Sydney Rainfall Forecast

Logan Sartain

12/7/2022

PROJECT SETUP

Install Required Libraries (If Necessary)

```
install.packages("fpp3", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'fpp3' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

```
install.packages("lubridate", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'lubridate' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'lubridate'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\logan\AppData\Local\R\win-library\4.2\00LOCK\lubridate\libs\x64\lubridate.dll
## to
## C:\Users\logan\AppData\Local\R\win-library\4.2\lubridate\libs\x64\lubridate.dll:
## Permission denied
```

```
## Warning: restored 'lubridate'
```

```
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

```
install.packages("fastDummies", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'fastDummies' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

```
install.packages("gplots", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'gplots' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

```
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

```
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/logan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\logan\AppData\Local\Temp\RtmpQ3abA8\downloaded_packages
```

Load Required Libraries

```
library(fpp3)
```

```
## ── Attaching packages ──────────────────────────────────────────── fpp3 0.4.0 ──
```

```
## ✓ tibble      3.1.8      ✓ tsibble     1.1.3
## ✓ dplyr       1.0.10     ✓ tsibbledata 0.4.1
## ✓ tidyr       1.2.1      ✓ feasts      0.3.0
## ✓ lubridate   1.9.0      ✓ fable       0.3.2
## ✓ ggplot2     3.4.0
```

```
## ── Conflicts ──────────────────────────────────────────────── fpp3_conflicts ──
## ✕ lubridate::date()    masks base::date()
## ✕ dplyr::filter()      masks stats::filter()
## ✕ tsibble::intersect() masks base::intersect()
## ✕ tsibble::interval()  masks lubridate::interval()
## ✕ dplyr::lag()         masks stats::lag()
## ✕ tsibble::setdiff()   masks base::setdiff()
## ✕ tsibble::union()     masks base::union()
```

```
library(lubridate)
library(fastDummies)
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching packages
## ───────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ readr   2.1.3     ✓ stringr 1.5.0
## ✓ purrr   0.3.5     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ lubridate::as.difftime() masks base::as.difftime()
## ✕ lubridate::date()        masks base::date()
## ✕ dplyr::filter()          masks stats::filter()
## ✕ tsibble::intersect()     masks lubridate::intersect(), base::intersect()
## ✕ tsibble::interval()      masks lubridate::interval()
## ✕ dplyr::lag()             masks stats::lag()
## ✕ tsibble::setdiff()       masks lubridate::setdiff(), base::setdiff()
## ✕ tsibble::union()         masks lubridate::union(), base::union()
```

Import Dataset

```
options(max.print = 175)
url <- "https://github.com/LoganSartain/Final-Project-Bana-4090/blob/main/weatherAUS.csv?raw=tru
e"
AUS <- read.csv(url, header = TRUE)
print(AUS)
```

```
##         Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir
## 1 2008-12-01   Albury    13.4    22.9      0.6          NA       NA           W
## 2 2008-12-02   Albury     7.4    25.1      0.0          NA       NA         WNW
## 3 2008-12-03   Albury    12.9    25.7      0.0          NA       NA         WSW
## 4 2008-12-04   Albury     9.2    28.0      0.0          NA       NA          NE
## 5 2008-12-05   Albury    17.5    32.3      1.0          NA       NA           W
## 6 2008-12-06   Albury    14.6    29.7      0.2          NA       NA         WNW
## 7 2008-12-07   Albury    14.3    25.0      0.0          NA       NA           W
##   WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am
## 1            44          W        WNW           20           24          71
## 2            44        NNW        WSW            4           22          44
## 3            46          W        WSW           19           26          38
## 4            24         SE          E           11            9          45
## 5            41        ENE         NW            7           20          82
## 6            56          W          W           19           24          55
## 7            50         SW          W           20           24          49
##   Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm
## 1          22      1007.7      1007.1        8       NA    16.9    21.8
## 2          25      1010.6      1007.8       NA       NA    17.2    24.3
## 3          30      1007.6      1008.7       NA        2    21.0    23.2
## 4          16      1017.6      1012.8       NA       NA    18.1    26.5
## 5          33      1010.8      1006.0        7        8    17.8    29.7
## 6          23      1009.2      1005.4       NA       NA    20.6    28.9
## 7          19      1009.6      1008.2        1       NA    18.1    24.6
##   RainToday RainTomorrow
## 1        No           No
## 2        No           No
## 3        No           No
## 4        No           No
## 5        No           No
## 6        No           No
## 7        No           No
##  [ reached 'max' / getOption("max.print") -- omitted 145453 rows ]
```

INTRODUCTION

This dataset has 10 years of weather data taken around multiple locations in Australia. I decided to focus on Temperature in Sydney, Australia. It includes many weather variables that would be useful in predicting and forecasting temperature.

The dataset is from Kaggle.

This dataset was created by Joe Young and Adam Young. They gathered data from the Australia government and compiled it to create this dataset.

Index: Date

Key: Location

Forecast Variable: MaxTemp

Predictor Variables: MinTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainToday, and RainTomorrow.

I chose this dataset because I have always found weather and storms interesting. I would love to be able to predict the weather for a meteorologist/news station as a future job. It is also interesting to me how hard it can be to accurately predict the weather so I thought it would be cool to see how accurate I could be.

The forecast on this data can be leveraged to make better decisions by a multitude of different organizations in Australia. An obvious one would be weather/news stations making more accurate predictions on temperature but, this forecast could also be useful for farmers, sporting events, wedding venues, outdoor concert coordinators,Uber drivers, restaurants with outdoor dining, and airlines just to name a few. This would allow all of these different types of organizations to plan better according to the weather. For example, a restaurant may want to schedule less waiters on a night where it is going to be too hot or too cold because they won't need anyone for outdoor dining. Or a wedding venue may need to prepare a backup plan in case of extreme heat or cold. The forecast would overall allow for better planning and decision making in this regard.

DATA WRANGLING

Convert to a tsibble

```
AUS$Date <- as.Date(AUS$Date , format="%Y-%m-%d")

AUS <- AUS %>%
  as_tsibble(index = Date, key = Location)
```

Deal with Missing Data

```
summary(Filter(is.numeric, AUS))
```

```
##     MinTemp          MaxTemp          Rainfall          Evaporation    
##  Min.   :-8.50   Min.   :-4.80   Min.   :  0.000   Min.   :  0.00  
##  1st Qu.: 7.60   1st Qu.:17.90   1st Qu.:  0.000   1st Qu.:  2.60  
##  Median :12.00   Median :22.60   Median :  0.000   Median :  4.80  
##  Mean   :12.19   Mean   :23.22   Mean   :  2.361   Mean   :  5.47  
##  3rd Qu.:16.90   3rd Qu.:28.20   3rd Qu.:  0.800   3rd Qu.:  7.40  
##  Max.   :33.90   Max.   :48.10   Max.   :371.000   Max.   :145.00  
##  NA's   :1485    NA's   :1261    NA's   :3261      NA's   :62790   
##     Sunshine     WindGustSpeed     WindSpeed9am     WindSpeed3pm  
##  Min.   : 0.00   Min.   :  6.00   Min.   :  0.00   Min.   : 0.00  
##  1st Qu.: 4.80   1st Qu.: 31.00   1st Qu.:  7.00   1st Qu.:13.00  
##  Median : 8.40   Median : 39.00   Median : 13.00   Median :19.00  
##  Mean   : 7.61   Mean   : 40.03   Mean   : 14.04   Mean   :18.66  
##  3rd Qu.:10.60   3rd Qu.: 48.00   3rd Qu.: 19.00   3rd Qu.:24.00  
##  Max.   :14.50   Max.   :135.00   Max.   :130.00   Max.   :87.00  
##  NA's   :69835   NA's   :10263    NA's   :1767     NA's   :3062   
##   Humidity9am      Humidity3pm      Pressure9am      Pressure3pm    
##  Min.   :  0.00   Min.   :  0.00   Min.   : 980.5   Min.   : 977.1  
##  1st Qu.: 57.00   1st Qu.: 37.00   1st Qu.:1012.9   1st Qu.:1010.4  
##  Median : 70.00   Median : 52.00   Median :1017.6   Median :1015.2  
##  Mean   : 68.88   Mean   : 51.54   Mean   :1017.6   Mean   :1015.3  
##  3rd Qu.: 83.00   3rd Qu.: 66.00   3rd Qu.:1022.4   3rd Qu.:1020.0  
##  Max.   :100.00   Max.   :100.00   Max.   :1041.0   Max.   :1039.6  
##  NA's   :2654     NA's   :4507     NA's   :15065    NA's   :15028   
##     Cloud9am        Cloud3pm         Temp9am         Temp3pm     
##  Min.   :0.00    Min.   :0.00    Min.   :-7.20   Min.   :-5.40  
##  1st Qu.:1.00    1st Qu.:2.00    1st Qu.:12.30   1st Qu.:16.60  
##  Median :5.00    Median :5.00    Median :16.70   Median :21.10  
##  Mean   :4.45    Mean   :4.51    Mean   :16.99   Mean   :21.68  
##  3rd Qu.:7.00    3rd Qu.:7.00    3rd Qu.:21.60   3rd Qu.:26.40  
##  Max.   :9.00    Max.   :9.00    Max.   :40.20   Max.   :46.70  
##  NA's   :55888   NA's   :59358   NA's   :1767    NA's   :3609   
```

```
# Replacing missing data with the median value of the predictor variable for numeric
AUS$MinTemp[is.na(AUS$MinTemp)] <- median(AUS$MinTemp,na.rm=TRUE)
AUS$MaxTemp[is.na(AUS$MaxTemp)] <- median(AUS$MaxTemp,na.rm=TRUE)
AUS$Rainfall[is.na(AUS$Rainfall)] <- median(AUS$Rainfall,na.rm=TRUE)
AUS$Evaporation[is.na(AUS$Evaporation)] <- median(AUS$Evaporation,na.rm=TRUE)
AUS$Sunshine[is.na(AUS$Sunshine)] <- median(AUS$Sunshine,na.rm=TRUE)
AUS$WindGustSpeed[is.na(AUS$WindGustSpeed)] <- median(AUS$WindGustSpeed,na.rm=TRUE)
AUS$WindSpeed9am[is.na(AUS$WindSpeed9am)] <- median(AUS$WindSpeed9am,na.rm=TRUE)
AUS$WindSpeed3pm[is.na(AUS$WindSpeed3pm)] <- median(AUS$WindSpeed3pm,na.rm=TRUE)
AUS$Humidity9am[is.na(AUS$Humidity9am)] <- median(AUS$Humidity9am,na.rm=TRUE)
AUS$Humidity3pm[is.na(AUS$Humidity3pm)] <- median(AUS$Humidity3pm,na.rm=TRUE)
AUS$Pressure9am[is.na(AUS$Pressure9am)] <- median(AUS$Pressure9am,na.rm=TRUE)
AUS$Pressure3pm[is.na(AUS$Pressure3pm)] <- median(AUS$Pressure3pm,na.rm=TRUE)
AUS$Cloud9am[is.na(AUS$Cloud9am)] <- median(AUS$Cloud9am,na.rm=TRUE)
AUS$Cloud3pm[is.na(AUS$Cloud3pm)] <- median(AUS$Cloud3pm,na.rm=TRUE)
AUS$Temp9am[is.na(AUS$Temp9am)] <- median(AUS$Temp9am,na.rm=TRUE)
AUS$Temp3pm[is.na(AUS$Temp3pm)] <- median(AUS$Temp3pm,na.rm=TRUE)

summary(Filter(is.numeric, AUS))
```

```
##     MinTemp         MaxTemp         Rainfall        Evaporation
##  Min.   :-8.50   Min.   :-4.80   Min.   :  0.000   Min.   :  0.00
##  1st Qu.: 7.70   1st Qu.:18.00   1st Qu.:  0.000   1st Qu.:  4.00
##  Median :12.00   Median :22.60   Median :  0.000   Median :  4.80
##  Mean   :12.19   Mean   :23.22   Mean   :  2.308   Mean   :  5.18
##  3rd Qu.:16.80   3rd Qu.:28.20   3rd Qu.:  0.600   3rd Qu.:  5.20
##  Max.   :33.90   Max.   :48.10   Max.   :371.000   Max.   :145.00
##     Sunshine      WindGustSpeed     WindSpeed9am     WindSpeed3pm
##  Min.   : 0.00   Min.   :  6.00   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 8.20   1st Qu.: 31.00   1st Qu.:  7.00   1st Qu.:13.00
##  Median : 8.40   Median : 39.00   Median : 13.00   Median :19.00
##  Mean   : 7.99   Mean   : 39.96   Mean   : 14.03   Mean   :18.67
##  3rd Qu.: 8.70   3rd Qu.: 46.00   3rd Qu.: 19.00   3rd Qu.:24.00
##  Max.   :14.50   Max.   :135.00   Max.   :130.00   Max.   :87.00
##   Humidity9am     Humidity3pm      Pressure9am      Pressure3pm
##  Min.   :  0.0   Min.   :  0.00   Min.   : 980.5   Min.   : 977.1
##  1st Qu.: 57.0   1st Qu.: 37.00   1st Qu.:1013.5   1st Qu.:1011.1
##  Median : 70.0   Median : 52.00   Median :1017.6   Median :1015.2
##  Mean   : 68.9   Mean   : 51.55   Mean   :1017.6   Mean   :1015.3
##  3rd Qu.: 83.0   3rd Qu.: 65.00   3rd Qu.:1021.8   3rd Qu.:1019.4
##  Max.   :100.0   Max.   :100.00   Max.   :1041.0   Max.   :1039.6
##     Cloud9am       Cloud3pm       Temp9am         Temp3pm
##  Min.   :0.00   Min.   :0.00   Min.   :-7.20   Min.   :-5.40
##  1st Qu.:3.00   1st Qu.:4.00   1st Qu.:12.30   1st Qu.:16.70
##  Median :5.00   Median :5.00   Median :16.70   Median :21.10
##  Mean   :4.66   Mean   :4.71   Mean   :16.99   Mean   :21.67
##  3rd Qu.:6.00   3rd Qu.:6.00   3rd Qu.:21.50   3rd Qu.:26.20
##  Max.   :9.00   Max.   :9.00   Max.   :40.20   Max.   :46.70
```

```r
# Removing missing data entirely for RainToday and RainTomorrow

AUS <- AUS %>%
  mutate(RainToday  = ifelse(is.na(RainToday), "Unknown", RainToday))

