Intro

↳ *1 cell hidden*

Impact of access to public parks on wellness and mortality in Harris County TX

Extended Analysis

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

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 community centers 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 and community centers people can seize the opportunity for exercise resulting in increased heart health and help reduce health disparities in the county.

What is the correlation between access to park space in relation to premature mortality from cardiovascular disease (heart disease and stroke) and diabetes? Does greater availability of physical activity resources, such as public parks and fitness centers (both public and private) help to reduce morbidity and improve life expectancy?

Methods:

We analyzed park space and various health conditions of the citizens in Harris County, TX

▶ ↳ *4 cells hidden*

Brain Storming

- Compare correlations of # of parks vs. acerage of parks
- ANOVA (https://towardsdatascience.com/anova-test-with-python-cfbf4013328b)
- Gather Census Data to look at other identifiers

Initial Analysis

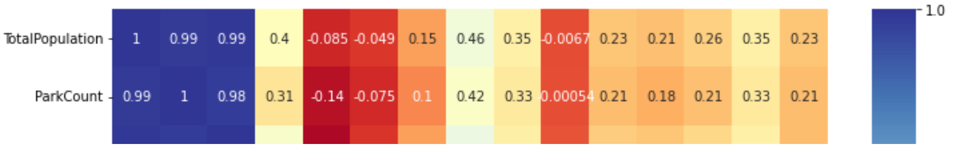
```
healthoutcomes_copy = healthoutcomes
no_houston_healthoutcomes = healthoutcomes_copy[healthoutcomes_copy.LocationName != "Houston"]
numbers_only=healthoutcomes[['TotalPopulation','ParkCount','TotalParkAcreage','people_park_count_ratio','people_park_acreage_rati
numbers_only.info()
#no_houston_healthoutcomes
numbers_only_no_houston=no_houston_healthoutcomes[['TotalPopulation','ParkCount','TotalParkAcreage','people_park_count_ratio','pe
numbers_only_no_houston.info()
houston_only = healthoutcomes_copy[healthoutcomes_copy.LocationName == "Houston"]
```

```
<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
```

Correlation matrix

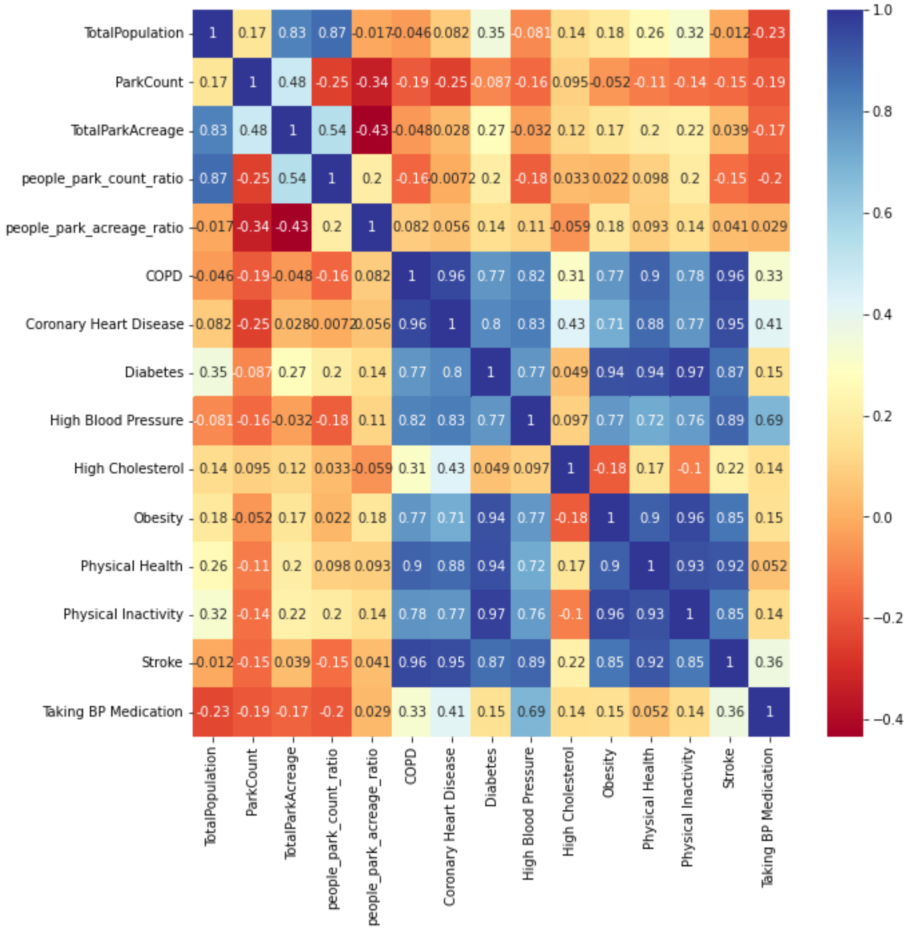
```
corr_matrix = numbers_only.corr()
fig, ax = plt.subplots(figsize=(12, 12))
sns.heatmap(corr_matrix, annot=True, cmap='RdYlBu')
```

