



HOSPITAL DISTANCE IN DALLAS BY INCOME

- **Access to hospitals can save lives.**
- **The physical distance patients have to travel to get to a hospital is crucially important**
- **In the United States, the quality of healthcare received often depends on the socioeconomic status of the patient.**
- **Hospitals have an incentive to locate in high income areas**

Introduction and Business Problem

Background

- I will explore whether there is any link between the physical distance one must travel to get to a hospital and the wealth of the person requiring treatment
- For this exercise, I will focus on hospitals in Dallas, Texas.
- I will explore whether any link exists between people's income and the physical distance they must travel to get to local hospitals

Introduction and Business Problem

Problem

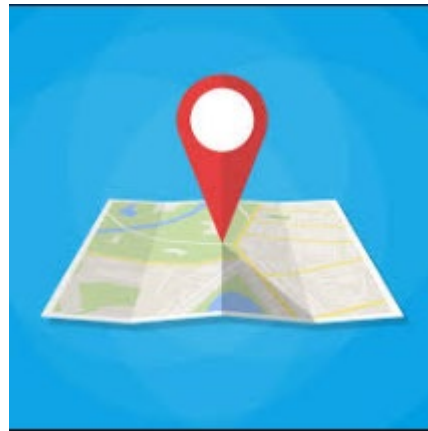
- **This information could be of interest to policymakers and hospital administrators who are looking to understand the scope of the issue and make decisions in the best interest of the public.**



Introduction and Business Problem

Interest

- I needed two primary pieces of information to complete this analysis: income and hospital locations.
- Then, we need a way to geo-locate this data



Data

Feature Selection

- The US Census publishes a report of income as reported in tax filings. Kaggle has taken this data and compiled it into a user-friendly .csv format. This data can be accessed with a Kaggle account at:
- <https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations>
- The locations of hospitals were taken using Foursquare's API.

Data

Data Sources

- Both datasets required cleaning and, eventually, merging.
- This was done using PANDAS software.
- I filtered out extraneous columns, as the datasets were quite large.
- I selected the subset of data that applied to the Dallas Texas region. I defined and removed "NaN" values (records in which no income data was available).

Data

Data Cleaning

- **First I imported PANDAS and cleaned the data.**
 - **This included dropping unnecessary columns, dropping rows with "NaN" values, and changing all data types to the correct format for analysis**
- **The following data table header summarizes the data from the Kaggle income table**

	City	Lat	Lon	Median
0	Aledo	32.696186	-97.663302	120366.0
1	Bridgeport	33.209319	-97.772440	41754.0
2	Carrollton	32.988360	-96.899770	79305.0
3	Corsicana	32.081960	-96.467579	38775.0
4	Keene	32.355614	-97.292037	50201.0

