ESS 330 Final Project Report - Lab 9

# Urbanization, Density and Access to Public Parks in the United States

## Abstract

### Introduction

Since 2008, the majority of the world’s population has lived in urban areas, a result of urbanization in developing countries (Beall et al., 2010; Kohlhase, 2013). The United States developed earlier than many nations, with more than 50 percent of the population living in urban areas.

|  |
| --- |
| Definition of Urban Areas in Census History |
| Prior to the 2020 Census urban areas were defined as any area with greater than 2500 people. Following the 2010 census urban clusters described areas with populations greater than 2,500 and less than 50,000; urbanized areas described areas with a population greater than 50,000. For the 2020 Census the threshold was changed to 5000 people (Ratcliffe, 2022). |

by the 14th Census in 1920. In the century since the 1920 census the percentage of individuals living in urban areas has increased to 80.7% (Slack & Jensen, 2020). As more people moved to urban areas, those areas expanded forming urbanized areas and large cities.

Urban planning has existed for centuries out of necessity, historically dominated by efficiency and utilitarianism, optimizing the world we live in for profitability, corporate productivity, and automobile-based mobility. This optimization came with sacrifices, which now impact an increasingly large majority of the population. In recent years, the discipline has begun to prioritize human factors over utilitarian efficiency. Thousands of years of living in rural settings makes urban living hard for most people’s biology. Connection to nature and time outdoors even in small amounts has been shown to be a vital part maintaining physical and mental health [DOTHIS:: Find additional source]. In an effort to make urban spaces more livable, planners are turning to parks and natural areas to connect people to nature.

Equity issues aside, overturning and correcting more than a century of bad planning is a daunting task. Many cities filled in and built up over the course of the 20th century as land became a premium commodity [DOTHIS:: Find additional source]. Does this density present significant challenges for today’s planning professionals? This research seeks to investigate the relationship between urban demographics like density and park access. In exploring this relationship, we hypothesize that there is an intermediate/sublinear relationship between urban population density and public open space availability.

### Data Overview

This report uses data from the UN-Habitat Urban Indicators Database and the ParkServe® Database maintained by the Trust for Public Land. The UN data relates to the UN-SDG 11.7.1 pertaining to access to open spaces and green areas.

The January 2025 version of the UN Open Spaces and Green Areas data includes the average share of urban areas allocated to streets and open public spaces as well as the share of the urban population with convenient access to an open public space.

|  |
| --- |
| UN Definition |
| In this case, the UN defines “convenient access to an open public space” as the “urban population withing 400 meters walking distance along the street network to an open public space” (May et al., 2000). |

These data collected by the UN were collected in 2020 and provided as a .xls format spreadsheet. These data were converted to .csv format with Microsoft Excel. The ParkServe® data selected for use is the 2020 data set to match the year the UN data was recorded. Specifically, this report uses elements of the City Park Facts: Acreage & Park System Highlights. The ParkServe® data is much less synthesized and was available as a .xml file. The file was structured for viewing as a spreadsheet rather than for further analysis and included multiple worksheets withing the workbook. In converting the file to a .csv file, the data spread across multiple worksheets was collated in a single worksheet and converted to a summarized dataset .csv file.

These data are lacking a shared numerical position data type but share a city name column formatted as “city\_name, two\_letter\_state\_abbreviation”. There is not perfect overlap between cities with data in each database however, there are 25 cities shared between the datasets. Cities present in only one data set will be culled when the data is joined.

### Methods

1. Clean the data. The raw data were downloaded as Excel spreadsheets, some reformatting in Excel was required to effectively exporting as a .csv file and importing the new summarized file to RStudio. Remaining data cleaning will occur in R as needed including any header changes or additional columns needed.
2. Conduct Exploratory Data Analysis (EDA).
3. Join datasets by “city name” to have a complete working dataset. These data will be combined into a single data frame with an inner join because there is a large number of cities listed in one data set but not the other. The new dataset will include only cities found in both datasets, with columns from both.

|  |
| --- |
| Limiting Scope |
| The cities found in only one dataset will be cut from the data to accommodate the limited scope of the project. With a bigger scope it is possible that additional data could be used to understand these patterns with more depth. |

1. Prep data and split it into training and testing datasets. Perform a 10-fold cross-validation on training data.
2. Create a recipe.
3. Set up several models in regression mode.
4. Create a workflow set including the previously written models and the recipe.
5. Map function over workflow using workflow map.
6. Using the highest performing model, fit the data and augment.
7. Plot and graph data to visually display test results.
8. Explore using the model to predict values for cities included in only one document (if time allows).

