Joe Fragala (Level 4000)
Professor Munasinghe
Data Analytics
20 April 2024

Assignment 6:

General assignment: Your term projects should fall within the scope of a data analytics problem of the type you have worked with in class/ labs, or know of yourself – the bigger the data the better. This means that the work must go beyond just making lots of figures. You should develop the project to indicate you are thinking of and exploring the relationships and distributions within your data to lead to optimized predictive models. Start with a hypothesis, claim, or questions. Think of one or more ways to construct model(s), find or collect the necessary data, and do both preliminary analysis, detailed modeling, validation, summary (interpretation) and (if any) resulting decisions.

Note: You do not have to come up with a positive result, i.e. disproving the hypothesis is just as good. Use the section numbering below for your written submission for this assignment.

1. Abstract and Introduction (2%) Describe your motivation, initial hypothesis/ idea that you wanted to investigate, and if applicable any prior work, interest in the topic (like an intro for a paper, with references), Minimum 1/2 page:

   Abstract: The health and vigor of vegetation are fundamental to our planet's ecosystem. This study explores the relationship between Land Surface Temperature (LST) and vegetation health, as indicated by the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI).

   This project originally embarked as a way to derive the temperature differences between urban areas and natural areas, but as I started coding I realized this would be extremely tedious to separate the urban areas from the natural areas and include all these areas from around the world to get a good mix of data. While exploring this, I found myself with an interesting compilation of results and data which ultimately changed the course of this project and changed my hypothesis.

   The hypothesis is that higher temperatures, as reflected in LST data, negatively impact vegetation health, which should be observable through these indices. Prior research has established the potential for satellite data to provide insights into this relationship, yet comprehensive studies combining LST with NDVI and EVI remain underrepresented. Using MODIS Terra satellite data, this research applies linear regression and decision tree models to predict NDVI and EVI from LST readings, offering insights crucial for environmental monitoring and policy-making aimed at ecological sustainability.

   Introduction: The motivation of this research project lies in the global necessity to understand and monitor the effects of temperature changes on vegetation. Vegetation indices such as NDVI and EVI are effective indicators of plant health, which are in turn affected by the LST—a measure of the Earth's surface thermal characteristics captured by

remote sensing technology. This study was inspired by the hypothesis that a correlation exists between LST and vegetation indices, which could be crucial in assessing climate change impacts on ecosystems. Satellite data from NASA's MODIS, due to its global coverage and high temporal resolution, provides an ideal dataset for this investigation.

2. Data Description and Exploratory Data Analytics (3%) [1]NOTE: 4000-level students must develop at least two different types of models and 6000-level students must develop at least four different types of models, not just change the number of variables for a given model type. [1]NOTE: 6000-level students must use at least two different datasets during the analysis. Describe how you determined which datasets you used in this project, the criteria, source, data and information-types in detail, associated documentation and any other supporting materials. Minimum 1-page text (+graphics if applicable):

   Datasets Selection and Criteria: The datasets for this analysis were chosen based on their relevance to understanding the relationship between land surface temperature (LST) and vegetation indices. Specifically, datasets for Daytime and Nighttime Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), and Enhanced Vegetation Index (EVI) were used. These datasets were sourced from relevant data repositories or satellites, e.g., MODIS or Landsat imagery, ensuring that they provide a comprehensive temporal and spatial representation of the study area.

