The Impacts of Urban Environments on Individual Well-being

Amanda Chen, Joanna Chen, Lily Feng, Vanessa Tam

Table of Contents

Introduction

Our team is intrigued by the impact of living in urban regions on various aspects of individuals' well-being and quality of life. We believe that the environment in which people reside plays a pivotal role in shaping their physical health, social interactions, and overall satisfaction with life. In our exploration, we aim to shed light on the multifaceted dynamics of urban living in comparison to non-urban areas across the United States.

Our research journey originated from a broader interest in regional disparities, prompting us to focus specifically on the contrasting experiences between urban and non-urban environments. Recognizing the significance of where people live and its potential influence on their daily lives, we formulated hypotheses and corresponding analyses to delve into key factors affecting well-being and quality of life.

Research Questions

What specific characteristics differentiate between urban and non-urban (suburban or rural) regions within cities, and how do these distinctions impact variations in well-being and quality of life among residents?

To answer this question, we use the following research questions (in order of appearance in this .ipynb file):

1. Does residing in urban regions increase residents' physical well-being and quality of life compared to non-urban areas?
    - Does the walkability of a region correlate with an increase in quality of life, as measured by physical activity levels (walkability score)?
    - Do variations in the distribution of commute modes (e.g., walking, bicycling, transit) correspond to differences in the well-being and quality of life among

residents?

Data Descriptions

All raw datasets can be found at this Google Drive folder link: https://drive.google.com/drive/folders/1VY3g8Pz9CxpGMqkxF-mwumK5PgyNjPXB?usp=sharing

SMART LOCATION DATABASE:

What are the observations (rows) and the attributes (columns)?

The observations (rows) correspond to different counties.

The attributes (columns) are as follows:

- COUNTYFP: County FIPS code
- Ac_Land: Land area (acres)
- Ac_Unpr: Land area that is not protected from development
- P_WrkAge: Percentage of the population that is working-aged (19 to 64 years)
- AutoOwn0: Number of households in the Census Block Group (CBG) that own 0 automobiles
- Pct_AO0: Percentage of zero-car households in CBG (2018)
- Pct_AO2p: Percentage of households with 2 or more cars in CBG (2018)
- D1A: Gross residential density (Housing Units per acre) on unprotected land
- D1B: Gross population density (Housing Units per acre) on unprotected land
- D1C: Gross employment density (Housing Units per acre) on unprotected land
- D2A_JPHH: Jobs per household
- D3A: Total road network density
- D4A: Distance from the population-weighted centroid to the nearest transit stop (meters)
- D4C: Aggregate frequency of transit service within 0.25 miles of the CBG boundary per hour during the evening peak period
- D5AR: Jobs within a 45-minute auto travel time, time-decay (network travel time) weighted
- D5CR: Proportional accessibility to regional destinations by car: Employment accessibility expressed as a ratio of total CBSA (Core-Based Statistical Area) accessibility.

Funded by the United States Environmental Protection Agency, the Smart Location Database was created in order to aggregate several different factors (demographics, employment, and build environment variables) for every census block (CBG) in the US, variables that are important in transportation research and behavior. It's useful for scenario planning studies and "characterizing the relative location efficiency of CBGs within US metropolitan areas."

The EPA received data for the smart location database with coverage limited to metropolitan regions served by transit agencies that share their service data in a standard format called GTFS. This means that any metropolitan areas not covered by such transit

agencies are not included in the data, and thus the dataset cannot be considered comprehensive of the entire geographic area in which the data was collected.

Preprocessing included summarizing and organizing the demographic, employment and environment variables for each CBG. These variables are transformed into indicators of the commonly cited "D" variables associated with transportation behavior, including density, land use diversity, built environment design, access to destinations, and distance to transit.

As a census, people were surveyed on a large scale but it is unclear of the extent of the survey transparency. According to the EPA, they used: "Geographic Information System (GIS) software and obtaining the latest Census data (such as through https://www.census.gov/programs-surveys/geography.html). In addition, the mapping application Mapping Applications for Response, Planning, and Local Operational Tasks (MARPLOT), from the Computer-Aided Management of Emergency Operations (CAMEO) software suite, and the Circular Area Profiles application (https://mcdc.missouri.edu/applications/caps2010.html) ."

Link to source site: https://catalog.data.gov/dataset/smart-location-database1

PHYSICAL ACTIVITY:

Columns: City name and state, walkability score (given the Stanford study's walkability index.)

This national dataset includes walkability scores for all block groups as well as the underlying attributes that are used to rank the block groups. Each row is a city in the United States and its walkability score. The dataset ranks block groups according to their relative walkability. The raw data was provided by Azumio, and the data process was done by a research team at Stanford University.

The processes that influence what data was observed and recorded include the city selection; this collection involve selecting cities that meet certain criteria, specifically cities with at least 20,000 weekdays of data. This suggests that the data selection process is influenced by a threshold for the amount of data available for analysis. Other attributes related to the cities or more cities may not be included in this particular analysis.

The preprocessing here involves sorting and ranking cities based on their walkability scores. The dataset appears to have been sorted according to the city's walkability score and presented in a tabular form with city names and their respective walkability scores.

The data was collected by the researchers through the subjects' smartphones via the Azumio Argus app, measuring distance traveled, steps walked, and basic biometric data for each person. Subjects were aware of the data collection and purpose.

Link to source site: https://cs.stanford.edu/people/jure/pubs/activity-inequality-nature17.pdf

LOW AND MODERATE INCOME AREAS:

Columns: State abbreviation, County name, number of residences considered low income, number of residences considered low to moderate income, proportion/percentage of total population considered to be low to moderate income in the block

This dataset provides information on the U.S. Housing and Urban Development's (HUD) low to moderate income areas. The observations are census block groups across the United States, and the attributes include the geographic location of the census block group, the number and percentage of low to moderate income population in the census block group, and the area of the census block group. In the Community Development Block Grant (CDBG) program, at least 70% of funds, chosen over 1-3 years, must benefit low- to moderate-income individuals, referred to as "low-mod." The dataset was created to visualize the distribution of low to moderate populations, thus helps identifying the “low-mod” communities to provide benefits to the groups.

The dataset was created by the Department of Housing and Urban Development, which provides data on U.S. housing and urban communities.

The data collection would focus on identifying areas where at least 51% of households have incomes at or below 80% of the area median income (AMI). Data regarding the number and percentage of low to moderate-income population in census block groups aligns with these programmatic criteria but data unrelated to these specific income thresholds or programmatic goals may not be included.

Preprocessing involves counting the low to moderate-income individuals within each census block group as well as determining the total population in each block group. This information is then used to calculate the percentage of low to moderate-income individuals in each area.

The data description does not explicitly mention whether people within these census block groups were aware of the data collection. However, residents and community organizations in these areas that benefit from CDBG funds would likely be aware of the data collection efforts. They would expect that the data is used for the purpose of identifying and allocating resources to communities that meet the "low-mod" criteria. The data helps ensure that the benefits provided through the CDBG program are directed to areas that serve low to moderate-income individuals and households.

Link to source site: https://catalog.data.gov/dataset/hud-low-and-moderate-income-areas

MEANS OF TRANSPORTATION TO WORK:

Columns: Mode of transportation, all other columns are estimated population count under specified state name (Including District of Columbia and Puerto Rico)

This dataset was conducted as a part of the American Community Survey by the US Census Bureau in 2022, looking at the distribution of the modes of transportation taken to commute to work by workers ages 16 and older in the United States.

Since the method of the data collection is that of a census survey, there is room for sampling variability due to certain people abstaining from participating, which translates to the 90 percent margin of error, as well as the website saying that the ACS estimates are subject to nonsampling error.

Preprocessing was accomplished through the recording of data from a sample size smaller than the original amount of potential subjects contacted (https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-

size/index.php), with different levels of response or response behavior also recorded and accounted for in margin of error (https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/index.php). It appears for there to be minimal absence of coverage for US states surveyed in 2022, with high percentages of participation generally: https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/coverage-rates/index.php

As a project of the US Census Bureau, participants were directly contacted as a part of the data collection process and understood the nature and purpose of the data collection as documentation under the government agency.

Link to source site: https://data.census.gov/table/ACSDT1Y2022.B08301?q=b08301&g=010XX00US$0400000&y=2022

URBAN POPULATION BY STATE:

Columns: State code/index, State abbreviation, state name, 2020 total population count, 2020 population count living in urban areas, 2020 percentage of population living in urban areas

This dataset provides census information for urban and rural areas at the state-level in the United States. It includes total populations, urban populations, and rural populations of each state for the 2020 census and 2010 census. The dataset was created by The Census Bureau to define urban and rural areas following each ten-year census by applying specific criteria to census and related data.

For preprocessing, this dataset has been aggregated and anonymized to protect the confidentiality of the individuals due to the Census Bureau's legal mandate. The dataset has also been aggregated to different geographic levels to create summary tables suitable for its applications. In cases where individual responses are missing or incomplete, the Census Bureau may use statistical techniques to estimate missing values to maintain data completeness and accuracy.

The individuals involved were aware of the data collection and are aware that this data will be used to inform the federal government, businesses, and other decision makers with things like funding distribution and political power.

Link to source site: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

TRANSPORTATION AND HEALTH TOOL DATA:

Columns: state name, commute mode raw scores based on mode of transportation, miles of travel for each mode of transportation (Refer to specific indicator profiles and the meaning of raw score here: https://www7.transportation.gov/mission/health/indicator-profiles. Will elaborate and clean further in Phase III)

The Transportation and Health Tool (THT) data was developed by the U.S. Department of Transportation and the Centers for Disease Control and Prevention to provide accessible data for practitioners to examine how transportation systems impact health. It offers information on transportation and public health indicators (eg. Commute Mode Shares,

Public Transportation Trips per Capita, Vehicle Miles Traveled (VMT) per Capita, and etc.) for U.S. states and metropolitan areas.

The process of indicator selection could have influenced the data recorded. According to the data provider, each indicator was selected by the CDC, USDOT, and the American Public Health Association (APHA) and then combined, so it's possible both for data to be left out and for there to be inconsistencies that they needed to smooth over.

Preprocessing: for each indicator, the dataset includes both the raw score and a calculated score (range from 0 to 100) for easier comparison. The process of data collection doesn't involve people, so no human expectation that might affect the data.

Link to source site: https://www7.transportation.gov/mission/health/transportation-and-health-tool-data-excel

QUALITY OF LIFE BY STATE:

Columns: State name, aggregate score across all score categories, QOL score, affordability score, economy score, education and health score, and safety score

The dataset ranks the states in the US for the quality of life, determined by both material factors and mental and physical considerations. The quality of life in this dataset is converted to a score that is calculated from factors that might affect the quality of life, including healthcare, education, economy, infrastructure, opportunity, fiscal stability, crime and corrections, and the natural environment. The dataset is created by worldpopulationreview.com, where it includes a wide range of data that summarizes aspects of populations worldwide. Topics include economics, health, environment, etc.

In terms of processing, all the quality of life variables have values ranked on a scale from 1 to 50, with 50 being the best possible score in the category (eg., income growth, education rate, and other "life metrics"). This is a scale of rating that was determined by World Population Review, and the factors considered to produce each score is unknown to those who don't read their methodology, and it's possible for nuance to be lost when different considerations are combined into a single numerical score.

Preprocessing for the quality of life indicators includes evaluating through all 50 states across 51 indicators, ranging from housing costs and income growth to the education rate and quality of hospitals. However, the ranking for the remaining variables remain unclear, which might influence the data in terms of how different factors are combined to create the overall total score.

Link to source site: https://worldpopulationreview.com/state-rankings/quality-of-life-by-state

URBAN POPULATION BY COUNTY:

Columns: State name, population count, population density, Housing unit density, population count living in urban areas, population percentage living in urban areas, overall percentage of urban area in the specified region, and population density in the urban area

This dataset provides census information for urban and rural areas at the county-level in the United States. It includes total populations, urban populations, and rural populations of each state for the 2020 census and 2010 census. The dataset was created by The Census Bureau to define urban and rural areas following each ten-year census by applying specific criteria to census and related data. This data collection process includes mail in forms and in-person interviews. Since the census is mandatory for all people living in the US, it ensures that the data collected is completed.

The processes employed by the US Census Bureau are the same as mentioned earlier for the [Urban population by state] dataset, and is subject to the same risks, margin of error, and level of sample variability.

For preprocessing, this dataset has been aggregated and anonymized to protect the confidentiality of the individuals due to the Census Bureau's legal mandate. The dataset has also been aggregated to different geographic levels to create summary tables suitable for its applications.

In cases where individual responses are missing or incomplete, the Census Bureau may use statistical techniques to estimate missing values to maintain data completeness and accuracy. The individuals involved were aware of the data collection and are aware that this data will be used to inform the federal government, businesses, and other decision makers with things like funding distribution and political power.

Link to source site: https://www.google.com/url?q=https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html&sa=D&source=docs&ust=1697771335144738&usg=AOvVaw0QWaCzfVguck8YdL3N

Preregistration statements

We pre-registered the following three analyses:

(1) Hypothesis: We expect that living in urban regions is more conducive to residents' quality of life, as measured by physical activity levels, walkability score, and proportion of various modes of transportation used to commute to work (referred throughout this report as 'commute mode'), in comparison to non-urban areas.

We believe that having a variety of options for transportation, which may be more conducive to residents' quality, is seen more in urbanized areas, since surburban areas are categorized by more spread out (lower population density) residential and commercial districts that heavily favor travel by car, with high road density. Thus, it is both unsafe if one were to travel by walkin due to proximity to roads/highways, or inconvenient by public transportation, and without sufficient accommodation for such other modes of transportation as seen in large cities, such as sidewalks and bike lanes.

Analysis on how quality of life and physical well-being differ in each state based on the percentage of urban population. We will input the the percentage of urban population of each state) and output the score for quality of life assoiciated with the state (QOL total score) and physical well-being (walkabiltiy and mode of transportation) associared with the state. This approach will allow us to examine the correlation between level of ubranization and these physcial well-being and quality of life measure.

(2) Hypothesis: We expect non-urban regions to be more likely to lead to geographic isolation.

Analysis on how geographic isoaltion differ in each state based on the percentage of urban population. Conduct linear regressions and scatter plots to examine the relationship between the level of urbanization and the variables of population density, commuting/transportation distance, and time it takes to commute to work for each state. Run linear regression to examine the relationship betweent the level of urbanization, physical mode of transpotation ,

(3) Hypothesis: We expect regions characterized by lower income levels are correlated with a diminished quality of life.

Analysis on how income levels differ in each state based on the quality of life. Run a linear regression where we input the income level of areas (dummy variables, inputted as the proportion of low to moderate income population within the census block group) and output the score for quality of life associated with the area (on the same county level). Both input and output are non-negative numbers, so we will test whether βincome > 0.

Data Collection & Cleaning

In [134…
```
import matplotlib.pyplot as plt
import matplotlib.ticker
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm

from scipy.stats import ttest_ind, binom, poisson, norm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from numpy import random
import scipy.stats as stats
```

In [135…
```
smartlocation_df_cleaned = pd.read_csv('cleaning/cleaned_smart_location.csv')
low_mod_income_df = pd.read_csv('cleaning/cleaned_low_mod_income.csv')
transportations_to_work_df = pd.read_csv('cleaning/cleaned_transportations_to_
qol_df = pd.read_csv('cleaning/cleaned_qol.csv')
urban_pop_df = pd.read_csv('cleaning/cleaned_urban_pop.csv')
urban_county_df = pd.read_csv('cleaning/cleaned_urban_county.csv')
physical_activity_df = pd.read_csv('cleaning/cleaned_physical_activity.csv')
transportation_health_df = pd.read_csv('cleaning/cleaned_transportation_health
```

```
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3797646237.p
y:2: DtypeWarning: Columns (2,3) have mixed types. Specify dtype option on imp
ort or set low_memory=False.
  low_mod_income_df = pd.read_csv('cleaning/cleaned_low_mod_income.csv')
```

Note: For any questions where we used QualityofLife (QOL) variable, please note that the QOL represents a 1 - 50 ranking of the states, 1 being the state with the highest QOL score, and it is not a numerical score. In this context, a smaller number signifies a better quality of life.

Similarly, the 'TotalScore' variable from the Quality of Life dataset (qol_df) refers to an aggregation of the different factors considered by the dataset (refer to data description for this dataset as mentioned above.) We use this definition whenever we refer to "quality of life" throughout this project.