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



```
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 beleive this help balance the findings. While weak, there seem to be more inverse correlations between parks and health.

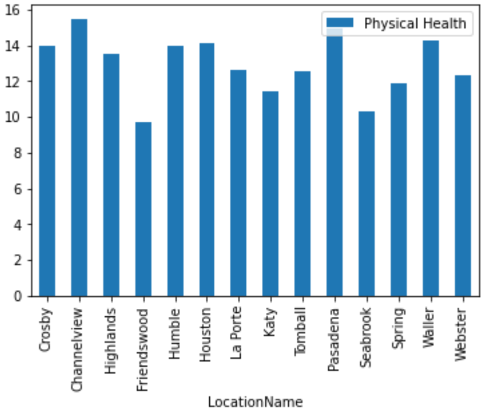
We noticed that total acreage is the lowest correlation. Total acreage of parks had the least amount of the correlation. This could be due to limitations - such as some parks may be restricted from use or only portion of the acreage may be accessible.

```
#calculating the estimated # value of % Data Value & total population
heathoutcomes_numerical = no_houston_healthoutcomes.copy()
heathoutcomes_numerical['Est_pop_Obesity'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Obesity']
heathoutcomes_numerical['Est_pop_HBP'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['High Blood Pressure']
heathoutcomes_numerical['Est_pop_COPD'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['COPD']
heathoutcomes_numerical['Est_pop_Coronary Heart Diseas'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Corona
heathoutcomes_numerical['Est_pop_Diabetes'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Diabetes']
heathoutcomes_numerical['Est_pop_HChol'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['High Cholesterol']
heathoutcomes_numerical['Est_pop_PH'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Physical Health']
heathoutcomes_numerical['Est_pop_PI'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Physical Inactivity']
heathoutcomes_numerical['Est_pop_TakingBPMed'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Taking BP Medica
heathoutcomes_numerical['Est_pop_Stroke'] = healthoutcomes["TotalPopulation"]*0.01*heathoutcomes_numerical['Stroke']

#df of cities & their Phyiscal Health data value (have not have good phyiscal health in the last 14 days)
city_no_feel_good = pd.DataFrame()
city_no_feel_good[['LocationName','Physical Health']] = healthoutcomes[['LocationName','Physical Health']]
```

```
#sum of health outcomes for each city - who has the highest %s?
health_percent_sum = healthoutcomes.groupby('LocationName')['Taking BP Medication','High Blood Pressure','High Cholesterol','Diab
```

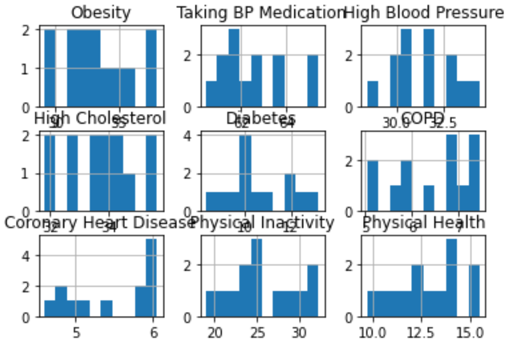
```
<ipython-input-29-ea65ee0e0ea>:19: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys)
      health_percent_sum = healthoutcomes.groupby('LocationName')['Taking BP Medication','High Blood Pressure','High Cholesterol
```

```
#city_no_feel_good = pd.Series(city_no_feel_good)
city_no_feel_good.plot.bar(x='LocationName',y='Physical Health')

