the average pressure in atm. Here is a glance at our data:

```
climate_change <- read.csv("DailyDelhiClimateTrain.csv")
climate_change <- as.data.frame(ts(climate_change))
head(climate_change)
```

```
##   date  meantemp humidity wind_speed meanpressure
## 1    1 10.000000 84.50000   0.000000     1015.667
## 2    2  7.400000 92.00000   2.980000     1017.800
## 3    3  7.166667 87.00000   4.633333     1018.667
## 4    4  8.666667 71.33333   1.233333     1017.167
## 5    5  6.000000 86.83333   3.700000     1016.500
## 6    6  7.000000 82.80000   1.480000     1018.000
```

1. Exploratory Data Analysis

a

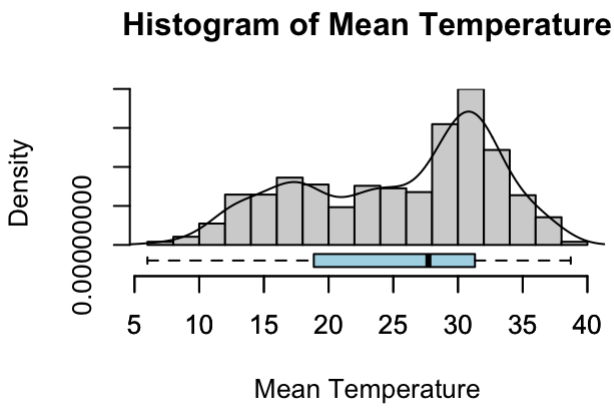
```
# seeing if there are any NA values
lapply(climate_change, function(x) {any(is.na(x))})
```

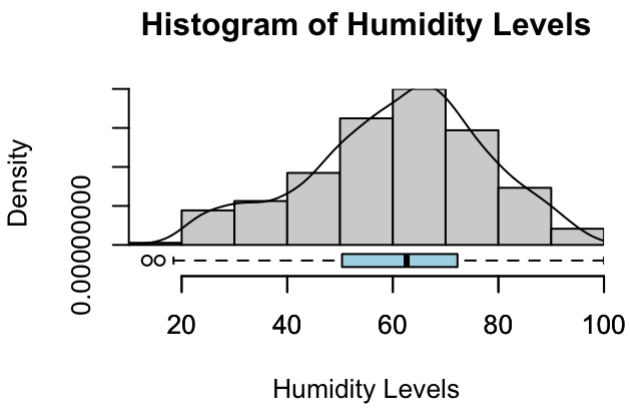
```
## $date
## [1] FALSE
##
## $meantemp
## [1] FALSE
##
## $humidity
## [1] FALSE
##
## $wind_speed
## [1] FALSE
##
## $meanpressure
## [1] FALSE
```

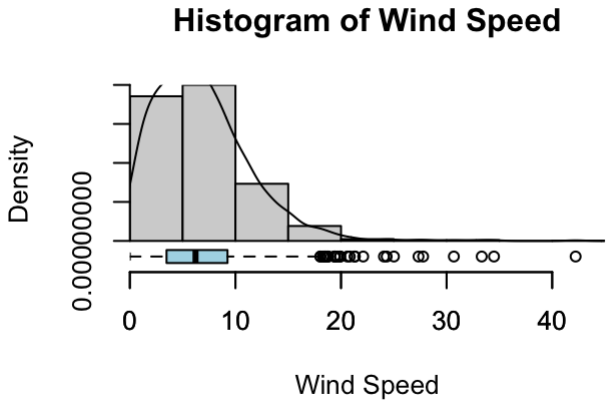
FALSE means that there are no NA values, since we used the function `any(is.na(x))` in which it is looking for any values that are NA and outputs TRUE if any values are NA and FALSE otherwise, since all the outputs are FALSE, we can assume they have no NA values.

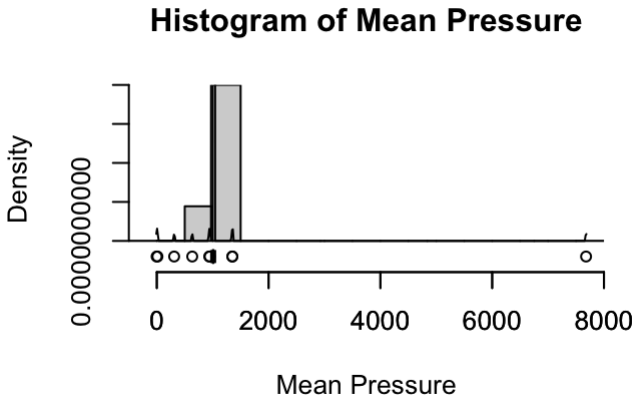
b

```
par(mfrow = c(2, 2))
hist_boxplot(climate_change$meantemp, freq = FALSE, main = "Histogram of Mean Tempe
rature",
             xlab = "Mean Temperature")
lines(density(climate_change$meantemp))

hist_boxplot(climate_change$humidity, freq = FALSE, main = "Histogram of Humidity L
evels",
             xlab = "Humidity Levels")
lines(density(climate_change$humidity))

hist_boxplot(climate_change$wind_speed, freq = FALSE, main = "Histogram of Wind Spe
ed",
             xlab = "Wind Speed")
lines(density(climate_change$wind_speed))

hist_boxplot(climate_change$meanpressure, freq = FALSE, main = "Histogram of Mean P
ressure",
             xlab = "Mean Pressure")
lines(density(climate_change$meanpressure))
```



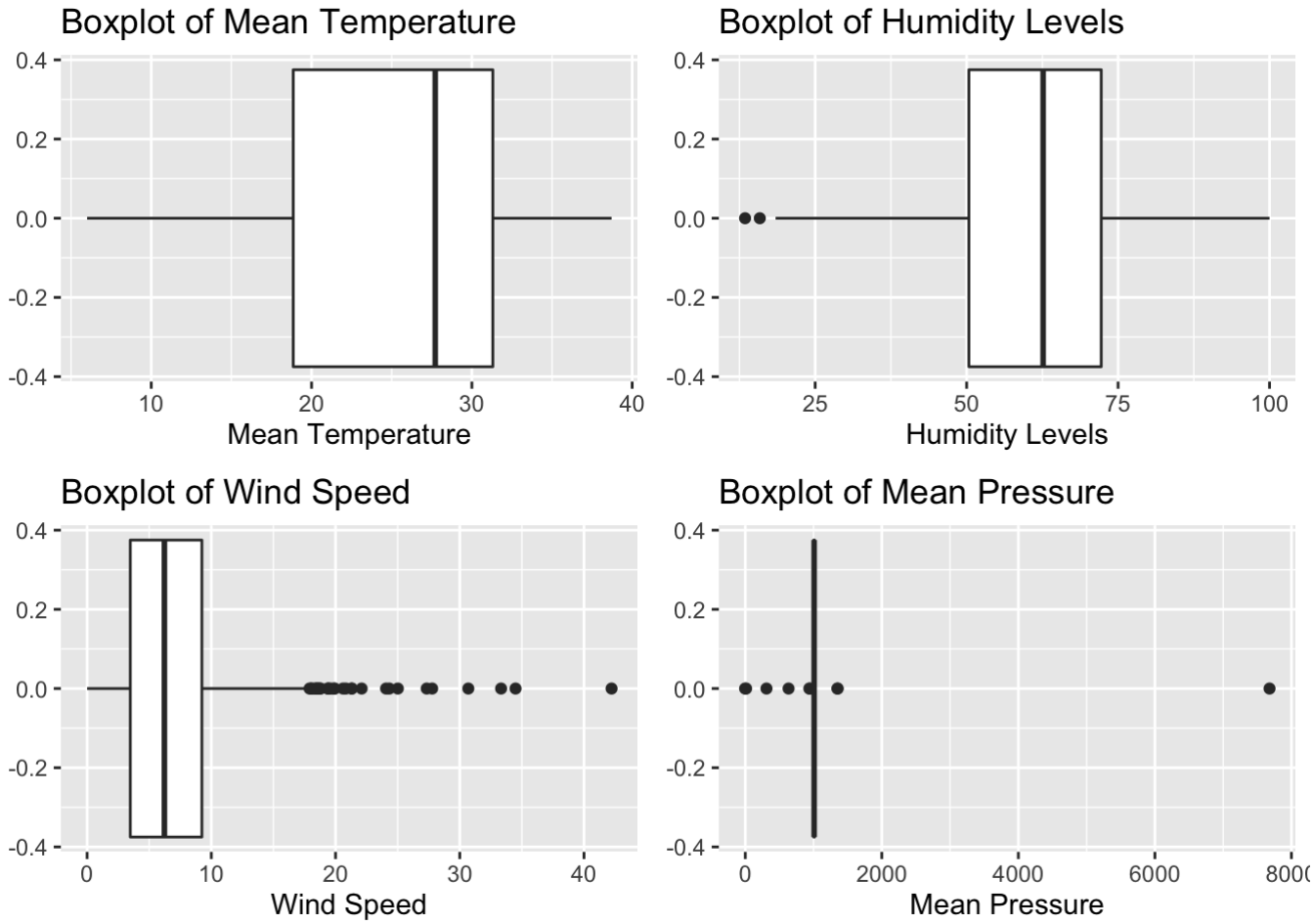






Here, we display a histogram and box plot for all of the variables as well as their fitted distribution line, we see that the mean temperature is very widely spread and peaking to the right, making it left skewed. The boxplot confirms the skewness of the data as the mean is toward the right and there are not outliers present. The humidity levels is somewhat centered but it does indeed have very fat tails. Looking at its boxplot, it is slightly skewed to the left and has some outliers. The wind speed is heavily skewed to the right and the boxplot shows many outliers in the data, this is problematic as this can interfere with the forecast of our data. Mean pressure is the most interesting as it is the most skewed and contains many outliers, this will be problematic in our forecast so we should remove outliers in the data

```
meantemp_bp <- climate_change %>%
  ggplot(mapping = aes(x = meantemp)) +
  geom_boxplot() +
  ggtitle("Boxplot of Mean Temperature") +
  labs(x = "Mean Temperature")

humidity_bp <- climate_change %>%
  ggplot(mapping = aes(x = humidity)) +
  geom_boxplot() +
  ggtitle("Boxplot of Humidity Levels") +
  labs(x = "Humidity Levels")

windspeed_bp <- climate_change %>%
  ggplot(mapping = aes(x = wind_speed)) +
  geom_boxplot() +
  ggtitle("Boxplot of Wind Speed") +
  labs(x = "Wind Speed")

meanpressure_bp <- climate_change %>%
  ggplot(mapping = aes(x = meanpressure)) +
  geom_boxplot() +
  ggtitle("Boxplot of Mean Pressure") +
  labs(x = "Mean Pressure")

plot_grid(meantemp_bp, humidity_bp, windspeed_bp, meanpressure_bp, ncol = 2, nrow =
2)
```



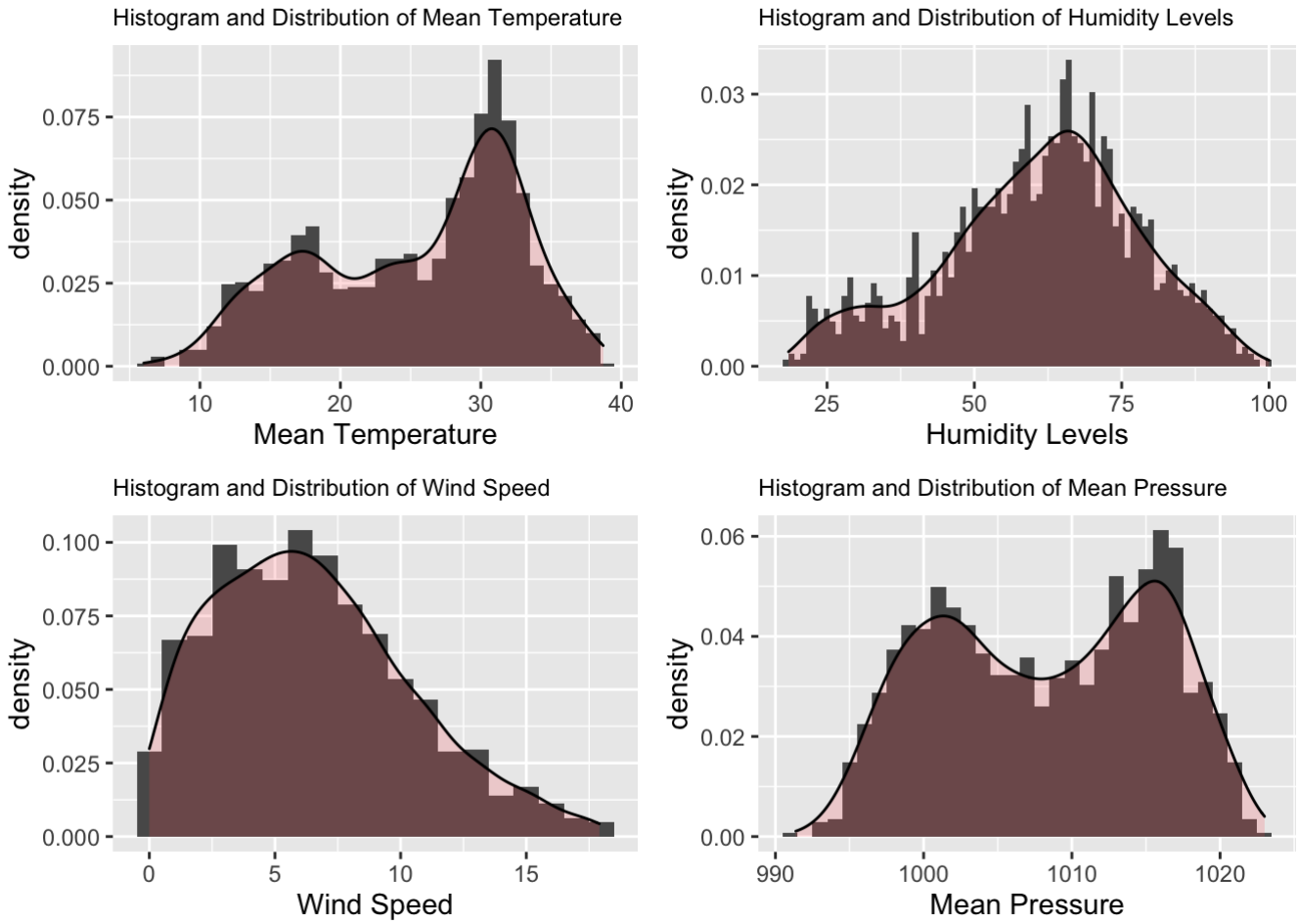
Here are the boxplots to give a closer look at the outliers, we can see the boxplot of mean pressure has many outliers and is heavily clustered, which is extremely alarming and using this data will cause our forecast to be wrong, so we should remove the outliers.

```
climate_change_f <- climate_change[
 -c(which(climate_change$meanpressure %in% boxplot.stats(climate_change$meanpressur
e)$out),
    which(climate_change$wind_speed %in% boxplot.stats(climate_change$wind_speed)$o
ut),
    which(climate_change$humidity %in% boxplot.stats(climate_change$humidity)$ou
t)), ]

nrow(climate_change_f)
```

```
## [1] 1422
```

After removing the outliers, we still have a large dataset to work with with $1,422$ observations, much larger than 30, so we can safely move on.

```
temp <- climate_change_f %>%
  ggplot(mapping = aes(x = meantemp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Mean Temperature") +
  labs(x = "Mean Temperature") +
  theme(plot.title = element_text(size = 9))

humid <- climate_change_f %>%
  ggplot(mapping = aes(x = humidity)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Humidity Levels") +
  labs(x = "Humidity Levels") +
  theme(plot.title = element_text(size = 9))

wind <- climate_change_f %>%
  ggplot(mapping = aes(x = wind_speed)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Wind Speed") +
  labs(x = "Wind Speed") +
  theme(plot.title = element_text(size = 9))

pressure <- climate_change_f %>%
  ggplot(mapping = aes(x = meanpressure)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Mean Pressure") +
  labs(x = "Mean Pressure") +
  theme(plot.title = element_text(size = 9))

plot_grid(temp, humid, wind, pressure, ncol = 2, nrow = 2)
```