Exploratory Data Analysis

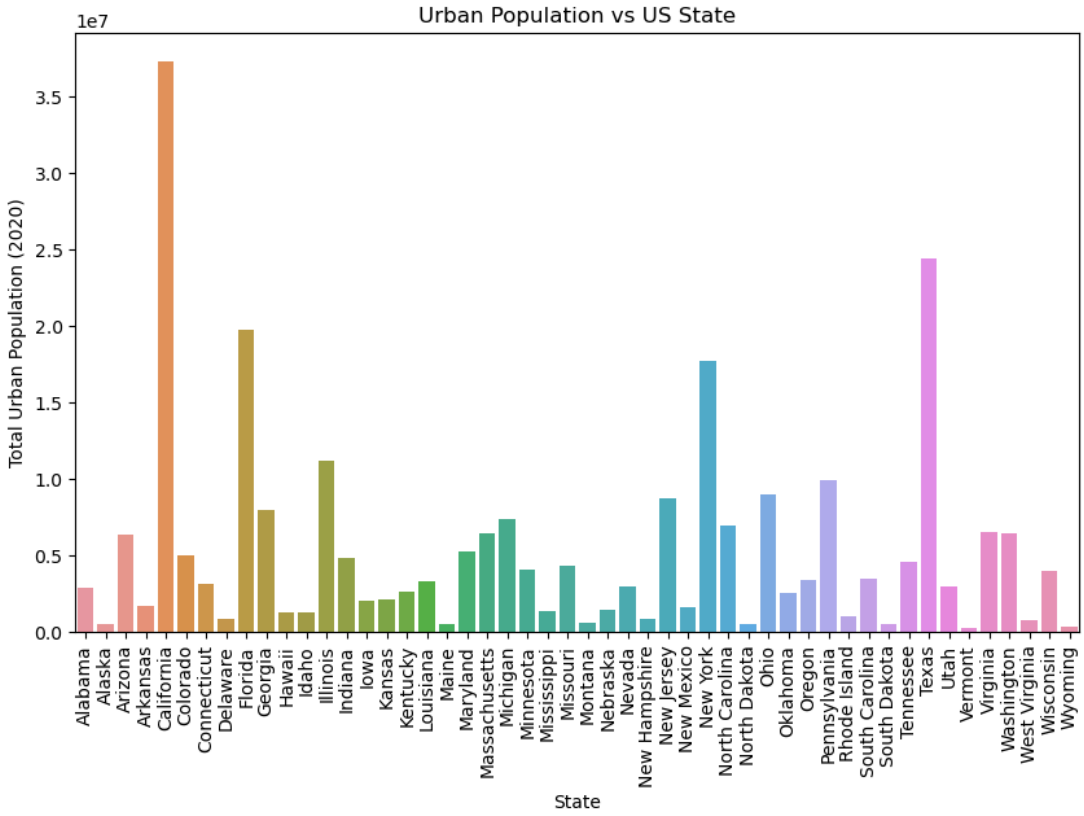
Descriptive statistics: Average road network density and average percentage of households without cars based on the county for the smartlocation_df_cleaned:

In [136…
```
data = {
    'Average road network density': smartlocation_df_cleaned['D3A'].mean(),
    'Average % households without cars': smartlocation_df_cleaned['Pct_AO0'].m
}

smart_location_avg_df = pd.DataFrame(data, index=[0])
print(smart_location_avg_df)
```

```
   Average road network density  Average % households without cars
0                  61215.662834                           0.087256
```

In [137…
```
urban_state_df = pd.merge(urban_pop_df, qol_df, \
        left_on='STATE NAME', right_on='state')

plt.figure(figsize=(10, 6))
sns.barplot(x='STATE NAME', y='2020 URBAN POP', data=urban_state_df)
plt.xlabel('State')
plt.ylabel('Total Urban Population (2020)')
plt.title('Urban Population vs US State')
plt.xticks(rotation=90)
plt.show()
```

```
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
```

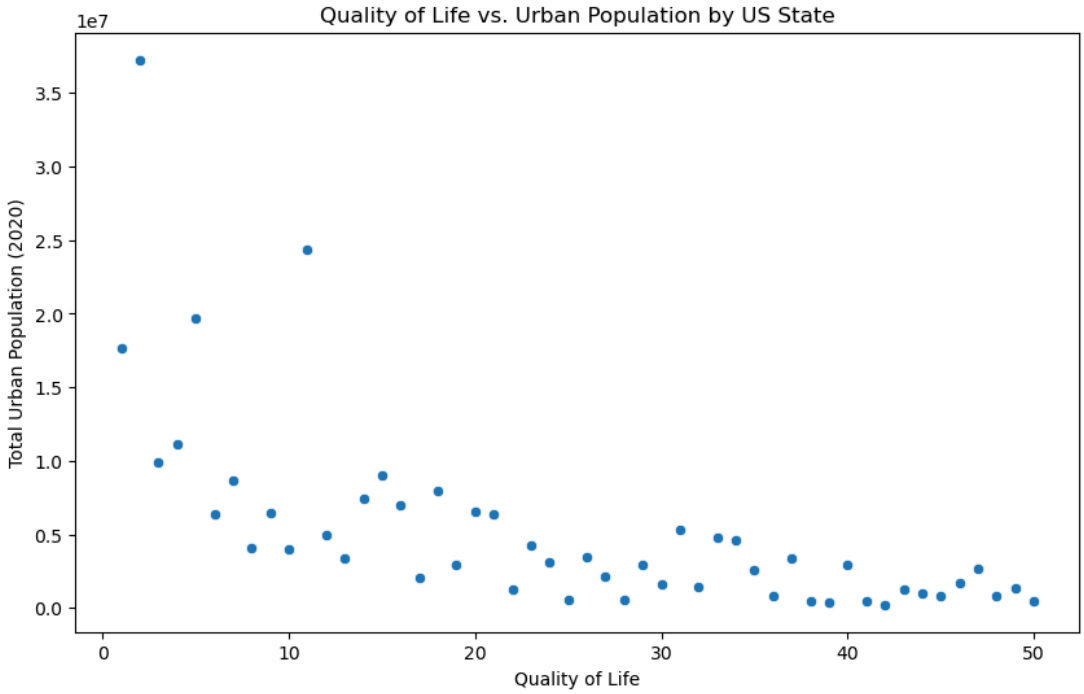


This bar plot displays the relationship between Total Urban Population and States. Though there's no specific correlation we can gather from this graph, it will be helpful for our future data explorations graph when we need to view the urban population of a specific state.

In [138…

```
urban_qol_df = pd.merge(urban_pop_df, qol_df, \
            left_on='STATE NAME', right_on='state')

plt.figure(figsize=(10, 6))
sns.scatterplot(x='QualityOfLife', y='2020 URBAN POP', data=urban_qol_df)
plt.xlabel('Quality of Life')
plt.ylabel('Total Urban Population (2020)')
plt.title('Quality of Life vs. Urban Population by US State')
plt.show()
```

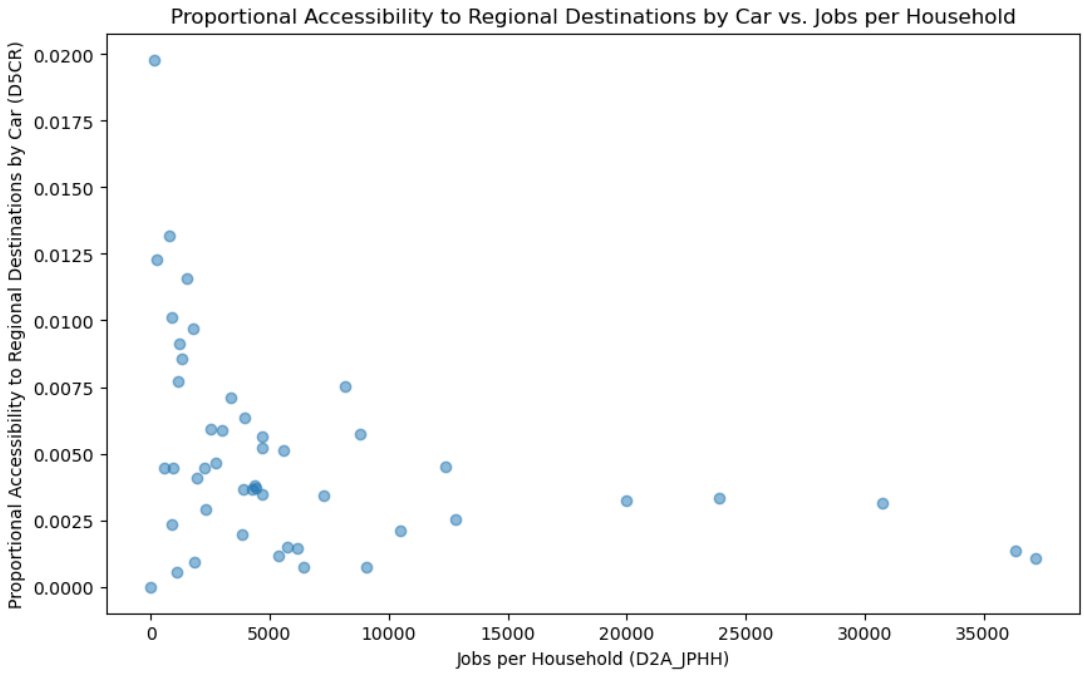
```
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
```



There seems to be a logarithmic relationship between the Quality of Life and the Total Urban Population by State. On the scatterplot, as "Quality of Life" increases, the decrease in "Total Urban Population" is not consistent and becomes more gradual as "Quality of Life" scores increase.

In [139…

```
x = smartlocation_df_cleaned['D2A_JPHH']
y = smartlocation_df_cleaned['D5CR']

plt.figure(figsize=(10, 6))
plt.scatter(x, y, alpha=0.5)

plt.title('Proportional Accessibility to \
Regional Destinations by Car vs. Jobs per Household')
plt.xlabel('Jobs per Household (D2A_JPHH)')
plt.ylabel('Proportional Accessibility to Regional Destinations \
by Car (D5CR)')

plt.show()
```

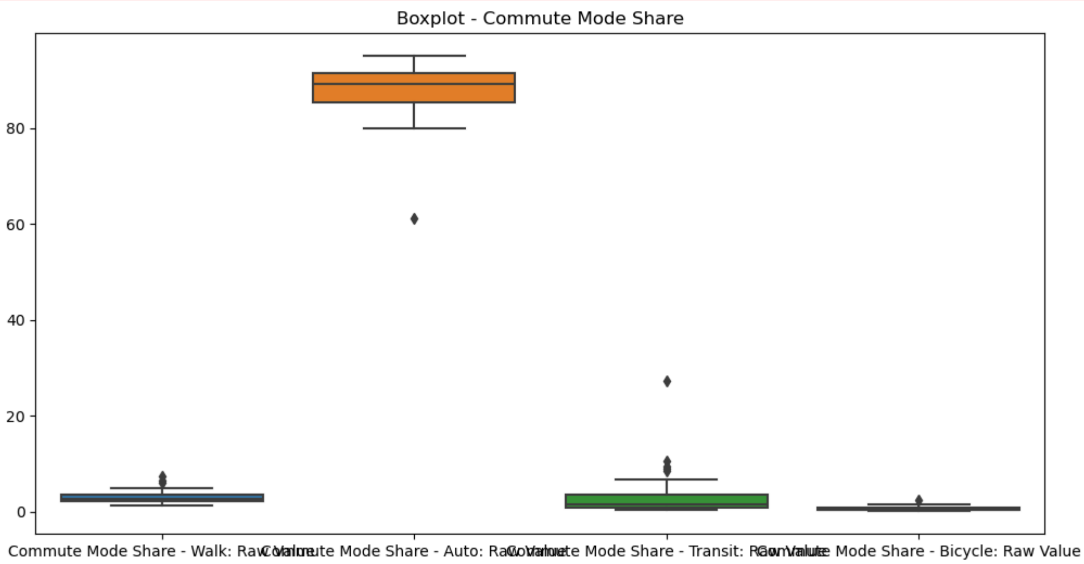


A scatter plot for the relationship between the proportional accessibility to regional destinations by car and the number of jobs per household from smartlocation_df. We expected there to be a potential correlation between jobs per household and regional accessibility but this relationship is not clear enough to strongly pursue in regards to urban and non-urban regions.

In [140…

```
# Clean up column names
transportation_health_df.columns = transportation_health_df.columns.str.strip(


# Merge transportation_health_df and qol_df on the 'State' column
merged_transport_qol_df = pd.merge(transportation_health_df, qol_df, how='inne
                                   left_on='State', right_on='state')


# Define the relevant columns for boxplots
columns_for_boxplots = ['Commute Mode Share - Walk: Raw Value',
                        'Commute Mode Share - Auto: Raw Value',
                        'Commute Mode Share - Transit: Raw Value',
                        'Commute Mode Share - Bicycle: Raw Value']
                       # Replace with your actual column names


# Convert the relevant columns to numeric (if they are not already)
merged_transport_qol_df[columns_for_boxplots] = \
   merged_transport_qol_df[columns_for_boxplots].apply(pd.to_numeric, errors='
transportation_health_df[columns_for_boxplots] = \
   transportation_health_df[columns_for_boxplots].apply(pd.to_numeric, errors=


# Boxplot for Quality of Life DataFrame
plt.figure(figsize=(12, 6))
sns.boxplot(data=merged_transport_qol_df[columns_for_boxplots])
plt.title('Boxplot - Commute Mode Share')
plt.show()
```

```
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
```



This box plot allows us to compare the distribution of values between the four commute modes: Auto, Bicycle, Walking, and Transit. From the boxplot, we can see that the distribution of Auto is drastically different than the remaining three with all of its values being in the higher end (greater than 60) and the distribution of Bicycle, Walking, and Transit is fairly similar. We plan on further analyzing these relationships in Hypothesis 1.

Hypothesis 1

1. Does residing in urban regions increase residents' physical well-being and quality of life compared to non-urban areas?