AUS <- AUS %>%
  mutate(RainTomorrow  = ifelse(is.na(RainTomorrow), "Unknown", RainTomorrow))

summary(AUS)
```

```
##       Date              Location            MinTemp         MaxTemp     
##  Min.   :2007-11-01   Length:145460      Min.   :-8.50   Min.   :-4.80  
##  1st Qu.:2011-01-11   Class :character   1st Qu.: 7.70   1st Qu.:18.00  
##  Median :2013-06-02   Mode  :character   Median :12.00   Median :22.60  
##  Mean   :2013-04-04                      Mean   :12.19   Mean   :23.22  
##  3rd Qu.:2015-06-14                      3rd Qu.:16.80   3rd Qu.:28.20  
##  Max.   :2017-06-25                      Max.   :33.90   Max.   :48.10  
##     Rainfall         Evaporation        Sunshine      WindGustDir       
##  Min.   :  0.000   Min.   :  0.00   Min.   : 0.00   Length:145460      
##  1st Qu.:  0.000   1st Qu.:  4.00   1st Qu.: 8.20   Class :character   
##  Median :  0.000   Median :  4.80   Median : 8.40   Mode  :character   
##  Mean   :  2.308   Mean   :  5.18   Mean   : 7.99                      
##  3rd Qu.:  0.600   3rd Qu.:  5.20   3rd Qu.: 8.70                      
##  Max.   :371.000   Max.   :145.00   Max.   :14.50                      
##  WindGustSpeed     WindDir9am         WindDir3pm         WindSpeed9am   
##  Min.   :  6.00   Length:145460      Length:145460      Min.   :  0.00  
##  1st Qu.: 31.00   Class :character   Class :character   1st Qu.:  7.00  
##  Median : 39.00   Mode  :character   Mode  :character   Median : 13.00  
##  Mean   : 39.96                                         Mean   : 14.03  
##  3rd Qu.: 46.00                                         3rd Qu.: 19.00  
##  Max.   :135.00                                         Max.   :130.00  
##   WindSpeed3pm    Humidity9am     Humidity3pm      Pressure9am    
##  Min.   : 0.00   Min.   :  0.0   Min.   :  0.00   Min.   : 980.5  
##  1st Qu.:13.00   1st Qu.: 57.0   1st Qu.: 37.00   1st Qu.:1013.5  
##  Median :19.00   Median : 70.0   Median : 52.00   Median :1017.6  
##  Mean   :18.67   Mean   : 68.9   Mean   : 51.55   Mean   :1017.6  
##  3rd Qu.:24.00   3rd Qu.: 83.0   3rd Qu.: 65.00   3rd Qu.:1021.8  
##  Max.   :87.00   Max.   :100.0   Max.   :100.00   Max.   :1041.0  
##   Pressure3pm        Cloud9am       Cloud3pm       Temp9am         Temp3pm     
##  Min.   : 977.1   Min.   :0.00   Min.   :0.00   Min.   :-7.20   Min.   :-5.40  
##  1st Qu.:1011.1   1st Qu.:3.00   1st Qu.:4.00   1st Qu.:12.30   1st Qu.:16.70  
##  Median :1015.2   Median :5.00   Median :5.00   Median :16.70   Median :21.10  
##  Mean   :1015.3   Mean   :4.66   Mean   :4.71   Mean   :16.99   Mean   :21.67  
##  3rd Qu.:1019.4   3rd Qu.:6.00   3rd Qu.:6.00   3rd Qu.:21.50   3rd Qu.:26.20  
##  Max.   :1039.6   Max.   :9.00   Max.   :9.00   Max.   :40.20   Max.   :46.70  
##   RainToday         RainTomorrow      
##  Length:145460      Length:145460     
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
## 
```

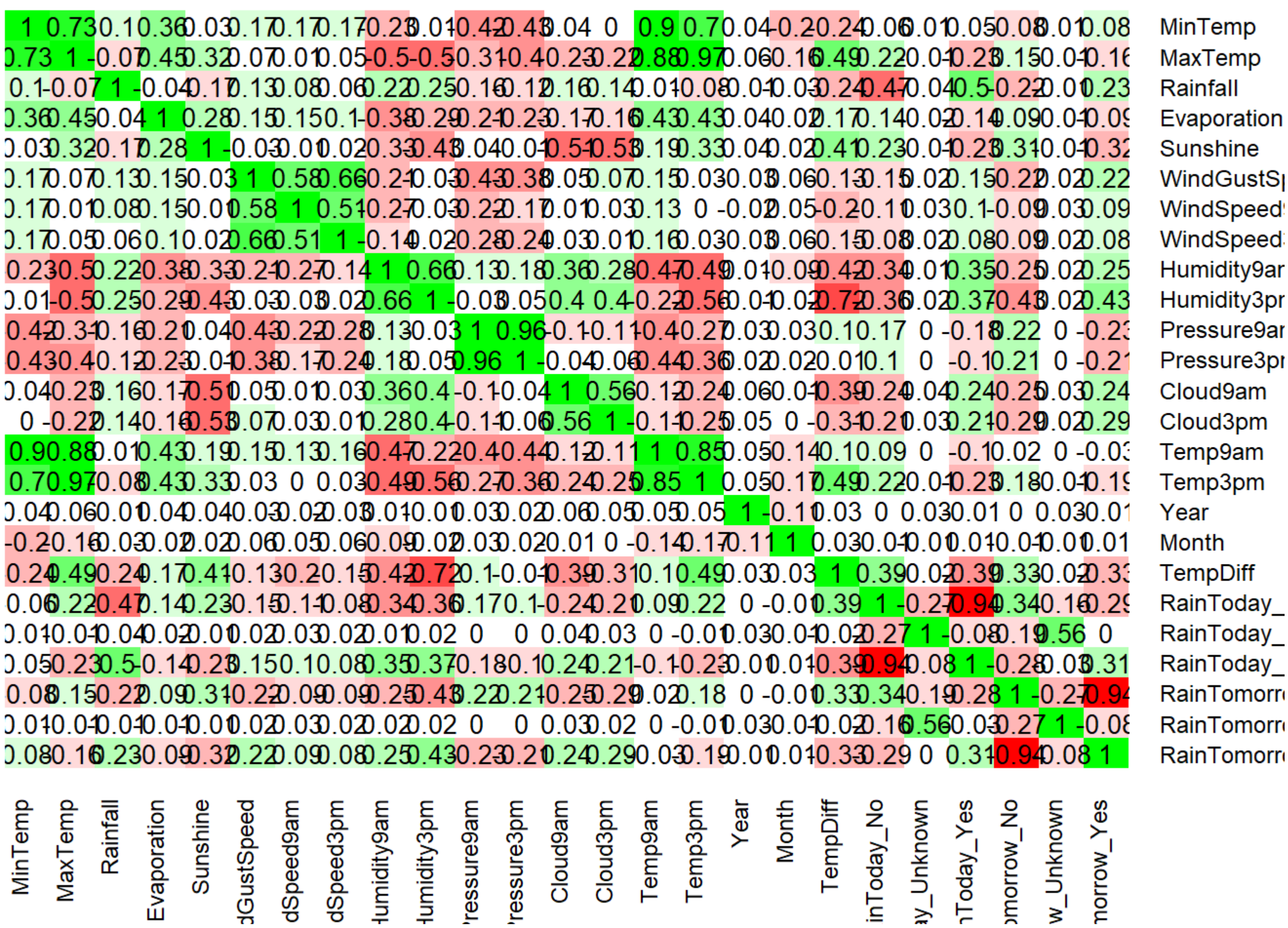
```
# Remove variables WindGustDir, WindDir9am, and WindDir3pm

AUS <- AUS[,!names(AUS) %in% c("WindGustDir", "WindDir9am", "WindDir3pm")]
```

Create New Variables to aid in Forecasting

```
AUS$Year <- year(ymd(AUS$Date)) # Add Year Column
AUS$Month <- month(ymd(AUS$Date)) # Add Month Column
AUS2 <- AUS %>% mutate(TempDiff = MaxTemp - MinTemp) # Temperature Difference Variable
```

Aggregate time series to desired format for forecasting

```
# Create Dummy Variables for RainToday and RainTomorrow

AUS3 <- dummy_cols(AUS2,
                      select_columns = c("RainToday","RainTomorrow"),
                      remove_selected_columns = TRUE)

# Checking for variables with autocorrelation to see if we want to remove any

colfunc <- colorRampPalette(c("red","white","green"))
heatmap.2(cor(Filter(is.numeric, AUS3), use = "complete.obs"), Rowv = FALSE,
          Colv = FALSE, dendrogram = "none", lwid=c(0.1,4), lhei=c(0.1,4),
          col = colfunc(15),
          cellnote = round(cor(Filter(is.numeric, AUS3), use = "complete.obs"),2),
          notecol = "black", key = FALSE, trace = 'none')
```

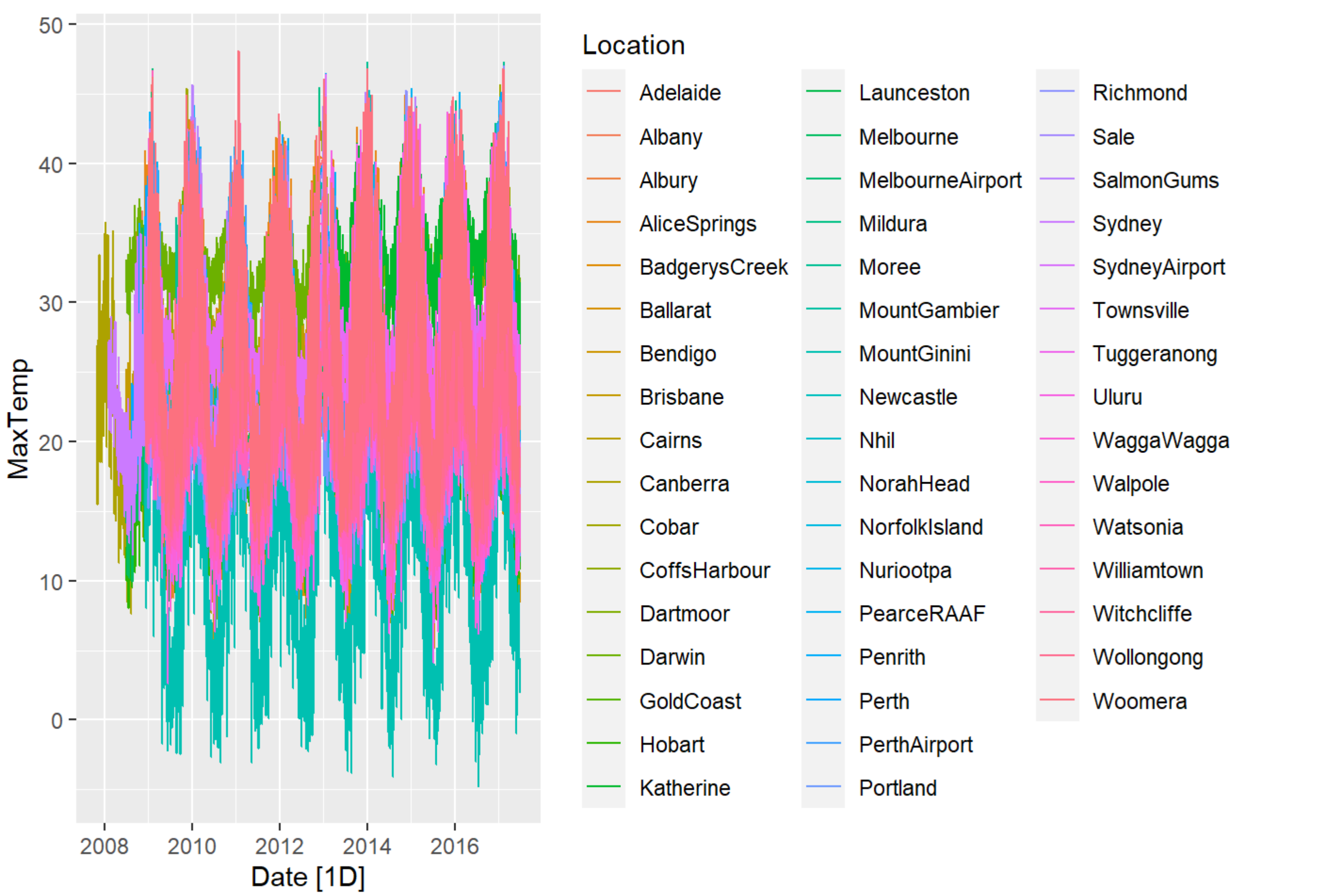


```
# Look to see which variables are highly correlated with MaxTemp.
# MinTemp, Temp9am, and Temp3pm are all highly positively correlated with MaxTemp
# We will remove these three variables as they may likely cause problems with autocorrelation.

drop <- c("MinTemp","Temp9am", "Temp3pm")
AUS4 = AUS3[,!(names(AUS3) %in% drop)]

# Convert processed dataset to tsibble again

AUS_Final <- AUS4 %>%
  as_tsibble(index = Date, key = Location)
```

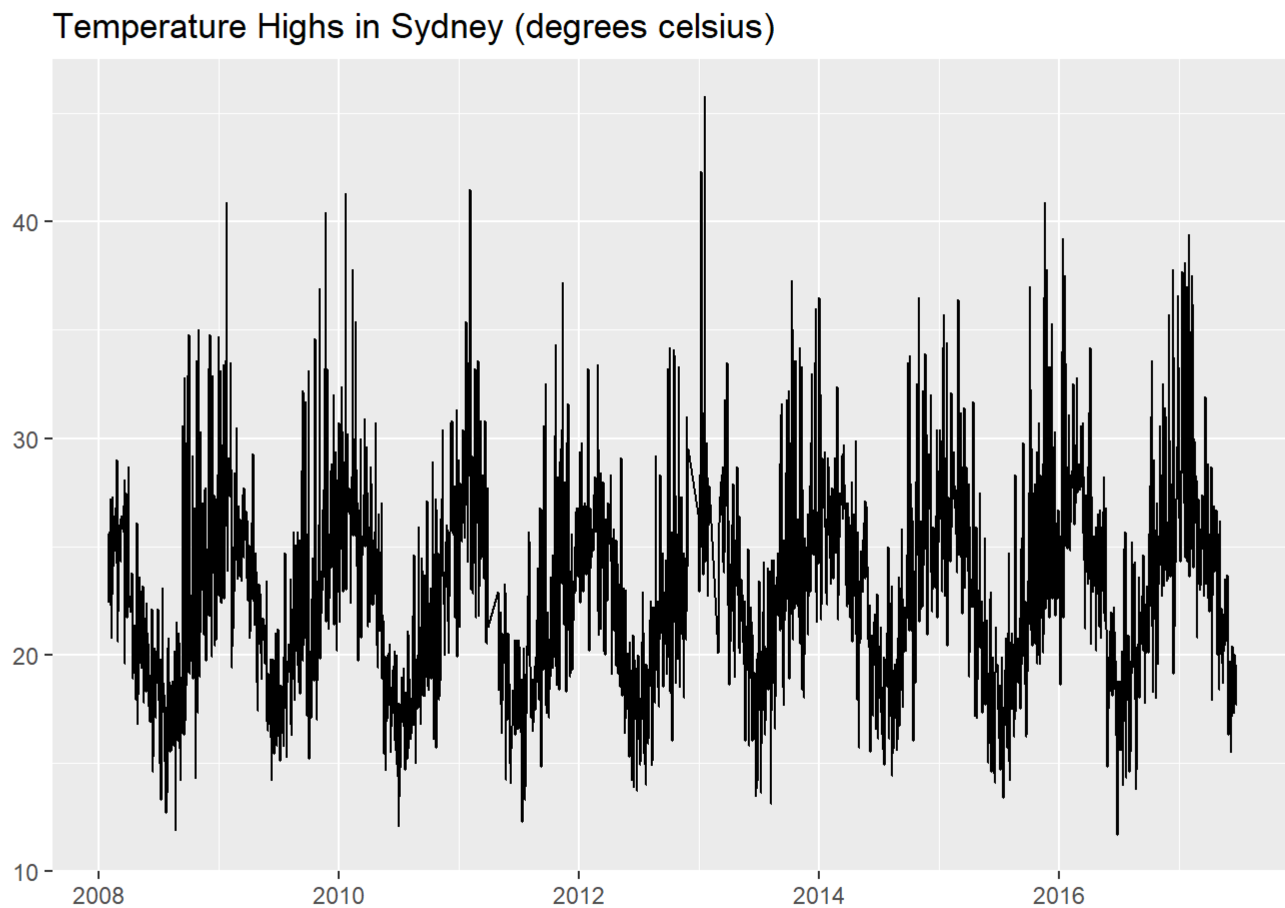
EXPLORATORY ANALYSIS AND VISUALIZATION FOR THE DATASET

Visualize the dataset and comment on characteristics of time series

```
AUS_Final %>% autoplot(MaxTemp)
```



```
# Since there are so many different locations, it makes the plot hard to read. To fix this probl
em
# we will turn our focus on the largest city, Sydney, to further look for seasonality.

