INTRODUCTION

In recent years, there has been a growing interest in understanding the factors that influence student choices in transferring between post-secondary institutions. One such factor is geographical distance, which has been highlighted by the American Council on Education’s 2016 study. This study revealed that 57% of incoming freshmen attending public four-year colleges enrolled within 100 miles of their permanent residence, suggesting that proximity plays a significant role in students’ college decisions. Furthermore, the same study proposed that some students may strategically use these nearby schools as “transit points” in their post-secondary education plans. Considering these findings, this research aims to explore the relationship between geographical distance and post-secondary transfer rates.

DATA INTRODUCTION: The 2020-2021 CollegeScorecard includes data for 6,644 post-secondary institutions for the following parameters:

- **INSTMA:** Name of the School.
- **LATITUDE:** The North-South position of the School.
- **LONGITUDE:** The East-West position of the School.
- **OMETRIP A.U.:** The proportion of students that withdraw from the institution originally stated at and enrolled in another institution within 8 years.
- **UGDS:** The number of undergraduate/degree-seeking students that start in the fall of 2010.
- **CONTROL:** Indicates whether a post-secondary institution is Publicly or Privately owned.
- **LEVEL:** Indicates whether a post-secondary institution is a four-year school or a two-year school.
- **CURRPER:** Indicates if a school is currently operating or not.
- **DISTANCEONLY:** Indicates if a school only offers online classes or not.

DATA WRANGLING: Public four-year post-secondary institutions that are currently operating and do not operate online were used in the analysis of this project. Schools missing values for either their latitude or longitude, or transfer rates prior to analysis were filtered, leaving 2,349 schools remained.

METHODS

- **DISTANCE BETWEEN SCHOOLS:** The latitude and longitude of each school were used to find the distance between schools using the great-circle distance function which computes the shortest distance between two schools on the surface of a sphere.
- **NUMBER OF SCHOOLS WITHIN 100 MILES:** After finding the distance between schools, we identified the schools within 100 miles of each other. This group represents a neighborhood of schools, giving students various options when considering a transfer.
- **CLUSTERING COEFFICIENT:** The clustering coefficient of a school is a measure that aids in understanding the degree to which schools are interconnected. This measure represents the number of neighbors a school has in common with its neighbors, indicating the strength of connections within the network of schools. The clustering coefficient is then calculated as the ratio of the actual connections to the maximum possible connections.
- **NETWORK TRANSITIVITY:** The network transitivity is a global view of the clustering coefficient and measures proportion of shared neighbors among the nearest neighbors within the neighborhood of schools and their tendency to form clusters or tightly-knit groups. An increasing transitivity coefficient can be a sign of interconnectivity and a better-quality transfer rate at that institution.
- **GEOPANDAS:** GeoPandas is a powerful and user-friendly library for working with spatial data, especially geospatial, by extending Pandas functionality with geospatial capabilities. This library provides spatial operations and features that the user-defined functions were used to create the network map displayed in Figure 2.
- **NETWORKX:** NetworkX is an open-source Python library designed for the creation, manipulation, and visualization of complex networks or graphs. It provides a wide range of tools and algorithms that can be applied to various types of networks, such as social networks, biological networks, transportation networks, and many others.
- **GGRAPH2:** Ggraph is a powerful and popular data visualization library for the R programming language. It is based on the Grammar of Graphics and is a tool which aid in creating consistent and expressive visualizations.

METHODS

- **DISTANCE BETWEEN SCHOOLS:** The latitude and longitude of each school were used to find the distance between schools using the great-circle distance function which computes the shortest distance between two schools on the surface of a sphere.
- **NUMBER OF SCHOOLS WITHIN 100 MILES:** After finding the distance between schools, we identified the schools within 100 miles of each other. This group represents a neighborhood of schools, giving students various options when considering a transfer.
- **CLUSTERING COEFFICIENT:** The clustering coefficient of a school is a measure that aids in understanding the degree to which schools are interconnected. This measure represents the number of neighbors a school has in common with its neighbors, indicating the strength of connections within the network of schools. The clustering coefficient is then calculated as the ratio of the actual connections to the maximum possible connections.
- **NETWORK TRANSITIVITY:** The network transitivity is a global view of the clustering coefficient and measures proportion of shared neighbors among the nearest neighbors within the neighborhood of schools and their tendency to form clusters or tightly-knit groups. An increasing transitivity coefficient can be a sign of interconnectivity and a better-quality transfer rate at that institution.
- **GEOPANDAS:** GeoPandas is a powerful and user-friendly library for working with spatial data, especially geospatial, by extending Pandas functionality with geospatial capabilities. This library provides spatial operations and features that the user-defined functions were used to create the network map displayed in Figure 2.
- **NETWORKX:** NetworkX is an open-source Python library designed for the creation, manipulation, and visualization of complex networks or graphs. It provides a wide range of tools and algorithms that can be applied to various types of networks, such as social networks, biological networks, transportation networks, and many others.
- **GGRAPH2:** Ggraph is a powerful and popular data visualization library for the R programming language. It is based on the Grammar of Graphics and is a tool which aid in creating consistent and expressive visualizations.

