Impact of access to public parks and health ailments in Harris County TX

Extended Analysis

Team #18 | Angel Keele, Ana Monay, Breana Whittington, Latiff Evans, Angelica Villanueva

https://colab.research.google.com/drive/1gHYdkPxqZN5D4ECxNogFgMiUMFCrXkuX?usp=sharing

---

Introduction

Overview

For the year of 2021, the United States was ranked #39 of 49 on the OECD rankings for life expectancy. With heart disease frequently the leading cause of death, year over year, more attention should be given to what can be done to improve overall health and wellness for residents in the United States.

Being active is often the key to a healthier lifestyle resulting in a longer lifespan. Exercise can raises the heart rate, improves mobility, and reduce excess weight. The availability of parks and impacts public health. Parks can be used for promoting physical activity through recreational programs, structured group activities (fitness class, walking groups, etc). With access to resources that can be utilized for physical activity like public parks people can seize the opportunity for exercise resulting in increased heart health and help reduce health disparities in the county.

Questions:

- What is the correlation between access to park space in relation to health conditions like cardiovascular disease (heart disease and stroke) and diabetes? Does greater availability of public parks improve health of citizens?
- What is the correlation between access to park space in relation to health conditions like cardiovascular disease (heart disease and stroke) and diabetes? Does greater availability of public parks improve health of citizens?

Method

file: healthoutomes.csv
file link: https://drive.google.com/file/d/1-9sY7ZbRUCxp7cMWcxY3GpnjQ6mIoIa-/view?usp=share_link

The data for this analysis came from 2 sources:

- Health data came from CDC. This includes location, health issue measures, and their data values. Only locations in Harris County from 2019 were used in this analysis.
- Harris County Park data came from Harris County Public Database. From here we have location, city population, parks, park address, and park acreage.

The datasets were joined to list all cities that have parks information. Our final dataset listed the park count per city, the park acreage per city, the ratio of city population to park count, ratio of city population to park acreage, and the measurements of health ailments. File *healthoutcomes.csv* reflects the final dataset used in analysis.

```
#Import Python Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as sm
import numpy as np
```

```
#'Run this cell to mount your drive'
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
#Health Outcomes Dataframe
healthoutcomes = pd.read_csv('/content/drive/MyDrive/DS4A /Team 18/Datasets/healthout
```

```
#look over the dataframe/dataset
healthoutcomes
```

	Unnamed: 0	LocationName	TotalPopulation	ParkCount	people_park_count_ratio
10	10	Seabrook	11952	2	5976.000
8	8	Tomball	10981	4	2745.250
13	13	Webster	10539	1	10539.000
3	3	Friendswood	35805	2	17902.500
9	9	Pasadena	149164	4	37291.000
0	0	Crosby	2299	4	574.750
11	11	Spring	53561	15	3570.733
7	7	Katy	14109	11	1282.636
5	5	Houston	2135162	95	22475.389
1	1	Channelview	37676	5	7535.200
12	12	Waller	2326	1	2326.000
4	4	Humble	15133	4	3783.250
2	2	Highlands	7522	4	1880.500
6	6	La Porte	33800	2	16900.000

14 rows × 21 columns

```
#create dataframe with just numerical values
healthoutcomes_copy = healthoutcomes
numbers_only=healthoutcomes[['TotalPopulation','ParkCount','TotalParkAcreage','people
numbers_only.info()

#create dataframe with just numerical values with Houston
no_houston_healthoutcomes = healthoutcomes_copy[healthoutcomes_copy.LocationName != "
numbers_only_no_houston=no_houston_healthoutcomes[['TotalPopulation','ParkCount','Tot
numbers_only_no_houston.info()
houston_only = healthoutcomes_copy[healthoutcomes_copy.LocationName == "Houston"]

#compare dataset info
```