Sydney <- AUS_Final %>%
  filter(Location == "Sydney")

Sydney %>%
  autoplot(MaxTemp) + labs(title = "Temperature Highs in Sydney (degrees celsius)", x = " ", y =
" ")
```



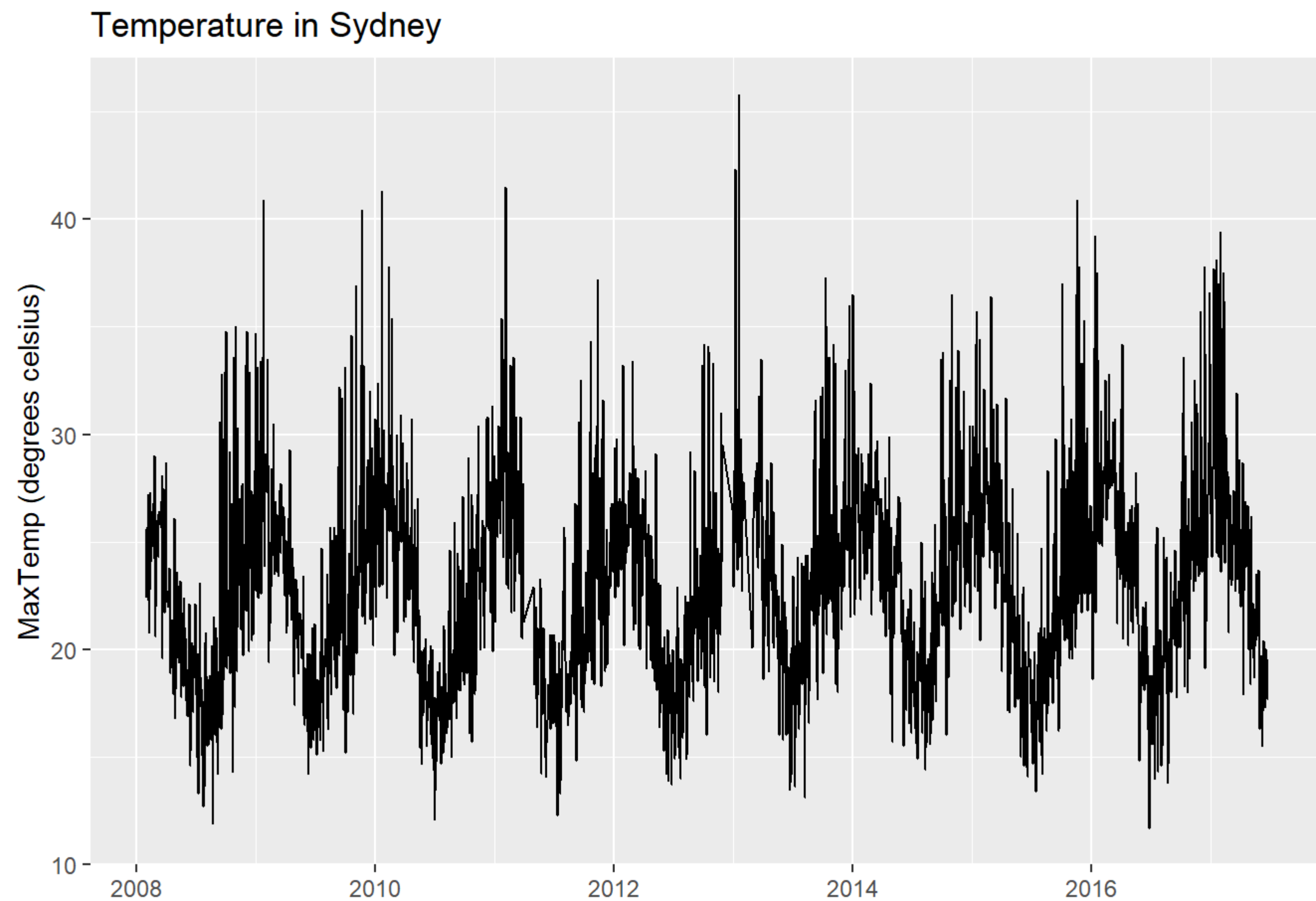
Comment on any anomalies in the data

It looks like there is a huge spike upwards in temperature in Jan 2013. There are also some abnormally low drops in temperature in may/june of 2016.

Describe trend/seasonality/cycles with supporting charts:

Trend

```
Sydney %>%
  autoplot(MaxTemp) + labs(y = "MaxTemp (degrees celsius)",x = " ", title = "Temperature in Sydn
ey")
```

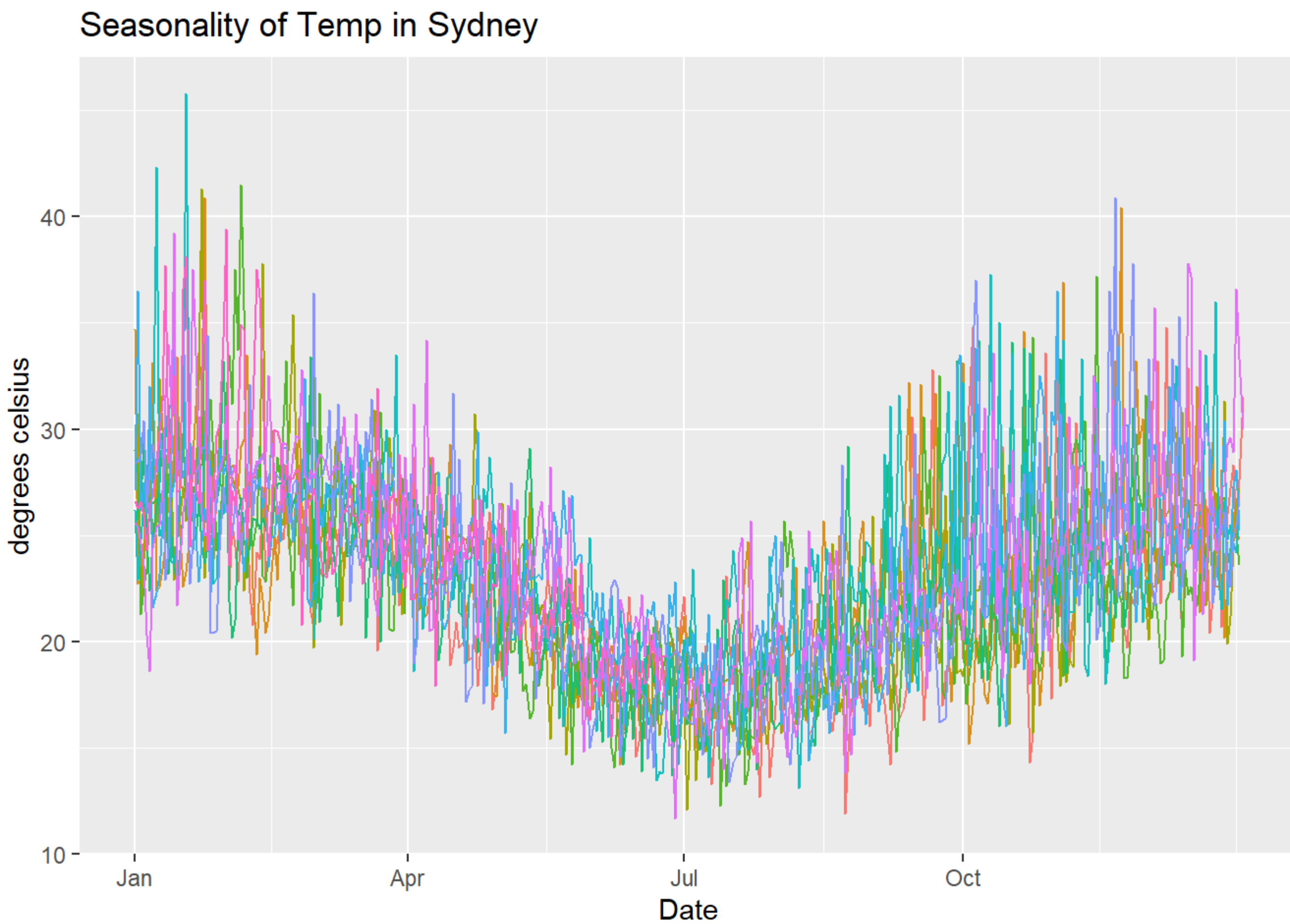


There is no apparent trend in Temperature.

Seasonality

```
Syd_Fill <- Sydney %>% fill_gaps()

Syd_Fill %>% gg_season(MaxTemp, period = "year") +
  theme(legend.position = "none") +
  labs(y="degrees celsius", title="Seasonality of Temp in Sydney")
```

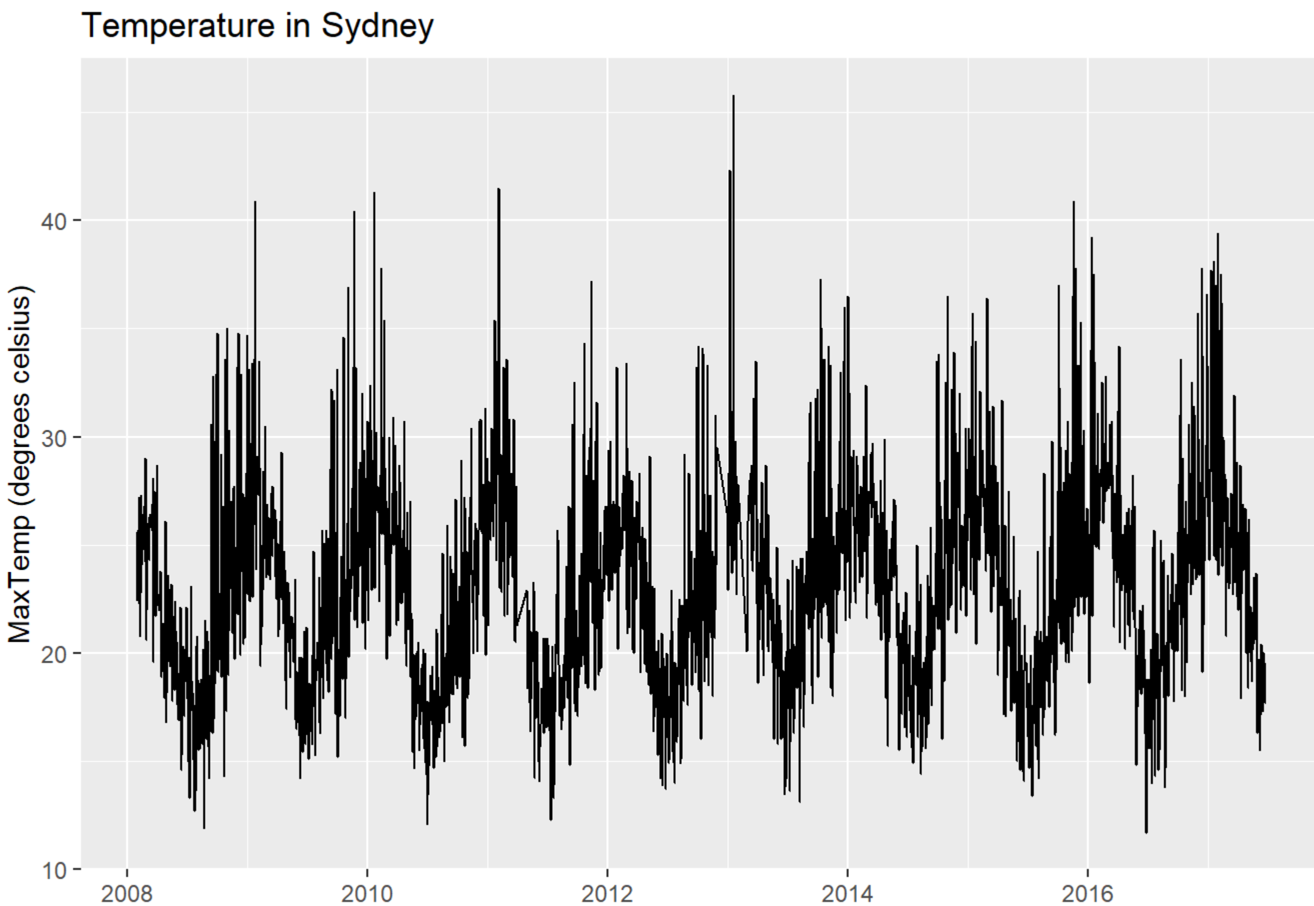
```
## Warning: Removed 31 rows containing missing values (`geom_line()`).
```



There is a very apparent season trend in the Temperature in Sydney. It starts off with the hottest temperatures in January and February, and then stays warm until the start of a slow decline in temperature in April. The decline in temperature continues until July where we see the lowest temperatures. The temperature then slowly increases until November where it stays very warm through December going into the next year.

Cycles

