CS 105 Final Project
Kelsey Borovinsky (klboro)
Megan Fantes (mfantes)
Rebecca Jahnke (rsjahnke)
Victor Kholod (vkholod)

Cat Videos Save Lives

A Study of the Relationship Between Internet Use and Homicide Rates Across the Globe

1. Introduction

Our project interprets internet use, homicide rates and population size for countries around the world from 1990-2014 to see what, if any, relationships exist between these factors. The central, initial question that piqued our curiosity was whether internet use is linked to homicide rates. Perhaps high internet use signified a more highly educated population, and so there would be less crime (or maybe people were spending more time watching cat videos than plotting a crime). Or, perhaps more internet use would mean more access to violent content and forums for violent communities, resulting in higher homicide rates. We set out to build a model that would allow us to determine what, if any, link exists.

Ultimately, when we ran our data in Weka, we found that the highest indicator of homicide rate was population. This makes sense: as population rises, you'd expect homicide to rise proportionally.

However, since this model did not answer our central curiosity as to whether there was a link between internet use and homicide rates, we also ran our data in R, another statistical tool. R presented us with a model showing higher internet use as linked with lower homicide rates. Our theory of high internet use indicating a more educated, civilized culture with less homicide held. However, the p-value of this model was quite high, indicating that this seemingly potential link - and theory - is insignificant.

2. Dataset Description

We denormalized three tables from the United Nations' collection of databases to create our dataset.

The first table we pulled from detailed the percent of individuals using the internet in each country for each year from 1990 to 2014 (http://data.un.org/Data.aspx?d=ITU&f=ind1Code%3aI99H).

The second table we pulled from gave homicide rate for each country in each year from 1990 to 2014 (http://data.un.org/Data.aspx?d=UNODC&f=tableCode%3a1).

Finally, the third table gave population size for each country from 1979 to 2014 (http://data.un.org/Data.aspx?d=POP&f=tableCode%3a22).

We combined data from all three data tables into one table describing the rate of internet use, the homicide count, the homicide rate and the population size for each country in each year from 1990 to 2014.

Key attributes present in the table we've created from these three datasets are population size per country, count (representing the number of homicides per country), homicide rate per country, and internet use (quantified as the percent of individuals using the internet per country). The primary key was the combination of (_Country_, _Year_), because countries will appear multiple times for multiple years with that year's data, however the combination of country and year is unique (all data for a given country in a given year will be in one record). See the table below for more information on each attribute:

Country (Primary Key)	Name of the country (string)
Year (Primary Key)	Year the population, homicide count, homicide rate and internet use numbers are for (String)
Population	An integer representing the population size of the country
Homicide Count	An integer representing the number of homicides per country
Homicide Rate	An value representing the homicide rate per country based on number of homicides and the population
Internet Use	An integer representing the percent of individuals using the internet per country

Our final relational table contained the year, country, population, internet use percent, homicide count and homicide rate.

3. Data Preparation

To begin our data preparation, we downloaded the internet, homicide, and population tables from the UN database website. However, the UN website only allows users to download 100,000 rows of a table at a time. The internet and homicide tables were small enough to

download in their entirety, but the population table contained more than 1.5 million rows, so we had to download the table 100,000 rows at a time.

After downloading all of the 100,000-row population subtables, we had to “clean” the population data and combine the sub-tables into one large table. The original table from the UN website was so large because it broke down the population of each country for each year into many different categories (total population, number of men, number of women, number of people living in urban/rural areas, population by age, etc.). We only needed the total population of each country in each year, so we wrote a Python code to extract only the total population values. This code read all CSV files containing the population sub-tables, and printed the total populations to a single output file.[1] Once the population was cleaned, we uploaded the population, internet, and homicide tables to SQLite, to create a SQLite database that we could access through Python.

Once the tables were uploaded to SQLite, we wrote a Python code to denormalize the tables, i.e. combine them into one relational table. Our code connected Python to the sqlite3 module, executed a join command, and printed the resulting table to an output file.[2] The output table contained attributes for Country, Year, Population, Percent of Population Using the Internet, Homicide Count per 100,000 People, and Homicide Rate per 100,000 People. After cleaning and combining all of our data, we ended with 731 data points.

