Project: Analyzing Socio-Economic Factors and their impact on Quality of Life

Team C:

Samipriya Vanga

Bharath Nanduri

Niveditha Manoharan

Praveena Pusuluri

Date: 5th December 2019.

Index

Executive Summary ---------- **3**

Approach ---------- **4**

Analysis & Insights ---------- **5**

Conclusion ---------- **15**

References ---------- **15**

Executive Summary:

Many dimensions of analyzing social, economic and demographic data have been explored previously by research scholars with a specific focus on research interests. This is an effort that builds upon the previously done research to find out the underlying factors that affect the quality of life in a general sense.

The purpose of this study is to analyze and find the underlying factors that impact the quality of life that is often not considered with high regard. The data involves a subset of Socioeconomic, Demographic and Geographic data for US counties. The data is retrieved from several sources such as Fact Finder, GitHub etc. Having extracted data from multiple sources, the number of variables involved in the dataset exceeded 100. But, the data cleaned for this specific purpose of analysis and only involved variables that are: Important in nature for a socio-economic outcome, strongly dependent on local geography, less ambiguity in their overall presence. After following these analysis specific rules, a total of 2248 counties were considered with 48 socioeconomic indicators which explain most of the socio-economic and geographic information.

With the data made ready for this specific analysis, different types of analyses were run in SAS - University Edition. The efforts made to find out the unexplored relations among variables involved running analysis such as Multiple Regression Analysis, Correlation Analysis, Interaction Effects, Factor Analysis, Cluster Analysis.
The results of the analyses are discussed in the further sections of this report.

Approach:

1. Factor Analysis is performed for summarizing the entire dataset. This has given the opportunity to see data from a different dimension and choose the variables that are important for further analysis.

2. Correlation between Economic Inequality and Race and Ethnicity Homogeneity.

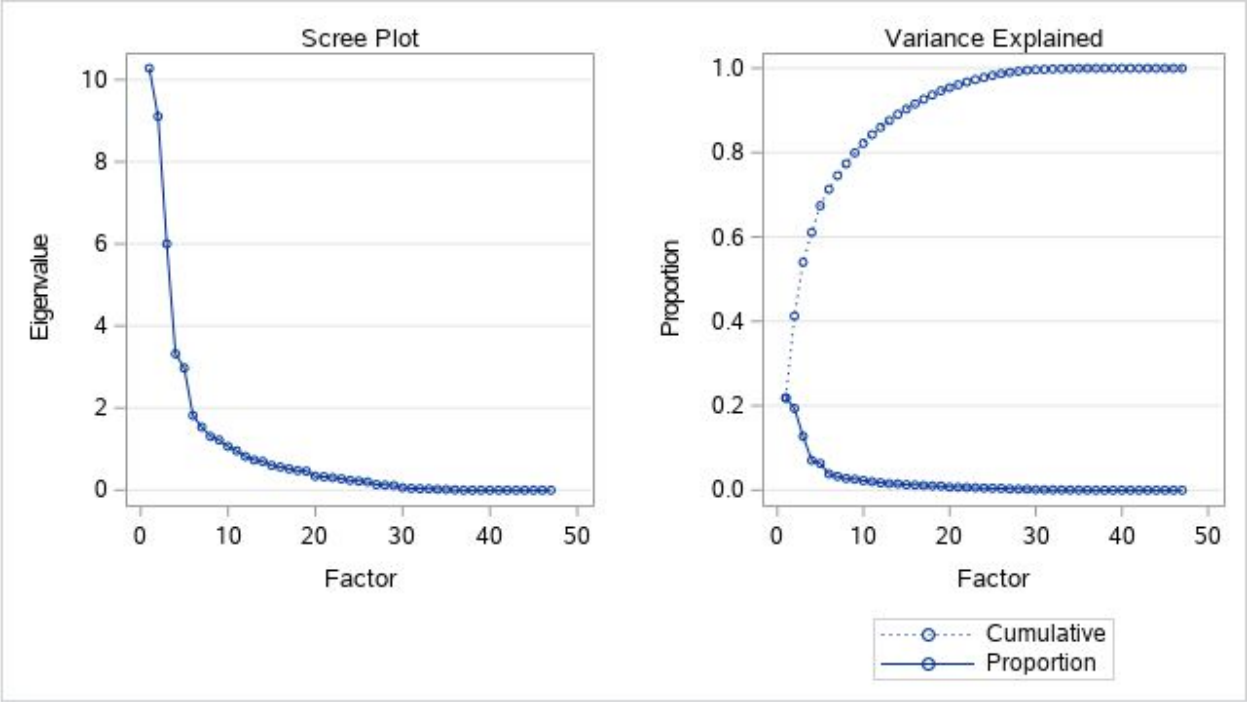
3. The impact of SIRE Homogeneity on the relation between Education Attainment and Family Life.

