**Invasive Species K-Means Cluster Model**

Ashley Deibler

Data Science | Bellevue University

DSC 630 – T303 Predictive Analytics

Professor Frank Neugebauer

November 16, 2024

**Introduction**

In wildlife ecology and management, invasive species have become one of the largest threats to biodiversity and ecosystem health throughout recent decades. Invasive species are defined as any living species that are “non-native (or alien) to the ecosystem under consideration and whose introduction causes or is likely to cause economic or environmental harm or harm to human health” (*What are invasive species?*). Spread of invasive species happens primarily due to human activity such as global transportation of goods (*What are invasive species?*). Threats to biodiversity and decline in population of native species occur through predation, herbivory, disease, resource competition, and habitat alteration (*Environmental and Ecological Impacts*). It is estimated that around forty-two percent of species listed as threatened or endangered on the International Union for Conservation of Nature and Natural Resources (IUCN) Red List are at risk as a result of biological invasions by non-native species (*Invasive Species)*. These increasing impacts on natural ecosystems result in trickle-down effects on human health and economies as we rely on healthy ecosystems for many of our commercial, agricultural, and recreational activities (*Invasive Species*). Studies have shown that between the 1960s and 2010s, invasive species studies, management, and control efforts have increased from $2 billion to over $26 billion per year, with a global economic cost of over $1.288 trillion over the past fifty years (*Economic and Social Impacts*).

In the United States, the issues of biological invasion on native ecosystems should be of concern to several groups including biologists, wildlife conservationists, environmental consultants, economists, elected officials, and healthcare professionals. Not only does this problem impact natural processes and ecosystem health, but it also has proven to have a significant impact on the economy and health of the human population. This is why it is important to conduct research and make an effort to find solutions to address increasing invasive species spread around the globe. To begin this process, machine learning and model production can be used to visually highlight geographic regions of highest concern where populations of invasive species are more densely located. This information can then be used further to research more intensely on unique areas, what species are invading, what species are affected as a result, and what other environmental variables to consider when formulating a management plan.

To create a model displaying this concept, a general dataset consisting of survey data for a wide variety of invasive species (flora and fauna) in the United States was obtained from Kaggle.com. This dataset was then preprocessed and formulated to fit visualizations that resulted in the species chosen to build this model off of. Bar charts were used to determine the scientific name of the species with the most available observations – *Myocastor coypus*, more commonly known as the “nutria” or the “coypu.” The nutria is a large, rat-like rodent commonly found near permanent water sources. While they primarily stay near freshwater sources, they have been found near coastal waterways. They are semiaquatic much like capybara and beavers, and their webbed feet allow them to be very strong swimmers. The nutria is native to South America, but was brought to the United States by fur farmers who raised them for their pelts. When the value of nutria pelts fell, farmers released their captive nutria into the wild, thus marking the beginning of their existence in United States ecosystems. The rodent has since come to create significant damage to native waterways and wetlands around the country by eating native vegetation that holds the soil together; in addition, they burrow, which ends up weakening flood control levees and therefore altering the topography of native waterways. Nutria are direct threats to species such as the native muskrat, who compete aggressively for similar resources, as well as freshwater mussels, which are a common food source to the nutria. Their presence also threatens other threatened and endangered species who rely on wetland habitats. Additionally, nutria can host tuberculosis or septicemia, which are threats to livestock, pets, and humans. Now, they can be found in at least fifteen different U.S. states.

After narrowing down the focus to *M. coypus*, a secondary data source was obtained from the United States Geological Survey, consisting of survey observation details including longitude and latitude, HUC 8 number, year of observation, as well as state, county, and drainage name of the observation (*Nutria – Collections*). This dataset is useful for this particular project because it provides geographic information revolving around observations of individuals of interest. The latitude and longitude information specifically is most useful in creating a k-means cluster model to illustrate “hotspots” for invasive species presence across the United States. Upon official deployment of the model, any data source offering latitude and longitude data can be used to obtain similar results for other species of interest.

**Methods**

Preparing the dataset included reducing the data to consist of only the variables necessary for the model, which included scientific name, specimen number, latitude, and longitude. While the rest of the variables available in the dataset provide valuable information regarding specific geographic position within a cluster, they are not useful for creating the k-means cluster model. The selected variables were then cleaned to address any missing values, typos, or outliers.