RESULTS

- **University Transfer Rates between Schools within 100 Miles (Figure 1):** Figure 1 shows the location of each school as a gold dot (or vertex) on a map. The size of each dot represents the number of undergraduate students at the school. When two schools are within 100 miles of each other, a red line (or edge) connects them. The width of this line reflects the school’s transfer rate, with wider lines indicating higher transfer rates. From this visualization, we can observe that areas with fewer schools seem to have fewer and thinner connecting lines, which might suggest lower transfer rates.
- **Histogram of Post-Secondary Transfer Rates (Figure 3):** Figure 3 displays the distribution of post-secondary transfer rates across different regions of the U.S. This plot shows the average transfer rate of each region does not significantly deviate from the overall average transfer rate observed in Figure 2 (visualized by the dotted red line). The region with the lowest average transfer rate is the Southeast, while the highest average transfer rate was observed by the Great Lakes region. Notably, although the average transfer rate of the Southeast region is less than the average transfer rate of the Great Lakes, the variability of transfer rates observed for the Southeast is much larger than the variability observed for the Great Lakes.
- **Scatterplot of the Number of Schools within 100 miles by Transfer Rate (Figure 4):** Figure 4 displays the number of schools within 100 miles of a school on the vertical axis and the transfer rate horizontally. For each panel, the overall average post-secondary transfer rate is plotted as a dotted red line while the region-specific average is plotted as dotted blue line. Although some regions (Great Lakes) display a slight positive relationship between the number of schools within 100 miles and post-secondary transfer rate, this plot suggests the relationship may be weak or nonexistent.
- **Boxplot of the Number of Schools within 100 miles by Region (Figure 5):** Figure 5 displays the distribution of the number of schools within 100 miles across different regions. This plot shows for each region the number of schools within 100 miles, as well as the range of values observed for schools in that region. The Rocky Mountain region was observed with the lowest number of schools within 100 miles. This suggests the Rocky Mountain region is very sparse in post-secondary institutions (which can be observed in Figure 1). In contrast, the Mid East Region was observed with the highest average suggesting the post-secondary density of this region is very high.

CONCLUSIONS

- **University Transfer Rates between Schools within 100 miles (Figure 1):** Figure 1 shows the location of each school as a gold dot (or vertex) on a map. The size of each dot represents the number of undergraduate students at the school. When two schools are within 100 miles of each other, a red line (or edge) connects them. The width of this line reflects the school’s transfer rate, with wider lines indicating higher transfer rates. From this visualization, we can observe that areas with fewer schools seem to have fewer and thinner connecting lines, which might suggest lower transfer rates.
- **Histogram of Post-Secondary Transfer Rates (Figure 3):** Figure 2 presents the distribution of transfer rates among post-secondary institutions. The average transfer rate, marked by a dotted red line, is 25%. We can see that Southeast region has the highest proportion of transfer students, while other regions have a wide range of transfer rates among schools, and certain institutions may have unique factors or circumstances contributing to such high transfer rates. Considering the wide range of transfer rates observed in Figure 2, further analysis was conducted to explore the variation in transfer rates among different schools and regions.
- **Scatterplot of the Number of Schools within 100 miles by Transfer Rate (Figure 4):** Figure 4 displays the number of schools within 100 miles of a school on the vertical axis and the transfer rate horizontally. For each panel, the overall average post-secondary transfer rate is plotted as a dotted red line while the region-specific average is plotted as dotted blue line. Although some regions (Great Lakes) display a slight positive relationship between the number of schools within 100 miles and post-secondary transfer rate, this plot suggests the relationship may be weak or nonexistent.
- **Boxplot of the Number of Schools within 100 miles by Region (Figure 5):** Figure 5 displays the distribution of the number of schools within 100 miles across different regions. This plot shows for each region the number of schools within 100 miles, as well as the range of values observed for schools in that region. The Rocky Mountain region was observed with the lowest number of schools within 100 miles. This suggests the Rocky Mountain region is very sparse in post-secondary institutions (which can be observed in Figure 1). In contrast, the Mid East Region was observed with the highest average suggesting the post-secondary density of this region is very high.