Topic: Selected Public Health Indicators by Chicago Community Area

INTRODUCTION

This dataset contains a selection of 21 indicators of public health significance by Chicago community area, with the most updated information available. The indicators are rates, percent's, or other measures related to natality, mortality, infectious disease, lead poisoning, and economic status.

The Chicago Department of Public Health (CDPH) calculated the indicators using a variety of sources, including:

- Geocoded annual birth and death certificate datasets supplied by the Illinois Department of Public Health (IDPH)
- Census tract-level counts and estimates obtained from the U.S. Census Bureau 2000 census, 2010 census, and 2006-2010 American Community Survey
- Case reports and laboratory reports received under the notifiable disease rules of the Illinois

It determines the Target of Healthy Chicago 2020 and the value of Chicago as a whole using the Life Expectancy Variable. Most indicators are percent's, crude rates, age-adjusted rates, or age-specific rates. A crude rate is the total number of events occurring among residents of a specified geographic area (e.g., community area, ZIP code) divided by the total population for the same geographic area, for a specified time period. Below is the table which shows the unit of measure for each sub-category:

Category	Measure	Units	U.S. Baseline	U.S. Year	H.P. 2020 Target	City year	City rate or %
NATALITY	Birth rate	Per 1,000 persons	13.5	2009	.	2009	16.4
	General fertility rate	Per 1,000 females aged 15-44	66.7	2009	.	2009	67.4
	Low birth weight	Percent of live births	8.2	2007	7.8	2009	9.7
	Prenatal care beginning in first trimester	Percent of females delivering a live birth	70.8	2007	77.9	2009	76.5
	Preterm births	Percent of live births	12.7	2007	11.4	2009	10.8
	Teen birth rate	Per 1,000 females aged 15-19	39.1	2009	.	2009	57
MORTALITY	Assault (homicide)	Per 100,000 persons (age adjusted)	6.1	2007	5.5	2005-2009	15.1
	Breast cancer in females	Per 100,000 females (age adjusted)	22.9	2007	20.6	2005-2009	26.6
	Cancer (all sites)	Per 100,000 persons (age adjusted)	178.4	2007	160.6	2005-2009	193.6
	Colorectal cancer	Per 100,000 persons (age adjusted)	17	2007	14.5	2005-2009	21.4
	Diabetes-related	Per 100,000 persons (age adjusted)	73.1	2007	65.8	2005-2009	70.1
	Firearm-related	Per 100,000 persons (age adjusted)	10.2	2007	9.2	2005-2009	13.8
	Infant mortality rate	Per 1,000 live births	6.7	2006	6	2005-2009	8.1
	Lung cancer	Per 100,000 persons (age adjusted)	50.6	2007	45.5	2005-2009	50.3
	Prostate cancer in males	Per 100,000 males (age adjusted)	23.5	2007	21.2	2005-2009	34
	Stroke (cerebrovascular disease)	Per 100,000 persons (age adjusted)	42.2	2007	33.8	2005-2009	44.6

Category	Measure	Units	U.S. Baseline	U.S. Year	H.P. 2020 Target	City year	City rate or %
LEAD	Childhood blood lead level screening	Per 1,000 children aged 0-6 years	.		.	2011	419.7
	Childhood lead poisoning	Per 100	0.9	2005-2008	0	2011	0.9
INFECTIOUS	Gonorrhea in females	Per 100,000 females aged 15 to 44 years	285	2008	257	2011	660.8
	Gonorrhea in males	Per 100,000 males aged 15 to 44 years	220	2008	198	2011	598.6
	Tuberculosis	Per 100,000 persons	3.0	2010	1	2007-2011	7.4
ECONOMIC	Below poverty level	Percent of households	13.3	2007-2011	.	2007-2011	19
	Crowded housing	Percent of occupied housing units	3.2	2007-2011	.	2007-2011	4.7
	Dependency	Percent of persons aged less than 16 or more than 64 years	37	2007-2011	.	2007-2011	33.8
	No high school diploma	Percent of persons aged 25 years and older	14.6	2007-2011	.	2007-2011	19.8
	Per capita income	2011 inflation-adjusted dollars	27915	2007-2011	.	2007-2011	27940
	Unemployment	Percent of persons in labor force aged 16 years and older	8.7	2007-2011	.	2007-2011	12

Our Categories of Evaluation:

NATALITY: Birth Rate, Low Birth Rate

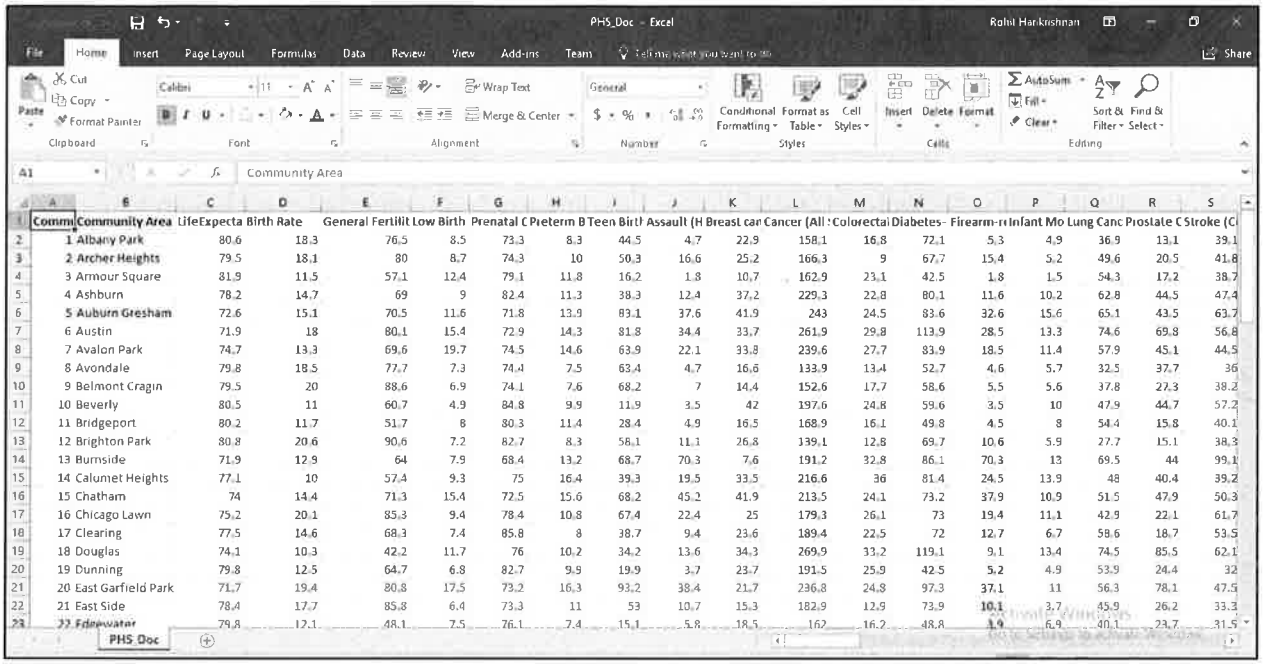
MORTALITY: Assault, Breast and Cancer and Lung Cancer

LEAD: Childhood Blood lead level screening and poisoning.

INFECTIOUS: Tuberculosis

ECONOMIC: Unemployed and Dependency.

