

Tracking COVID-19 Cases Across Different Settlement Types in New York State

https:
[//github.com/sgm22/MiltonZouKhalsa___ENV872_EDA_FinalProject](https://github.com/sgm22/MiltonZouKhalsa___ENV872_EDA_FinalProject)

SMiltonDZouGKhalsa

2022-11-22

Contents

0.1	Rationale and Research Questions	3
0.2	Dataset Information	6
0.3	Exploratory Analysis and Data Visualization	8
0.4	Analysis	22
0.5	Summary and Conclusions	34
0.6	Conclusion	35
0.7	References	36

List of Figures

1	Cumulative Number of Positive COVID-19 Tests	10
2	Log of Cumulative Number of Positive COVID-19 Tests	11
3	Cumulative Number of COVID-19 Tests Conducted	12
4	Log of Cumulative Number of Positive COVID-19 Tests Conducted	12
5	Population of Counties Across New York State	13
6	Log of Population of Counties Across New York State	14
7	Density of Counties Across New York State	15
8	Log of Density of Counties Across New York State	15
9	Log of Absolute Population Size of Counties by Settlement Type	16
10	Log of Population Density of Counties by Settlement Type	17
11	Population of New York Counties	19
12	Settlement Types of New York Counties	19
13	Cumulative Number of Tests Administered in Each New York County	20
14	Cumulative Number of Positive Test Results in Each New York County	21
15	Correlation Between Absolute Population Across Counties and COVID-19 Cases	23
16	Correlation Between Population Densities Across Counties and COVID-19 Cases	25
17	Correlation Between Absolute Population Across Counties and COVID-19 Tests Conducted	30
18	Correlation Between Absolute Population Densities Across Counties and COVID-19 Tests Conducted	31
19	Log of Cumulative COVID-19 Cases Across Settlement Types	33

0.1 Rationale and Research Questions

0.1.1 Rationale

Coronavirus disease (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Coronavirus, n.d.). The virus is primarily spread from person to person through respiratory droplets and as such, population density plays a role in the transmission of the virus with many experts suggesting that physical distancing is an effective way to combat its spread (CDCMMWR, 2020; What Is Coronavirus?, 2022). Since COVID-19 was declared a global pandemic in March 2020 by the World Health Organization (WHO) in March 2020, there has been a resurgence in COVID-19 infections across the globe as a result of new strains of the virus (CDC, 2020). Policy decisions to protect human health leverages data that such as case numbers, hospitalizations, and death resulting from COVID-19 to project COVID-19 spread (Truelove et al., n.d.). Very few epidemiological COVID-19 models incorporate population density explicitly even though population density is thought to be an important factor (Hamidi et al., 2020; Wong & Li, 2020).

In New York State, since the beginning of the pandemic there has been over 6 million cases of COVID-19. In January 2022, the state recorded the highest number of cases while in April 2020, it recorded the highest record of death (Times, 2020). The density of people in New York State varies drastically across the state with 43% of the state's population living in the five boroughs of New York City (New York State Tracking Program | Tracking | NCEH | CDC, 2019). Of the 62 counties in the State, Kings County is the most populated with over 2 million residents, while Hamilton County is the least populated with about 4,485 residents (New York Population 2022 (Demographics, Maps, Graphs), n.d.). With physical distancing playing an important role in reducing the likelihood of infection, this paper explores how well absolute population and population density predict COVID-19 case numbers in the diversely sized New York State. Further, we explore whether the relationship between absolute population and population density changes among different settlement types.

Settlement types are defined as population centers of humans who have developed a long-term community in a specific area (Settlements Overview & Types | What Are Settlements?, n.d.). In this study, settlement types are classified as rural, micropolitan, and urban. We define rural settlements as counties with a population size of less than 25,000 people, micropolitans as counties with a population size of between 25,000 and 49,999 people, and an urban settlement as a county with more than 50,000 people (adapted from USDA ERS - What Is Rural?, n.d.). We hope that by better understanding the importance of population size and density on COVID-19 prevalence and how that varies across settlement type, we are able to better understand if population density and/or absolute population can be used as a proxy when formulating policy decisions around COVID-19.