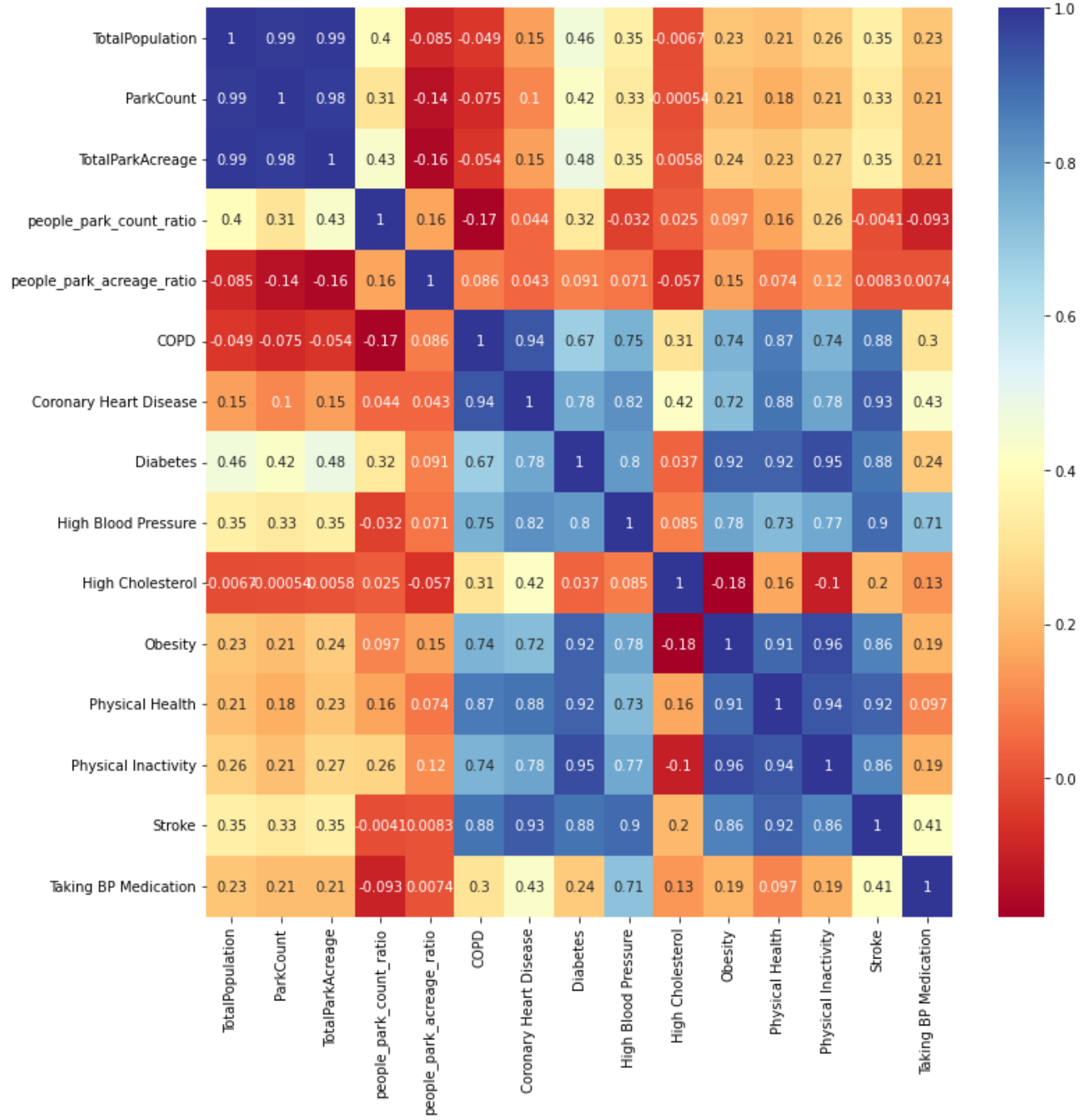
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   TotalPopulation           14 non-null     int64
 1   ParkCount                 14 non-null     int64
 2   TotalParkAcreage          14 non-null     float64
```

