**Exploration and Preprocessing of Weather Data**

Corey St Jean

Speed School of Engineering, University of Louisville

CSE-632: Data Mining

Dr. Mehmed Kantardzic

September 12, 2024

**Abstract:**

To study the data mining process, specifically data preprocessing, a study into the effects and impacts of climate change was conducted. Preprocessing was performed on six different datasets holding weather data from two different locations: Fayetteville, Arkansas, and Sofia, Bulgaria. For each area, three datasets were collected, each from a different decade (1995, 2005, 2015), leaving us with six datasets to preprocess. But, before preprocessing, different research articles discussing data mining and climate science and how they can be used together, were used to gain an initial understanding of climate change and how current problems are solved using data mining strategies. Then, three hypotheses were generated based on the datasets gathered about how climate change will impact various weather features. The first hypothesis states that wind gusts will trend higher in three separate years showing that temperatures have increased due to climate change. The second hypothesis is that temperature ranges during each month trend higher in three separate years showing climate change through increased temperatures. The final hypothesis states that cloud ceilings should trend upward in three different years due to an increase of trapped heat within the atmosphere heating clouds and causing them to rise. To test said hypotheses, six different data preprocessing steps were conducted starting with data cleaning, followed by filling of missing data, outlier detection and removal using Z-scores, min-max normalization, data aggregation, and regression smoothing. Based on our results, we reject all three hypotheses, but this does not outright imply climate change is not happening. Our results instead show that weather patterns have become more variable and a sort of “seasonal shift” has occurred in both locations with all seasons starting later and ending later.

**Problem Description:**

Climate change is here, and to ensure our survival on this planet, we need to invest time and resources into reducing the adverse effects caused by this condition. But what is climate change and what are some ways we can prove it? Climate change, according to the United Nations, is the long-term shifting of temperature and weather patterns caused by human activities (“What is climate change?”, 2021). Climate change is not new to the earth, with many instances of climate change being discovered throughout the history of the earth, but this is the first time climate change is being caused by humans and not some other natural cause. Humans have caused climate change primarily through the release of greenhouse gases via resource production (coal, oils, plastics, energy, etc.). These greenhouse gases remain trapped in the atmosphere, acting as a blanket trapping the heat energy from the sun within the atmosphere, thus increasing temperatures. Temperature plays a major role in weather, high temperatures evaporate water creating clouds that will cool, condense, and precipitate, warm air travels to cooler spots pushing cooler air into warmer spots creating wind patterns, and so much more. This is why a change in the normal temperatures of the earth can cause hurricanes, floods, droughts, and other extreme weather phenomena that have major impacts on the earth as well as its inhabitants (“What is climate change?”, 2021)

Data mining is a tool that can have major impacts in the climate science field due to its ability to draw insights from large amounts of data spanning large periods. This makes data mining the perfect tool for climate studies because we have been collecting mass amounts of weather data since the mid-twentieth century due to increases in understanding of weather patterns as well as the development of tools to measure weather. We can now use this data to understand relationships within the weather as well as prove the existence of climate change, but to do so we must also understand data mining.

Data mining is defined by IBM as the use of machine learning and statistical analysis to discover patterns and other useful insights from large amounts of data (SOURCE). Within data mining, two different underlying goals will impact the tools used. Some data mining focuses on gaining insights into trends and relationships within the data. This can be accomplished via a plethora of techniques but some of the most popular include decision trees and clustering. The other focus of data mining is prediction where a model is created to take various factors as input and from these predict a value as output. This process is typically a form of artificial intelligence, specifically machine-learning where a machine learning model is trained on a smaller data set containing super accurate data with results, then after training it is tested with data in which the expected output is known. These values are then compared and, based on how close they are, we can tell how accurate our machine learning model is. Sometimes projects only require one of the goals of data mining, but often time the data is fully explored and then a model is created using the insights gained from relationships within the data (IBM source).