## Exploratory Data Analysis (EDA)

The data has already been discussed in general terms in the data overview section. There will be a readme file created to elaborate on the sources, formatting and manipulation required for each dataset before joining them into the urban\_parks\_data data frame. In general terms, prior to importing into RStudio, the .xml files the data came in were opened in MS Excel; the sheets were formatted to be converted to .csv files including condensing multiple worksheets of the ParkServe® data into a sinigle sheet for easier conversion to a .csv file. While some of the cleaning done in Excel could have been completed in RStudio, it was not efficient to do so. Using Excel was faster and more flexible for that use case. Similar reformatting was required with the UN data as the headers were unreasonably long by default. The readme that will be created for each (or both) files will include a more detailed summary of what each variable means.

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(here)

Warning: package 'here' was built under R version 4.4.3

here() starts at C:/Users/Zacha/github/CSU/ESS 330/ess\_330\_final\_project

library(flextable)

Attaching package: 'flextable'  
  
The following object is masked from 'package:purrr':  
  
 compose

library(patchwork)

Warning: package 'patchwork' was built under R version 4.4.3

# Create some visualizations and descriptions of what data you have, where you got it, and how and if you need to clean and manipulate it for your project  
  
# Import data from csvs and clean NAs  
parkserve\_data <- read\_csv(here("data", "clean\_data/parkserve\_summarized\_facts\_2020.csv")) %>%   
 drop\_na()

Rows: 100 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (10): city\_name, city\_pop, land\_area, revised\_area, percent\_designed\_par...  
dbl (5): parkland\_area, designed\_park\_area, natural\_park\_area, parkland\_per...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

un\_land\_use\_data <- read\_csv(here("data", "clean\_data/un\_land\_use.csv")) %>%   
 drop\_na()

Rows: 59 Columns: 3  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): city\_name  
dbl (2): mean\_percent\_built\_open\_space, mean\_percent\_open\_space\_access  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Add columns and finish cleaning parkserve data  
clean\_parkserve\_data <- parkserve\_data %>%  
 mutate(  
 across(-city\_name, ~ as.numeric(.x)), # Convert all columns to numeric except for city name  
 parkland\_percent = parkland\_percent \* 100, # Convert parkland percent from ratio  
 # Fix design/natural park area percentage calculations  
 percent\_designed\_parks = ifelse(parkland\_area == 0, NA, (designed\_park\_area / parkland\_area) \* 100),  
 percent\_natural\_parks = ifelse(parkland\_area == 0, NA, (natural\_park\_area / parkland\_area) \* 100),  
 #New calculation for designed-natural area ratio  
 dn\_area\_ratio = ifelse(percent\_natural\_parks == 0, NA, percent\_designed\_parks / percent\_natural\_parks)  
 )  
  
   
   
# Join data removing cities found in only one of the two datasets  
urban\_parks\_data <- clean\_parkserve\_data %>%   
 inner\_join(un\_land\_use\_data, by = "city\_name")   
  