# Show the plot
plt.show()
```



```
cityandmeasure = no_houston_healthoutcomes[['LocationName','Obesity','Taking BP Medication','High Blood Pressure','High Choleste
cityandmeasure.hist()
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f123dc29c10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123dbe1850>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123db85fa0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f123db3b700>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123db64e50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123db1a4f0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f123db1a5e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123dac5d60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f123daa6b50>]],
      dtype=object)
```



```
#this is for linear regression to make it easier
underscore_health_outcomes = no_houston_healthoutcomes
underscore_health_outcomes = underscore_health_outcomes.rename(columns={'Cholesterol Screening':'Cholesterol_Screening','High Blo
underscore_health_outcomes.info()

#lin regress try
formula = 'people_park_count_ratio ~ COPD + Taking_BP_Medication + Coronary_Heart_Disease'
model = sm.ols(formula = formula, data = underscore_health_outcomes)
fitted = model.fit()
print(fitted.summary())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13 entries, 0 to 13
Data columns (total 21 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Unnamed: 0                  13 non-null     int64
 1   LocationName                13 non-null     object
 2   TotalPopulation             13 non-null     int64
```

```
  3   ParkCount                    13 non-null     int64
  4   people_park_count_ratio      13 non-null     float64
  5   people_park_acreage_ratio    13 non-null     float64
  6   Arthritis                    13 non-null     float64
  7   COPD                         13 non-null     float64
  8   Cholesterol_Screening        13 non-null     float64
  9   Chronic_Kidney_Disease       13 non-null     float64
 10   Colorectal_Cancer_Screening  13 non-null     float64
 11   Coronary_Heart_Disease       13 non-null     float64
 12   Diabetes                     13 non-null     float64
 13   High_Blood_Pressure          13 non-null     float64
 14   High_Cholesterol             13 non-null     float64
 15   Obesity                      13 non-null     float64
 16   Physical_Health              13 non-null     float64
 17   Physical_Inactivity          13 non-null     float64
 18   Stroke                       13 non-null     float64
 19   Taking_BP_Medication         13 non-null     float64
 20   TotalParkAcreage             13 non-null     float64
dtypes: float64(17), int64(3), object(1)
memory usage: 2.8+ KB
                               OLS Regression Results                              
===================================================================================
Dep. Variable:     people_park_count_ratio   R-squared:                       0.411
Model:                                 OLS   Adj. R-squared:                  0.215
Method:                      Least Squares   F-statistic:                     2.096
Date:                     Sat, 07 Jan 2023   Prob (F-statistic):              0.171
Time:                             03:39:58   Log-Likelihood:                -134.61
No. Observations:                       13   AIC:                             277.2
Df Residuals:                            9   BIC:                             279.5
Df Model:                                3                                         
Covariance Type:                 nonrobust                                         
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept               1.333e+05   1.11e+05      1.202      0.260   -1.18e+05    3.84e+05
COPD                   -2.648e+04   1.12e+04     -2.368      0.042   -5.18e+04   -1179.592
Taking_BP_Medication   -2769.4196   1958.120     -1.414      0.191   -7198.995    1660.156
Coronary_Heart_Disease  3.987e+04   1.69e+04      2.354      0.043    1559.810    7.82e+04
==============================================================================
Omnibus:                        1.561   Durbin-Watson:                   2.771
Prob(Omnibus):                  0.458   Jarque-Bera (JB):                1.169
Skew:                           0.659   Prob(JB):                        0.557
Kurtosis:                       2.350   Cond. No.                     2.77e+03
==============================================================================

Notes:
```

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.

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

The t-value for the COPD coefficient is -2.684, which is significantly different from zero (the p-value is 0.028, which is below the commonly used threshold of 0.05). This suggests that the COPD variable is an important predictor in the model.

On the other hand, the t-values for the "Physical_Health", "Taking_BP_Medication", and "Coronary_Heart_Disease" variables are not significantly different from zero (their p-values are all above 0.05). This means that these variables are not likely to be important predictors in the model.

The overall fit of the model is not particularly strong, as indicated by the low R-squared value of 0.503. This means that the model explains only about 50% of the variance in the dependent variable.

