# Datathon Presentation

Krishita Laungani
Daniella Mangibin

# CV_Mape Explanation

**Input:**
‘error’ - this is a vector that contains the forecast errors and the differences between predicted and actual values.

**‘actual’** - contains the time series data that will be compared with the forecasts.

**‘actual_table’ initialization:**
a function that initializes a data frame to store a table of actual values with the same number of rows as the actual vector.

**for loop over time points:**
A loop that iterates from 1 to the length of (actual - value of window). This will construct a table of actual values
**Cross Validation MAPE Calculation:**
Function determines the Mean Absolute Percentage Error

- divides error with the corresponding ‘actual_table’ values
- takes the absolute value and multiplies it by 100 to make a percentage
- na.rm = TRUE is used to remove missing (NA) values.

```
cv_mape = function(error, actual){

  actual_table = data.frame(matrix(NA, nrow =
length(actual), ncol =h))

  for(i in 1:(length(actual)-window)){
if((i+window+h-1)<=length(actual)){actual_table[i+window
-1,]=actual[(i+window):(i+window+h-1)]}


else{actual_table[i+window-1,1:(length(actual)-(i+window-
1))]=actual[(i+window):(length(actual))]}

  }


return(100*mean(abs(as.matrix(error)/as.matrix(actual_ta
ble)),na.rm=T))

}
```

# Performance Table Explanation

**Calculates 3 Metrics:**

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- MAPE (Mean Absolute Percentage Error)

**Metric Formatting**:
formatC() - function to format the calculated metrics to a specific number of decimal places (5) and as a character of strings

- this is to make consistent formatting

**Combining Metrics:**
cbind()- combines the metrics to a single matric and labeled
**Data Frame:**
data.frame() - creates a data frame fro combined metrics matrix into perf_st`
**Format Table:**
kable() - perf_st1 is reformatted to be organized
**Then the table is displayed in an R environment**
.

```
perf_stl=data.frame(rbind(

  cbind('rmse',
        formatC(round(accuracy(data_stl12)[,'RMSE'],5),format='f',digits=5),
        formatC(round(sqrt(mean((data_test-data_stl12$mean)^2)),5),format='f',digits=5),
        formatC(round(sqrt(mean(error^2,na.rm=T)),5),format='f',digits=5)),

  #'mae'=MeanAbsoluteError

  cbind('mae',
        formatC(round(accuracy(data_stl12)[,'MAE'],5),format='f',digits=5),
        formatC(round(mean(abs(data_test-data_stl12$mean)),5),format='f',digits=5),
        formatC(round(mean(abs(error),na.rm=T),5),format='f',digits=5)),
  #'mape'=MeanAbsolutePercentError
  cbind('mape',
        formatC(round(accuracy(data_stl12)[,'MAPE'],5),format='f',digits=5),

formatC(round(mean(100*(abs(data_test-data_stl12$mean))/data_test),5),format='f',digi
ts=5),
        formatC(round(cv_mape(error,data),5),format='f',digits=5))),
  stringsAsFactors=F)
 #Codeforturningthedataframeaboveintoaniceplotusingkable.

kable(perf_stl,caption='Performance-TemperatureAnomalieshorizon=12,window=36',alig
n='r',col.names=c('','train','test','cv'))%>%
  kable_styling(full_width=F,position='l')%>%
  column_spec(2,width='7em')%>%
  column_spec(3,width='4.5em')%>%
  column_spec(4,width='4.5em')
```