```
Sydney %>%
  autoplot(MaxTemp) + labs(y = "MaxTemp (degrees celsius)",x = " ", title = "Temperature in Sydn
ey")
```



There is no evidence of any cyclic behavior here.

MODEL FITTING

Split dataset into training and test sets

```
View(Sydney)
train <- Syd_Fill %>%
  filter(year(Date) < '2015-01-01')
test <- Syd_Fill %>%
  filter(year(Date)  >= '2015-01-01')
```

```
train$MaxTemp[is.na(train$MaxTemp)] <- median(train$MaxTemp,na.rm=TRUE)
train$Rainfall[is.na(train$Rainfall)] <- median(train$Rainfall,na.rm=TRUE)
train$Evaporation[is.na(train$Evaporation)] <- median(train$Evaporation,na.rm=TRUE)
train$Sunshine[is.na(train$Sunshine)] <- median(train$Sunshine,na.rm=TRUE)
train$WindGustSpeed[is.na(train$WindGustSpeed)] <- median(train$WindGustSpeed,na.rm=TRUE)
train$WindSpeed9am[is.na(train$WindSpeed9am)] <- median(train$WindSpeed9am,na.rm=TRUE)
train$WindSpeed3pm[is.na(train$WindSpeed3pm)] <- median(train$WindSpeed3pm,na.rm=TRUE)
train$Humidity9am[is.na(train$Humidity9am)] <- median(train$Humidity9am,na.rm=TRUE)
train$Humidity3pm[is.na(train$Humidity3pm)] <- median(train$Humidity3pm,na.rm=TRUE)
train$Pressure9am[is.na(train$Pressure9am)] <- median(train$Pressure9am,na.rm=TRUE)
train$Pressure3pm[is.na(train$Pressure3pm)] <- median(train$Pressure3pm,na.rm=TRUE)
train$Cloud9am[is.na(train$Cloud9am)] <- median(train$Cloud9am,na.rm=TRUE)
train$Cloud3pm[is.na(train$Cloud3pm)] <- median(train$Cloud3pm,na.rm=TRUE)
train$Year[is.na(train$Year)] <- median(train$Year,na.rm=TRUE)
train$Month[is.na(train$Month)] <- median(train$Month,na.rm=TRUE)
train$TempDiff[is.na(train$TempDiff)] <- median(train$TempDiff,na.rm=TRUE)
train$RainToday_1[is.na(train$RainToday_1)] <- median(train$RainToday_1,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainToday_1`.
## Unknown or uninitialised column: `RainToday_1`.
## Unknown or uninitialised column: `RainToday_1`.
```

```
train$RainToday_2[is.na(train$RainToday_2)] <- median(train$RainToday_2,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainToday_2`.
```

```
## Warning: Unknown or uninitialised column: `RainToday_2`.
## Unknown or uninitialised column: `RainToday_2`.
```

```
train$RainToday_Unknown[is.na(train$RainToday_Unknown)] <- median(train$RainToday_Unknown,na.rm=
TRUE)
train$RainTomorrow_1[is.na(train$RainTomorrow_1)] <- median(train$RainTomorrow_1,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_1`.
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_1`.
## Unknown or uninitialised column: `RainTomorrow_1`.
```

```
train$RainTomorrow_2[is.na(train$RainTomorrow_2)] <- median(train$RainTomorrow_2,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_2`.
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_2`.
## Unknown or uninitialised column: `RainTomorrow_2`.
```

```
train$RainTomorrow_Unknown[is.na(train$RainTomorrow_Unknown)] <- median(train$RainTomorrow_Unkno
wn,na.rm=TRUE)

summary(train)
```

```
##       Date              Location            MaxTemp         Rainfall      
##  Min.   :2008-02-01   Length:2891        Min.   :11.90   Min.   :  0.000  
##  1st Qu.:2010-01-23   Class :character   1st Qu.:19.60   1st Qu.:  0.000  
##  Median :2012-01-16   Mode  :character   Median :22.60   Median :  0.000  
##  Mean   :2012-01-16                      Mean   :22.78   Mean   :  3.059  
##  3rd Qu.:2014-01-07                      3rd Qu.:25.60   3rd Qu.:  1.000  
##  Max.   :2015-12-31                      Max.   :45.80   Max.   :119.400  
##                                                                           
##   Evaporation        Sunshine      WindGustSpeed    WindSpeed9am  
##  Min.   : 0.000   Min.   : 0.000   Min.   :17.00   Min.   : 0.00  
##  1st Qu.: 3.200   1st Qu.: 4.400   1st Qu.:37.00   1st Qu.:11.00  
##  Median : 4.800   Median : 8.300   Median :39.00   Median :15.00  
##  Mean   : 5.084   Mean   : 7.203   Mean   :40.77   Mean   :15.01  
##  3rd Qu.: 6.800   3rd Qu.:10.100   3rd Qu.:43.00   3rd Qu.:20.00  
##  Max.   :18.400   Max.   :13.600   Max.   :96.00   Max.   :54.00  
##                                                                   
##   WindSpeed3pm    Humidity9am      Humidity3pm     Pressure9am    
##  Min.   : 0.00   Min.   : 19.00   Min.   :10.00   Min.   : 986.7  
##  1st Qu.:15.00   1st Qu.: 59.00   1st Qu.:45.00   1st Qu.:1014.1  
##  Median :19.00   Median : 70.00   Median :56.00   Median :1018.6  
##  Mean   :19.31   Mean   : 68.78   Mean   :55.01   Mean   :1018.4  
##  3rd Qu.:24.00   3rd Qu.: 80.00   3rd Qu.:64.00   3rd Qu.:1023.1  
##  Max.   :50.00   Max.   :100.00   Max.   :99.00   Max.   :1038.8  
##                                                                   
##   Pressure3pm        Cloud9am        Cloud3pm          Year     
##  Min.   : 989.8   Min.   :0.000   Min.   :0.000   Min.   :2008  
##  1st Qu.:1011.6   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2010  
##  Median :1016.3   Median :5.000   Median :5.000   Median :2012  
##  Mean   :1016.1   Mean   :4.338   Mean   :4.379   Mean   :2012  
##  3rd Qu.:1020.8   3rd Qu.:7.000   3rd Qu.:6.000   3rd Qu.:2014  
##  Max.   :1036.7   Max.   :9.000   Max.   :8.000   Max.   :2015  
##                                                                 
##      Month           TempDiff       RainToday_No    RainToday_Unknown
##  Min.   : 1.000   Min.   : 0.200   Min.   :0.0000   Min.   :0.000000  
##  1st Qu.: 4.000   1st Qu.: 6.100   1st Qu.:0.0000   1st Qu.:0.000000  
##  Median : 7.000   Median : 8.000   Median :1.0000   Median :0.000000  
##  Mean   : 6.608   Mean   : 8.116   Mean   :0.7402   Mean   :0.002421  
##  3rd Qu.: 9.000   3rd Qu.:10.000   3rd Qu.:1.0000   3rd Qu.:0.000000  
##  Max.   :12.000   Max.   :24.100   Max.   :1.0000   Max.   :1.000000  
##                                    NA's   :89                         
##  RainToday_Yes    RainTomorrow_No  RainTomorrow_Unknown RainTomorrow_Yes
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000     Min.   :0.000   
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000     1st Qu.:0.000   
##  Median :0.0000   Median :1.0000   Median :0.000000     Median :0.000   
##  Mean   :0.2573   Mean   :0.7405   Mean   :0.002421     Mean   :0.257   
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000     3rd Qu.:1.000   
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000000     Max.   :1.000   
##  NA's   :89       NA's   :89                            NA's   :89      
```

```
test$MaxTemp[is.na(test$MaxTemp)] <- median(test$MaxTemp,na.rm=TRUE)
test$Rainfall[is.na(test$Rainfall)] <- median(test$Rainfall,na.rm=TRUE)
test$Evaporation[is.na(test$Evaporation)] <- median(test$Evaporation,na.rm=TRUE)
test$Sunshine[is.na(test$Sunshine)] <- median(test$Sunshine,na.rm=TRUE)
test$WindGustSpeed[is.na(test$WindGustSpeed)] <- median(test$WindGustSpeed,na.rm=TRUE)
test$WindSpeed9am[is.na(test$WindSpeed9am)] <- median(test$WindSpeed9am,na.rm=TRUE)
test$WindSpeed3pm[is.na(test$WindSpeed3pm)] <- median(test$WindSpeed3pm,na.rm=TRUE)
test$Humidity9am[is.na(test$Humidity9am)] <- median(test$Humidity9am,na.rm=TRUE)
test$Humidity3pm[is.na(test$Humidity3pm)] <- median(test$Humidity3pm,na.rm=TRUE)
test$Pressure9am[is.na(test$Pressure9am)] <- median(test$Pressure9am,na.rm=TRUE)
test$Pressure3pm[is.na(test$Pressure3pm)] <- median(test$Pressure3pm,na.rm=TRUE)
test$Cloud9am[is.na(test$Cloud9am)] <- median(test$Cloud9am,na.rm=TRUE)
test$Cloud3pm[is.na(test$Cloud3pm)] <- median(test$Cloud3pm,na.rm=TRUE)
test$Year[is.na(test$Year)] <- median(test$Year,na.rm=TRUE)
test$Month[is.na(test$Month)] <- median(test$Month,na.rm=TRUE)
test$TempDiff[is.na(test$TempDiff)] <- median(test$TempDiff,na.rm=TRUE)
test$RainToday_1[is.na(test$RainToday_1)] <- median(test$RainToday_1,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainToday_1`.
```

```
## Warning: Unknown or uninitialised column: `RainToday_1`.
## Unknown or uninitialised column: `RainToday_1`.
```

```
test$RainToday_2[is.na(test$RainToday_2)] <- median(test$RainToday_2,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainToday_2`.
```

```
## Warning: Unknown or uninitialised column: `RainToday_2`.
## Unknown or uninitialised column: `RainToday_2`.
```

```
test$RainToday_Unknown[is.na(test$RainToday_Unknown)] <- median(test$RainToday_Unknown,na.rm=TRU
E)
test$RainTomorrow_1[is.na(test$RainTomorrow_1)] <- median(test$RainTomorrow_1,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_1`.
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_1`.
## Unknown or uninitialised column: `RainTomorrow_1`.
```

```
test$RainTomorrow_2[is.na(test$RainTomorrow_2)] <- median(test$RainTomorrow_2,na.rm=TRUE)
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_2`.
```

```
## Warning: Unknown or uninitialised column: `RainTomorrow_2`.
## Unknown or uninitialised column: `RainTomorrow_2`.
```

```
test$RainTomorrow_Unknown[is.na(test$RainTomorrow_Unknown)] <- median(test$RainTomorrow_Unknown,
na.rm=TRUE)

summary(test)
```

```
##       Date              Location            MaxTemp         Rainfall
##  Min.   :2016-01-01   Length:542         Min.   :11.70   Min.   : 0.000
##  1st Qu.:2016-05-15   Class :character   1st Qu.:20.60   1st Qu.: 0.000
##  Median :2016-09-27   Mode  :character   Median :24.20   Median : 0.000
##  Mean   :2016-09-27                      Mean   :24.11   Mean   : 4.154
##  3rd Qu.:2017-02-09                      3rd Qu.:27.00   3rd Qu.: 1.400
##  Max.   :2017-06-25                      Max.   :39.40   Max.   :94.400
##   Evaporation       Sunshine      WindGustSpeed    WindSpeed9am
##  Min.   : 0.00   Min.   : 0.000   Min.   :19.00   Min.   : 0.00
##  1st Qu.: 3.40   1st Qu.: 4.500   1st Qu.:31.00   1st Qu.:11.00
##  Median : 5.20   Median : 8.400   Median :39.00   Median :15.00
##  Mean   : 5.65   Mean   : 7.272   Mean   :41.28   Mean   :15.29
##  3rd Qu.: 7.80   3rd Qu.:10.100   3rd Qu.:49.50   3rd Qu.:20.00
##  Max.   :15.80   Max.   :13.500   Max.   :96.00   Max.   :44.00
##   WindSpeed3pm    Humidity9am     Humidity3pm     Pressure9am
##  Min.   : 2.00   Min.   :21.00   Min.   :14.00   Min.   : 998.3
##  1st Qu.:15.00   1st Qu.:56.00   1st Qu.:43.00   1st Qu.:1013.2
##  Median :19.00   Median :66.00   Median :54.00   Median :1018.0
##  Mean   :19.36   Mean   :65.57   Mean   :53.14   Mean   :1017.9
##  3rd Qu.:24.00   3rd Qu.:76.00   3rd Qu.:62.75   3rd Qu.:1022.6
##  Max.   :57.00   Max.   :92.00   Max.   :91.00   Max.   :1039.0
##   Pressure3pm      Cloud9am        Cloud3pm          Year          Month
##  Min.   : 994   Min.   :0.000   Min.   :0.000   Min.   :2016   Min.   : 1.000
##  1st Qu.:1011   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2016   1st Qu.: 3.000
##  Median :1016   Median :5.000   Median :4.500   Median :2016   Median : 5.000
##  Mean   :1015   Mean   :4.332   Mean   :4.304   Mean   :2016   Mean   : 5.515
##  3rd Qu.:1020   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:2017   3rd Qu.: 8.000
##  Max.   :1036   Max.   :8.000   Max.   :8.000   Max.   :2017   Max.   :12.000
##     TempDiff       RainToday_No    RainToday_Unknown RainToday_Yes
##  Min.   : 0.400   Min.   :0.0000   Min.   :0         Min.   :0.0000
##  1st Qu.: 6.400   1st Qu.:0.0000   1st Qu.:0         1st Qu.:0.0000
##  Median : 8.150   Median :1.0000   Median :0         Median :0.0000
##  Mean   : 8.232   Mean   :0.7325   Mean   :0         Mean   :0.2675
##  3rd Qu.:10.200   3rd Qu.:1.0000   3rd Qu.:0         3rd Qu.:1.0000
##  Max.   :17.100   Max.   :1.0000   Max.   :0         Max.   :1.0000
##  RainTomorrow_No  RainTomorrow_Unknown RainTomorrow_Yes
##  Min.   :0.0000   Min.   :0            Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0            1st Qu.:0.0000
##  Median :1.0000   Median :0            Median :0.0000
##  Mean   :0.7325   Mean   :0            Mean   :0.2675
##  3rd Qu.:1.0000   3rd Qu.:0            3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :0            Max.   :1.0000
```