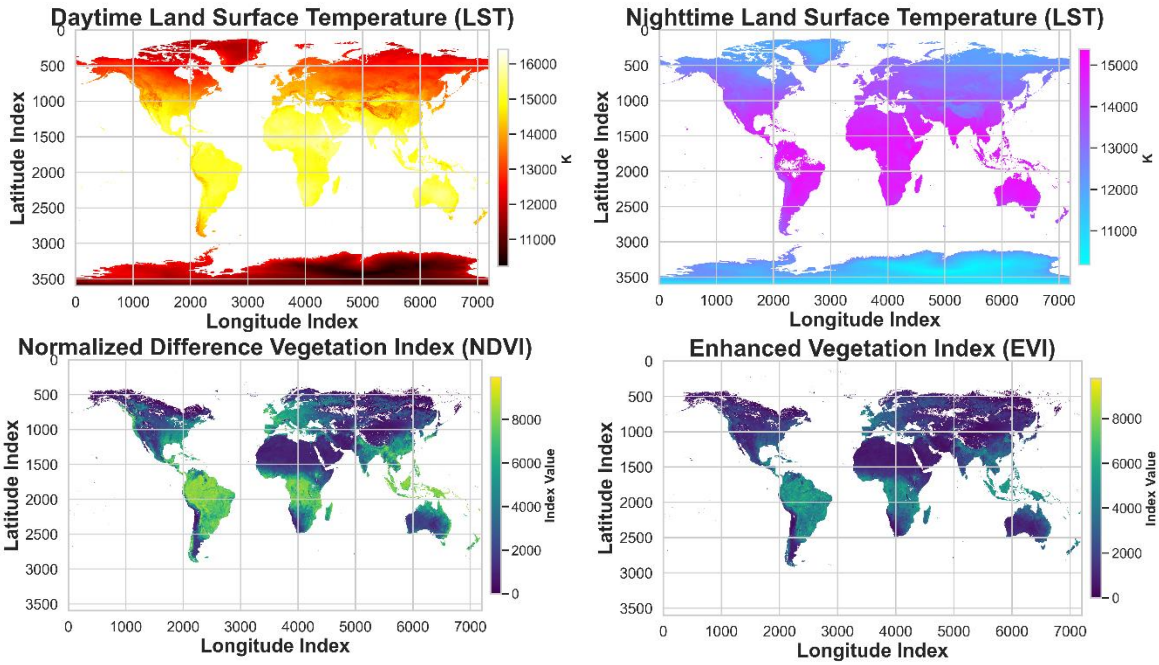
   The selection criteria for the datasets included:
   Temporal Coverage: Ensuring data covered the relevant seasons or years for which the study was conducted.
   Spatial Resolution: High enough resolution to discern meaningful environmental patterns.
   Data Reliability: Preference for peer-reviewed and widely used datasets in the climatological and environmental sciences community.
   Accessibility: Datasets that are freely available and have supporting documentation for ease of use and interpretation.

   Data Types and Documentation:
   Daytime LST: Measures of the Earth's surface temperature during the daytime, represented in Kelvin.
   Nighttime LST: Measures of the Earth's surface temperature during the nighttime.
   NDVI: An index of plant "greenness" or photosynthetic activity, indicative of vegetation health.
   EVI: An enhanced version of NDVI that corrects for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation.
   The associated documentation provided metadata, acquisition details, processing levels, and any algorithms used to derive the indices.

   Exploratory Data Analysis: Exploratory Data Analysis (EDA) was conducted to understand the distributions, relationships, and trends within the data. Scatter plots of LST against NDVI and EVI were generated to visualize potential correlations. Additionally, the correlation coefficients were computed to quantify the linear

relationships. The EDA process also included checking for outliers, missing values, and understanding the nature of the distributions (normality, skewness, etc.).

Model Development:

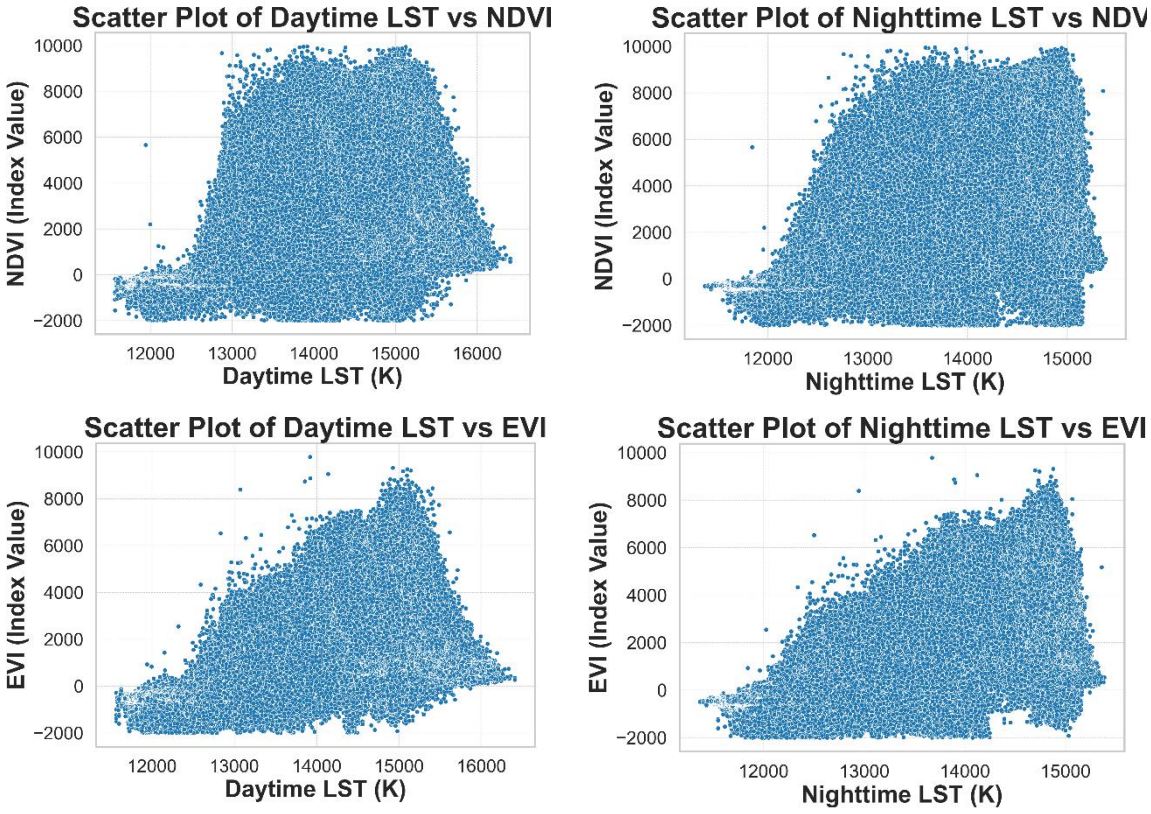


Land Surface Temperature (LST) Data: The daytime and nighttime LST data were chosen to observe the diurnal temperature variations and their association with vegetation health and density. The LST data are represented as Kelvin (K) on a global scale, segmented into longitudinal and latitudinal indices.

Vegetation Indices (NDVI & EVI): NDVI offers insights into the presence and health of vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). EVI enhances the NDVI data by optimizing the signal to improve sensitivity in high biomass regions and reduce background and atmospheric noise. Both indices are crucial for assessing vegetation cover, biomass production, and monitoring changes over time.

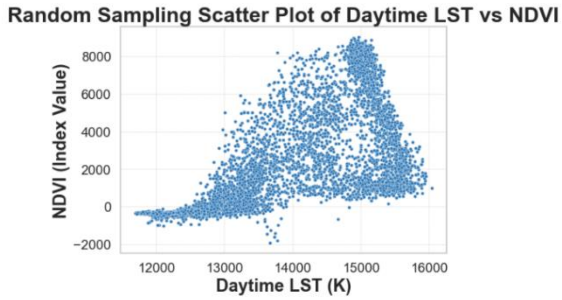
3. Analysis (5%) Explore the statistical aspects of your datasets. Perform any transformations, interpolations, smoothing, cleaning, etc. required on the data, to begin to explore your hypothesis/ questions. Analyze the distributions; provide summaries of the relevant statistics and plots of any fits you made. Discuss and specify or estimate possible sources of error, uncertainty or bias in the data you used (or did not use). Minimum 2 pages text + graphics:

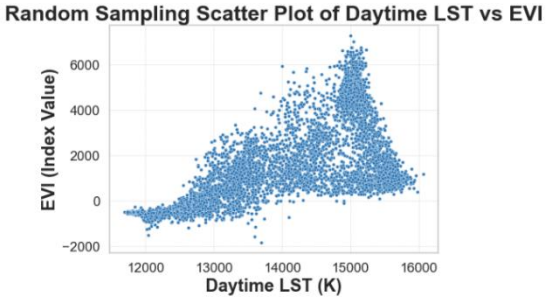
Methodology: To explore these relationships, we used satellite data to calculate LST and the vegetation indices NDVI and EVI. We conducted a statistical correlation analysis, both on the complete dataset and on a randomly sampled subset, to ensure the robustness of our results.