DATASET OVERVIEW



Comm	Community Area	LifeExpecta	Birth Rate	General Fertilit	Low Birth	Prenatal C	Preterm B	Teen Birth	Assault (H	Breast can	Cancer (All	Colorectal	Diabetes-	Firearm-re	Infant Mo	Lung Canc	Prostate C	Stroke (Cl
1	Albany Park	80.6	18.3	76.5	8.5	73.3	8.3	44.5	4.7	22.9	158.1	16.8	72.1	5.3	4.9	36.9	13.1	39.1
2	Archer Heights	79.5	18.1	80	8.7	74.3	10	50.3	16.6	25.2	166.3	9	67.7	15.4	5.2	49.6	20.5	41.8
3	Armour Square	81.9	11.5	57.1	12.4	79.1	11.8	16.2	1.8	10.7	162.9	23.1	42.5	1.8	1.5	54.3	17.2	38.7
4	Ashburn	78.2	14.7	69	9	82.4	11.3	38.3	12.4	37.2	229.3	22.8	80.1	11.6	10.2	62.8	44.5	47.4
5	Auburn Gresham	72.6	15.1	70.5	11.6	71.8	13.9	83.1	37.6	41.9	243	24.5	83.6	32.6	15.6	65.1	43.5	63.7
6	Austin	71.9	18	80.1	15.4	72.9	14.3	81.8	34.4	33.7	261.9	29.8	113.9	28.5	13.3	74.6	69.8	56.8
7	Avalon Park	74.7	13.3	69.6	19.7	74.5	14.6	63.9	22.1	33.8	239.6	27.7	83.9	18.5	11.4	57.9	45.1	44.5
8	Avondale	79.8	18.5	77.7	7.3	74.4	7.5	63.4	4.7	16.6	133.9	13.4	52.7	4.6	5.7	32.5	37.7	36
9	Belmont Cragin	79.5	20	88.6	6.9	74.1	7.6	68.2	7	14.4	152.6	17.7	58.6	5.5	5.6	37.8	27.3	38.2
10	Beverly	80.5	11	60.7	4.9	84.8	9.9	11.9	3.5	42	197.6	24.8	59.6	3.5	10	47.9	44.7	57.2
11	Bridgeport	80.2	11.7	51.7	8	80.3	11.4	28.4	4.9	16.5	168.9	16.1	49.8	4.5	8	54.4	15.8	40.1
12	Brighton Park	80.8	20.6	90.6	7.2	82.7	8.3	58.1	11.1	26.8	139.1	12.8	69.7	10.6	5.9	27.7	15.1	38.3
13	Burnside	71.9	12.9	64	7.9	68.4	13.2	68.7	70.3	7.6	191.2	32.8	86.1	70.3	13	69.5	44	99.1
14	Calumet Heights	77.1	10	57.4	9.3	75	16.4	39.3	19.5	33.5	216.6	36	81.4	24.5	13.9	48	40.4	39.2
15	Chatham	74	14.4	71.3	15.4	72.5	15.6	68.2	45.2	41.9	213.5	24.1	73.2	37.9	10.9	51.5	47.9	50.3
16	Chicago Lawn	75.2	20.1	85.3	9.4	78.4	10.8	67.4	22.4	25	179.3	26.1	73	19.4	11.1	42.9	22.1	61.7
17	Clearing	77.5	14.6	68.3	7.4	85.8	8	38.7	9.4	23.6	189.4	22.5	72	12.7	6.7	58.6	18.7	53.5
18	Douglas	74.1	10.3	42.2	11.7	76	10.2	34.2	13.6	34.3	269.9	33.2	119.1	9.1	13.4	74.5	85.5	62.1
19	Dunning	79.8	12.5	64.7	6.8	82.7	9.9	19.9	3.7	23.7	191.5	25.9	42.5	5.2	4.9	53.9	24.4	32
20	East Garfield Park	71.7	19.4	80.8	17.5	73.2	16.3	93.2	38.4	21.7	236.8	24.8	97.3	37.1	11	56.3	78.1	47.5
21	East Side	78.4	17.7	85.8	6.4	73.3	11	53	10.7	15.3	182.9	12.9	73.9	10.1	3.7	45.9	26.2	33.3
22	Edgewater	79.8	12.1	48.1	7.5	76.1	7.4	15.1	5.8	18.5	162	16.2	48.8	1.9	6.9	40.1	23.7	31.5

Above is the dataset overview with the variable which we are going to use for the analysis.

METHODS AND TOOLS

CLUSTERING	**R STUDIO**
DECISION TREES	**SAS ENTERPRISE MINER**
LINEAR REGRESSION	**SAS STUDIO**

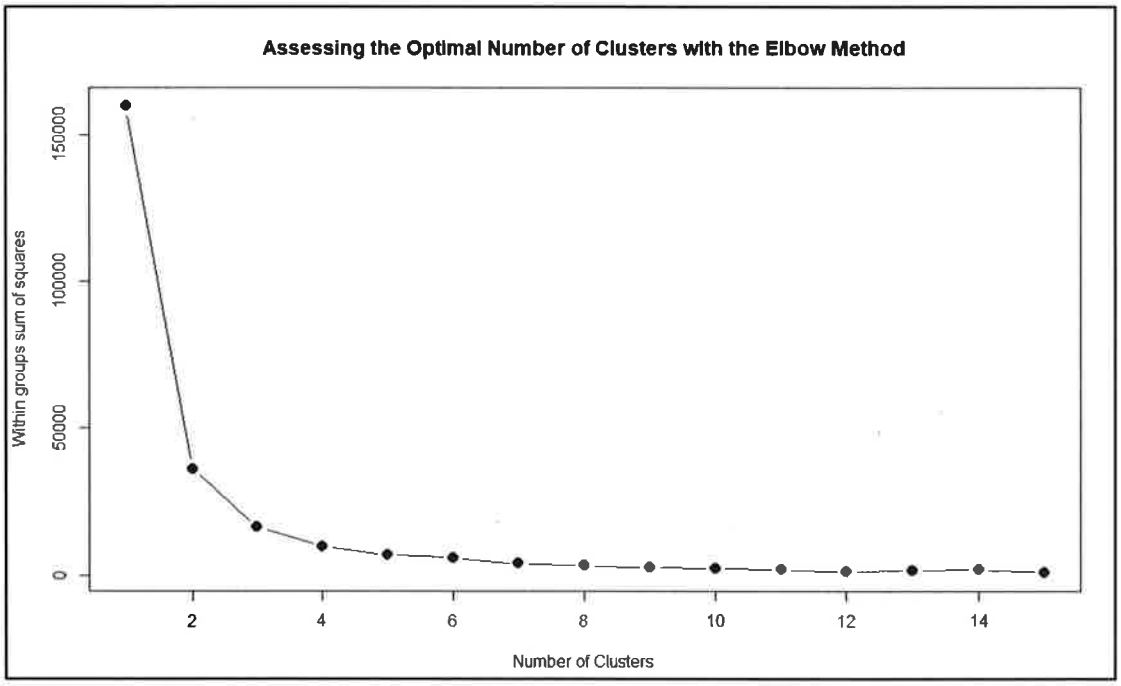
CLUSTERING

To get a deeper understanding of our dataset we used the non-inferential method Clustering.

We used K means clustering in R-studio to group the communities in Chicago. We categorized the communities based on the percentage of cancer rate, per capita income and their dependency on Life Expectancy Rate.

Cancer Vs Life expectancy

We thought of considering a health issue as a factor in our analysis. After running the analysis in R-studio and taking the variables as Life Expectancy and Cancer rate with K value = 3, we could obtain the WSS plot given below.