Analysis: t-test for difference in means

- H_0: There is no significant difference in residents' physical well-being and quality of life between those residing in urban regions and those in non-urban areas. μ(urban_qol) = μ(non_urban_qol)

- H_A: Residing in urban regions is associated with a significant increase in residents' physical well-being and quality of life compared to non-urban areas. μ(urban_qol) < μ(non_urban_qol)

In [141…

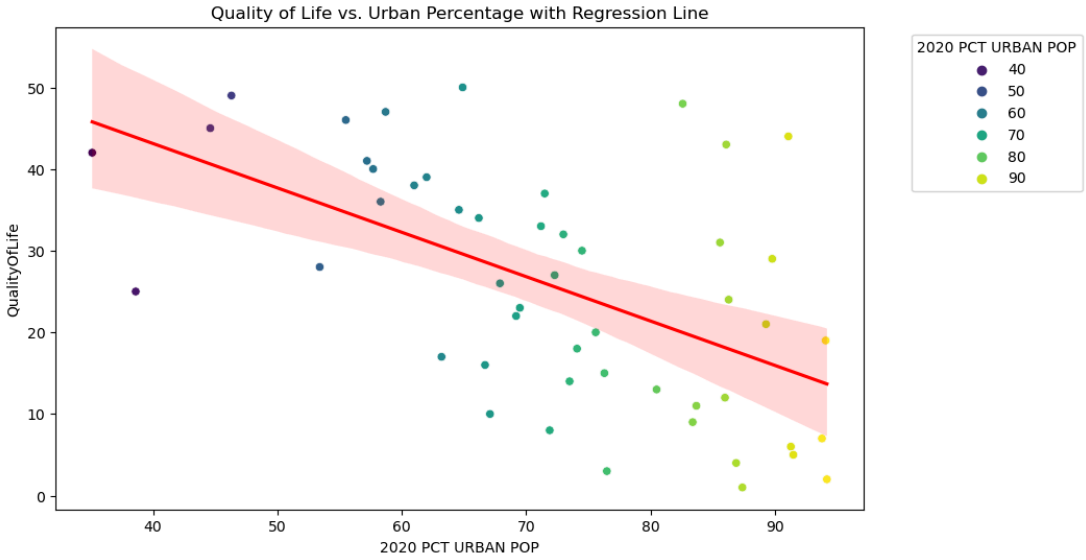
```
# Merge QOL and Urban Population DataFrames on State
merged_df = pd.merge(qol_df, urban_pop_df, how='inner', left_on='state', \
                     right_on='STATE NAME')

# Drop unnecessary columns
columns_to_drop = ['STATE ABBREV', 'STATE NAME', 'Affordability', 'Economy', \
                   'EducationAndHealth', 'Safety', 'STATEFP']
merged_df = merged_df.drop(columns=columns_to_drop, axis=1)

# Display the merged DataFrame
print(merged_df.head())
```

```
           state  TotalScore  QualityOfLife  2020 TOTAL POP  2020 URBAN POP  \
0  Massachusetts       62.65              6         7029917         6416895
1     New Jersey       62.01              7         9288994         8708779
2       New York       60.64              1        20201249        17665166
3          Idaho       58.73             22         1839106         1273437
4       Virginia       58.73             20         8631393         6528313

   2020 PCT URBAN POP
0                91.3
1                93.8
2                87.4
3                69.2
4                75.6
```

In [142…

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=merged_df, x='2020 PCT URBAN POP', y='QualityOfLife', \
                hue='2020 PCT URBAN POP', palette='viridis')
plt.title('Quality of Life vs. Urban Percentage with Regression Line')
plt.xlabel('Percentage of Urban Population')
plt.ylabel('Quality of Life Score')

# Plotting the regression line
sns.regplot(x='2020 PCT URBAN POP', y='QualityOfLife', data=merged_df, scatter
            color='red')
plt.legend(title='2020 PCT URBAN POP', bbox_to_anchor=(1.05, 1), loc='upper le
plt.show()
```

```
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
```



In order to test this hypothesis, we plotted a regression line between quality of life and the 2020 Urban Population. We merged two datasets and ran a scatterplot. This plot shows a negative correlation between quality of life and the 2020 Urban Population Percentages, suggesting that a more urbanized environment might come at the cost of a less desirable quality of life. We see a smaller confidence interval between 60-70%, suggesting a more reliable estimate of the regression line in that range. However, outliers are evident, particularly in instances of a high urban population percentage, such as in the high 90s. But in general, it's safe to assume that with higher urban population percentages, there's a better (lower numerically) quality of life.

In [143…
```
urban_qol = merged_df[merged_df['2020 PCT URBAN POP'] > 50]['QualityOfLife']
non_urban_qol = merged_df[merged_df['2020 PCT URBAN POP'] <= 50]['QualityOfLif

# One-sided t-test (greater) for the alternative hypothesis
t_stat, p_value = stats.ttest_ind(urban_qol, non_urban_qol, alternative='less'

print(f'T-statistic: {t_stat}\nP-value: {p_value}')
```

```
T-statistic: -2.1900410538242174
P-value: 0.016706272106550613
```

When we conducted a one-sided t-test on the percentage of urban population and residents' quality of life, we got a p-value of 0.9832. This is not significant at a 5% significance level, so we fail to reject the null hypothesis. In other words the data we used is not strong enough to claim a statistically significant difference in physical well-being between urban and non-urban residents. While the t-statistic of -2.19 suggests a potential negative association between urban population percentage and well-being, the p-value doesn't provide enough confidence to definitively say that one group has lower well-being than the other. The analysis suggests that there might be a trend of lower well-being in more urban areas, but the evidence is inconclusive.

1.1 Does the walkability of a region correlate with an increase in quality of life, as measured by physical activity levels (walkability score)?

Analysis: OLS Regression

- H_0: There is no correlation between the walkability of a region and the quality of life, as measured by physical activity levels (walkability score). β(walkability_qol)=0
- H_A: There is a correlation between the walkability of a region and the quality of life, as measured by physical activity levels (walkability score). β(walkability_qol)≠0

In [144…

```
# Create a mapping dictionary to convert abbreviated states to full names
state_mapping = {
    'NY': 'New York',
    'NJ': 'New Jersey',
    'CA': 'California',
    'MA': 'Massachusetts',
    'PA': 'Pennsylvania',
    'FL': 'Florida',
    'IL': 'Illinois',
    'DC': 'District of Columbia',
    'WA': 'Washington',
    'VA': 'Virginia',
    'MD': 'Maryland',
    'MN': 'Minnesota',
    'TX': 'Texas',
    'OR': 'Oregon',
    'HI': 'Hawaii',
    'MO': 'Missouri',
    'WI': 'Wisconsin',
    'OH': 'Ohio',
    'LA': 'Louisiana',
    'CO': 'Colorado',
    'GA': 'Georgia',
    'CA': 'California',
    'AZ': 'Arizona',
    'NE': 'Nebraska',
    'NV': 'Nevada',
    'IA': 'Iowa',
    'AK': 'Alaska',
    'VT': 'Vermont',
    'WY': 'Wyoming',
    'PA': 'Pennsylvania',
    'IL': 'Illinois',
    'VA': 'Virginia',
    'CA': 'California',
    'WI': 'Wisconsin',
    'OH': 'Ohio',
    'FL': 'Florida',
    'TX': 'Texas',
    'CA': 'California',
    'CA': 'California',
    'WI': 'Wisconsin',
    'FL': 'Florida',
    'GA': 'Georgia',
    'TX': 'Texas',
    'CA': 'California',
    'CA': 'California',
    'WI': 'Wisconsin',
    'FL': 'Florida',
    'AZ': 'Arizona',
    'NM': 'New Mexico',
    'FL': 'Florida',
    'AZ': 'Arizona',
    'TX': 'Texas',
    'CA': 'California',
    'CO': 'Colorado',
    'MO': 'Missouri',
    'TX': 'Texas',
    'OK': 'Oklahoma',
```

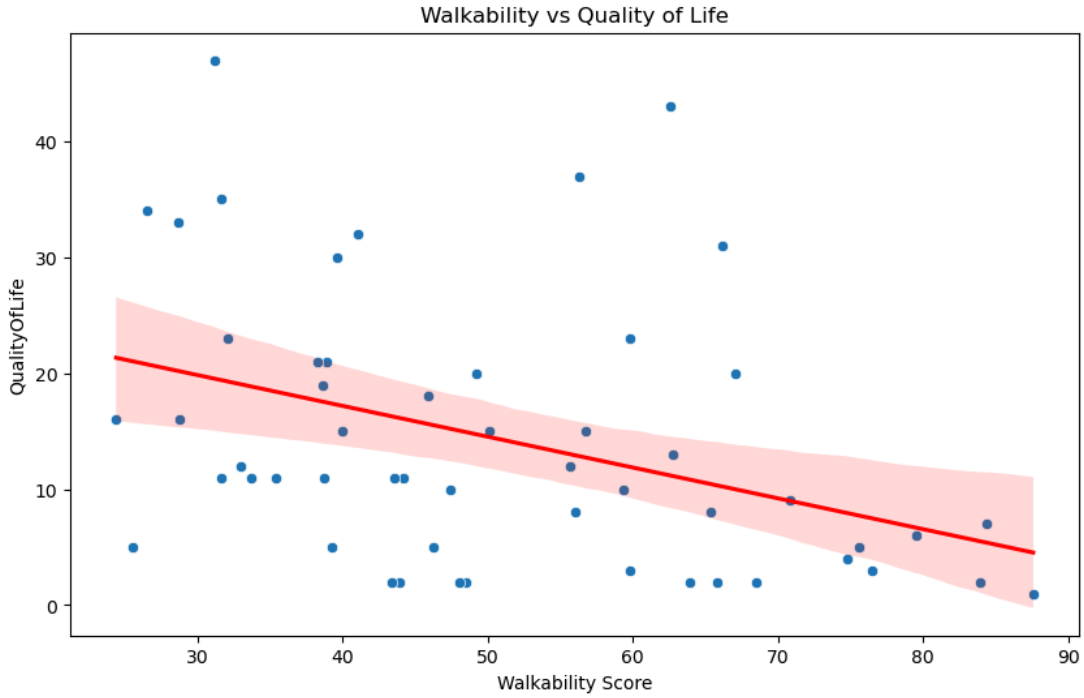
```
    'KY': 'Kentucky',
    'NC': 'North Carolina',
    'IN': 'Indiana',
    'TN': 'Tennessee',
    'FL': 'Florida',
    'NC': 'North Carolina'
}

# Convert abbreviated states to full names in the 'State' column of
# physical_activity_df
physical_activity_df['State'] = \
    physical_activity_df['City'].str.split(',').str[1].str.strip().map(state_m

# Merge datasets on the state
merged_df = pd.merge(physical_activity_df, qol_df, how='inner', left_on='State
                     right_on='state')

# Drop unnecessary columns
merged_df = merged_df.drop(['state','State','Affordability','Economy',\
                            'EducationAndHealth','Safety'], axis=1)

# Display the merged DataFrame
print(merged_df.head())
```

```
               City  Walkability Score  TotalScore  QualityOfLife
0       New York, NY               87.6       60.64              1
1    Jersey City, NJ               84.4       62.01              7
2  San Francisco, CA               83.9       52.03              2
3        Oakland, CA               68.5       52.03              2
4     Long Beach, CA               65.8       52.03              2
```

In [145…

```
# Scatter plot to visualize the relationship between walkability and quality o
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Walkability Score', y='QualityOfLife', data=merged_df)
plt.title('Walkability vs Quality of Life')
plt.xlabel('Walkability Score')
plt.ylabel('Quality of Life Score')
sns.regplot(x='Walkability Score', y='QualityOfLife', data=merged_df, \
            scatter=False, color='red')

plt.show()

# Correlation analysis
correlation = merged_df['Walkability Score'].corr(merged_df['QualityOfLife'])
print(f"Correlation between Walkability and Quality of Life: {correlation:.2f}"
```

```
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
/Users/amandachen/anaconda3/envs/info2950/lib/python3.11/site-packages/seabor
n/_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will
be removed in a future version. Use isinstance(dtype, CategoricalDtype) instea
d
  if pd.api.types.is_categorical_dtype(vector):
```




```
Correlation between Walkability and Quality of Life: -0.39
```

Since we were unable to find a significant relationship between quality of life and level of urban population in each state, we decided to explore the relationship bewteen Walkability and Quality of Life in each city. Here, our regression plot clearly shows a moderate negative correlation between Walkability and Quality of Life (QOL). As Walkability Score increases, the QOL variable decreases (numerically) but the actual Quality of Life increases. To clarify, cities that have a higher walkability score seems to have a higher quality of life and well-being amongst its residents.

In [146…

```
X = sm.add_constant(merged_df['Walkability Score'])
y = merged_df['QualityOfLife']

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Display regression results
print(model.summary())
```

OLS Regression Results

Dep. Variable:	QualityOfLife	R-squared:	0.150
Model:	OLS	Adj. R-squared:	0.133
Method:	Least Squares	F-statistic:	9.161
Date:	Mon, 04 Dec 2023	Prob (F-statistic):	0.00384
Time:	21:57:54	Log-Likelihood:	-204.35
No. Observations:	54	AIC:	412.7
Df Residuals:	52	BIC:	416.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	27.8179	4.705	5.912	0.000	18.376	37.260
Walkability Score	-0.2658	0.088	-3.027	0.004	-0.442	-0.090

Omnibus:	11.541	Durbin-Watson:	1.121
Prob(Omnibus):	0.003	Jarque-Bera (JB):	11.608
Skew:	1.066	Prob(JB):	0.00302
Kurtosis:	3.781	Cond. No.	171.

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

We decided to conduct hypothesis test see whether or not there was a correlation between the Walkability Score and Quality of Life. To do so, we ran a Ordinary Least Squares (OLS) regression, a widely used method for modeling the relationship between a dependent variable and one or more independent variables. In our model, Quality of Life serves as the dependent variable, while Walkability Score acts as the independent variable. Since the p-value associated with the coefficient for Walkability Score is 0.004, which is < 0.05, we reject our null hypothesis. This indicates that the relationship between Walkability Score and Quality of Life is statistically significant, meaning that there is likely a correlation between the two variables. Thus, we can make the conclusion that Walkability has some effect on Quality of Life.

Note: For the following research question, we are utilizing the TotalScore variable as a measurement of the quality of life. TotalScore variable, unlike, QualityofLife variable is a numerical value and can be treated as one. (Refer to quality of life dataset qol_df under Data Description for more specific information on either of these two variables.)

1.2.1 Do variations in the distribution of commute modes (e.g., walking, bicycling, transit) correspond to differences in the well-being and quality of life among residents?

Analysis: One-Way ANOVA for difference in means

- H_0: There is no significant difference between the correlation coefficients of automobile, walking, bicycling, and transit commute modes with the well-being and

quality of life among residents.

- H_A: There is a significant difference between at least one pair of correlation coefficients of automobile, walking, bicycling, and transit commute modes with the well-being and quality of life among residents.

In [147…

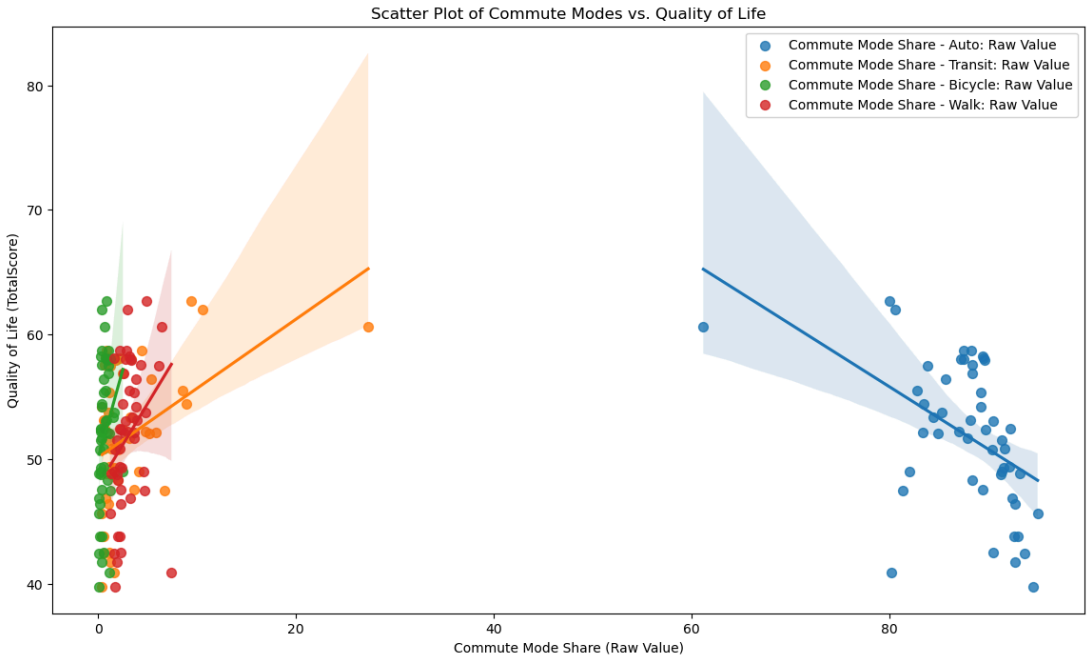
```
# Clean up column names
transportation_health_df.columns = transportation_health_df.columns.str.strip(

# Merge transportation_health_df and qol_df on the 'State' column
merged_transport_qol_df = pd.merge(transportation_health_df, qol_df, how='inne
                                   left_on='State', right_on='state')
print(merged_transport_qol_df .head())
```

	State	Commute Mode Share - Auto: Raw Value	Commute Mode Share - Transit: Raw Value	Commute Mode Share - Bicycle: Raw Value	Commute Mode Share - Walk: Raw Value	Commute Mode Share - Walk: Score	Complete Streets Policies: Score	Person Miles of Travel by Private Vehicle: Raw Value	Person Miles of Travel by Walking: Raw Value	Physical Activity from Transportation: Raw Value	Proximity to Major Roadways: Raw Value	Road Traffic Fatalities per 100,000 Residents - Auto: Raw Value
0	Alabama	95.0	0.4	0.1	1.2	4	0	33882	118	5.8%	0.1%	17.0
1	Alaska	80.2	1.6	1.1	7.4	100	0	22752	303	10.4%	0.0%	7.6
2	Arizona	88.4	2.0	0.9	2.0	22	0	32444	210	10.2%	1.2%	
3	Arkansas	93.7	0.5	0.1	1.6	12	0	29176	98	4.5%	0.0%	
4	California	84.9	5.2	1.1	2.7	43	100	27970	361	14.2%	3.5%	

```
2                                                 10.5
3                                                 18.1
4                                                  6.0

