Estimation of Poverty Rate of the US Counties

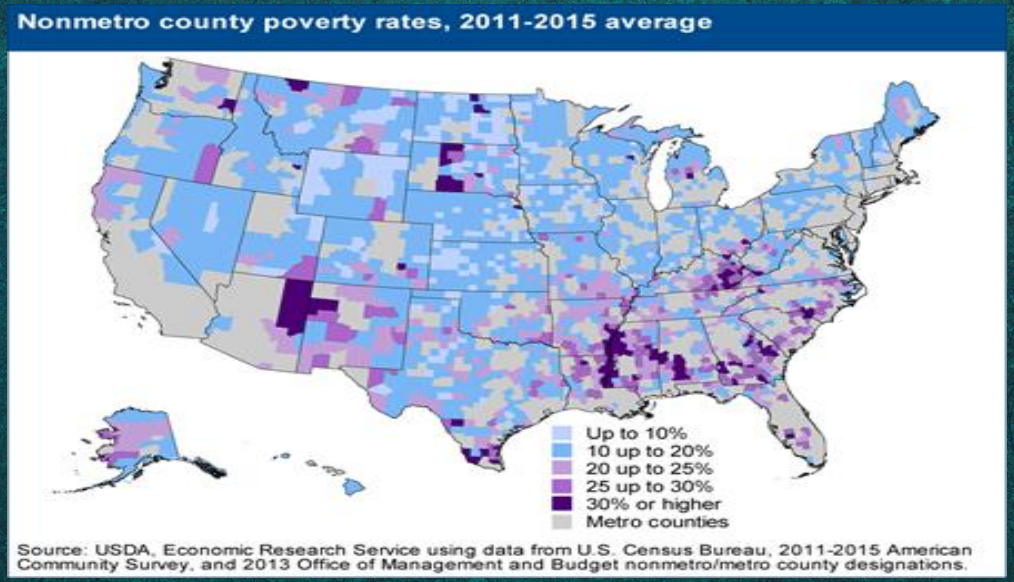


Image Source [6]

Index

Goals and Objective

- Primary objective is to study variability in the poverty rate in the US counties by means of one or more of independent or control variable and provide best suitable model to quantify relationships in determining target value
- Goal is to design various models to take into consideration the effect of various factors like employment, population and education to predict the poverty rate in all US Counties
- Furthermore, wish to analyze the status of a county based on whether it is metropolitan or not

Data Sources

a. List of datasets:

- Socioeconomic indicators like poverty rates, population change, unemployment rates, and education levels vary geographically across U.S. States and counties[1]

Dataset name	# of variables	# of records
Unemployment.xlsx	48	3274
PovertyEstimates.xlsx	30	3194
PopulationEstimates.xlsx	117	3273
Education.xlsx	47	3283
Final Dataset after merging		
MergedDataset.xlsx	17	3283

b. Source Link:

- https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Merging of Datasets based on unique common variable

FIPS	State	Area_Name
00000	US	United States
01000	AL	Alabama
01001	AL	Autauga County
01003	AL	Baldwin County
01005	AL	Barbour County
01007	AL	Bibb County
01009	AL	Blount County
01011	AL	Bullock County
01013	AL	Butler County
01015	AL	Calhoun County
01017	AL	Chambers County
01019	AL	Cherokee County
01021	AL	Chilton County
01023	AL	Choctaw County

Individual dataset - sample data

FIPStxt	State	Area_name
01000	AL	Alabama
02000	AK	Alaska
04000	AZ	Arizona
05000	AR	Arkansas
06000	CA	California
08000	Co	Colorado
09000	CT	Connecticut
10000	DE	Delaware
11000	DC	District of Columbia
12000	FL	Florida
13000	GA	Georgia
15000	HI	Hawaii
16000	ID	Idaho

List of all states falling under the USA

All the four individual datasets have common unique id FIPS Code defined as State-County FIPS Code. It is unique for each county falling under the states. In our dataset, we are covering all 52 USA states including federal district DC and Puerto Rico.

