

Pennsylvania Economic Index

*A review of the economic metrics from the 2010 US Census for each county in Pennsylvania. Purpose of improving economic distribution across the state.*



**Kyle Gill**

Messer Americas

February 2021

1. Introduction

Background

Pennsylvania has always been considered the ‘Keystone’ of the U.S. economy. Rich is coal, natural gas, and lumber has allowed PA to grow into a strong industrial producer. During the founding of the American steel industry, Andrew Carnegie helped develop Pittsburgh into the ‘Steel City’…later on the rise of Bethlehem Steel helped grow the currently bustling Lehigh Valley area. As time has passed, PA has retained its Keystone nature; however, its industry focus and distribution of economic prosperity has shifted from the resource-intensive industrial of steel and coal and has transitioned into a serviced-based path. As business hubs in the Lehigh Valley, Philadelphia, and Pittsburgh have flourish, regions of the coal belt and steel-centric areas has dropped. This has left PA in a paradigm where some counties are prospering; while others are seeing its constituents struggle.

Statement of Intent

While the scope of this analysis will not look at solutions to bring new industries into the Keystone state, it will instead look to find correlations in the economic index of each county. In this report we will look at the economic distribution or the state and attempt to find commonalities in those with high and low metrics. With this, we hope to provide the PA Economic Development Association with a better understanding as to how we can better distribute the economic wealth of our state with the current employment numbers.

1. Data Description

Sources

There are three main source of the data used in this analysis:

1. Pennsylvania State Economic Data - *2010*
2. PA State County centroid locations
3. Venues in each county

Pennsylvania State Economic Data (1)

The Pennsylvania State Economic data was retried off of a table on a *Wikipedia* page:

<https://en.wikipedia.org/wiki/List_of_Pennsylvania_counties_by_per_capita_income>

The information cited here is taken from the 2010 US Census report. This table is broken down into each county and provides:

* Per Capita Income – income per person of population
* Median Household Income – middle income for each listed address
* Median Family Income – middle income per family (families can be broke up into several households)
* Population
* Number of Households

PA County Centroids (2)

We defined the centroid of each county as its middle point. Purpose of this was to be able to easily search for nearby venues. This latitude and longitude numbers (2) were retrieved off the PA.GOV website and downloaded as an excel file from the web address below:

<https://opendatanetwork.herokuapp.com/dataset/data.pa.gov/dvjn-d63b>

This dataset lists the centroid county location (latitude, longitude) by county name.

PA County Venues (3)

Final data source was nearby venue data from Foursquare. We found the top 100 nearby venues from each county’s centroid location over a 20 mile radius. Since many of these locations were similar categories but named differnet, we grouped by venue category.

1. Methodology

Overview

To identify which venues help improve the economic distribution of the county, we used the three economic metrics of *Per Capita Income, Median Household Income, and Median Family Income*. The income metrics are deemed best indicators of the economic prosperity of a family or individual as it helps gauge total income and income per person. For example, a family of four which a household income of $100,000 would create a per capita income of only $20,000. To smooth out this, we created the economic metric of **Economic Index**.