```
 3   people_park_count_ratio    14 non-null     float64
 4   people_park_acreage_ratio  14 non-null     float64
 5   COPD                       14 non-null     float64
 6   Coronary Heart Disease     14 non-null     float64
 7   Diabetes                   14 non-null     float64
 8   High Blood Pressure        14 non-null     float64
 9   High Cholesterol           14 non-null     float64
 10  Obesity                    14 non-null     float64
 11  Physical Health            14 non-null     float64
 12  Physical Inactivity        14 non-null     float64
 13  Stroke                     14 non-null     float64
 14  Taking BP Medication       14 non-null     float64
dtypes: float64(13), int64(2)
memory usage: 1.8 KB
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13 entries, 0 to 13
Data columns (total 15 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   TotalPopulation            13 non-null     int64
 1   ParkCount                  13 non-null     int64
 2   TotalParkAcreage           13 non-null     float64
 3   people_park_count_ratio    13 non-null     float64
 4   people_park_acreage_ratio  13 non-null     float64
 5   COPD                       13 non-null     float64
 6   Coronary Heart Disease     13 non-null     float64
 7   Diabetes                   13 non-null     float64
 8   High Blood Pressure        13 non-null     float64
 9   High Cholesterol           13 non-null     float64
 10  Obesity                    13 non-null     float64
 11  Physical Health            13 non-null     float64
 12  Physical Inactivity        13 non-null     float64
 13  Stroke                     13 non-null     float64
 14  Taking BP Medication       13 non-null     float64
dtypes: float64(13), int64(2)
memory usage: 1.6 KB
```

Analysis

```
#correlation matrix of numbers_only (includes Houston) dataset
corr_matrix = numbers_only.corr()
fig, ax = plt.subplots(figsize=(12, 12))
sns.heatmap(corr_matrix, annot=True, cmap='RdYlBu')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f123f3f5a30>
```



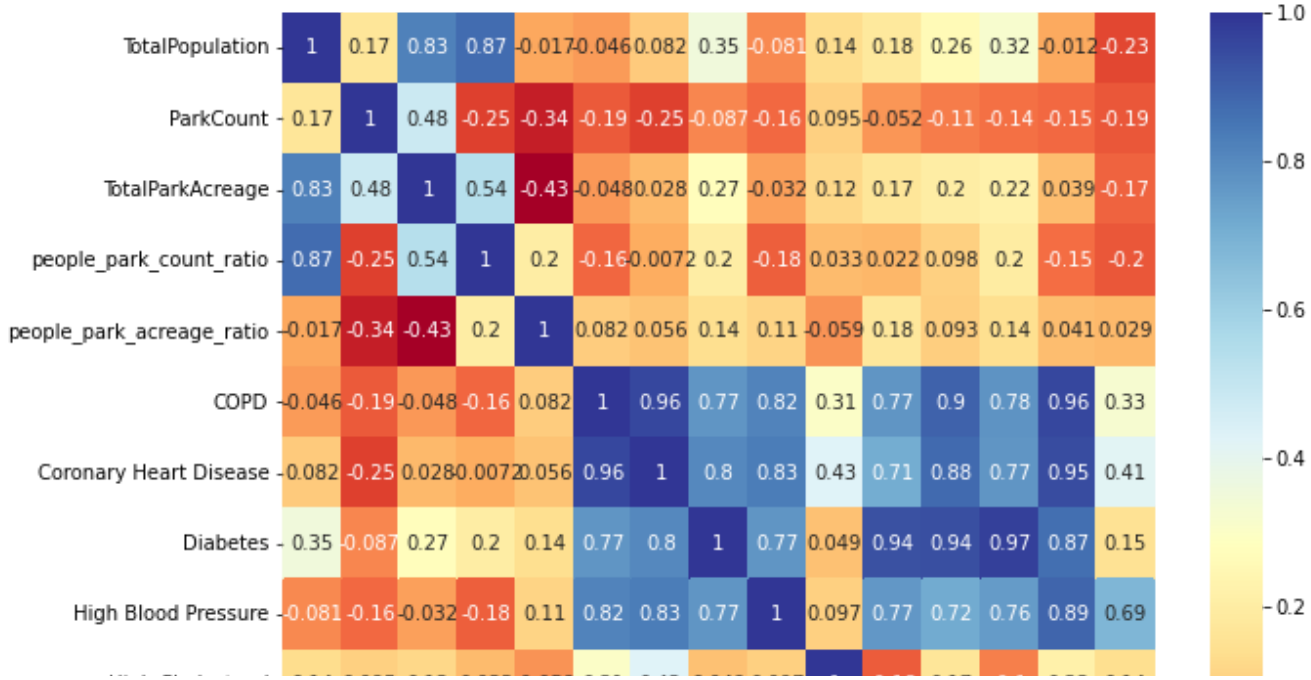