The emergence of new data mining technology has inspired many climate and data scientists to start working together to gain more insights into climate change. This includes weather prediction, proof of climate change, and insight into certain changes in weather patterns that are most damaging to the world climate. To gain a better understanding of how the field of data science, specifically data mining, is used in climate science, three academic journal articles were reviewed. The first journal article reviewed is “Data Mining for Climate Change and Impacts” by Auroop R Ganguly and Karsten Steinhaeuser (SOURCE). This paper is mostly a proof of concept for data mining to be used in climate science, so they explore what tools can be used for the best insights as well as how this technology can be used to inform decisions that may have lasting impacts on the climate. Weather data has some unique features that require adaptations to the typical data mining process. This includes features like long-range spatial dependence, long-memory processes, and nonlinear behavior. This essentially means that weather data has extremes, seasonal effects, and correlations that need to be considered and dealt with before data mining. Weather data is most like spatial and spatiotemporal data mining, meaning techniques for dealing with SSTDM (spatial/spatiotemporal data mining) can also be used for weather with some minor changes and tweaks. The paper ends with a case study showing that even basic data mining techniques can bring about new insights into climate change. This case study focuses on using weather data from the past to predict heat waves and changes in temperature because of climate change, and they show a local example dealing with the Africanization of Spain (increasing temperatures and decreasing precipitation) with results found in the study pointing to the Africanization effect being true (Data Mining for Climate Change).

The next research article is a specific study into the creation of Artificial Neural Networks and Decision Trees to gain insights into climate change as well as predict future impacts. The neural network was trained using ten years of weather data collected in Nigeria to be able to predict weather trends like wind, temperature, precipitation, evaporation, and more. A decision tree was trained on the same data but was used to gain insights into the correlation between various features of the weather data instead. This model was able to predict said weather patterns based on the current weather trends with reasonable accuracy, meaning that with more training, this model could be used to accurately predict weather patterns worldwide (Application of Data Mining Techniques).

The final research article once again aims at making a model or technique for predicting weather in a specific region that can be scaled to predict worldwide climate change. Chinese and Mongolian weather data was studied to gain insight into how various biomes with different physical features are impacted differently by climate change. Empirical mode decomposition was used to find trends in the data which was visualized using 3D surface maps and 2D contour maps. The results gained from the study include a mass temperature increase throughout all biomes studied, temperature increased faster in dry regions (deserts) than in wet regions (wetlands), and precipitation increased around water features but decreased everywhere else. A small study into climate changes between stations was also conducted with less variation in temperature results but more variation in precipitation results. This study showed some of the factors complicating data mining for climate change, like physical characteristics, and how they can be used to gain greater insights (investigating long-term trends of climate change).

For this project, we are tasked with gathering six different datasets to explore the effects and proofs of climate change. These six datasets include three from one city in the US and three from one city in Europe. Each dataset from each city occurs within different decades to see the change in weather patterns over time. With these datasets, three hypotheses will be created to correlate changes in weather patterns with the effects of climate change. Various preprocessing techniques will be implemented to first clean the data and then draw conclusions about these hypotheses through visualizations.

**Data Gathering:**

To begin the data gathering process, NOAA was data was explored through the National Centers for Environmental Information website. This website holds weather data dating back to the 1960s for various locations across the world, but only two locations, one in the US and one in Europe, need to be selected. To start with the US, a local location was chosen because of personal knowledge about the climate as well as personal impacts of climate change. The US location chosen was Fayetteville, Arkansas, a local college town with hot-humid summers, and short dry winters. The years of data to be collected were simply taken from ten-year gaps starting with the year 1995 meaning our three years of data will be 1995, 2005, and 2015. This is to ensure that we have similar data between years as well as data from a relatively recent period, but not too recent where some data may be missing or yet to be included. For the European city, research into similar climates across the world was done to find cities in Europe with a climate most like Fayetteville, Arkansas. Not many European cities get as warm and humid as the US south, so this limits our region to the Balkans region (Greece, Bulgaria, Bosnia, etc.). The city in the Balkans with a climate relatively like Fayetteville is Sofia, Bulgaria. This is the capital city of Bulgaria located in a very mountainous region making the summers much cooler and less humid, but winters are similar, both being short and barely dropping below freezing.