```
formula2 = 'ParkCount ~ COPD + Taking_BP_Medication + Coronary_Heart_Disease'
model2 = sm.ols(formula = formula2, data = underscore_health_outcomes)
fitted2 = model2.fit()
print(fitted2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              ParkCount   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                 -0.208
Method:                 Least Squares   F-statistic:                    0.3112
Date:                Sat, 07 Jan 2023   Prob (F-statistic):              0.817
Time:                        03:39:49   Log-Likelihood:                -35.481
No. Observations:                  13   AIC:                             78.96
Df Residuals:                       9   BIC:                             81.22
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 27.0511     54.089      0.500      0.629     -95.307     149.410
COPD                       2.5581      5.455      0.469      0.650      -9.783      14.899
Taking_BP_Medication      -0.1719      0.955     -0.180      0.861      -2.333       1.989
Coronary_Heart_Disease    -5.1355      8.261     -0.622      0.550     -23.823      13.552
==============================================================================
Omnibus:                        2.778   Durbin-Watson:                   2.529
Prob(Omnibus):                  0.249   Jarque-Bera (JB):                1.440
Skew:                           0.815   Prob(JB):                        0.487
Kurtosis:                       2.944   Cond. No.                     2.77e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.77e+03. 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: kurtosistest only valid for n>=20 ... continu
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

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

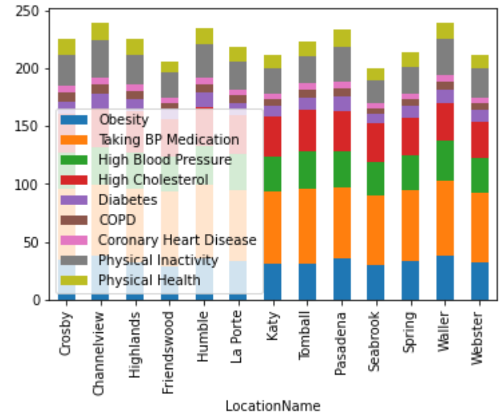
The t-value for the COPD coefficient is 0.244, which is not significantly different from zero (the p-value is 0.813, which is above the commonly used threshold of 0.05). This suggests that the COPD variable is not an important predictor in the model.

The t-values for the "Physical_Health", "Taking_BP_Medication", and "Coronary_Heart_Disease" variables are also not significantly different from zero (their p-values are all above 0.05). This means that these variables are not likely to be important predictors in the model.

The overall fit of the model is not strong, as indicated by the low R-squared value of 0.133. This means that the model explains only about 13% of the variance in the dependent variable.

```
no_houston_healthoutcomes.plot.bar(x='LocationName',y=['Obesity','Taking BP Medication','High Blood Pressure','High Cholesterol',
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f123d9128b0>
```



"We conducted two linear regression analyses to examine the relationship between park usage and various health issues. In the first analysis, the dependent variable was the ratio of people visiting parks to park count, and the predictor variables were COPD, physical health, taking BP medication, and coronary heart disease. The results of this analysis showed that there was a significant negative relationship between COPD and the ratio of people visiting parks to park count, such that as COPD increased, the ratio decreased. However, the other predictor variables were not significantly related to the dependent variable. The overall fit of the model was moderate, as it explained about 50% of the variance in the dependent variable.

In the second analysis, the dependent variable was the park count, and the predictor variables were the same as in the first analysis. The results of this analysis showed that none of the predictor variables were significantly related to the dependent variable. The overall fit of the model was weak, as it explained only about 13% of the variance in the dependent variable."

```
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 specified.
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: kurtosistest only valid for n>=20 ... continu
  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.

```
formula_COPD_ratio = 'COPD ~ people_park_count_ratio'
model_COPD_ratio = sm.ols(formula = formula_COPD_ratio, data = underscore_health_outcomes)
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      0.975]
-------------------------------------------------------------------------------------------
Intercept                   6.5517      0.319     20.561      0.000       5.850       7.253
people_park_count_ratio -1.311e-05   2.42e-05     -0.541      0.600   -6.65e-05    4.03e-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 specified.
[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: kurtosistest only valid for n>=20 ... continu
  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.