For our data analysis, we were curious about the *change* in internet use, homicide rate, and population from year to year. We planned to conduct numeric estimation using the numeric change in internet, homicide, and population, so we needed to calculate the change in each attribute from year to year. We also planned to conduct classification learning on indicators for internet use, homicide rate, and population -- 1 for increase, -1 for decrease, 0 for no change -- to see if there were any relationships simply between changes in internet use, homicide, and population without knowing specific numbers. We wrote a Python code to calculate the average change in each attribute from year to year and determine the correct indicator.[3] Our code made pairwise comparisons between sequential years for each country, calculated averages and indicators, and printed the resulting table to an output file.

Finally, we wanted to add a column for the region each country is in to see if there were trends in different areas of the world. We found a list of each country and its region,[4] created a table in Excel, and wrote a Python code to join our table of internet use, homicide, and population with the table indicating region,[5] and print it to an output file.[6]

For our data analysis, we planned to use both categorical and numeric algorithms in Weka. In Excel, we created two separate tables: one table with region and the categorical indicators for change in population, internet use, and homicide count per 100,000 people,[7] and one table with the numeric changes in population, internet use, and homicide count per 100,000 people.[8] Once these tables were uploaded to Weka, we split them into training and test sets using an 80/20 split, respectively.

4. Data Analysis

To analyze our categorical data, we first conducted a 1R analysis on the indicator variable for change in homicide count per 100,000 people, which chose the single attribute that best predicted the change in homicide. The 1R algorithm acts as a baseline test providing a baseline accuracy to which we can compare the accuracy of the rest of our categorical tests. If the accuracy of a more complicated algorithm is lower than 1R, then we know the more complicated algorithm is not efficient or worthwhile for our data. Once we ran a 1R algorithm, we conducted a J48 analysis on our categorical attributes, which created a decision tree from the attributes to best predict change in homicide count.

To further explore the relationships between our attributes, we ran a numeric estimation algorithm on our numeric attributes (Avg. change in population, avg. change in internet use) to find the best prediction model for change in homicide count. The numeric estimate algorithm conducts a linear regression on the data and outputs a formula for the regression. We used the M5 method, which removes the least significant attributes from the model, to create our model.

Weka does not give significance values for its linear models, nor does it allow us to create a model for our original question: are internet use and homicide count correlated? We decided to use the R statistical computing package to find significance values and independently create our own linear models. Once we imported our data set into R, we could write our own formulas for linear models because the R interface is much more similar to Python than Weka, in that we can type commands into the console to get output. R gives a much more detailed summary of the model, its accuracy, and its significance, giving us even better insight into the true relationship between our attributes than numeric estimation in Weka.

5. Results

The 1R algorithm chose region as the most accurate indicator of change in homicide count, with 55% accuracy.[9] When we ran the 1R model on the test set, the model still yielded 55% accuracy, indicating that the 1R model did not overfit the training data and generalized well.

The J48 algorithm created a decision tree that predicted change in homicide count directly from region, like the 1R model, for all regions except Asia.[10] If a country is in Asia, the decision tree then looks at the indicator for change in internet use in order to predict change in homicide rate. The decision tree had 52% accuracy,[11] which is lower than the accuracy of 1R and indicates that the simpler 1R model is sufficient to describe the data. When we ran the J48 model on the test set, the model yielded 54% accuracy, which is *higher* than the training accuracy and indicates that the model did not overfit the training data and generalizes well.

When we conducted our numeric estimation test, with the M5 method, the algorithm created the following formula:

Avg. Change in Hom. Count = -12.0074 + (0.0001) * Avg. Change in Pop.

This formula indicates that Average Change in Homicide Count is more significantly correlated with Average Change in Population than Average Change in Internet Use, because the M5 method removed the Change in Internet Use variable. The positive coefficient for the Average Change in Population variable indicates that homicide and population are positively correlated, i.e. as population increases, homicide increases. Note that this makes intuitive sense - as population rises, you'd expect homicide to rise proportionally.