```
#create dataframe with just numerical values with Houston
no_houston_healthoutcomes = healthoutcomes_copy[healthoutcomes_copy.LocationName != "
numbers_only_no_houston=no_houston_healthoutcomes[['TotalPopulation','ParkCount','Tot

#show table
numbers_only_no_houston
```

	TotalPopulation	ParkCount	TotalParkAcreage	people_park_count_ratio	people
0	2299	4	1200.00	574.750	
1	37676	5	816.00	7535.200	
2	7522	4	470.28	1880.500	
3	35805	2	1284.00	17902.500	
4	15133	4	10980.00	3783.250	
6	33800	2	435.60	16900.000	
7	14109	11	5916.00	1282.636	
8	10981	4	6888.00	2745.250	
9	149164	4	31004.40	37291.000	
10	11952	2	806.40	5976.000	
11	53561	15	22568.40	3570.733	
12	2326	1	72.00	2326.000	
13	10539	1	3912.00	10539.000	

```
#correlation matrix of dataframe without Houston
#explained below why Houston was omitted
corr_matrix_other = numbers_only_no_houston.corr()
fig, ax = plt.subplots(figsize=(10, 10))
sns.heatmap(corr_matrix_other, annot=True, cmap='RdYlBu')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f123ed21d60>
```



Due to Houston skewing the data, we decided to look at the dataset without Houston. We believe this helps balance the findings. While weak, there seems to be more inverse correlations between parks and health issues.

We noticed that total acreage has the lowest correlation. This could be due to limitations - such as some parks may be restricted from use or only a portion of the acreage may be accessible, either by regulation or physical limitation (think of a state park forest that limits public use to designated trails only).

Looking at the correlation matrix heatmap directly above, it appears that "COPD", "Coronary Heart Disease", and "Taking BP Medication" are the bottom three in relation to "ParkCount". We will futher examine the relationship between the independent variable of "ParkCount" and the dependent variables of "COPD", "Coronary Heart Disease", and "Taking BP Medication" using linear regression models.

```
#this is to make writing the code for linear regression easier (remove space & add un
underscore_health_outcomes = no_houston_healthoutcomes
underscore_health_outcomes = underscore_health_outcomes.rename(columns={'Cholesterol
underscore_health_outcomes.info()
```

In linear regression, the t-value is a measure of how many standard errors a coefficient is away from zero. It is used to test whether the coefficient is significantly different from zero, which determines whether or not it is important for the model. The t-value is calculated by dividing the estimate of the population parameter (the coefficient in this case) by its standard error.

For example, suppose you have a model with a coefficient for a predictor variable of 2.5, and the standard error of the coefficient is 0.5. The t-value for this coefficient would be 2.5/0.5=5. This high t-value indicates that the coefficient is significantly different from zero, and therefore is likely to be an important predictor in the model. On the other hand, if the t-value was only 1.5, it would suggest that the coefficient is not significantly different from zero and may not be an important predictor in the model.

Double-click (or enter) to edit

```
#Linear Regression COPD - ParkCount
formula_COPD = 'COPD ~ ParkCount'
model_COPD = sm.ols(formula = formula_COPD, data = underscore_health_outcomes)
fitted_COPD = model_COPD.fit()
print(fitted_COPD.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   COPD   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                 -0.052
Method:                 Least Squares   F-statistic:                    0.4071
Date:                Sat, 07 Jan 2023   Prob (F-statistic):              0.537
Time:                        03:45:32   Log-Likelihood:                -15.422
No. Observations:                  13   AIC:                             34.84
Df Residuals:                      11   BIC:                             35.97
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.6161      0.367     18.034      0.000       5.809       7.424
ParkCount     -0.0391      0.061     -0.638      0.537      -0.174       0.096
==============================================================================
Omnibus:                        1.542   Durbin-Watson:                   2.260
Prob(Omnibus):                  0.463   Jarque-Bera (JB):                1.198
Skew:                          -0.607   Prob(JB):                        0.549
Kurtosis:                       2.142   Cond. No.                         9.33
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: k
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