Our datasets contain a variable number of features, and many of these features contain very few samples due to measurements being taken more sparingly. These features will not be considered to hypothesize about their changes due to climate change. Many features fit these characteristics, but some examples are AA1, AA2, AG1, AJ1, GD1, GD2, etc. Some features are static like physical features and locations that repeat the same value for every sample. Finally, there is also some data stating information about how the data was collected, how good the quality is, and so on, this data will not be helpful. The data we will consider includes three specific features: WND, CIG, and TMP. WND describes wind measurements, and five different values are included in the order: WD, WDQC, WTC, WS, and WSQC. WD is the Wind Direction giving the direction the wind is blowing in degrees, WDQC is the Wind Direction Quality Code giving a report of the quality of wind direction data with a range of 0-9, 0 being the best, WTC is the wind type code describing the character of the wind with N being Normal and most commonly present. The final two values are Wind Speed (WS) measured in meters per second and Wind Speed Quality Code (WSQC) once again reporting on the quality of data collection. CIG describes the cloud coverage and height and five different values are included in the order: CH, CQC, CDC, and CC. CH is the ceiling height which is the height of the lowest layer of clouds measurement in meters, CQC is the Ceiling Quality Code once again describing the quality with a range of 0-9, CDC is the Ceiling Determination Code describing what tool or device was used to collect the cloud data, and the final value CC is the CAVOK Code asking whether the ‘Ceiling And Visibility O.K.’ with Y meaning yes and N meaning no as typical values. The ceiling height category also has values of ‘22000’ which represent infinite cloud height or the days with no clouds, so these values will have to be dealt with specifically. Finally, TMP describes temperature data with two different values in the order: T and TQC. T is temperature measured in degrees Celsius, but with the tens decimal place included (for example, 14.5 degrees Celsius is saved as 145.0 degrees) and TQC is Temperature Quality Code describing the quality of measurement with the range of 0-9.

Our CIG feature has a lot of its data missing, which is most likely due to recording occurring less frequently. Since weather, wind, and cloud data tend to have high variance, we should be okay with replacing missing values with a static value that is calculated from the included data. This value will be the monthly median for that specific value, this is because weather tends to reflect the current season, so by finding the median of all values within a month, we should get a reliable estimate within a normal range of values for that time of the year. This is to see if using the median value produces a skew to the results obtained and if so, how much of a skew is present. Our WND feature also contains some missing values, but these values occur much less frequently and just like the CIG feature will be replaced with a monthly median value. The TMP has very few missing data entries, but when present they will be dealt with the same.

When it comes to the noisiness of the three features we wish to explore, all these features exhibit some amount of it. This is because weather data will always have outliers, everyone has experienced a super warm day in December or a snowstorm in the spring which are both examples of outliers that create noise in the data. To help differentiate between what data is typical and what data might be an outlier a Z-score will be used. This will help us understand how certain values relate to the mean or average and how far away the values are from the average. We can set a specific distance from the average in the form of standard deviations past which we consider values as outliers, but as mentioned previously, an entire year’s average of weather patterns will not accurately reflect seasonal changes. So, we will need to perform a Z-score analysis either monthly or seasonally that will help us determine if a value is an outlier for that specific season of that specific year.

From here, we also need to ensure that weather measurements are consistent regardless of the year they were taken or how measurement devices have changed. To do this, all values are to be compared with values from another dataset (Arkansas vs Sofia, Bulgaria, 1995 vs 2005, etc). This means that the WND, CIG, and TMP features need to be normalized so that they can be effectively compared. This normalization will also help us deal with the unconventional units used for the TMP feature as well.

With data cleaned, missing data filled, outliers removed, and data normalized, data can now be aggregated to help easily compare out datasets and help prove or disprove hypotheses. The specific aggregations required include averages, standard deviations, and variance measures for temperatures, wind speeds, and cloud ceiling heights for every month and year. Comparing averages can quickly let us know if things are trending upwards or downwards, standard deviation and variance will give us insights into how dispersed data is compared to the mean, which should give us insights into how the stability of specific weather measurements has changed. We can also consider things like the range of values (Max – Min) for each month and year that can show us if temperatures are all around trending upwards (warmer winters and summers), or if the range of weather is changing (colder winters and warmer summers).

Another necessary preprocessing step is to remove data collection indicators. These are samples included within the data to signal the start or end of a period of data collection. Within the data sets, these signals are easily noticeable with all samples having the values of 9. These rows need to be removed from the datasets to ensure their values will not impact the quality of the results obtained.

The final step to be completed during preprocessing will be regression smoothing. This will be included to reduce variance further because weather data is prone to variation. Weather patterns are unpredictable with large variances between days and sometimes even hours within the day, regression smoothing will help reduce this variance within graphs to return trends within the data so we can conclude our hypotheses. But to ensure that data skewing is not happening, figures and graphs will be shown before and after exponential smoothing.