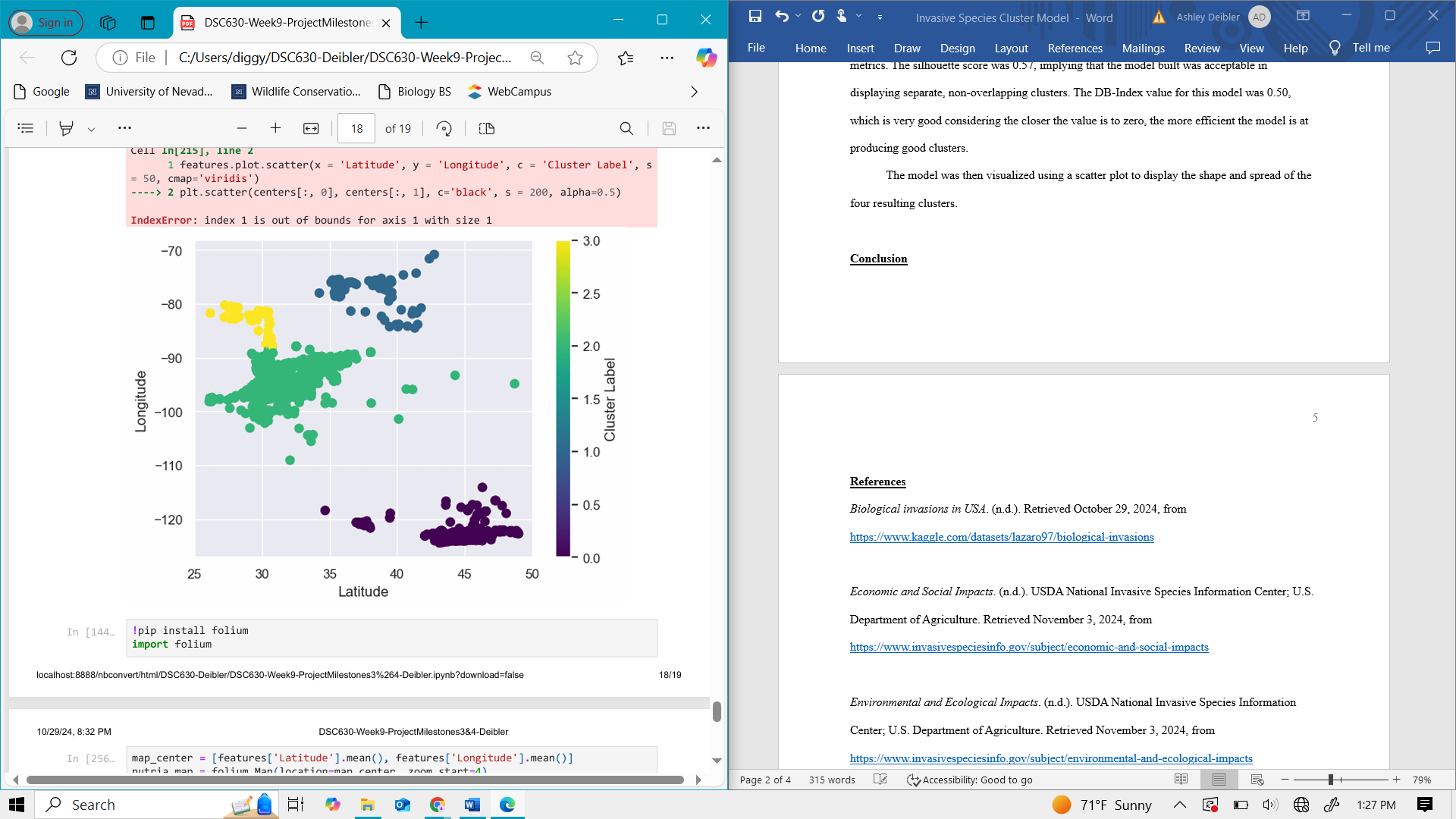
The k-means cluster model was created using Python. The x- and y-variables were assigned to latitude and longitude, respectively. Prior to building the model, an Elbow Curve was used to determine that the optimal number of clusters for this particular dataset was four. To evaluate the model, silhouette score will be used to assess the model’s ability to create clusters that do not overlap, while the Davies-Bouldin Index (DB-Index) will be used to determine the ability to produce good, accurate clusters.