Exploration and Summary of dataset

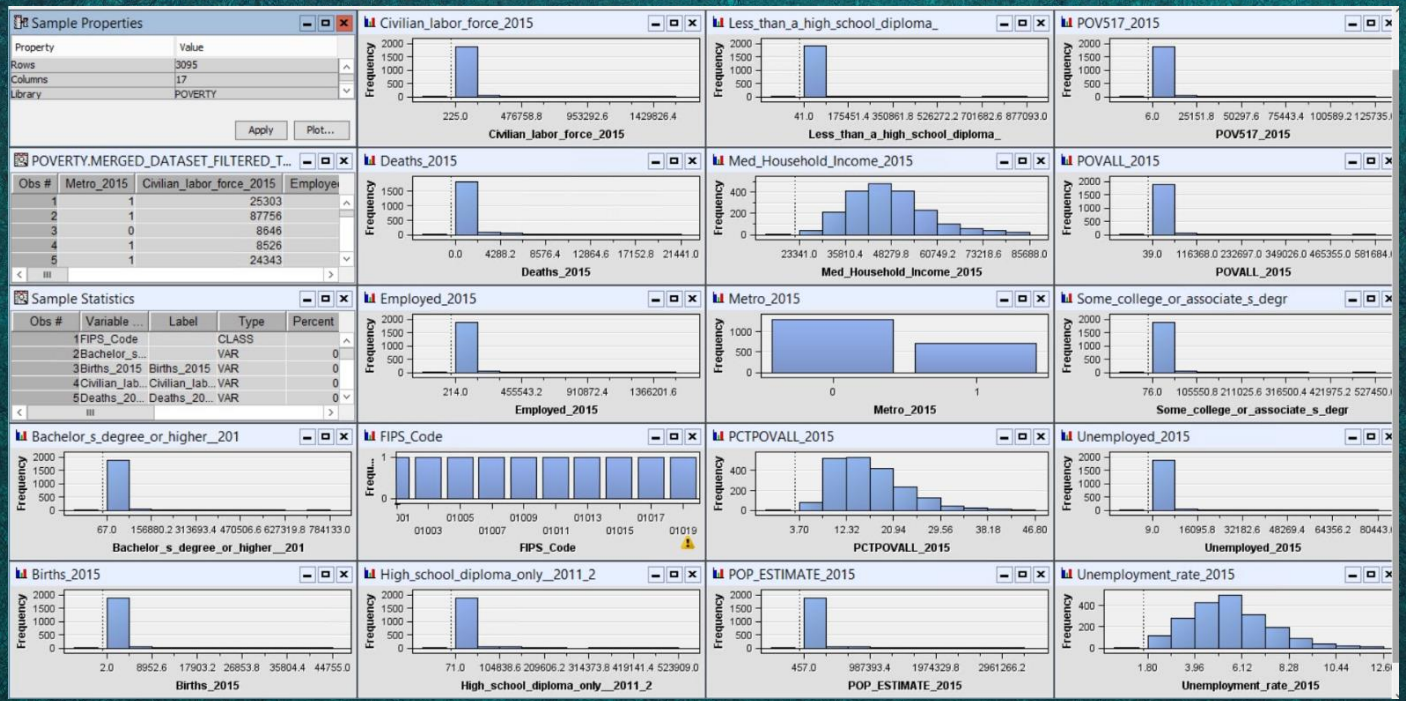
- **Summary Statistics**

Proportion of missing values is less than 1% which is viable option.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Bachelor_s_degree_or_higher__201	INPUT	16599.61	50982.32	3092	3	2	2854	784133	7.44894	76.2376
Births_2015	INPUT	1106.61	3015.029	3068	27	1	293	44755	7.149487	68.54157
Civilian_labor_force_2015	INPUT	43236.73	112111.2	3092	3	77	11331	1588671	6.604188	58.54194
Deaths_2015	INPUT	776.1444	1729.825	3068	27	0	269	21441	5.786117	43.61475
Employed_2015	INPUT	40936.13	106293.9	3092	3	73	10671	1517978	6.615205	58.81576
High_school_diploma_only__2011_2	INPUT	16994.08	36878.61	3092	3	55	6315	523909	6.264789	53.63575
Less_than_a_high_school_diploma_	INPUT	15114.9	41864.27	3057	38	33	5866	877093	10.23664	148.1596
Med_Household_Income_2015	INPUT	47879.16	10754.84	3068	27	22894	46593	85688	0.755292	0.683789
Metro_2015	INPUT	0.366397	0.481898	3095	0	0	0	1	0.554844	-1.69324
POP_ESTIMATE_2015	INPUT	89271.25	226791.6	3092	3	115	25559	3290245	6.741525	61.3364
POV517_2015	INPUT	2929.499	8407.062	3068	27	6	901	127547	8.218986	87.03105
POVALL_2015	INPUT	13212.75	36329.97	3068	27	16	4059	581684	7.918349	82.5261
Some_college_or_associate_s_degr	INPUT	10956.59	30221	3088	7	16	2712	527450	7.546956	82.40459
Unemployed_2015	INPUT	2300.602	5987.101	3092	3	4	653	80443	6.69171	58.98378
Unemployment_rate_2015	INPUT	5.535058	1.871977	3092	3	1.8	5.3	13	0.778777	0.842455
PCTPOVALL_2015	TARGET	16.36617	6.305925	3068	27	3.7	15.3	47.4	1.123165	1.988783