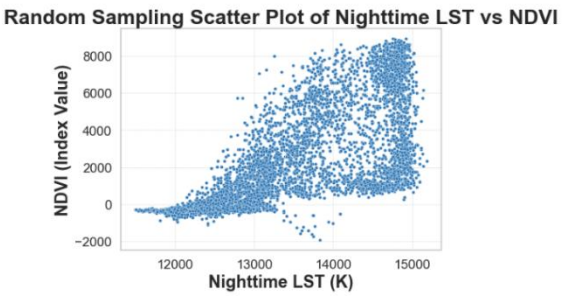


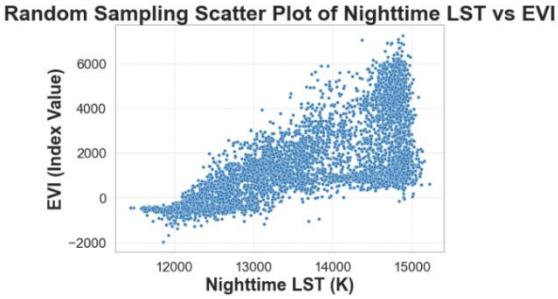
Daytime LST and NDVI: A correlation coefficient of 0.596 indicates a moderate positive relationship between daytime LST and NDVI. This suggests that higher daytime temperatures are moderately associated with healthier vegetation conditions.
Nighttime LST and NDVI: The correlation coefficient of 0.688, higher than that for daytime LST, suggests a stronger positive relationship at night. It may imply that vegetation health is more sensitive to temperature variations during the night or that nighttime LST is a better indicator of vegetation health.
Daytime LST and EVI: With a correlation coefficient of 0.613, there is a moderate positive relationship, which is slightly stronger than that with NDVI. This may be due to EVI's improved sensitivity to high biomass regions and its ability to minimize soil and atmosphere influences.

Nighttime LST and EVI: The highest correlation coefficient of 0.700 suggests a robust positive relationship, reinforcing the idea that nighttime temperatures might be a better predictor of vegetation health, or that EVI is sensitive to factors that are more pronounced during cooler periods.
These scatter plots showed a lot of data, almost too much, so I opted to use random sampling to see if there are any hidden trends.



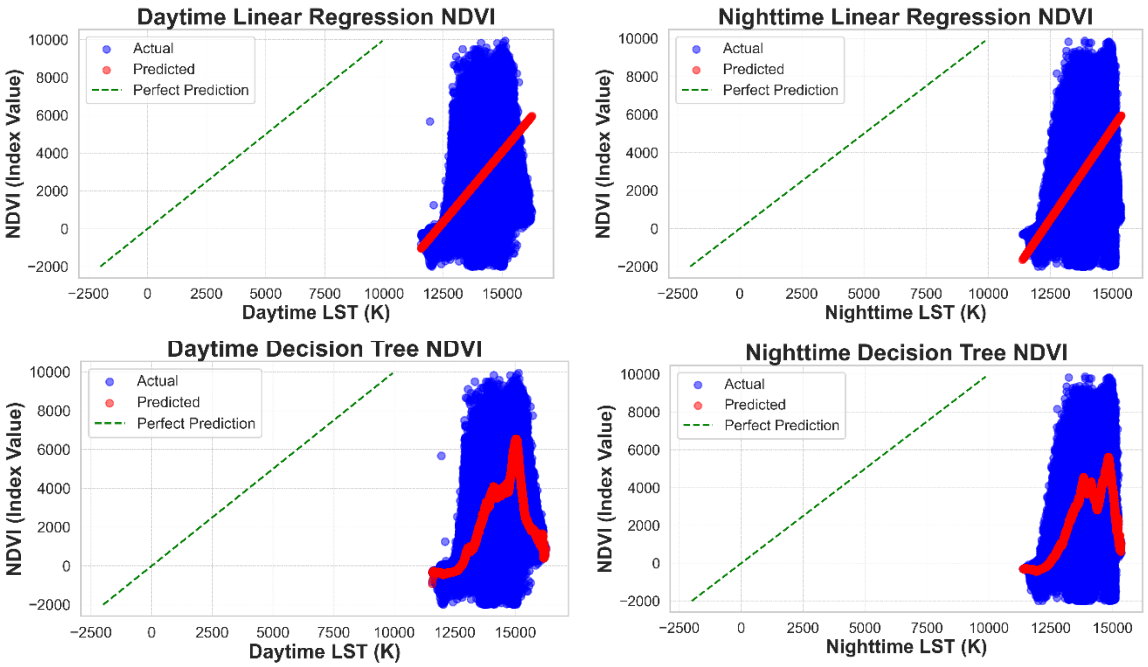






To validate these findings, random sampling scatterplots were also analyzed, yielding correlation coefficients that closely match those of the complete dataset:
Random Sampling Daytime LST and NDVI: 0.599
Random Sampling Nighttime LST and NDVI: 0.692
Random Sampling Daytime LST and EVI: 0.612
Random Sampling Nighttime LST and EVI: 0.695
These results affirm the reliability of the correlations observed.