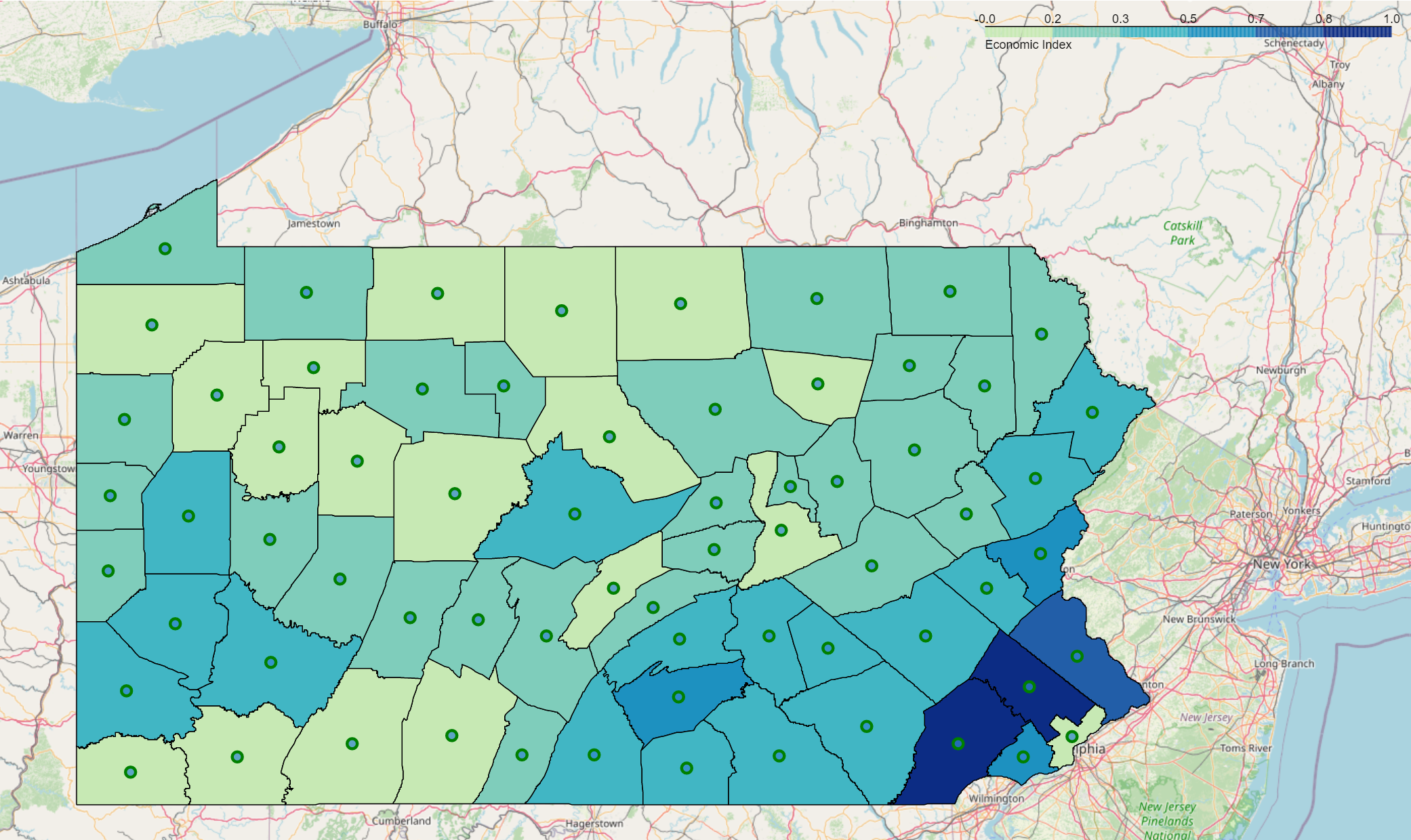
Economic Index

The economic index of a county is defined as the weighted average ranking of the county’s economic metrics in relation to its constituent counties. TO obtain this metric, we first normalized the three economic metrics of *Per Capita Income, Median Household Income, and Median Family Income* using a minimum-maximum scaler: the county with the highest per capita income will have a metric of 1.0.

Once we normalized these metrics, we found the economic index as the average of the normalized economic metrics. This all allow normalization of the economic metrics and allow us a single KPI to utilize for modeling. The scalar used will enable us to divide the county into over-performing counties and under-performing which is simplify the venue classification system.

Exploratory Analysis

Before tackling how we can equalize the economic distribution, we should first look to see how its currently distributed! Using our newly created economic metric of the economic index, we can see how it is currently fouced throughout Pennsylvania.



From the folium chart above, we see that the counties with the highest economic index (dark blue) are centered in the southeast portion of the state. This aligns with the Philadelphia region, which has quickly grown into a financial hub of the US. Much of the eastern portion of the state is well distributed, but we have some improvement in the southwest and northern sections of the state!

Machine Learning Analysis

The over-arching question here is *which local venues improve the economic index and which tend to decrease it*. To answer this, we utilized two different machine learning approaches:

1. Multiple Linear Regression
2. K-Means Clustering

Multiple Linear Regression Model

Our first machine learning method was multiple linear regression (MLS) analysis. Purpose here was to find which venues drove economic index higher at the strongest weight; will help identify which venues are best to add to counties with lower indexes for most improvement.

To start, we found the number of each venue category per county and took the overall state mean and median. We then removed high-frequency venues from this analysis due to fact that probability they occurred in high and low economic index counties was high (and likely and little factor into the economic index). High-frequency venues were defined as those with state medians of 1 or state means above 0.55.

The model itself fit the number of each venue per county as the independent variable (X) to produce the counties scaled economic index (Y) with independent variable coefficients of the model as the economic index factors per venue.

K-Means Clustering

Our second machine learning analysis utilized a cluster method of k-means clustering. With this, our aim was to group the venue categories into three distinct groups:

* Venues common to high economic index counties
* Venues common to low economic index counties
* Venues that are common to average or have no impact to economic index

The dataset was a bit different this time as we did not filter the venue categories based on high-frequency. Instead, we multiplied each counties economic index by the number of venues per category in each county to create an economic index factor.