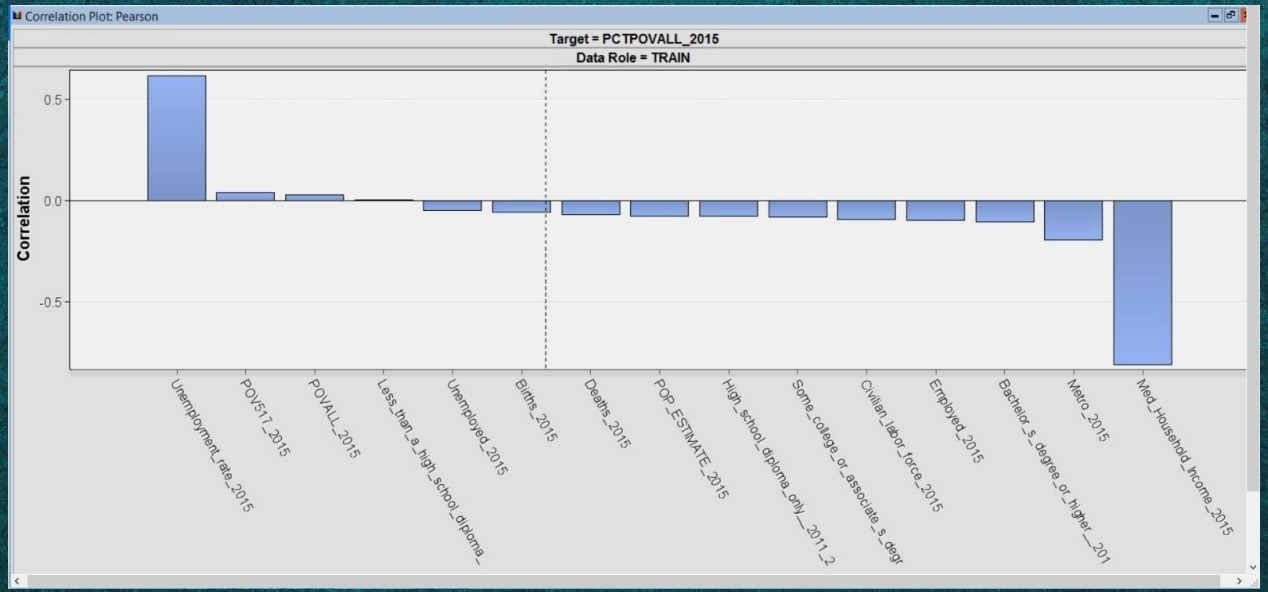
Exploration and Summary of dataset

- **Histograms**



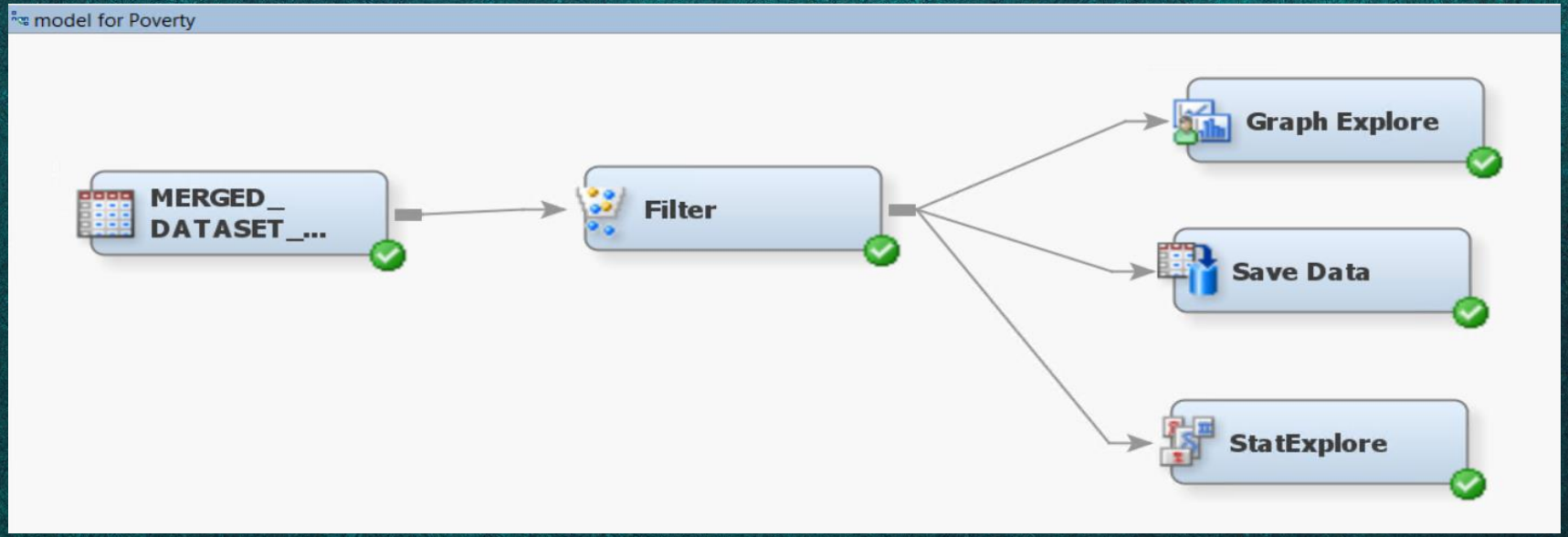
Exploration and Summary of dataset

- **Correlation**
  Unemployment and poverty are negatively correlated with education, income and population growth



Data Filtering

- Metro_2015 variable lists all the counties (1-metro and 0-non-metro) and states (null-states). Hence, we filtered out 'states' by using Filter Node while cleaning the data.



Data Modelling (Primary Model)