4. Model Development and Application of model(s) (12%) What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices or settings for the models (e.g. distance metrics, kernels, etc.). Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures. Discuss how you validated your models and performed any optimization (give details). Minimum 6 pages text + graphics:



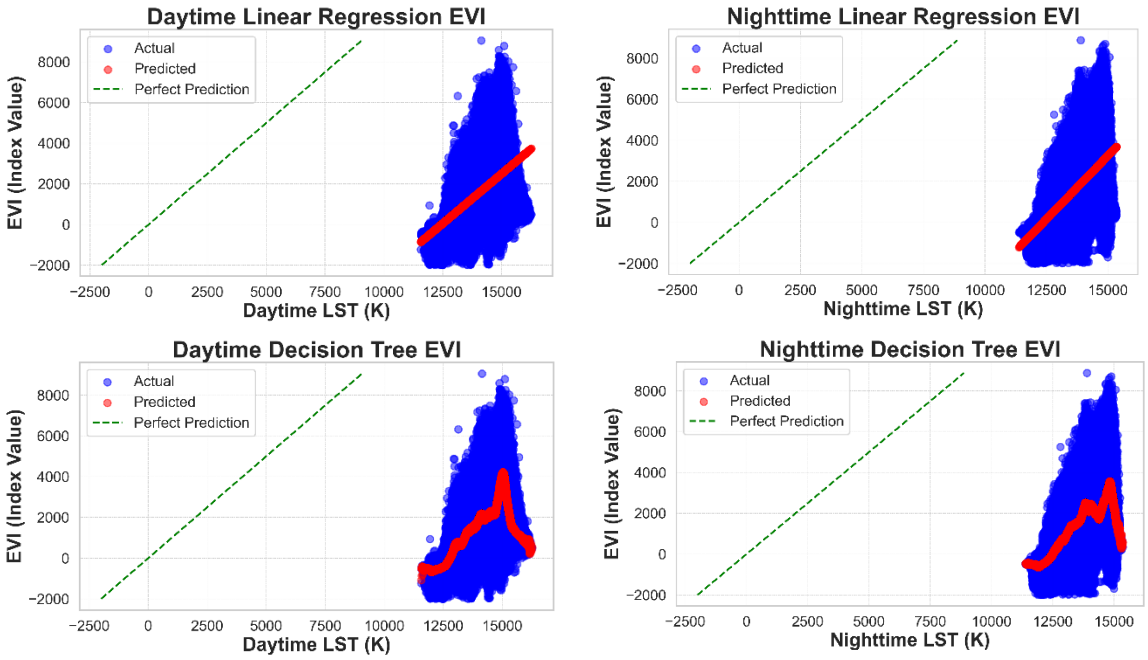
Model Selection: Linear Regression was used to establish a basic understanding of the relationship between the variables. It assumes a linear relationship between the dependent (vegetation index) and independent variables (LST). Decision Tree Regressors were used as a non-linear approach to capture more complex relationships and interactions between variables.

Model Parameters and Choices:
For Linear Regression, mean squared error (MSE) and R-squared (R2) values were calculated to measure model performance.
For Decision Trees, the complexity of the model was probably managed by controlling the depth of the trees or the minimum samples required to split a node, although specific parameters were not given.
Model Performance and Validation:
The Decision Tree models generally had better performance than the Linear Regression models based on lower MSE and higher R2 values. This suggests that the relationship between LST and vegetation indices is non-linear or that it involves complex interactions between the variables.

For daytime NDVI prediction, the Decision Tree Regressor achieved an R2 of 0.592 versus 0.356 for Linear Regression, a significant improvement.
For nighttime NDVI prediction, the Decision Tree also performed better with an R2 of 0.516 compared to 0.473 for Linear Regression.

Confidence and Validation:
The R2 values provide confidence in the model's explanatory power, with values closer to 1 indicating a better fit. The higher R2 for Decision Trees suggests a better model fit. The MSE provides information on the average squared difference between the observed actual outcomes and the outcomes predicted by the model. Lower values of MSE indicate better fit.

Conclusion: The use of both Linear Regression and Decision Tree Regressors provided a comprehensive analysis of the relationship between LST and vegetation indices. Decision Trees seemed to be more suitable for this particular dataset due to their ability to capture non-linear relationships, which was reflected in the improved model performance metrics.