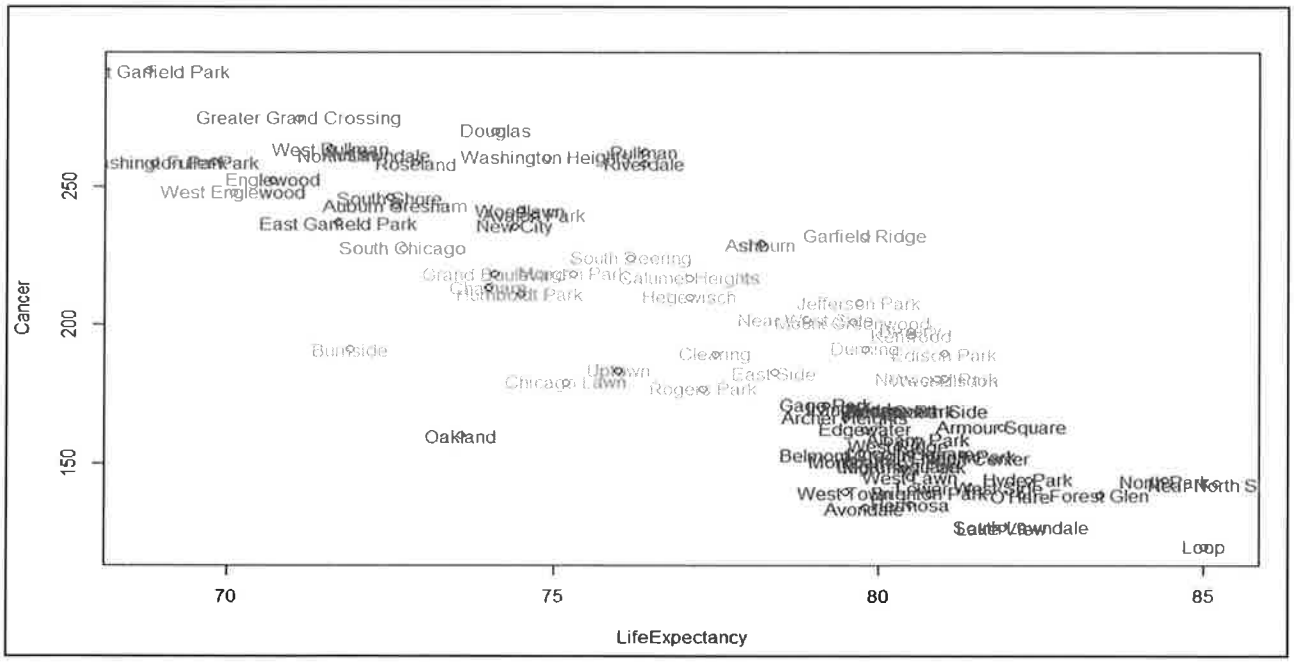
Let us take a look at the graph and understand the optimal number of clusters we have used in our analysis. From the elbow method, the optimal number we have obtained is 3. As you can see after the number of cluster =3, the curve is having a flat slope.

With this we ran our clustering model and could obtain the results given below:

```
K-means clustering with 3 clusters of sizes 23, 22, 32

Cluster means:
  LifeExpectancy   Cancer
1       73.07826 252.6000
2       77.56818 198.3000
3       80.86875 149.5937

Clustering vector:
 [1] 3 3 3 1 1 1 1 3 3 2 3 3 2 2 2 2 2 1 2 1 2 3 2 1 3 1 3 1 2 1 2 3 2 3 3 2 2 3 3 3 3 3 3
[44] 3 3 2 2 3 3 2 1 3 1 3 2 3 3 3 1 1 2 1 1 2 3 1 2 1 1 2 1 1 3 1 3 3 1

Within cluster sum of squares by cluster:
[1] 5670.399 4747.148 5950.107
 (between_SS / total_SS =  89.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

In the above analysis we can see the 3 clusters are divided of sizes 23, 22 and 32.

This analysis gave us the Cluster means of each groups and in which cluster does the observations belong to. Always a visual plot could explain the above given analysis in a better way. So by plotting the variables, we could get the below given plot.



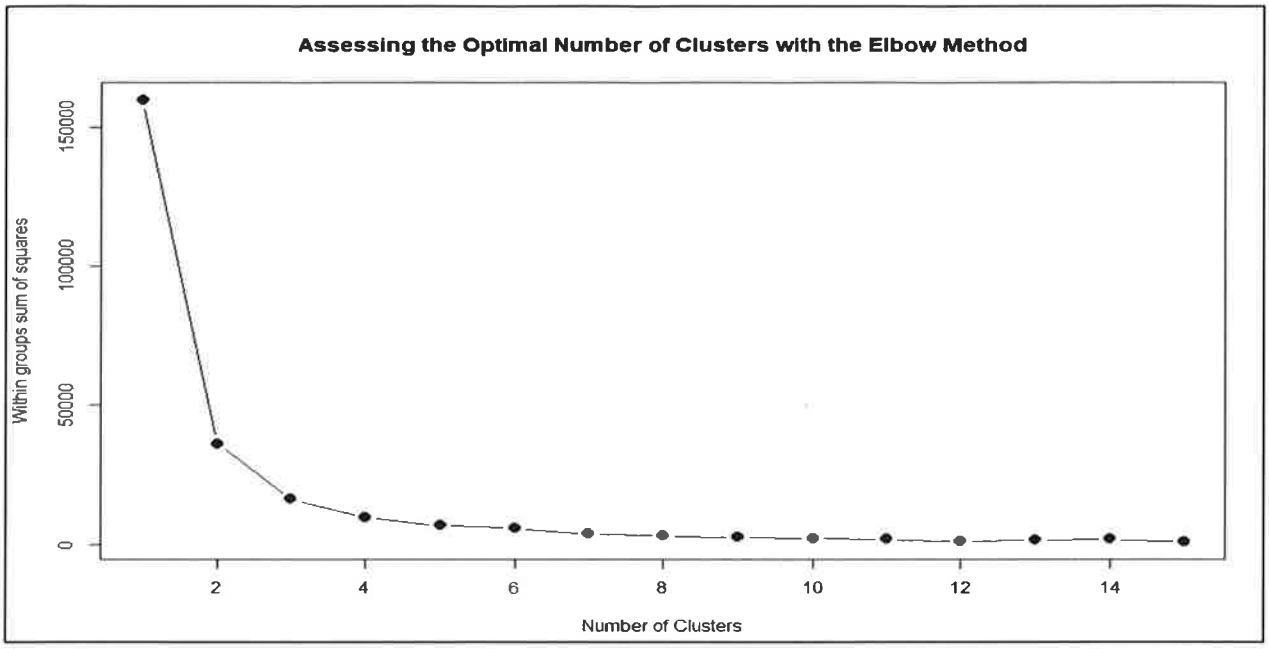
Interpretations

- Red observations in the graph denotes the areas/ communities in Chicago with very low Life Expectancy rate. Ex: South Chicago, Douglas etc.
- From our analysis, though we are only looking at a single health issue as a factor. We can interpret that these communities are dealing with a lot of other health related issues that resulted in poor life expectancy rate. This could be because of the poor facilities and living conditions.
- Blue observations are the communities with high life expectancy rate. Ex: Loop, Forest Glen etc.
- These are the posh communities in Chicago with very high life expectancy rate, since they have better medical facilities and living condition.

Now to see if these group of communities have better socio-economic condition. We thought of selecting an economic factor in our dataset as our clustering variable and its dependency on Life Expectancy.

Per Capita Income vs Life Expectancy

We ran the k-means clustering analysis in R studio with per capita income and life expectancy rate as our variables. Again, by looking at wss plot, we could understand the optimal number of clusters to be included in our analysis.



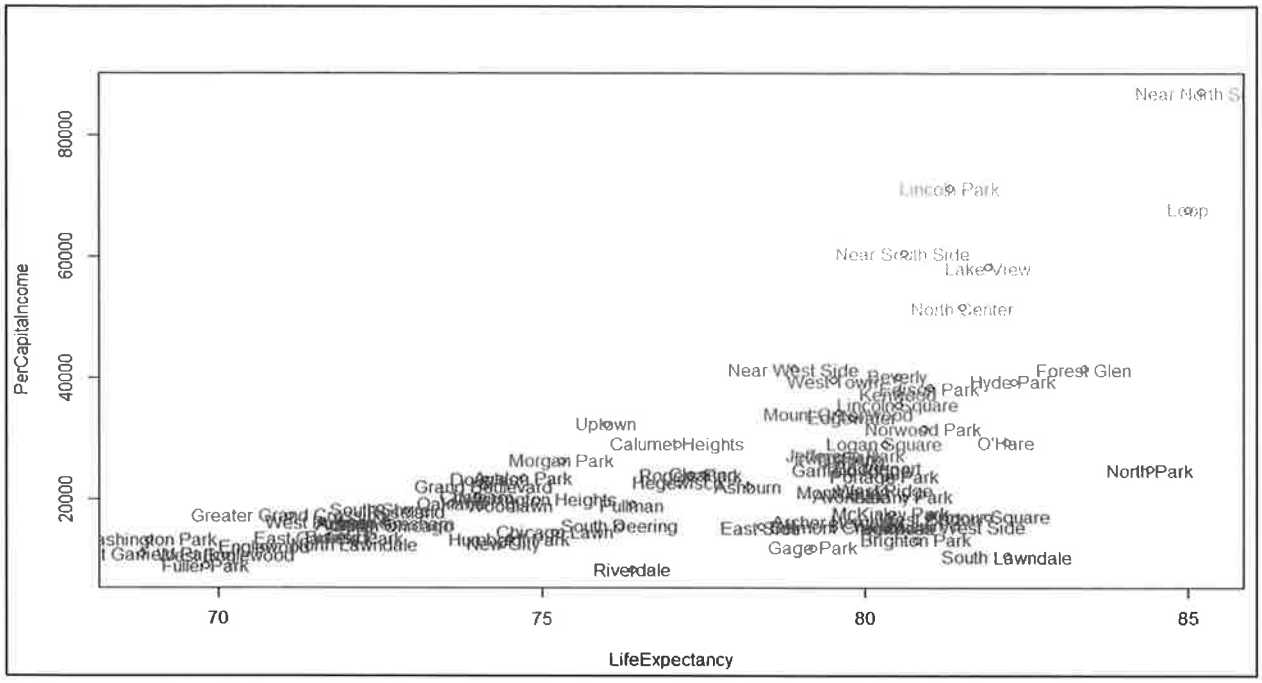
From the elbow method, the optimal number we have obtained is again 3. As you can see after the number of cluster =3, the curve is having a flat slope.

After running our clustering model, we could obtain the results given below.