**Hypotheses:**

With the data being collected and explored, hypotheses can now be introduced that we can prove or disprove through data preprocessing. Our first hypothesis is that wind gusts will trend higher in three separate years showing that temperatures have increased due to climate change. The logic behind this hypothesis has to do with the logic behind wind in the first place. Wind is created by warm air full of energy that wants to move to a cooler area to help it reach equilibrium, this cold air is then pushed by the warm air moving in which moves the cold air to warmer areas where it heats until it eventually pushes cold air out of a cooler area. This looping cycle is what moves clouds, releases precipitation, and is the wind we feel every day. So, if temperatures increase, we should expect more aggressive movement of this hotter air to the cooler areas which should be measurable through an increased wind gust speed.

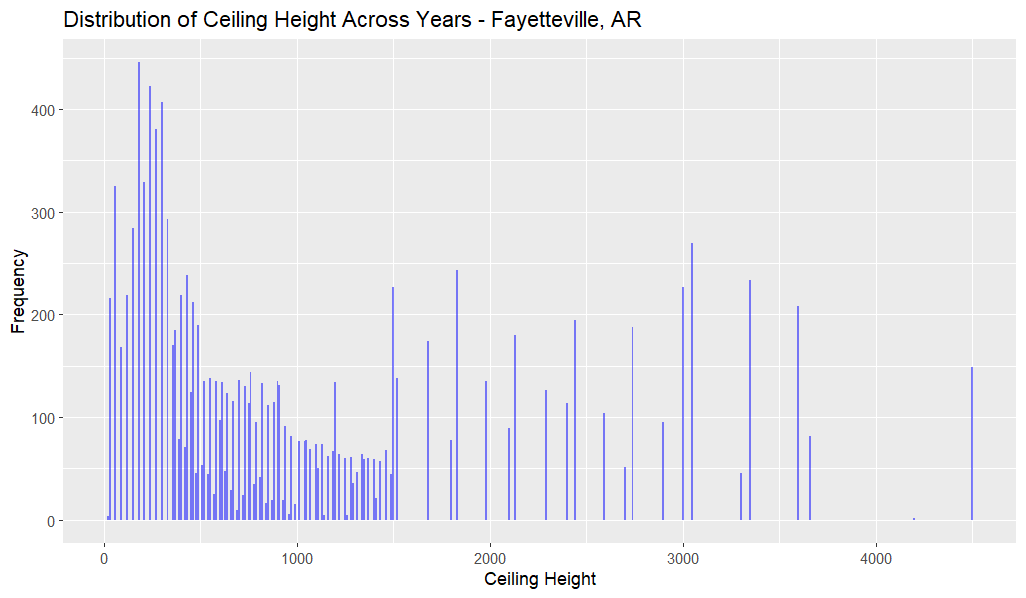
Our next hypothesis is that the temperature ranges during each month trend higher in three separate years showing climate change through increased temperatures. While on a day-to-day basis, climate change might not be extremely clear, it should be present through increased averages and more weather oddities like cold and heat waves. So, if climate change is present, we should see the average monthly temperatures be higher in later years as well as much more extreme maximum and minimum temperatures per month and entire year.

The final hypothesis is that the cloud ceiling should trend upward in three different years due to an increase of trapped heat within the atmosphere heating clouds and causing them to rise. The logic behind this hypothesis is very similar to the logic behind hot air balloons. Just like how the hot air moves to cooler areas to create winds, this phenomenon also happens in the vertical direction with the hot air rising, but the physics behind this is different. When hot air rises, it has more to do with the density of warm air compared to cold air, warm air is less dense, and cold air is denser, so the warm air will essentially “float” on top of the cooler air. This effect also occurs in clouds, with warmer clouds being less dense than the air surrounding them causing the clouds to lift, so with an increased temperature due to climate change, we should see cloud ceiling heights increase as time goes on and climate change worsens.

**Preprocessing:**

To start preprocessing, the data to be processed needs to be isolated to not waste time and resources processing all the data while only using a couple of the features. This will also allow us to focus on our preprocessing steps to ensure they are applied correctly while also focusing on data that will help us conclude the hypotheses. All features other than the four main ones will be removed within Excel, this includes static data like the physical location of the weather data collection center, tools used, etc., as well as data collected that will not help in our hypotheses testing (visibility, rainfall, pressure, ocean measurements, dew point, etc.). These features or columns are simply deleted within Excel. The data features we are left with are the DATE, WND (wind), CIG (cloud ceiling), and TMP (temperature) features. These features contain flags like indicators of measurement type and quality as well as other measurements like cardinal directions, but these measurements are not important to our hypotheses testing, so they will be removed in Excel by delimiting these features and then deleting columns not containing the necessary measurements. Finally, to make data processing and replacement of missing data easier in R, our next tool, all missing values will have the placeholder removed within Excel so they are replaced with ‘NA’ in the .csv file and easier to target in R. The data uses a placeholder of 9’s where data is simply filled with 9’s in the case it is missing, these columns are simply replaced with NA within Excel using the Find and Replace tool.