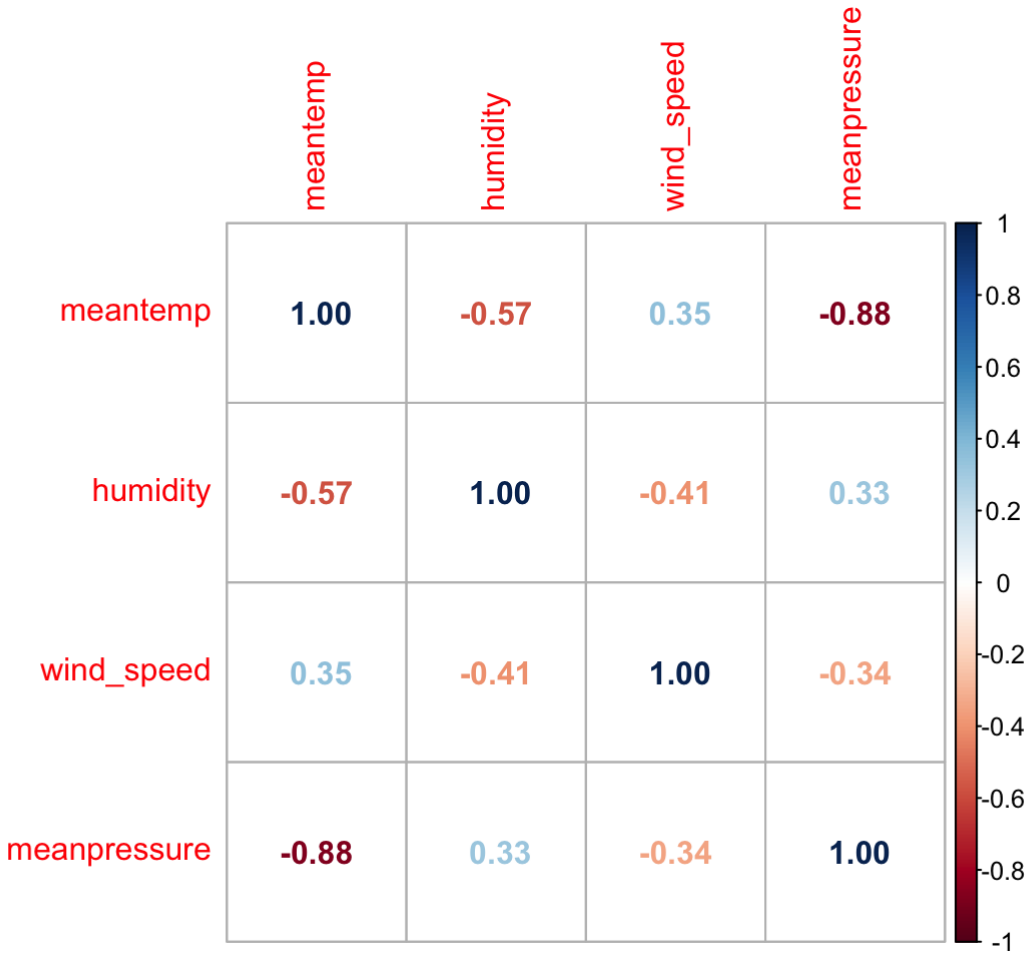


Plotting the histograms for our variables after removing the outliers, we can see that they have a better spread, but skewness is still an issue and mean pressure is now bimodal, meaning that it has two peaks so there are two points where the data is clustered. This can present different problems since we want our data to follow a normal distribution (with one peak and even spread) for better forecasting and so our standard errors can be more accurate. Because of this, we decided to differentiate the data as differencing the data helped make it unimodal and follow a normal distribution.

```
corr_mroz <- climate_change_f[, (names(climate_change_f) %in%
                                 c('meantemp', 'humidity', 'wind_speed', 'meanpressu
re'))]
res <- cor(corr_mroz)
round(res, 2)
```

```
##              meantemp humidity wind_speed meanpressure
## meantemp         1.00    -0.57       0.35        -0.88
## humidity        -0.57     1.00      -0.41         0.33
## wind_speed       0.35    -0.41       1.00        -0.34
## meanpressure    -0.88     0.33      -0.34         1.00
```

```
corrplot(res, method = 'number')
```



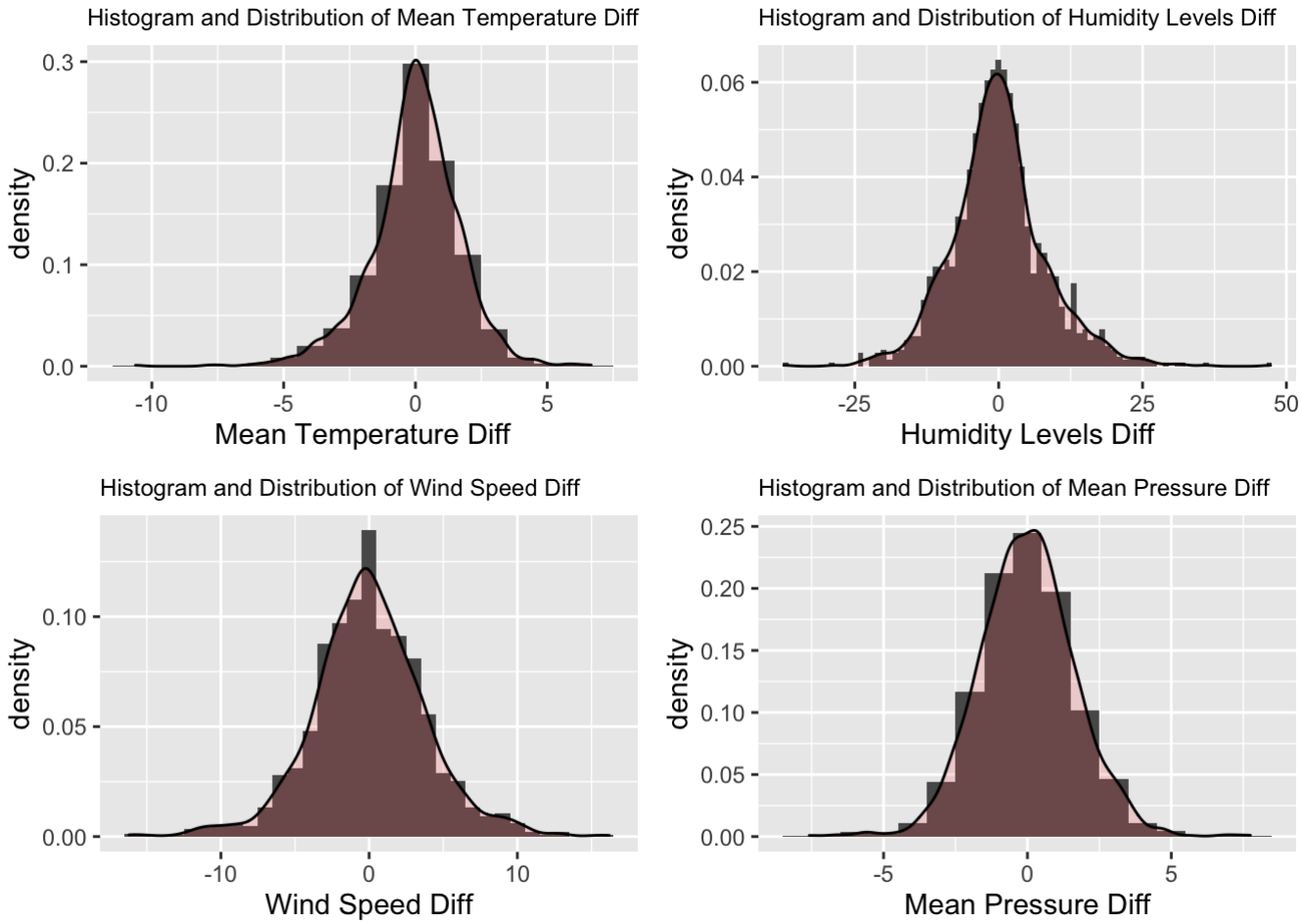
Even though we plan to differentiate the data, we believe it's still important to examine the correlation between the variables to see if there are any strong correlation between the variables. Examining the correlation plot between the variables, we see that *meanpressure* and *meantemp* has a correlation of -0.88, indicating that they are highly correlated and much of the variation in *meantemp* is explained by *meanpressure*. The correlation between *meantemp* and *humidity* is -0.57 so some of the variation in *meantemp* is explained by *humidity*. The correlation between *meantemp* and *wind_speed* is only 0.35 so they do have a weaker correlation.

```
climate_change_f <-
  climate_change_f %>%  mutate(meantemp_df = meantemp - lag(meantemp),
                           humidity_df = humidity - lag(humidity),
                           windspeed_df = wind_speed - lag(wind_speed),
                           meanpressure_df = meanpressure - lag(meanpressure))

climate_change_f$meantemp_df[1] <- 0
climate_change_f$humidity_df[1] <- 0
climate_change_f$windspeed_df[1] <- 0
climate_change_f$meanpressure_df[1] <- 0

head(climate_change_f)
```

```
##   date  meantemp humidity wind_speed meanpressure meantemp_df humidity_df
## 1    1 10.000000 84.50000   0.000000     1015.667   0.0000000    0.000000
## 2    2  7.400000 92.00000   2.980000     1017.800  -2.6000000    7.500000
## 3    3  7.166667 87.00000   4.633333     1018.667  -0.2333333   -5.000000
## 4    4  8.666667 71.33333   1.233333     1017.167   1.5000000  -15.666667
## 5    5  6.000000 86.83333   3.700000     1016.500  -2.6666667   15.500000
## 6    6  7.000000 82.80000   1.480000     1018.000   1.0000000   -4.033333
##   windspeed_df meanpressure_df
## 1     0.000000       0.0000000
## 2     2.980000       2.1333333
## 3     1.653333       0.8666667
## 4    -3.400000      -1.5000000
## 5     2.466667      -0.6666667
## 6    -2.220000       1.5000000
```

Now, we have 4 new variables where *meantemp_df* is the *meantemp* variable differentiated once on its lag, *humidity_df* is the *humidity* variable differentiated on its lag, *windspeed_df* is the *wind_speed* variable differentiated on its lag, and *meanpressure_df* is the *meanpressure* differentiated on its lag.

```
temp_df <- climate_change_f %>%
  ggplot(mapping = aes(x = meantemp_df)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Mean Temperature Diff") +
  labs(x = "Mean Temperature Diff") +
  theme(plot.title = element_text(size = 9))

humid_df <- climate_change_f %>%
  ggplot(mapping = aes(x = humidity_df)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Humidity Levels Diff") +
  labs(x = "Humidity Levels Diff") +
  theme(plot.title = element_text(size = 9))

wind_df <- climate_change_f %>%
  ggplot(mapping = aes(x = windspeed_df)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Wind Speed Diff") +
  labs(x = "Wind Speed Diff") +
  theme(plot.title = element_text(size = 9))

pressure_df <- climate_change_f %>%
  ggplot(mapping = aes(x = meanpressure_df)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  geom_density(alpha= .2, fill="#FF6666") +
  ggtitle("Histogram and Distribution of Mean Pressure Diff") +
  labs(x = "Mean Pressure Diff") +
  theme(plot.title = element_text(size = 9))

plot_grid(temp_df, humid_df, wind_df, pressure_df, ncol = 2, nrow = 2)
```



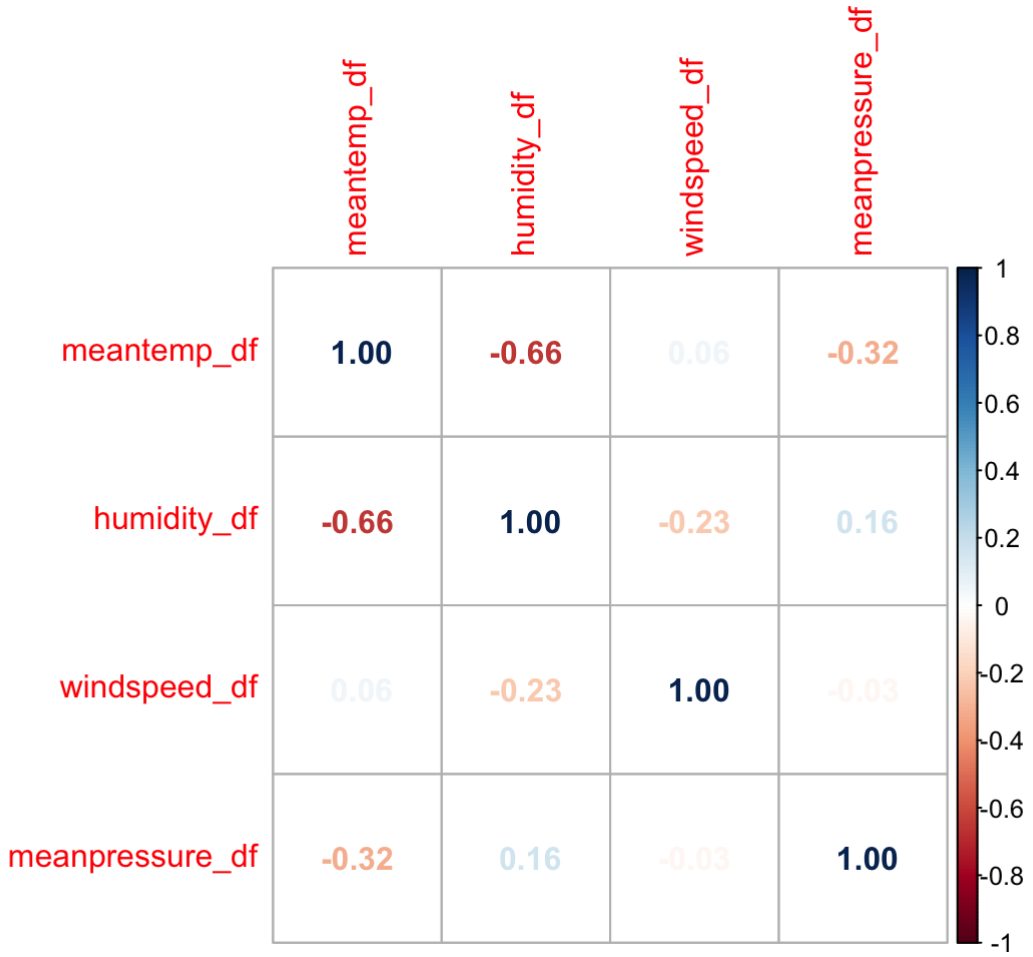
These are the histograms of our dataset after we took out the outliers and differentiated them as indicated by “Diff” in the labels:

- The probability histogram for the mean temperature after differentiation seems to have normality because it is bell-shaped, unimodal, and pretty symmetric around the mean, except for a longer tail to the left. This unimodal graph allows our forecast to be easier to graph.
- The probability histogram for humidity levels after differentiation seems to have normality because it is bell-shaped, unimodal, and pretty symmetric around the mean, except for a longer tail to the right. This unimodal graph allows our forecast to be easier to graph.
- The probability histogram for the wind speed after differentiation has normality because although it is bell-shaped, unimodal, and symmetric around the mean. This histogram with normality allows our forecast to be easier to graph.
- The probability histogram for the mean pressure has normality because it is unimodal, bell-shaped, and pretty symmetric around the mean. This histogram with normality allows our forecast to be easier to graph. We removed one very high outlier to the right, and removing it created a better representation of our data.

```
corr_mroz_diff <- climate_change_f[, (names(climate_change_f) %in%
                                c('meantemp_df', 'humidity_df',
                                  'windspeed_df', 'meanpressure_df'))]
res_diff <- cor(corr_mroz_diff)
round(res_diff, 2)
```

```
##                 meantemp_df humidity_df windspeed_df meanpressure_df
## meantemp_df            1.00       -0.66         0.06           -0.32
## humidity_df           -0.66        1.00        -0.23            0.16
## windspeed_df           0.06       -0.23         1.00           -0.03
## meanpressure_df       -0.32        0.16        -0.03            1.00
```

```
corrplot(res_diff, method = 'number')
```