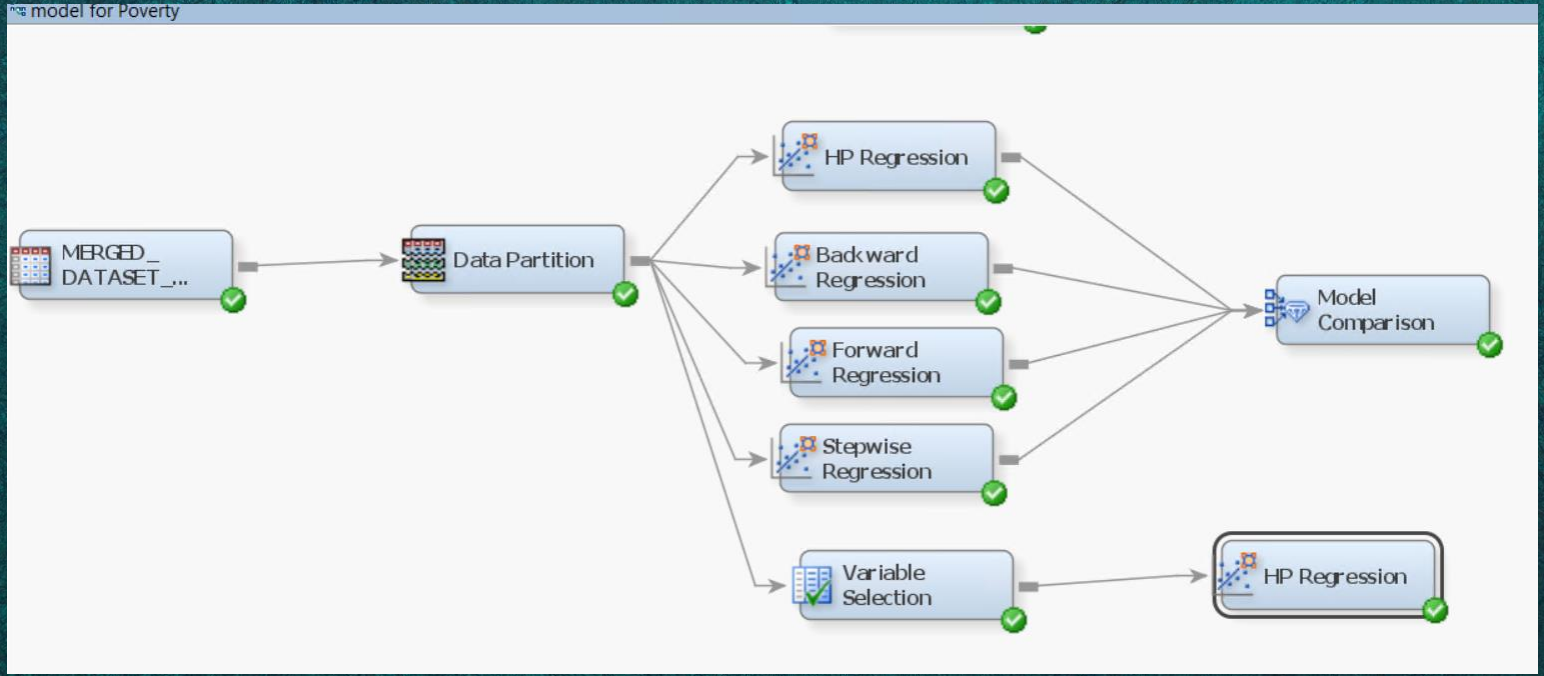
- **Multiple Linear Regression**

Target Variable: PCT_POVALL_2015 – This variable represents % estimate of people of all ages in poverty 2015.

- The linear regression model will determine the **regression equation** that helps in predicting the estimation of poverty of people of all ages.
- Dataset will be partitioned into **training, validation** datasets.
- When the linear regression is performed on target and independent variables, significant variables will be selected based on the **confidence level** and **p-value**.
- Strength of the model will be determined by comparing **Adj. R-Square** value which shows how strongly the coefficients predict the value of target variable.

Data Modelling (Primary Model)

- **Multiple Linear Regression**



Data Modelling (Primary Model)

- **Model Comparison**

Fit Statistics

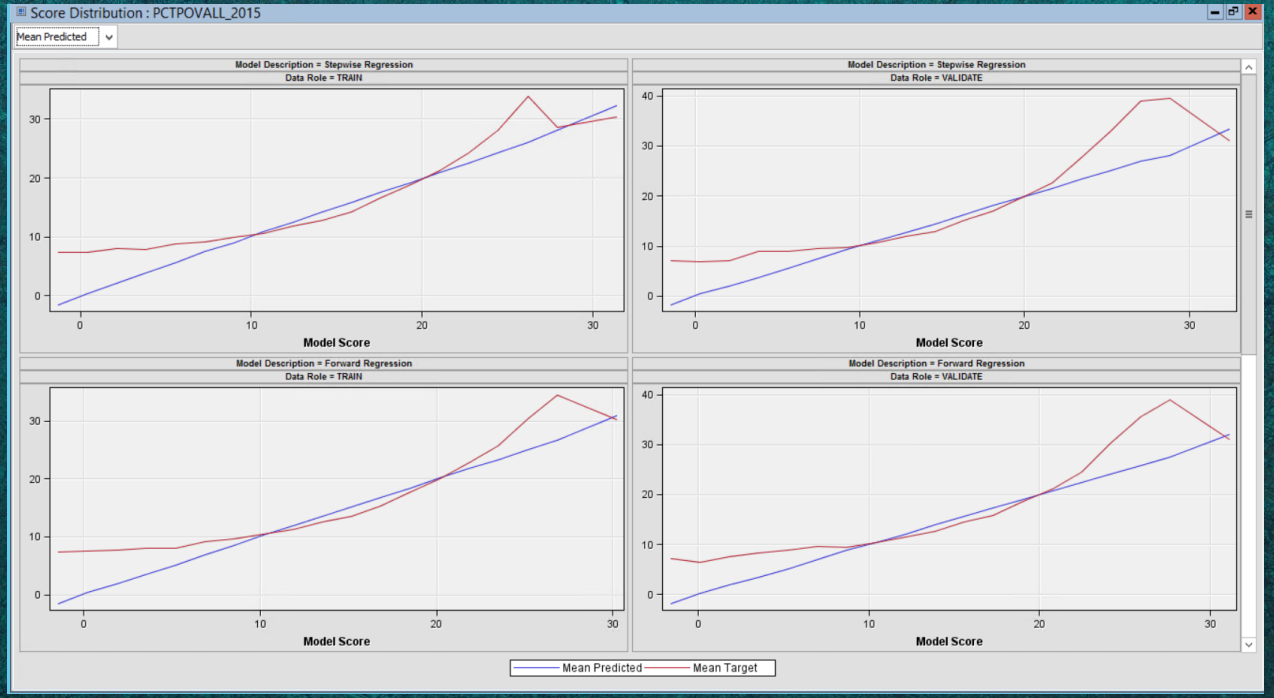
Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Valid: Average Squared Error	Valid: Divisor for ASE	Valid: Maximum Absolute Error	Valid: Sum of Frequencies	Valid: Root Average Squared Error	Valid: Sum of Squared Errors
Y	HPReg2	HPReg2	Backward ...	PCTPOVAL...	PCTPOVAL...	14.35521	10.45085	1542	22.4066	1542	3.232778	16115.21	14.35521	1526	24.07643	1526	3.788826	21906.04
	HPReg4	HPReg4	Forward Re...	PCTPOVAL...	PCTPOVAL...	14.41841	10.35669	1542	22.42844	1542	3.218181	15970.01	14.41841	1526	24.0753	1526	3.797158	22002.49
	HPReg	HPReg	HP Regres...	PCTPOVAL...	PCTPOVAL...	14.4542	10.42622	1542	22.43927	1542	3.228966	16077.24	14.4542	1526	24.09702	1526	3.801867	22057.1
	HPReg5	HPReg5	Stepwise R...	PCTPOVAL...	PCTPOVAL...	14.49915	10.37559	1542	22.44838	1542	3.221116	15999.16	14.49915	1526	24.10193	1526	3.807775	22125.71

Model Description	Valid: Average Squared Error
Backward ...	14.35521
Forward Re...	14.41841
HP Regres...	14.4542
Stepwise R...	14.49915

- Fit statistics calculated from validation data select the best model from the sequence
- Average Squared Error of Validation dataset is minimum for Backward Linear Regression model which is 14.35, which shows highest accuracy in predicting the target variable. Thus, this model can be considered as strong model.