```
test <- test %>%
  as_tsibble(index = Date, key = NULL)
train <- train %>%
  as_tsibble(index = Date, key = NULL)
```

I chose to split the data at the year 2015 because it is close to 80% of the records in the training set and 20% of the records into the test set.

Fit TSLM, ETS and ARIMA model(s):

TSLM and ETS

```
TSLM_ETS_Models <- train %>%
  model(
    TSLM = TSLM(MaxTemp ~ trend()),
    SES = ETS(log(MaxTemp) ~ error("A") + trend("N") + season("N")),
    Holt = ETS(log(MaxTemp) ~ error("A") + trend("A") + season("N")),
    Damped = ETS(log(MaxTemp) ~ error("A") + trend("Ad") + season("N")),
    Additive = ETS(log(MaxTemp) ~ error("A") + trend("A") + season("A")),
    Multiplicative = ETS(log(MaxTemp) ~ error("M") + trend("A") + season("M"))
  )

glance(TSLM_ETS_Models)
```

```
## # A tibble: 6 × 18
##   .model  r_squa…¹ adj_r_…²  sigma2 stati…³  p_value    df log_lik    AIC   AICc
##   <chr>      <dbl>    <dbl>   <dbl>   <dbl>    <dbl> <int>   <dbl>  <dbl>  <dbl>
## 1 TSLM     0.00535  0.00500 1.89e+1    15.5  8.32e-5     2  -8352.  8507.  8507.
## 2 SES     NA       NA       1.66e-2    NA   NA          NA  -5597. 11200. 11200.
## 3 Holt    NA       NA       1.67e-2    NA   NA          NA  -5606. 11221. 11221.
## 4 Damped  NA       NA       1.67e-2    NA   NA          NA  -5605. 11222. 11222.
## 5 Additi… NA       NA       1.67e-2    NA   NA          NA  -5598. 11220. 11220.
## 6 Multip… NA       NA       1.72e-3    NA   NA          NA  -5587. 11198. 11199.
## # … with 8 more variables: BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>, MSE <dbl>, AMSE <dbl>, MAE <dbl>, and
## #   abbreviated variable names ¹r_squared, ²adj_r_squared, ³statistic
```

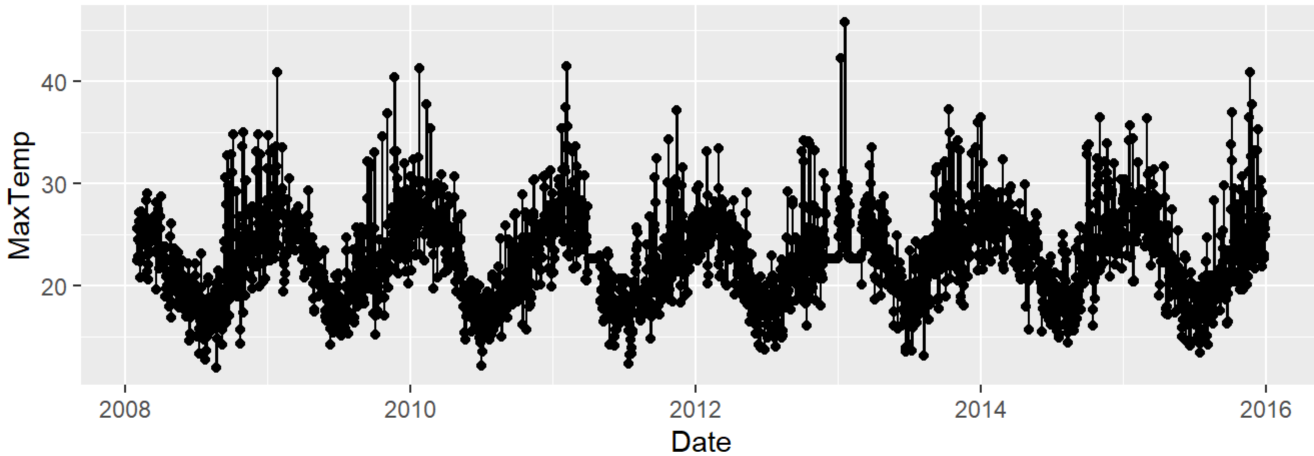
The lowest AICc of the TSLM and different ETS models is the TSLM model at 8507.

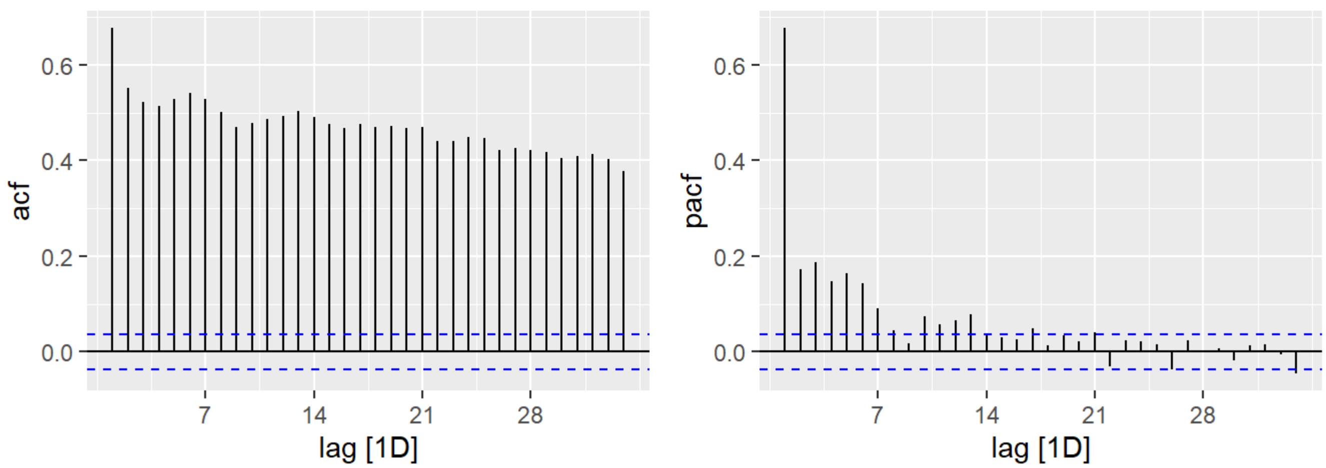
ARIMA

```
# Check for stationarity

train %>% features(MaxTemp, unitroot_nsdiffs)
```

```
## # A tibble: 1 × 1
##   nsdiffs
##     <int>
## 1       0
```

```
# 0 ndsdiffs recommended therefore data is stationary and we can continue with ARIMA model

# Plot ACF and PACF

train %>% gg_tsdisplay(MaxTemp, plot_type = 'partial')
```






```
# PACF dies in somewhat sine wave manner but acf does not die out at all. Therefore there is no clear ar or
# ma choice based on the ACF and PACF plot.

# Create ARIMA models

ARIMA_Models <- train %>%
  model(
     arima_auto = ARIMA(MaxTemp),
     automatic_exhaustive = ARIMA(MaxTemp, stepwise = FALSE), #exhaustive search
     automatic_no_seas_exhaustive = ARIMA(MaxTemp ~ PDQ(0, 0, 0), stepwise = FALSE), #exhaustive search no seasonal differences
     automatic_no_seas = ARIMA(MaxTemp ~ PDQ(0,0,0)) #fable algorithm no seasonal differencing
  )

glance(ARIMA_Models)
```

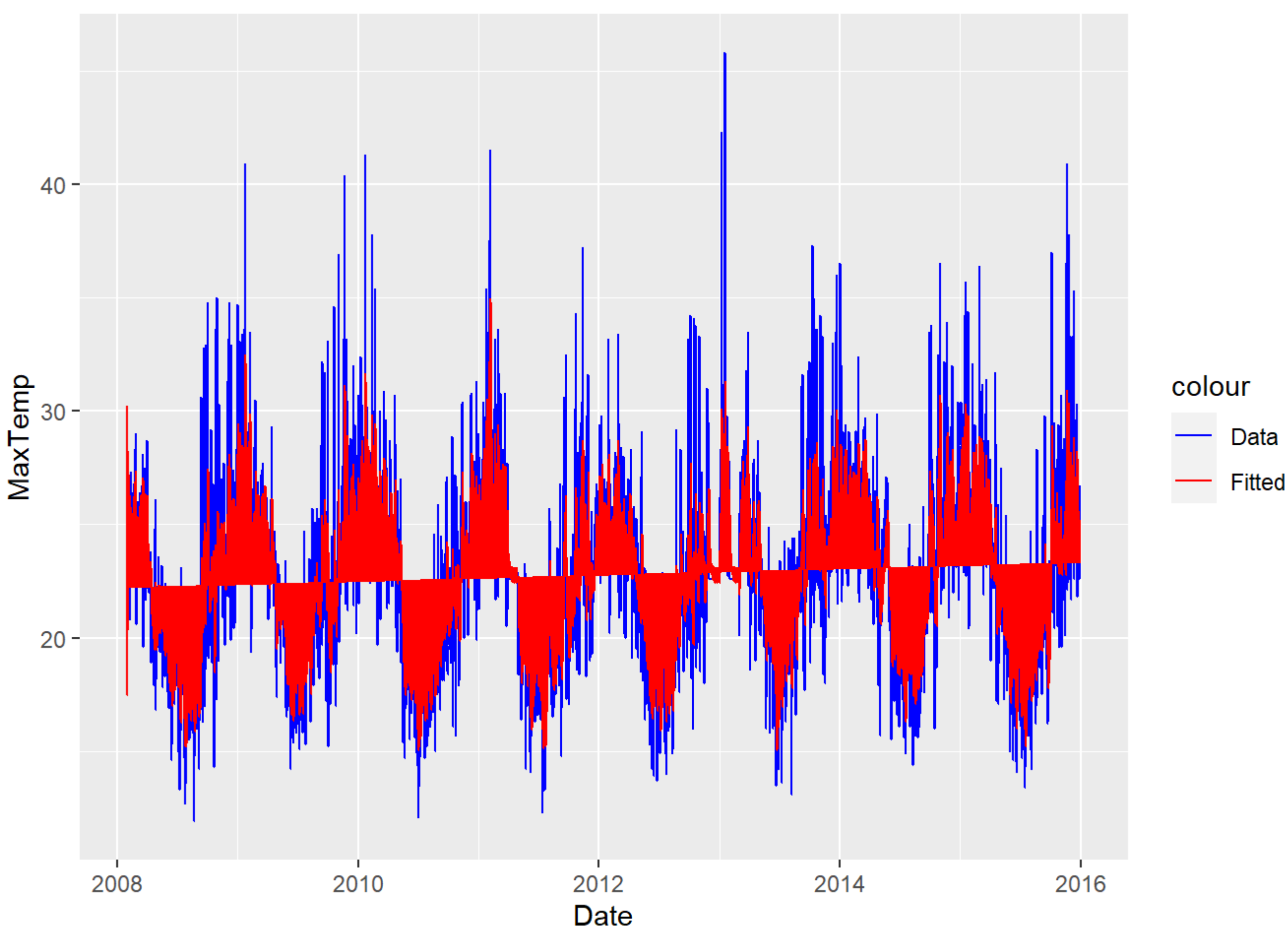
```
## # A tibble: 4 × 8
##   .model                     sigma2 log_lik    AIC   AICc    BIC ar_ro…¹ ma_ro…²
##   <chr>                       <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>  <list>
## 1 arima_auto                   8.77  -7238. 14491. 14491. 14539. <cpl>   <cpl>
## 2 automatic_exhaustive         8.77  -7238. 14491. 14491. 14539. <cpl>   <cpl>
## 3 automatic_no_seas_exhaust…   8.78  -7239. 14495. 14495. 14543. <cpl>   <cpl>
## 4 automatic_no_seas            9.41  -7341. 14691. 14691. 14714. <cpl>   <cpl>
## # … with abbreviated variable names ¹ar_roots, ²ma_roots
```

Lowest AICc of ARIMA models is arima_auto with 14491

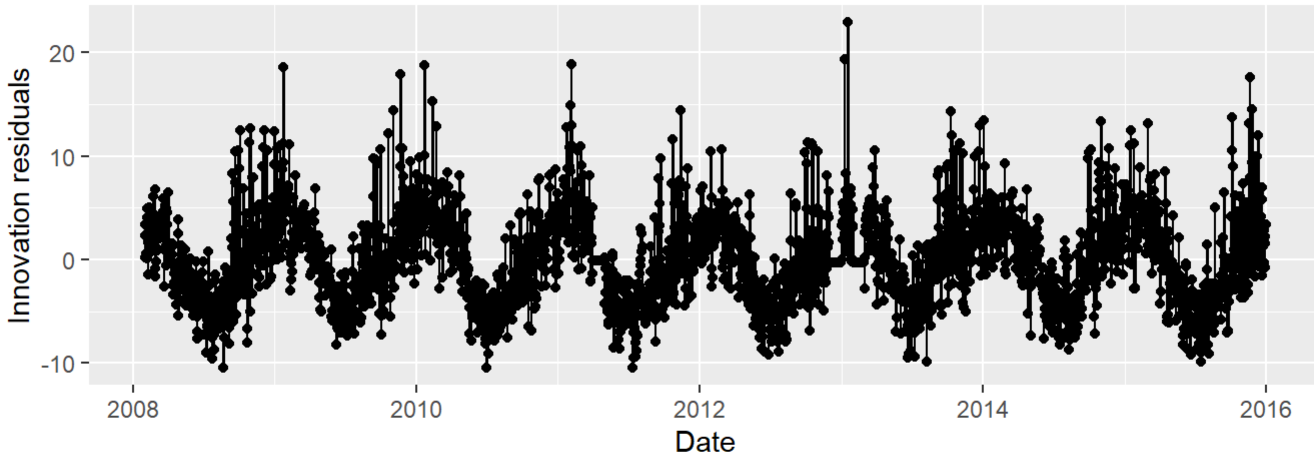
Evaluate residuals of TSLM, ETS, and ARIMA models:

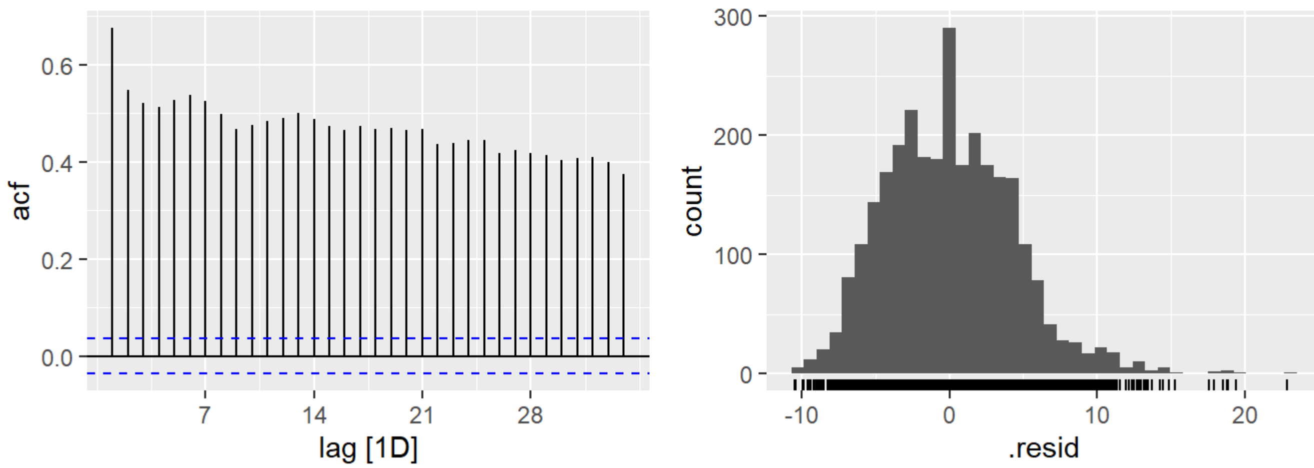
TSLM and ETS residuals

