Data exploration included looking at the similarity between multiple industries over a 5-year and 20-year periods of time to help find relative correlations between our variables (features) and identity relationships that may not be as apparent when viewing the data tables on their own. The first pair of correlation matrices using pearson correlation, a 5-year and 20-year, were constructed using Heatmaps including their correlation significance p-values using specific libraries including ggcorrplot, gplots, and tidyr. p-value provides whether the correlation coefficient is significantly different from 0. R programing language in the R Studio environment with approximately 80 lines of code was used to construct our correlation heatmaps and heatmaps with dendrograms. Windows() and pdf() operations were utilized to render all the correlation matrices in a separate window outside R and then to generate/save a pdf version of each visualization. The higher correlation centers around trade transportation and utilities (Trade\_Trans\_Util), Manufacturing, leisure, and hospitality (Leisure\_Hosp) and education and health (Ed\_Health) industries. It is noted that all industries appear more highly correlated linear relationship compared to the last five years, though the highest correlations in the 20-year data set remain similar to those identified in the 5-year data.

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

**5-year Industry Heatmap 20-year Industry Heatmap**

**Pearson Correlation Pearson Correlation**

The second pair were constructed using Heatmaps with dendrograms using Euclidean distances and Clustering using Ward.D2 linkage, with calculated z-scores for columns in the key. These required using hearmap.2, dist and hclust functions, where hclust orders based on correlation coefficients including setting the Ward linkage. These clearly pulled out the groupings of industries within our data covering both a 5-year timeframe and a 20-year timeframe. Clusters are well identifiable using this approach as can be seen in both especially trade transportation and utilities (Trade\_Trans\_Util), Manufacturing, leisure, and hospitality (Leisure\_Hosp) and education and health (Ed\_Health) industries. It is important to note that the distances and clusters are clearly identified relative to trade transportation and utilities (Trade\_Trans\_Util), Manufacturing, leisure, and hospitality (Leisure\_Hosp) and education and health (Ed\_Health) industries as were in the simpler correlation matrices. Another observation is how the dendrogram show the relationships and how they differ when looking at the last 5 years relative to 20 years of data.

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

Chart

Description automatically generated

Text

Description automatically generated with low confidence

The last pair of correlation matrices are to help better understand and identified relationships between features (industries) using the 5-year and 20-year data, we constructed correlation plots with Loess regression and smooth/curve fitted representations of the data sets for 5 years and 20 years. Loess regression is a local weighted regression/local polynomial regression to fit a smooth curve within the data plot and can be useful to reveal trends and cycles in data that might be difficult to model with a parametric curve. The main advantage using LOESS has over many other methods is that it does not require the specification of a function to fit a model to all of the data in the sample.[NIST1] These clearly show more linear relationships in the 20-year data set relative to those in the 5-year data set, which aligns with what we were seeing in the heatmaps. This clearly helps see the impact looking over a longer period. Both show us the relationships and amounts of data points as they present for trade transportation and utilities (Trade\_Trans\_Util), Manufacturing, leisure, and hospitality (Leisure\_Hosp) and education and health (Ed\_Health) industries, and then as well all other industries in our data set variables (industries). Diving into each plot shows relationships that are not as apparent with a single linear regression line, as seen in the relationships between the red line and the blue curves. This combined correlations with plots, one for 5-year and one for 20-year, were constructed using specific libraries including GGally and ggplot with ggpair functions. R programing language in the R Studio environment was again used needing approximately 35 lines of code.

[NIST1] Engineering Statistics Handbook, 4.1.4.4 Loess (aka Lowess), retrieved at <https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm>, last accessed November 29, 2021.



**5-year Correlation with Plots using Loess Regression with Smoother**



**20-year Correlation with Plots using Loess Regression with Smoother**

The coding was straight forward for each approach to provide each pair of correlation matrices, this included code testing some re-coding and making sure the operations were working as expected. Ingestion of data went well, the only concern was making sure the data being analyzed was the correct data file, 5 years, or 20 years. We also needed to assure that the feature naming scheme was consistent between files. Another issue was to assure that the column or row feature names fit within the visualization so as to make sure it was readable; similarly making sure the title and labels were meaningful as were the colors selected for each visualization.