Data Modelling (Primary Model)

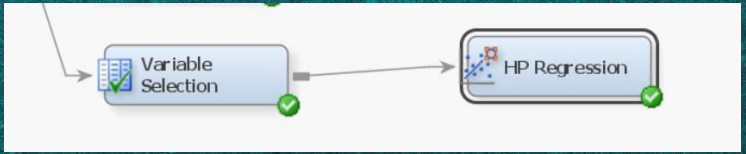
- Model Score Comparison



Final Model
Multiple Backward Linear Regression

- **Variable Selection**

Based on R-square values, variable selection node selects significant variables for prediction of target values.



Effects Chosen for Target: PCTPOVALL_2015

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: Med_Household_Income_2015	1	0.655048	2924.392650	<.0001	34955	11.952886
Var: Bachelor_s_degree_or_higher__201	1	0.030674	150.209483	<.0001	1636.844234	10.897077
Class: Metro_2015	1	0.008700	43.786903	<.0001	464.242476	10.602314
Var: High_school_diploma_only__2011_2	1	0.001480	7.482702	0.0063	79.000965	10.557812
Var: POV517_2015	1	0.003486	17.810537	<.0001	186.005915	10.443588
Var: Some_college_or_associate_s_degr	1	0.000661	3.380733	0.0662	35.252341	10.427426
Var: POP_ESTIMATE_2015	1	0.001990	10.245274	0.0014	106.192245	10.364998
Var: Less_than_a_high_school_diploma_	1	0.001334	6.895549	0.0087	71.198712	10.325315

Final Model
Multiple Backward Linear Regression

- **Regression Analysis**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	37247	4138.57835	403.26	<.0001
Error	1526	15661	10.26280		
Corrected Total	1535	52908			

Root MSE	3.20356
R-Square	0.70400
Adj R-Sq	0.70225
AIC	5124.58229
AICC	5124.75552
SBC	3639.95166
ASE (Train)	10.19598
ASE (Validate)	14.23287

Backward Regression Output

Adj. R-Sq = 70.22%
ASE (Validate) = 14.23

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	37128	6187.92771	599.55	<.0001
Error	1529	15781	10.32091		
Corrected Total	1535	52908			

Root MSE	3.21262
R-Square	0.70174
Adj R-Sq	0.70056
AIC	5130.27162
AICC	5130.36593
SBC	3629.63018
ASE (Train)	10.27387
ASE (Validate)	14.30266

HP Node Regression Output (Using Variable Selection Node)

Adj. R-Sq = 70.05%
ASE (Validate) = 14.30

Final Model
Multiple Backward Linear Regression

Prediction Equation:

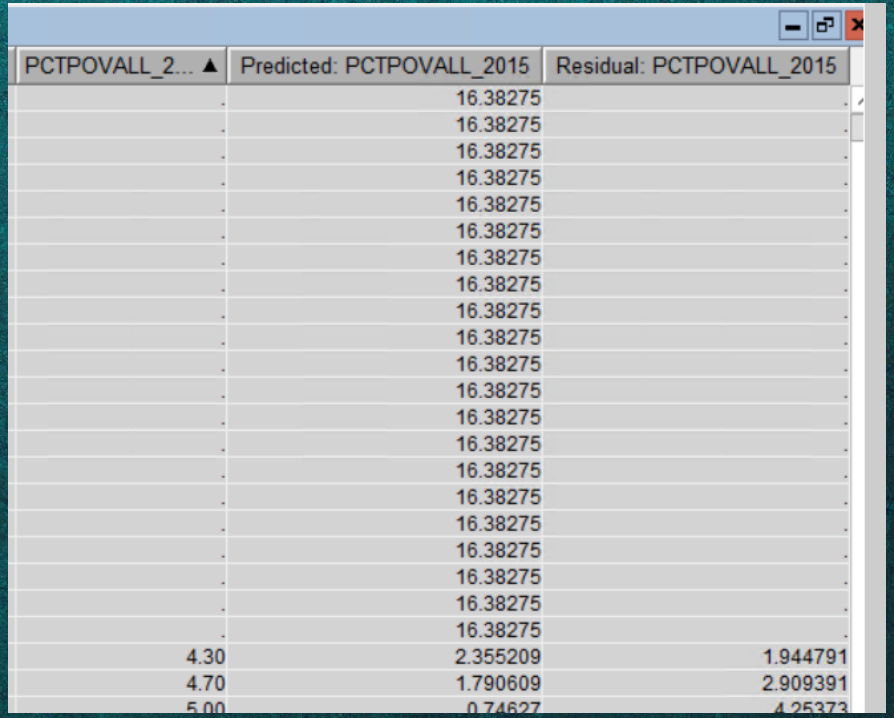
Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	41.473086	0.500621	82.84	<.0001	0
Metro_2015 0	1	-1.404613	0.199016	-7.06	<.0001	1.33483
Metro_2015 1	0	0	.	.	.	.
Civilian_labor_force_2015	1	0.000250	0.000085697	2.92	0.0036	13933
Employed_2015	1	-0.000261	0.000085172	-3.07	0.0022	12321
High_school_diploma_only__2011_2	1	-0.000087962	0.000011834	-7.43	<.0001	28.28836
Less_than_a_high_school_diploma_	1	0.000009217	0.000004002	2.30	0.0214	4.76708
Med_Household_Income_2015	1	-0.000512	0.000009312	-55.02	<.0001	1.34831
POP_ESTIMATE_2015	1	0.000025853	0.000009633	2.68	0.0074	743.10508
POV517_2015	1	-0.000089969	0.000043309	-2.08	0.0379	23.27825
Some_college_or_associate_s_degr	1	-0.000062666	0.000015978	-3.92	<.0001	36.68586