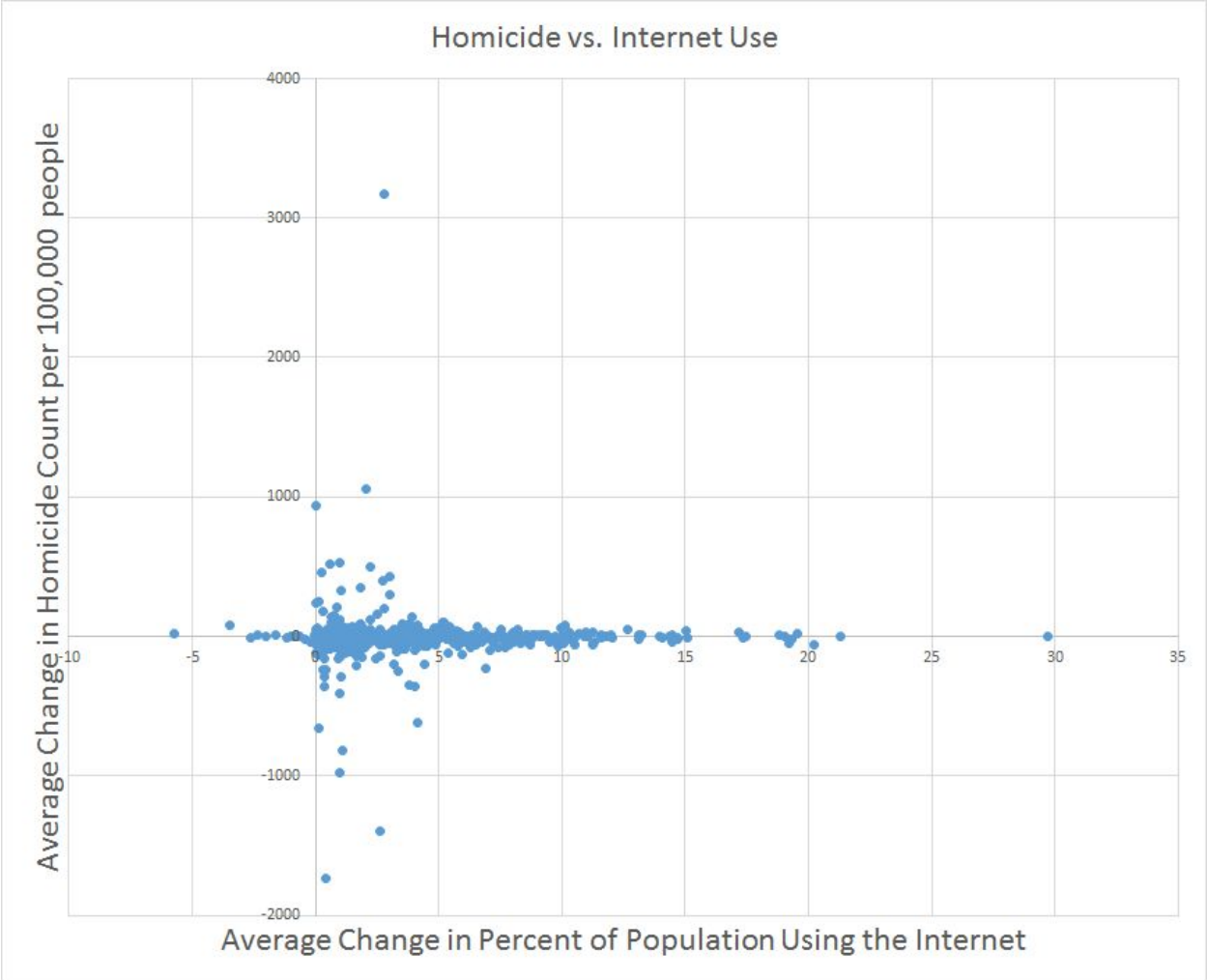
We found the significance of our linear model by running a regression with the same formula in R.[12] The significance of a linear model is given as a p-value, and a very significant model has a very low p-value. (A p-value essentially indicates the probability that our model is wrong -- so the lower the p-value, the lower the probability that our model is wrong). In general, a p-value less than 0.05 indicates a significant model, and a p-value less than 0.001 indicates a very significant model. When we ran our model in R, we found that its p-value is less than 0.001, meaning homicide count is significantly correlated with change in population.