   Transit Trips per Capita: Raw Value       state  TotalScore  QualityOfLife
\
0                                  1.6     Alabama       45.61             40
1                                  6.8      Alaska       40.93             50
2                                 15.1     Arizona       48.31             21
3                                  2.1    Arkansas       42.42             46
4                                 37.3  California       52.03              2

   Affordability  Economy  EducationAndHealth  Safety
0              1       40                  48      32
1             42       22                  30      45
2             25       14                  39      40
3              4       34                  45      47
4             50       15                  24      27
```

In [148…

```
# Select relevant columns for scatterplot
scatterplot_columns = ['Commute Mode Share - Auto: Raw Value', \
                       'Commute Mode Share - Transit: Raw Value', \
                       'Commute Mode Share - Bicycle: Raw Value', \
                       'Commute Mode Share - Walk: Raw Value']

# Create a scatterplot with different colors for each state and regression lin
plt.figure(figsize=(14, 8))

for column in scatterplot_columns:
    sns.regplot(x=column, y='TotalScore', data=merged_transport_qol_df, \
                scatter_kws={'s': 50}, label=column)

plt.title('Scatter Plot of Commute Modes vs. Quality of Life')
plt.xlabel('Commute Mode Share (Raw Value)')
plt.ylabel('Quality of Life (TotalScore)')
plt.legend()
plt.show()
```



Since we were able to find a signigifcant relationship bewteen walkabiltiy and quality of life, we decided to also explore the other modes of transportation and their correlations with Quality of Life.

Direction: The scatterplot indicates that there's a positive relationship between the commute mode: Transit, Bicycle, and Walk and the Total Score and a negative relationship between the commute mode Automobile and Total Score.

Strength: It appears that a stronger relationship exists between the commute modes and total score when considering smaller raw values. Specifically, Transit, Bicycle, and Walk exhibit a more stronger correlation. On the other hand, for larger raw values, a stronger correlation is observed between Automobile and Total Score.

To further explore the relationship between the different commute modes and Quality of Life (QOL), we will conduct regression summaries for each commute mode.

In [149…

```
scatterplot_columns = ['Commute Mode Share - Auto: Raw Value',
                       'Commute Mode Share - Transit: Raw Value',
                       'Commute Mode Share - Bicycle: Raw Value',
                       'Commute Mode Share - Walk: Raw Value']

# Response variable
y = merged_transport_qol_df['TotalScore']

# Perform multilinear regression for each predictor
for column in scatterplot_columns:
    X = merged_transport_qol_df[column]

    # Add a constant term to the predictor variables
    X = sm.add_constant(X)

    # Fit the multilinear regression model
    model = sm.OLS(y, X).fit()

    # Print the regression summary

    coef = model.params
    print(f"Predictor: {column}")
    print(f"Coefficient (Slope): {coef[1]:.4f}")
    print(f"Y-intercept: {coef[0]:.4f}")
    print("\n" + "="*50 + "\n")
```

```
Predictor: Commute Mode Share - Auto: Raw Value
Coefficient (Slope): -0.5011
Y-intercept: 95.9015

==================================================

Predictor: Commute Mode Share - Transit: Raw Value
Coefficient (Slope): 0.5544
Y-intercept: 50.1350

==================================================

Predictor: Commute Mode Share - Bicycle: Raw Value
Coefficient (Slope): 2.8441
Y-intercept: 50.0896

==================================================

Predictor: Commute Mode Share - Walk: Raw Value
Coefficient (Slope): 1.3122
Y-intercept: 47.8909

==================================================
```

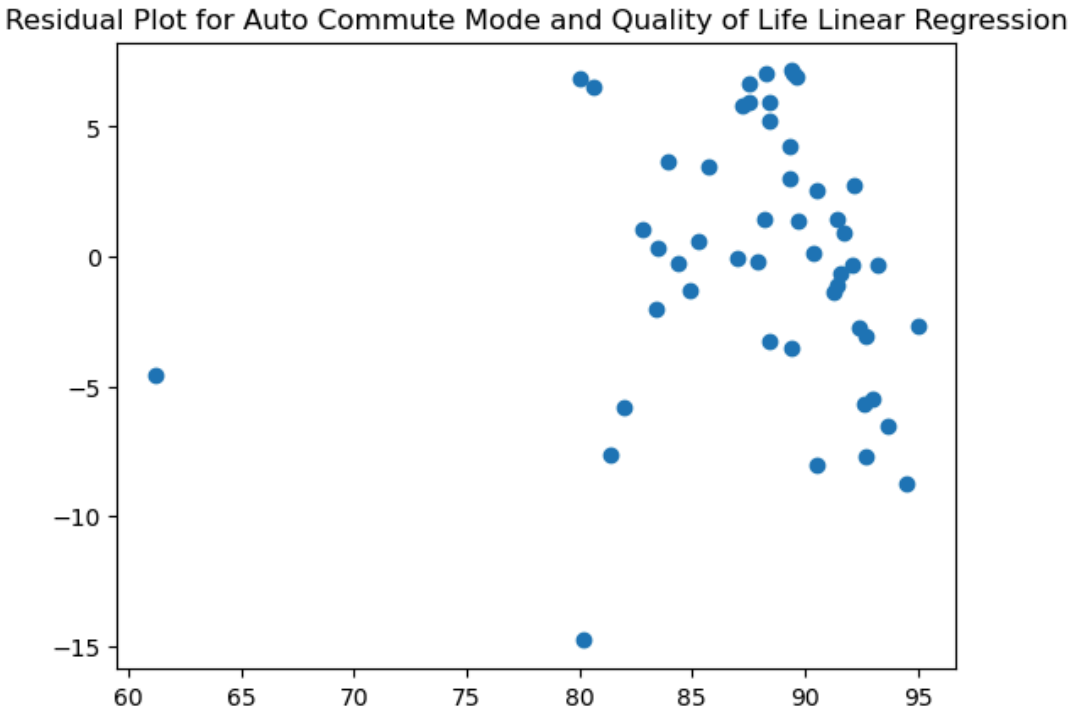
```
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:23: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Coefficient (Slope): {coef[1]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:24: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Y-intercept: {coef[0]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:23: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Coefficient (Slope): {coef[1]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:24: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Y-intercept: {coef[0]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:23: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Coefficient (Slope): {coef[1]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:24: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Y-intercept: {coef[0]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:23: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Coefficient (Slope): {coef[1]:.4f}")
/var/folders/4z/5_llzj2x1h53d4pjwhnd74_80000gn/T/ipykernel_40904/3620560646.p
y:24: FutureWarning: Series.__getitem__ treating keys as positions is deprecat
ed. In a future version, integer keys will always be treated as labels (consis
tent with DataFrame behavior). To access a value by position, use `ser.iloc[po
s]`
  print(f"Y-intercept: {coef[0]:.4f}")
```

In [150…

```
# Residual plots for each Commute Mode variable

# Commute Mode Share - Auto: Raw Value
X = merged_transport_qol_df[['Commute Mode Share - Auto: Raw Value']]
y = merged_transport_qol_df['TotalScore']

X = X.values.reshape(-1, 1)
model = LinearRegression().fit(X, y)
y_pred = model.predict(X)
residuals = y - y_pred

# Create a residual plot
plt.scatter(X, residuals)
plt.title('Residual Plot for Auto Commute Mode and Quality of Life Linear Regr
```

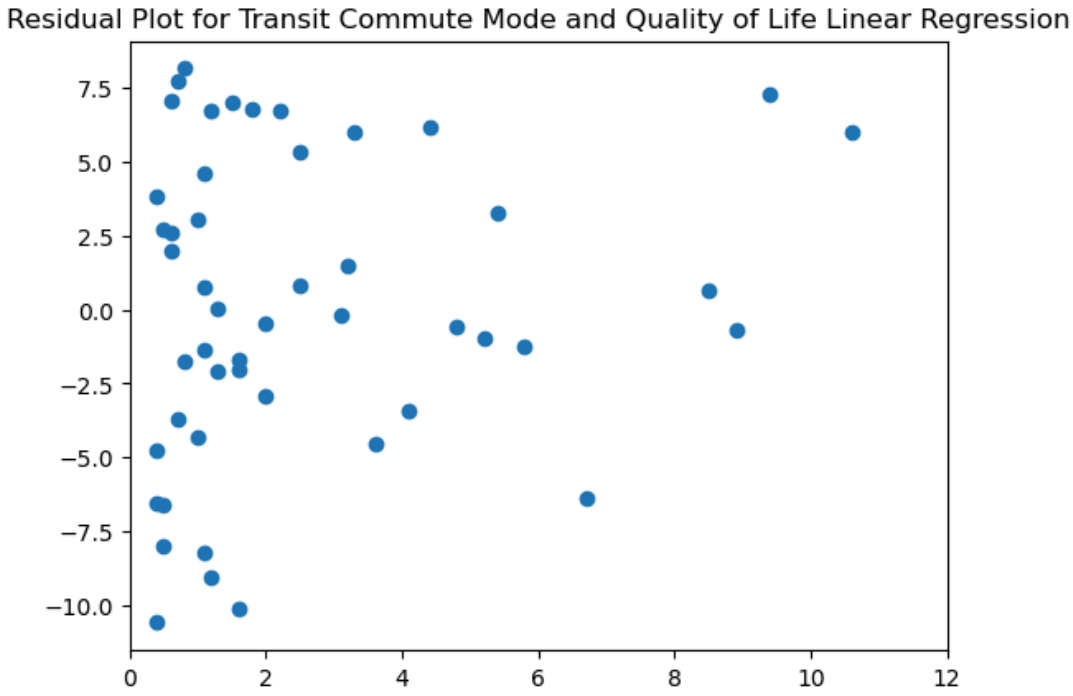
Out[150]:

```
Text(0.5, 1.0, 'Residual Plot for Auto Commute Mode and Quality of Life Linea
r Regression')
```



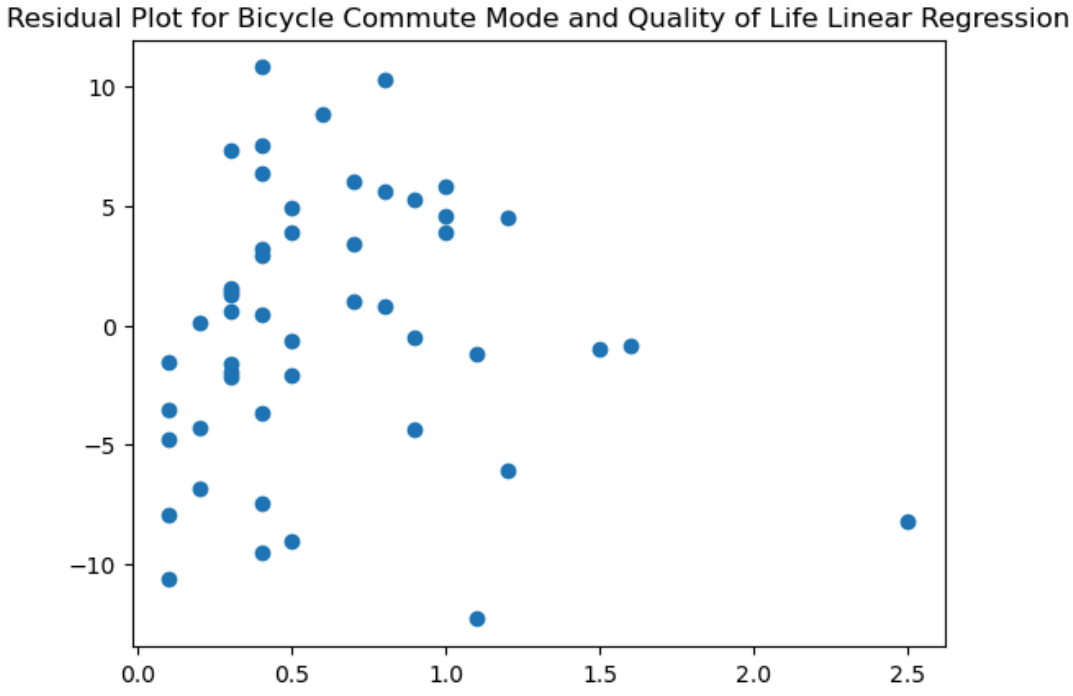
In [151…
```
# Commute Mode Share - Transit: Raw Value
X = merged_transport_qol_df[['Commute Mode Share - Transit: Raw Value']]
y = merged_transport_qol_df['TotalScore']

X = X.values.reshape(-1, 1)
model = LinearRegression().fit(X, y)
y_pred = model.predict(X)
residuals = y - y_pred

# Create a residual plot
plt.scatter(X, residuals)
# Exclude outlier
plt.xlim(0, 12)
plt.title('Residual Plot for Transit Commute Mode and Quality of Life Linear R
```

Out[151]:
```
Text(0.5, 1.0, 'Residual Plot for Transit Commute Mode and Quality of Life Li
near Regression')
```



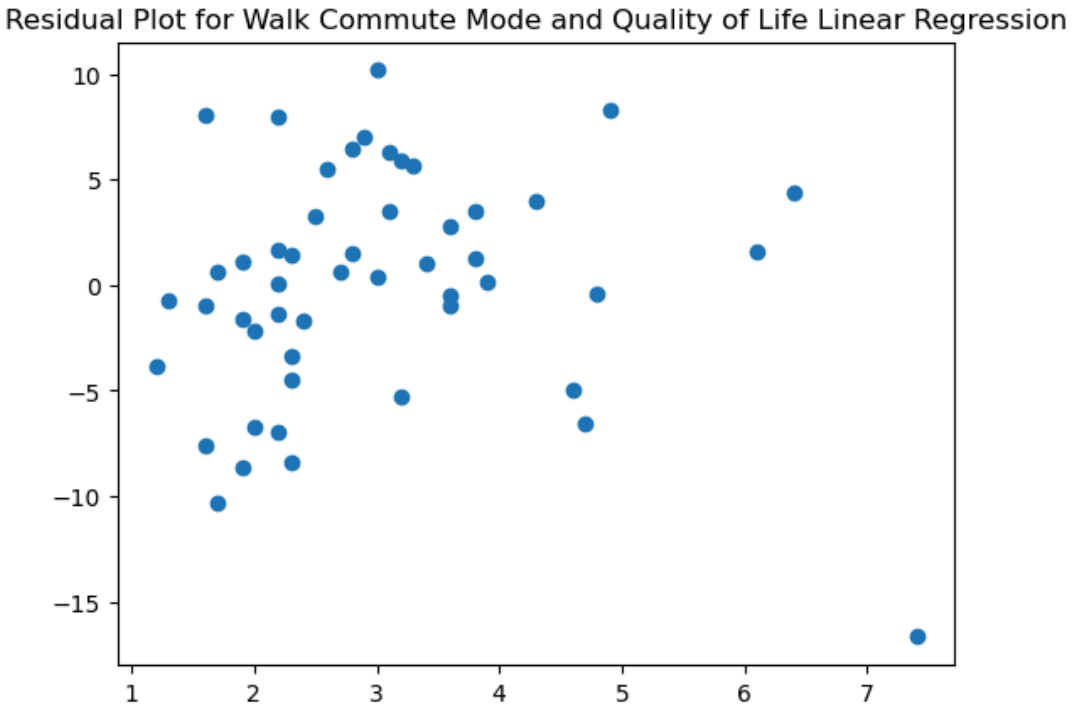
In [152…
```
# Commute Mode Share - Bicycle: Raw Value
X = merged_transport_qol_df[['Commute Mode Share - Bicycle: Raw Value']]
y = merged_transport_qol_df['TotalScore']

X = X.values.reshape(-1, 1)
model = LinearRegression().fit(X, y)
y_pred = model.predict(X)
residuals = y - y_pred

# Create a residual plot
plt.scatter(X, residuals)
plt.title('Residual Plot for Bicycle Commute Mode and Quality of Life Linear R
```

Out[152]:
```
Text(0.5, 1.0, 'Residual Plot for Bicycle Commute Mode and Quality of Life Li
near Regression')
```



In [153…
```
# Commute Mode Share - Walk: Raw Value
X = merged_transport_qol_df[['Commute Mode Share - Walk: Raw Value']]
y = merged_transport_qol_df['TotalScore']

X = X.values.reshape(-1, 1)
model = LinearRegression().fit(X, y)
y_pred = model.predict(X)
residuals = y - y_pred