# Basic data structure exploration  
glimpse(urban\_parks\_data)

Rows: 25  
Columns: 18  
$ city\_name <chr> "Anchorage, AK", "Atlanta, GA", "Boston…  
$ city\_pop <dbl> 299100, 498059, 687725, 2744859, 377963…  
$ land\_area <dbl> 1090997, 85217, 30897, 145686, 49726, 2…  
$ revised\_area <dbl> 1086019, 84250, 29175, 136796, 46880, 2…  
$ parkland\_area <dbl> 914138, 5293, 5072, 13609, 3170, 20352,…  
$ designed\_park\_area <dbl> 2417, 3864, 2556, 8593, 1792, 10974, 40…  
$ natural\_park\_area <dbl> 911721, 1429, 2516, 4430, 1378, 9378, 1…  
$ parkland\_percent <dbl> 84.173297, 6.282493, 17.384747, 9.94839…  
$ percent\_designed\_parks <dbl> 0.2644021, 73.0020782, 50.3943218, 63.1…  
$ percent\_natural\_parks <dbl> 99.735598, 26.997922, 49.605678, 32.551…  
$ pop\_density <dbl> 0.28, 5.91, 23.57, 20.07, 8.06, 6.39, 7…  
$ parkland\_per\_1k\_pop <dbl> 3056.295553, 10.627255, 7.375041, 4.744…  
$ park\_units <dbl> 228, 416, 373, 645, 179, 397, 302, 70, …  
$ park\_units\_per\_10k\_pop <dbl> 7.622869, 8.352424, 5.423680, 2.349847,…  
$ percent\_half\_mile\_walk <dbl> 0.75091, 0.72447, 0.99790, 0.98220, 0.8…  
$ dn\_area\_ratio <dbl> 0.00265103, 2.70398880, 1.01589825, 1.9…  
$ mean\_percent\_built\_open\_space <dbl> 22.0, 13.8, 19.8, 17.2, 20.0, 20.4, 20.…  
$ mean\_percent\_open\_space\_access <dbl> 71.9, 21.2, 68.2, 47.8, 29.8, 41.9, 43.…

# Descriptive Stats  
  
 # Write function to round numeric columns to two decimal places  
 round\_numeric <- function(df) {  
 df %>%   
 mutate(across(where(is.numeric), ~round(.x, 2)))  
 }  
  
 # Summarize stats by variable  
 desc\_stats\_parks <- urban\_parks\_data %>%   
 select(where(is.numeric)) %>%   
 pivot\_longer(cols = everything(), names\_to = "variable", values\_to = "value") %>%   
 group\_by(variable) %>%   
 summarize(mean = mean(value, na.rm = TRUE),   
 median = median(value, na.rm = TRUE),   
 sd = sd(value, na.rm = TRUE),  
 Q1 = quantile(value, 0.25, na.rm = TRUE),  
 Q3 = quantile(value, 0.75, na.rm = TRUE)) %>%   
 round\_numeric()  
   
 # Print descriptive stats with flextable  
 desc\_stats\_flex <- flextable(desc\_stats\_parks) %>%  
 set\_caption("Summarized Urban Parks Statistics") %>%   
 set\_header\_labels(  
 variable = "Variable",  
 mean = "Mean",  
 median = "Median",  
 sd = "Standard Deviation",  
 Q1 = "1st Quartile (Q1)",  
 Q3 = "3rd Quartile (Q3)") %>%   
 autofit()  
  
# Find Top/Bottom cities for percent parkland  
   
 # Select relevant columns  
 simplified\_vars <- c("city\_name", "city\_pop", "revised\_area", "pop\_density", "parkland\_area", "dn\_area\_ratio", "parkland\_percent", "parkland\_per\_1k\_pop", "percent\_half\_mile\_walk")  
   
 # Filter top/bottom 10 cities  
 top10\_park\_percent <- urban\_parks\_data %>%   
 arrange(desc(parkland\_percent)) %>%   
 slice\_head(n = 10) %>%   
 select(all\_of(simplified\_vars)) %>%   
 round\_numeric()  
   
 bottom10\_park\_percent <- urban\_parks\_data %>%   
 arrange(parkland\_percent) %>%   
 slice\_head(n = 10) %>%   
 select(all\_of(simplified\_vars)) %>%   
 round\_numeric()  
   
