MI2 Sea Monsters Checkpoint
Group: The Sea Monsters
Group Leader: Alexa Owen Group Members: Audrey Himes, Sharaf Tariq
DS 4002
April 18, 2023

Hypothesis: The data will predict with 75% accuracy the average global temperatures beyond 2015.

Research Question: Can we predict global average temperatures using data collected from 1850-2015?

Model Approach: We will be using the SARIMA (seasonal autoregressive integrated moving average) model for time series forecasting. The SARIMA model uses seasonality as a feature on previous data to make predictions about future values. The SARIMA model is an offshoot of the ARIMA model, which is a very popular tool for time series forecasting.

Executive Summary: This document will detail the details of our dataset and analysis plan. We will discuss our plans to use the SARIMA model for time series forecasting to predict global average temperatures for 2016-present. Additionally, we will conduct some preliminary exploratory data analysis to understand our data before heading into the analysis.

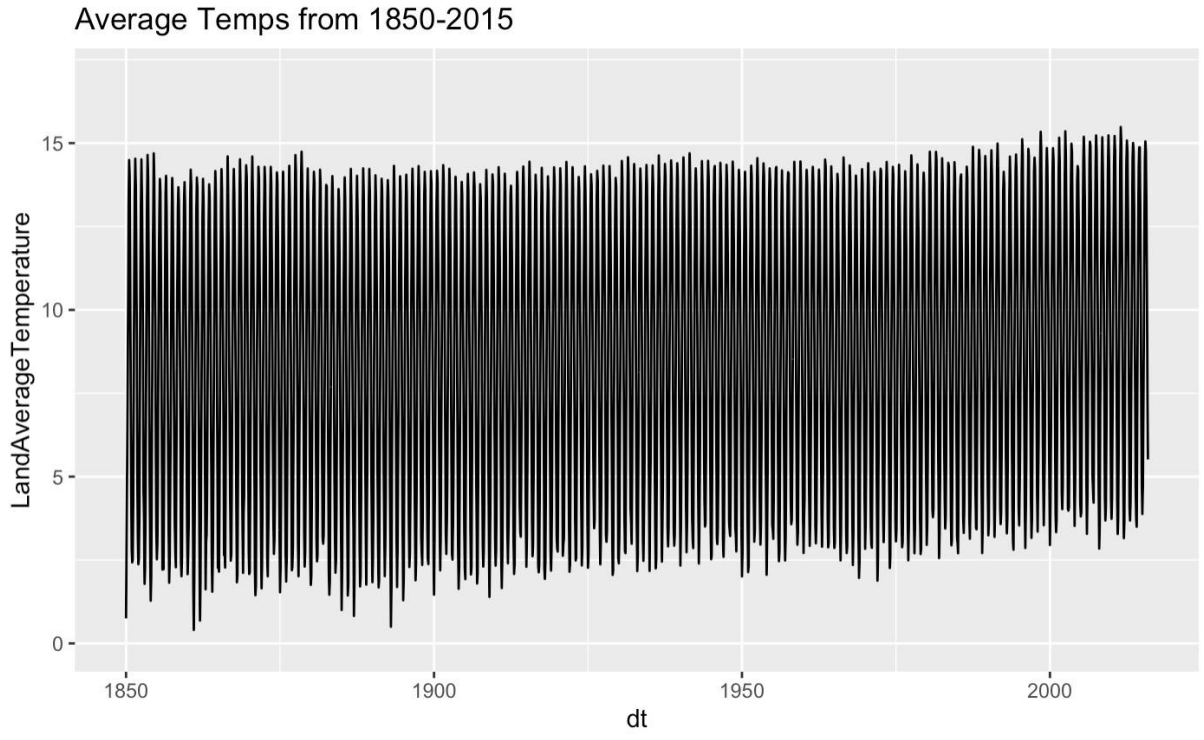
Data Set Establishment Details: The data set contains 3,192 rows and 9 columns of data on global surface temperatures. The columns are the date, average land temperature, average land temperature uncertainty, land max temperature, land max temperature uncertainty, land min temperature, min land temperature uncertainty, land and ocean average temperature, and land and ocean average temperature uncertainty. The data was recorded the first of every month from 1750 to 2015. Only data for the average land temperature and average land temperature uncertainty were recorded from 1750 to 1849. Data for the remaining columns were collected beginning in 1850.