However, the correlation between homicide and population size was not our research question. We wanted to know the correlation between homicide and internet use, so we ran a regression on the data predicting homicide from internet use.[13] R produced the following formula:

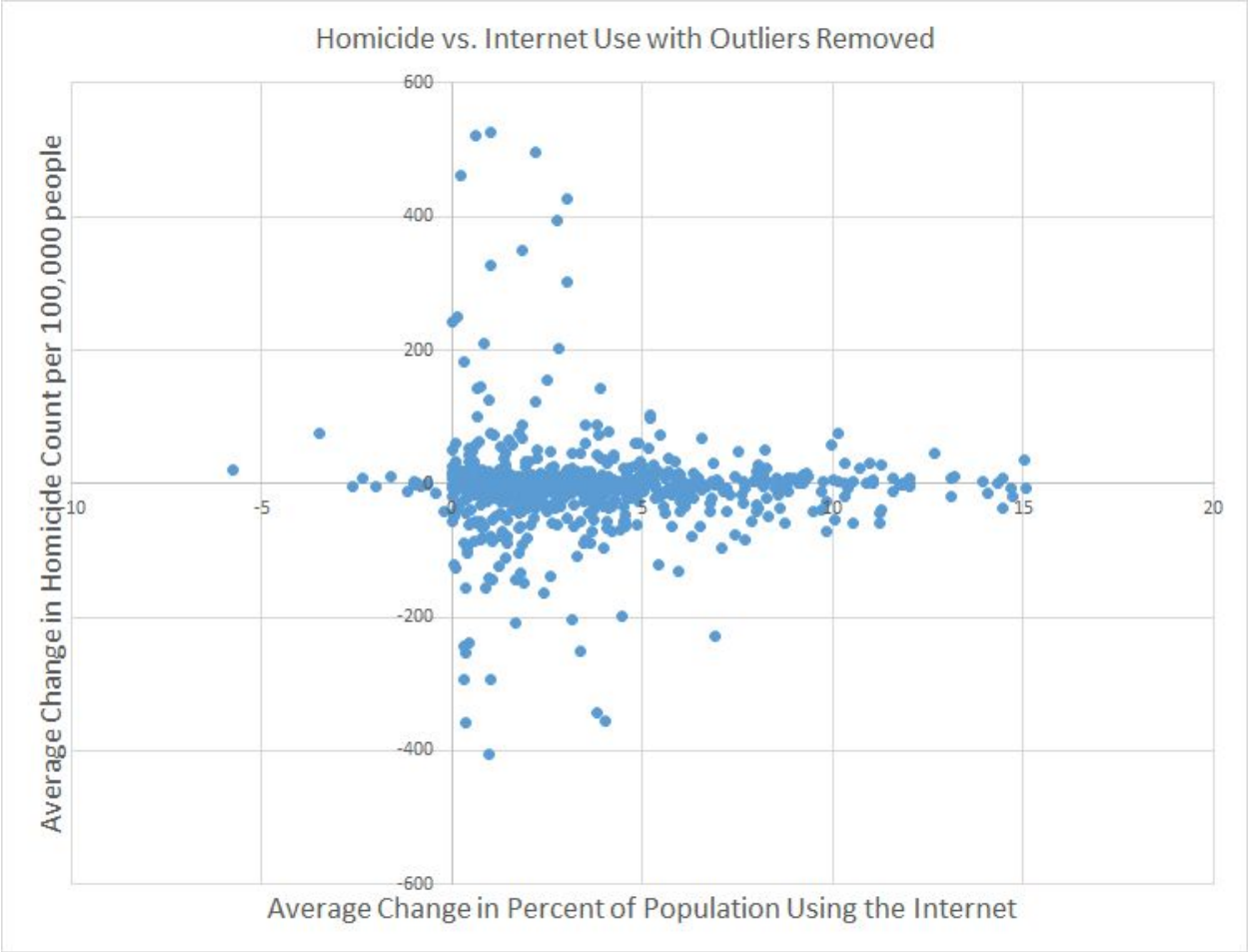
Avg. Change in Hom. Count = 0.5490 + (-0.4601) * Avg. Change in Internet Use

The negative coefficient for the Average Change in Internet Use variable indicates that homicide and internet use are negatively correlated, i.e. as internet use increases, homicide decreases. However, the p-value for this model is 0.825, which is very high and indicates that our model is not significant. Thus we may conclude that homicide and internet use are not correlated.