Examining the correlation plot between the differentiated data, we see that *humidity_df* and *meantemp_df* has a correlation of -0.66, indicating that they are highly correlated and much of the variation in *meantemp_df* is explained by *humidity_df*. The correlation between the other variables are lighter in color, indicating a weaker correlation. Comparing this to the non-differentiated data, *humidity_df* is now a stronger candidate for correlation.

```
lapply(climate_change_f, summary)[-1]
```

```
## $meantemp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   18.75   27.62   25.42   31.25   38.71
##
## $humidity
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.47   50.70   62.81   61.01   72.38  100.00
##
## $wind_speed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.437   6.025   6.460   8.921  17.908
##
## $meanpressure
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   991.4  1001.6  1008.7  1008.3  1014.9  1023.0
##
## $meantemp_df
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.6250  -0.8750   0.0625   0.0000   1.0000   6.6667
##
## $humidity_df
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -37.3333  -4.8089  -0.2500   0.0109   4.0870  47.2500
##
## $windspeed_df
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -16.20000  -2.31250  -0.05058   0.00000   2.31250  16.21250
##
## $meanpressure_df
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -7.583333 -1.119167  0.000000  0.000234  1.000000  7.750000
```

These are the 5-number summary data for each of the variables:

- For the average temperature's 5-number summary, the minimum is 6 degrees celsius, 1st quartile value is 18.75 degrees celsius, the median is 27.62 degrees celsius, the mean is 25.42 degrees celsius, the 3rd quartile value is 31.25 degrees celsius, and the maximum is 38.71 degrees celsius. Since our **median is higher than our mean, our data for mean temperature would be skewed to the left**, with a longer tail of low scores pulling the mean down more than the median.
- For humidity's 5-number summary, the minimum is 18.47 grams per cubic meter, 1st quartile value is 50.70 grams per cubic meter, the median is 62.81 grams per cubic meter, the mean is 61.01 grams per cubic meter, the 3rd quartile value is 72.38 grams per cubic meter, and the maximum is 100 grams per cubic meter. Since our **median is higher than our mean, our data for humidity would be skewed to the left**, with a longer tail of low scores pulling the mean down more than the median.
- For wind speed's 5-number summary, the minimum is 0 kmph, 1st quartile value is 3.437 kmph, the median is 6.025 kmph, the mean is 6.460 kmph, the 3rd quartile value is 8.921 kmph, and the maximum is 17.908 kmph. Since our **median is slightly lower than our mean, our data for wind speed would be slightly skewed to the right**, with a longer tail of high scores pulling the mean up more than the median.
- For mean pressure's 5-number summary, the minimum is 991.4 atm, 1st quartile value is 1001.6 atm, the median is 1008.7 atm, the mean is 1008.3 atm, the 3rd quartile value is 1014.9 atm, and

the maximum is 1023 atm. Since our **median is slightly higher than our mean, our data for mean pressure would be slightly skewed to the left**, with a longer tail of low scores pulling the mean down more than the median.

We also made the 5-number summary data for each of the variables after we differentiated them:

- For the average temperature's differentiated 5-number summary, we would set our mean to 0. So our minimum is -10.6250 degrees celsius, 1st quartile value is -0.8750 degrees celsius, the median is 0.0625 degrees celsius, the mean is 0 degrees celsius, the 3rd quartile value is 1 degrees celsius, and the maximum is 6.6667 degrees celsius. Since our **median is higher than our mean, our differentiated data for mean temperature would be skewed to the left**, with a longer tail of low scores pulling the mean down more than the median.
- For humidity's differentiated 5-number summary, we would set our mean to 0. So our minimum is -37.3333 grams per cubic meter, 1st quartile value is -4.8089 grams per cubic meter, the median is -0.25 grams per cubic meter, the mean is 0.0109 grams per cubic meter, the 3rd quartile value is 4.0870 grams per cubic meter, and the maximum is 47.25 grams per cubic meter. Since our **median is now lower than our mean, our differentiated data for humidity would be slightly skewed to the right**, with a longer tail of high scores pulling the mean up more than the median. But they are very close, indicating symmetry.
- For wind speed's differentiated 5-number summary, we would set our mean to 0. So our minimum is -16.2 kmph, 1st quartile value is -2.31 kmph, the median is -0.05 kmph, the mean is 0 kmph, the 3rd quartile value is 2.31 kmph, and the maximum is 16.21 kmph. Since our **median is slightly lower than our mean, our differentiated data for wind speed would be slightly skewed to the right**, with a longer tail of high scores pulling the mean up more than the median.
- For mean pressure's differentiated 5-number summary, we would set our mean to 0. So our minimum is -7.58 atm, 1st quartile value is -1.11 atm, the median is 0 atm, the mean is around 0 atm, the 3rd quartile value is 1 atm, and the maximum is 7.75 atm. **Since our median is equal to our mean, our differentiated data for mean pressure would be symmetric.**

2. Data pre-processing

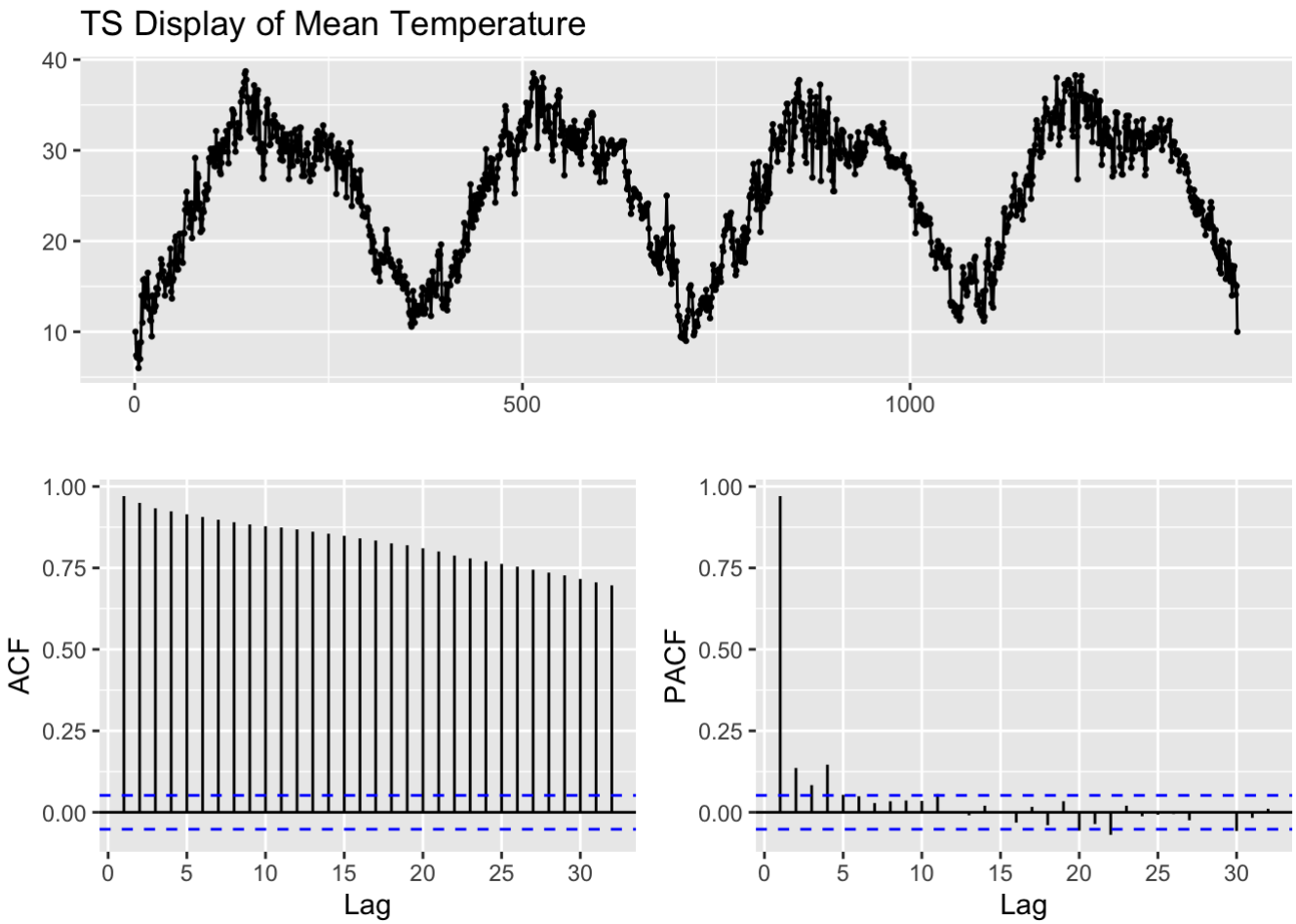
a

Before diving into our differenced data, we can examine the variables to see if un-differenced variables are stationary are not in order to see if they can be used as a proper model to forecast change in mean temperature.

For the structure of the time series, the lag order here is the number of lags used in the regression and the code is `trunc((length(x)-1)^(1/3))` which corresponds to the suggested upper bound on the rate at which the number of lags should be made to grow with the sample size. It is basically just the number of lags used to test if it's stationary.

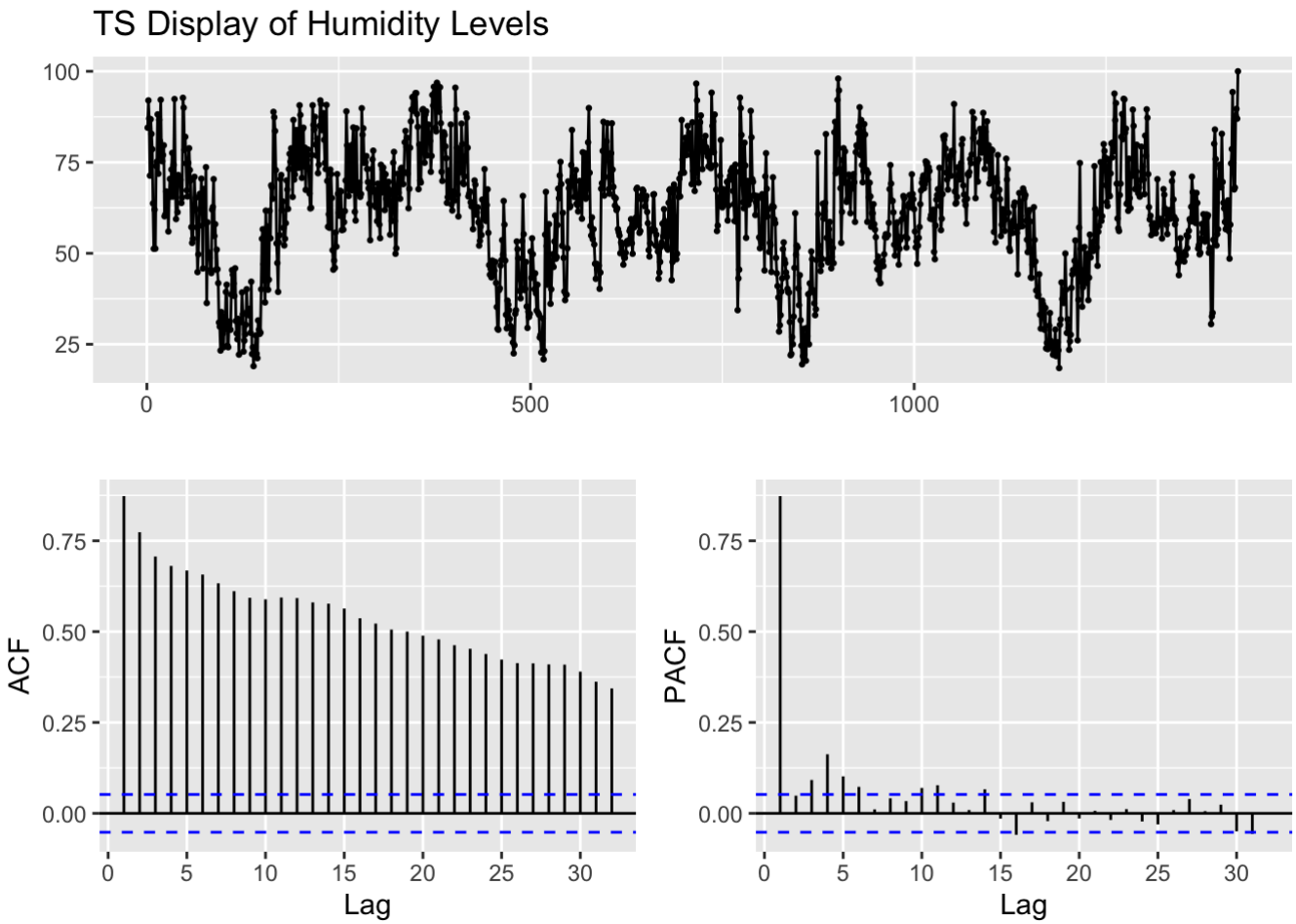
The observations appear to follow each other. We are using seasonal data, so the observations follow a trend to go up and down. The variance for all of these variables does not seem to be increasing; it looks pretty constant. Also, all of our graphs have seasonality. The existence of stationary for each of our variables is indicated below:

```
ggtsdisplay(climate_change_f$meantemp, main = "TS Display of Mean Temperature")
```