4. The Impact of Family Environment and Socio-Economic Condition on Violent Crimes.

5. Cluster Analysis is performed to finally divide all the counties into different clusters.

Analysis 1: Factor analysis on the data to summarize the variables.

After performing a factor analysis to figure out into how many factors these variables can be summarised into, we got the following results.



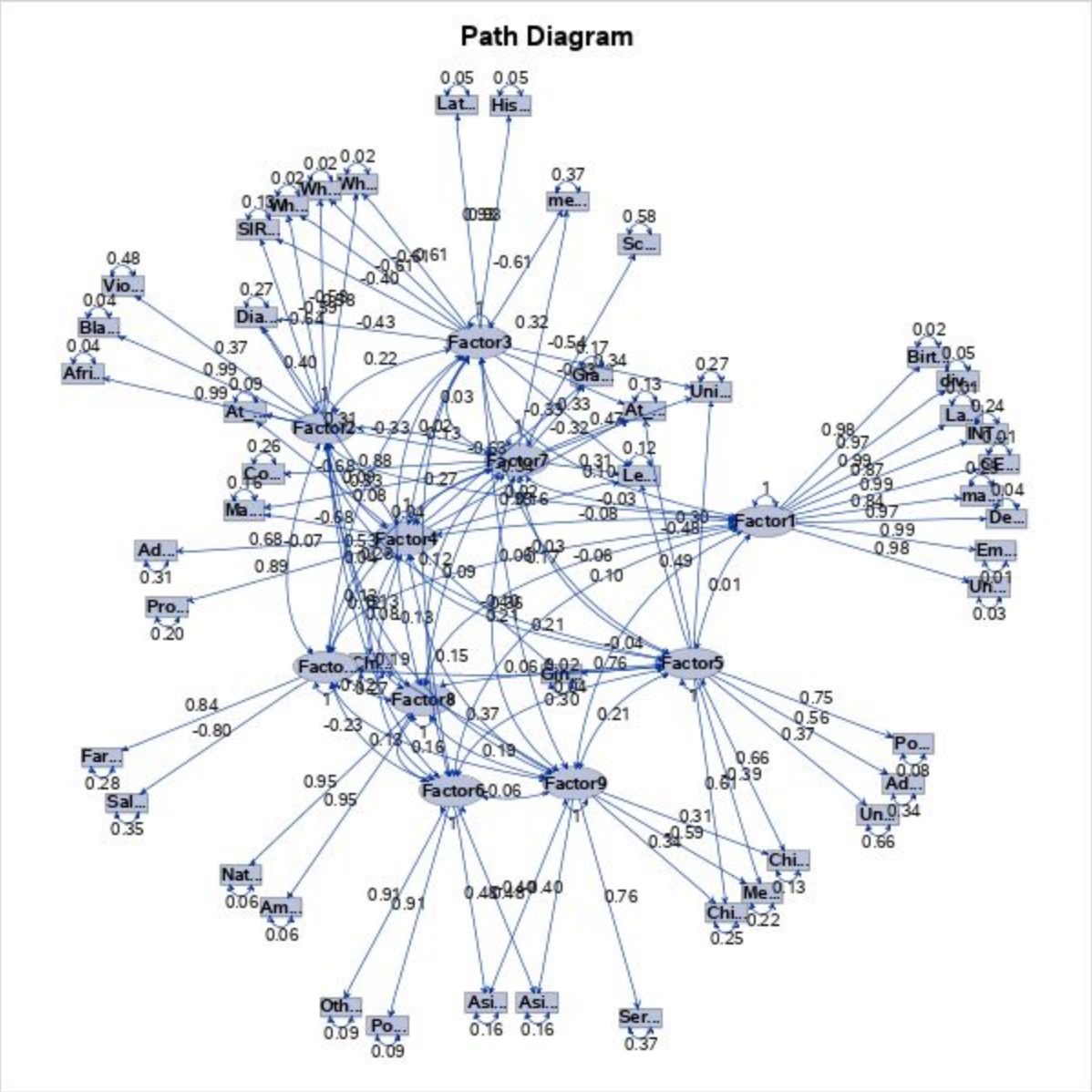
From the scree plot, we were not able to clearly infer how many factors were involved in explaining the total variation. Hence, we look at the Eigenvalues.