0.1.2 Research Questions

1. What is the correlation between population density and absolute population and the number of COVID-19 cases across New York State?
2. What is the relationship between the predictor variable: settlement types and the response variable: cumulative COVID-19 cases?
3. What is the correlation between absolute population and population density and the number of COVID-19 tests does done? Does the stronger correlated relationship vary among settlement types?

Set-up Working Directory, Import and Merge Data sets

In order to answer our research question, we imported data that contains information about New York (NY) state population size and density by county in 2022 from the U.S. Census Bureau. This information was then combined with data regarding number of positive COVID-19 cases, quantity of COVID-19 tests conducted and other relevant information collected by the New York Department of Health from March 1, 2020 to November 22, 2022.

0.2 Dataset Information

We acquired our 2020 to 2022 COVID-19 data from the New York Department of Health and population data from the U.S. Census Bureau on November 24th, 2022. County names did not match between the census data and NY Department of Health data. We wrangled our data by removing the word “county” from each county in the census data. In addition to the 62 counties, the NY Health Department broke counties down into smaller sections resulting in more than 62 counties. Since we only want to work with the original counties, we left-merged the NY health data onto the county data on November 24th, 2022. We are using the merged version as our main data source. Additionally on November 24th, 2022, we uploaded the geospatial shape file that was incorporated in class for ENV872. Using the shape file, we selected for counties in New York State. Data types and variable types for the combined data were set on November 27th, 2022.

Data Information: <https://www.census.gov/library/stories/state-by-state/new-york-population-change-between-census-decade.html> COVID-19 Variant Data | Department of Health (ny.gov) – (March 1, 2020, to November 22, 2022) cb_2018_us_county_20m.shp state FIPS code 36

Keywords

ENV872, COVID-19, COVID, virus, virus spread, New York, New York State

```
##                               Variable
## 1                             County
## 2                             pop2022
## 3                area_milessquared
## 4            density_milessquared
## 5                sum_new_positives
## 6 sum_cumulative_no_of_positives
## 7            sum_total_number_of_tests
## 8            sum_cumulative_no_of_tests
## 9                median_test_positive
## 10                settlement_type
```

```
##
```

```
## 1
```

```
## 2
```

```
## 3
```

```
## 4
```

```
## 5
```

```
## 6
```

```
## 7
```

```
## 8
```

```
## 9 For the dates before 4/4/2022: The median percentage of all COVID tests reported t
```

```
## 10
```

The

Running total for the number

##	Coding	Units	Minimum.Value	Maximum.Value	Missing
## 1	character	N/A	N/A	N/A	0
## 2	integer	N/A	5161	2782348	0
## 3	numeric miles squared		22.6558	2680.379	0
## 4	numeric miles squared		3.0052	75738.89	0
## 5	integer	N/A	1054	2908643	0
## 6	integer	N/A	432991	1224453000	0
## 7	integer	N/A	20261	60433650	0
## 8	integer	N/A	10356620	27893290000	0
## 9	character	N/A	N/A	N/A	0
## 10	character	N/A	N/A	N/A	0

0.3 Exploratory Analysis and Data Visualization

The merged data set was exploring for missing values and abnormal values using the R functions, “head”, “summary” and “dim”. Further, to better understand the distribution of the variables that we use in our analysis, we plotted a series of histograms. The graphical exploration of the distribution of the variables showed that the variables, “sum_cumulative_no_of_positive”, “sum_cumulative_no_of_tests”, “pop2022” and “density_permilessquared” are right-skewed. That is, most of the data points were located on the left side of the histograms plotted for each variable. For improved visualization, we log transformed the variables to get their distributions closer to a normal distribution. After each log-transformation, we carried out Shapiro–Wilk tests to check the normality of the transformed distributions. The p-value of each log-transformed distribution was less than 0.05 indicating that the evidence supported the fact that log-transformed distributions were still not normally distributed. We also explored and visualized the distribution of the absolute population and population density across the various settlement types using box plots.