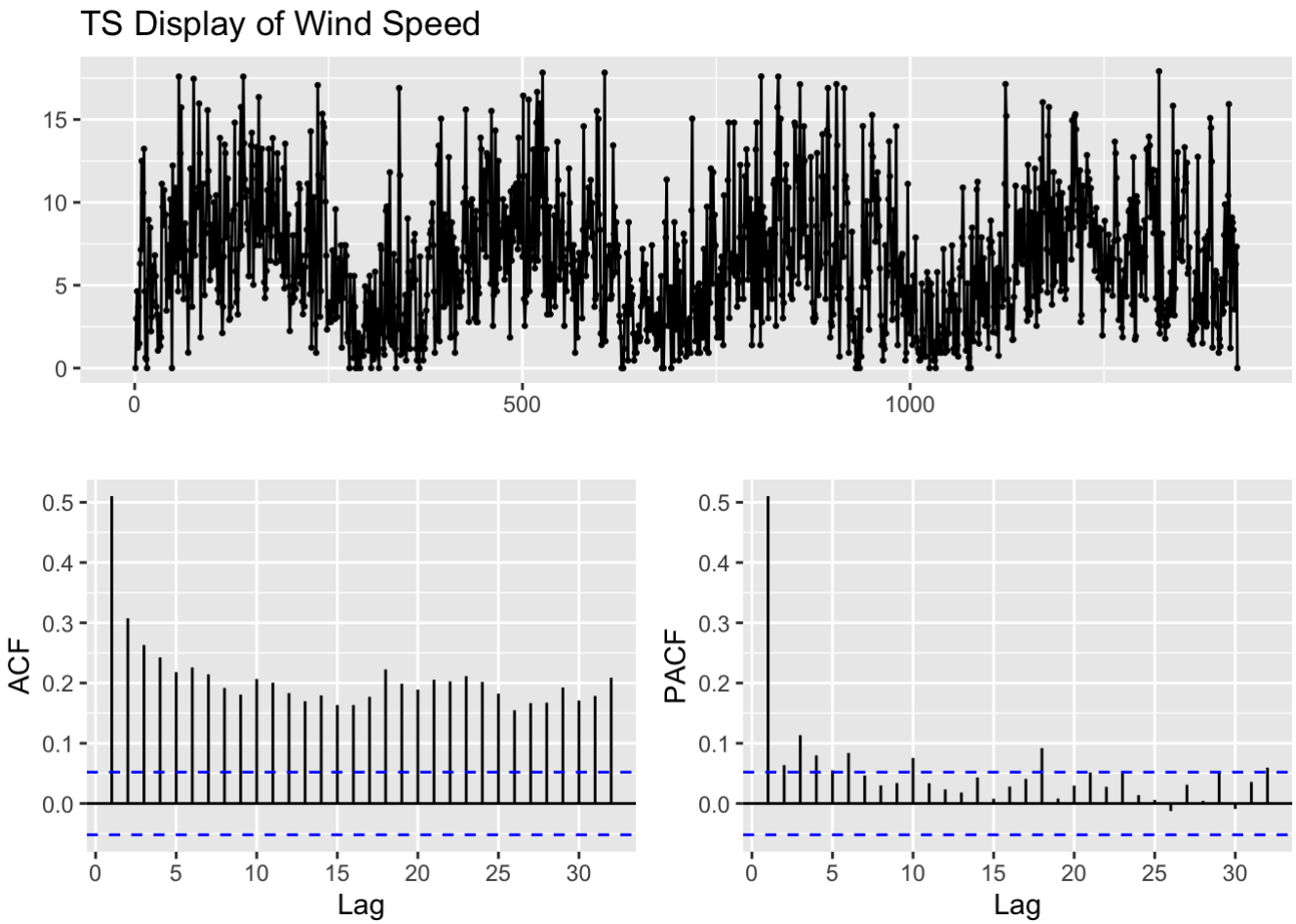
Looking at the `ggtsdisplay` of mean temperature, we can see that the time display at the top is reverberating, but it doesn't seem to be reverberating enough as the mean can change if you take different subsets of the data. The data also shows that this is cyclical as the rise and fall trend of the data is clear and they always seem to go in the same places. Mean temperature also seems to follow an $AR(4)$ model as the PACF of the model shows significance at lag 4. The ACF is also declining, but it is declining very slowly, indicating that it is non-stationary.

```
ggtsdisplay(climate_change_f$humidity, main = "TS Display of Humidity Levels")
```



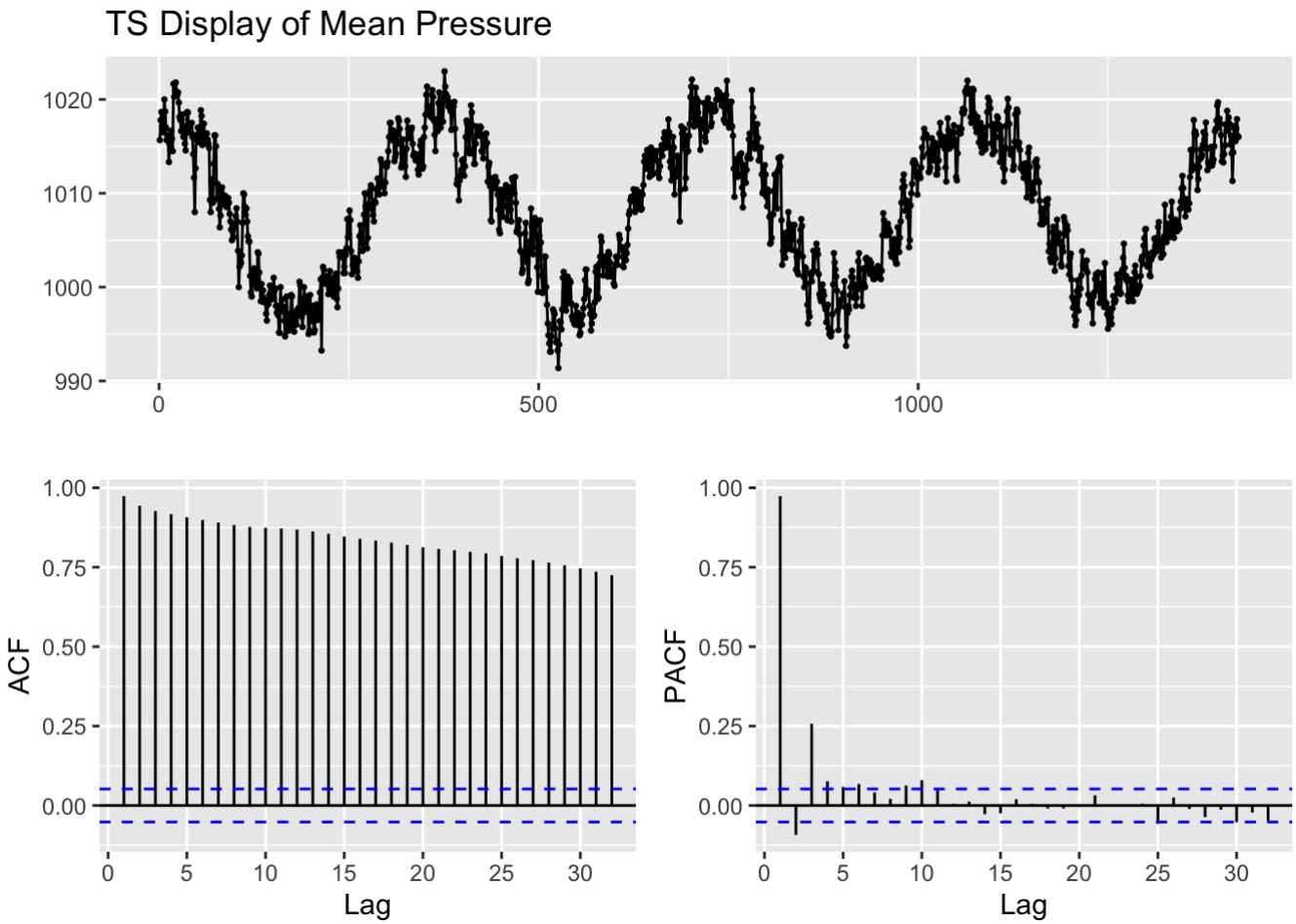
For the `ggtsdisplay` of humidity, we see that this follows an $AR(4)$ model as the ACF graph shows declining lags, and the PACF graph shows significance at lag 4. The time display at the top seems to be reverberating, and there seems to be a pattern of large and small ups and downs. Also, the variance does not seem to be constant and not increasing. Although it is reverberating, it is not reverberating enough to indicate that it is stationary. Moreover, the ACF graph is slowly declining, indicating that humidity may be non-stationary.

```
ggtsdisplay(climate_change_f$wind_speed, main = "TS Display of Wind Speed")
```



We used the `ggtsdisplay` method to generate ACF and PACF plots for mean temperature, wind speed, mean pressure, and humidity. The time series for wind speed displays seasonality as we see variations occur at specific intervals. The oscillations seem too pronounced to describe the data as mean-riveting. Wind speed seems to follow an $AR(18)$ model since the lags tail off in the ACF plot and cut off at lag 18. Furthermore, the ACF plot for wind speed shows a sharp decline early on, however, it stops decaying soon after. With this in mind, wind speed appears to be non-stationary.

```
ggtsdisplay(climate_change_f$meanpressure, main = "TS Display of Mean Pressure")
```



Looking at the `ggtsdisplay` for mean pressure, we see that this follows an $AR(2)$ model as the ACF graph shows declining lags and the PACF shows significance at lag 2. The time display at the top is also reverberating, and the variance does seem to be constant, but we do think that it's not reverberating enough to indicate that it is stationary. However, the ACF graph is slowly declining, indicating that mean pressure may be non-stationary.

The ACF graphs are taking a significant amount of time to decay, but they all do indeed look like AR models, moving forward, I believe using the differentiated series would be best. All variables have their ACF that is taking a long time to decay, so I believe approaching this and using the differentiated series would help immensely.

Here is the `adf.test` for each of the variables to test for stationary.

```
adf.test(climate_change_f$meantemp)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$meantemp
## Dickey-Fuller = -1.8886, Lag order = 11, p-value = 0.6255
## alternative hypothesis: stationary
```

```
adf.test(climate_change_f$humidity)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$humidity
## Dickey-Fuller = -3.7309, Lag order = 11, p-value = 0.02245
## alternative hypothesis: stationary
```

```
adf.test(climate_change_f$wind_speed)
```

```
## Warning in adf.test(climate_change_f$wind_speed): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$wind_speed
## Dickey-Fuller = -6.5266, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(climate_change_f$meanpressure)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$meanpressure
## Dickey-Fuller = -2.1196, Lag order = 11, p-value = 0.5277
## alternative hypothesis: stationary
```

Since the p value for *meantemp*, *humidity*, and *meanpressure* are all above 0.05, we fail to reject the null that it is non-stationary, so there is significance that the variable is non-stationary. The variable *wind_speed* is indicated to be stationary, but the ACF graph of it is slowly decaying, which is alarming to us so we decided to continue with the differentiated data.

```
ndiffs(climate_change_f$meantemp)
```

```
## [1] 1
```

```
ndiffs(climate_change_f$humidity)
```

```
## [1] 0
```

```
ndiffs(climate_change_f$wind_speed)
```

```
## [1] 0
```

```
ndiffs(climate_change_f$meanpressure)
```

```
## [1] 0
```

By using the `ndiffs` function, we see that *meantemp* should be differentiated once in order to make it stationary but everything else is 0 which is still alarming as the ACF of *humidity*, and *meanpressure* indicates non-stationary and the `adf.test()` also indicates that they are non-stationary, so there should be at least some differencing to make it stationary.

```
ndiffs(climate_change_f$meantemp_df)
```

```
## [1] 0
```

```
ndiffs(climate_change_f$humidity_df)
```

```
## [1] 0
```

```
ndiffs(climate_change_f$windspeed_df)
```

```
## [1] 0
```

```
ndiffs(climate_change_f$meanpressure_df)
```

```
## [1] 0
```

If we look at the differentiated data, it indicates that no differencing is needed, this may be because we already differentiated it once and now they should be stationary. We can use the `adf.test` to examine if they are stationary.

```
adf.test(climate_change_f$meantemp_df)
```

```
## Warning in adf.test(climate_change_f$meantemp_df): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$meantemp_df
## Dickey-Fuller = -13.192, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(climate_change_f$humidity_df)
```

```
## Warning in adf.test(climate_change_f$humidity_df): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$humidity_df
## Dickey-Fuller = -15.124, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(climate_change_f$windspeed_df)
```

```
## Warning in adf.test(climate_change_f$windspeed_df): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$windspeed_df
## Dickey-Fuller = -17.038, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

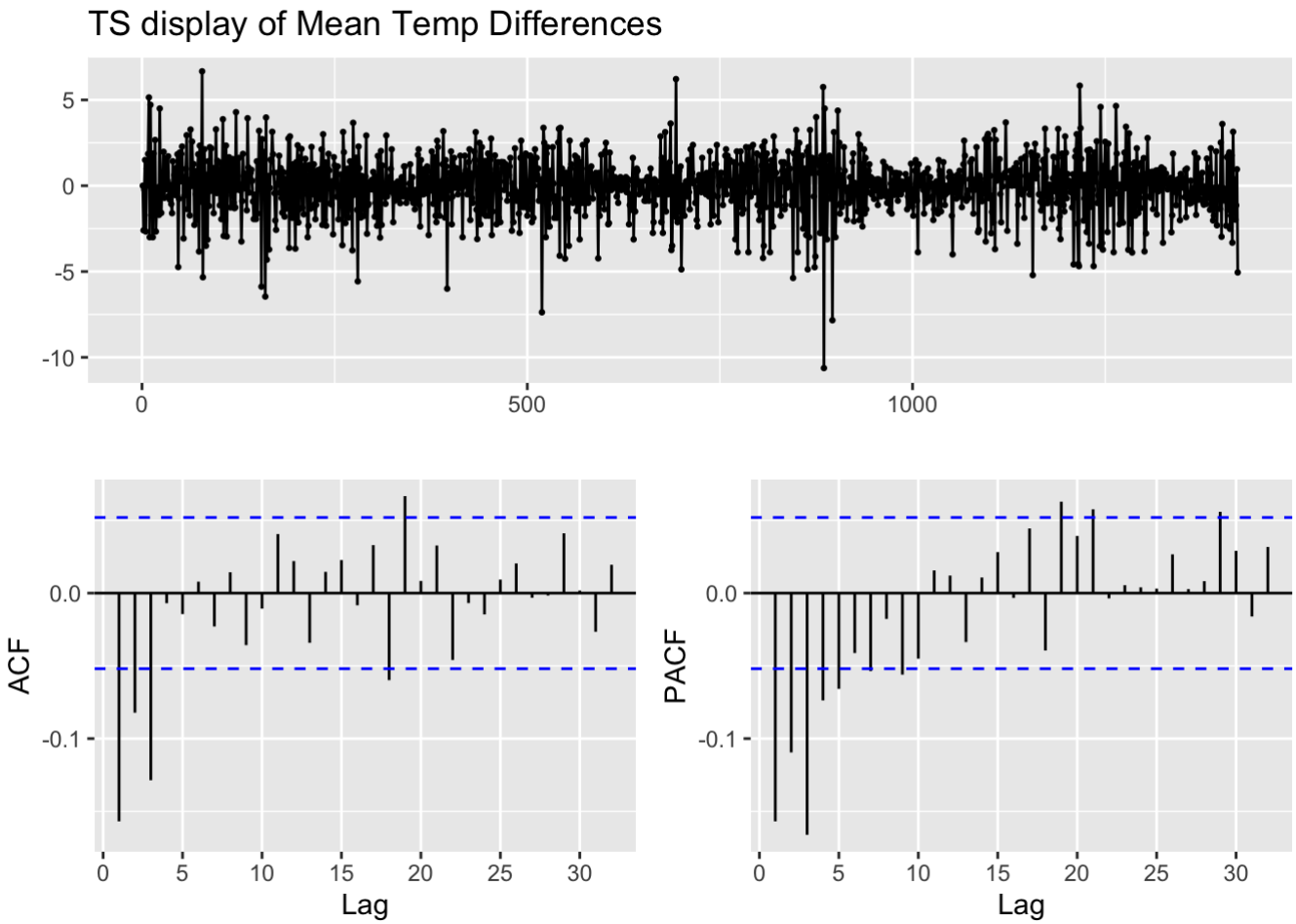
```
adf.test(climate_change_f$meanpressure_df)
```

```
## Warning in adf.test(climate_change_f$meanpressure_df): p-value smaller than
## printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  climate_change_f$meanpressure_df
## Dickey-Fuller = -14.401, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