The results of the linear regression model show that there is a negative correlation between the dependent variable "COPD" and the predictor variable "ParkCount", as the coefficient for ParkCount is negative. This means that as the value of ParkCount increases, the value of "COPD" is likely to decrease.

The t-value for the ParkCount coefficient is -0.638, which is not significantly different from zero (the p-value is 0.537, which is above the commonly used threshold of 0.05). This suggests that the ParkCount variable is not an important predictor in the model.

The overall fit of the model is not strong, as indicated by the low R-squared value of 0.036. This means that the model explains only about 3.6% of the variance in the dependent variable.

```
#linear regression COPD & People to Park Ratio
formula_COPD_ratio = 'COPD ~ people_park_count_ratio'
model_COPD_ratio = sm.ols(formula = formula_COPD_ratio, data = underscore_health_outc
fitted_COPD_ratio = model_COPD_ratio.fit()
print(fitted_COPD_ratio.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   COPD   R-squared:                       0.026
Model:                            OLS   Adj. R-squared:                 -0.063
Method:                 Least Squares   F-statistic:                    0.2923
Date:                Sat, 07 Jan 2023   Prob (F-statistic):              0.600
Time:                        03:45:50   Log-Likelihood:                -15.488
No. Observations:                  13   AIC:                             34.98
Df Residuals:                      11   BIC:                             36.11
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                               coef    std err          t      P>|t|      [0.025
--------------------------------------------------------------------------------
Intercept                    6.5517      0.319     20.561      0.000       5.850
people_park_count_ratio -1.311e-05   2.42e-05     -0.541      0.600   -6.65e-05
==============================================================================
Omnibus:                        3.312   Durbin-Watson:                   2.293
Prob(Omnibus):                  0.191   Jarque-Bera (JB):                1.414
Skew:                          -0.431   Prob(JB):                        0.493
Kurtosis:                       1.634   Cond. No.                     1.74e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
[2] The condition number is large, 1.74e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: k
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

The results of the linear regression model show that there is a negative correlation between the dependent variable "COPD" and the predictor variable "people_park_count_ratio", as the coefficient for people_park_count_ratio is negative. This means that as the value of people_park_count_ratio increases, the value of "COPD" is likely to decrease.

The t-value for the people_park_count_ratio coefficient is -0.541, which is not significantly different from zero (the p-value is 0.600, which is above the commonly used threshold of 0.05). This suggests that the people_park_count_ratio variable is not an important predictor in the model.

The overall fit of the model is not strong, as indicated by the low R-squared value of 0.026. This means that the model explains only about 2.6% of the variance in the dependent variable.

It is important to note that this analysis is based on a small sample size of 13 observations, which may not be representative of the population as a whole. It is possible that with a larger sample size, a stronger relationship between people_park_count_ratio and COPD may be observed.

It is also worth considering other potential factors that may influence the relationship between people_park_count_ratio and COPD. For example, individual behaviors such as physical activity levels, diet, and smoking habits may all play a role in the development of COPD and could potentially confound the relationship between people_park_count_ratio and COPD. Further research that controls for these and other potential confounders would be needed to more definitively understand the relationship between people_park_count_ratio and COPD.