Methodology

Cleaning

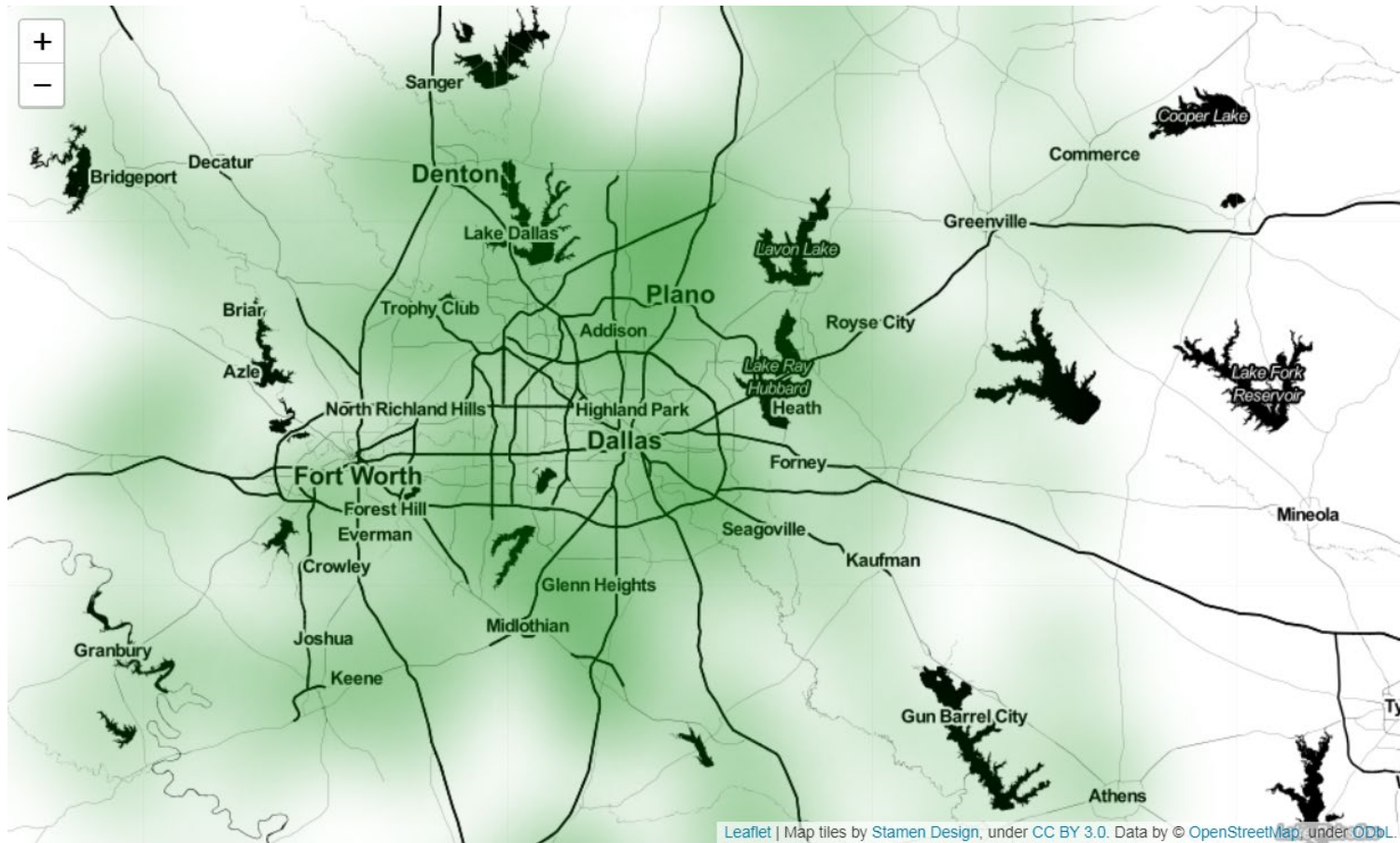
- I then normalized the income data

Methodology

Normalizing Income Data

	City	Lat	Lon	Median
0	Aledo	32.696186	-97.663302	0.152540
1	Bridgeport	33.209319	-97.772440	-0.648920
2	Carrollton	32.988360	-96.899770	-0.266082
3	Corsicana	32.081960	-96.467579	-0.679291
4	Keene	32.355614	-97.292037	-0.562801

- Using Folium and the Folium heatmap plugin, I created a heatmap showing the dispersion of median incomes around the Dallas metro area