 # Create top/bottom 10 flextables w/ function  
 # Create function  
 make\_best\_worst\_flextbl <-function(df, caption) {  
 flextable(df) %>%   
 set\_caption(caption) %>%  
 set\_header\_labels(  
 city\_name = "City Name",  
 city\_pop = "City Population",  
 revised\_area = "City Land Area (Revised) (Acres)",  
 pop\_density = "Population Density (People/Acre)",  
 parkland\_area = "City Parkland Area (Acres)",  
 parkland\_percent = "Percent Parkland",  
 parkland\_per\_1k\_pop = "Parkland Per (1000) Capita",  
 percent\_half\_mile\_walk = "Percent of Residents within 0.5 Miles of a Park",  
 dn\_area\_ratio = "Designed-Natural Park Area Ratio (Designed Park (%) / Natural Park (%)") %>%   
 autofit()  
 }  
   
 top10\_park\_percent\_flex <- make\_best\_worst\_flextbl(top10\_park\_percent, "Top 10 Cities for Parkland Percentage")  
   
 bottom10\_park\_percent\_flex <- make\_best\_worst\_flextbl(bottom10\_park\_percent, "Top 10 Cities for Parkland Percentage")  
   
# Make plots to visualize the data  
# Histogram: Land Area  
land\_area\_plot <- ggplot(urban\_parks\_data, aes(x = as.numeric(land\_area))) +  
 geom\_histogram(bins = 20, fill = "steelblue", color = "white") +  
 labs(x = "Land Area (Acres)", y = "Frequency", title = "Distribution of City Land Areas") +  
 theme\_minimal()  
  
# Scatterplot: Land Area vs Parkland Area  
land\_vs\_park\_area\_plot <- ggplot(urban\_parks\_data, aes(x = as.numeric(land\_area), y = parkland\_area)) +  
 geom\_point(color = "forestgreen") +  
 labs(x = "Land Area (Acres)", y = "Parkland Area\n(Acres)", title = "Land Area vs Parkland Area") +  
 scale\_x\_continuous(labels = scales::label\_comma()) +  
 scale\_y\_continuous(labels = scales::label\_comma()) +  
 theme\_minimal()  
  
# Scatterplot: Population Density vs Parkland Percent  
density\_vs\_park\_percent\_plot <- ggplot(urban\_parks\_data, aes(x = as.numeric(pop\_density), y = parkland\_percent)) +  
 geom\_point(color = "darkorange") +  
 labs(x = "Population Density (People/Acre)", y = "Percent Parkland", title = "Population Density vs\nParkland Percent") +  
 scale\_x\_continuous(labels = scales::label\_comma()) +  
 scale\_y\_continuous(labels = scales::label\_comma()) +  
 theme\_minimal()  
  
# Scatterplot: Designed-Natural Park Area Ratio vs Parkland Percent  
dn\_area\_ratio\_vs\_park\_percent\_plot <- ggplot(urban\_parks\_data, aes(x = dn\_area\_ratio, y = parkland\_percent)) +  
 geom\_point(color = "mediumvioletred") +  
 labs(x = "Designed-Natural Park Area Ratio", y = "Percent Parkland", title = "Designed-Natural Park Area\nRatio vs Parkland Percent") +  
 scale\_x\_continuous(labels = scales::label\_comma()) +  
 scale\_y\_continuous(labels = scales::label\_comma()) +  
 theme\_minimal()  
  
# Scatterplot: Percent Designed Parks vs Percent Natural Parks  
designed\_vs\_natural\_parks\_plot <- ggplot(urban\_parks\_data, aes(x = percent\_designed\_parks, y = percent\_natural\_parks)) +  
 geom\_point(color = "cadetblue") +  
 labs(x = "Percent Designed Parks", y = "Percent Natural\nParks", title = "Percent Designed vs Natural Parks") +  
 theme\_minimal()  
  
# Scatterplot: Percent Open Space Access vs Percent Built Open Space  
open\_space\_access\_vs\_built\_plot <- ggplot(urban\_parks\_data, aes(x = mean\_percent\_open\_space\_access, y = mean\_percent\_built\_open\_space)) +  
 geom\_point(color = "seagreen") +  
 labs(x = "Mean % Open Space Access", y = "Mean % Built\nOpen Space", title = "Accessibility of Built Open Space") +  
 theme\_minimal()  
  