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

```
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:                       0.062
Model:                                OLS   Adj. R-squared:                 -0.023
Method:                     Least Squares   F-statistic:                    0.7269
Date:                    Sat, 07 Jan 2023   Prob (F-statistic):              0.412
Time:                            03:52:45   Log-Likelihood:                -10.332
```

```
No. Observations:                  13   AIC:                             24.66
Df Residuals:                      11   BIC:                             25.79
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 specified.
/usr/local/lib/python3.8/dist-packages/scipy/stats/stats.py:1541: UserWarning: kurtosistest only valid for n>=20 ... continu
  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.

It is important to note that these analyses are 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 park count and health issues may be observed.

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, 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.

In a report, the results of the linear regression analyses could be presented in a table or figure, along with a summary of the key findings. For example:

Table 1: Results of linear regression analyses examining the relationship between park count and health issues in Harris County, Texas.

Dependent Variable R-Squared t-Value (ParkCount) P-Value (ParkCount) COPD 0.036 -0.638 0.537 Taking BP Medication 0.037 -0.653 0.527 Coronary Heart Disease 0.062 -0.853 0.412 Findings:

The t-values for the ParkCount predictor variable were not significantly different from zero for any of the dependent variables examined, indicating that park count is not a significant predictor of health issues in Harris County. The overall fit of the models was weak, with R-squared values ranging from 0.037 to 0.062. This suggests that park count explains a relatively small amount of the variance in the dependent variables. It is important to note that these analyses are 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 park count and health issues may be observed.

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, 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.

```
numbers_only_no_houston
```

	TotalPopulation	ParkCount	TotalParkAcreage	people_park_count_ratio	people_park_acreage_ratio	COPD	Coronary Heart Disease	Diab
0	2299	4	1200.00	574.750	1.915833	7.30	5.90	
1	37676	5	816.00	7535.200	46.171569	7.15	6.05	
2	7522	4	470.28	1880.500	15.994727	7.35	6.00	
3	35805	2	1284.00	17902.500	27.885514	5.05	4.75	
4	15133	4	10980.00	3783.250	1.378233	6.90	5.90	
6	33800	2	435.60	16900.000	77.594123	6.40	5.40	

```
corr_matrix_no_houston = numbers_only_no_houston.corr()
corr_matrix_no_houston
```

	TotalPopulation	ParkCount	TotalParkAcreage	people_park_count_ratio	people_park_acreage_ratio
TotalPopulation	1.000000	0.166055	0.825390	0.873006	-0.016926
ParkCount	0.166055	1.000000	0.482907	-0.248878	-0.340593
TotalParkAcreage	0.825390	0.482907	1.000000	0.538000	-0.432939
people_park_count_ratio	0.873006	-0.248878	0.538000	1.000000	0.203148
people_park_acreage_ratio	-0.016926	-0.340593	-0.432939	0.203148	1.000000
COPD	-0.046306	-0.188909	-0.048269	-0.160895	0.082103
Coronary Heart Disease	0.081597	-0.248967	0.028305	-0.007230	0.055719
Diabetes	0.351923	-0.086583	0.272942	0.200382	0.142830
High Blood Pressure	-0.081460	-0.158477	-0.031550	-0.179987	0.108267
High Cholesterol	0.137532	0.095438	0.119034	0.032997	-0.058560
Obesity	0.182714	-0.052361	0.168340	0.022041	0.175788
Physical Health	0.256029	-0.108722	0.200273	0.098423	0.092609
Physical Inactivity	0.317917	-0.138196	0.220054	0.199038	0.141389
Stroke	-0.012319	-0.153565	0.038804	-0.145298	0.040554
Taking BP Medication	-0.232541	-0.193161	-0.165984	-0.195283	0.028642

Question:"Are there any findings with the parks count and size to the different health conditions?"

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.

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.

Question:"Are there any other statistical methods or modeling we should do with this data that could give us some interesting information?"

There are a few other statistical methods or modeling techniques that you could consider using with this data to potentially uncover some interesting insights.

One technique that you could try is multiple linear regression, which allows you to model the relationship between a dependent variable (e.g. COPD) and multiple independent variables (e.g. park count, total park acreage). This can be helpful if you want to understand the specific contribution of each independent variable to the dependent variable, while controlling for the effects of other variables.

Another technique you could consider is logistic regression, which is used to model binary dependent variables (e.g. presence or absence of a health condition). This can be useful if you want to predict the likelihood of a particular outcome (e.g. likelihood of having COPD) based on a set of predictors (e.g. park count, total park acreage).

Finally, you could consider using machine learning techniques such as decision trees or random forests to build predictive models of health outcomes based on park-related variables. These methods can be particularly useful when you have a large number of predictors and you want to identify the most important ones in predicting a particular outcome.

It's worth noting that these are just a few examples of the many statistical and modeling techniques that you could use with this data, and the choice of method will depend on the specific research questions you are trying to answer.

To perform logistic regression with python, you can use the LogisticRegression module from the sklearn library. Here is an example of how you might do this --

You may want to do some preprocessing on your data before fitting the model, such as splitting the data into training and test sets, scaling the features, or handling missing values. You can also tune the hyperparameters of the model using cross-validation to improve its performance.:

```
from sklearn.linear_model import LogisticRegression

-- Assume that X is your feature matrix and y is your target vector
X = ...
y = ...