Methodology

Mapping Median Income

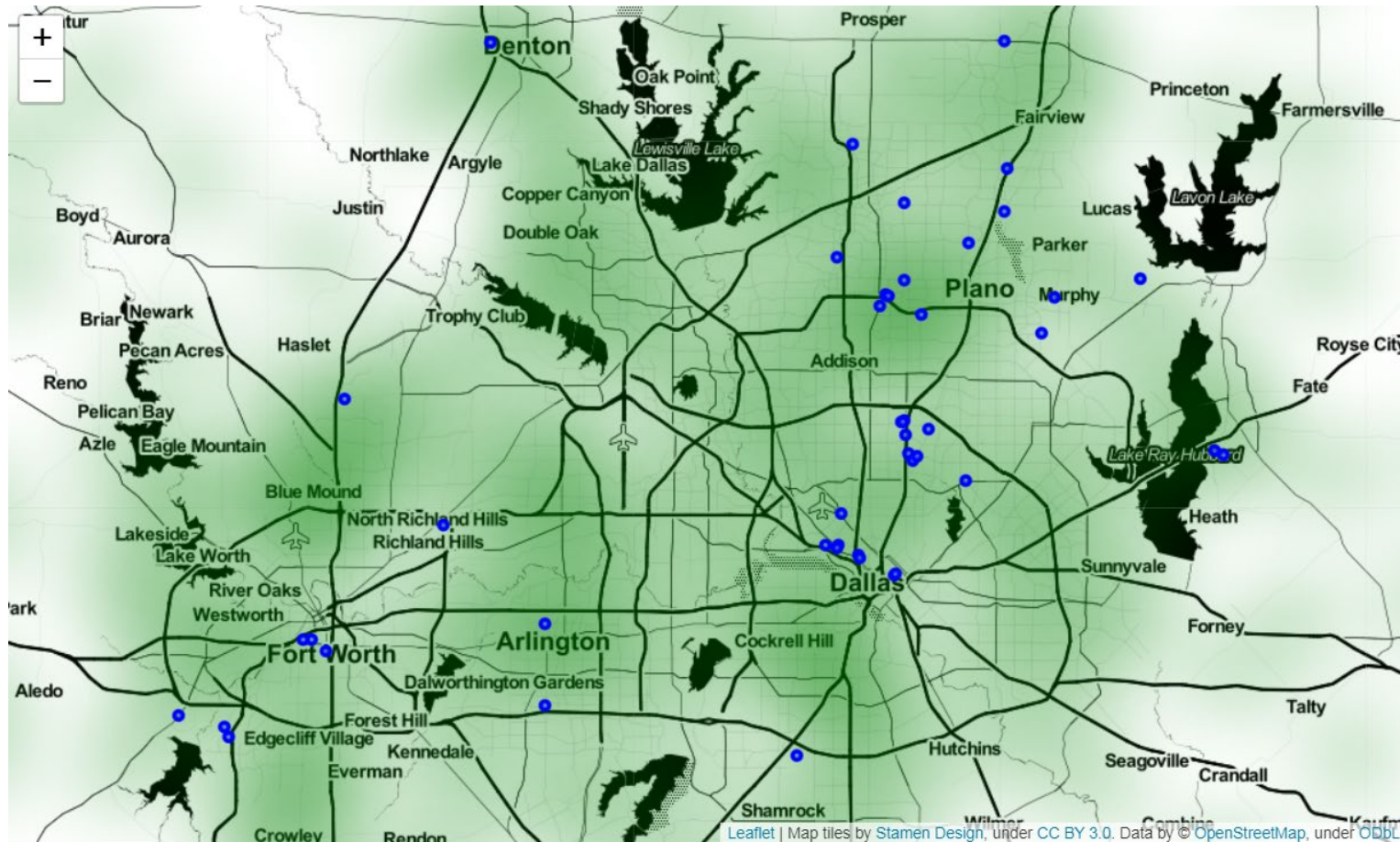
- Next, I imported the locations of Dallas-area hospitals using the Foursquare API.
- I imported the data to a PANDAS dataframe.
- I then cleaned the data, dropping unneeded columns. The following table header shows a sample of the data extracted

	name	location.lat	location.lng
0	Medical City Children's Hospital	32.911987	-96.774639
1	Texas Health Presbyterian Hospital Dallas	32.881813	-96.763005
2	Parkland Health & Hospital System	32.813022	-96.835212
3	Green Oaks Hospital	32.913076	-96.772017
4	JPS Hospital	32.727128	-97.326021

Methodology

Importing and Cleaning Hospital Data from Foursquare

- Next, I overlaid the hospital locations on top of our income heatmap.
- A visual inspection shows that hospitals appear to be clustered around areas of high income.



Methodology

Hospital Locations Overlaying Income Map

- Next, I combined the two dataframes.
- I created a new column in the income dataframe to serve as a record number.