# Combine all plots in one figure using patchwork (optional)  
all\_eda\_plots <- (land\_area\_plot | land\_vs\_park\_area\_plot) /   
 (density\_vs\_park\_percent\_plot | dn\_area\_ratio\_vs\_park\_percent\_plot) /   
 (designed\_vs\_natural\_parks\_plot| open\_space\_access\_vs\_built\_plot) +  
 plot\_layout(guides = "collect")  
  
# Display data summary and visualization  
  
 # Display flextables  
 desc\_stats\_flex

| Variable | Mean | Median | Standard Deviation | 1st Quartile (Q1) | 3rd Quartile (Q3) |
| --- | --- | --- | --- | --- | --- |
| city\_pop | 1,113,559.56 | 655,061.00 | 1,696,534.55 | 377,963.00 | 1,006,142.00 |
| designed\_park\_area | 5,559.20 | 3,864.00 | 5,020.97 | 2,652.00 | 5,785.00 |
| dn\_area\_ratio | 2.42 | 1.02 | 4.90 | 0.26 | 1.56 |
| land\_area | 178,285.40 | 88,800.00 | 224,466.29 | 53,723.00 | 201,635.00 |
| mean\_percent\_built\_open\_space | 18.71 | 18.20 | 3.23 | 17.20 | 20.60 |
| mean\_percent\_open\_space\_access | 42.06 | 40.00 | 17.04 | 29.80 | 52.80 |
| natural\_park\_area | 48,113.72 | 4,538.00 | 180,621.84 | 1,429.00 | 22,527.00 |
| park\_units | 460.80 | 302.00 | 811.06 | 179.00 | 416.00 |
| park\_units\_per\_10k\_pop | 4.55 | 4.54 | 1.72 | 3.28 | 5.42 |
| parkland\_area | 53,824.36 | 9,478.00 | 180,158.96 | 5,075.00 | 28,312.00 |
| parkland\_per\_1k\_pop | 143.17 | 12.09 | 607.26 | 9.15 | 28.14 |
| parkland\_percent | 15.31 | 12.54 | 15.58 | 6.76 | 17.38 |
| percent\_designed\_parks | 44.66 | 50.28 | 27.85 | 20.43 | 60.90 |
| percent\_half\_mile\_walk | 0.76 | 0.80 | 0.20 | 0.63 | 0.96 |
| percent\_natural\_parks | 53.78 | 49.61 | 28.29 | 32.55 | 79.57 |
| pop\_density | 9.52 | 6.39 | 9.84 | 3.59 | 12.41 |
| revised\_area | 175,389.60 | 87,844.00 | 222,631.64 | 52,765.00 | 196,098.00 |