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11
10.280732	9.109799	6.000567	3.323054	2.976316	1.819949	1.537571	1.310705	1.222604	1.066457	0.960300

From the Eigenvalues, we can clearly tell that 10 factors were having values greater than 1 and will suffice in summarizing the overall factors.

Now, we would run the analysis again, but this time, selecting the Varimax rotation method, with oblique rotation technique, and only 10 factors, so that we would be able to clearly see under which factor a particular variable falls.

After running this analysis by adjusting the options, the following is obtained as the Path Diagram.



From the path diagram, we can infer which variables fall under which factor.

Factor 1: Birthrate, Deathrate, Marriages, Employment, Unemployment

Factor 2: Latino, Hispanic, African-American, Asian, SIRE, White, White-American, White-Asian

Factor 3: At least Highschool, Undergraduate, Graduate Degree, Earnings

Factor 4: Fishing and Farming, Business professionals, Construction, sales, and office occupations, Production Transportation

Factor 5: Children< 5 living in poverty, Adults>65 living in poverty, Children from poor families, children from single parent house.

Factor 6: Diabetes, Adult obesity, Deaths, Median age

Factor 7: Income, divorced, un-insured, violent crimes, less than high school

Factor 8: school enrollment, graduate degree, median salary

Factor 9: Population, International Mig, Unemployed rate

Factor 10: labor force, population

Analysis 2: Correlation between Economic Inequality and Race and Ethnicity Homogeneity.

Ho: There is no Correlation between Economic Inequality and Race/Ethnicity Homogeneity

H1: There is a Correlation between Economic Inequality and Race/Ethnicity Homogeneity

Pearson Correlation Coefficients, N = 2248 Prob > |r| under H0: Rho=0	
	Gini_Coefficient
SIRE_homogeneity	-0.33151
SIRE_homogeneity	<.0001

The model is a significant model as it has P-Value less than Alpha.

The correlation coefficient is observed as -0.33.

There is a correlation between Economic Inequality and Race/Ethnicity Homogeneity but the relation is negative between them.

Analysis 3: Does SIRE Homogeneity have an impact on the relationship between Education Attainment and Family Life?

Moderation/Interaction Effect:

Ho: There is No Interaction

H1: There is Interaction effect present.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	7368.62080	1228.10347	113.41	<.0001
Error	2241	24268	10.82887		
Corrected Total	2247	31636			

As P-value < alpha, Model is good.

Root MSE	3.29073
Dependent Mean	6.43536
R-Square	0.2329
Adj R-Sq	0.2309
AIC	7612.21047
AICC	7612.27479
SBC	5402.23505

The model explains 23.09% variation in Graduate Degree Attainment.

As there are many variables that can be taken into consideration, we would like to see if there is a moderation effect of SIRE on Children in a single-parent household on Graduation Degree achievement.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.914793	1.062284	16.86	<.0001
Poverty_Rate_below_f	1	0.197181	0.029986	6.58	<.0001
Child_Poverty_living	1	-0.289385	0.025706	-11.26	<.0001
Children_in_single_p	1	-12.245159	3.068043	-3.99	<.0001
Children_Under_6_Liv	1	-0.009469	0.013435	-0.70	0.4810
SIRE_homogeneity	1	-11.750630	1.420980	-8.27	<.0001
Children_*SIRE_homog	1	18.908249	4.284964	4.41	<.0001

The equation is:

Graduate_Degree =

0.19(Poverty_rate_below_federal_frequency)-0.289(Child_Poverty_Living)-12.24(Childr en_in_single_parent_household)-11.75(SIRE)+

18.9(Children_in_single_parent_household*SIRE)

SIRE has a positive moderation effect on the relation between Children_in_single_parent_household & Graduation Degree.

Analysis 4: How Socio-Economic & Family Environment affect Violent Crimes?

Multiple Linear Regression:

Ho: Not a Significant Predictor

H1: Significant Predictor

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	26261329	6565332	260.50	<.0001
Error	2243	56529134	25202		
Corrected Total	2247	82790463			

Root MSE	158.75285	R-Square	0.3172
Dependent Mean	245.89773	Adj R-Sq	0.3160
Coeff Var	64.56052		

This is a good model as the P-value is less than Alpha.