Now that our data is properly cleaned, we can export the datasets as a .csv file from Excel and import them into R using RStudio. Here, the replacement of missing data can begin. As mentioned previously, we will replace missing data with a monthly median to account for seasonal weather changes. We use a median because only using an average leads to unintentional skews in the data since outliers have not been dealt with. If a month had an uncommonly cold week, this would drastically decrease the mean which could introduce a skew if we fill missing values with this mean. Using a median value will help avoid this skew that would occur if averages were used. This will tell us which value is directly in the middle of the range of weather experienced, reducing the skew because the presence of an outlier has little to no impact on the median value. To accomplish this, an additional dataset was created for each dataset, this new dataset simply contains the months of the year (1-12) along with the calculated monthly median value for that specific year. Then, this new dataset is joined with the dataset it was created from, so for example, if the monthly medians were calculated for Fayetteville, Arkansas in 1995, this dataset will be joined via a left join with the original dataset. With these datasets joined, we now can replace missing data using an if-else statement to check if the data is missing, and if it is then replace said missing data with the monthly median value for that feature. This was done individually to all three datasets, but these datasets were combined into a single histogram to show the before and after of each feature for each region (Arkansas and Bulgaria), but only the visualizations for Fayetteville will be shown in the text to save space, these visualizations are A graph of a ceiling height

Description automatically generatedincluded in Appendix B.

*Cloud ceiling height distribution before and after missing data is filled*

A graph of a graph showing the temperature

Description automatically generatedA graph showing the temperature of years

Description automatically generated*A graph of a wind speed

Description automatically generatedA graph of a wind speed

Description automatically generatedWind speed distribution before and after missing data is filled*

*Temperature distribution before and after missing data is filled*

From here, we can now figure out which values are outliers and remove them to reduce their impact on our results. To calculate outliers, z-scores will be calculated for all values within each dataset. To calculate these Z-scores, a new dataset was created for each dataset we are preprocessing, this dataset holds the means and standard deviations for each of the main features we are exploring (WND, CIG, TMP) that are calculated within R. The mean and standard deviation calculations were grouped by month to avoid seasonal changes being counted as outliers. From here, this dataset is joined with our original dataset allowing us to associate Z-scores with the values they were calculated from. Now we have one dataset containing values as well as their calculated Z-scores so we can now figure out what counts as an outlier and remove them. The threshold Z-score determined was three. This means that all values above three and less than negative three are considered outliers. These values were removed simply by a conditional statement replacing values with a Z-score above three with ‘NA’ so that they are not graphed and not considered in calculations. The threshold was determined by first seeing the range of values using an R statement and then basing the threshold on that range as well as industry standards. To visually express the changes made during outlier removal, scatterplots A graph of a number of blue dots

Description automatically generatedshowing distributions before and after outlier removal are shown below. In these scatterplots, A graph of a number of blue dots

Description automatically generatedyou can see the outliers are removed based on their calculated Z-score.

*A graph of a number of blue dots

Description automatically generated*A graph of a number of blue dots

Description automatically generated*Cloud ceiling height distribution before and after outlier removal*

*Wind speed distribution before and after outlier removal*

A graph of blue dots

Description automatically generatedA graph of a number of blue dots

Description automatically generated with medium confidence*Temperature distribution before and after outlier removal*

Now data has been cleaned, missing data has been filled, and outliers have been removed and we can move on to normalizing the data. To normalize the data, first, we must pick how we plan on normalizing the data. For weather data, min-max normalization should provide us with quality data on a constant scale regardless of when data was collected or where. The scale of this normalization will be zero to one with zero being the smallest (min) value and one being the largest (max) value. To accomplish this, a function was defined within R to allow for easy application of said function. This function takes a feature as input, finds its max and min values, and then applies the formula: [x – min(x)]/[max(x) – min(x)] where x is the current sample. This function was applied to all data features within all six datasets adding three columns to each dataset for the normalized data. To show how the scale changed without the actual values being impacted within said scale, line charts are shown below before and after normalization. By focusing on the scale between these graphs, one can tell that trends within the graph are conserved while the data is scaled between 0 and 1:

A graph of time and time

Description automatically generated with medium confidence*Line plots of features over time before normalization for Fayetteville, AR 1995*