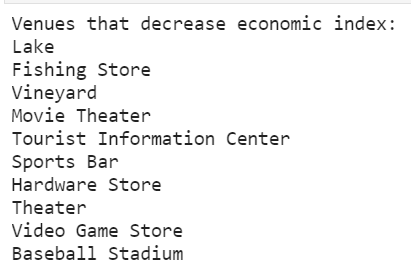
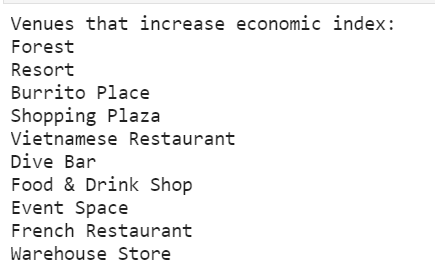
Unlike past instances, these economic index factors were not scaled across each county, but instead across each venue category (summation of economic index factors per venue category would equal 1 whereas sum across each count was equal to economic index multiplied by number of total venues). This enabled us to give more weight to venues in high economic index counties. For high-frequency venues, the overall weight of that venue category will not be a factor.

1. Results

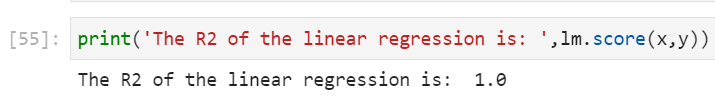
Overall, the results of this analysis yielded mix results.

Linear Regression Results

In performing the multiple linear regression analysis we found the following ten venues that improve and decrease economic index of each county in Pennsylvania:



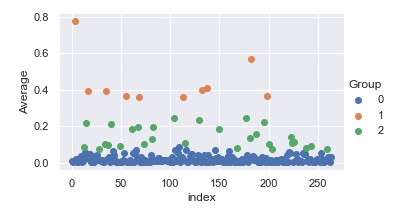
In terms of the model fit, we yielded a R2 of:



This indicates the residual square error is very low, as expected due to large amount of data fitted.

Clustering Results

The results for the clustering method yielded much better results as we were able to cluster the venues into three fairly distinct groups as seen by the graph below.



In this graph we yielded groups of integer numbers as:

* 0 = Group 1 clustered at the lowest economic index
* 1 = Group 2 clustered at the highest economic index
* 2 = Group 3 cluster within the middle range

Looking at the median and mean scaled economic index factors:

* Group 1 has median metric of: 0.01097596433417329 and average metric of 0.01718721687378404
* Group 2 has median metric of: 0.3937132519222071 and average metric of 0.4392683767310633
* Group 3 has median metric of: 0.11589293790786331 and average metric of 0.14075953639772476

From this, we determine that Group 1 venues are likely to decrease economic index, Group 2 are likely to increase economic index, and Group 3 are likely to not impact economic index.