Once the model is completed, the results will be visualized using the *folium* package, to overlay scatter plot results of the cluster data onto a geographic map of the United States.

**Results**

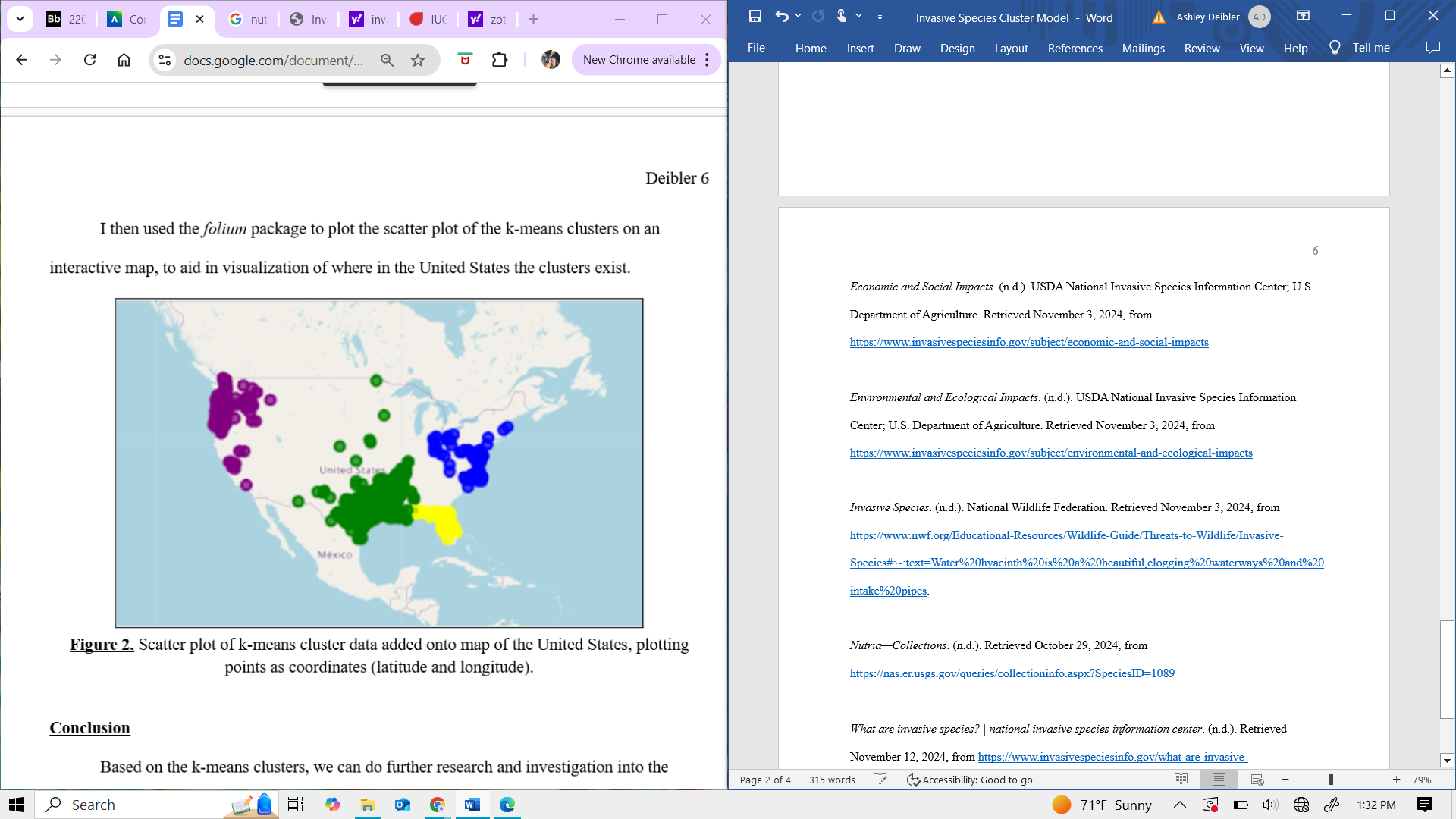
The k-means cluster model was created using four clusters and evaluated using two metrics. The silhouette score was 0.57, implying that the model built was acceptable in displaying separate, non-overlapping clusters. The DB-Index value for this model was 0.50, which is very good considering the closer the value is to zero, the more efficient the model is at producing good clusters.