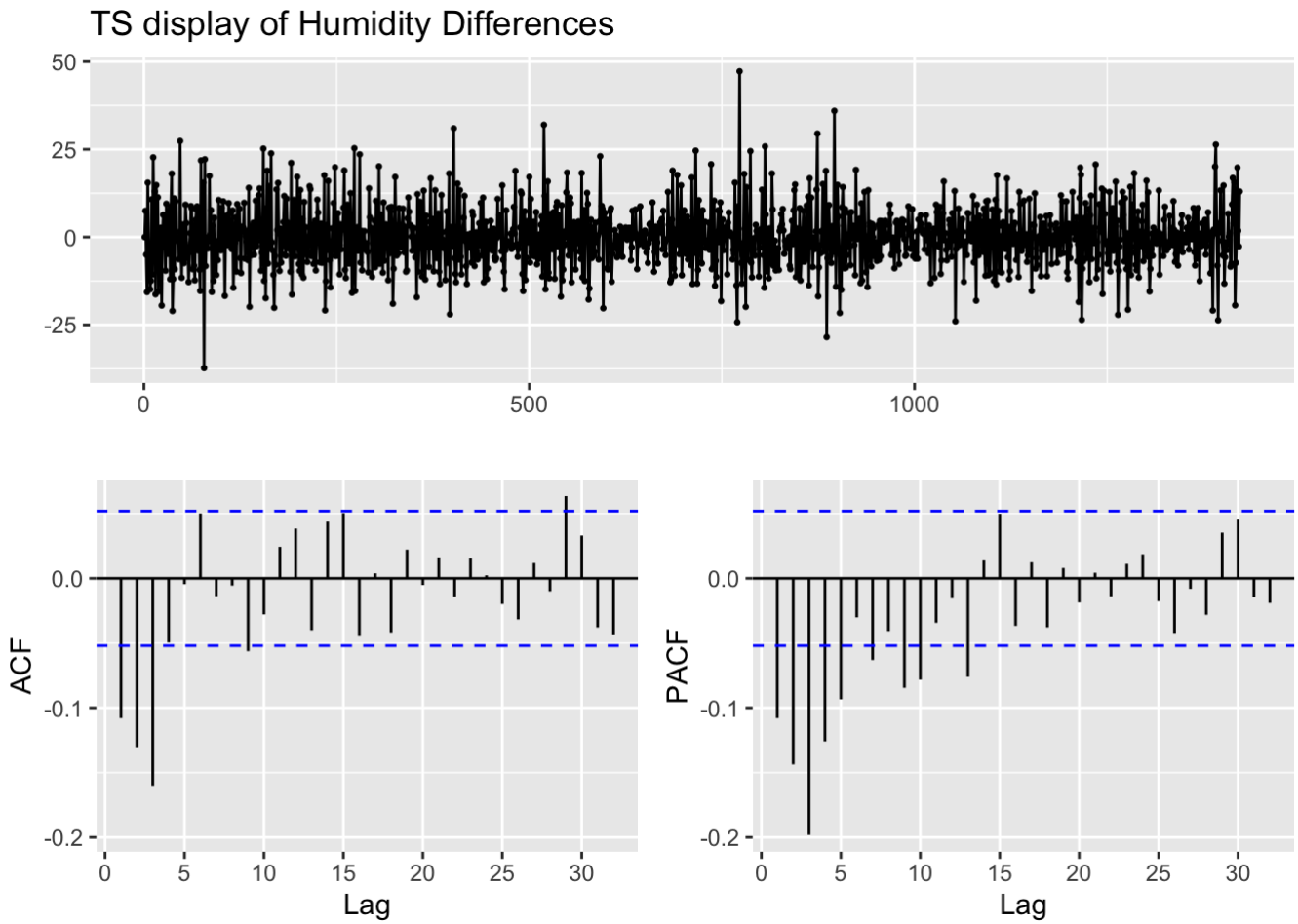
Since now all the p values are lower than 0.05 now, we reject the null that it is non-stationary.

```
ggtsdisplay(climate_change_f$meantemp_df, main = "TS display of Mean Temp Differenc
es")
```



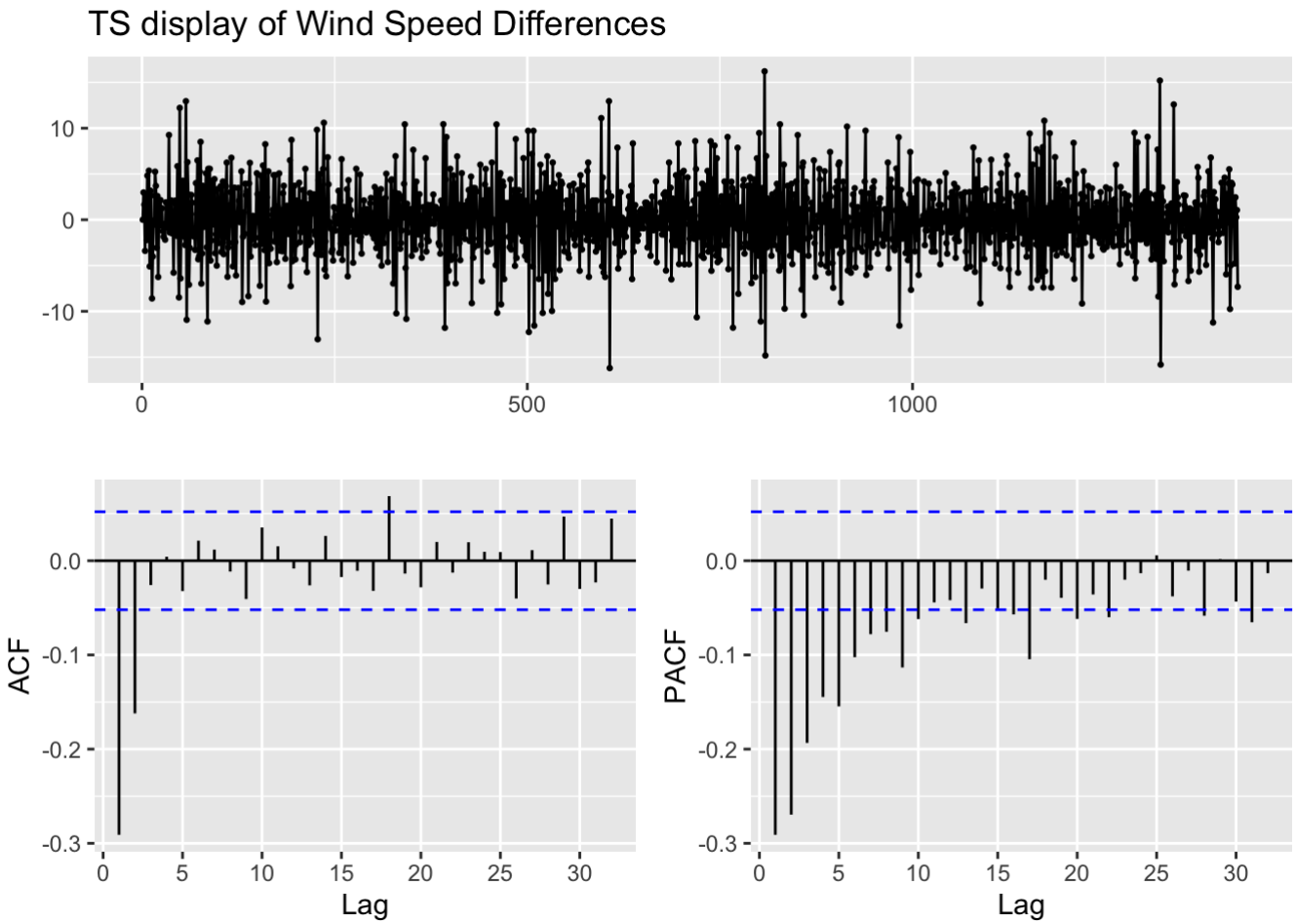
Examining the Mean Temp after differencing shows that for the time display, there doesn't seem to be a trend and the mean and variance are quite constant. It also follows more of a MA trend as the PACF is gradually declining and the ACF looks like an AR's PACF model. If this was the case, this would be $MA(19)$ as the 19th lag in the ACF model is the furthest and is also significant.

```
ggtsdisplay(climate_change_f$humidity_df, main = "TS display of Humidity Difference
s")
```



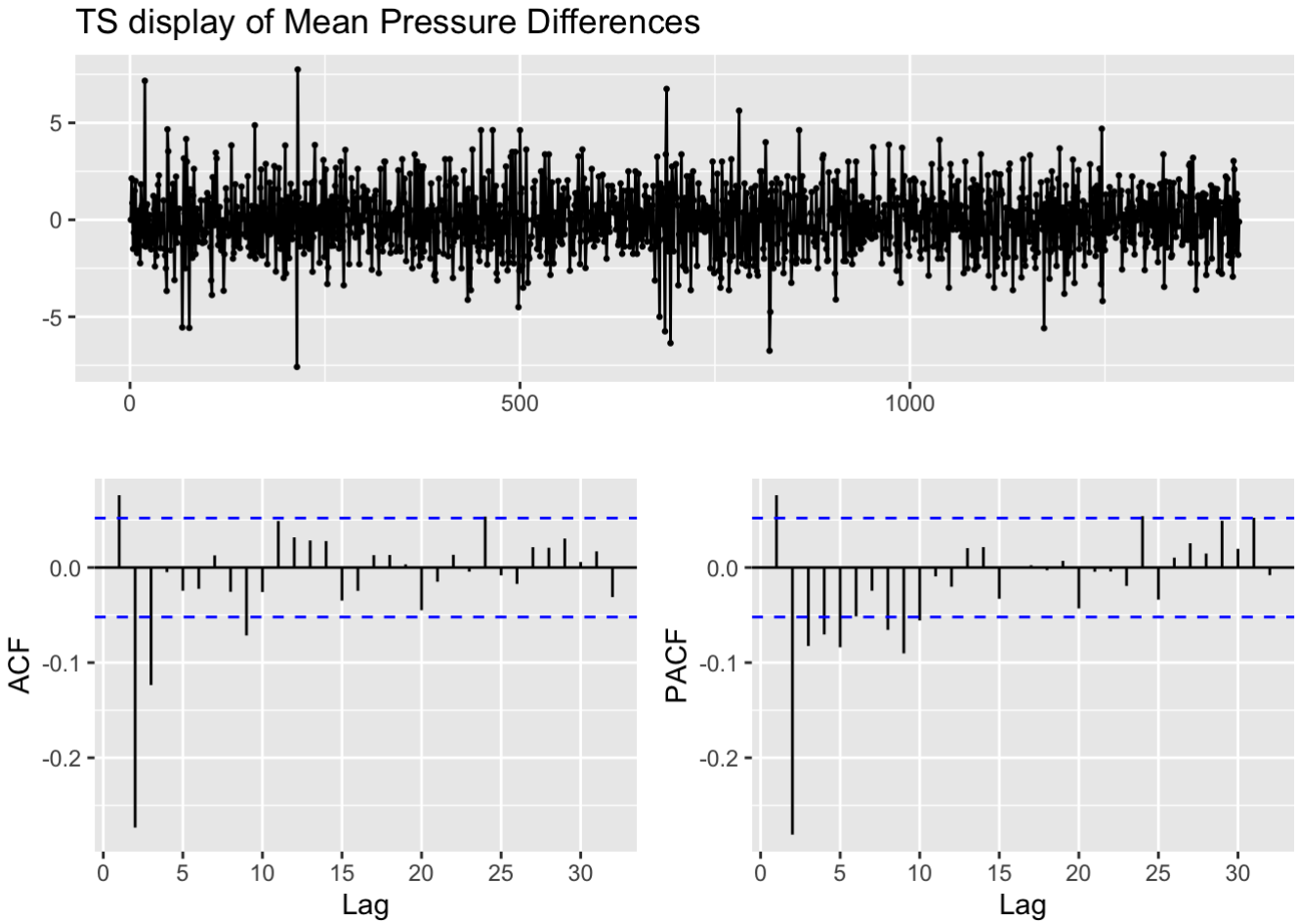
After differencing, the TS display of humidity differences seems to reverberate much more than before differencing, and the variance seems not to change widely than before. The graph overall does not have seasonal variations nor trends, thus indicating stationary. Also, for the ACF model after differencing, a $MA(29)$ model appears to be a better fit for humidity than an $AR(13)$ model since there is a drastic decline in the lags in the PACF graph while the ACF graph cuts off after lag 29. Also, the PACF graph is gradually declining while the ACF graph looks like an AR's PACF model.

```
ggtsdisplay(climate_change_f$windspeed_df, main = "TS display of Wind Speed Differe
nces")
```



Looking at the `ggtsdisplay` for wind speed after first-differencing it, the TS display appears to be mean reverting as there are no longer any seasonal variations or trends, indicating stationary. After differencing wind speed, a MA model appears to be a better fit for wind speed than an $AR(18)$ model since we see a geometric decline in the lags in the PACF plot, while the ACF plot cuts off after lag 18.

```
ggtsdisplay(climate_change_f$meanpressure_df, main = "TS display of Mean Pressure D
ifferences")
```



Examining the `ggtsdisplay` for the mean pressure differences, we see that the time display is now more reverberating and that the mean and variance do not seem to change widely. There seems to be more trend in the time display which indicates stationary. Examining the ACF and PACF, this mean pressure differences seems to resemble a MA model instead as the PACF has gradually declining lags and the ACF seems to look like the PACF of an AR model. If this was a MA model, then it would be $MA(2)$.

3. Feature Generation, Model Testing and Forecasting

a

Base on the tsdisplay of the $meantemp\_df$ variable, we shall create three different AR models of it. Base on the PACF, we saw that lags 5, 9, and 19 were all significant so we shall be building $AR(5)$, $AR(9)$, and $AR(19)$ where the $meantemp\_df$ is regressed against itself.

```
ar_5 <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) + lag(meantem
p_df, 3) +
              lag(meantemp_df, 4) + lag(meantemp_df, 5), data = climate_change_f)

ar_9 <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) + lag(meantem
p_df, 3) +
              lag(meantemp_df, 4) + lag(meantemp_df, 5) + lag(meantemp_df, 6) +
              lag(meantemp_df, 7) + lag(meantemp_df, 8) + lag(meantemp_df, 9),
              data = climate_change_f)

ar_19 <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) + lag(meante
mp_df, 3) +
              lag(meantemp_df, 4) + lag(meantemp_df, 5) + lag(meantemp_df, 6) +
              lag(meantemp_df, 7) + lag(meantemp_df, 8) + lag(meantemp_df, 9) +
              lag(meantemp_df, 10) + lag(meantemp_df, 11) + lag(meantemp_df, 12) +
              lag(meantemp_df, 13) + lag(meantemp_df, 14) + lag(meantemp_df, 15) +
              lag(meantemp_df, 16) + lag(meantemp_df, 17) + lag(meantemp_df, 18) +
              lag(meantemp_df, 19), data = climate_change_f)
```

```
model_score <- data.frame("Model" = paste("ar", c(5, 9, 19), sep = ""),
                          "BIC" = c(BIC(ar_5), BIC(ar_9), BIC(ar_19)),
                          "AIC" = c(AIC(ar_5), AIC(ar_9), AIC(ar_19)))

sorted_model <- model_score[order(model_score$BIC), ]
sorted_model
```

```
##   Model      BIC      AIC
## 2   ar9 5439.657 5381.868
## 3  ar19 5445.055 5334.882
## 1   ar5 5445.897 5409.103
```

The BIC for ar9, or $AR(9)$ is the lowest and since the BIC penalizes the model more for more variables, we believe following the BIC would be better. The AIC is still a good measure of the model, but our model for $AR(9)$ has a AIC score of 5381.868, which is the second smallest out of the tree, 50 off the smallest AIC and 18 away from the biggest AIC. However, the BIC for ar19, $AR(19)$ and ar5, $AR(5)$ is extremely close together so that is a warning that both of these models similarly fit the data.

```
summary(ar_9)
```

```
##
## Time series regression with "numeric" data:
## Start = 1, End = 1413
##
## Call:
## dynlm(formula = meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df,
##     2) + lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
##     5) + lag(meantemp_df, 6) + lag(meantemp_df, 7) + lag(meantemp_df,
##     8) + lag(meantemp_df, 9), data = climate_change_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4174 -0.9075  0.0812  1.0233  6.8010
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.002859   0.043052   0.066 0.947070
## lag(meantemp_df, 1) -0.220341   0.026655  -8.266 3.17e-16 ***
## lag(meantemp_df, 2) -0.175410   0.027262  -6.434 1.70e-10 ***
## lag(meantemp_df, 3) -0.212424   0.027608  -7.694 2.66e-14 ***
## lag(meantemp_df, 4) -0.111346   0.028147  -3.956 8.01e-05 ***
## lag(meantemp_df, 5) -0.096996   0.028173  -3.443 0.000593 ***
## lag(meantemp_df, 6) -0.067893   0.028122  -2.414 0.015895 *
## lag(meantemp_df, 7) -0.064001   0.027619  -2.317 0.020631 *
## lag(meantemp_df, 8) -0.030701   0.027268  -1.126 0.260404
## lag(meantemp_df, 9) -0.056283   0.026674  -2.110 0.035036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.618 on 1403 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.08197,    Adjusted R-squared:  0.07608
## F-statistic: 13.92 on 9 and 1403 DF,  p-value: < 2.2e-16
```