0.3.1 Tabular Exploration

```
head(combined_dat_ny_covid_pop)
```

##	County	pop2022	area_milessquared	density_permilessquared
## 1	Kings	2782348	69.8126	39854.529
## 2	Queens	2440412	108.7681	22436.837
## 3	New York City	1715927	22.6558	75738.895
## 4	Suffolk	1532434	911.7219	1680.813
## 5	Bronx	1490164	42.0506	35437.365
## 6	Nassau	1407022	284.8100	4940.212
##	sum_new_positives	sum_cumulative_no_of_positive	sum_total_number_of_tests	
## 1	860573	362966136	18713833	
## 2	813384	345088691	14735060	
## 3	2908643	1224453265	60433651	
## 4	532466	238589467	8496337	
## 5	480247	214795050	8847438	
## 6	515433	225331004	8400598	
##	sum_cumulative_no_of_tests	median_test_positive	settlement_type	
## 1	8370878686	2.45%	Urban	
## 2	6902671989	2.49%	Urban	
## 3	27893288263	2.51%	Urban	
## 4	4261845093	3.43%	Urban	
## 5	4167554932	2.62%	Urban	


```
## 6                4188172601                2.99%                Urban
```

```
summary(combined_dat_ny_covid_pop) #View Data
```

```
##      County                pop2022      area_milessquared density_permilessquared
## Length:62      Min.   :   5161      Min.   :  22.66      Min.   :    3.01
## Class :character 1st Qu.:  48289      1st Qu.: 441.55      1st Qu.:   73.02
## Mode  :character Median :  86998      Median : 653.65      Median :  113.46
##                Mean   : 328482      Mean   : 760.09      Mean   : 3279.30
##                3rd Qu.: 236910      3rd Qu.:1022.40      3rd Qu.:  418.14
##                Max.   :2782348      Max.   :2680.38      Max.   :75738.89
## sum_new_positives sum_cumulative_no_of_positive sum_total_number_of_tests
## Min.   :   1054      Min.   :4.330e+05      Min.   :   20261
## 1st Qu.:  12905      1st Qu.:5.329e+06      1st Qu.:  240066
## Median :   23804      Median :9.156e+06      Median :   406342
## Mean   :  138799      Mean   :5.937e+07      Mean   :  2723930
## 3rd Qu.:   64600      3rd Qu.:2.758e+07      3rd Qu.: 1462341
## Max.   :2908643      Max.   :1.224e+09      Max.   :60433651
## sum_cumulative_no_of_tests median_test_positive settlement_type
## Min.   :1.036e+07      Length:62      Length:62
## 1st Qu.:1.241e+08      Class :character  Class :character
## Median :2.024e+08      Mode  :character  Mode  :character
## Mean   :1.304e+09
## 3rd Qu.:7.539e+08
## Max.   :2.789e+10
```

```
dim(combined_dat_ny_covid_pop)
```

```
## [1] 62 10
```

```
#View mean COVID cases and mean COVID test conducted by settlement types
```

```
collapsed_dat_settlement_types <- combined_dat_ny_covid_pop %>%
  group_by(settlement_type) %>%
  summarise(Mean_COVID_cases = mean(sum_cumulative_no_of_positive),
            Mean_Tests_Conducted = mean(sum_cumulative_no_of_tests),
            Mean_Absolute_Population = mean(pop2022),
            Mean_Settlement_Density = mean(density_permilessquared))

print(collapsed_dat_settlement_types) #View table
```

```
## # A tibble: 3 x 5
##   settlement_type Mean_COVID_cases Mean_Tests_Conducted Mean_Absolute_~1 Mean_~2
##   <chr>           <dbl>           <dbl>           <dbl>   <dbl>
## 1 Micropolitan    4166072.           98127075.        41448.    64.2
## 2 Rural           1274503.           34469996.        15877     43.4
## 3 Urban           80411031.          1763706740.       438622.  4495.
## # ... with abbreviated variable names 1: Mean_Absolute_Population,
## #   2: Mean_Settlement_Density
```

0.3.2 Graphical Exploration

```
# Visualizing Distribution of Variables
```

```
##### COVID-19 Cases #####
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(sum_cumulative_no_of_positive)) +
  geom_histogram() +
  xlab("Cumulative Number of Positive COVID-19 Tests") +
  ggtitle("Histogram of Cumulative Number of Positive COVID-19 Tests")
```