	Cit_City	Cit_Lat	Cit_Lon	Median	Record Number	Hospital	Hos_Lat	Hos_Lon
0	Aledo	32.696186	-97.663302	0.15254	0	Medical City Children's Hospital	32.911987	-96.774639
1	Aledo	32.696186	-97.663302	0.15254	0	Texas Health Presbyterian Hospital Dallas	32.881813	-96.763005
2	Aledo	32.696186	-97.663302	0.15254	0	Parkland Health & Hospital System	32.813022	-96.835212
3	Aledo	32.696186	-97.663302	0.15254	0	Green Oaks Hospital	32.913076	-96.772017
4	Aledo	32.696186	-97.663302	0.15254	0	JPS Hospital	32.727128	-97.326021

Methodology

Combining Dataframes

- Next, I use Python's "geopy.distance" library to calculate the distance between each respondant and each hospital.
- The header of the resulting table is shown below.

	Cit_City	Cit_Lat	Cit_Lon	Median	Record Number	Hospital	Hos_Lat	Hos_Lon	Distance
0	Aledo	32.696186	-97.663302	0.15254	0	Medical City Children's Hospital	32.911987	-96.774639	53.812973
1	Aledo	32.696186	-97.663302	0.15254	0	Texas Health Presbyterian Hospital Dallas	32.881813	-96.763005	53.941990
2	Aledo	32.696186	-97.663302	0.15254	0	Parkland Health & Hospital System	32.813022	-96.835212	48.886519
3	Aledo	32.696186	-97.663302	0.15254	0	Green Oaks Hospital	32.913076	-96.772017	53.980066
4	Aledo	32.696186	-97.663302	0.15254	0	JPS Hospital	32.727128	-97.326021	19.764394

Methodology

Calculating Hospital Distance

- Finally, I normalized the calculated distances.

	Cit_City	Cit_Lat	Cit_Lon	Median	Record Number	Hospital	Hos_Lat	Hos_Lon	Distance
0	Aledo	32.696186	-97.663302	0.15254	0	Medical City Children's Hospital	32.911987	-96.774639	0.471591
1	Aledo	32.696186	-97.663302	0.15254	0	Texas Health Presbyterian Hospital Dallas	32.881813	-96.763005	0.472721
2	Aledo	32.696186	-97.663302	0.15254	0	Parkland Health & Hospital System	32.813022	-96.835212	0.428418
3	Aledo	32.696186	-97.663302	0.15254	0	Green Oaks Hospital	32.913076	-96.772017	0.473055
4	Aledo	32.696186	-97.663302	0.15254	0	JPS Hospital	32.727128	-97.326021	0.173205

Methodology

Normalizing Data

- **Applying a statistical analysis to the distances and median incomes, we find that in fact, there is a relationship.**
- **Lower median income is significantly correlated with hospital distance.**

Results

Statistical Analysis

- We see that using least squares, the analysis returns a standard error of 0.002
 - This implies statistical significance.
- The coefficient value is negative, implying that the lower one's income, the farther they live from a hospital

OLS Regression Results

Dep. Variable:	Distance	R-squared (uncentered):	0.006
Model:	OLS	Adj. R-squared (uncentered):	0.006
Method:	Least Squares	F-statistic:	164.0
Date:	Thu, 26 Mar 2020	Prob (F-statistic):	1.86e-37
Time:	21:17:27	Log-Likelihood:	-5853.0
No. Observations:	29424	AIC:	1.171e+04
Df Residuals:	29423	BIC:	1.172e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Median	-0.0221	0.002	-12.808	0.000	-0.025	-0.019
Omnibus:	3295.185	Durbin-Watson:	0.203			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4562.191			
Skew:	0.898	Prob(JB):	0.00			
Kurtosis:	3.706	Cond. No.	1.00			

Results

Statistical Analysis

- On the other hand, we calculated a rather low R-squared value of 0.006.
- This implies that although income may be one factor, other factors also influence hospital location.