**Group 1 venues tend to decrease the economic index of a county. These venues are:**

Accessories Store

Airport

Airport Terminal

Antique Shop

Apres Ski Bar

Arcade

Argentinian Restaurant

Art Museum

Arts & Crafts Store

Asian Restaurant

Athletics & Sports

Australian Restaurant

Bagel Shop

Bank

Baseball Field

Baseball Stadium

Basketball Court

Beach

Bed & Breakfast

Beer Bar

Beer Garden

Beer Store

Big Box Store

Bistro

Botanical Garden

Boutique

Bowling Alley

Bowling Green

Brazilian Restaurant

Buffet

Burrito Place

Butcher

Cajun / Creole Restaurant

Campground

Candy Store

Caribbean Restaurant

Casino

Cave

Cemetery

Cheese Shop

Chocolate Shop

Churrascaria

Climbing Gym

Clothing Store

Cocktail Bar

College Auditorium

College Basketball Court

Community Center

Concert Hall

Construction & Landscaping

Cosmetics Shop

Cupcake Shop

Deli / Bodega

Department Store

Dessert Shop

Distillery

Dive Bar

Doctor's Office

Dog Run

Drive-in Theater

Eastern European Restaurant

Electronics Store

Event Space

Eye Doctor

Fabric Shop

Fair

Farm

Fish Market

Fishing Store

Flower Shop

Fondue Restaurant

Food

Food & Drink Shop

Food Court

Food Service

Food Stand

Food Truck

Football Stadium

Forest

Fountain

French Restaurant

Fried Chicken Joint

Frozen Yogurt Shop

Fruit & Vegetable Store

Furniture / Home Store

Garden

Garden Center

Gastropub

General Entertainment

German Restaurant

Gift Shop

Golf Course

Golf Driving Range

Gourmet Shop

Greek Restaurant

Gun Range

Gym / Fitness Center

Gym Pool

Gymnastics Gym

Harbor / Marina

Hardware Store

Historic Site

History Museum

Hobby Shop

Hockey Arena

Hockey Rink

Home Service

Hookah Bar

Hot Dog Joint

Hotel Bar

IT Services

Indian Restaurant

Indie Movie Theater

Insurance Office

Irish Pub

Japanese Restaurant

Jewelry Store

Kitchen Supply Store

Korean Restaurant

Lake

Laser Tag

Lingerie Store

Liquor Store

Lounge

Market

Massage Studio

Mediterranean Restaurant

Middle Eastern Restaurant

Miscellaneous Shop

Mobile Phone Shop

Mongolian Restaurant

Monument / Landmark

Moroccan Restaurant

Motel

Motorcycle Shop

Mountain

Movie Theater

Moving Target

Multiplex

Museum

Music Store

Music Venue

National Park

Neighborhood

Nightclub

Noodle House

Optical Shop

Organic Grocery

Other Great Outdoors

Outdoors & Recreation

Outlet Mall

Paella Restaurant

Performing Arts Venue

Peruvian Restaurant

Pet Store

Playground

Plaza

Pool

Post Office

Racetrack

Rafting

Record Shop

Recreation Center

Rental Car Location

Resort

Rest Area

Salad Place

Salon / Barbershop

Scandinavian Restaurant

Science Museum

Sculpture Garden

Shipping Store

Shoe Store

Shop & Service

Shopping Mall

Shopping Plaza

Skate Park

Skating Rink

Ski Area

Ski Lodge

Smoke Shop

Snack Place

Soup Place

South Indian Restaurant

Southern / Soul Food Restaurant

Spa

Speakeasy

Spiritual Center

Sporting Goods Shop

Sports Bar

Summer Camp

Supplement Shop

Sushi Restaurant

Taco Place

Tanning Salon

Tapas Restaurant

Tattoo Parlor

Tea Room

Tex-Mex Restaurant

Thai Restaurant

Theater

Theme Park

Thrift / Vintage Store

Tourist Information Center

Town

Toy / Game Store

Train Station

Truck Stop

Tunnel

Turkish Restaurant

Vegetarian / Vegan Restaurant

Video Game Store

Video Store

Vietnamese Restaurant

Vineyard

Warehouse Store

Water Park

Waterfall

Whisky Bar

Wine Bar

Wine Shop

Wings Joint

Women's Store

Yoga Studio

Zoo

**Group 2 venues help improve the economic index of a county. These venues are:**

American Restaurant

Bar

Brewery

Coffee Shop

Discount Store

Grocery Store

Ice Cream Shop

Italian Restaurant

Pizza Place

Sandwich Place

**Group 3 venues tend to have no impact on the economic index of a county. These venues are:**

BBQ Joint

Bakery

Bookstore

Breakfast Spot

Burger Joint

Café

Chinese Restaurant

Convenience Store

Diner

Donut Shop

Farmers Market

Fast Food Restaurant

Gas Station

Gym

Hotel

Mexican Restaurant

New American Restaurant

Park

Pharmacy

Pub

Restaurant

Scenic Lookout

Seafood Restaurant

State / Provincial Park

Steakhouse

Supermarket

Theme Park Ride / Attraction

Trail

Winery

1. Discussion

As stated in the opening line of the results section, both forms of analysis yielded mixed results. There are several assumptions and inferences made that are likely causing this and will be discussed.

Using linear regression to sort through venues by highest coefficients told us having venues such as resorts, burrito places, shopping plazas, Vietnamese restaurants, dive bars, food & drink shops, French restaurants, and warehouse stores help drive economic index. If we think about this logically – shopping plazas will likely promote increased spending habits and improve number of tourists (as will resorts and local food spots). In addition, higher number of international-inspired restaurants is indicative of a culturally-diverse population and one that has a broad range of working individuals with different backgrounds.

The clustering method did not indicate similar results as it grouped American & Italian restaurants, discount stores, brewery, and pizza/sandwich places. From an economic growth perspective, I would not lean towards these venues are attractions to pull in tourists or high-earning career professionals. Likely what occurred here was including high-frequency venues in the model skewed the weighting even though it was originally not expected to.

1. Conclusion

The two machine learning analysis methods did produce differencing results that was not expected. We are very comfortable with the results produced in the linear regression model and is highly suggested to increase shopping plazas, resorts, and international style restaurants in counties with lower economic indexes. This will not only attract tourists (bosting local businesses), but will offer incentive to high-earning professionals and larger corporations to shift their industrial facilities or corporate headquarters towards these counties.

The likely issue with the clustering method was the inclusion of the high-frequency venues and the economic index factor method we utilized to assign venue weights. If we were to re-do the clustering aspect, we would likely look at using a different clustering method and developing a more accurate metric to assign economic index values to the number of venues per category within a county. Suggested improvement would be to scale the number of venues within a county, and then multiply by the economic index (yielding a more reflective metric as a venue category that is more frequent in a high economic index county would produce higher accuracy results).