% estimate of people of all ages in poverty 2015 (**PCT_POVALL_2015)** = 41.473086 – 1.404613 (Metro_2015 0) + 0.000250 (Civilian_labor_force_2015) -0.000261 (Employed_2015) -0.000087962 (High_school_diploma_only_2011_2) + 0.000009217 (Less_than_a_high_school_diploma) -0.000512 (Med_household_Income_2015) + 0.000025853 (POP_ESTIMATE_2015) -0.000089969 (POV17_2015) -0.000062666 (Some_college_or_associate_s_degr)

Final Model
Multiple Backward Linear Regression

- **Strengths**
  - Selected regression model predicts missing values for target variable there by estimating % Poverty in the year of 2015
  - Higher adjusted R-square values provide enough evidence to support the conclusions derived
  - Regression models make their best predictions for cases near the centers of the input distributions in case of unusual inputs



PCTPOVALL_2... ▲	Predicted: PCTPOVALL_2015	Residual: PCTPOVALL_2015
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
.	16.38275	.
4.30	2.355209	1.944791
4.70	1.790609	2.909391
5.00	0.74627	4.25373

Final Model
Multiple Backward Linear Regression

- **Strengths**

  - Low residual values represents the strength of model in determining accurate % Poverty rate.

  - Intercept and parameter estimates are chosen to minimize the squared error between the predicted and observed target values (least squares estimation)

PCTPOVALL_2015	Predicted: PCTPOVALL_2015	Residual: PCTPOVALL_2... ▲
21.40	21.40367	-0.00367
5.30	5.28815	0.01185
19.20	19.18794	0.012062
20.50	20.48705	0.012945
9.70	9.681733	0.018267
14.80	14.77807	0.02193
12.80	12.77164	0.028365
17.00	16.96848	0.031523
12.70	12.65688	0.043122
19.50	19.45135	0.048649
17.50	17.45005	0.049953
16.50	16.44653	0.053466
22.10	22.03881	0.061186
13.80	13.73508	0.064918
22.70	22.63332	0.066675
17.50	17.43169	0.068312
9.20	9.12208	0.07792
9.40	9.319876	0.080124
24.40	24.31802	0.08198
20.50	20.41478	0.085216
16.00	15.91444	0.085559
16.60	16.50658	0.093417
17.60	17.50349	0.096514
19.40	19.30215	0.09785
19.10	18.99959	0.10041
16.20	16.09844	0.101556
17.10	16.99813	0.101874

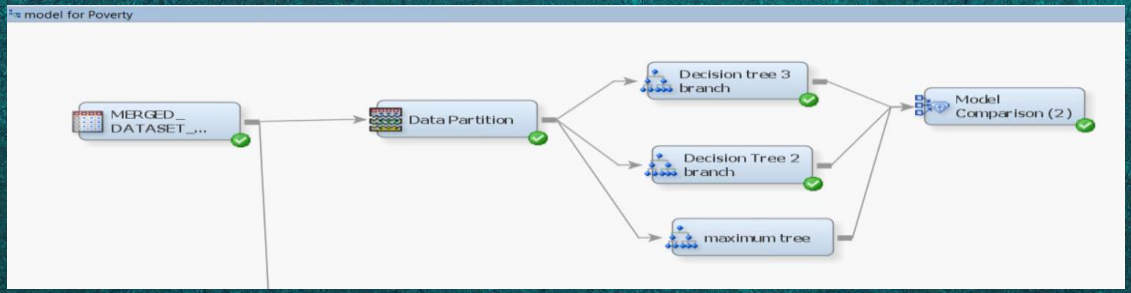
Final Model
Multiple Backward Linear Regression

- **Limitations**
  - While this model can generate a prediction, this prediction can be biased beyond reason if there are missing values in the input dataset
  - Few of the records show high residual values which suggests this model is not able to determine accurate target variable values, which could be a result of manual error in capturing data or unconventional factors not considered during analysis.

PCTPOVALL_2015	Predicted: PCTPOVALL_2015	Residual: PCTPOVALL_2...
47.10	22.99876	24.10124
47.40	23.8324	23.5676
46.30	26.22744	20.07256
44.20	24.55472	19.64528
46.80	27.71489	19.08511
44.00	25.14815	18.85185
44.70	27.73716	16.96284
42.30	26.45682	15.84318
35.10	19.28833	15.81167
43.30	27.69206	15.60794

Data Modelling (Secondary Models)

- **Decision Tree**



Target Variable: Metro_2015 – This binary variable shows status of County as Metro or Non-Metro

•A decision tree model designed using Metro_2015 as target variable will efficiently determine the **classification** of population into **Metro and Non-metro** counties.

•Dataset will be partitioned into **training and validation** datasets before implementing decision tree rules.

•The attributes that will be considered in selecting best model will be **fit statistics, misclassification rate and average square error**.

Data Modelling (Secondary Models)

- **Decision Tree**

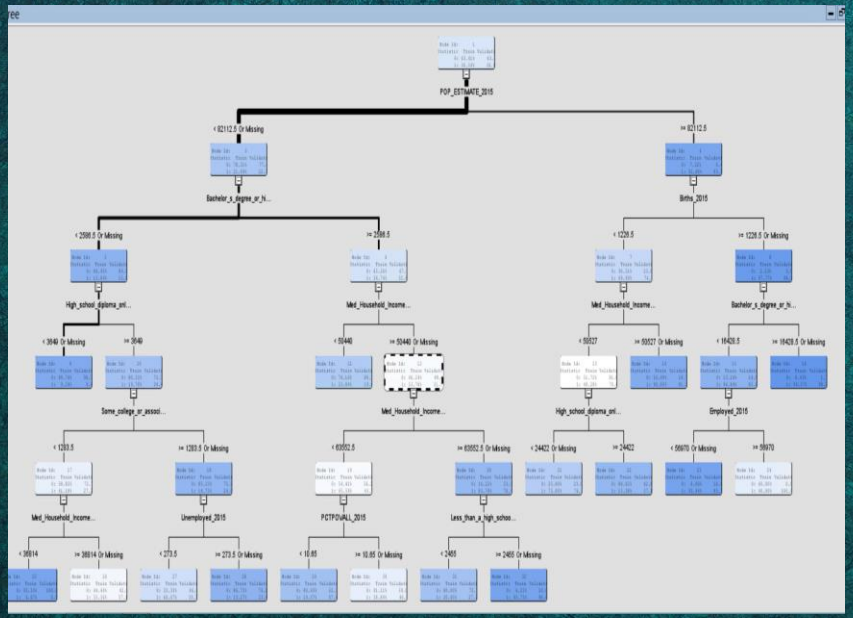
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision tree 3 branch	0.17636	0.12408	0.16548	0.13373
	Tree2	Decision Tree 2 branch	0.18282	0.11207	0.14609	0.13442
	Tree3	maximum tree	0.18282	0.11207	0.14609	0.13442