OLS Regression Results

Dep. Variable:	Distance	R-squared (uncentered):	0.006
Model:	OLS	Adj. R-squared (uncentered):	0.006
Method:	Least Squares	F-statistic:	164.0
Date:	Thu, 26 Mar 2020	Prob (F-statistic):	1.86e-37
Time:	21:17:27	Log-Likelihood:	-5853.0
No. Observations:	29424	AIC:	1.171e+04
Df Residuals:	29423	BIC:	1.172e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Median	-0.0221	0.002	-12.808	0.000	-0.025	-0.019
Omnibus:	3295.185		Durbin-Watson:	0.203		
Prob(Omnibus):	0.000		Jarque-Bera (JB):	4562.191		
Skew:	0.898		Prob(JB):	0.00		
Kurtosis:	3.706		Cond. No.	1.00		

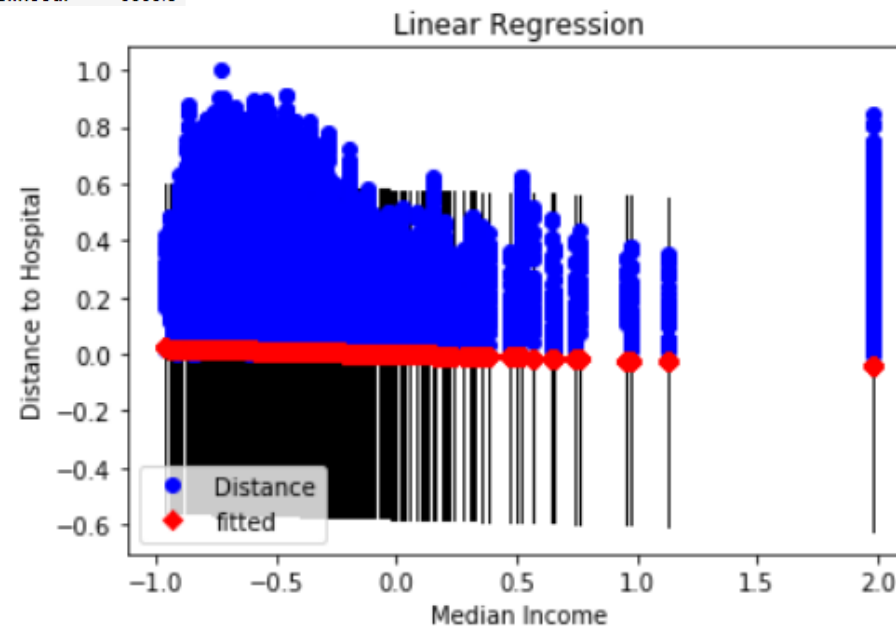
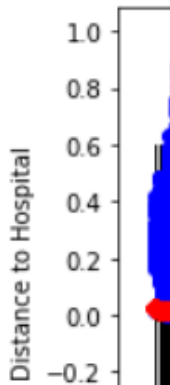
Results

Statistical Analysis

- On the other hand, we calculated a rather low R-squared value of 0.006.
- This implies that although income may be one factor, other factors also influence hospital location.

OLS Regression Results

Dep. Variable:	Distance	R-squared (uncentered):	0.006			
Model:	OLS	Adj. R-squared (uncentered):	0.006			
Method:	Least Squares	F-statistic:	164.0			
Date:	Thu, 26 Mar 2020	Prob (F-statistic):	1.86e-37			
Time:	21:17:27	Log-Likelihood:	-5853.0			
No. Observations:	29424					
Df Residuals:	29423					
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Median	-0.0221	0.002	-12.808	0.000	-0.025	-0.019
Omnibus:	3295.185	Durbin-Watson:	0.203			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4562.191			
Skew:	0.898	Prob(JB):	0.00			
Kurtosis:	3.706	Cond. No.	1.00			



Results

Statistical Analysis

- **We continue this analysis by looking at the distance of each income respondent to the hospital nearest to them.**
- **To do this, I used PANDAS groupby method.**

Results

Statistical Analysis

- We see that using least squares, the analysis returns a standard error of 0.005.
- This implies that the relationship between income and distance to hospitals is statistically significant

OLS Regression Results

Dep. Variable:	Distance	R-squared (uncentered):	0.018
Model:	OLS	Adj. R-squared (uncentered):	0.016
Method:	Least Squares	F-statistic:	11.27
Date:	Thu, 26 Mar 2020	Prob (F-statistic):	0.000838
Time:	21:52:06	Log-Likelihood:	416.62
No. Observations:	613	AIC:	-831.2
Df Residuals:	612	BIC:	-826.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Median	-0.0167	0.005	-3.357	0.001	-0.026	-0.007
Omnibus:	311.930	Durbin-Watson:	0.035			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1455.240			
Skew:	2.378	Prob(JB):	0.00			
Kurtosis:	8.862	Cond. No.	1.00			