Element/Variable Display Name	Description	Data Type
dt	Date of temperature recording	Character
LandAverageTemperature	Average land surface temperature in celsius	Integer
LandAverageTemperature Uncertainty	95% confidence interval for LandAverageTemperature	Integer
LandMaxTemperature	Max land surface temperature in celsius	Integer
LandMaxTemperatureUncertainty	95% confidence interval for LandMaxTemperature	Integer

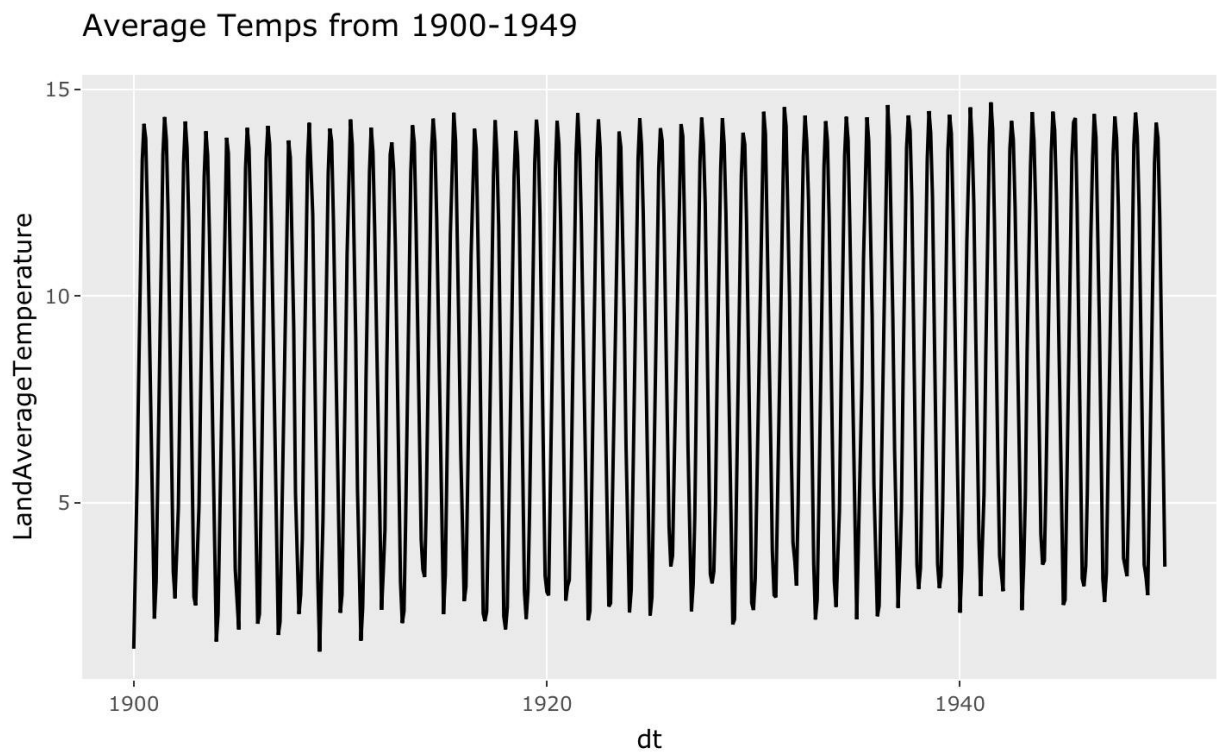
LandMinTemperature	Min land surface temperature in celsius	Integer
LandMinTemperatureUncertainty	95% confidence interval for LandMinTemperature	Integer
LandAndOceanAverageTemperature	Average land and ocean temperature in celsius	Integer
LandAndOceanAverageTemperatureUncertainty	95% confidence interval for LandAndOceanAverageTemperature	Integer

This dataset is available on Data World [1].