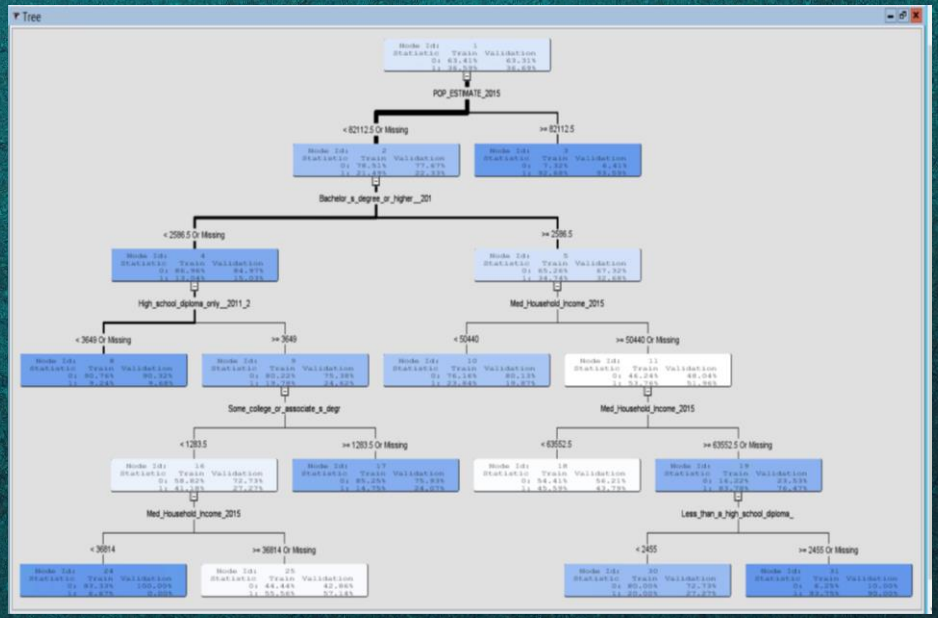
Based on lowest misclassification rate and average squared error for validation dataset, we conclude that decision tree with 3 branches provides more accurate classification.

Data Modelling (Secondary Models)

- **Decision Tree**



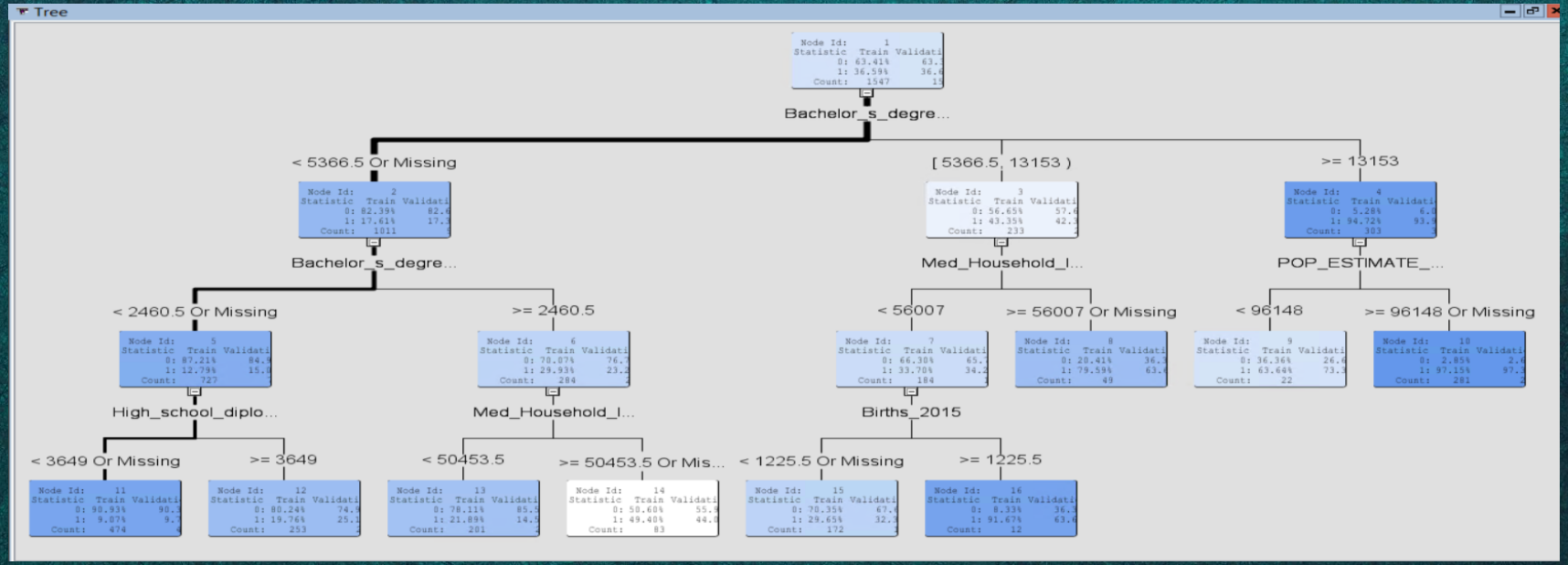
Interactive Tree : max branches



Decision Tree : 2 branches

Data Modelling (Secondary Models)