A graph of time and time

Description automatically generated with medium confidence*Line plots of features over time after normalization for Fayetteville, AR 1995*

With data normalized, we can now begin to aggregate data to help conclude the hypotheses. This will include monthly averages, standard deviations, and variances that can be compared across years to see the effects of climate change and how they impact our hypotheses. To hold these aggregate values, new datasets were created to hold the normalized data grouped by month and then also calculate a mean, standard deviation, and variance for the three main features within each dataset. This will leave us with six new datasets, one for each original dataset, that now hold the mean, standard deviation, and variance values grouped by month. This will allow us to easily compare monthly averages between years to measure the effects of climate change. All formulas for calculating these values are the sample formulas using tools native to the language R. To show the difference in data before and after aggregation, line charts are graphed before and after showing the feature (Wind, Temperature, Ceiling Height) plotted against time A graph of a graph showing the height of a number of people

Description automatically generated with medium confidence(monthly).

*Monthly CIG averages, standard deviations, and variances plotted against time for Fayetteville, AR*

A graph of different colored lines

Description automatically generatedA graph with numbers and lines

Description automatically generated*Monthly TMP averages, standard deviations, and variances plotted against time for Fayetteville, AR*

*Monthly WND averages, standard deviations, and variances plotted against time for Fayetteville, AR*

A graph of a graph showing the number of the stock market

Description automatically generated with medium confidenceThe final necessary step is to smooth the visualizations created in the previous step to extract clear trends and the overall big picture from the data. This will essentially take choppy, variable data and smooth out the trend line so that one can easily interpret results. To smooth said visualizations, a local regression model will be used. Within R, this implementation is done within the visualization code, but instead of using a line chart, a smoothed line is added with the method set to ‘loess’ which is local polynomial regression fitting. Additional parameters include se (standard error) being set to false, and span being set to 0.3.

*Monthly averages of TMP plotted against time for Fayetteville, AR*

A graph of a graph showing the number of the year

Description automatically generated with medium confidence*Monthly averages of CIG plotted against time for Fayetteville, AR*

A graph with different colored lines

Description automatically generated

*Monthly averages of WND plotted against time for Fayetteville, AR*

**Visualizing patterns:**

Hypothesis #1:

The first hypothesis states that wind gusts will trend higher in three separate years showing that temperatures have increased due to climate change. Based on the Fayetteville graph below, one can infer that wind has not necessarily increased because of climate change, meaning that we will reject our first hypothesis based on this graph. This does not mean that the effects of climate change are not displayed in our results though, we can see that on average, summers tend to have low wind and winters tend to have high wind, but as time went on winters became more variable when it comes to wind speeds and summers became less windy. This could imply that the air has become more stagnant in the summer due to a decrease of cold temperatures in surrounding areas decreasing the total amount of wind during the summer. When it comes to the graph for Sofia, Bulgaria, one can see an increase in wind speeds during the summer months, with early winter having a reduced wind speed average. Based on this graph we could conditionally accept our first hypothesis with the condition being only within summer months. Regardless of our hypothesis, one can see the impacts of climate change with irregular patterns beginning to emerge more and more. Between both graphs, it almost appears that a shift has occurred with summers lasting later into the year and winter weather lasting well into the spring.

A graph of different colored lines

Description automatically generatedA graph with different colored lines

Description automatically generated*Monthly averages of WND plotted against time for Fayetteville, AR*

*Monthly averages of WND plotted against time for Sofia, BU*

Hypothesis #2:

A graph of a graph showing the number of the stock market

Description automatically generated with medium confidence The second hypothesis states that the temperature ranges during each month trend higher in three separate years showing climate change through increased temperatures. Based on the data from Fayetteville, AR, one can see that this effect mostly depends on the specific time of the year. During the summer, temperatures are most definitely higher, additionally, the summer heat seems to last much longer with summer temperatures lasting well into October in 2015. This is drastically different from 1995 and 2005 where summer temperatures decreased by September. It can also be noted that winters seem to be getting colder with minimum temperatures lowering as time passes. These two effects are both accepted and recorded impacts of climate change meaning that our hypothesis can be accepted for this graph. The next region, Sofia, Bulgaria, had contradicting results though. Sofia had a decrease in temperature during the summer month with each decade having slightly reduced temperatures compared to the previous. This means that for this graph we would reject the hypothesis because temperatures did not trend higher. But this does not disprove climate change since the climate is still changing, just in a different way to Fayetteville, AR.