Even though we found that homicide and internet use are not significantly correlated, we still wanted to fully understand any relationship they may have and give our original research question due diligence. We began by making a scatter plot of homicide count vs. internet use:



When we looked at the graph, we noticed that there were a few extreme outliers in the data. Such outliers prevent us from understanding the relationship between the majority of the data points, because they stretch the scale of the graph and squish the rest of the data points together. We were interested to see the graph with the outliers removed, so we removed all data points that were more than 3 standard deviations away from the mean of either the average change in homicide count or the mean of average change in internet use. We removed 20 outliers and created a new scatter plot:

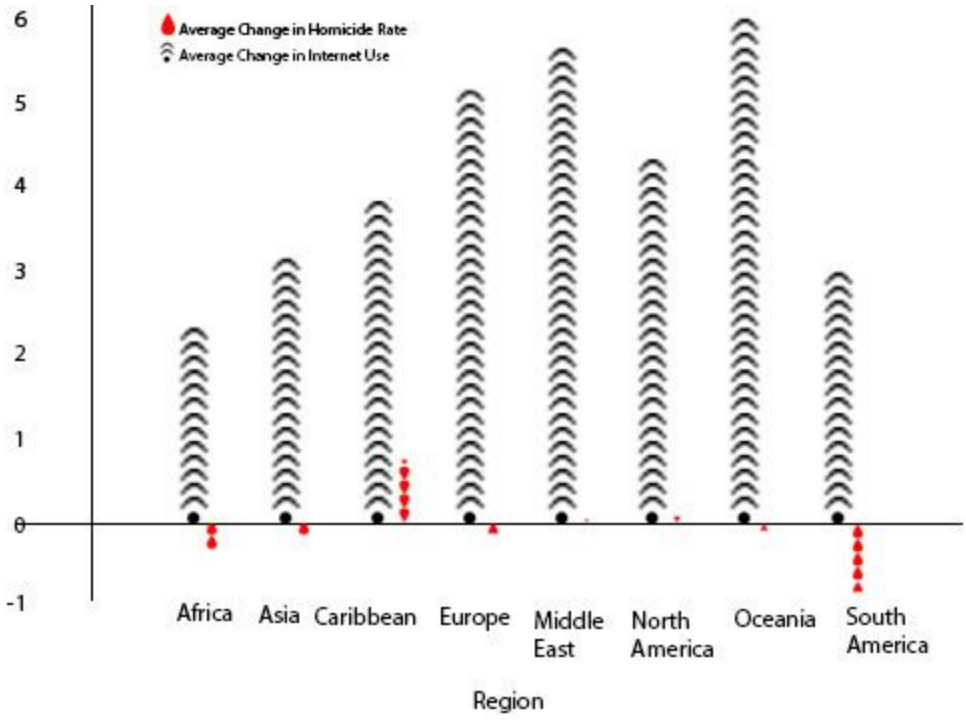


In the graph with outliers removed, we can more easily see general trends in the data. From the new graph, we see that lower values in change in internet use have much higher variability in change in homicide count, and higher values in change in internet use have less variability in change in homicide count (i.e. are more concentrated around zero).

Data points with change in internet use less than 5% have the highest variability in change in homicide rate. This suggests that countries with a slower rate of increase in internet use are much less predictable. In general, countries that are slower to advance in technology are more unstable, so their social behavior (like tendency to commit murder) is more erratic. Data points with change in internet use greater than 5% have low variability in change in homicide count. This suggests that countries with a higher rate of increase in internet use are more predictable. Countries that are rapidly advancing in technology are also likely rapidly modernizing and becoming more organized. Having a more organized society that is more globally connected through technology yields behavior that is much less erratic and much less likely to drastically change over time.

While we could not quantize the correlation between homicide and internet use using a formula, a simple scatter plot of the data revealed likely relationships between the variables.