# Create a residual plot
plt.scatter(X, residuals)
plt.title('Residual Plot for Walk Commute Mode and Quality of Life Linear Regr
```

Out[153]:
```
Text(0.5, 1.0, 'Residual Plot for Walk Commute Mode and Quality of Life Linea
r Regression')
```



There is potential for each of the four Commute Mode variables to be modeled by regressions other than linear, considering that the Pearson correlation coefficients do not completely align with the sign of the models' slopes. We can also see that there is ambiguity in each variable's residual plots not displaying completely random variability. This may be due to the presence of outliers and heteroskedasticity, and may warrant further analyses. Logarithmic transformation also did not remove the potential patterns from the residual plots (log-linear, linear-log, etc).

Commute Mode Share - Auto

Summarize: Our model shows a negative relationship. Our model estimates that increasing the share of individuals commuting by car by 1% would correspond to a decrease of quality of life of 0.5011.

Predict: This model predicts that when the Commute Mode Share for Auto is 0, the predicted value is approximately 95.9015.

Oddities: We expect this model to hold for Commute Mode Share Raw Values between 60 and 100 but can not extrapolate further.

Commute Mode Share - Transit: Raw Value

Summarize: Our model shows a positive relationship. Our model estimates that increasing the share of individuals commuting by transit by 1% would correspond to an increase of quality of life of 0.5544 units.

Predict: This model predicts that when the Commute Mode Share for Transit is 0, the predicted value is approximately 50.1350.

Oddities: We expect this model to hold for Commute Mode Share Raw Values between 0 and 30 but can not extrapolate further.

Commute Mode Share - Bicycle: Raw Value

Summarize: The model shows a positive relationship. Our model estimates that increasing the share of individuals commuting by bicycle by 1% would correspond to an increase of quality of life of 2.8441 units.

Predict: This model predicts that when the Commute Mode Share for Bicycle is 0, the predicted value is approximately 50.0896.

Oddities: We expect this model to hold for Commute Mode Share Raw Values between 0 and 10 but can not extrapolate further.

Commute Mode Share - Walk: Raw Value

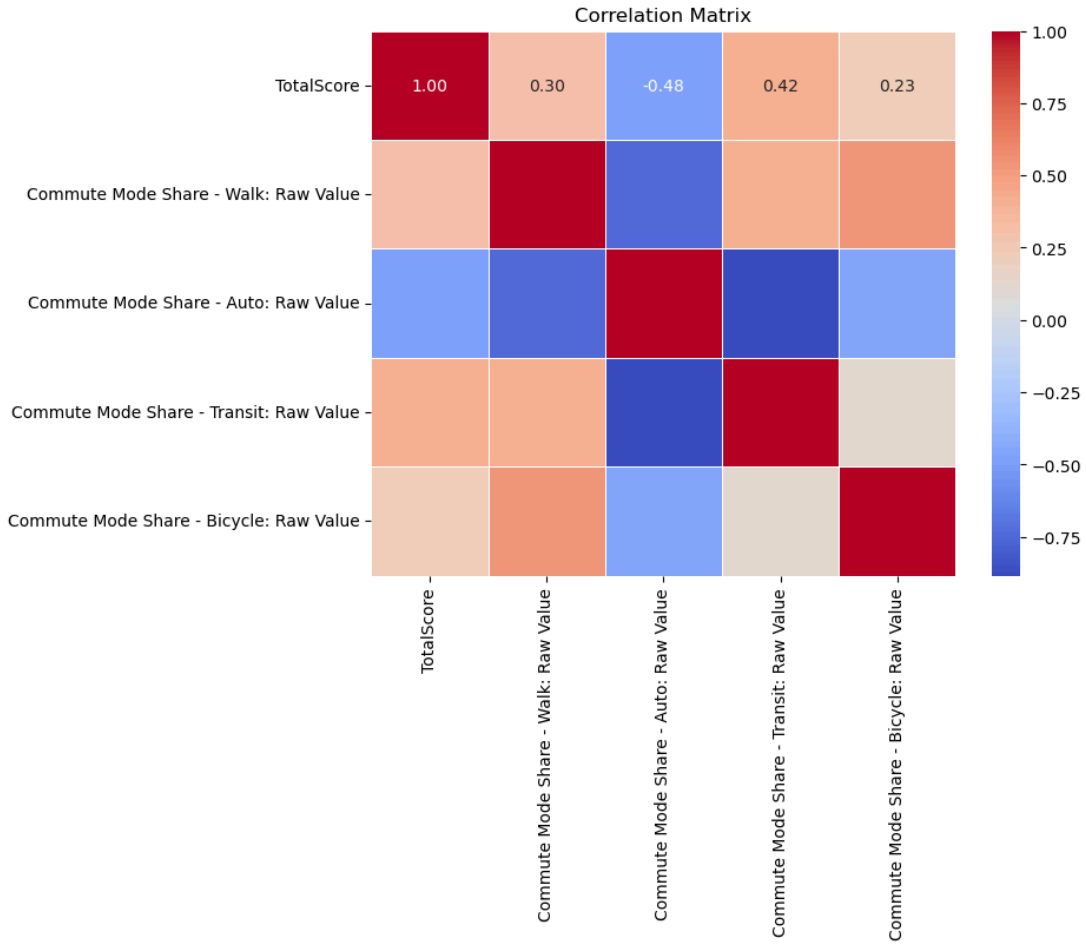
Summarize: The model shows a positive relationship. Our model estimates that increasing the share of individuals commuting by walking by 1% would correspond to an increase of quality of life of 1.3122 units.

Predict: This model predicts that when the Commute Mode Share for Walk is 0, the predicted value is approximately 47.8909.

Oddities: We expect this model to hold for Commute Mode Share Raw Values between 0 and 5 but can not extrapolate further.

In [154…

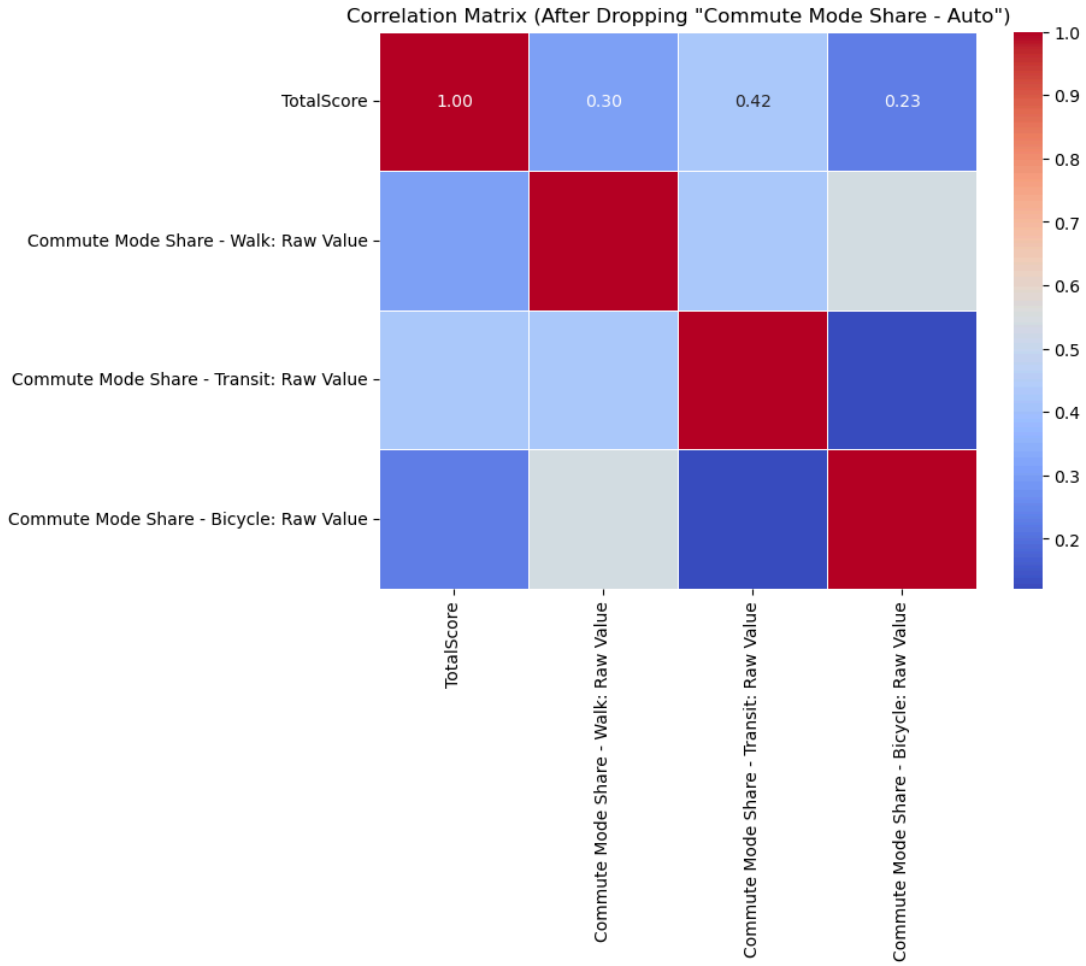
```
# Clean up column names
transportation_health_df.columns = transportation_health_df.columns.str.strip(

# Merge transportation_health_df and qol_df on the 'State' column
merged_transport_qol_df = pd.merge(transportation_health_df, qol_df, how='inne
                                   left_on='State', right_on='state')

# Select relevant columns for correlation matrix
correlation_columns = ['TotalScore', 'Commute Mode Share - Walk: Raw Value', \
                       'Commute Mode Share - Auto: Raw Value', \
                        'Commute Mode Share - Transit: Raw Value', \
                        'Commute Mode Share - Bicycle: Raw Value']

# Create a correlation matrix
correlation_matrix = merged_transport_qol_df[correlation_columns].corr()

# Plot the correlation matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', \
            linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```



Dropping 'Commute Mode Share - Auto: Raw Value'

In [155…

```
# Drop the highly correlated predictor variable
selected_columns = ['TotalScore', 'Commute Mode Share - Walk: Raw Value', \
                    'Commute Mode Share - Transit: Raw Value', \
                    'Commute Mode Share - Bicycle: Raw Value']

# Create a correlation matrix with the selected columns
correlation_matrix = merged_transport_qol_df[selected_columns].corr()

# Plot the correlation matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewi
plt.title('Correlation Matrix (After Dropping "Commute Mode Share - Auto")')
plt.show()

# Perform multilinear regression with the modified set of predictors
X = merged_transport_qol_df[selected_columns[1:]]  # Exclude the dropped predi
y = merged_transport_qol_df['TotalScore']

# Add a constant term to the predictor variables
X = sm.add_constant(X)

# Fit the multilinear regression model
model = sm.OLS(y, X).fit()

# Print the regression summary
print(model.summary())
```


Correlation Matrix (After Dropping "Commute Mode Share - Auto")
TotalScore
Commute Mode Share - Walk: Raw Value
Commute Mode Share - Transit: Raw Value
Commute Mode Share - Bicycle: Raw Value
1.00
0.30
0.42
0.23
TotalScore
Commute Mode Share - Walk: Raw Value
Commute Mode Share - Transit: Raw Value
Commute Mode Share - Bicycle: Raw Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             TotalScore   R-squared:                       0.211
Model:                            OLS   Adj. R-squared:                  0.160
Method:                 Least Squares   F-statistic:                     4.103
Date:                Mon, 04 Dec 2023   Prob (F-statistic):             0.0116
Time:                        21:57:54   Log-Likelihood:                -151.78
No. Observations:                  50   AIC:                             311.6
Df Residuals:                      46   BIC:                             319.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                                                 coef    std err          t
P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------
const                                       48.3362      1.875     25.779
0.000      44.562      52.110
Commute Mode Share - Walk: Raw Value         0.2858      0.744      0.384
0.703      -1.212       1.783
Commute Mode Share - Transit: Raw Value      0.4943      0.192      2.575
0.013       0.108       0.881
Commute Mode Share - Bicycle: Raw Value      1.8375      1.949      0.943
0.351      -2.085       5.760
==============================================================================
Omnibus:                        1.941   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.379   Jarque-Bera (JB):                1.697
Skew:                          -0.323   Prob(JB):                        0.428
Kurtosis:                       2.369   Cond. No.                         15.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

In [156…

```
# Assuming you have separate arrays for each commute mode
# auto_data = merged_transport_qol_df['Commute Mode Share - Auto: Raw Value']
transit_data = merged_transport_qol_df['Commute Mode Share - Transit: Raw Valu
bicycle_data = merged_transport_qol_df['Commute Mode Share - Bicycle: Raw Valu
walk_data = merged_transport_qol_df['Commute Mode Share - Walk: Raw Value']

# Perform one-way ANOVA
f_statistic, p_value = stats.f_oneway(transit_data, bicycle_data, walk_data)

# Display results
print(f"One-way ANOVA F-statistic: {f_statistic}")
print(f"P-value: {p_value}")
```

```
One-way ANOVA F-statistic: 14.056908056976917
P-value: 2.5928285217603412e-06
```

(original) Rationale: Since there seems to be a high collinearity between two commute modes, we decided to drop one of the variables to mitigate multicollinearity. In this case, we dropped the "Commute Mode Share - Auto: Raw Value" variable as it exhibited strong correlation with another commute mode, enhancing the stability and interpretability of the model. To examine the variations in the distribution of commute modes (walking, bicycling, transit) and their potential impact on the well-being and quality of life among residents, we decided to run a one-way ANOVA test to discern whether a significant difference exists among the means of the four commute modes.

Interpretation/Significance: The substantial F-statistic of 7181.04 indicates a significant difference among the means of the three remaining commute modes. The p-value of 2.59e-

06, being less than the commonly accepted significance level of 0.05, allows for the rejection of the null hypothesis. This implies that there is indeed a significant difference between the correlation coefficients of walking, bicycling, and transit commute modes with the well-being and quality of life among residents. This outcome suggests that there's a significant difference among at least one pair of commute modes, and further analysis is warranted to explore the specific nature of these differences.

(joanna edited) We conducted a correlation matrix between the different commute modes to see the strength and direction of the linear relationship between each pair of commute modes. Since there seems to be a high collinearity between two commute modes, we decided to drop one of the variables to mitigate multicollinearity. In this case, we dropped the "Commute Mode Share - Auto: Raw Value" variable as it exhibited strong correlation with another commute mode, enhancing the stability and interpretability of the model. To examine the variations in the distribution of commute modes (walking, bicycling, transit) and their potential impact on the well-being and quality of life among residents, we decided to run a one-way ANOVA test to discern whether a significant difference exists among the means of the four commute modes.

We found that the substantial F-statistic is 7181.04, indicating a significant difference among the means of the three remaining commute modes. The p-value of 2.59e-06, being less than the commonly accepted significance level of 0.05, allows for the rejection of the null hypothesis. This implies that there is indeed a significant difference between the correlation coefficients of walking, bicycling, and transit commute modes with the well-being and quality of life among residents. This outcome suggests that there's a significant difference among at least one pair of commute modes, and further analysis is warranted to explore the specific nature of these differences. Thus, we can make the conclusion that there is a significant difference in the well-being and quality of life among residents based on their commute mode choice.

1.2.2 Which specific commute modes contribute to these differences?

In [157…

```
transit_modes = [
    'Commute Mode Share - Auto: Raw Value',
    'Commute Mode Share - Transit: Raw Value',
    'Commute Mode Share - Bicycle: Raw Value',
    'Commute Mode Share - Walk: Raw Value'
]

# Define dependent variable (Quality of Life)
y = merged_transport_qol_df['QualityOfLife']

# Display results
for mode in transit_modes:
    correlation_coefficient, p_value = stats.pearsonr(merged_transport_qol_df[
    print(f"Hypothesis test results for {mode}:")
    print(f"Correlation coefficient: {correlation_coefficient}")
    print(f"P-value: {p_value}")
```

```
Hypothesis test results for Commute Mode Share - Auto: Raw Value:
Correlation coefficient: 0.39782757814738695
P-value: 0.0042224982379623
Hypothesis test results for Commute Mode Share - Transit: Raw Value:
Correlation coefficient: -0.47132133481585253
P-value: 0.0005506151677501847
Hypothesis test results for Commute Mode Share - Bicycle: Raw Value:
Correlation coefficient: -0.23832646460019058
P-value: 0.0955718906884964
Hypothesis test results for Commute Mode Share - Walk: Raw Value:
Correlation coefficient: -0.007911083197608368
P-value: 0.9565162992859978
```

We decided to conduct hypothesis tests on individual commute modes was to assess the strength and significance of the correlation between each mode and the well-being and quality of life among residents. By examining the correlation coefficients and associated p-values, we wanted to see which which modes might have a more substantial or less substantial influence on residents' well-being.