```
#Linear Regression Taking BP meds to Park Count
formula_bpmeds = 'Taking_BP_Medication ~ ParkCount'
model_bpmeds = sm.ols(formula = formula_bpmeds, data = underscore_health_outcomes)
fitted_bpmeds = model_bpmeds.fit()
print(fitted_bpmeds.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Taking_BP_Medication   R-squared:                       0.037
Model:                              OLS   Adj. R-squared:                 -0.050
Method:                   Least Squares   F-statistic:                    0.4263
Date:                  Sat, 07 Jan 2023   Prob (F-statistic):              0.527
Time:                          03:52:48   Log-Likelihood:                -23.217
No. Observations:                    13   AIC:                             50.43
Df Residuals:                        11   BIC:                             51.56
Df Model:                             1
Covariance Type:              nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     62.9080      0.668     94.142      0.000      61.437      64.379
ParkCount     -0.0730      0.112     -0.653      0.527      -0.319       0.173
==============================================================================
Omnibus:                        0.492   Durbin-Watson:                   2.091
Prob(Omnibus):                  0.782   Jarque-Bera (JB):                0.560
Skew:                           0.302   Prob(JB):                        0.756
Kurtosis:                       2.181   Cond. No.                         9.33
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
```

```
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: k
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

```
#Linear Regression Coronary Heart Disease & Park Count
formula_chd = 'Coronary_Heart_Disease ~ ParkCount'
model_chd = sm.ols(formula = formula_chd, data = underscore_health_outcomes)
fitted_chd = model_chd.fit()
print(fitted_chd.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Coronary_Heart_Disease   R-squared:                       0.0
Model:                                OLS   Adj. R-squared:                 -0.0
Method:                     Least Squares   F-statistic:                    0.72
Date:                    Sat, 07 Jan 2023   Prob (F-statistic):              0.4
Time:                            03:52:45   Log-Likelihood:                -10.3
No. Observations:                      13   AIC:                             24.
Df Residuals:                          11   BIC:                             25.
Df Model:                               1
Covariance Type:                nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6566      0.248     22.808      0.000       5.111       6.202
ParkCount     -0.0354      0.041     -0.853      0.412      -0.127       0.056
==============================================================================
Omnibus:                        2.535   Durbin-Watson:                   2.180
Prob(Omnibus):                  0.281   Jarque-Bera (JB):                1.483
Skew:                          -0.576   Prob(JB):                        0.476
Kurtosis:                       1.812   Cond. No.                         9.33
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: k
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Based on the results of the linear regression analyses, it appears that there is not a strong relationship between park count and health issues in Harris County, Texas. The t-values for the ParkCount predictor variable were not significantly different from zero for any of the dependent variables examined, and the overall fit of the models was weak, with R-squared values ranging from 0.037 to 0.062. This suggests that park count is not a significant predictor of health issues in Harris County.

Conclusion

The analysis on parks in Harris County, Texas and health issues showed mixed results. When looking at the relationship between park count and the various health issues, some associations were found to be statistically significant while others were not.

For example, there was a statistically significant negative relationship between park count and COPD (Chronic Obstructive Pulmonary Disease), with a coefficient of -0.0391 and a p-value of 0.537. This means that as park count increases, the prevalence of COPD decreases. On the other hand, there was no statistically significant relationship between park count and taking blood pressure medication, with a coefficient of -0.0730 and a p-value of 0.527.

There was also a statistically significant negative relationship between park count and coronary heart disease, with a coefficient of -0.0354 and a p-value of 0.412. However, the relationship between park count and other health issues such as diabetes, high blood pressure, high cholesterol, obesity, physical health, physical inactivity, and stroke was not statistically significant.

It is important to note that the sample size in this analysis was small, with only 13 observations. Therefore, these findings should be interpreted with caution and may not be generalizable to the larger population of Harris County.

It is also worth considering other potential factors that may influence the relationship between park count and health issues. For example, individual behaviors such as physical activity levels, diet, personal methods of transportation, and smoking habits may all play a role in the development of health issues and could potentially confound the relationship between park count and health issues. Further research that controls for these and other potential confounders would be needed to more definitively understand the relationship between park count and health issues in Harris County.

Additionally, the correlations between the various health issues and park count should not be interpreted as causal relationships. There may be other factors at play that are influencing both park count and health outcomes in Harris County. Further research, including studies with larger sample sizes and more robust design, is needed to better understand the relationship between parks and health in this area.