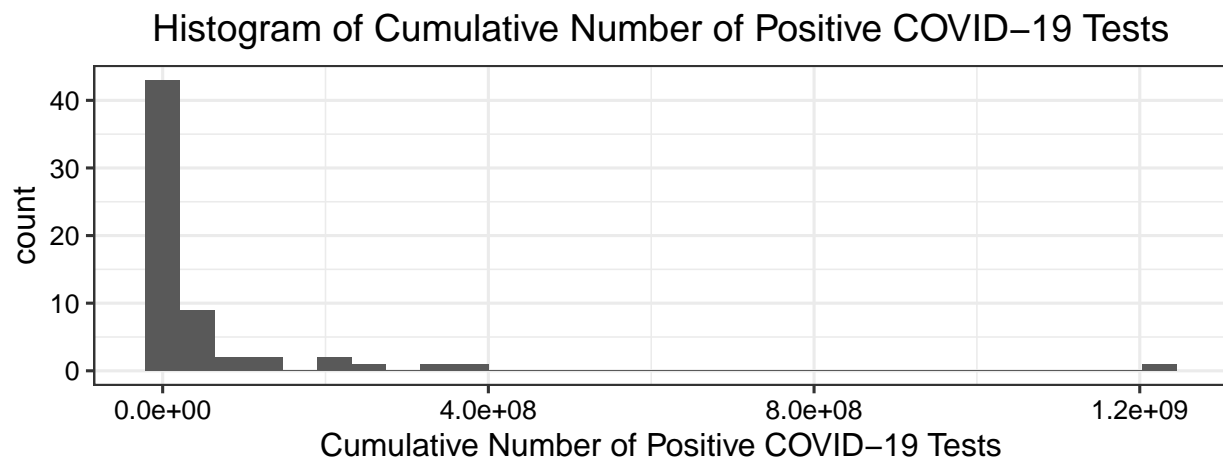


Figure 1: Cumulative Number of Positive COVID-19 Tests

```
## Log Transform
```

```
ggplot(data = combined_dat_ny_covid_pop,
  aes(log(sum_cumulative_no_of_positive))) +
  geom_histogram() +
  xlab("Log of Cumulative Number of Positive COVID-19 Tests") +
  ggtitle("Histogram of Log of Cumulative Number of Positive COVID-19 Tests")
```

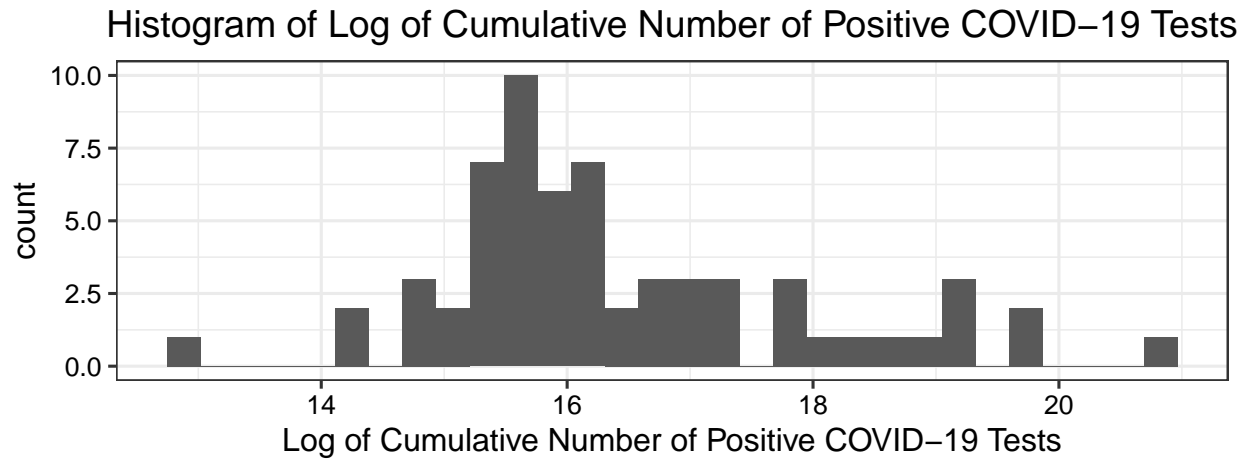


Figure 2: Log of Cumulative Number of Positive COVID-19 Tests

```
shapiro.test(log(combined_dat_ny_covid_pop$sum_cumulative_no_of_positive))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(combined_dat_ny_covid_pop$sum_cumulative_no_of_positive)
## W = 0.92854, p-value = 0.001407
```