```
aug_TSLM_ETS <- augment(TSLM_ETS_Models)
aug_TSLM_ETS %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y = MaxTemp, color = "Data")) +
  geom_line(aes(y = .fitted, color = "Fitted")) +
  scale_color_manual(values = c(Data = "Blue", Fitted = "Red"))
```




```
# Using best model for gg_tsresiduals()
TSLM_ETS_Models %>% select(TSLM) %>%  gg_tsresiduals()
```





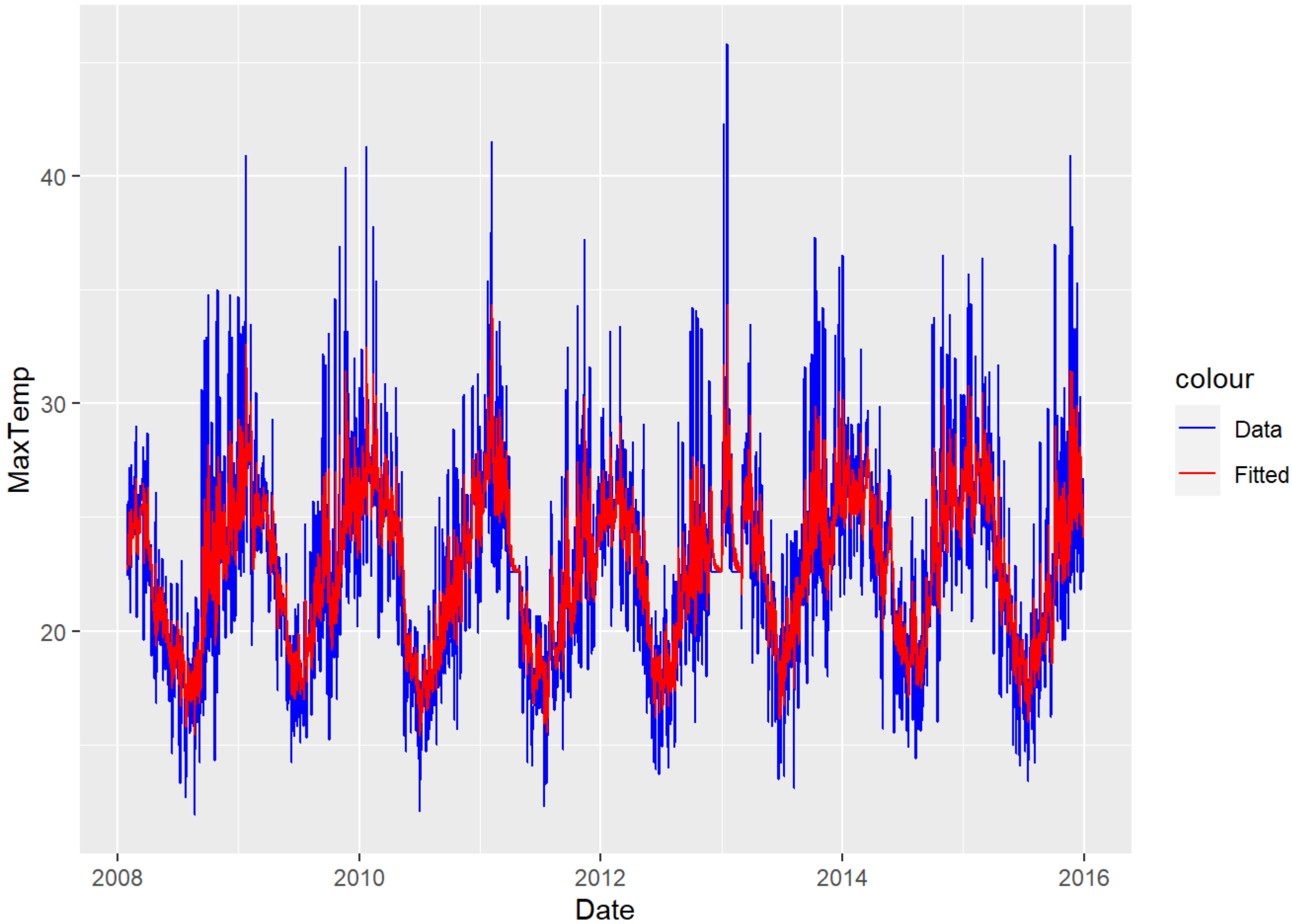

```
# Check if residuals are stationary
aug_TSLM_ETS %>% features(.innov, unitroot_kpss)
```

```
## # A tibble: 6 × 3
##   .model         kpss_stat kpss_pvalue
##   <chr>              <dbl>       <dbl>
## 1 Additive          0.0198         0.1
## 2 Damped            0.117          0.1
## 3 Holt              0.0145         0.1
## 4 Multiplicative    0.0398         0.1
## 5 SES               0.0140         0.1
## 6 TSLM              0.123          0.1
```

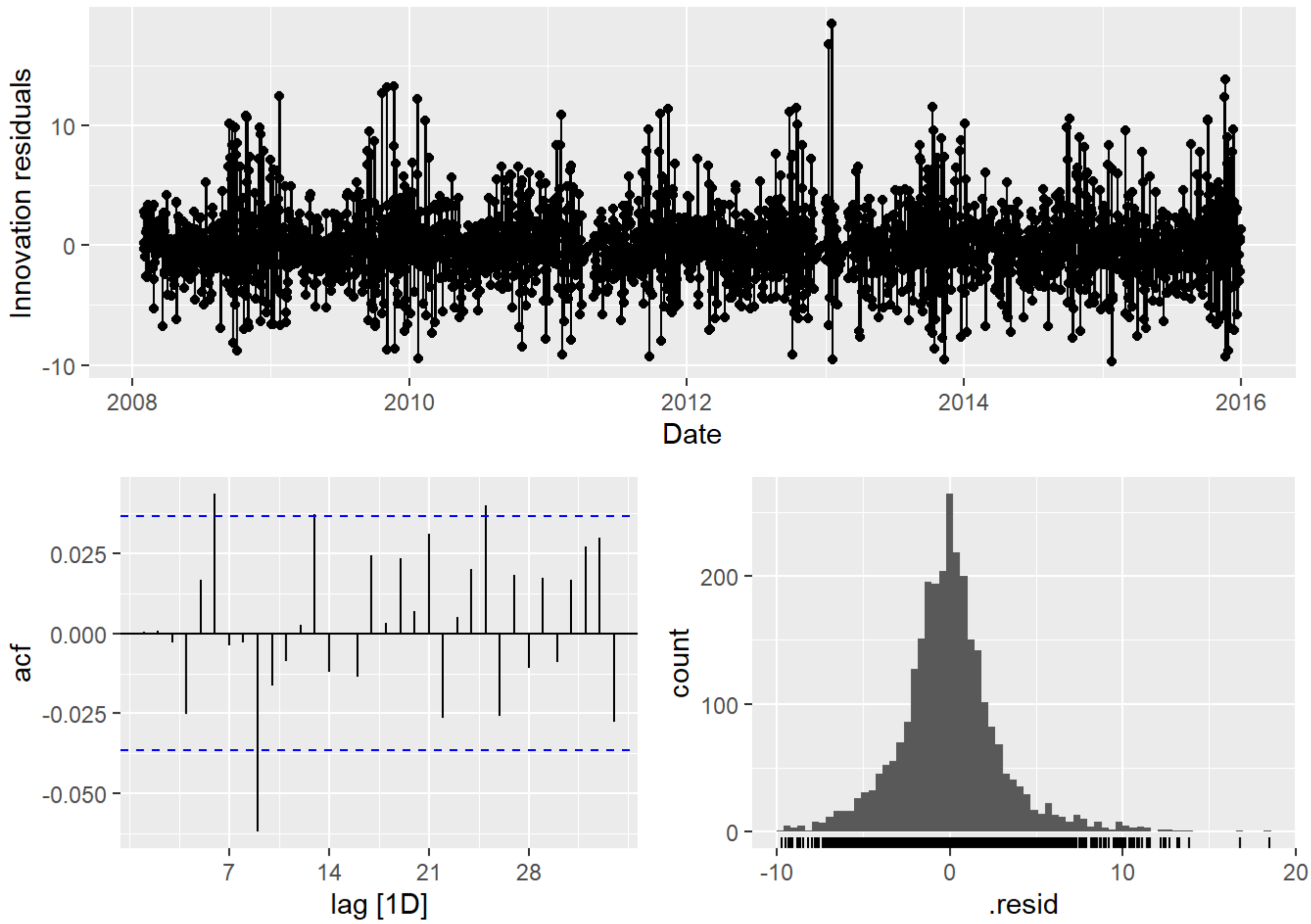
Because the model residuals each have a p-value of 0.1, this means that they are all stationary.

ARIMA residuals

```
aug_ARIMA <- augment(ARIMA_Models)
aug_ARIMA %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y = MaxTemp, color = "Data")) +
  geom_line(aes(y = .fitted, color = "Fitted")) +
  scale_color_manual(values = c(Data = "Blue", Fitted = "Red"))
```




```
# Using best model for gg_tsresiduals()
ARIMA_Models %>% select(arima_auto) %>%  gg_tsresiduals()
```




```
# Check if residuals are stationary
aug_ARIMA %>% features(.innov, unitroot_kpss)
```

```
## # A tibble: 4 × 3
##   .model                          kpss_stat kpss_pvalue
##   <chr>                               <dbl>       <dbl>
## 1 arima_auto                         0.0553         0.1
## 2 automatic_exhaustive               0.0553         0.1
## 3 automatic_no_seas                  0.0765         0.1
## 4 automatic_no_seas_exhaustive      0.0549         0.1
```

Because the model residuals each have a p-value of 0.1, this means that they are all stationary.

If there are predictor variables – fit a TSLM with predictor variables, Regression with ARIMA errors:

TSLM with Predictor Variables

```
TSLM_Predictors <- train %>%
  model(
    lm = TSLM(MaxTemp ~ TempDiff),
    lm2 = TSLM(MaxTemp ~ TempDiff + Rainfall),
    lm3 = TSLM(MaxTemp ~ Evaporation + Humidity3pm + Cloud3pm),
    lm4 = TSLM(MaxTemp ~ Humidity9am + Humidity3pm + Pressure9am + Pressure3pm + TempDiff),
    lm5 = TSLM(MaxTemp ~ Humidity9am + Humidity3pm + Pressure9am + Pressure3pm),
    lm6 = TSLM(MaxTemp ~ Rainfall + Evaporation + Humidity9am + Humidity3pm),
    lm7 = TSLM(MaxTemp ~ Sunshine + Humidity9am + Humidity3pm),
    lm8 = TSLM(MaxTemp ~ Sunshine + Cloud9am + Cloud3pm + TempDiff),
    lm9 = TSLM(MaxTemp ~ TempDiff + Sunshine + Evaporation + Humidity9am + Humidity3pm + Pressur
e9am +
                 Pressure3pm + Rainfall)
  )

glance(TSLM_Predictors)
```

```
## # A tibble: 9 × 15
##   .model r_squared adj_r_sq…¹ sigma2 stati…²   p_value    df log_lik   AIC  AICc
##   <chr>      <dbl>      <dbl>  <dbl>   <dbl>     <dbl> <int>   <dbl> <dbl> <dbl>
## 1 lm         0.106      0.106  17.0     343. 1.69e- 72     2  -8198. 8198. 8198.
## 2 lm2        0.112      0.112  16.9     183. 1.60e- 75     3  -8188. 8179. 8179.
## 3 lm3        0.262      0.261  14.1     341. 1.11e-189     4  -7921. 7649. 7649.
## 4 lm4        0.289      0.288  13.5     235. 7.69e-211     6  -7867. 7543. 7543.
## 5 lm5        0.200      0.199  15.2     181. 2.03e-138     5  -8037. 7882. 7882.
## 6 lm6        0.282      0.281  13.7     283. 1.34e-205     5  -7882. 7571. 7571.
## 7 lm7        0.114      0.113  16.9     124. 1.33e- 75     4  -8185. 8176. 8176.
## 8 lm8        0.181      0.180  15.6     160. 1.11e-123     5  -8071. 7950. 7950.
## 9 lm9        0.533      0.531   8.92    410. 0             9  -7261. 6337. 6337.
## # … with 5 more variables: BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>, and abbreviated variable names
## #   ¹adj_r_squared, ²statistic
```

Lowest AICc is lm9 with 6337.

ARIMA with Errors