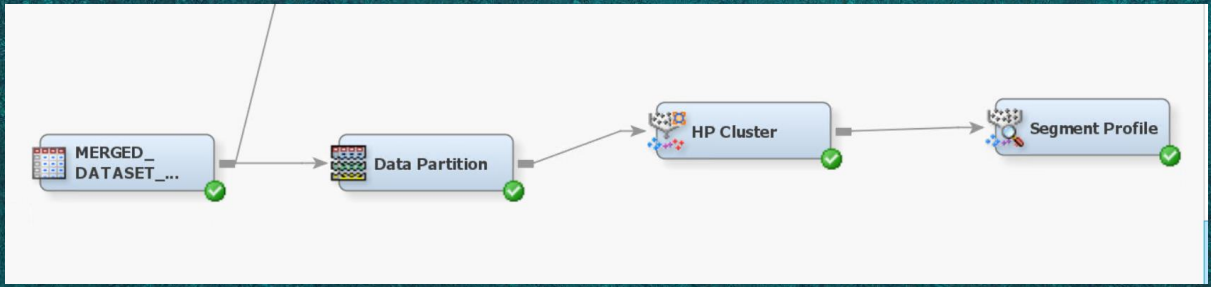
- **Decision Tree**



Decision Tree : 3 branches

Data Modelling (Secondary Models)

- **Clustering**

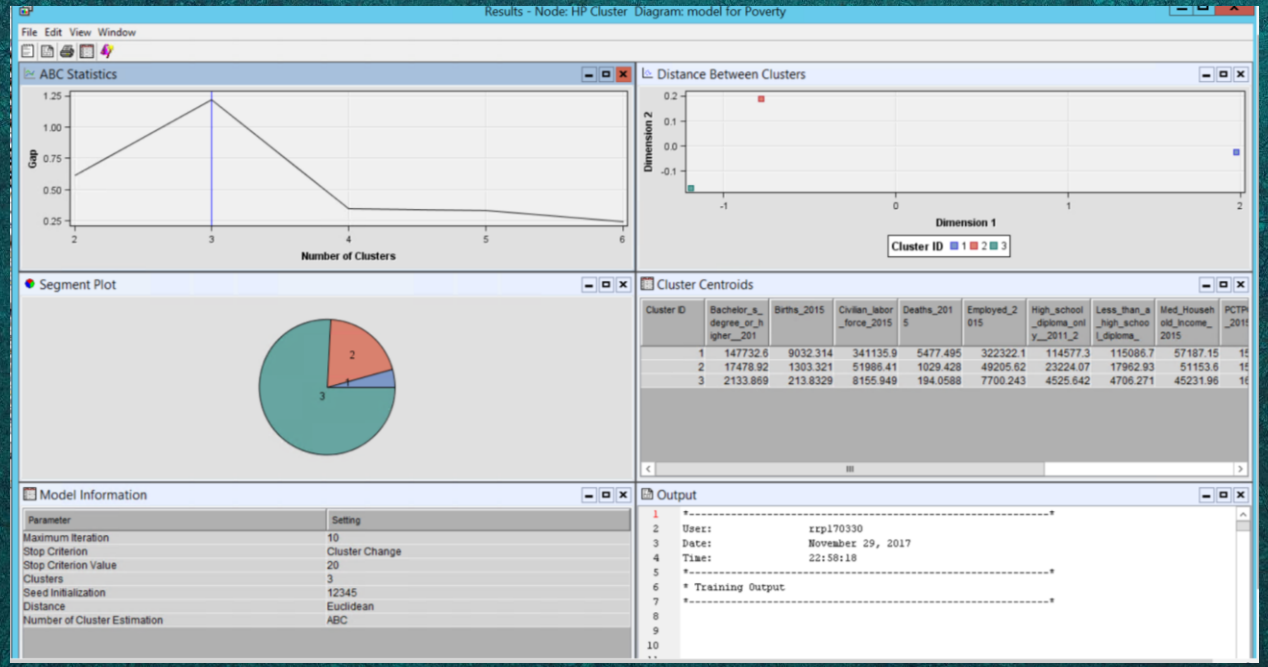


• Clustering is performed to create collection of objects similar to each other which will give insight into data distribution.

• Dataset is **partitioned** into training and validation dataset before performing clustering. **Segment Profile node is used** to compare the distribution of a variable in an individual segment to the distribution of the variable overall.

Data Modelling (Secondary Models)

- Clustering



Recommendation

- The regression equation determines % Poverty rate in a particular county based on significant factors. It can be used to determine the estimation of poverty rate for new dataset values.
- This regression model can also be used for causal inference and studies.
- This model can be used by education boards to increase or decrease the funds spent on education system in different counties in order to lower the poverty rate.
- Census board can use this model in identifying poverty line index based on population estimate and average household income.
- By estimating the poverty rate and considering factors like unemployment and education, an analysis can be done to set up employment opportunities in targeted counties.

References

- [1] https://www.ers.usda.gov/data-products/county-level-data-sets/
- [2] https://www.youtube.com/watch?v=TnWRJQb5z4c&list=PLVBcK_IpFVi-xzvJiOlf33UvVbRoLRu0z&index=4
- [3]http://support.sas.com/documentation/cdl/en/emgsj/67981/HTML/default/viewer.htm#n1cpd0rgpneqwqn16mfcxp4sbjsb.htm
- [4] Applied Analytics using SAS Enterprise Miner ebook
- [5] Data Mining for Business Intelligence, 2nd Edition, by Shmueli, Patel and Bruce. Wiley, ISBN-10: 0470526823, ISBN-13: 978-0470526828
- [6]https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/geography-of-poverty.aspx