Our $AR(9)$ model would be $AR(9) : meantemp\_df_t = 0.002859 +$

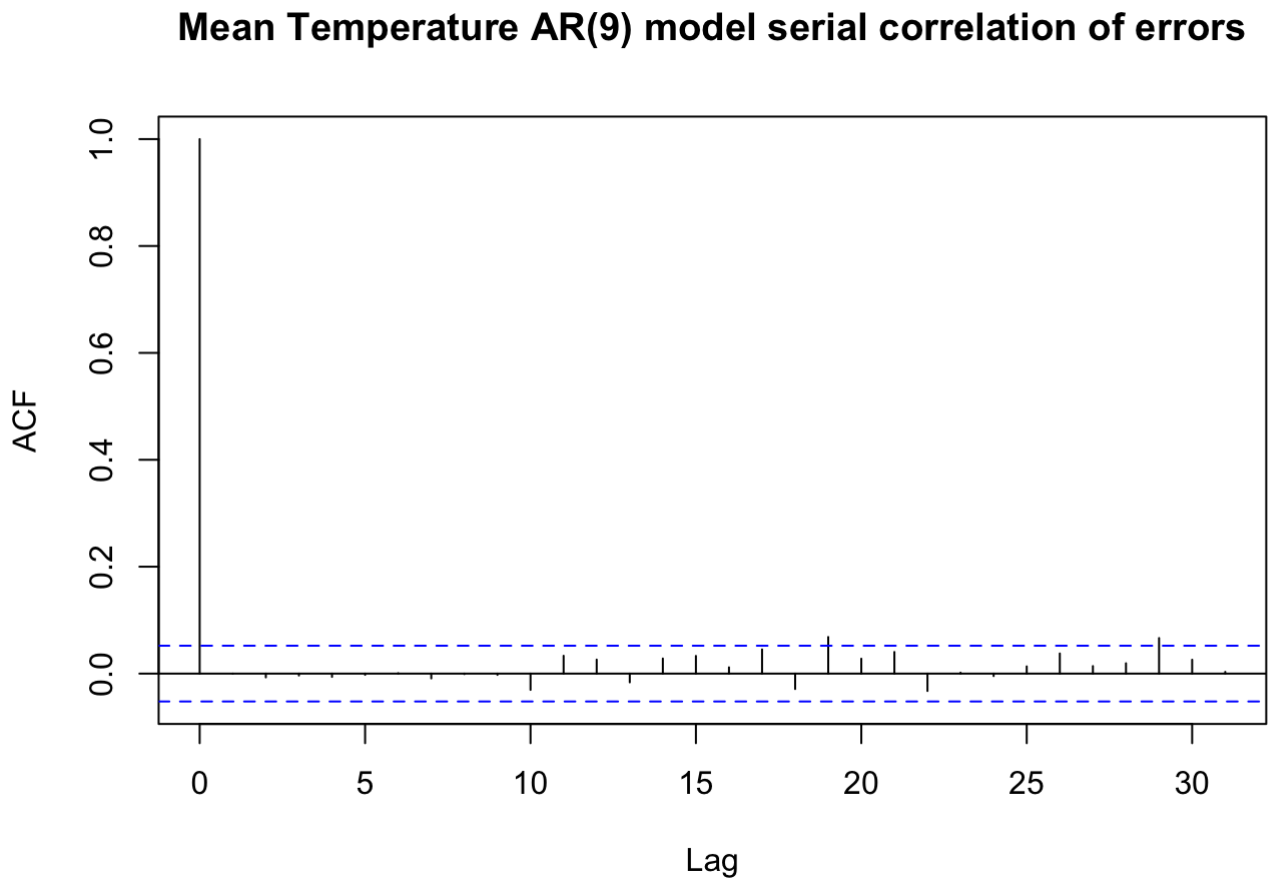
$$-0.220341 meantemp\_df_{t-1} + -0.175410 meantemp\_df_{t-2} +$$
$$-0.212424 meantemp\_df_{t-3} + -0.111346 meantemp\_df_{t-4} +$$
$$-0.096996 meantemp\_df_{t-5} + -0.067893 meantemp\_df_{t-6} +$$
$$-0.064001 meantemp\_df_{t-7} + -0.030701 meantemp\_df_{t-8} +$$
$$-0.056283 meantemp\_df_{t-9}$$

What this model is telling us is that it is predicting the change in mean temperature at time t base on 9 previous lags. The immediate multiplier is -0.220341 while the total effect at time q is $\sum_{s=0}^{q} B_s$ where B represents the coefficients of the $AR(9)$ model excluding the intercept, 0.002859. Examining the model further, the intercept has an extremely large p value that is almost close to 1, indicating that the intercept is very not significant. Lags 6, 7, and 9 also show moderate significance as it's significant at the 1 level. Lag 8, however, is shown not to be significant enough. The R^2 is also very low, around 0.08197 so much of the variation in $meantemp\_df$ is not explained by itself. Interpreting the coefficients, the intercept is 0.002859, which does not mean the $meantemp\_df$ start at 0.002859, but rather it is used as a baseline for the model

to predict future forecasts. We can estimate the change at certain lags using these coefficients, for example, at $t-1$, the change in $meantemp\_df$ is -0.220341 with respect to $meantemp\_df_{t-1}$, which means for every unit of $meantemp\_df_{t-1}$, there is a change of -0.220341 in $meantemp\_df$.

b

```
acf(ar_9$residuals, main = "Mean Temperature AR(9) model serial correlation of erro
rs")
```



Examining the residuals to see if any errors show any sign of serial correlation, we see that in the correlogram, they seem to show little significance. Lag 19 and 29 are a bit alarming as they reach over the interval where it is statistically significance that serial correlation is present, but since they only reach over it by such a little amount, we decided to ignore them.

c

In order to create an ARDL model, we need to choose one of the three predictor variables: $humidity\_df$, $windspeed\_df$, and $meanpressure\_df$. In order to choose the best predictor variables, we shall use the Boruta algorithmn, VIF model, and Mallows' CP to choose for feature selection.

```
library(Boruta)
bor_res <- Boruta(meantemp_df ~., data = climate_change_f[, 6:9], doTrace = 1)
```

```
## After 9 iterations, +7.7 secs:
```

```
##  confirmed 3 attributes: humidity_df, meanpressure_df, windspeed_df;
```

```
##  no more attributes left.
```

```
attStats(bor_res)
```

```
##                    meanImp  medianImp     minImp     maxImp normHits  decision
## humidity_df     103.855129 104.639230 101.146764 106.483682        1 Confirmed
## windspeed_df      5.833455   5.560666   3.759606   7.593633        1 Confirmed
## meanpressure_df  24.633443  24.095514  22.392832  27.200551        1 Confirmed
```

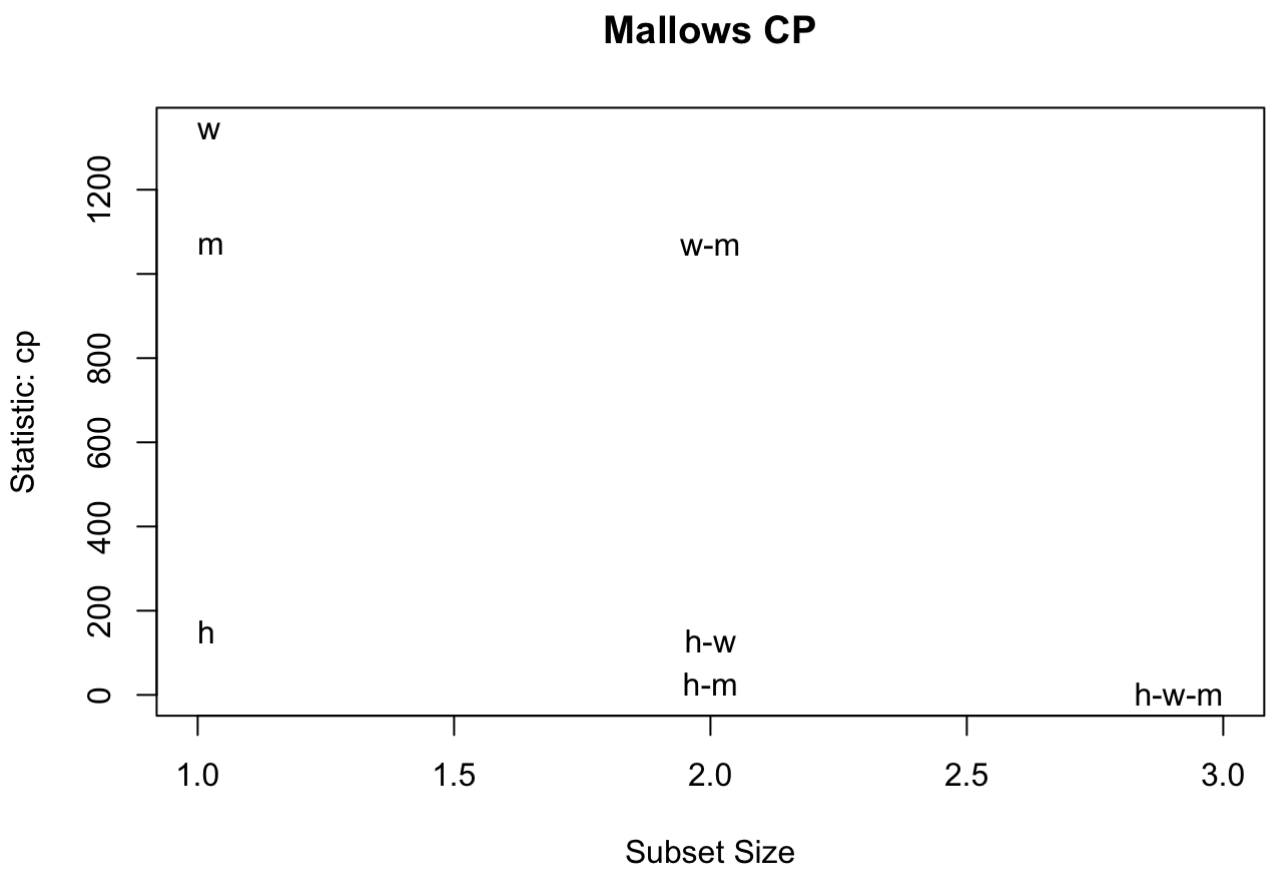
To measure the importance of the variables we would choose for the ARDL model, we first used the boruta algorithm in which it duplicates the dataset and shuffles each column and uses individual decision trees to spit out a class prediction and the higher the votes (or score), the more important the prediction. Here, we see that *humidity_df* has the highest prediction with a score of 103 and *meanpressure_df* with a score of 24 and *windspeed_df* at 6.

```
vif_model <- dynlm(meantemp_df ~ humidity_df + windspeed_df + meanpressure_df, data
= climate_change_f[, 6:9])
vif(vif_model)
```

```
##     humidity_df    windspeed_df meanpressure_df
##        1.081704        1.053914        1.027539
```

The VIF model here estimates for collinearity between the independent and the explanatory variable and looks for multicollinearity. The rule of thumb is that as long as the score of below 5, the variable does not shown signs of multicollinearity and is acceptable. So since all of the potential predicator variables are very much below 5, they are good predictors.

```
ss <- regsubsets(meantemp_df ~ humidity_df + windspeed_df + meanpressure_df, method
= c("exhaustive"), nbest = 3, data = climate_change_f)

subsets(ss, statistic = "cp", legend = F, main = "Mallows CP")
```




```
##                 Abbreviation
## humidity_df                h
## windspeed_df               w
## meanpressure_df            m
```

We decided to use Mallow's CP in order to find which predictor is best used to explain the explanatory variable. It basically creates a full model and compares it with a smaller model with a certain number of parameters and determines how much error is left unexplained by the partial model. The rule of thumb is that the smaller the score, the better the fit. Here, we see $humidity\_df$ is yet again the best fit, so our ARDL model would be using the $humidity\_df$ variable. We also want to point out that h-w-m, which means all three predictors, is lower than just h, so a VAR model containing all three predictors may better fit the forecast.

```
demo1 <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) +
                       lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
5) +
                       lag(meantemp_df, 6) + lag(meantemp_df, 7) + lag(meantemp_df,
8) +
                       lag(meantemp_df, 9) + humidity_df + lag(humidity, 1) +
                       lag(humidity_df, 2) + lag(humidity_df, 3) + lag(humidity_df,
4) +
                       lag(humidity_df, 5), data = climate_change_f)

summary(demo1)
```

```
##
## Time series regression with "numeric" data:
## Start = 1, End = 1413
##
## Call:
## dynlm(formula = meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df,
##     2) + lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
##     5) + lag(meantemp_df, 6) + lag(meantemp_df, 7) + lag(meantemp_df,
##     8) + lag(meantemp_df, 9) + humidity_df + lag(humidity, 1) +
##     lag(humidity_df, 2) + lag(humidity_df, 3) + lag(humidity_df,
##     4) + lag(humidity_df, 5), data = climate_change_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4609 -0.6562  0.0710  0.7313  6.1530
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.3377376  0.1380353   2.447  0.01454 *
## lag(meantemp_df, 1) -0.0900247  0.0214245  -4.202 2.81e-05 ***
## lag(meantemp_df, 2) -0.1864019  0.0268699  -6.937 6.10e-12 ***
## lag(meantemp_df, 3) -0.1975406  0.0270822  -7.294 5.02e-13 ***
## lag(meantemp_df, 4) -0.0875840  0.0273909  -3.198  0.00142 **
## lag(meantemp_df, 5) -0.0753320  0.0273946  -2.750  0.00604 **
## lag(meantemp_df, 6) -0.0448693  0.0217711  -2.061  0.03949 *
## lag(meantemp_df, 7) -0.0303219  0.0211831  -1.431  0.15254
## lag(meantemp_df, 8) -0.0005086  0.0208239  -0.024  0.98052
## lag(meantemp_df, 9) -0.0285684  0.0203076  -1.407  0.15972
## humidity_df         -0.1360437  0.0042039 -32.361  < 2e-16 ***
## lag(humidity, 1)    -0.0054473  0.0022011  -2.475  0.01345 *
## lag(humidity_df, 2) -0.0280023  0.0056464  -4.959 7.94e-07 ***
## lag(humidity_df, 3) -0.0280821  0.0054373  -5.165 2.76e-07 ***
## lag(humidity_df, 4) -0.0150141  0.0053774  -2.792  0.00531 **
## lag(humidity_df, 5) -0.0028055  0.0054915  -0.511  0.60952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.224 on 1397 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.4773, Adjusted R-squared:  0.4717
## F-statistic: 85.05 on 15 and 1397 DF,  p-value: < 2.2e-16
```