This model explains 31.6% of the variation in the dependent variable-Violent Crimes.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	285.49301	23.67896	12.06	<.0001
Children_in_single_parent_househ	Children_in_single_parent_households	1	636.19000	50.25535	12.66	<.0001
Children_Under_6_Living_in_Pover	Children_Under_6_Living_in_Poverty	1	-0.52581	0.64730	-0.81	0.4167
Child_Poverty_living_in_families	Child_Poverty_living_in_families_below_the_poverty_line	1	1.17633	0.84644	1.39	0.1647
SIRE_homogeneity	SIRE_homogeneity	1	-345.87847	20.71269	-16.70	<.0001

Here, SIRE_Homogeneity and Children_in_Sigle_Parent_household are significant predictors of the Violent_Crimes. Interestingly, contradicting the common belief, the

economic conditions have much less effect on Violent Crimes when compared to conditions that they are being raised.

The homogeneity of a race/ethnicity in a place actually brings down the violent crime rates.

The regression equation can be written as:

Violent_Crimes=285.4+636.19(Children_in_single_parent_house)-345.87(SIRE_Homogeneity)

Analysis 5: Cluster Analysis to divide all the counties into different clusters.

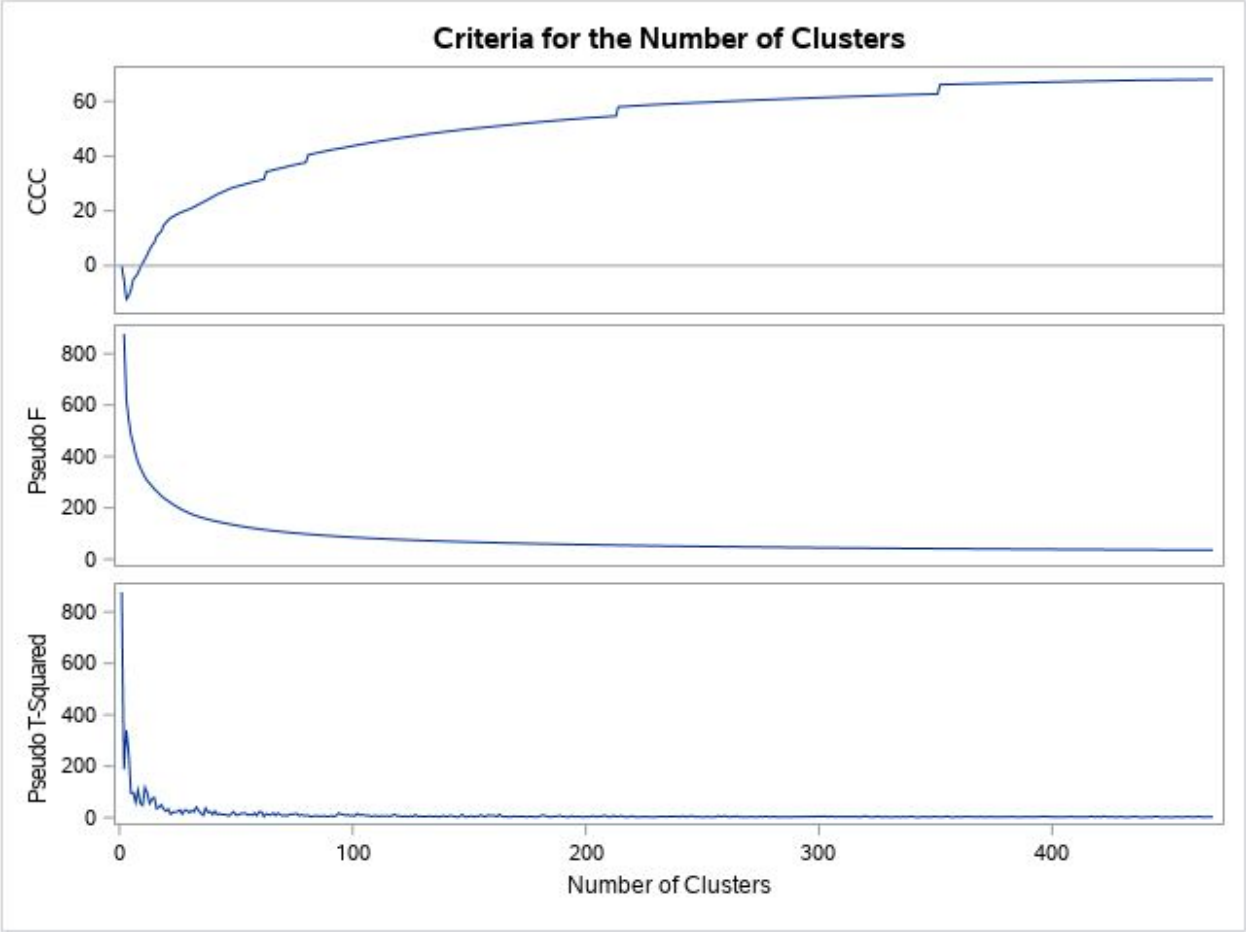
We perform two different types of clustering here :

Hierarchical Clustering - We can perform it under cluster observation in SAS. This helps us find the no. of clusters.