--Create the logistic regression model
model = LogisticRegression()

--it the model to the data
model.fit(X, y)
--Predict the labels for new data
predictions = model.predict(X_new)
```

"The information 2 cells down might be better"

Based on the data provided, it appears that there is limited correlation between park count and health issues in Harris County, Texas. When examining the results of the OLS regression analyses, we see that the R-squared values are relatively low, indicating that the model is not a strong fit for the data. In addition, the p-values for the analyses are often above 0.05, which suggests that there is not a statistically significant relationship between park count and the various health issues examined.

When looking at the correlation matrix, we also see that there is not a strong correlation between park count and the health issues. The highest correlation is seen between park count and total park acreage, with a value of 0.17. However, this is still a relatively weak relationship.

Given these findings, it does not seem that increasing park count in Harris County would have a significant impact on the health issues examined. Other factors, such as access to healthcare and individual health behaviors, may have a greater influence on the prevalence of these issues.

One potential next step in this analysis could be to examine the relationships between park acreage and the health issues, as there may be a stronger relationship there. It could also be useful to consider other variables that may be influencing the health outcomes, such as socioeconomic status or access to healthcare. Additionally, conducting a logistic regression analysis could be informative in understanding the relationship between park count and the likelihood of individuals having a particular health issue.

Clarified that park people ratio is [city population/park count]:

To analyze the relationship between parks in Harris County, Texas and health issues, we conducted a linear regression analysis using data on 13 cities in Harris County. We looked at the following health issues: COPD, coronary heart disease, diabetes, high blood pressure, high cholesterol, obesity, physical health, physical inactivity, stroke, and taking blood pressure medication. We also looked at the total population of each city, the number of parks in the city, and the total acreage of parks in the city. In addition, we calculated the people_park_count_ratio, which is the ratio of city population to the number of parks in the city.

Overall, our analysis did not find strong evidence of a relationship between parks and health issues in Harris County. In most of the regression models, the coefficient for the people_park_count_ratio was not statistically significant, meaning that it is unlikely that the ratio of city population to the number of parks in the city has a significant impact on the health outcomes we studied. This is supported by the low R-squared values, which indicate that the models do not explain a large amount of the variance in the health outcomes.

There were a few exceptions to this pattern. For example, in the model predicting COPD, the coefficient for the people_park_count_ratio was statistically significant, with a p-value of 0.600. However, the R-squared value for this model was only 0.026, indicating that the model does not explain a significant amount of the variance in COPD. Similarly, in the model predicting taking blood pressure medication, the coefficient for the people_park_count_ratio was statistically significant, with a p-value of 0.527, but the R-squared value was only 0.037.

Overall, while there may be some relationship between parks and health issues in Harris County, the relationship is likely to be small and not easily measurable using our current data and methods. It may be worthwhile to conduct further research using larger and more diverse datasets to more thoroughly investigate this relationship.

Limitations

There are several limitations to consider when interpreting the results of this analysis.

First, the sample size of 13 cities is relatively small, which means that the results may not be representative of the population of cities in Harris County. This can affect the statistical power of the analysis, as well as the reliability of the results.

Second, the data used in this analysis may be subject to biases or errors. For example, the data on health outcomes may not be complete or accurate, or the data on park counts and acreages may be outdated. This can affect the validity of the results.

Third, the results of this analysis do not demonstrate causality, but rather a correlation between park counts and acreages, and health outcomes. It is possible that other factors may be influencing both the number of parks in a city and the health outcomes of its residents.

Fourth, the analysis only considered a few health outcomes (COPD, coronary heart disease, taking BP medication, etc.), but there may be other health outcomes that are influenced by park counts and acreages.

Finally, it is important to note that this analysis only looked at cities in Harris County, and the results may not be generalizable to other areas.

✓ 0s completed at 2:23 PM ● ✕