```
K-means clustering with 3 clusters of sizes 19, 6, 52

Cluster means:
  LifeExpectancy      PCI
1       79.83684 33622.68
2       82.58333 66116.67
3       76.20577 17263.23

Clustering vector:
 [1] 3 3 3 3 3 3 3 3 3 1 3 3 3 1 3 3 3 3 1 3 3 1 1 3 1 3 3 3 3 3 3 3 3 1 1 1 1 2 2 1 1 2 3
[44] 3 3 1 1 2 2 1 3 2 3 3 1 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 1 3

Within cluster sum of squares by cluster:
[1]  536955828  776453335 1010992385
 (between_SS / total_SS =  86.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
```

In the above analysis we can see the 3 clusters are divided of sizes 19, 6 and 52.

This analysis gave us the Cluster means of each groups and in which cluster does the observations belong to. So by plotting the variables, we could get the below given plot.



Interpretations:

- Blue observations in the graph denotes the communities in Chicago with very low Life Expectancy rate. Ex: South Chicago, Douglas etc.
- From our analysis, we are looking at the income of people in these communities. We can interpret that poor economic state of the people in these areas could be one of the reasons behind low life expectancy rate.
- Green observations are the communities with high life expectancy rate. Ex: Loop, Forest Glen etc.
- Again, these posh communities in Chicago with very high life expectancy rate, can afford their medical bills and medical treatment. And that explains their high life expectancy rate.

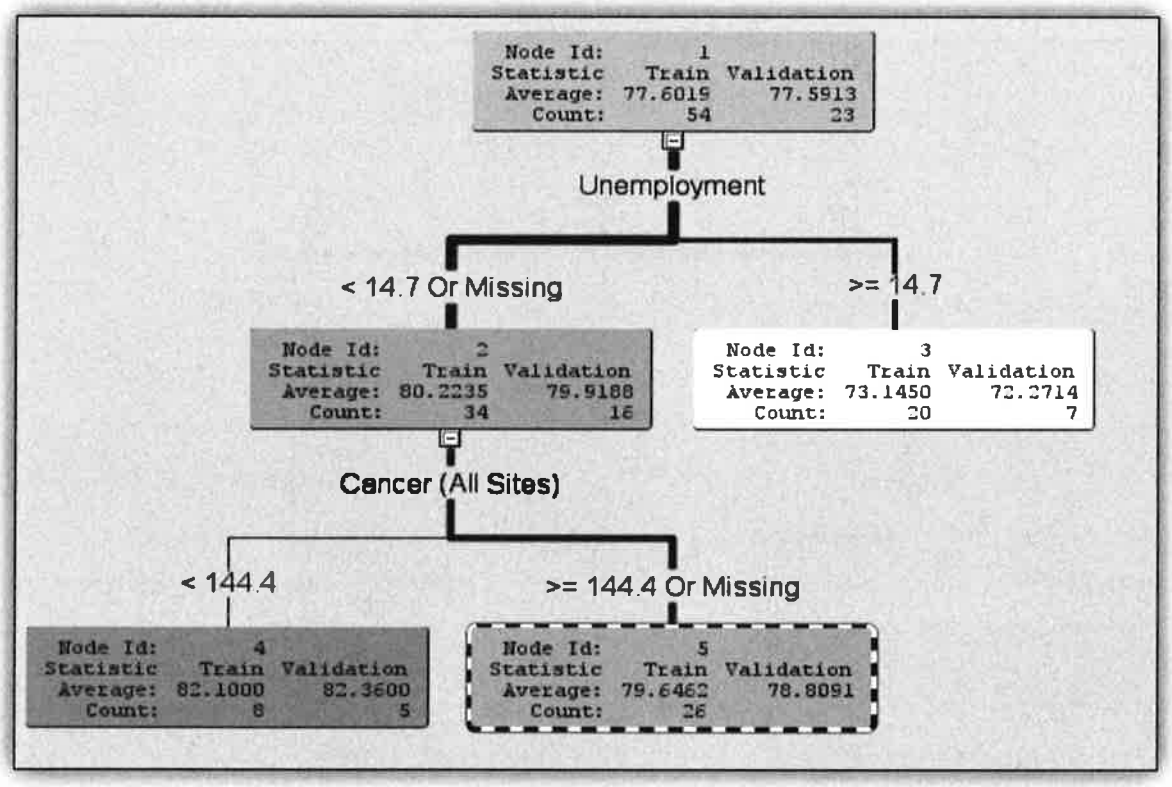
The comparison of both the visual plots obtained gave us a better idea on how economic factors and health related issues in the Chicago are related to each other. With these results as our, we started our descriptive analysis that follows.

DECISION TREE

A **Decision Tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

We used **Decision Trees** in **SAS Enterprise Miner**, to classify the data based on Life Expectancy for different communities of Chicago. We input all the variables in our dataset taking Life Expectancy as our target variable. Decision tree allows the addition of new possible outcomes and helps determine worst, best and expected values for different scenarios.

By taking all the variables in Decision Tree, we obtained the following in SAS EM:



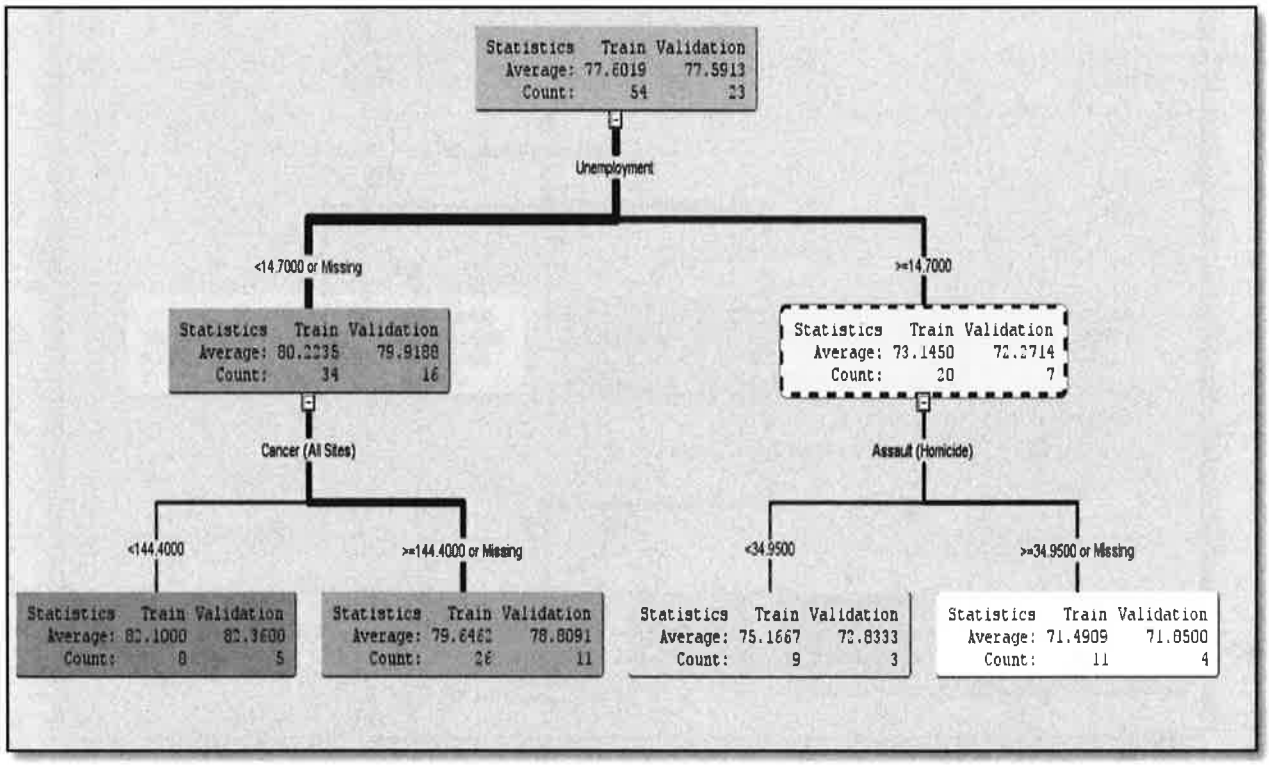
Node 0- Life Expectancy split at 30-70

Node 1- Unemployment split at >= 14.7 and <14.7 (Node1 was split on the basis of Information Gain i.e. the attribute with the highest information gain is used first)

Node 2- Cancer (9All Sites) split at >= 144.4 and <144.4

So what we can conclude from this is that in counties where the unemployment rate is less than 14.7, the average age of the person increases than the node value (77.6 years). Which means that more people employed, healthier lifestyle, more the number of years they live. Similarly, when we see the number of cases of cancer, we can say that in counties where the cancer cases are less than 144.4, the average life of a person increases to 82.1 years as compared to the root value.

After this, we ran the interactive decision tree and got the following results:



We wanted to split the right side of the tree (unemployment >= 14.70) and see what is the next best

attribute which we can use to split that node. We saw that Assault (Homicide) has the next highest Information Gain value. So the node was split into two: Assault <34.95 and >= 34.95.

Node 0- Life Expectancy split at 30-70

Node 1- Unemployment split at >= 14.7 and <14.7