*Monthly averages of TMP plotted against time for Fayetteville, AR*

*A graph of a graph of a number of numbers

Description automatically generated with medium confidenceMonthly averages of TMP plotted against time for Sofia, BU*

Hypothesis #3:

The third and final hypothesis states that the cloud ceiling should trend upward in three different years due to an increase of trapped heat within the atmosphere heating clouds and causing them to rise. Cloud height typically follows a pattern of low clouds during the winter and season transitions (spring to summer, summer to fall) and high clouds during each respective season. This is true of both Fayetteville and Sofia and can be seen in the 1995 graphs. But as you can see this pattern quickly breaks down as the decades continue, showing drastic changes in cloud height patterns. In 2005, cloud height was much higher in the spring than in 1995, and this trend continued with spring 2015 having a cloud height comparable to summer cloud height. During the summer cloud height did not trend upward and stayed quite consistent, but during the wintertime, cloud height was much higher as time went on for Fayetteville. The graph for Sofia shows a similar breakdown of typical patterns, but much more drastic. Every year, summer cloud height trended upward, and winter cloud height had a much more interesting trend, with it increasing in 2005 and decreasing during 2015 back to 1995 levels. Based on these graphs, we can accept our hypothesis conditionally within the summer, but cannot accept our hypothesis for winter, spring, or fall. Once again, these graphs are displaying the impacts of climate change, but A graph of a graph showing the number of the year

Description automatically generated with medium confidenceour hypothesis did not correctly cover the expected and measured effects.

A graph with numbers and lines

Description automatically generated*Monthly averages of CIG plotted against time for Fayetteville, AR*

*Monthly averages of CIG plotted against time for Sofia, BU*

**Conclusions:**

Climate change is here, and to ensure our survival on this planet, we need to be investing time and resources into reducing the adverse effects caused by this condition. We have explored weather data to discover some of these impacts to see the current, measurable impacts of climate change. Based on the data gathered, three hypotheses were generated, the first stating there should be an increase in wind, the second for an increase in temperature ranges, and the third that cloud height should increase over time. Based on the findings and visualizations created from the data, we now know that all hypotheses can be rejected at face value. This is because the hypothesized features did not express said increases, but the impacts of climate change are still evidently clear. For the first hypothesis, we saw summer winds trend downwards in Fayetteville and trend upward in Sofia, but even separate from the trends, one can see that the wind is much more variable as time goes on. This means that if I were to restate the hypothesis in a form that would be accepted, I would restate the hypothesis as so: There should be an increase in the variability of wind speeds during each year along with seasonal shifting (summer lasting into October, winter lasting to March, etc.). For the second hypothesis, we did not see the expected rise in temperatures across the board as expected, instead, we once again saw an increase in variability as well as seasonal shifting. This is why there was a big change in the climate science community to change the term from “global warming” to “climate change” because there will not always be an across-the-board increase in temperature in every location, weather is complicated, and we should expect to see different results in different locations. The common characteristic between these locations is the increase in variability and shifting temperature ranges. If I were to restate this hypothesis, I would state that there should be a measurable increase in temperature variance as well as a shifting pattern emerging. For the final hypothesis, we saw that cloud height was another factor impacted by climate change, but not in the way expected. Just like how temperature did not increase across the board, the cloud height did not as well, meaning that the final hypothesis is rejected as well. But once again, this is mostly due to it being too general to truly describe the effects of climate change, and if it were restated, it would state that cloud ceiling height variability should trend upward in three different years due to impacts from climate change and temperature shifts. So, if retesting was possible, the steps would most likely look the same, but with hypotheses being more focused on how each season is impacted differently or how variability in weather has increased.

**Appendix A:**

Ganguly, A. R., & Steinhaeuser, K. (2008, December). Data mining for climate change and impacts. In 2008 IEEE international conference on data mining workshops (pp. 385-394). IEEE.

Holdsworth, J. (2024, June 28). What is data mining?. IBM. https://www.ibm.com/topics/data-mining

Olaiya, F., & Adeyemo, A. B. (2012). Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1), 51.

United Nations. (2021). What is climate change?. United Nations. <https://www.un.org/en/climatechange/what-is-climate-change>

Xie, Y., Zhang, Y., Lan, H., Mao, L., Zeng, S., & Chen, Y. (2018). Investigating long-term trends of climate change and their spatial variations caused by regional and local environments through data mining. Journal of Geographical Sciences, 28, 802-818.