```
ARIMA_Errors <- train %>%
  model(
  ARIMA1 = ARIMA(MaxTemp ~ TempDiff),
  ARIMA2 = ARIMA(MaxTemp ~ TempDiff + Rainfall),
  ARIMA3 = ARIMA(MaxTemp ~ Evaporation + Humidity3pm + Cloud3pm),
  ARIMA4 = ARIMA(MaxTemp ~ Humidity9am + Humidity3pm + Pressure9am + Pressure3pm + TempDiff),
  ARIMA5 = ARIMA(MaxTemp ~ Humidity9am + Humidity3pm + Pressure9am + Pressure3pm),
  ARIMA6 = ARIMA(MaxTemp ~ Rainfall + Evaporation + Humidity9am + Humidity3pm),
  ARIMA7 = ARIMA(MaxTemp ~ Sunshine + Humidity9am + Humidity3pm),
  ARIMA8 = ARIMA(MaxTemp ~ Sunshine + Cloud9am + Cloud3pm + TempDiff),
  ARIMA9 = ARIMA(MaxTemp ~ TempDiff + Sunshine + Evaporation + Humidity9am + Humidity3pm + Press
ure9am +
             Pressure3pm + Rainfall)
  )
glance(ARIMA_Errors)
```

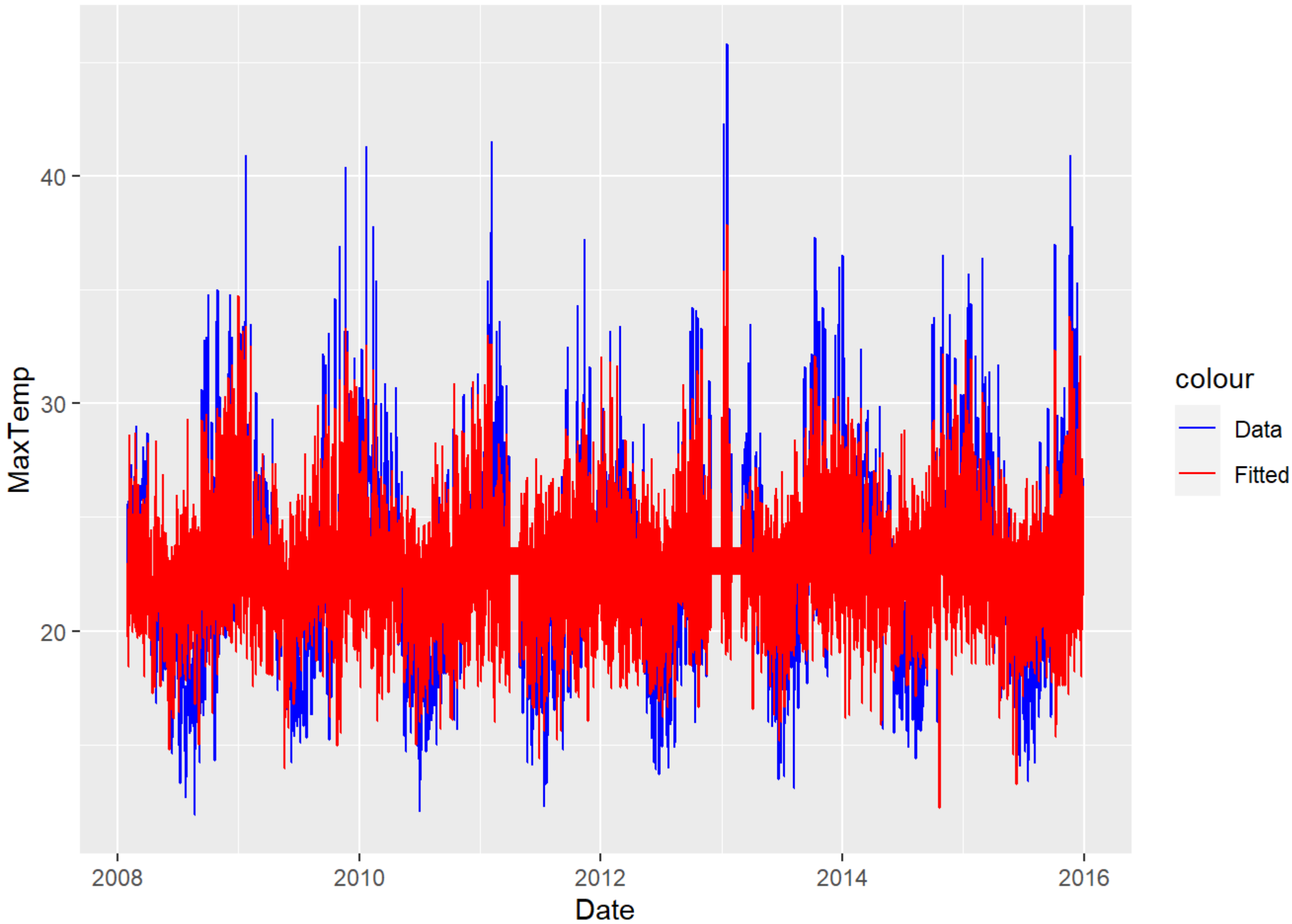
```
## # A tibble: 9 × 8
##   .model sigma2 log_lik    AIC   AICc    BIC ar_roots  ma_roots
##   <chr>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>    <list>
## 1 ARIMA1   3.14  -5753. 11522. 11522. 11570. <cpl [1]> <cpl [16]>
## 2 ARIMA2   3.13  -5749. 11515. 11516. 11569. <cpl [1]> <cpl [16]>
## 3 ARIMA3   6.71  -6851. 13724. 13724. 13789. <cpl [8]> <cpl [16]>
## 4 ARIMA4   2.43  -5380. 10785. 10785. 10857. <cpl [1]> <cpl [4]>
## 5 ARIMA5   5.41  -6537. 13095. 13095. 13154. <cpl [3]> <cpl [2]>
## 6 ARIMA6   6.68  -6843. 13708. 13708. 13774. <cpl [8]> <cpl [9]>
## 7 ARIMA7   6.50  -6805. 13630. 13630. 13690. <cpl [8]> <cpl [9]>
## 8 ARIMA8   3.09  -5728. 11478. 11478. 11543. <cpl [1]> <cpl [16]>
## 9 ARIMA9   2.39  -5354. 10738. 10738. 10827. <cpl [1]> <cpl [4]>
```

The lowest AICc is ARIMA9 with 10738

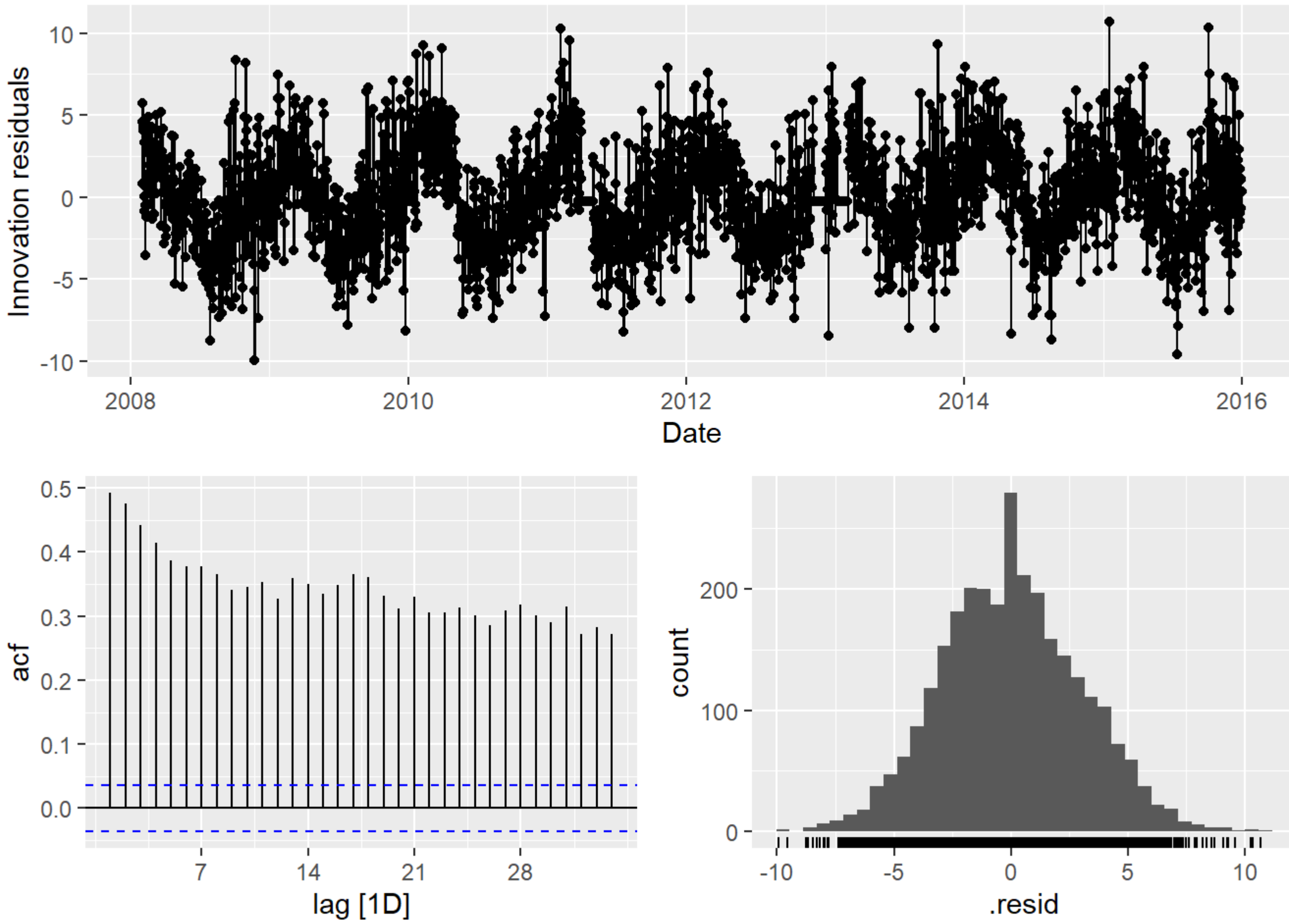
Evaluate Residuals of TSLM w/ predictors and ARIMA with errors:

TSLM w/ Predictors residuals

```
aug_TSLM_Predictors <- augment(TSLM_Predictors)
aug_TSLM_Predictors %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y = MaxTemp, color = "Data")) +
  geom_line(aes(y = .fitted, color = "Fitted")) +
  scale_color_manual(values = c(Data = "Blue", Fitted = "Red"))
```




```
# Using best model for gg_tsresiduals()
TSLM_Predictors %>% select(lm9) %>%  gg_tsresiduals()
```



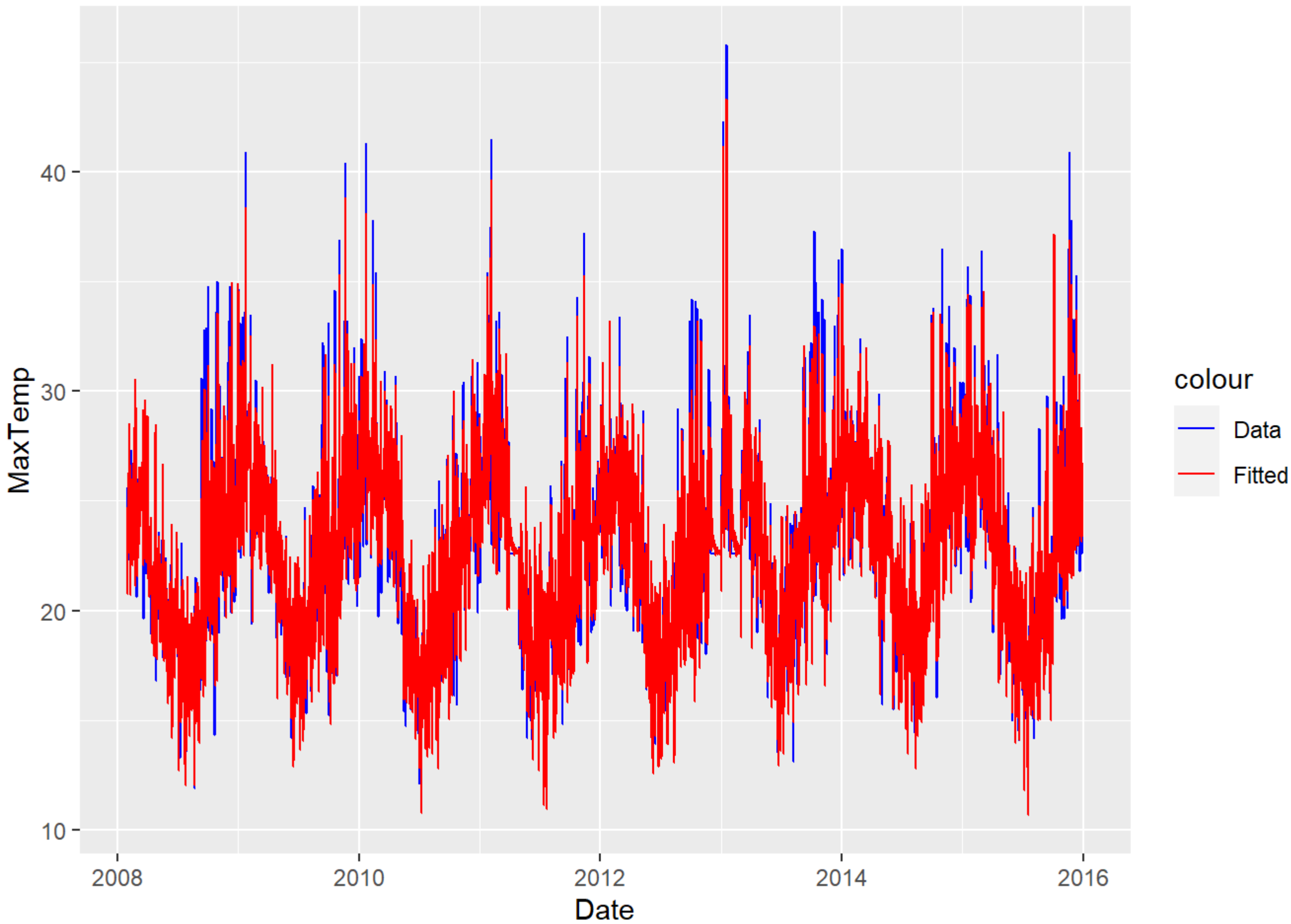

```
# Check if residuals are stationary
aug_TSLM_Predictors %>% features(.innov, unitroot_kpss)
```