Node 2- Cancer (All Sites) split at >= 144.4 and <144.4

Node 2- Assault (Homicide) split at >= 34.95 and <34.95

So what we can conclude from this is that in counties where there are more number of assault (homicide cases), the average age of the person decreases than the node value (77.6 years).

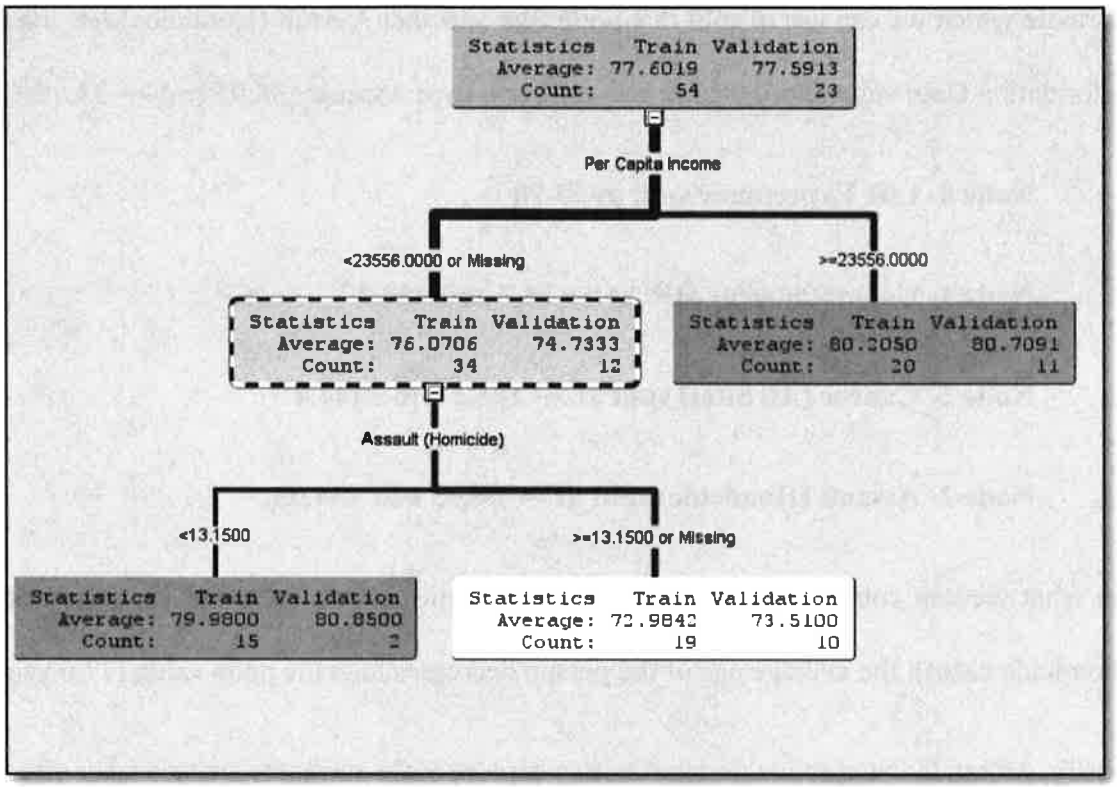
Lastly, we ran the interactive decision tree to explore more attributes and see what effect they have. So we chose Per Capita Income as our choice of attribute to split the Node 0.

We saw that we got two nodes: Per Capita Income >= 23556.0 and Per Capita Income <23556.0. Again based on the Information gain we got our next attribute as Assault (Homicide).

Node 0- Life Expectancy split at 30-70

Node 1- Per Capita Income split at >= 23556.0 and <23556.0

Node 2- Assault (Homicide) split at >= 13.15 and <13.15



So we can conclude that counties in which the per capita income is high (>= 23556.0), the average life of a person is more (80.2 years). Which means that more income, healthier lifestyle, better facilities and thus more number of years they live.

LINEAR REGRESSION

We all know that regression is one of the supervised learning method to predict the relationship between the target variable and independent variable.

As in our dataset we have all continuous variable, so we had taken multiple linear regression to find out the relationship between variables. Our dataset contains five category race, housing, natality, mortality and economic.

In our analysis, our target variable is life expectancy which is the average age of the person living and our independent variable are the various factors of Natality and Mortality.

Life Expectancy VS Race:

We have data of 3 Race i.e. Whites, African American and Asian. So, when we ran the regression to these factors we got the following results.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	453.27739	151.09246	13.49	<.0001
Error	73	817.37248	11.19688		
Corrected Total	76	1270.64987			

Root MSE	3.34617	R-Square	0.3567
Dependent Mean	77.59870	Adj R-Sq	0.3303
Coeff Var	4.31215		

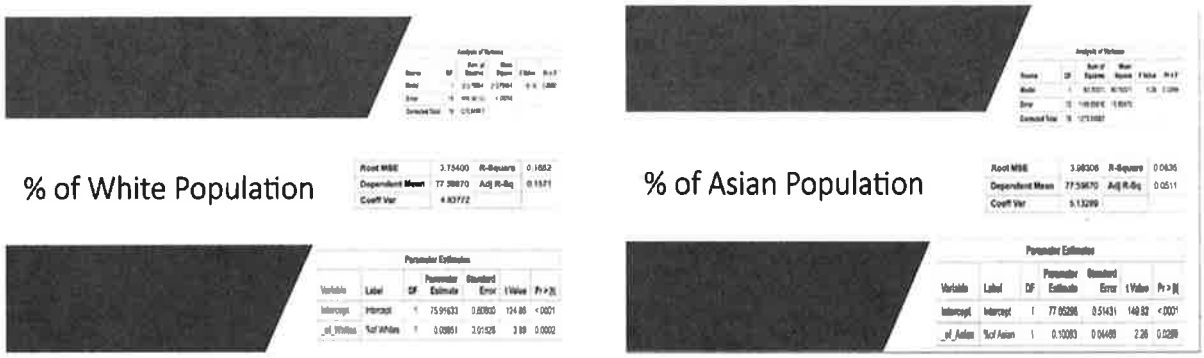
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	79.88069	1.12134	71.24	<.0001
_of_Whites	%of Whites	1	-0.00239	0.01921	-0.12	0.9014
_of_African_Amerian	%of African Amerian	1	-0.05964	0.01392	-4.28	<.0001
_of_Asian	%of Asian	1	0.02254	0.03992	0.56	0.5740

If we see at the above results we can say that the model is significant from the F- Statistic. As we had taken multiple linear regression we will look at the adjusted R-square and we can say that 33.03% of the variance in life expectancy is been explained by our various factors of Race like

Whites, African American and Asian. Now, if we look at the p value of each factor we can see that only African American is the significant variable as the p value is less that the significant level (0.05). So we can say that African American has significant effect on Life Expectancy. So, to avoid the multi collinearity effect we will again run the simple regression on Whites and Asians.

Simple Regression on Whites and Asian:



From the above analysis we could say that Whites and Asian do have significant effects on the Life expectancy but looking at the Adjusted-R square we can say that Asians have less impact on life expectancy that is about 5% while on the other hand Whites have more impact on life expectancy to 15%.

Life Expectancy VS Housing Factor:

We have data of 5 Housing factor Renter Occupied, Occupied Housing, Vacant Housing, Owned with a loan and owned without loan. So, when we ran the regression to these factors we got the following results.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	438.42853	109.60713	9.48	<.0001
Error	72	832.22134	11.55863		
Corrected Total	76	1270.64987			

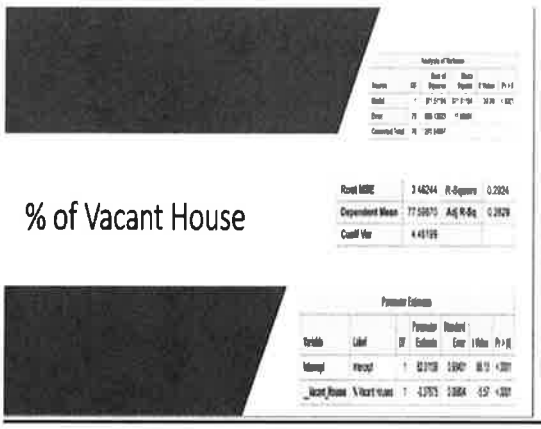
Root MSE	3.39960	R-Square	0.3450
Dependent Mean	77.59870	Adj R-Sq	0.3087
Coeff Var	4.38126		

Parameter Estimates

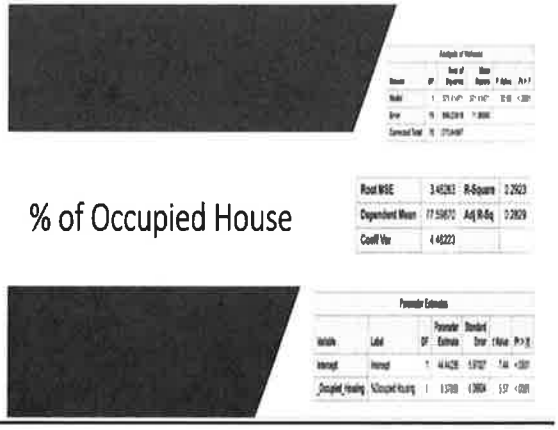
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	17027	13703	1.24	0.2161
_Renter_Occupied	%Renter Occupied	B	0.06099	0.09198	0.66	0.5094
_Occupied_Housing	%Occupied Housing	B	-169.46174	137.02393	-1.24	0.2202
__Vacant_Houses	% Vacant Houses	B	-169.92558	137.03516	-1.24	0.2190
_Owned_with_a_loan	%Owned with a loan	B	0.00413	0.13437	0.03	0.9756
_Owned_free__clear	%Owned free &clear	0	0	.	.	.

If we see at the above results we can say that the model is significant from the F- Statistic. As we had taken multiple linear regression we will look at the adjusted R-square and we can say that 30.87% of the variance in life expectancy is been explained by our various factors of Housing like Renter Occupied, Occupied Housing, Vacant Housing, and Owned with a loan and owned without loan. Now, if we look at the p value of each factor we can see that none of the variable are significant as the p value is not less than that the significant level (0.05). So, to avoid the multi collinearity effect we will again run the simple regression on all the variables.

Simple Regression on all Housing Factors:



% of Vacant House



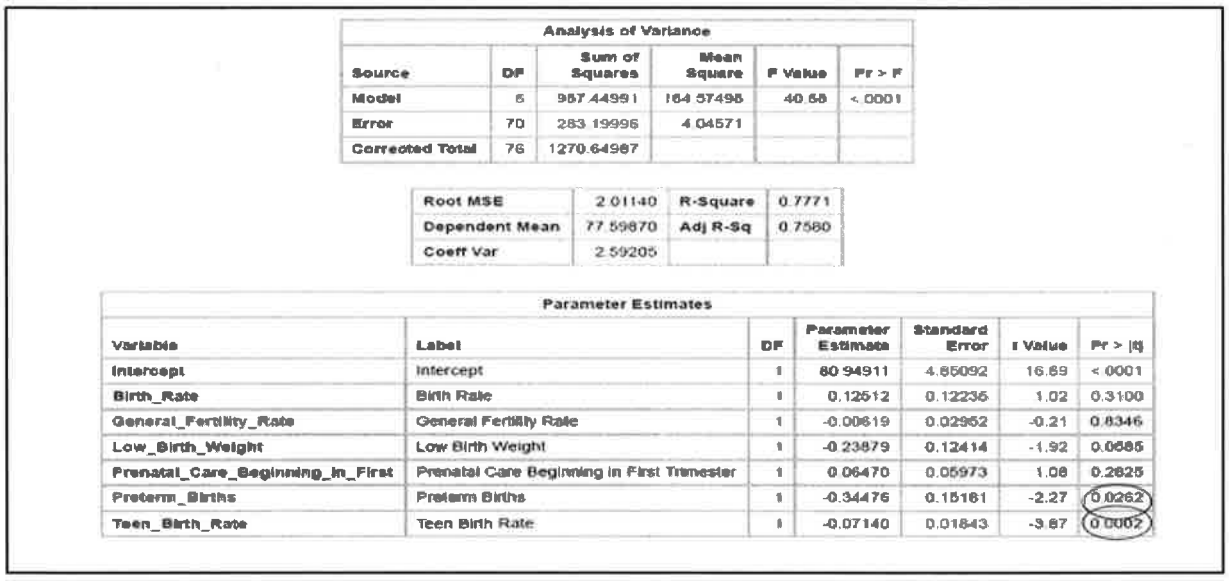
% of Occupied House

After running regression on each variable, we could see that only vacant and occupied house has significant effect on the life expectancy. Looking at the Adjusted-R square we can say that both the variables have almost same impact on life expectancy to 28%. But looking at the coefficient we can say that Occupied house has positive relationship and Vacant Housing has negative relationship with Life Expectancy, which logically also make sense as better facilities are seen in where people are more residing.

Life Expectancy VS Natality and Mortality:

Basically, Natality is the birth rate of small children. So, in natality we have analyzed various factors like Birth rate, Fertility rate, low birth weight, preterm birth, and teen birth rate affecting the life expectancy rate.

So, when we ran the regression to these factors we got the following results.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	957.44991	164.57498	40.68	<.0001
Error	70	283.19996	4.04571		
Corrected Total	76	1270.64987			

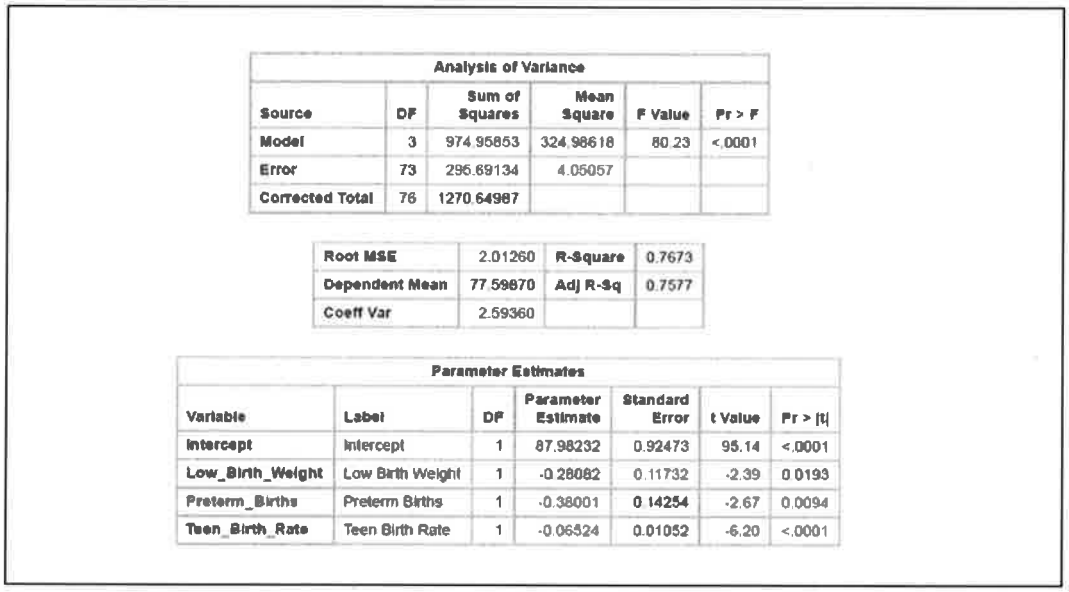
Root MSE	2.01140	R-Square	0.7771
Dependent Mean	77.59870	Adj R-Sq	0.7580
Coeff Var	2.59205		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	80.94911	4.85092	16.69	<.0001
Birth_Rate	Birth Rate	1	0.12512	0.12235	1.02	0.3100
General_Fertility_Rate	General Fertility Rate	1	-0.00619	0.02952	-0.21	0.8346
Low_Birth_Weight	Low Birth Weight	1	-0.23879	0.12414	-1.92	0.0585
Prenatal_Care_Beginning_In_First	Prenatal Care Beginning in First Trimester	1	0.06470	0.05973	1.08	0.2825
Preterm_Births	Preterm Births	1	-0.34476	0.15161	-2.27	0.0262
Teen_Birth_Rate	Teen Birth Rate	1	-0.07140	0.01843	-3.87	0.0002

If we see at the above results we can say that the model is significant. As we had taken multiple linear regression we will look at the adjusted R-square and we can say that 75.80 % of the variance in life expectancy is been explained by our various factors like birth rate, fertility rate, low