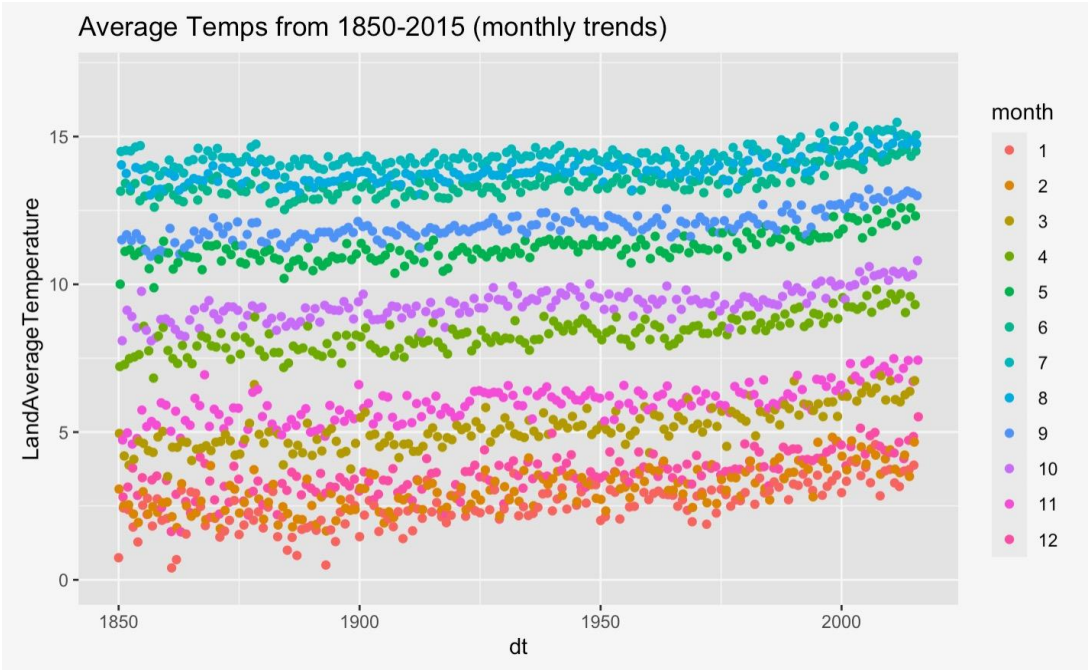
Since the data has been collected from 1850-2015, is there too much data?
At first, this was a concern, but it is important to have thorough and extensive data that can show a long-term trend. When we graph the data in a line graph, however, it is hard to see because the lines are so densely packed in the graph. An overall trend can be seen with the darker areas in the middle of the graph, but it is still very hard to see. We plan on showing the data by 50-year intervals so that the changes in the line graph and the overall trend can be seen clearer. We can see an overall increase in the land average temperature over time based on this graph, but it is important that we can see it clearer.
Below is the line graph of the entire data from 1850-2015.



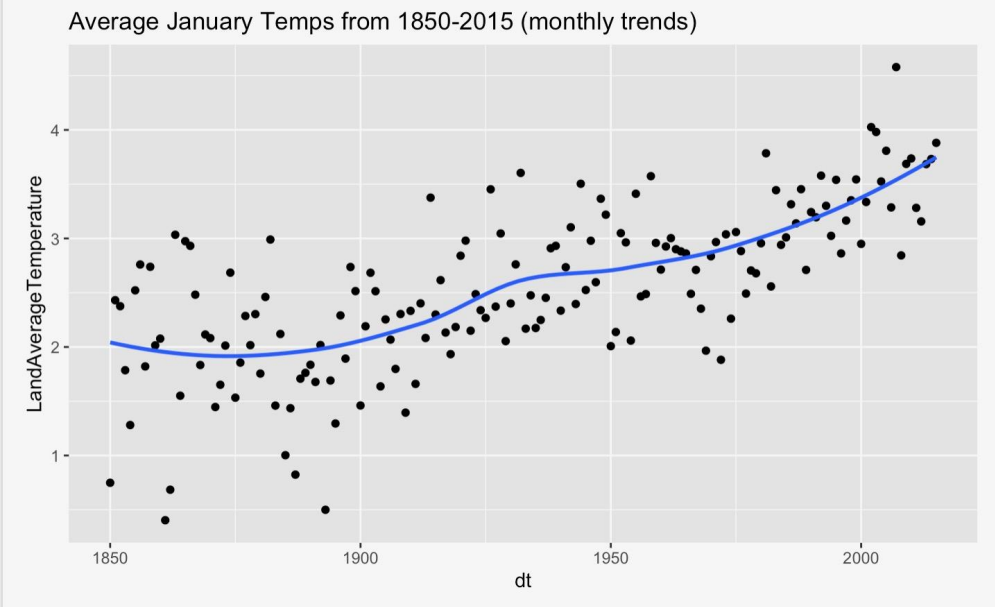
Below is a line graph showing the land temperature by month from 1900-1949. As you can see, the lines are much clearer and it is easier to see the changes of temperature. However, the overall trend is still not very clear.