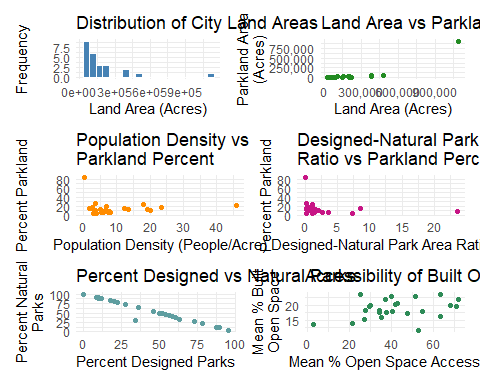
top10\_park\_percent\_flex

| City Name | City Population | City Land Area (Revised) (Acres) | Population Density (People/Acre) | City Parkland Area (Acres) | Designed-Natural Park Area Ratio (Designed Park (%) / Natural Park (%) | Percent Parkland | Parkland Per (1000) Capita | Percent of Residents within 0.5 Miles of a Park |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Anchorage, AK | 299,100 | 1,086,019 | 0.28 | 914,138 | 0.00 | 84.17 | 3,056.30 | 0.75 |
| New Orleans, LA | 386,105 | 107,655 | 3.59 | 27,775 | 0.11 | 25.80 | 71.94 | 0.80 |
| Washington, DC | 702,321 | 38,955 | 18.03 | 9,478 | 1.09 | 24.33 | 13.50 | 0.98 |
| New York, NY | 8,627,852 | 187,946 | 45.91 | 40,190 | 1.01 | 21.38 | 4.66 | 0.99 |
| San Diego, CA | 1,399,844 | 205,918 | 6.80 | 39,385 | 0.29 | 19.13 | 28.14 | 0.81 |
| Virginia Beach, VA | 457,832 | 159,341 | 2.87 | 28,312 | 0.26 | 17.77 | 61.84 | 0.68 |
| Boston, MA | 687,725 | 29,175 | 23.57 | 5,072 | 1.02 | 17.38 | 7.38 | 1.00 |
| Honolulu, HI | 1,006,142 | 379,885 | 2.65 | 57,141 | 0.09 | 15.04 | 56.79 | 0.79 |
| Minneapolis, MN | 421,339 | 33,958 | 12.41 | 5,075 | 8.47 | 14.94 | 12.04 | 0.98 |
| Jacksonville, FL | 925,142 | 467,298 | 1.98 | 67,707 | 0.14 | 14.49 | 73.19 | 0.35 |

bottom10\_park\_percent\_flex

| City Name | City Population | City Land Area (Revised) (Acres) | Population Density (People/Acre) | City Parkland Area (Acres) | Designed-Natural Park Area Ratio (Designed Park (%) / Natural Park (%) | Percent Parkland | Parkland Per (1000) Capita | Percent of Residents within 0.5 Miles of a Park |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Durham, NC | 275,758 | 68,678 | 4.02 | 2,665 | 0.11 | 3.88 | 9.66 | 0.51 |
| Memphis, TN | 655,061 | 196,098 | 3.34 | 9,194 | 1.10 | 4.69 | 9.15 | 0.46 |
| Winston-Salem, NC | 248,839 | 83,917 | 2.97 | 4,263 | 7.34 | 5.08 | 17.13 | 0.37 |
| Detroit, MI | 660,960 | 87,844 | 7.52 | 5,102 | 3.63 | 5.81 | 7.72 | 0.80 |
| Toledo, OH | 277,467 | 51,169 | 5.42 | 3,175 | 1.41 | 6.20 | 11.44 | 0.81 |
| Atlanta, GA | 498,059 | 84,250 | 5.91 | 5,293 | 2.70 | 6.28 | 10.63 | 0.72 |
| Cleveland, OH | 377,963 | 46,880 | 8.06 | 3,170 | 1.30 | 6.76 | 8.39 | 0.83 |
| Dallas, TX | 1,378,903 | 215,676 | 6.39 | 20,352 | 1.17 | 9.44 | 14.76 | 0.71 |
| St. Louis, MO | 310,144 | 39,090 | 7.93 | 3,749 | 23.66 | 9.59 | 12.09 | 0.98 |
| Chicago, IL | 2,744,859 | 136,796 | 20.07 | 13,609 | 1.94 | 9.95 | 4.74 | 0.98 |

# Display patchwork plots  
 all\_eda\_plots



Source: [Article Notebook](https://zcramton.github.io/ess_330_final_project/index.qmd.html)

## References

Beall, J., Guha-Khasnobis, B., & Kanbur, R. (2010). *Urbanization and development: Multidisciplinary perspectives*. Oxford University Press.

Kohlhase, J. E. (2013). The new urban world 2050: Perspectives, prospects and problems. *Regional Science Policy & Practice*, *5*(2), 153–166.

May, R., Rex, K., Bellini, L., Sadullah, S., Nishi, E., James, F., & Mathangani, A. (2000). UN habitat indicators database: Evaluation as a source of the status of urban development problems and programs. *Cities*, *17*(3), 237–244.

Ratcliffe, M. (2022). Redefining Urban Areas following the 2020 Census. In *Census.gov*. <https://www.census.gov/newsroom/blogs/random-samplings/2022/12/redefining-urban-areas-following-2020-census.html>

Slack, T., & Jensen, L. (2020). The changing demography of rural and small-town america. *Population Research and Policy Review*, *39*(5), 775–783.