birth weight, preterm birth, teen birth rate. Now, if we look at the p-value of each factor we can see that pre-term birth and teen birth rate are the significant variables as the p value are less that the significant level (0.05). So, to avoid the multi collinearity effect we will again run the regression on just the significant variables i.e. Low birth weight, preterm birth, teen birth rate.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	974.95853	324.98618	80.23	<.0001
Error	73	295.69134	4.05057		
Corrected Total	76	1270.64987			

Root MSE	2.01260	R-Square	0.7673
Dependent Mean	77.59870	Adj R-Sq	0.7577
Coeff Var	2.59360		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	87.98232	0.92473	95.14	<.0001
Low_Birth_Weight	Low Birth Weight	1	-0.28082	0.11732	-2.39	0.0193
Preterm_Births	Preterm Births	1	-0.38001	0.14254	-2.67	0.0094
Teen_Birth_Rate	Teen Birth Rate	1	-0.06524	0.01052	-6.20	<.0001

So again, looking at the above results model is significant and 75.77% of the variance is been explained by the factors.

Linear regression equation:

Life expectancy = 87.98 -0.28(Low birth weight) – 0.38(preterm birth) – 0.065(teen birth)

Now let's say we want to interpret how various factors affect the Life expectancy.

It is obvious that as the coefficient of the variable are negative, life expectancy is going to decrease.

We will assume preterm birth and teen birth as constant variable, so we can say that per unit increase in low birth weight, life expectancy is going to decrease by 0.28 units. Similarly, we can interpret other variables.

Now, let's do similar analysis for **Mortality** category. Mortality is basically the causes of death rate.

So, from our dataset we have various factors like Assault, Cancer, Diabetes, and Stroke. So, let's analyze which factors are causing more effect on life expectancy rate.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1124.01043	374.67014	186.52	<.0001
Error	73	146.63944	2.00876		
Corrected Total	76	1270.64987			

Root MSE	1.41731	R-Square	0.8846
Dependent Mean	77.59870	Adj R-Sq	0.8799
Coeff Var	1.82646		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	88.46035	0.90830	97.39	<.0001
Assault__Homicide_	Assault (Homicide)	1	-0.11694	0.01565	-7.47	<.0001
Diabetes_related	Diabetes-related	1	-0.05450	0.01169	-4.66	<.0001
Cancer__All_Sites_	Cancer (All Sites)	1	-0.02485	0.00600	-4.14	<.0001

The results above are found after running regression on various factors of mortality.

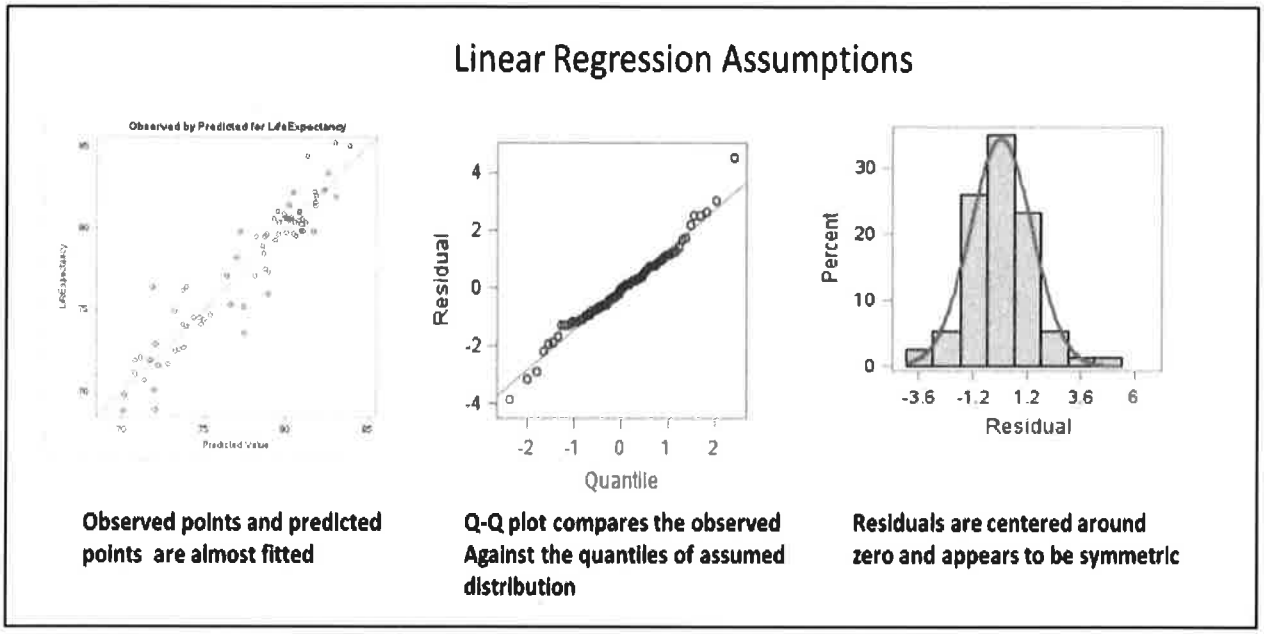
We found that model is significant. 87.99% of the variance in life expectancy is been explained by various factors of mortality.

Our linear equations:

Life Expectancy = 88.46 - 0.11(Assault) – 0.05(Diabetes) – 0.02(Cancer)

Here also we can see that all the coefficient has negative values. So as any of the mortality factors increases life expectancy will be decreased.

So, if we assume Diabetes and Cancer as constant variable we can interpret as per unit increase in assault, life expectancy will decrease by 0.11 units.



Looking at the above graphs we can say that all the assumptions we took into consideration while running regression analysis holds true.

Life Expectancy VS Economic Factors:

Similarly, we will now analyze for **economic factors** category. So, from our dataset we have various factors like Unemployment, Per Capita Income, Below Poverty Level, Crowded Housing, Dependency and No high School Diploma. So, let's analyze which factors are causing more effect on life expectancy rate.