```
## # A tibble: 9 × 3
##   .model kpss_stat kpss_pvalue
##   <chr>      <dbl>       <dbl>
## 1 lm         0.202      0.1
## 2 lm2        0.209      0.1
## 3 lm3        0.306      0.1
## 4 lm4        0.343      0.1
## 5 lm5        0.492      0.0434
## 6 lm6        0.400      0.0772
## 7 lm7        0.340      0.1
## 8 lm8        0.237      0.1
## 9 lm9        0.427      0.0656
```

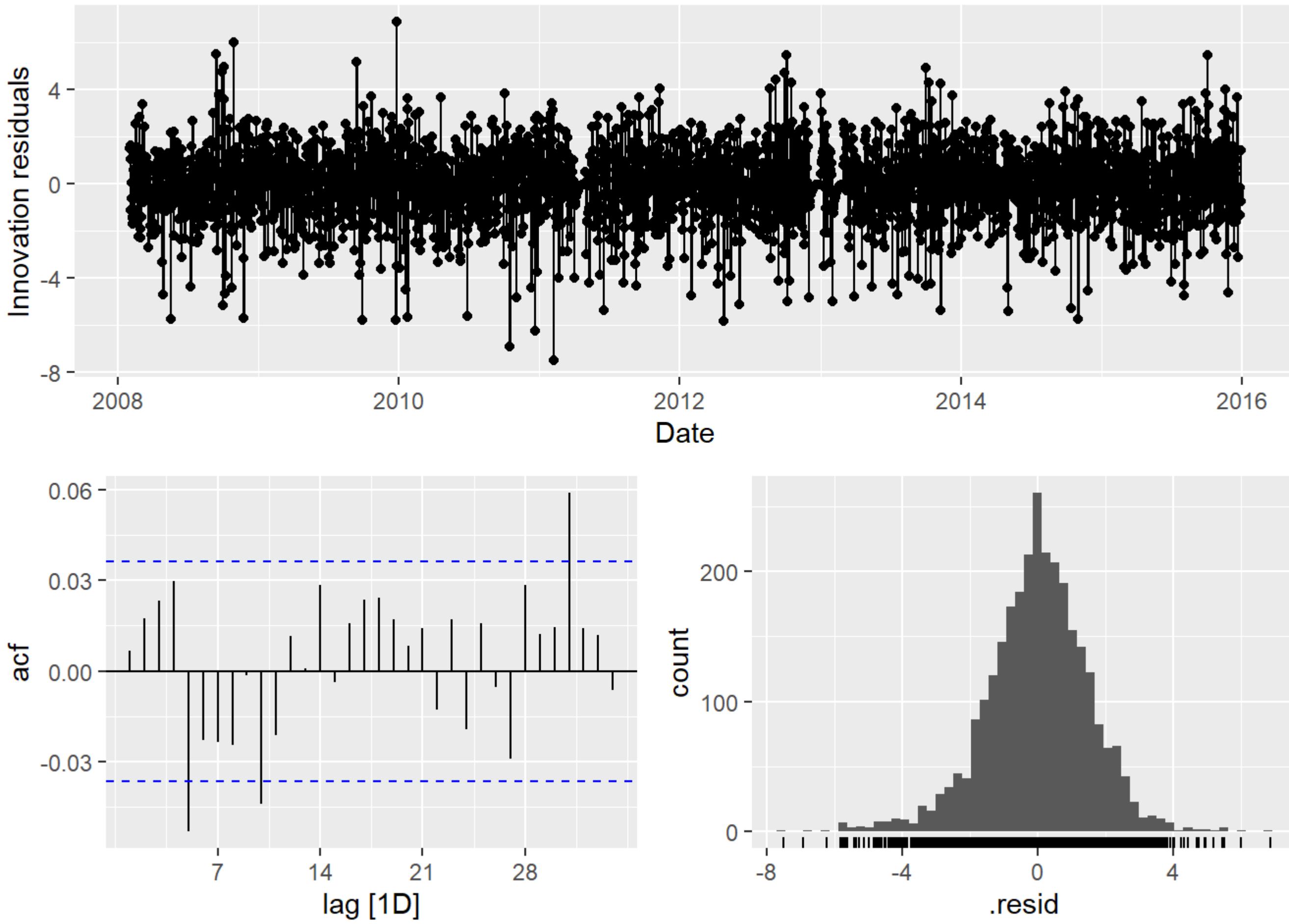
The residuals are stationary for lm1, lm2, lm3, lm4, lm7, and lm8 because they are the only ones with a p-value of 0.1. The others have smaller p-values making them not stationary.

ARIMA w/ Errors residuals

```
aug_ARIMA_e <- augment(ARIMA_Errors)
aug_ARIMA_e %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y = MaxTemp, color = "Data")) +
  geom_line(aes(y = .fitted, color = "Fitted")) +
  scale_color_manual(values = c(Data = "Blue", Fitted = "Red"))
```




```
# Using best model for gg_tsresiduals()
ARIMA_Errors %>% select(ARIMA9) %>%  gg_tsresiduals()
```




```
# Check if residuals are stationary
aug_ARIMA_e %>% features(.innov, unitroot_kpss)
```

```
## # A tibble: 9 × 3
##   .model kpss_stat kpss_pvalue
##   <chr>      <dbl>       <dbl>
## 1 ARIMA1    0.0377         0.1
## 2 ARIMA2    0.0382         0.1
## 3 ARIMA3    0.0462         0.1
## 4 ARIMA4    0.0460         0.1
## 5 ARIMA5    0.0408         0.1
## 6 ARIMA6    0.0479         0.1
## 7 ARIMA7    0.0489         0.1
## 8 ARIMA8    0.0370         0.1
## 9 ARIMA9    0.0454         0.1
```

All of the models have stationary residuals here because they all have a p-value of 0.1.

Benchmark Methods

```
benchmark <- train %>%
  model(
    mean = MEAN(MaxTemp),
    naive = NAIVE(MaxTemp),
    s_naive = SNAIVE(MaxTemp),
    drift = RW(MaxTemp ~ drift())
  )
glance(benchmark)
```

```
## # A tibble: 4 × 2
##   .model  sigma2
##   <chr>    <dbl>
## 1 mean      19.0
## 2 naive     12.3
## 3 s_naive   18.0
## 4 drift     12.3
```

ACCURACY

Used glance from each of the 4 model families I built

```
glance(TSLM_ETS_Models)
```

```
## # A tibble: 6 × 18
##   .model  r_squa…¹ adj_r_…²  sigma2 stati…³  p_value    df log_lik    AIC   AICc
##   <chr>      <dbl>    <dbl>   <dbl>   <dbl>    <dbl> <int>   <dbl>  <dbl>  <dbl>
## 1 TSLM     0.00535  0.00500 1.89e+1    15.5  8.32e-5     2  -8352.  8507.  8507.
## 2 SES     NA       NA       1.66e-2    NA   NA          NA  -5597. 11200. 11200.
## 3 Holt    NA       NA       1.67e-2    NA   NA          NA  -5606. 11221. 11221.
## 4 Damped  NA       NA       1.67e-2    NA   NA          NA  -5605. 11222. 11222.
## 5 Additi… NA       NA       1.67e-2    NA   NA          NA  -5598. 11220. 11220.
## 6 Multip… NA       NA       1.72e-3    NA   NA          NA  -5587. 11198. 11199.
## # … with 8 more variables: BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>, MSE <dbl>, AMSE <dbl>, MAE <dbl>, and
## #   abbreviated variable names ¹r_squared, ²adj_r_squared, ³statistic
```

```
glance(ARIMA_Models)
```

```
## # A tibble: 4 × 8
##   .model                     sigma2 log_lik    AIC   AICc    BIC ar_ro…¹ ma_ro…²
##   <chr>                       <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>  <list>
## 1 arima_auto                   8.77  -7238. 14491. 14491. 14539. <cpl>   <cpl>
## 2 automatic_exhaustive         8.77  -7238. 14491. 14491. 14539. <cpl>   <cpl>
## 3 automatic_no_seas_exhaust…   8.78  -7239. 14495. 14495. 14543. <cpl>   <cpl>
## 4 automatic_no_seas            9.41  -7341. 14691. 14691. 14714. <cpl>   <cpl>
## # … with abbreviated variable names ¹ar_roots, ²ma_roots
```

```
glance(TSLM_Predictors)
```

```
## # A tibble: 9 × 15
##   .model r_squared adj_r_sq…¹ sigma2 stati…²   p_value    df log_lik   AIC  AICc
##   <chr>      <dbl>      <dbl>  <dbl>   <dbl>     <dbl> <int>   <dbl> <dbl> <dbl>
## 1 lm         0.106      0.106  17.0     343. 1.69e- 72     2  -8198. 8198. 8198.
## 2 lm2        0.112      0.112  16.9     183. 1.60e- 75     3  -8188. 8179. 8179.
## 3 lm3        0.262      0.261  14.1     341. 1.11e-189     4  -7921. 7649. 7649.
## 4 lm4        0.289      0.288  13.5     235. 7.69e-211     6  -7867. 7543. 7543.
## 5 lm5        0.200      0.199  15.2     181. 2.03e-138     5  -8037. 7882. 7882.
## 6 lm6        0.282      0.281  13.7     283. 1.34e-205     5  -7882. 7571. 7571.
## 7 lm7        0.114      0.113  16.9     124. 1.33e- 75     4  -8185. 8176. 8176.
## 8 lm8        0.181      0.180  15.6     160. 1.11e-123     5  -8071. 7950. 7950.
## 9 lm9        0.533      0.531   8.92    410. 0             9  -7261. 6337. 6337.
## # … with 5 more variables: BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>, and abbreviated variable names
## #   ¹adj_r_squared, ²statistic
```

```
glance(ARIMA_Errors)
```

```
## # A tibble: 9 × 8
##   .model sigma2 log_lik    AIC   AICc    BIC ar_roots  ma_roots
##   <chr>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>    <list>
## 1 ARIMA1   3.14  -5753. 11522. 11522. 11570. <cpl [1]> <cpl [16]>
## 2 ARIMA2   3.13  -5749. 11515. 11516. 11569. <cpl [1]> <cpl [16]>
## 3 ARIMA3   6.71  -6851. 13724. 13724. 13789. <cpl [8]> <cpl [16]>
## 4 ARIMA4   2.43  -5380. 10785. 10785. 10857. <cpl [1]> <cpl [4]>
## 5 ARIMA5   5.41  -6537. 13095. 13095. 13154. <cpl [3]> <cpl [2]>
## 6 ARIMA6   6.68  -6843. 13708. 13708. 13774. <cpl [8]> <cpl [9]>
## 7 ARIMA7   6.50  -6805. 13630. 13630. 13690. <cpl [8]> <cpl [9]>
## 8 ARIMA8   3.09  -5728. 11478. 11478. 11543. <cpl [1]> <cpl [16]>
## 9 ARIMA9   2.39  -5354. 10738. 10738. 10827. <cpl [1]> <cpl [4]>
```

MODELS BEST MODEL AICc TSLM and ETS TSLM 8507 ARIMA arima_auto 14491 TSLM_Predictors lm9 6337 ARIMA Errors ARIMA9 10738

The overall model with the lowest AICc is the lm9 model from TSLM_Predictors.

I selected AICc because it was an easy method to compare accuracy of my models performance while also accounting for model complexity.

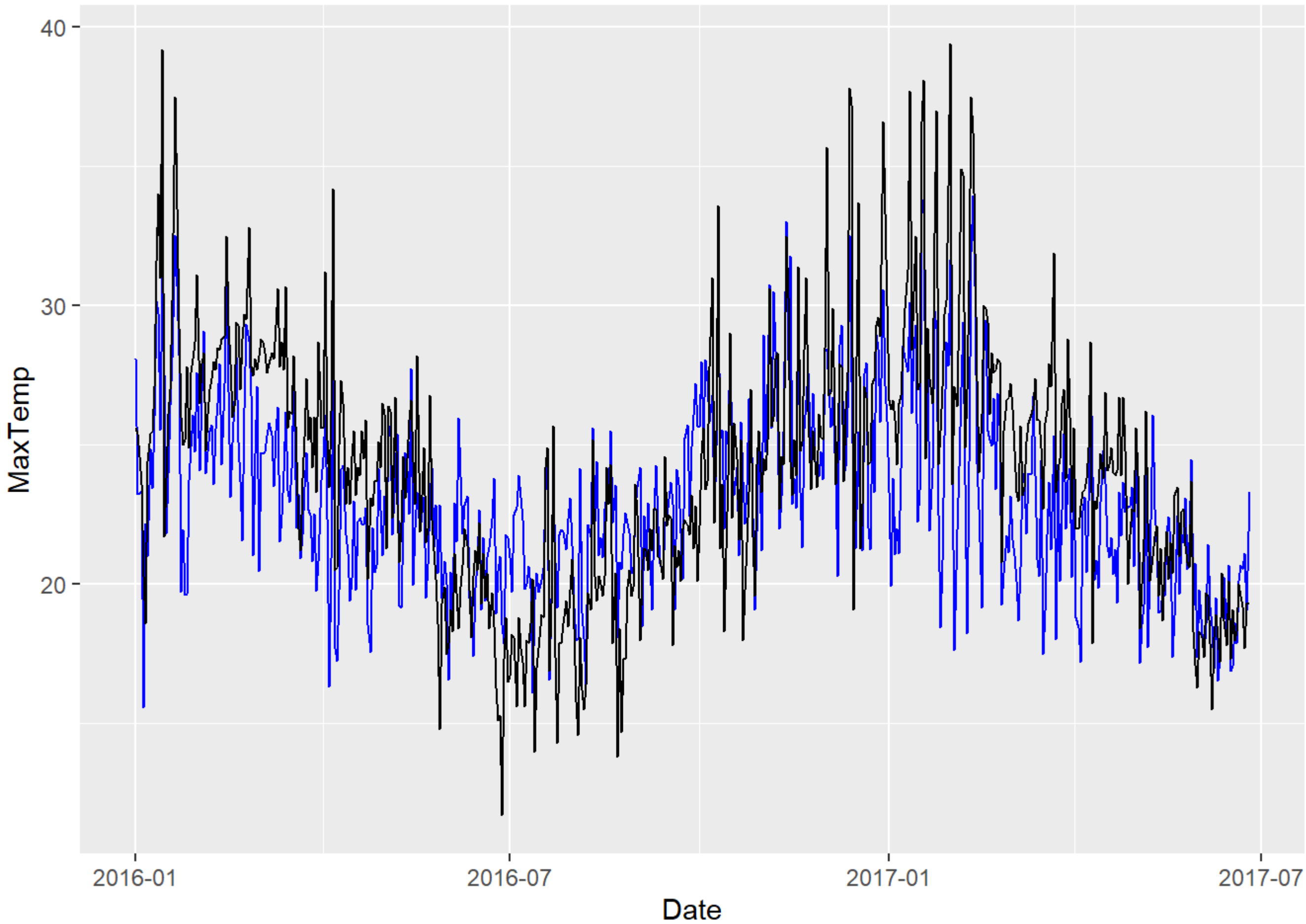
Final Model

```
final_model <- TSLM_Predictors %>% select(lm9)
```

The final model that I will use is lm9 because it had the lowest AICc of any of the models I was able to build.

FORECAST

```
fc <- final_model %>%
  forecast(new_data = test)
fc %>% autoplot(test, level = NULL)
```




```
fc %>% accuracy(test)
```

```
## # A tibble: 1 × 10
##   .model .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 lm9    Test  0.986  3.31  2.65  2.34  11.2   NaN   NaN 0.530
```

I am forecasting about 2 years into the future. This is because the test dataset contains about 20% of the records which ends up being about 2 years out of the almost 10 years of data.

Some considerations when implementing this dataset is that this is the peak temperature recorded of each day, not the average temperature of the day. It is also important to consider that I only focused on temperature in Sydney, not all of Australia. As the many different locations have very different climates and temperatures in the country/continent. So using this forecast to predict on another location would not result in accurate results despite them both being in Australia.