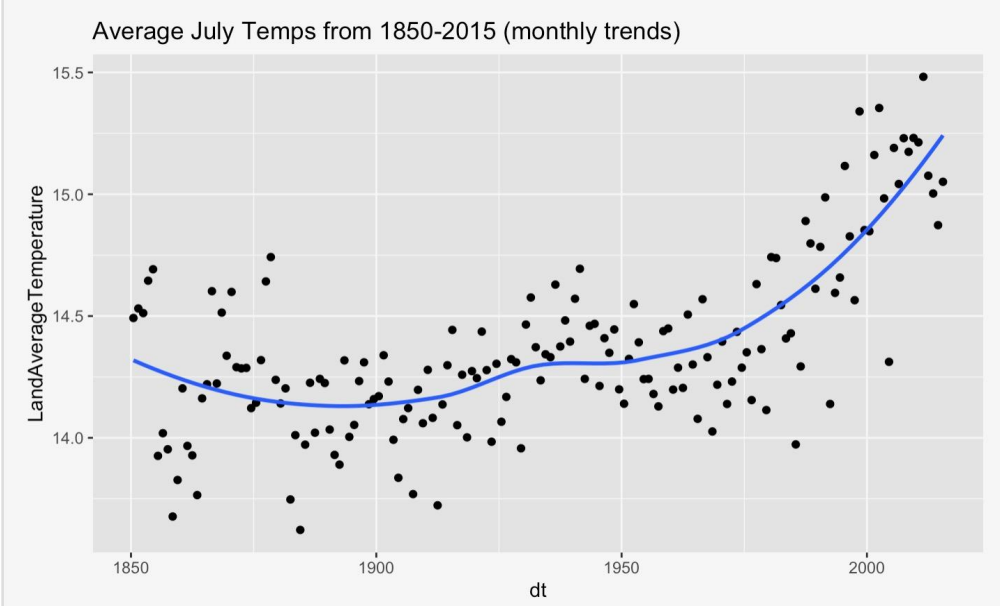


When we separated the data by month and graphed it, an overall trend is much clearer. An overall upward trend is much more obvious in this graph.



Finally, we looked at the trends of every January, the coldest month, and just July, the hottest month. The overall upward trend over time is the clearest in these graphs.





What is the original source of this data?

The original source comes from Berkeley Earth, a non-profit organization based in Berkeley, California. Berkeley Earth was created in 2010, and the organization published their initial findings in 2012. This organization focuses on the environment and provides open-source data on global land temperatures and air pollution data. Their main goal is to provide accessible data [2].

Analysis Plan:

EDA and Data Cleaning:
When we remove na values the data is reduced to 1992 rows.

STL Decomposition:
After cleaning the data and removing na values, we will perform an STL Decomposition (Seasonal-Trend Decomposition using LOESS) on our data to determine the seasonality and to smooth our seasonal and trend data. This model will then minimize the effects of any outliers. Our output will be a series of graphs, one showing the graph our data creates, a trend graph, a seasonal graph, and a remainder graph. The first step is to calculate the seasonal component, and then that is removed to calculate the trend component. The remainder is calculated by subtracting the seasonality and the trends from the time series data [3].

Create ACF and PACF Plots:
We will then create an Autocorrelation Function plot and a Partial Autocorrelation Function plot using the plot_acf and plot_pacf functions respectively. These functions come from the statsmodels.graphics.tsaplots library. The ACF plot will observe whether our data is completely random or if there is a correlation and if it can be modeled with a Moving Average (MA) model. The PACF plot will help us determine if our data can be modeled with an Autoregressive (AR) model and, if so, in what order [4].

Fitting to SARIMA Model:
We will fit this data to the Seasonal-Autoregressive Integrated Moving Average (SARIMA) model to observe the seasonal historic data to forecast future data [5]. We will import the SARIMAX class from the statsmodels library from above and will be tuned by values of p (order of the AR part), d (degree of first differencing involved), and q (order of the moving average part) [6].

Summarize and Visualize Results:
Finally, we will create a plot that compares the actual data with our predicted data from the SARIMA model. The two will be overlaid on top of each other to see how similar or different they are. The accuracy of the model will be determined by calculating the Pearson's coefficient and R squared metrics.

References:
[1] Data Society, "Global Climate Change Data," *data.world,* 2016. [Online]. Available: https://data.world/data-society/global-climate-change-data/discuss/global-climate-change-data/gntggylf. [Accessed Apr. 13, 2023].
[2] Berkeley Earth, "About Berkeley Earth," *Berkeley Earth*, 2012. [Online]. Available: https://berkeleyearth.org/about/. [Accessed Apr. 13, 2023].

[3] ArcGIS, “Seasonal-Trend decomposition using LOESS,” *Esri*, 2023. [Online]. Available: https://doc.arcgis.com/en/insights/latest/analyze/stl.htm.
[4] L. Monigatti, “Interpreting ACF and PACF Plots for Time Series Forecasting,” *Medium*, Aug. 2, 2022. [Online]. Available: https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c.
[5] A. Bajaj, “Arima & Sarima: Real-world time series forecasting,” *neptune.ai*, 26-Jan-2023. [Online]. Available: https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide. [Accessed: 11-Apr-2023].
[6] A. L. Duca, “How to model a time series through a SARIMA model,” *Medium*, Sept. 9, 2020. [Online]. Available: https://towardsdatascience.com/how-to-model-a-time-series-through-a-sarima-model-e7587f85929c.