# Hypothesis 1: CO2 emissions and temperature anomalies have high positive correlation.

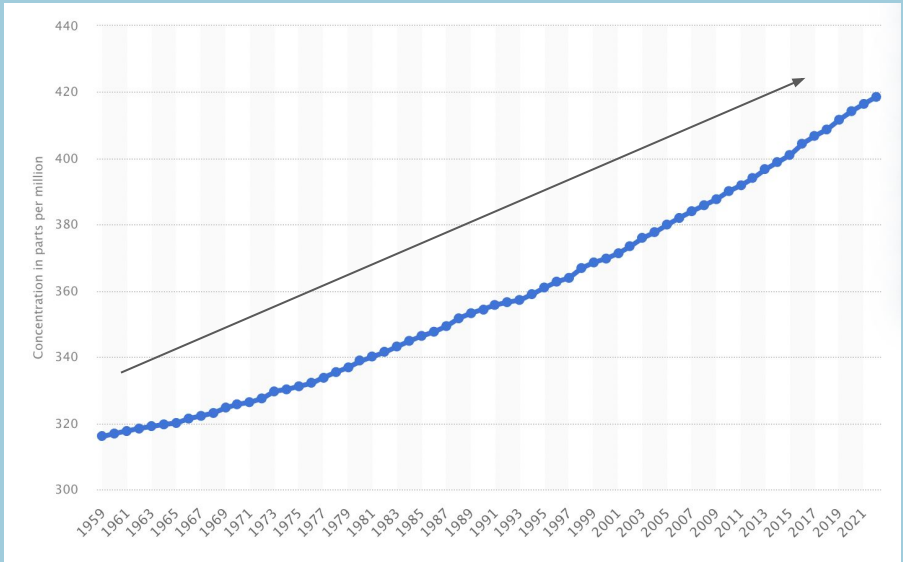
Link for CO2 emissions dataset 1959 - 2022:
https://www.statista.com/statistics/1091926/atmospheric-concentration-of-co2-historic/

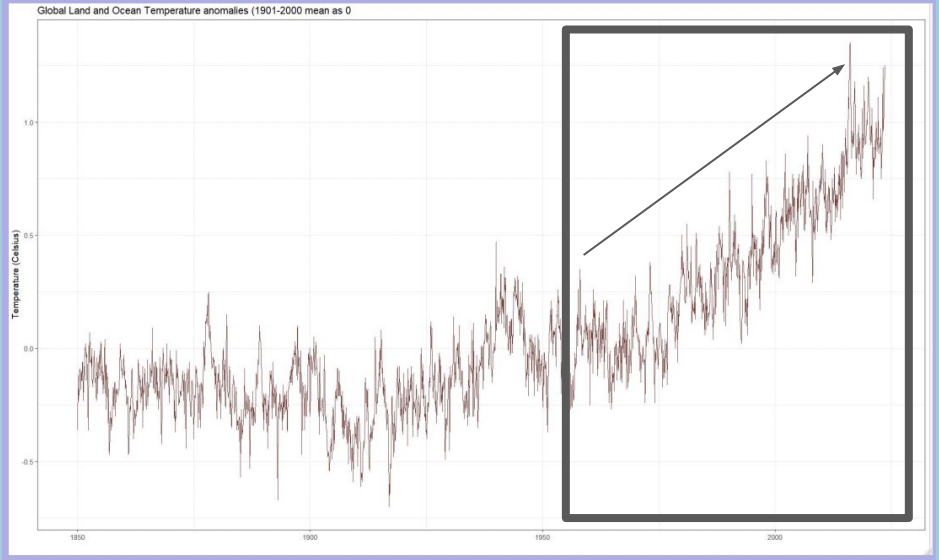
R Code to find Correlation:

```r
# getting data
temp_data_filtered <- subset(temp_data, Year >= 195900 & Year <= 202200)
co2_data_filtered <- subset(co2_data, Year >= 1959 & Year <= 2022)

# yearly global surface temperature
mean_temp <- aggregate(Value ~ Year, data = temp_data_filtered, FUN = mean)

# correlation between CO2 emissions and mean temperature
correlation <- cor(mean_temp$Value, co2_data_filtered$Value)
```

# Findings:

Since 1959, Temperature anomalies have been consistent with a rising amount of CO2 in the atmosphere
Hypothesis conclusion: true.




(r graphs were not loading so used dataset website and time series)

# Further Research:

As CO2 emissions continue to rise, they may pose a threat to nature and subsequently humanity. The NASA article linked below discusses the significant differences in the impacts of global temperature increases of 1.5 degrees Celsius. It highlights that limiting warming to 1.5 degrees Celsius can substantially reduce the risks associated with climate change. The article covers diverse areas such as temperature extremes, droughts, water availability, extreme precipitation, impacts on biodiversity and ecosystems, ocean effects, and human-related impacts. It underscores that taking action to limit warming to 1.5 degrees Celsius is crucial to mitigate the severe consequences of climate change across the globe. Scientists have deemed 1.5 degrees the “point of no return”.

This poses the question of, with current rises in co2, how long before we hit the “point of no return” ?

https://climate.nasa.gov/news/2865/a-degree-of-concern-why-global-temperatures-matter/

# Hypothesis 2: The world will reach a global temperature anomaly of 1.5 by the end of the century(year 2100).

Steps:
First, to build a time series forecast BASED UPON co2 levels, we need to have a forecast of co2 levels till the end of the century.