Our scatter plot only reveals global trends. We were curious if there was a difference in the relationship between homicide and internet use between the different regions of the world[14]. We created a bar graph comparing change in internet use from 1990 - 2014 and the change in homicide rate from 1990 - 2014:



We decided to compare homicide rates instead of homicide counts so that the units of change (percent of population) are the same for both homicide rate and internet use (and as homicide rate tended to follow the trends of homicide count in our data, the switch between rate and count was inconsequential). From our bar graph, we see that internet use increases for every region, as expected. However, the change in homicide rate is very different for every region. In South America, for example, the average change in homicide rate from 1990 - 2014 significantly decreases. If we had run a linear regression on just South America, we might have found more significant evidence that an increase in internet use is correlated with a decrease in homicide rate (or homicide count). Looking back, we can also see from this graph why it made intuitive sense for 1R to have chosen region; the change in homicide rate varies from region to region, so region is indicative of what that change will look like. Future studies on the relationship between homicide and internet use could analyze different regions in more detail and create linear models for each individual region. Perhaps such a future study will find significant evidence for vastly different trends in each region, and the variability between regions is the reason for the lack of significant evidence in the global trends of our study.

6. Conclusions

We set out to see if there was a correlation between Internet use and homicide rates throughout the world from 1990-2014. After running our data on global populations, homicide counts, and internet rates through various data mining algorithms in Weka, we found that the greatest numeric indicator of homicide count was population. This made sense considering if population increased, then homicide counts should increase as well. To answer our initial question about the correlation between homicide and internet use, we ran our data in R, and discovered that when internet use increased, homicide counts decreased. We cannot make this claim in confidence, though, due to a high p-value indicating that our theory is insignificant. Thus we conclude that there is no direct correlation linking internet usage with homicide counts.

7. Appendix

[1]see CodeToCleanPop.py

[2]see CodeToJoinTables.py

[3]see CodeToAdd_Indicators_Averages.py

[4]on www.internetworldstats.com

[5]see CodeToAddRegion.py

[6]see Full Table with Regions

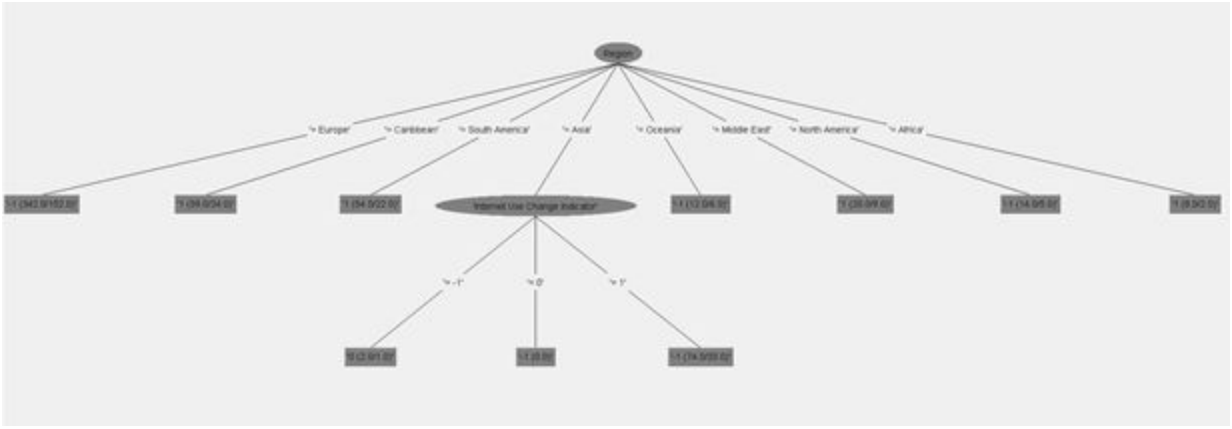
[7]see Classification Table

[8]see Numeric Table

[9]Confusion matrix:

	Predicted Value:			
Actual Value:		**-1**	**0**	**1**
	-1	242	0	55
	0	21	0	6
	1	179	0	82

[10]Decision Tree:



[11]Confusion matrix:

	Predicted Value:			
Actual Value:		**-1**	**0**	**1**
	-1	178	0	119
	0	11	0	16
	1	137	0	124

[12]Abbreviated R output:
Coefficients: Estimate
(Intercept) -1.201e+01
Avg. Change in Pop. 1.350e-04
p-value: 3.022e-11

[13]Abbreviated R output:
Coefficients: Estimate
(Intercept) 0.5490
Avg. Change in Internet Use -0.4601
p-value: 0.825

[14]see CodeToCalcAvgs.py (includes SQL command to calculate averages)