A graph of a wind speed

Description automatically generatedA graph of a graph

Description automatically generated**Appendix B:**

A graph showing the temperature

Description automatically generatedA graph showing the temperature

Description automatically generated*[1] Distribution of wind speeds before and after data filling plotted against time for Sofia, BU*

A graph of a ceiling height

Description automatically generatedA graph of a ceiling height

Description automatically generated*[2] Distribution of temperatures before and after data filling plotted against time for Sofia, BU*

*[3] Distribution of ceiling height before and after data filling plotted against time for Sofia, BU*

A graph of red dots

Description automatically generatedA graph with red dots

Description automatically generatedA graph with red lines

Description automatically generatedA graph with red lines

Description automatically generatedA graph of red dots

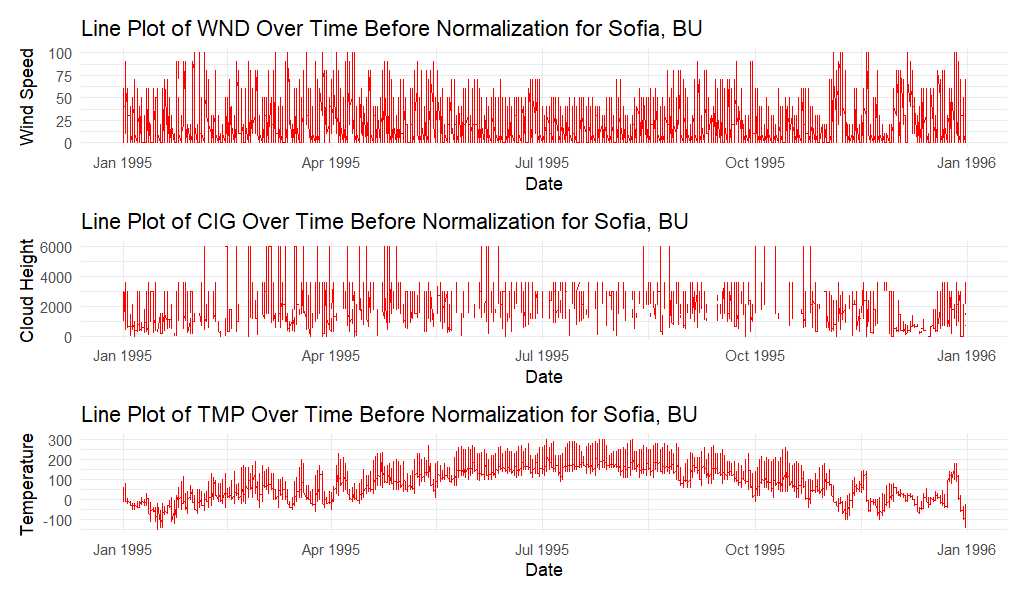
Description automatically generatedA graph of a number of red dots

Description automatically generated*[4] Distribution of wind before and after outlier removal for Sofia, BU*

*[5] Distribution of temperature before and after outlier removal for Sofia, BU*

*[6] Distribution of ceiling height before and after outlier removal for Sofia, BU*

A graph of time and time

Description automatically generated with medium confidence

*[7] Line graph of Sofia, BU features from 1995 before normalization*

*[8] Line graph of Sofia, BU features from 1995 after normalization*

A graph of time and date

Description automatically generated with medium confidenceA graph of time and date

Description automatically generated with medium confidence*[9] Line graph of Sofia, BU features from 2005 before normalization*

*[10] Line graph of Sofia, BU features from 2005 after normalization*

A graph of time and date

Description automatically generated with medium confidenceA graph of time and date

Description automatically generated with medium confidence*[11] Line graph of Sofia, BU features from 2015 before normalization*

*[12] Line graph of Sofia, BU features from 2015 after normalization*

A graph of a graph showing the height of a ceiling height

Description automatically generated with medium confidenceA graph showing different colored lines

Description automatically generated*[13] Monthly averages, standard deviations, and variance for temperature in Sofia, BU*

*[14] Monthly averages, standard deviations, and variance for ceiling height in Sofia, BU*

A graph of a graph of a number of numbers

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generated*[15] Monthly averages, standard deviations, and variances for wind speed in Sofia, BU*

*[16] Smoothed version of graph [13]*

A graph of different colored lines

Description automatically generatedA graph with numbers and lines

Description automatically generated*[17] Smoothed version of graph [14]*

*[18] Smoothed version of graph [15]*