In our first model, we decided to go with an $ARDL(9, 5)$ model since the $AR(9)$ model was the best fit for $meantemp\_df$ and the PACF for the $humidity\_df$ model has lags 5, 10, and 13 that are significant. Observing our first model, we see that $meantemp\_df$ at lag 7 is where it stops being significant and for $humidity\_df$, it stops at lag 5. We should re-estimate this ARDL with $ARDL(6, 4)$.

```
demo2 <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) +
                      lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
5) +
                      lag(meantemp_df, 6) + humidity_df + lag(humidity, 1) +
                      lag(humidity_df, 2) + lag(humidity_df, 3) + lag(humidity_df,
4),
                      data = climate_change_f)

summary(demo2)
```

```
##
## Time series regression with "numeric" data:
## Start = 1, End = 1416
##
## Call:
## dynlm(formula = meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df,
##     2) + lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
##     5) + lag(meantemp_df, 6) + humidity_df + lag(humidity, 1) +
##     lag(humidity_df, 2) + lag(humidity_df, 3) + lag(humidity_df,
##     4), data = climate_change_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5138 -0.6718  0.0627  0.7191  5.9899
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.312477   0.136046   2.297  0.02177 *
## lag(meantemp_df, 1) -0.085419   0.021272  -4.016 6.25e-05 ***
## lag(meantemp_df, 2) -0.180303   0.026556  -6.790 1.66e-11 ***
## lag(meantemp_df, 3) -0.190938   0.026762  -7.135 1.55e-12 ***
## lag(meantemp_df, 4) -0.079535   0.026873  -2.960  0.00313 **
## lag(meantemp_df, 5) -0.057730   0.020805  -2.775  0.00560 **
## lag(meantemp_df, 6) -0.034357   0.020295  -1.693  0.09071 .
## humidity_df         -0.136707   0.004184 -32.671  < 2e-16 ***
## lag(humidity, 1)    -0.005012   0.002167  -2.313  0.02089 *
## lag(humidity_df, 2) -0.027974   0.005513  -5.075 4.40e-07 ***
## lag(humidity_df, 3) -0.027872   0.005307  -5.252 1.73e-07 ***
## lag(humidity_df, 4) -0.014402   0.005371  -2.681  0.00742 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.226 on 1404 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.4761, Adjusted R-squared:  0.472
## F-statistic:   116 on 11 and 1404 DF,  p-value: < 2.2e-16
```

Here, we see that the lag at 6 for $humididty\_df$ is only significant at the 10% level, so we re-estimate this with $ARDL(5,4)$.

```
mean_humid <- dynlm(meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df, 2) +
                      lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
5) +
                      humidity_df + lag(humidity, 1) + lag(humidity_df, 2) +
                      lag(humidity_df, 3) + lag(humidity_df, 4), data = climate_cha
nge_f)

summary(mean_humid)
```

```
##
## Time series regression with "numeric" data:
## Start = 1, End = 1417
##
## Call:
## dynlm(formula = meantemp_df ~ lag(meantemp_df, 1) + lag(meantemp_df,
##     2) + lag(meantemp_df, 3) + lag(meantemp_df, 4) + lag(meantemp_df,
##     5) + humidity_df + lag(humidity, 1) + lag(humidity_df, 2) +
##     lag(humidity_df, 3) + lag(humidity_df, 4), data = climate_change_f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5300 -0.6660  0.0768  0.7261  5.9757
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.286003   0.135080   2.117 0.034411 *
## lag(meantemp_df, 1) -0.081782   0.021164  -3.864 0.000116 ***
## lag(meantemp_df, 2) -0.178333   0.026534  -6.721 2.62e-11 ***
## lag(meantemp_df, 3) -0.183980   0.026453  -6.955 5.38e-12 ***
## lag(meantemp_df, 4) -0.074193   0.026681  -2.781 0.005497 **
## lag(meantemp_df, 5) -0.049862   0.020286  -2.458 0.014094 *
## humidity_df         -0.136723   0.004186 -32.665  < 2e-16 ***
## lag(humidity, 1)    -0.004584   0.002151  -2.131 0.033246 *
## lag(humidity_df, 2) -0.028610   0.005493  -5.209 2.19e-07 ***
## lag(humidity_df, 3) -0.028125   0.005304  -5.302 1.33e-07 ***
## lag(humidity_df, 4) -0.014755   0.005368  -2.749 0.006058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 1406 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.4752, Adjusted R-squared:  0.4715
## F-statistic: 127.3 on 10 and 1406 DF,  p-value: < 2.2e-16
```

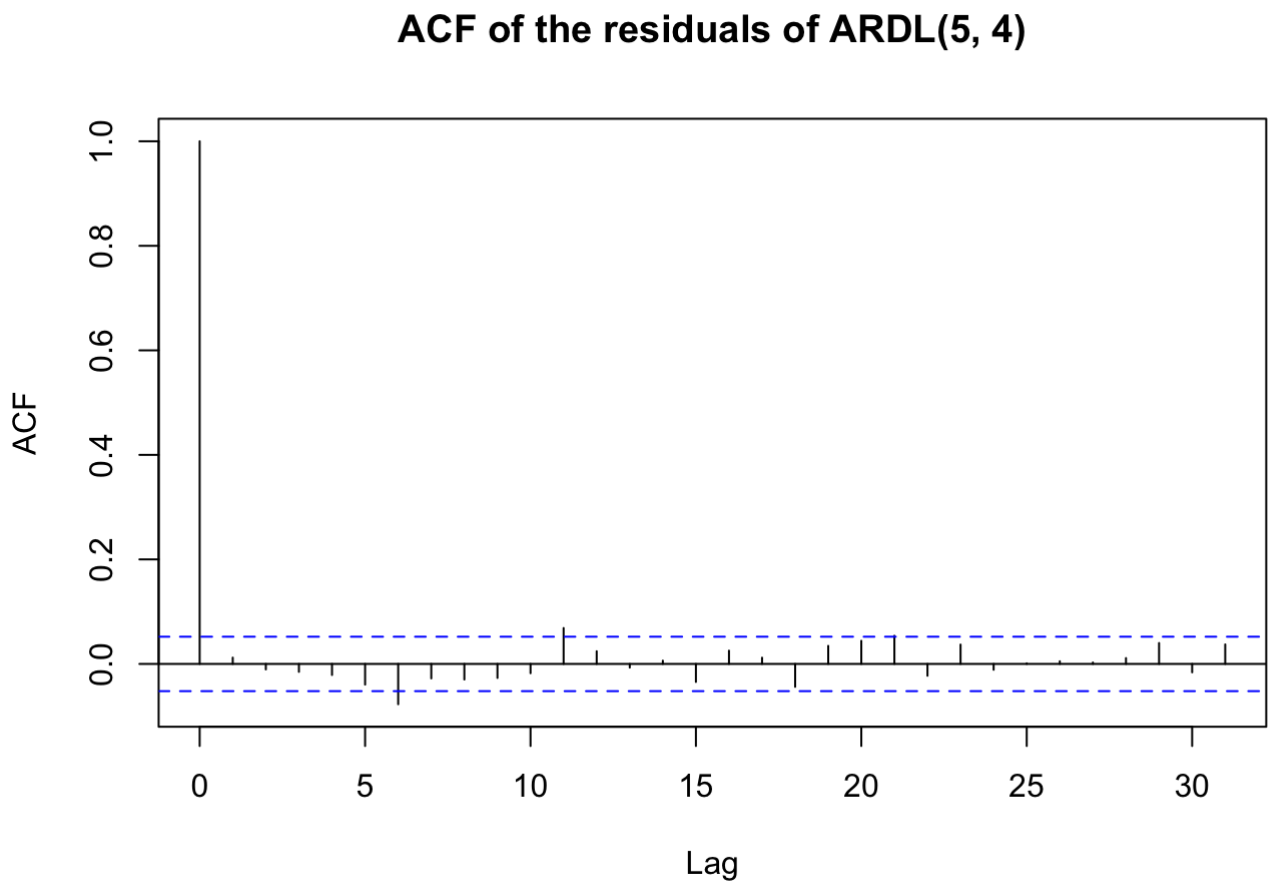
We see that all variables are relatively significant, so $ARDL(5,4)$ is our best fit.

Our $ARDL(5,4)$ model is: $meantemp\_df_t = 0.286003 +$
$$-0.081782 meantemp\_df_{t-1} + -0.178333 meantemp\_df_{t-2} +$$
$$-0.183980 meantemp\_df_{t-3} + -0.074193 meantemp\_df_{t-4} +$$
$$-0.049862 meantemp\_df_{t-5} + -0.136723 humidity\_df_t +$$
$$-0.004584 humidity\_df_{t-1} + -0.028610 humidity\_df_{t-2} +$$
$$-0.028125 humidity\_df_{t-3} + -0.014755 humidity\_df_{t-4}$$

which shows that the change in mean temperature is a model of past values of $meantemp\_df$ and $humidity\_df$.

The impact multiplier, B_0 would be -0.136723 and the first delay multiplier, B_1 would be $-0.00458 + -0.136723 \times -0.081782$ and each multiplier after would be $B_j = B_{j-1} \times -0.081782$. The R^2 is also 0.4752, which isn't very high but it does hold more significant coefficients so we shall continue with this model. Interpreting the coefficients, the intercept is 0.286003 which does not mean that the $meantemp\_df$ starts at 0.286003, it's just the way the equation is forecasted. We see that for the coefficients, most of them are negative implying a negative relationship and we can also estimate the change at certain lags. For example, at lag $t-1$, the change in $meantemp\_df$ would be -0.081782 for every unit of $meantemp\_df_{t-1}$ and -0.004584 with for every unit of $humidity\_df_{t-1}$.

```
acf(mean_humid$residuals, main = "ACF of the residuals of ARDL(5, 4)")
```



We see that there are barely any significant correlations in the errors for our model and although lag 11 does show that there is correlation there, we decided to ignore it since it is at a long lag and is barely passes the test of serial error correlation.

4. Summary

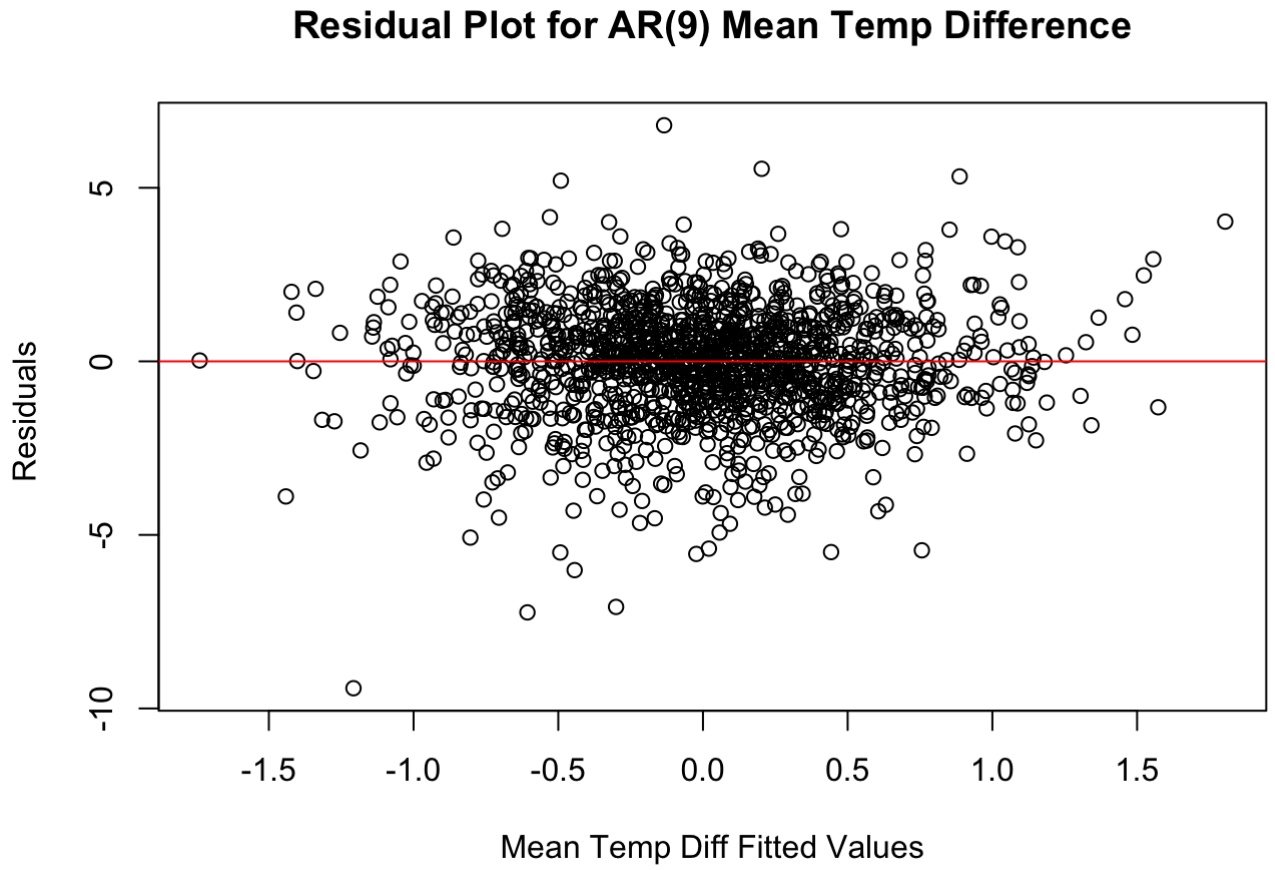
```
AIC(ar_9)
```

```
## [1] 5381.868
```

```
BIC(ar_9)
```

```
## [1] 5439.657
```

```
plot(fitted(ar_9), resid(ar_9), main = "Residual Plot for AR(9) Mean Temp Differenc
e",
     xlab = "Mean Temp Diff Fitted Values", ylab = "Residuals ")
abline(a = 0, b = 0, col = "red")
```



Examining the residual plot for the $AR(9)$ model, we see there there is no pattern, but there is a lot of noise here, which is a good sign it fits relatively well. The residuals are not increasing, meaning that there is no heteroskedasticity present.

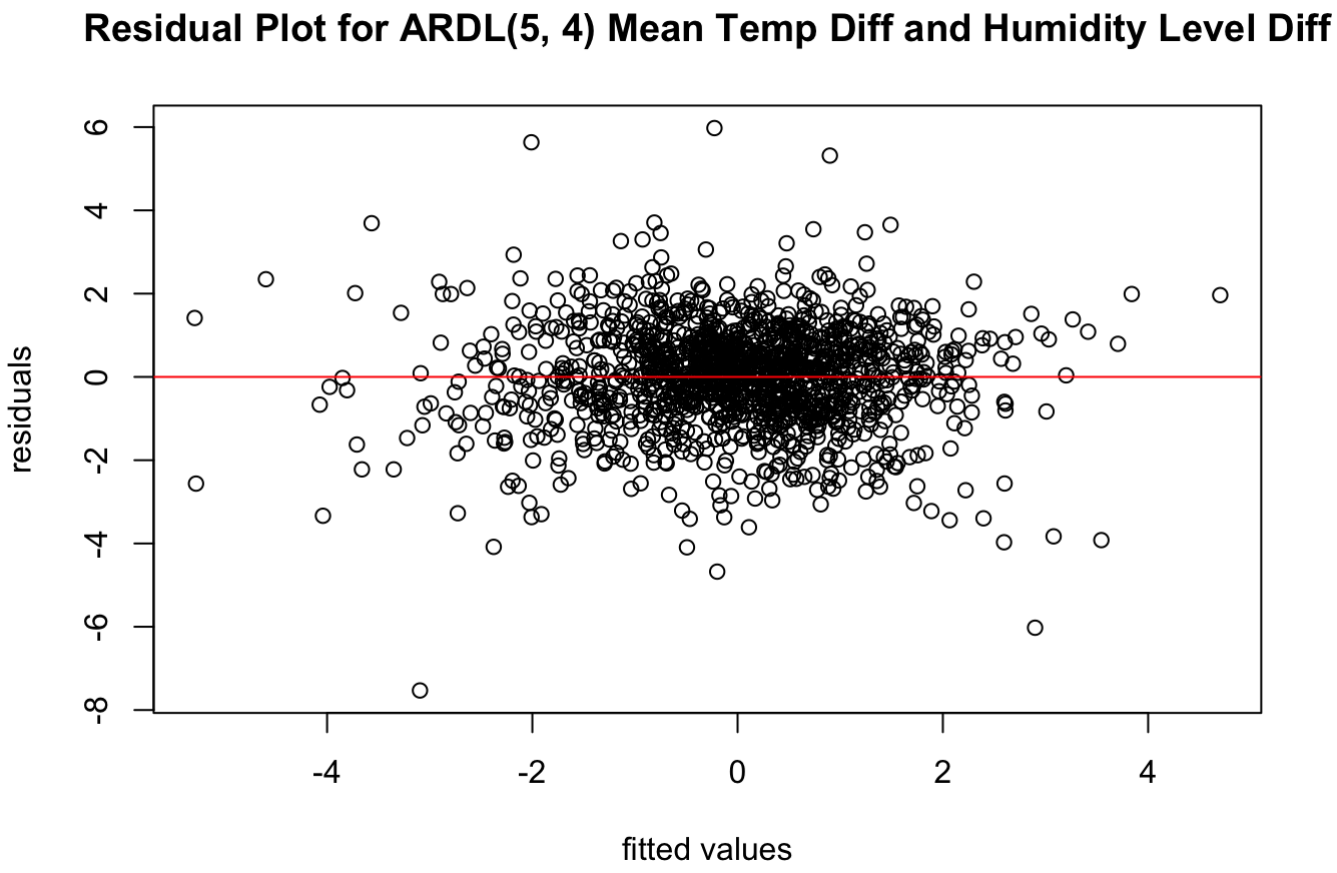
```
AIC(mean_humid)
```

```
## [1] 4613.535
```

```
BIC(mean_humid)
```

```
## [1] 4676.61
```

```
plot(fitted(mean_humid), resid(mean_humid), main = "Residual Plot for ARDL(5, 4) Me
an Temp Diff and Humidity Level Diff", xlab = "fitted values", ylab = "residuals")
abline(a = 0, b = 0, col = "red")
```