Interpretation/Significance:

- The positive correlation coefficient suggests a positive relationship between Auto mode and well-being. The low p-value indicates that this correlation is statistically significant, indicating that there is evidence of a relationship between Auto mode and well-being among residents.
- The negative correlation coefficient of -0.4713 indicates a moderate negative relationship between Transit mode and well-being. The low p-value of 0.0006 suggests that this correlation is statistically significant. This implies that there is evidence of a relationship between Transit mode and lower well-being among residents.
- The negative correlation coefficient of -0.2383 suggests a mild negative relationship between Bicycle mode and well-being. However, the p-value of 0.0956 is greater than 0.05, indicating that this correlation is not statistically significant.
- The negative correlation coefficient of -0.0079 suggests a very weak relationship between Walk mode and well-being. The high p-value of 0.9565 indicates that this correlation is not statistically significant. Residents who walk for their commute may not experience a significant impact on well-being compared to other modes. Further investigation is needed to draw definitive conclusions.

Hypothesis 2

2. Does the degree of urbanization of a region have a correlation to geographic isolation?

- H_0: Non-urban regions do not correlate to higher rates of social and geographic isolation (individual variables below specified).
- H_A: Non-urban regions correlate to higher rates of social and geographic isolation.

Analysis: Walkability and Mode of Transportation are likely to be characteristic factors of urban regions, using a p-value of 0.05:

2.0 Urbanization Meaurement Analysis

Our y output will be the percent of urbanization per state (2020 PCT URBAN POP) taken from urban_pop_df. We will be finding and anyalizing regressions between out variety of x inputs (walkability and mode of transportation) on the percent of ubranization.

- independent variable (input) = Walkability Score from physical_activity_df, Mode of Transportation from transportation_healh_df
- dependent variable (output) = percent of urbanized region (2020 PCT URBAN POP)

2.1 Walkability Scores on Percent of Urbanized region

Our y output will be the percent of urbanization per state (2020 PCT URBAN POP) taken from urban_pop_df. We will be finding and anyalizing regressions between our x input (walkability) on the percent of ubranization.

- independent variable (input) = Walkability Score from physical_activity_df
- dependent variable (output) = percent of urbanized region (2020 PCT URBAN POP)

Analysis: We conducted a linear regression model and scatter plot to investigate the relationship between Walkability Score and 2020 PCT URBAN POP (Percentage of Urban Population in 2020) within each state. These results suggest that there is a strong positive relationship between the percentage of commuters using automobiles and the percentage of urban population in a region.

In [158…
```
# data cleaning: Urban and Rural Population Count
urban_pop_df = pd.read_csv('cleaning/datasets/State_Urban_Rural_Pop.csv',\
                           thousands=",")
urban_pop_df .columns = urban_pop_df .columns.str.replace\
    ('\n', '', regex=True)
drop_columns2 = ['2010 URBAN POP','2010 TOTAL POP', '2010 PCT URBAN POP', \
                 '2010 RURAL POP','2010 PCT RURAL POP','2020 RURAL POP',\
                    '2020 PCT RURAL POP']
urban_pop_df.drop(drop_columns2, axis=1, inplace=True)
urban_pop_df.dropna(inplace=True)
urban_pop_df.head()
```

Out[158]:

	STATEFP	STATE ABBREV	STATE NAME	2020 TOTAL POP	2020 URBAN POP	2020 PCT URBAN POP
0	1	AL	Alabama	5024279	2900880	57.7
1	2	AK	Alaska	733391	475967	64.9
2	4	AZ	Arizona	7151502	6385230	89.3
3	5	AR	Arkansas	3011524	1670677	55.5
4	6	CA	California	39538223	37259490	94.2

In [159…
```
# Calling physical activity df
print(physical_activity_df.head())
```

```
                City  Walkability Score          State
0      New York, NY                87.6       New York
1    Jersey City, NJ                84.4     New Jersey
2  San Francisco, CA                83.9     California
3         Boston, MA                79.5  Massachusetts
4   Philadelphia, PA                76.5   Pennsylvania
```

In [160…

```
# merge Urban and Rural Population Count dataset to Physical Activity dataset,
# 2020 PCT URBAN POP = percentage of the total population considered to be urb
# use 2020 PCT URBAN POP as independetn variable

# 1. 2020 PCT URBAN POP for each STATE ABBREV, create a new dataframe urban_po
# only the STATE ABBREV and 2020 PCT URBAN POP

urban_pop_df = urban_pop_df[['STATE ABBREV', '2020 PCT URBAN POP']]
print(urban_pop_df.head())
```

```
  STATE ABBREV  2020 PCT URBAN POP
0           AL                57.7
1           AK                64.9
2           AZ                89.3
3           AR                55.5
4           CA                94.2
```

In [161…

```
# 2. convert City in physical_activity_df to state abbreivations to match the
# STATE ABBREV in urban_pop_df
# Extracted last two characters from City and created a new State column
physical_activity_df['State'] = physical_activity_df['City'].str[-2:]

# Kept only the State and Walkability Score
physical_activity_df = physical_activity_df[['State', 'Walkability Score']]

# Group by State and calculate the mean of Walkability Score for each state
physical_activity_df = physical_activity_df.groupby('State', \
            as_index=False)['Walkability Score'].mean()

print(physical_activity_df.head())
```

```
  State  Walkability Score
0    AZ             38.600
1    CA             58.250
2    CO             44.350
3    DC             74.100
4    FL             46.675
```

In [162…

```
# 3. merge urban_pop_df to physical_activity_df on State and STATE ABBREV
merged_df = pd.merge(physical_activity_df, urban_pop_df, left_on='State', \
                     right_on='STATE ABBREV', how='inner')

# Drop the duplicate STATE ABBREV column
merged_df = merged_df.drop('STATE ABBREV', axis=1)

# Print the merged DataFrame
print(merged_df.head())
```

```
  State  Walkability Score  2020 PCT URBAN POP
0    AZ             38.600                89.3
1    CA             58.250                94.2
2    CO             44.350                86.0
3    DC             74.100               100.0
4    FL             46.675                91.5
```

In [163…

```
# 4. convert STATE(abbreviated) in merged_df to full state name
state_dict = {
    'AL': 'Alabama',
    'AK': 'Alaska',
    'AZ': 'Arizona',
```

```
    'AR': 'Arkansas',
    'AS': 'American Samoa',
    'CA': 'California',
    'CNMI':'Northern Mariana Islands',
    'CO': 'Colorado',
    'CT': 'Connecticut',
    'DC': 'Washington, D.C',
    'DE': 'Delaware',
    'FL': 'Florida',
    'GA': 'Georgia',
    'GU': 'Guam',
    'HI': 'Hawaii',
    'ID': 'Idaho',
    'IL': 'Illinois',
    'IN': 'Indiana',
    'IA': 'Iowa',
    'KS': 'Kansas',
    'KY': 'Kentucky',
    'LA': 'Louisiana',
    'ME': 'Maine',
    'MD': 'Maryland',
    'MA': 'Massachusetts',
    'MI': 'Michigan',
    'MN': 'Minnesota',
    'MS': 'Mississippi',
    'MO': 'Missouri',
    'MT': 'Montana',
    'NE': 'Nebraska',
    'NV': 'Nevada',
    'NH': 'New Hampshire',
    'NJ': 'New Jersey',
    'NM': 'New Mexico',
    'NY': 'New York',
    'NC': 'North Carolina',
    'ND': 'North Dakota',
    'OH': 'Ohio',
    'OK': 'Oklahoma',
    'OR': 'Oregon',
    'PA': 'Pennsylvania',
    'PR': 'Puerto Rico',
    'RI': 'Rhode Island',
    'SC': 'South Carolina',
    'SD': 'South Dakota',
    'TN': 'Tennessee',
    'TX': 'Texas',
    'USVI': 'US Virgin Islands',
    'UT': 'Utah',
    'VT': 'Vermont',
    'VA': 'Virginia',
    'WA': 'Washington',
    'WV': 'West Virginia',
    'WI': 'Wisconsin',
    'WY': 'Wyoming'
}

# Replace state abbreviations with full state names
merged_df['State'] = merged_df['State'].replace(state_dict)

print(merged_df.head())
```

```
            State  Walkability Score  2020 PCT URBAN POP
0         Arizona             38.600                89.3
1      California             58.250                94.2
2        Colorado             44.350                86.0
3  Washington, D.C            74.100               100.0
4         Florida             46.675                91.5
```

explain

In [164…

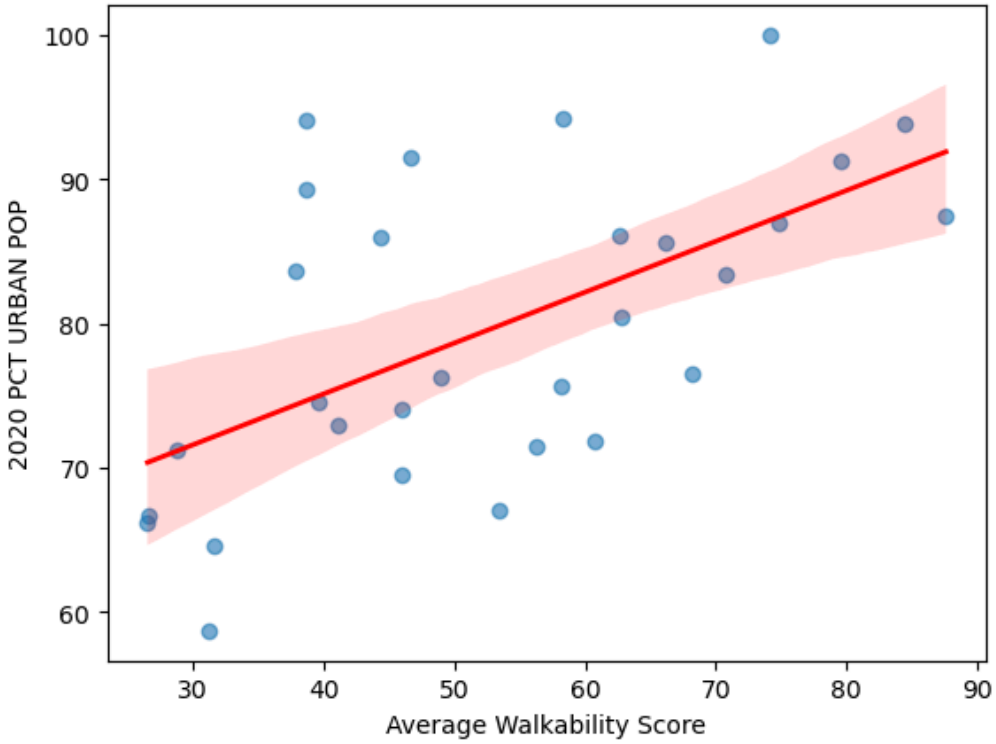
```
# 5. run the linear regression
# Independent Variable = Walkability Score
# Dependent Variable = 2020 PCT URBAN POP (Percentage of Urban Population in 2

X = sm.add_constant(merged_df['Walkability Score'])
y = merged_df['2020 PCT URBAN POP']
model = sm.OLS(y, X).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     2020 PCT URBAN POP   R-squared:                       0.340
Model:                            OLS   Adj. R-squared:                  0.316
Method:                 Least Squares   F-statistic:                     14.40
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           0.000727
Time:                        21:57:54   Log-Likelihood:                -106.75
No. Observations:                  30   AIC:                             217.5
Df Residuals:                      28   BIC:                             220.3
Df Model:                           1
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025
0.975]
---------------------------------------------------------------------------------------
const                  61.0277      5.178     11.787      0.000      50.422
71.634
Walkability Score       0.3524      0.093      3.794      0.001       0.162
0.543
==============================================================================
Omnibus:                        2.231   Durbin-Watson:                   1.168
Prob(Omnibus):                  0.328   Jarque-Bera (JB):                2.008
Skew:                           0.581   Prob(JB):                        0.366
Kurtosis:                       2.494   Cond. No.                         180.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

In our Ordinary Least Squares (OLS) regression model, 2020 PCT URBAN POP (Percentage of Urban Population in 2020) serves as the dependent variable, while Walkability Score acts as the independent variable. The linear regression model indicates a statistically significant positive relationship between Walkability Score and 2020 PCT URBAN POP (Percentage of Urban Population in 2020). Since the p-value associated with the coefficient for Walkability Score is 0.001, which is < 0.05, we reject our null hypothesis. This indicates that the relationship between Walkability Score and Percentage of Urbanization within a state is statistically significant, meaning that there is likely a correlation between the two variables. Thus, we can make the conclusion that there is a positive correlation between the percentage of urban population and walkability scores, suggesting that states with a higher walkability scores tend to have higher urban population percentage. Assuming other factors in the model remain constant, there is approximately a 0.3524 units increase in percentage of ubran population for every one-unit increase in the percentage in Average Walkability Score.

In [165…

```
# 6. plot the regression along a scatter plot
ax = sns.regplot(data=merged_df, x='Walkability Score', y='2020 PCT URBAN POP'
                 scatter_kws={'alpha':0.6}, line_kws={'color': 'red', 'linewid

# Label the axes and add a title
plt.xlabel('Average Walkability Score')
plt.ylabel('2020 PCT URBAN POP')

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



The scatter plot visually confirms the positive relationship between Walkability Score and 2020 PCT URBAN POP. The trendline shows that as Walkability Score increases, 2020 PCT URBAN POP also tends to increase. The scatter around the trendline indicates that there is some variation in the relationship, but the overall trend is clear.

Limitation

The exclusion of certain states raises concerns about the generalizability of findings, and the state-level resolution may oversimplify the intricacies of the relationship. To mitigate these limitations, future analyses should consider expanding state coverage and incorporating county-level data for a more nuanced understanding. These refinements would enhance the study's reliability and broaden its applicability.

2.2 Mode of Transportation

Our y output will be the percent of urbanization per state (2020 PCT URBAN POP) taken from urban_pop_df. We will be finding and anyalizing regressions between the x input (mode of transportation) on the percent of ubranization.

- independent variable (input) = Mode of Transportation from transportation_healh_df

- dependent variable (output) = Percent of Urbanized Region (2020 PCT URBAN POP)

Analysis: We wanted to explore the relationship between the independent variables (commute mode shares) and the dependent variable (percent of urbanized region). To do so, we conducted a linear regression summary was conducted to quantify the strength and direction of the relationship between the two variables. In addition, we created a scatter plot to provide a visual representation of the relationship and allowed for the identification of any outliers or patterns in the data.