So, when we ran the regression to these factors we got the following results.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	968.34295	161.39049	37.37	<.0001
Error	70	302.30692	4.31867		
Corrected Total	76	1270.64987			

Root MSE	2.07814	R-Square	0.7621
Dependent Mean	77.59870	Adj R-Sq	0.7417
Coeff Var	2.67806		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	76.62296	2.86164	26.78	<.0001
Unemployment	Unemployment	1	-0.34170	0.06518	-5.24	<.0001
Per_Capita_Income	Per Capita Income	1	0.00012557	0.00003385	3.71	0.0004
Below_Poverty_Level	Below Poverty Level	1	-0.05989	0.03480	-1.72	0.0896
Crowded_Housing	Crowded Housing	1	-0.21392	0.17050	-1.25	0.2138
Dependency	Dependency	1	0.01098	0.05476	0.20	0.8417
No_High_School_Diploma	No High School Diploma	1	0.19642	0.05723	3.43	0.0010

If we see at the above results we can say that the model is significant. As we had taken multiple linear regression we will look at the adjusted R-square and we can say that 74.1 % of the variance in life expectancy is been explained by our various factors Unemployment, poverty level etc. Now, if we look at the p-value of each factor we can see that below poverty level, crowded housing and dependency are not significant variables as the p-value are more that the significant level (0.05). So, to avoid the multi collinearity effect we will again run the regression on just the significant variables i.e. Unemployment, Per capita Income and No high school diploma.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	943.11970	314.37323	70.07	<.0001
Error	73	327.53017	4.48671		
Corrected Total	76	1270.64987			

Root MSE	2.11819	R-Square	0.7422
Dependent Mean	77.59870	Adj R-Sq	0.7316
Coeff Var	2.72967		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	77.60586	1.57512	49.27	<.0001
Per_Capita_Income	Per Capita Income	1	0.00011106	0.00002801	3.97	0.0002
Unemployment	Unemployment	1	-0.39913	0.04448	-8.97	<.0001
No_High_School_Diploma	No High School Diploma	1	0.11643	0.02836	4.10	0.0001

We found that model is significant. 73.16 % of the variance in life expectancy is been explained by various economic factors.

Our linear equations:

Life Expectancy = 77.6 + 0.0001 (Per Capita Income) – 0.039 (Unemployment) + 0.11 (No high school diploma)

Here, we can see that one coefficient has negative values. So as unemployment increases life expectancy decreases.

So, if we assume other variables as constant variable we can interpret as per unit increase in unemployment rate, life expectancy will decrease by 0.039 units.

CONCLUSION

By **Clustering**, we can see that grouping the communities based on health and economic factors, we could identify the less developed areas in Chicago.

By **Decision Tree analysis**, we can see that indicators like unemployment, cancer and per capita income are good predictors for life expectancy.

By **Regression analysis**, we can see that Housing, Race, Natality, Mortality and Economic factors have significant impact on Life Expectancy of community areas in Chicago.

RECOMMENDATION:

Focusing on the less developed areas and improving their facilities, life expectancy could be increased. We can see that how life expectancy variable is affected by every indicator and how each community area can reach an average baseline value for a need to have the life expectancy variable on a targeted baseline value (Healthy Chicago 2020).