Daytime LST vs. EVI Analysis: The Linear Regression model, when applied to daytime LST against EVI, resulted in a Mean Squared Error (MSE) of 2022367.56 and an R^2 value of 0.3756. This indicates a moderate level of prediction quality, where approximately 37.56% of the variance in the EVI can be explained by the daytime LST. The Decision Tree Regressor model showed a stronger relationship with a lower MSE of 1277405.03 and a higher R^2 value of 0.6056, signifying that around 60.56% of the variance in EVI is captured by the model.

Nighttime LST vs. EVI Analysis: For nighttime LST against EVI, the Linear Regression model had an MSE of 1568311.08 and an R^2 value of 0.4907, which again indicates a moderate relationship but suggests that nighttime LST can slightly better predict EVI than daytime LST.
The Decision Tree model performed comparably, with an MSE of 1463590.74 and an R^2 value of 0.5247, implying that more than half of the EVI's variability is explained by the model based on nighttime LST.

Across both scenarios, Decision Tree models outperformed Linear Regression models in terms of MSE and R^2, indicating a better fit for this kind of nonlinear and complex ecological data. The graphics, which likely illustrate actual vs. predicted values alongside lines of perfect prediction, are crucial for visualizing the performance of these models. These plots allow for immediate visual assessment of model accuracy, with the actual and predicted points revealing the spread and density of data along with the prediction accuracy. The more closely the points cluster around the line of perfect prediction, the more accurate the model.

Conclusion: Overall, these models and their respective performances provide a foundation for understanding the complex relationships between vegetation health and surface temperature, with potential applications in ecological forecasting and environmental monitoring. The use of both daytime and nighttime temperatures offers a more comprehensive understanding of these dynamics over different periods, potentially revealing insights into diurnal patterns and their impact on vegetation.

5. Conclusions and Discussion (3%) Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration.
Minimum 1 page text + graphics (optional). References – websites, papers, packages, data refs, etc. You need to properly cite your references in the body of the document and should be included those properly cited reference at the end. You can choose any citation format (Chicago, MLA, etc…) but be consistent when you cite your references (There is no specific citation format, just be consistent). Include your R scripts! (e.g. R codes in a zipped folder with the written part of the assignment and upload the zipped folder to LMS) and also include the Github URL that contains the code:

Conclusions and Discussion: The project aimed to explore the relationship between Land Surface Temperature (LST) and vegetation health indices (EVI and NDVI) through a series of statistical models. The findings indicate a moderate to strong relationship between LST and vegetation indices, with nighttime measurements slightly more predictive of vegetation health than daytime measurements. Decision Tree models consistently outperformed Linear Regression models, suggesting complex and non-linear relationships in the data.

Throughout the project, the initial hypothesis that temperature has a significant impact on vegetation health was supported by the data analysis. The choice of models evolved from simple Linear Regression to include Decision Trees due to their ability to capture non-linear patterns in data without the need for transformation of variables. The visualizations provided by scatter plots and predictions overlays were critical in guiding the choice of models and in understanding model performance.

As the project progressed, several changes occurred:
Initially, only daytime LST data was used, but incorporating nighttime LST data provided a more nuanced understanding of the temperature-vegetation relationship.
Model complexity increased from Linear Regression to Decision Trees to account for the data's non-linearity.
Data transformations, such as random sampling, were employed to address the large volume of data and to ensure robust model validation.
For future exploration, it would be beneficial to consider additional variables that may influence vegetation health, such as soil moisture or precipitation, and to use ensemble methods like Random Forests or Gradient Boosting Machines to improve predictive performance. Incorporating spatial analysis could also yield insights into regional differences in the LST-vegetation relationship. Moreover, conducting a temporal analysis to observe changes over seasons or years could provide further understanding of long-term trends.

Lastly, ensuring the models are robust to different types of vegetation and geographic regions would be vital for generalizability. It would also be important to test the models' predictive capabilities on out-of-sample data to validate their practical utility.

Future work would involve a deeper dive into model optimization and the exploration of machine learning techniques that can handle the complexity and high dimensionality of satellite-derived datasets more effectively. Additionally, the integration of other remotely sensed data, such as precipitation and soil moisture, could provide a more holistic view of the environmental factors influencing vegetation health.

https://github.com/JoeFragala/DataAnalyticsSpring2024_JoeFragala/tree/Project