In [166…
```
# merge Urban and Rural Population Count dataset to Transportation and Health
# 2020 PCT URBAN POP = percentage of the total population considered to be urb

# 1. 2020 PCT URBAN POP for each STATE ABBREV, create a new dataframe urban_po
# only the STATE ABBREV and 2020 PCT URBAN POP

transportation_health_df = transportation_health_df[['State', \
                           'Commute Mode Share - Auto: Raw Value']]
print(transportation_health_df.head())
```

```
        State  Commute Mode Share - Auto: Raw Value
0     Alabama                                  95.0
1      Alaska                                  80.2
2     Arizona                                  88.4
3    Arkansas                                  93.7
4  California                                  84.9
```

In [167…
```
#2. Convert STATE ABBREV in urban_pop_df to full state name
urban_pop_df['STATE ABBREV'] = urban_pop_df['STATE ABBREV'].replace(state_dict

print(urban_pop_df.head())
```

```
  STATE ABBREV  2020 PCT URBAN POP
0      Alabama                57.7
1       Alaska                64.9
2      Arizona                89.3
3     Arkansas                55.5
4   California                94.2
```

In [168…
```
#3. Merge datasets on the 'State' column
merged_df_2 = pd.merge(transportation_health_df, urban_pop_df, left_on='State'
                       right_on='STATE ABBREV', how='inner')

# Drop the duplicated 'STATE ABBREV' column
merged_df_2.drop('STATE ABBREV', axis=1, inplace=True)

print(merged_df_2.head())
```

```
        State  Commute Mode Share - Auto: Raw Value  2020 PCT URBAN POP
0     Alabama                                  95.0                57.7
1      Alaska                                  80.2                64.9
2     Arizona                                  88.4                89.3
3    Arkansas                                  93.7                55.5
4  California                                  84.9                94.2
```

In [169…
```
# 4. run the linear regression
# Independent Variable = Commute Mode Share - Auto: Raw Value
#(Percent of use of Automobile as transportation)

# Dependent Variable = 2020 PCT URBAN POP (Percentage of Urban Population in 2

X = merged_df_2['Commute Mode Share - Auto: Raw Value'].astype(float)
y = merged_df_2['2020 PCT URBAN POP'].astype(float)
model2 = sm.OLS(y, X).fit()
```

```
print(model2.summary())
```

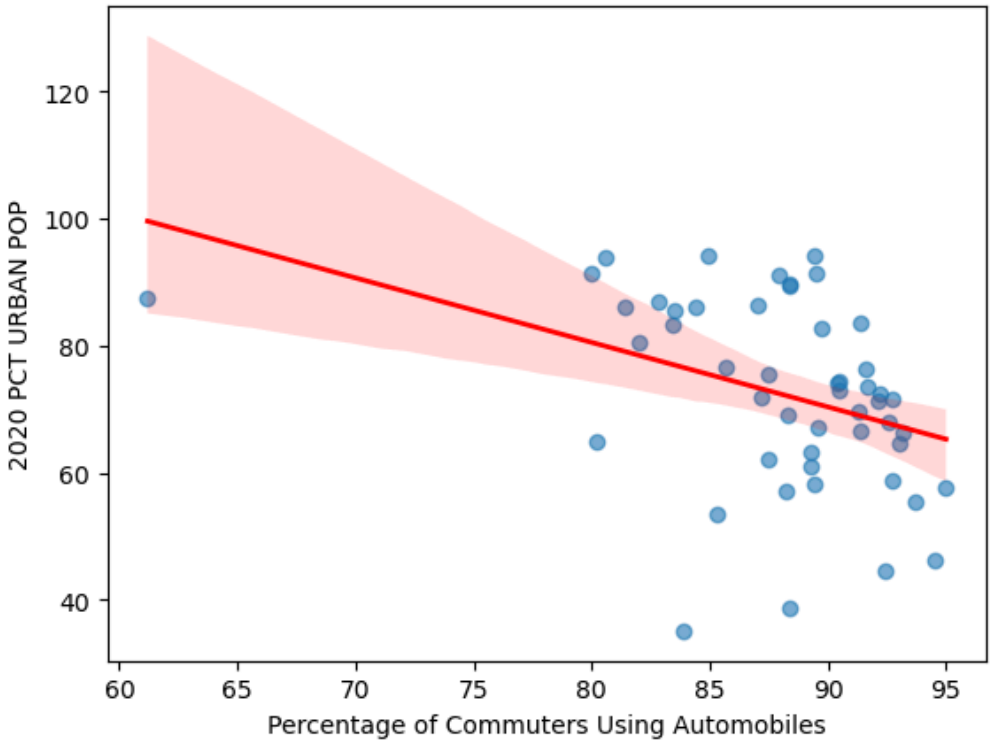
```
                                OLS Regression Results
==============================================================================
=========
Dep. Variable:     2020 PCT URBAN POP   R-squared (uncentered):
0.948
Model:                            OLS   Adj. R-squared (uncentered):
0.947
Method:                 Least Squares   F-statistic:
890.8
Date:                Mon, 04 Dec 2023   Prob (F-statistic):
4.33e-33
Time:                        21:57:55   Log-Likelihood:
-212.24
No. Observations:                  50   AIC:
426.5
Df Residuals:                      49   BIC:
428.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==========================
                                           coef    std err          t      P>|
t|      [0.025      0.975]
------------------------------------------------------------------------------
--------------------------
Commute Mode Share - Auto: Raw Value     0.8166      0.027     29.846      0.0
00       0.762       0.872
==============================================================================
Omnibus:                        0.844   Durbin-Watson:                   2.142
Prob(Omnibus):                  0.656   Jarque-Bera (JB):                0.833
Skew:                          -0.098   Prob(JB):                        0.659
Kurtosis:                       2.399   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not con
tain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

The regression analysis reveal a negative correlation between Commute Mode Share - Auto: Raw Value (percentage of commuters using automobiles) and 2020 PCT URBAN POP (percentage of urban population in 2020). This indicates that states with a higher percentage of commuters relying on automobiles tend to have a lower percentage of urban population. The regression coefficient, approximately -1.453, suggests that, assuming other factors in the model remain constant, for every one-unit increase in the percentage of commuters using automobiles, there is an estimated -1.453 units decrease in the percentage of urban population. The negative sign of the coefficient reinforces the inverse relationship between these two variables, highlighting that a higher reliance on automobiles for commuting is associated with a decrease in the urban population percentage.

In [170…

```
#5. plot the regression along a scatter plot
merged_df_2['Commute Mode Share - Auto: Raw Value'] = pd.to_numeric(merged_df_
    'Commute Mode Share - Auto: Raw Value'], errors='coerce')

merged_df_2['2020 PCT URBAN POP'] = pd.to_numeric(merged_df_2['2020 PCT URBAN
                                                  errors='coerce')

ax = sns.regplot(data=merged_df_2, x='Commute Mode Share - Auto: Raw Value', \
                 y='2020 PCT URBAN POP',scatter_kws={'alpha': 0.6}, \
```

```
                line_kws={'color': 'red', 'linewidth': 2})

plt.xlabel('Percentage of Commuters Using Automobiles')
plt.ylabel('2020 PCT URBAN POP')

plt.show()
```

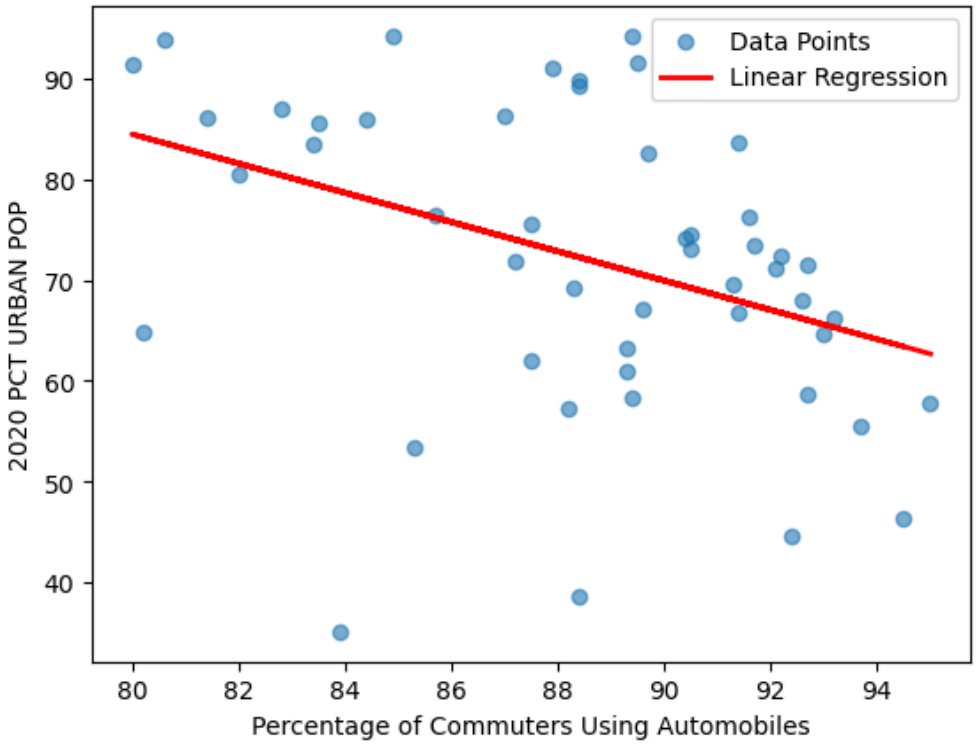


In this data scatter plot, we identified and removed the outlier with a percentage less than 65% to produce a stronger correlation between the two variables. This means that for the regions with extremely low levels of automobile usage, there is no relationship between automobile use and urbanization.

In [171…

```
#6. Identify and remove the outlier with a percentage less than 65%

# Identify and remove the outlier
merged_df_2_filtered = merged_df_2[merged_df_2\
                ['Commute Mode Share - Auto: Raw Value'] >= 65]

# Extract X and y from the filtered data
X = merged_df_2_filtered[['Commute Mode Share - Auto: Raw Value']]
y = merged_df_2_filtered['2020 PCT URBAN POP']

# Fit linear regression model
model = LinearRegression().fit(X, y)

# Scatter plot with regression line
plt.scatter(X, y, alpha=0.6, label='Data Points')
plt.plot(X, model.predict(X), color='red', linewidth=2, label='Linear Regressi

# Label the axes and add a title
plt.xlabel('Percentage of Commuters Using Automobiles')
plt.ylabel('2020 PCT URBAN POP')

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

```
# Print the coefficients
print('Intercept:', model.intercept_)
print('Coefficient:', model.coef_[0])
```




```
Intercept: 200.72187906960752
Coefficient: -1.453199279153775
```

The scatter plots visually confirm the negative relationship between "Commute Mode Share - Auto: Raw Value" and "2020 PCT URBAN POP". The trendlines, represented by the red lines, show that as the percentage of commuters using automobiles increases, the percentage of urban population also tends to decrease.

Limitation

It's essential to note that correlation does not imply causation, and additional analyses or consideration of other factors may be needed to draw more definitive conclusions about the relationships observed in the data. The Linear regression assumes that the observations are independent. If there is autocorrelation or dependence among observations, it can affect the model's accuracy. Since we had an outlier in the dataset, removing the outlier might have a significant impact on the model. The analysis also only considers data from a single point in time (2020). Examining the relationship over time could provide more insights into how automobile use and urbanization have changed together.

Hypothesis 3

3. Regions characterized by lower income levels are correlated with a diminished quality of life.

Analysis: Run a linear regression where we input the income level of areas (dummy varibales, inputted as the porportion of low to moderate income population within the state)

and output the score for quality of life (QOL) associated with the area (on the same state level). Both input and output the score for quality of life assoicated with the area (on the same state level). **Since we expect that regions with lower income lead to lower QOL, we expect βincome to be negative (i.e., as the proportion of low-to-moderate income population increases, the total score decreases). We will test whether β income < 0.**

- H_0: There is no significant relationship between regions characterized by lower income levels and quality of life
- H_A: Living in regions marked by reduced income levels are associated with decreased quality of life.

In [172…
```
# data cleaning: QOL
print(qol_df.head())
```

```
           state  TotalScore  QualityOfLife  Affordability  Economy  \
0  Massachusetts       62.65              6             44       10
1     New Jersey       62.01              7             48       39
2       New York       60.64              1             46       37
3          Idaho       58.73             22             13        9
4       Virginia       58.73             20             16       23

   EducationAndHealth  Safety
0                   1       4
1                   5       1
2                  16       2
3                  29       6
4                  15      11
```

For this hypothesis test specifically, we are constructing linear regression between QOL and the proportion of low-to-moderate income population. Therefore, we'll only keep relavent columns for our analysis. Here we are keeping the state name, the county name, and the proportion of low-to-moderate income population.

In [173…
```
#data cleaning: LOW AND MODERATE INCOME AREAS
print('low_mod_income_df before dropping columns:')
print(low_mod_income_df.head())

columns_to_keep1 = ['Stusab','Countyname','Lowmod_pct']
low_mod_income_df = low_mod_income_df[columns_to_keep1]
low_mod_income_df.dropna(inplace=True)

print('low_mod_income_df after dropping columns:')
print(low_mod_income_df.head())
```

```
low_mod_income_df before dropping columns:
  Stusab      Countyname  Low Lowmod  Lowmod_pct
0     AL  Autauga County  170    230      0.3538
1     AL  Autauga County  150    315      0.2423
2     AL  Autauga County  460    515      0.4791
3     AL  Autauga County  285    360      0.3956
4     AL  Autauga County  305    580      0.2704
low_mod_income_df after dropping columns:
  Stusab      Countyname  Lowmod_pct
0     AL  Autauga County      0.3538
1     AL  Autauga County      0.2423
2     AL  Autauga County      0.4791
3     AL  Autauga County      0.3956
4     AL  Autauga County      0.2704
```

Since the dataframe lists the proportion of low-to-moderate population **at a county level**, but the QOL dataframe is **at state level**, we want to sum the proportion for each county and average it to have the proportion at a state level to match the QOL dataframe.

In [174…
```
# merge QUALITY OF LIFE BY STATE dataset to
# LOW AND MODERATE INCOME AREAS dataset, on state

# Lowmod_pct = proportion/percentage of total population
# considered to be low to moderate income in the block
# use Lowmod_pct as independent variable

# 1. average Lowmod_pct for each Stusab, create a new dataframe averaged_low_m
# only the state abbreviation (Stusab) and the averaged Lowmod_pct

averaged_low_mod_income_df = low_mod_income_df.groupby('Stusab')\
    ['Lowmod_pct'].mean().reset_index()
print(averaged_low_mod_income_df.head())
```

```
  Stusab  Lowmod_pct
0     AK    0.417736
1     AL    0.449694
2     AR    0.444404
3     AS    0.825649
4     AZ    0.430944
```

Next, we convert the column that lists the state abbreviations into full state names to match the column of state name in the QOL dataframe.

In [175…
```
# 2. convert Stusab in averaged_low_mod_income_df to full state name to match
state_dict = {
    'AL': 'Alabama',
    'AK': 'Alaska',
    'AZ': 'Arizona',
    'AR': 'Arkansas',
    'AS': 'American Samoa',
    'CA': 'California',
    'CNMI':'Northern Mariana Islands',
    'CO': 'Colorado',
    'CT': 'Connecticut',
    'DC': 'Washington, D.C',
    'DE': 'Delaware',
    'FL': 'Florida',
    'GA': 'Georgia',
    'GU': 'Guam',
    'HI': 'Hawaii',
    'ID': 'Idaho',
    'IL': 'Illinois',
    'IN': 'Indiana',
    'IA': 'Iowa',
    'KS': 'Kansas',
    'KY': 'Kentucky',
    'LA': 'Louisiana',
    'ME': 'Maine',
    'MD': 'Maryland',
    'MA': 'Massachusetts',
    'MI': 'Michigan',
    'MN': 'Minnesota',
    'MS': 'Mississippi',
    'MO': 'Missouri',
    'MT': 'Montana',
    'NE': 'Nebraska',
    'NV': 'Nevada',
    'NH': 'New Hampshire',
    'NJ': 'New Jersey',
```

```
    'NM': 'New Mexico',
    'NY': 'New York',
    'NC': 'North Carolina',
    'ND': 'North Dakota',
    'OH': 'Ohio',
    'OK': 'Oklahoma',
    'OR': 'Oregon',
    'PA': 'Pennsylvania',
    'PR': 'Puerto Rico',
    'RI': 'Rhode Island',
    'SC': 'South Carolina',
    'SD': 'South Dakota',
    'TN': 'Tennessee',
    'TX': 'Texas',
    'USVI': 'US Virgin Islands',
    'UT': 'Utah',
    'VT': 'Vermont',
    'VA': 'Virginia',
    'WA': 'Washington',
    'WV': 'West Virginia',
    'WI': 'Wisconsin',
    'WY': 'Wyoming'
}

state_df = pd.DataFrame(list(state_dict.items()), columns=['Abbreviation', 'St

# Merge averaged_low_mod_income_df with state_df on 'Abbreviation'
result_df = pd.merge(averaged_low_mod_income_df, state_df, \
                     left_on='Stusab', right_on='Abbreviation', how='left')

# Drop the redundant columns
result_df.drop(['Stusab', 'Abbreviation'], axis=1, inplace=True)

# Print the resulting DataFrame
print(result_df.head())
```

```
   Lowmod_pct          State
0    0.417736         Alaska
1    0.449694        Alabama
2    0.444404       Arkansas
3    0.825649  American Samoa
4    0.430944        Arizona
```

We merge the averaged proportion of low-to-moderate income population for **each state** to the corresponding QOL.