K - means clustering - This clustering technique helps us assign each observation under a specific cluster. On performing cluster observation, we can identify the no. of clusters from the Cluster history table. In this table, under Cluster cubic criterion, when we come across the first positive value that would be the no. of clusters identified.

Cluster History											
Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
12	CL27	CL145	87	0.1032	0.0099	.580	.565	4.53	280	34.1	
11	CL21	CL13	736	0.0612	0.0118	.568	.558	3.14	294	94.0	
10	CL14	CL24	184	0.0999	0.0126	.555	.549	1.86	311	40.6	
9	CL35	CL18	192	0.0901	0.0173	.538	.539	-.42	326	78.9	
8	CL15	CL16	256	0.0938	0.0179	.520	.528	-2.3	347	67.6	
7	CL8	CL10	440	0.1030	0.0225	.498	.514	-4.2	370	63.9	
6	CL20	CL11	1027	0.0665	0.0233	.474	.496	-5.7	405	161	
5	CL6	CL17	1173	0.0748	0.0403	.434	.473	-10	430	226	
4	CL7	CL12	527	0.1155	0.0545	.380	.441	-16	458	135	
3	CL5	CL19	1529	0.0782	0.0585	.321	.391	-16	531	302	
2	CL4	CL9	719	0.1188	0.0593	.262	.287	-5.2	797	131	
1	CL2	CL3	2248	0.1084	0.2619	.000	.000	0.00	.	797	

Here, we encounter the first positive value in the 10th cluster and hence we go with the assumption that we have 10 clusters.



Above plots give us a descriptive way to identify the no. of clusters. On performing K-means clustering with no. of clusters as 10, we get the below cluster summary table.

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	4	0.1238	1.0241		4	1.0046
2	194	0.0985	1.4797		8	0.5778
3	88	0.0919	1.0109		2	0.7622
4	42	0.1160	1.3422		1	1.0046
5	590	0.0828	1.3102		10	0.3993
6	10	0.1187	0.9641		2	1.1713
7	160	0.0981	1.4597		10	0.8586
8	506	0.1046	1.3508		2	0.5778
9	1	.	0		4	1.7306
10	750	0.0736	1.2278		5	0.3993

We find that under cluster 9 there is only one observation and cluster 1 only 4 observations. Hence, we re-run the analysis with 8 clusters and obtain the below output.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	631	0.0853	1.4090		2	0.5044
2	870	0.0742	1.2493		1	0.5044
3	206	0.0986	1.5162		2	0.7223
4	419	0.1024	1.4034		1	0.7556
5	2	0.1954	0.8848		7	1.0871
6	112	0.0934	1.0429		4	0.8661
7	92	0.1209	1.5214		4	1.0048
8	13	0.1259	1.2612		4	1.0143

Here, we find that cluster 5 has only 2 observations and hence we reduce the no. of clusters as 7 and re-run the analysis. However, the variation explained by the clusters have not significantly improved.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	10	0.1187	1.0083		6	1.1947
2	1153	0.0799	1.2632		4	0.5905
3	26	0.1180	1.5138		6	1.0616
4	592	0.0965	1.3081		2	0.5905
5	191	0.0944	1.2481		4	0.9071
6	116	0.0984	1.4135		4	0.9357
7	257	0.1087	1.4807		4	0.8008

From the above table, we can see that the distance between clusters 1 and 10 is maximum. Also, cluster 4 is near to clusters 5,6 and 7.

Approximate Expected Over-All R-Squared =	0.17615

17.6 % of the variation in data is explained by the model.

Conclusion

In the different type of analyses performed, we were able to find some interesting insights that are often not thought about. But, broadly classifying, the quality of life depends on the fundamental elements like Education, Income, Security and Social culture.

In our analyses, we were able to find that the SIRE_Homogeneity will really tell a lot about a county as it affects Education, Income, and crimes also.

The places where income inequality prevails is not the place where crimes happen. It actually depends on the SIRE_Homogeneity.

Overall, the analyses performed resulted in models that are a good fit and explained the intended purpose of running them.

References

- Kaggle
- GitHub
- Google Scholars
- Wikipedia