Results

Statistical Analysis

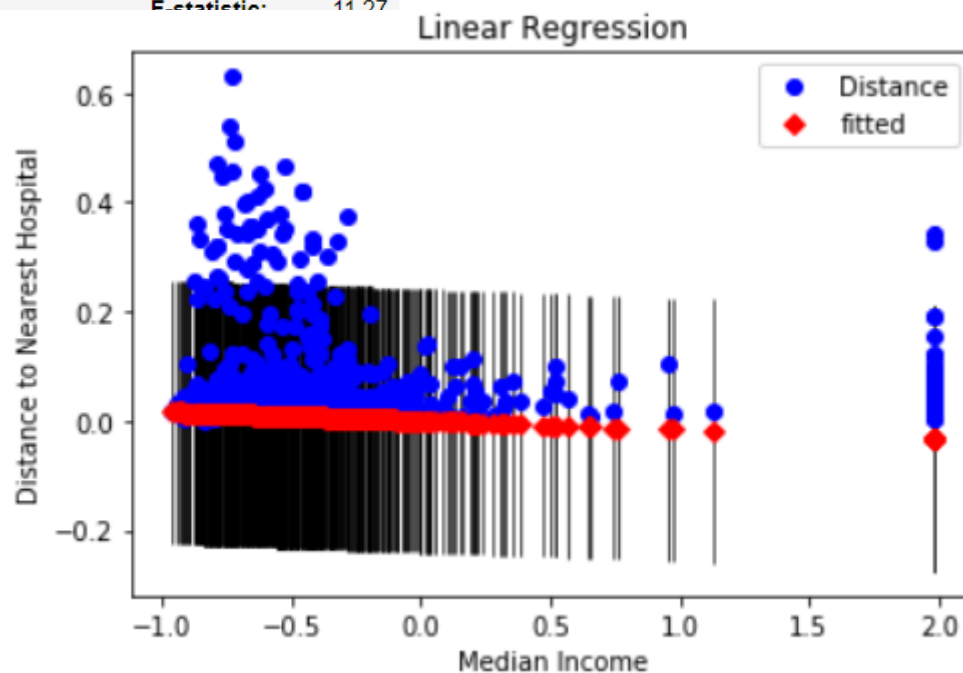
- R-squared value in this case was a bit higher at 0.018, this is still low.
- This implies that while income is important, other factors also influence hospital location

OLS Regression Results

Dep. Variable:	Distance	R-squared (uncentered):	0.018
Model:	OLS	Adj. R-squared (uncentered):	0.016
Method:	Least Squares	F-statistic:	11.27

Date:	Thu, 26 Mar 2020
Time:	21:52:06
No. Observations:	613
Df Residuals:	612
Df Model:	1
Covariance Type:	nonrobust

	coef	std err	t	P> t	[
Median	-0.0167	0.005	-3.357	0.001	-
Omnibus:	311.930				Durbin-Watson
Prob(Omnibus):	0.000				Jarque-Bera (JB)
Skew:	2.378				Prob(JB)
Kurtosis:	8.862				Cond. N



Results

Statistical Analysis

- **There is a statistically significant relationship between median income and hospital distance.**
- **However, as our low r-squared value shows, income is only one factor that models hospital location.**
- **In order to build a stronger model, we would want to look at other factors that could influence hospital location such as population, tax incentives to build hospitals, highway access, and other things.**
- **We must also question whether Dallas is a good representation of the United States as a whole**
- **Further study would have to be done in other cities to see whether this is the case.**

Discussion

Further Study

- **Distance to hospital is a critical factor in likelihood of patient recovery.**
- **We reviewed whether there was a link between median income and distance of travel to hospitals in Dallas, TX**
- **We found that the relationship was statistically significant.**
- **We found that income only explains a small part of why hospitals are located where they are**
- **Further study should be done to examine other factors**

Conclusion

Significant, but not the whole story.