In [176…

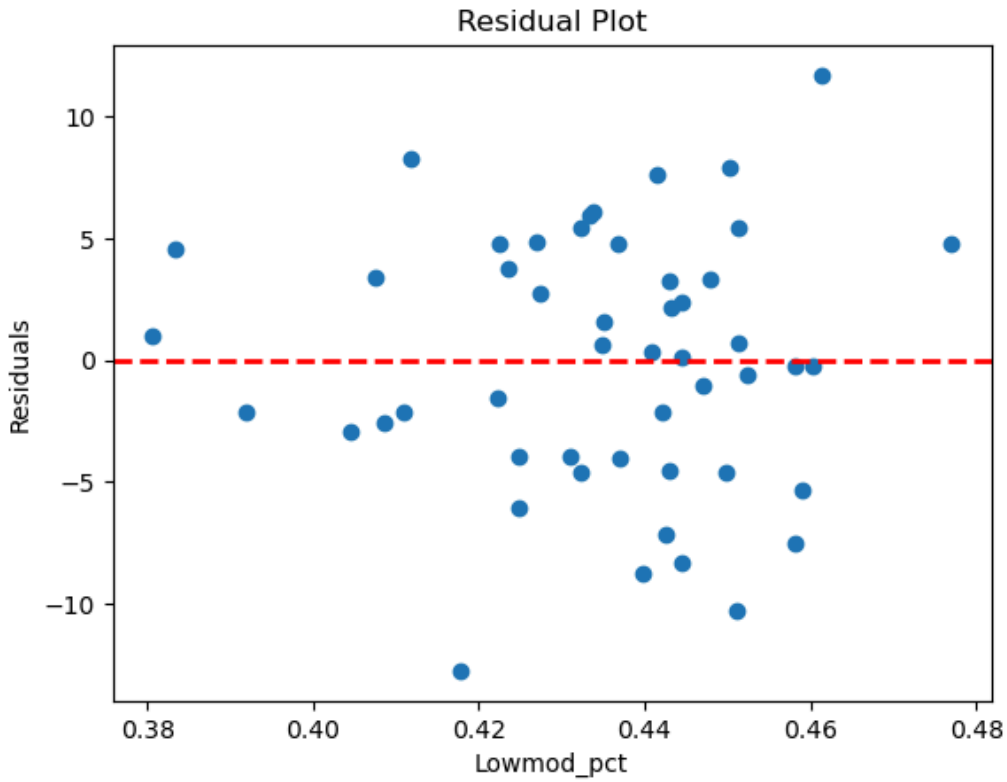
```
# 3. merge averaged_low_mod_income_df to qol_df on state
# Select the 'TotalScore' column from qol_df
total_score_column = qol_df[['state', 'TotalScore']]

# Merge result_df with the 'TotalScore' column on 'State'
merged_total_score_df = pd.merge(result_df, total_score_column, left_on='State
                                 right_on='state', how='inner')

# Drop the redundant 'state' column
merged_total_score_df.drop('state', axis=1, inplace=True)

# Print the resulting DataFrame
print(merged_total_score_df.head())
```

```
   Lowmod_pct       State  TotalScore
0    0.417736      Alaska       40.93
1    0.449694     Alabama       45.61
2    0.444404    Arkansas       42.42
3    0.430944     Arizona       48.31
4    0.476940  California       52.03
```

To check the feasibility of the regression, we plot out the residual plot to check the randomness of residuals.

In [177…

```
# run the linear regression
# independent variable is the averaged Lowmod_pct for each state
# dependent variable would be TotalScore for each state

X = merged_total_score_df[['Lowmod_pct']]
y = merged_total_score_df['TotalScore']

# Reshape X to make it a 2D array
X = X.values.reshape(-1, 1)

# Initialize the linear regression model
model = LinearRegression()

# Fit the model to the entire dataset
model.fit(X, y)

# Make predictions on the entire dataset
y_pred = model.predict(X)

# Calculate the residuals
residuals = y - y_pred

# Plot the residuals
plt.scatter(X, residuals)
plt.axhline(y=0, color='r', linestyle='--', linewidth=2)  # Add a horizontal l
plt.title('Residual Plot')
plt.xlabel('Lowmod_pct')
plt.ylabel('Residuals')
plt.show()
```



To see the randomness of residuals, we conducted a residual plot. Since there is no pattern within the residual plot, we can conclude that the residuals are randomly scattered around

zero. This indicates that the linear regression model is a good fit for the data and that there are no systematic deviations from the model.

In [178…
```
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

# Print the results summary
print(model.summary())
```

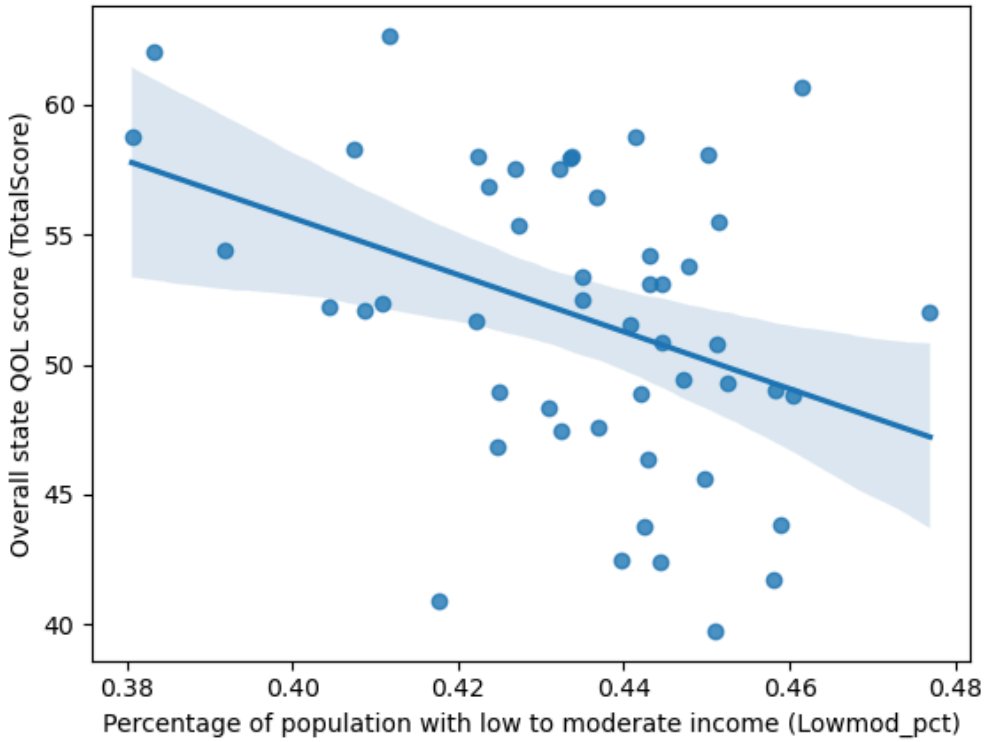
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             TotalScore   R-squared:                       0.145
Model:                            OLS   Adj. R-squared:                  0.127
Method:                 Least Squares   F-statistic:                     8.112
Date:                Mon, 04 Dec 2023   Prob (F-statistic):            0.00646
Time:                        21:57:55   Log-Likelihood:                -153.81
No. Observations:                  50   AIC:                             311.6
Df Residuals:                      48   BIC:                             315.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         99.4686     16.743      5.941      0.000      65.804     133.133
x1          -109.5654     38.469     -2.848      0.006    -186.913     -32.218
==============================================================================
Omnibus:                        0.623   Durbin-Watson:                   1.965
Prob(Omnibus):                  0.732   Jarque-Bera (JB):                0.746
Skew:                          -0.212   Prob(JB):                        0.689
Kurtosis:                       2.577   Cond. No.                         60.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

The linear regression model indicates a statistically significant negative relationship between the proportion of low-to-moderate income population (x1) and the Total Quality of Life (QOL) score (dependent variable). This means that as the proportion of lower-income individuals in a region increases, the overall quality of life would decrease for that region tends to decrease. The p-value for the coefficient of x1 is less than 0.006, which means that there is strong evidence to reject the null hypothesis of no relationship between the two variables.

The coefficient for x1 is -109.5654, which means that for every one-unit increase in the proportion of low-to-moderate income population, the Total QOL score is expected to decrease by 109.5654 units. This suggests that the proportion of lower-income individuals is a significant factor in determining the overall QOL of a region.

In [179…
```
# plot out the regression
ax = sns.regplot(data=merged_total_score_df, x="Lowmod_pct", y="TotalScore", \
                 n_boot=1000)
ax.set_ylabel("Overall state QOL score (TotalScore)")
ax.set_xlabel("Percentage of population with low to moderate income (Lowmod_pc
```

Out[179]:
```
Text(0.5, 0, 'Percentage of population with low to moderate income (Lowmod_pc
t)')
```



The scatter plot visually confirms the negative correlation between the proportion of low-to-moderate income population and the Total QOL score. The downward trend of the data points indicates that as the proportion of lower-income individuals increases, the QOL score tends to decrease. The spread of the data points around the trendline suggests that there is some variation in the relationship, but the overall trend is clear. This relationship could be due to several factors, such as the fact that lower-income individuals may have less access to quality education, healthcare, and other amenities that contribute to QOL.

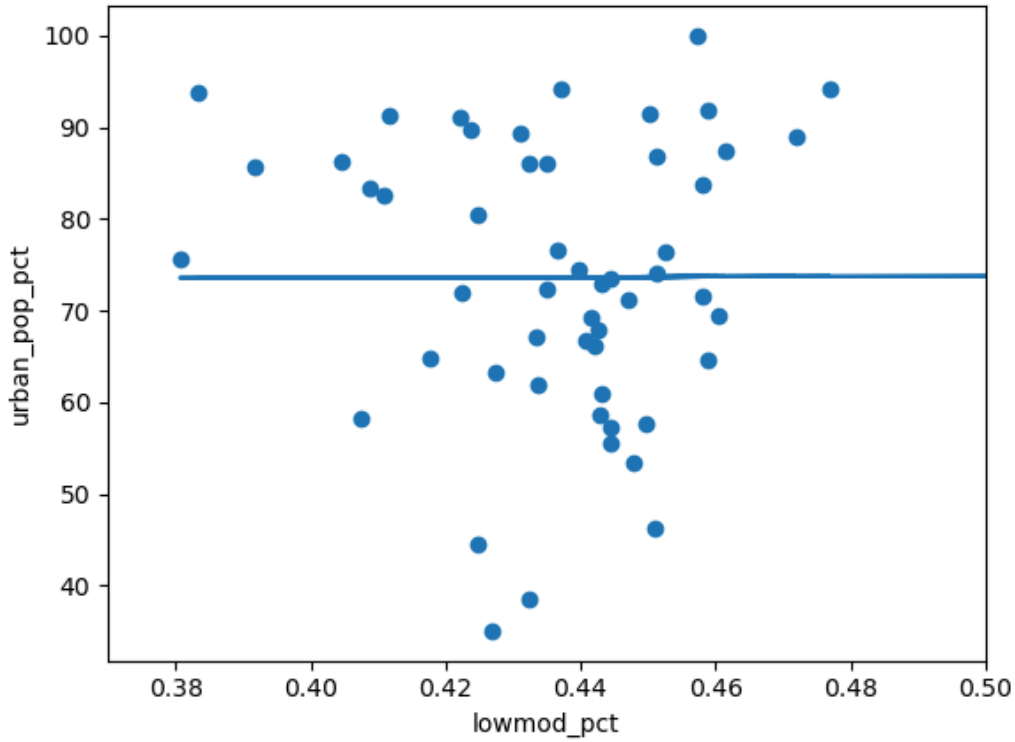
Since we found that there is a strong correlation between lower-income population and Quality of Life, we want to explore the relationship bewteen the proportion of low-to-moderate income population and proportion of urbanized area in-state. To do so, we decided to conduct a linear regression model.

In [182…

```
# plot for correlation between proportion of low-to-moderate income population

urban_pop_df.rename(columns={'2020 PCT URBAN POP': 'URBAN POP %', \
                             'STATE ABBREV':'State'}, inplace=True)
urban_pop_income_df = pd.merge(left=result_df, right=urban_pop_df, \
                               on='State', how='outer')

urban_pop_income_df.dropna(inplace=True)
print(urban_pop_income_df.head())

lowmod_pct = urban_pop_income_df['Lowmod_pct'].values.reshape(-1,1)
urban_pop_pct = urban_pop_income_df['URBAN POP %'].values.reshape(-1,1)

model = LinearRegression().fit(lowmod_pct, urban_pop_pct)
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)

plt.scatter(lowmod_pct, urban_pop_pct)
plt.plot(lowmod_pct, model.predict(lowmod_pct), label='Linear Regression Line'
plt.xlim(0.37,0.5)
```

```
plt.xlabel('lowmod_pct')
plt.ylabel('urban_pop_pct')
plt.show()
```

```
   Lowmod_pct           State  URBAN POP %
0    0.417736          Alaska         64.9
1    0.449694         Alabama         57.7
2    0.444404        Arkansas         55.5
3    0.825649  American Samoa         75.7
4    0.430944         Arizona         89.3
Coefficients: [[1.63074223]]
Intercept: [72.94885365]
```



Since there is no clear correlation between the two variables, we tried transformations on each variable, or both variable. There was still no clear correlation even after the transformations. Therefore, we conclude that the proportion of low-to-moderate income population is not related to the proportion of urbanzied population.

Overall Interpretations

The coefficient of Lowmod (-109.57) represents the estimated change in TotalScore for each one-unit increase in Lowmod. The significance of the coefficient is highly significant, as indicated by the associated p-value (Prob (F-statistic)), which is less than 0.05 (0.00646), suggesting that the overall model is significant. Therefore, combining the results from the plot, these imply a negative relationship between the percentage of low-to-moderate population and the score of quality of life.

However, the proportion of the variance in the dependent variable (TotalScore) that is predictable from the independent variable (Lowmod), which is indicated by the R-squared value (0.145), is relatively low. This indicates that, although the percentage of low-to-moderate income population explains some of the variance in the quality of life, there is still a substantial amount of unexplained variance in the dependent variable.

This result is consistent with our alternative hypothesis, where living in regions marked by reduced income levels are associated with a decreased quality of life. It also suggests that more aspects of quality of life should be considered.

Limitations

The resolution of this analysis is too large (by the state level) since we don't have more refined data for quality of life. If we have the data for quality of life by the county level, the conclusion drawn from the analysis will be more accurate.

Conclusion

In conclusion, several specific characteristics were identified differentiating between urban and non-urban (suburban or rural) regions within cities, and these factor are found associated with the well-being of residents. First, commute mode of residents are shown to be associated with the quality of life. In our hypothesis 1, we concluded that waking, transit, and bicycle show positive relationship to the quality of life, whereas automobile has a negative association.

Second, we found a correlation between the distribution of commute modes and the percentage of urbanization. Our analysis reveals a positive correlation between urban population percentage and walkability scores, as well as negative correlations between urban population percentage and the percentage of commuters using automobiles in the analyzed states, implying that decrease in urban population shows higher reliaance on automobiles. Given the correlation results from testing various commute modes in Hypothesis 1 (negative for automobiles, positive for walking) and considering that commuting by automobiles negatively correlates with quality of life while walking has a positive correlation, we can infer that urban regions exhibit increased walkability scores and a reduced reliance on automobiles. This suggests that residing in urban areas might be advantageous for overall quality of life by mitigating the negative impacts associated with automobile commuting.

Lastly, although we found that living in regions marked by reduced income levels are associated with a decreased quality of life, we didn't find clear association between urbanization to income level. Therefore, we couldn't conclude if living in urban areas are associated with better quality of life in terms of the difference in income level population.

Final Thoughts

Overall, determining the beneficial and detrimental effects that living in urban or non-urban environments is a complex issue that is confounded by a wide variety of factors, much more than we could sufficiently explore in this project. Generally speaking, the defining characteristics expected of either urban or non-urban areas, such as extent of population density, road network density, availability of alternative transportation besides automobiles, etc. will vary depending on which city or state that region is a part of. Governing bodies and local community cultures may also dictate the growth, well-being and quality of life of that region's residents.

In fact, what may be considered adverse effects of living in an urban area (for example, noise pollution) or in a surburban area (for example, social isolation and lack of pedestrian amenities) may not necessarily be dealbreaking considerations for different individuals, when weighing the benefits between different residential areas. Ultimately, however, it is a worthwhile topic to explore under the context of sociology, urban planning, and human geography as we strive to continue modeling our cities and neighborhoods around our needs, instead of the other way around.

Appendix

- data-cleaning.ipynb for pre-cleaning of original datasets
- `cleaning` folder contains both original datasets and cleaned datasets