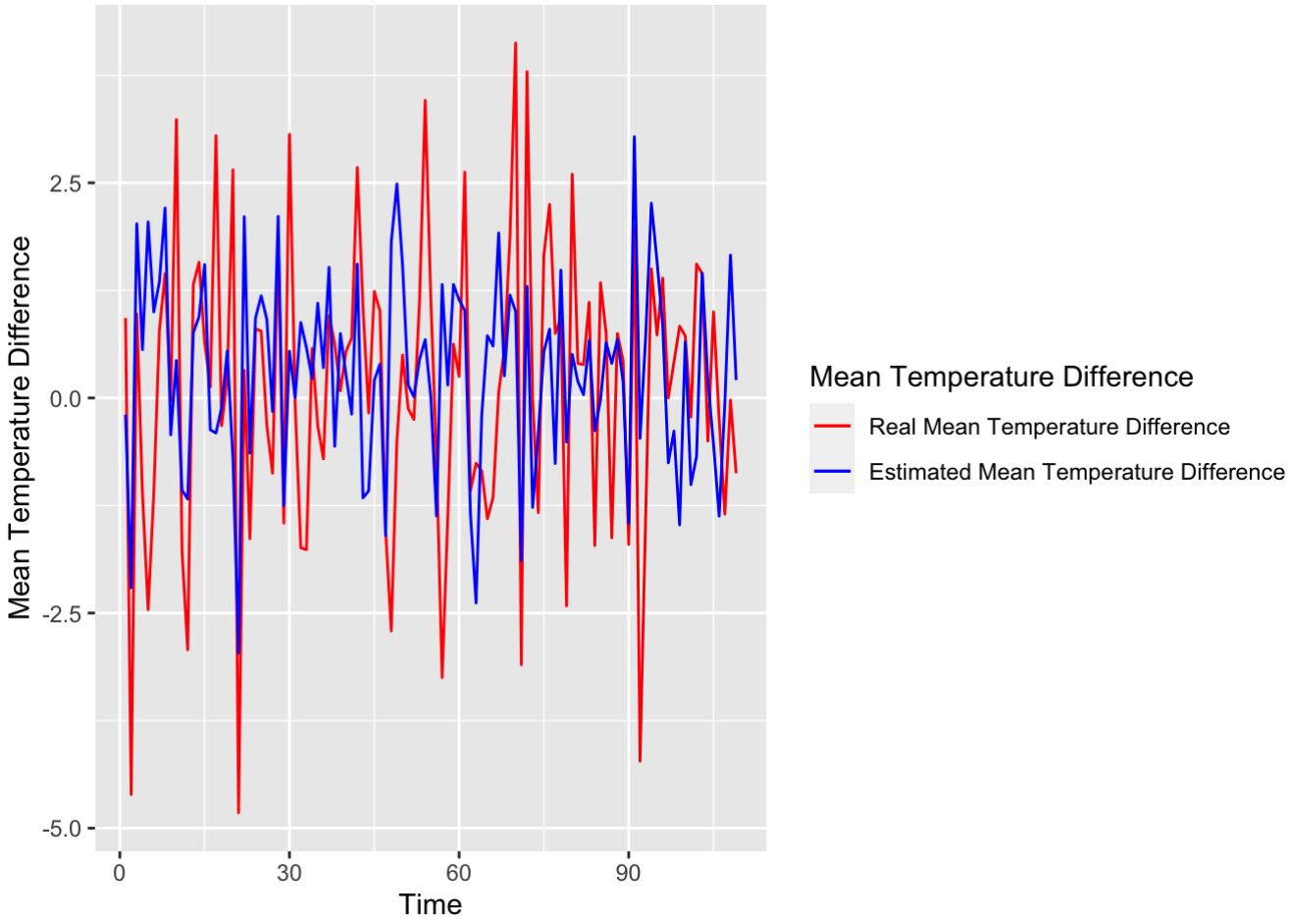


Examining the residual plot for the $ARDL(5,4)$ model, we also see that there there is no pattern and there is a lot of white noise, which is a good sign it fits relatively well. There are also no patterns in the residuals as well, meaning that there is no heteroskedasticity present.

Comparing our AIC and BIC models, we see that the AIC score for our $AR(9)$ $meantemp\_df$ is 5381.868414 and the BIC is 5439.6565882. As for our $ARDL(5,4)$ with $meantemp\_df$ and $humidity\_df$, the AIC is 4613.5349275 and the BIC is 4676.6104944. Since both AIC and BIC score for $ARDL(5,4)$ is lower than the $AR(9)$ model, we can say that the $ARDL(5,4)$ is a better fit.

```
test <- read.csv("DailyDelhiClimateTest.csv")
test$meantemp_df <- c(0, diff(test$meantemp, lag = 1))
test$humidity_df <- c(0, diff(test$humidity, lag = 1))
predicted_values <- function(a, b){
  results <-
    0.286003 + -0.081782*lag(a, 1) + -0.178333*lag(a, 2) + -0.183980*lag(a, 3) +
    -0.074193*lag(a, 4) + -0.049862*lag(a, 5) + -0.136723*b + -0.004584*lag(b, 1) +
    -0.028610*lag(b, 2) + -0.028125*lag(b, 3) + -0.014755*lag(b, 4)
  results
}

test$model_results <- predicted_values(test$meantemp_df, test$humidity_df)

test_results <- test[, c(6, 8)][-(1:5), ]
test_results$t <- 1:nrow(test_results)
```

```
test_results %>%
  ggplot() +
  geom_line(aes(x = t, y = meantemp_df, color = "Real Mean Temperature Difference"
)) +
  geom_line(aes(x = t, y = model_results, color = "Estimated Mean Temperature Diffe
rence")) +
  xlab("Time") +
  ylab("Mean Temperature Difference") +
  scale_color_manual(name = "Mean Temperature Difference",
                     breaks=c("Real Mean Temperature Difference",
                              "Estimated Mean Temperature Difference"),
                     values=c("Real Mean Temperature Difference" = "red",
                              "Estimated Mean Temperature Difference" = "blue"))
```



Results of the model compared with the actual model.

5. Improvements

In observing what we can do to improve the model, we are suggesting that an MA model for the differentiated data would help immensely as observing the PACF and ACF graphs in the `ggtsdisplay`, we saw that it followed the pattern of a MA model. The ACF had sharp cut-off and the PACF should have a gradual decay.

We also observed in the Mallows' CP that `h-w-m`, which stands for $humidity\_df$, $windspeed\_df$, and $meanpressure\_df$, had the lowest score so we can see that a VAR(p) model can also work and may fit the data better.

```
VARselect(climate_change_f[, 6:9], lag.max = 12)
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     11      5      3     11
##
## $criteria
##                  1           2           3           4           5           6
## AIC(n)    8.101387    7.881151    7.772935    7.741076    7.710533    7.701100
## HQ(n)     8.129223    7.931254    7.845307    7.835715    7.827441    7.840275
## SC(n)     8.175875    8.015228    7.966602    7.994332    8.023379    8.073535
## FPE(n) 3299.042523 2646.919579 2375.438684 2300.957297 2231.751721 2210.810340
##                  7           8           9          10          11          12
## AIC(n)    7.690369    7.690773    7.671788    7.664307    7.652709    7.655193
## HQ(n)     7.851812    7.874485    7.877767    7.892554    7.903225    7.927977
## SC(n)     8.122394    8.182388    8.222993    8.275101    8.323093    8.385167
## FPE(n) 2187.231669 2188.140202 2147.019257 2131.054452 2106.526729 2111.818675
```

Using a VAR model, we see that $VAR(11)$ has the lowest AIC score at 7.652709, so a $VAR(11)$ would be best.

```
VAR_11 <- VAR(climate_change_f[, 6:9], p = 11)
VAR_11summ <- summary(VAR_11)
VAR_11summ$varresult$meantemp_df
```

```
##
## Call:
## lm(formula = y ~ -1 + ., data = datamat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3758 -0.8613  0.0874  1.0230  6.2086
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## meantemp_df.l1      -0.1918028  0.0381586  -5.026 5.66e-07 ***
## humidity_df.l1       0.0088528  0.0077772   1.138  0.25519
## windspeed_df.l1     -0.0292010  0.0139211  -2.098  0.03612 *
## meanpressure_df.l1  -0.0254545  0.0295611  -0.861  0.38935
## meantemp_df.l2      -0.2059749  0.0384834  -5.352 1.02e-07 ***
## humidity_df.l2      -0.0046072  0.0078592  -0.586  0.55783
## windspeed_df.l2     -0.0311424  0.0160005  -1.946  0.05182 .
## meanpressure_df.l2  -0.0388271  0.0295571  -1.314  0.18919
## meantemp_df.l3      -0.2028746  0.0393374  -5.157 2.87e-07 ***
## humidity_df.l3       0.0006753  0.0080608   0.084  0.93325
## windspeed_df.l3     -0.0156044  0.0174376  -0.895  0.37101
## meanpressure_df.l3   0.0132273  0.0311600   0.424  0.67127
## meantemp_df.l4      -0.1247149  0.0399659  -3.121  0.00184 **
## humidity_df.l4      -0.0019369  0.0082850  -0.234  0.81519
## windspeed_df.l4     -0.0222179  0.0183620  -1.210  0.22649
## meanpressure_df.l4  -0.0312341  0.0313025  -0.998  0.31855
## meantemp_df.l5      -0.0723127  0.0401744  -1.800  0.07209 .
## humidity_df.l5       0.0089191  0.0083651   1.066  0.28651
## windspeed_df.l5     -0.0297808  0.0188368  -1.581  0.11411
## meanpressure_df.l5  -0.0063451  0.0313213  -0.203  0.83949
## meantemp_df.l6      -0.0746924  0.0402371  -1.856  0.06363 .
## humidity_df.l6      -0.0028000  0.0083531  -0.335  0.73752
## windspeed_df.l6      0.0108347  0.0190617   0.568  0.56986
## meanpressure_df.l6   0.0080565  0.0313501   0.257  0.79723
## meantemp_df.l7      -0.0404370  0.0401491  -1.007  0.31403
## humidity_df.l7       0.0083489  0.0083543   0.999  0.31780
## windspeed_df.l7     -0.0080107  0.0188646  -0.425  0.67117
## meanpressure_df.l7   0.0048370  0.0312518   0.155  0.87702
## meantemp_df.l8      -0.0719847  0.0399012  -1.804  0.07144 .
## humidity_df.l8      -0.0123895  0.0082200  -1.507  0.13198
## windspeed_df.l8      0.0094975  0.0183807   0.517  0.60544
## meanpressure_df.l8   0.0280097  0.0311818   0.898  0.36920
## meantemp_df.l9      -0.0625861  0.0391615  -1.598  0.11024
## humidity_df.l9       0.0017949  0.0080146   0.224  0.82283
## windspeed_df.l9      0.0077417  0.0174389   0.444  0.65716
## meanpressure_df.l9  -0.0129408  0.0309030  -0.419  0.67546
## meantemp_df.l10     -0.0423848  0.0383238  -1.106  0.26894
## humidity_df.l10      0.0001174  0.0078311   0.015  0.98804
## windspeed_df.l10     0.0029343  0.0159064   0.184  0.85367
## meanpressure_df.l10 -0.0041647  0.0294295  -0.142  0.88749
## meantemp_df.l11      0.0463401  0.0379003   1.223  0.22166
## humidity_df.l11      0.0102829  0.0076660   1.341  0.18003
```

```
## windspeed_df.l11     0.0054985  0.0137375   0.400  0.68903
## meanpressure_df.l11  0.0054660  0.0293846   0.186  0.85246
## const                0.0004090  0.0429168   0.010  0.99240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.612 on 1366 degrees of freedom
## Multiple R-squared:  0.1064, Adjusted R-squared:  0.07762
## F-statistic: 3.697 on 44 and 1366 DF,  p-value: 1.92e-14
```

Looking at the $VAR(11)$ model, we can see the summary statistics but the R^2 is relatively low at 0.1064 and many of the coefficients is insignificant. We would have to do more testing and modeling in order to see which model is truly the best model.

But first, we test for reverse causality, where our explanatory variable is actually caused by our dependent variable.

```
grangertest(climate_change_f$humidity_df ~ climate_change_f$meantemp_df, order = 11
)
```

```
## Granger causality test
##
## Model 1: climate_change_f$humidity_df ~ Lags(climate_change_f$humidity_df, 1:11)
+ Lags(climate_change_f$meantemp_df, 1:11)
## Model 2: climate_change_f$humidity_df ~ Lags(climate_change_f$humidity_df, 1:11)
##   Res.Df  Df      F  Pr(>F)
## 1   1388
## 2   1399 -11 1.9202 0.03305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we reject the null that *meantemp_df* does not granger-cause *humidity_df* since the p value is 0.03305, much lower than the significant level at 0.05, so it is significant that *meantemp_df* explains *humidity_df* so perhaps a model trying to predict *meantemp_df* from *humidity_df* may not be ideal, but rather *humidity_df* is better predicted by *meantemp_df*.

```
grangertest(climate_change_f$meantemp_df ~ climate_change_f$humidity_df, order = 11
)
```

```
## Granger causality test
##
## Model 1: climate_change_f$meantemp_df ~ Lags(climate_change_f$meantemp_df, 1:11)
+ Lags(climate_change_f$humidity_df, 1:11)
## Model 2: climate_change_f$meantemp_df ~ Lags(climate_change_f$meantemp_df, 1:11)
##   Res.Df  Df      F Pr(>F)
## 1   1388
## 2   1399 -11 1.3897 0.1712
```

Since our p value is 0.1712, which is above the significance level of 0.05, we fail to reject the null $humidity\_df$ does not granger-cause $meantemp\_df$. Therefore, our model could be improved by predicting $humidity\_df$ from $meantemp\_df$ and use that model to reverse calculate the average temperature change.

Base on all of our resultse, we believe we can safely say that our $ARDL(5, 4)$ regressed on the $meantemp\_df$ lags and the $humidity\_df$ lags may not be the best model in predicting the difference in $meantemp$ in the next period and we may want to either consider a MA or $VAR(11)$ model. We also need to test a reverse casuality model where $humidity\_df$ is predicted from $meantemp\_df$.