#Check normality

Number of COVID Tests Conducted

```
ggplot(data = combined_dat_ny_covid_pop, aes(sum_cumulative_no_of_tests))+
  geom_histogram() +
  xlab ("Cumulative Number of COVID-19 Tests Conducted") +
  ggtitle("Histogram of Cumulative Number of COVID-19 Tests Conducted")
```

Log Transform

```
ggplot(data = combined_dat_ny_covid_pop, aes(log(sum_cumulative_no_of_tests))) +
  geom_histogram() +
  xlab ("Log of Cumulative Number of COVID-19 Tests Conducted") +
  ggtitle("Histogram of Log of Cumulative Number of COVID-19 Tests Conducted")
```

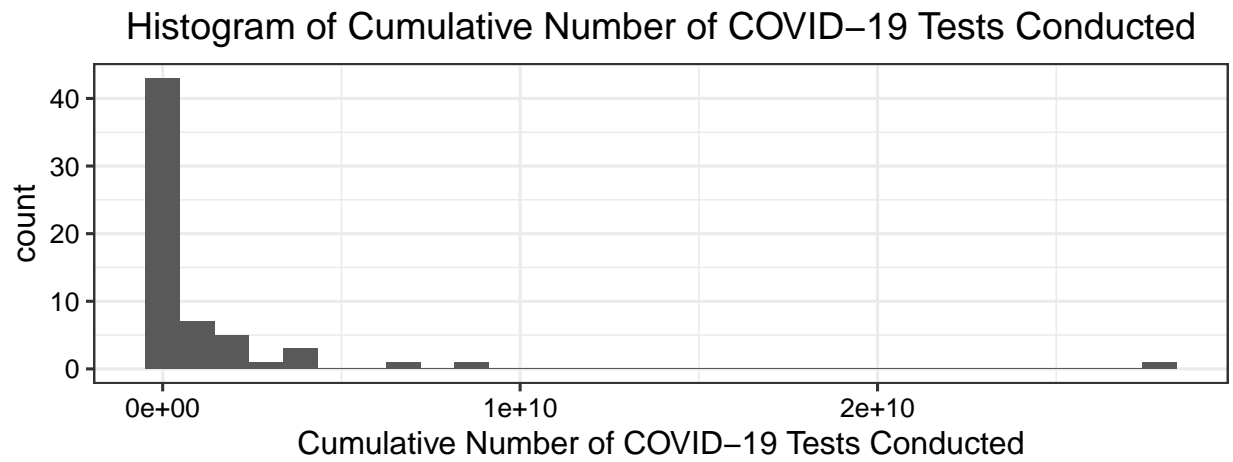


Figure 3: Cumulative Number of COVID-19 Tests Conducted

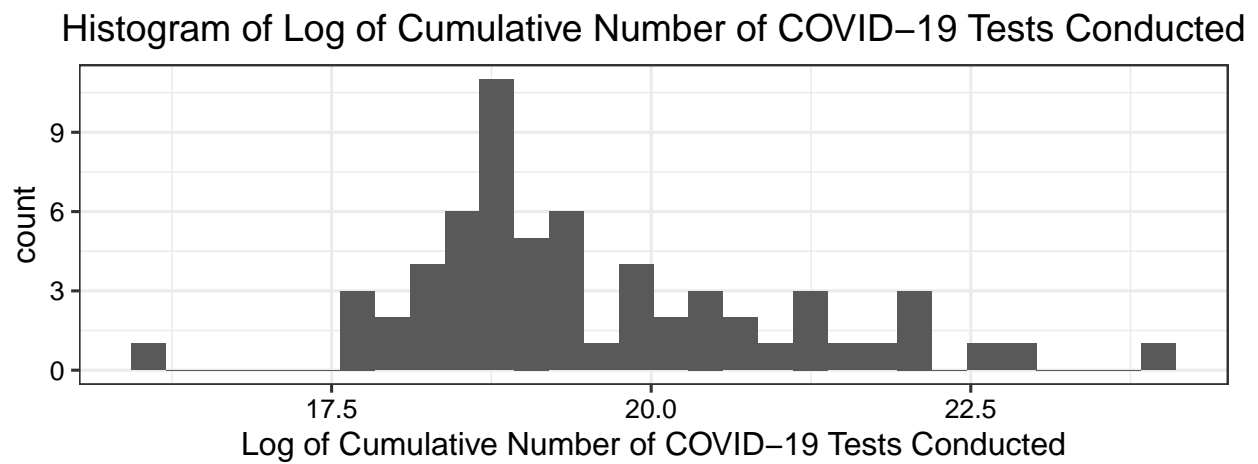


Figure 4: Log of Cumulative Number of Positive COVID-19 Tests Conducted

```