After making a time series forecast for co2, we used a GLM model to combine the yearly mean anomalies from 1959 to 2022 (as that is where the information stopped on the co2 dataset), and created a time series forecast of anomalies based on this information

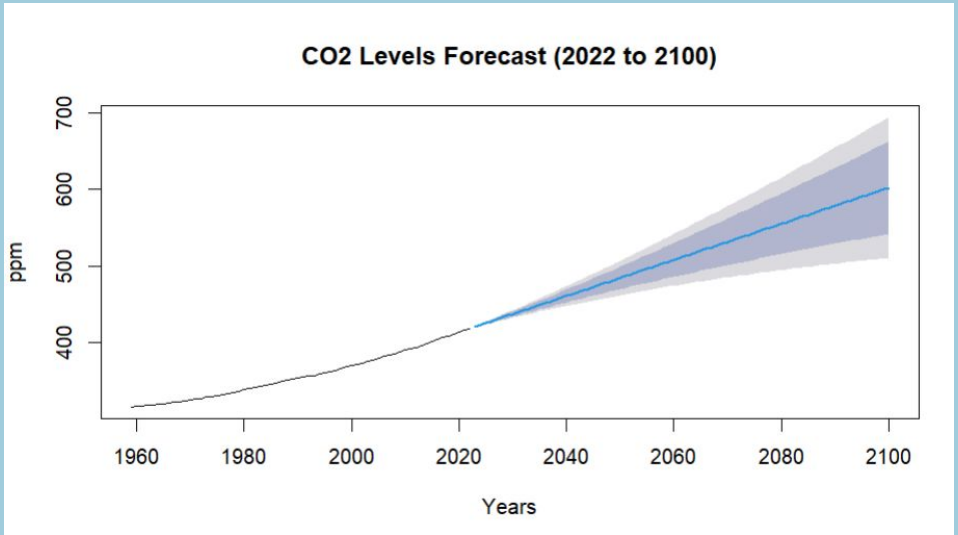
```r
#co2 time series
co2_ts <- ts(co2_data$Value, start = 1959, frequency = 1)

# Forecast CO2 levels from 2022 to 2100
co2_forecast <- forecast(auto.arima(co2_ts), h = 78 )

#forecasted CO2 levels
plot(co2_forecast, main = "CO2 Levels Forecast (2022 to 2100)")
```

```r
# Create a GLM model to predict temperature anomalies based on CO2 emissions
model <- glm(Value.x ~ Value.y, data = merged_data)

# Summary of the GLM model
summary(model)

# Extract the coefficients
coefficients <- coef(model)

# Calculate the correlation between CO2 emissions and temperature anomalies
correlation <- cor(merged_data$Value.x, merged_data$Value.y)

# Plot the relationship and regression line
ggplot(merged_data, aes(x = Value.y, y = Value.x)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "CO2 Emissions", y = "Mean Global Surface Temperature Anomalies") +
  geom_text(aes(label = paste("Correlation =", round(correlation, 2), "\n",
                              "Intercept =", round(coefficients[1], 2), "\n",
                              "CO2 Coefficient =", round(coefficients[2], 2))),
            x = max(merged_data$Value.y), y = max(merged_data$Value.x),
            hjust = 1, vjust = -1)
```

# Findings:

Hypothesis result: Inconclusive because of package and csv error in R Studio.

However, to the right is a forecast of co2 levels made from the code provided in the previous slide. As visible, co2 levels reach almost double the amount of ppm by the end of the century. Because of the positive correlation between the co2 levels in the atmosphere and temperature anomalies, and inference can be made that the temperature will continue to rise. By the end of the century, places near the equator will be almost uninhabitable because of the heat they will endure and the biodiversity that will be shattered.




Source: https://nymag.com/intelligencer/2017/07/climate-change-earth-too-hot-for-humans.html

# Use of Chat GPT

Chat GPT was extremely helpful in looking for datasets to use. Datasets are hard to come by, especially in the format that is needed. Chatgpt provided ideas of where to find datasets and what type of information they would provide. In addition, the AI helped us figure out the GLM model. Because of our newness to R, we needed  way to create a forecast based upon another dataset. In order to learn how to do that we asked for base code and different models in R that could help us.