The model was then visualized using a scatter plot to display the shape and spread of the four resulting clusters.



**Figure 1.** Scatter plot displaying cluster results of k-means cluster model based on latitude and longitude data.

The scatter plot displaying the k-means cluster results were then overlayed onto a United States map to aid in determination of where these clusters exist throughout the country.



**Figure 2.** Scatter plot of k-means cluster data added onto map of the United States, plotting points as coordinates of latitude and longitude.

**Conclusion**

Creating this model is effective in displaying the ways that data science can be used in wildlife conservation and management. Using this model, we are able to group observations of invasive individuals and establish where geographically they are most abundant to implement management strategies effective to that particular region. In this specific case, using the k-means clustering model allows us to observe that there are four regions around the United States where invasion of *Myocastor coypus* is abundant and requiring the implementation of some management strategy. However, one can tell that these clusters are located across the country and therefore are present in areas that may have very different conditions and features from one another. For example, the requirements and conditions for management in the Pacific Northwest (purple cluster in Figure 2) may be very different from those in the Southeastern portion of the country (yellow cluster in Figure 2). For this reason, it’s important to acknowledge that this model does not provide specific management solutions, as management strategies may vary between species, region, and ecosystem. However, this model is efficient in displaying clusters that can allow wildlife biologists to better identify where they should focus their attention.

There is definitely the ability to use machine learning to create models that could easily supplement this model, or act as an ensemble to provide management strategy recommendations based on specific geographic data. Creating such an addition would require additional model building and training, but may be a beneficial next step in furthering this project.

In terms of ethical considerations for this project, it is important to note that in wildlife biology there are a myriad of ways in which surveyors collect observation data, including radio telemetry, visual observation, or camera traps. This variation in ways observations are made may lead to the potential of skewed data. In addition, there needs to be consideration for the natural behavior of wildlife and their ability to move between ecosystems, hibernate, or migrate depending on time of year. This information is important to keep in mind as it may create duplicate observations of an individual which could also skew data and impact accuracy of the results. Finally, upon implementing management strategies biologists need to juggle the balance between important ecosystem function and human health and wellbeing. There may be instances where strategies used to eradicate an invasive population may cause more harm than good due to how the ecosystem may have adapted over time to the presence of that species, and whether those adaptations are used in human activities such as recreation or agriculture.

**References**

*Biological invasions in USA*. (n.d.). Retrieved October 29, 2024, from <https://www.kaggle.com/datasets/lazaro97/biological-invasions>

*Economic and Social Impacts*. (n.d.). USDA National Invasive Species Information Center; U.S. Department of Agriculture. Retrieved November 3, 2024, from <https://www.invasivespeciesinfo.gov/subject/economic-and-social-impacts>

*Environmental and Ecological Impacts*. (n.d.). USDA National Invasive Species Information Center; U.S. Department of Agriculture. Retrieved November 3, 2024, from <https://www.invasivespeciesinfo.gov/subject/environmental-and-ecological-impacts>

*Invasive Species*. (n.d.). National Wildlife Federation. Retrieved November 3, 2024, from <https://www.nwf.org/Educational-Resources/Wildlife-Guide/Threats-to-Wildlife/Invasive-Species#:~:text=Water%20hyacinth%20is%20a%20beautiful,clogging%20waterways%20and%20intake%20pipes>.

*Nutria—Collections*. (n.d.). Retrieved October 29, 2024, from <https://nas.er.usgs.gov/queries/collectioninfo.aspx?SpeciesID=1089>

*What are invasive species? | national invasive species information center*. (n.d.). Retrieved November 12, 2024, from <https://www.invasivespeciesinfo.gov/what-are-invasive-species#:~:text=Definitions%29%20an%20%22invasive%20species%22%20is%20a%20species%20that,in%20areas%20where%20they%20do%20not%20naturally%20exist>.