shapiro.test(log(combined_dat_ny_covid_pop$sum_cumulative_no_of_tests))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(combined_dat_ny_covid_pop$sum_cumulative_no_of_tests)
## W = 0.94091, p-value = 0.004985
```

```
#Check normality
```

```
##### County Population #####
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(pop2022)) +
  geom_histogram() +
  xlab("Absolute Population Size of Counties Across New York State 2022") +
  ggtitle ("Histogram of Population of Counties Across New York State")
```

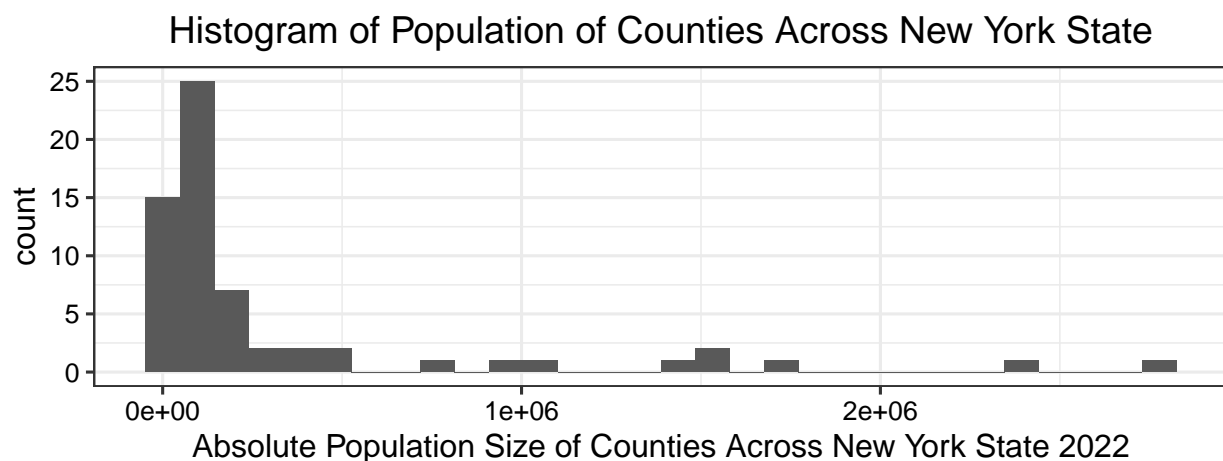


Figure 5: Population of Counties Across New York State

```
## Log Transform
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(log(pop2022))) +
  geom_histogram() +
  xlab("Log of Absolute Population of Counties Across New York State 2022 ") +
  ggtitle ("Log of Histogram of Population of Counties Across New York State")
```

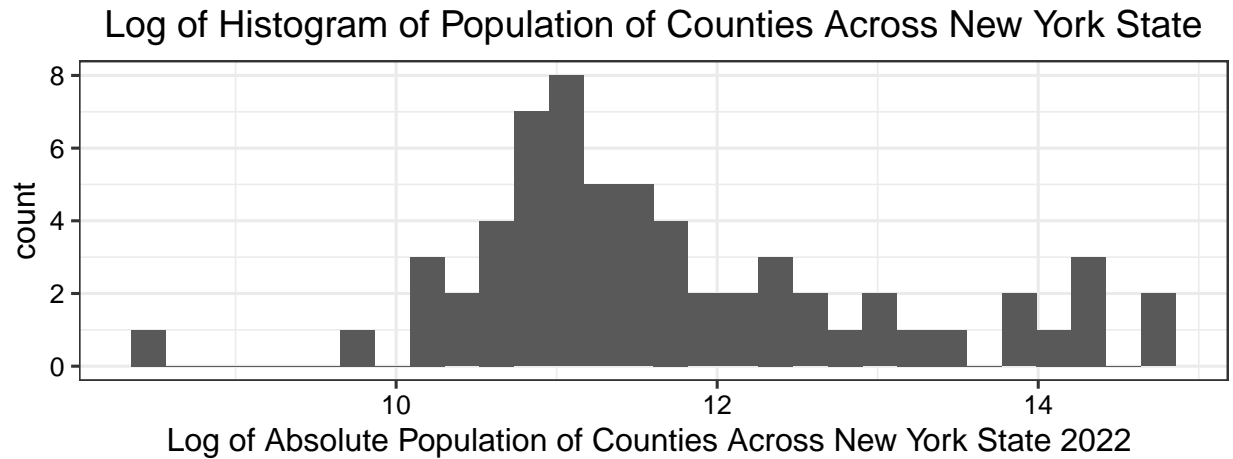


Figure 6: Log of Population of Counties Across New York State

```
shapiro.test(log(combined_dat_ny_covid_pop$pop2022)) #Check normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(combined_dat_ny_covid_pop$pop2022)
## W = 0.93542, p-value = 0.002814
```

```
##### County Density #####
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(density_permilessquared)) +
  geom_histogram() +
  xlab("Density of Counties Across New York State") +
  ggtitle (" Histogram of Density of Counties Across New York State")
```

```
## Log Transform
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(log(density_permilessquared))) +
  geom_histogram() +
  xlab("Log of Density of Counties Across New York State") +
  ggtitle (" Log of Histogram of Density of Counties Across New York State")
```

```
shapiro.test(log(combined_dat_ny_covid_pop$density_permilessquared))
```

```
##
```

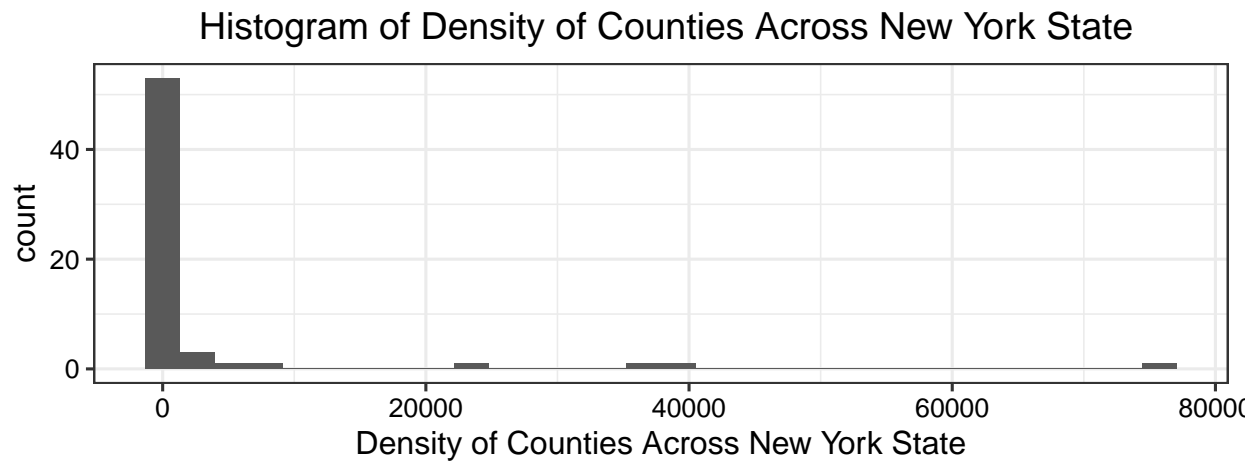


Figure 7: Density of Counties Across New York State

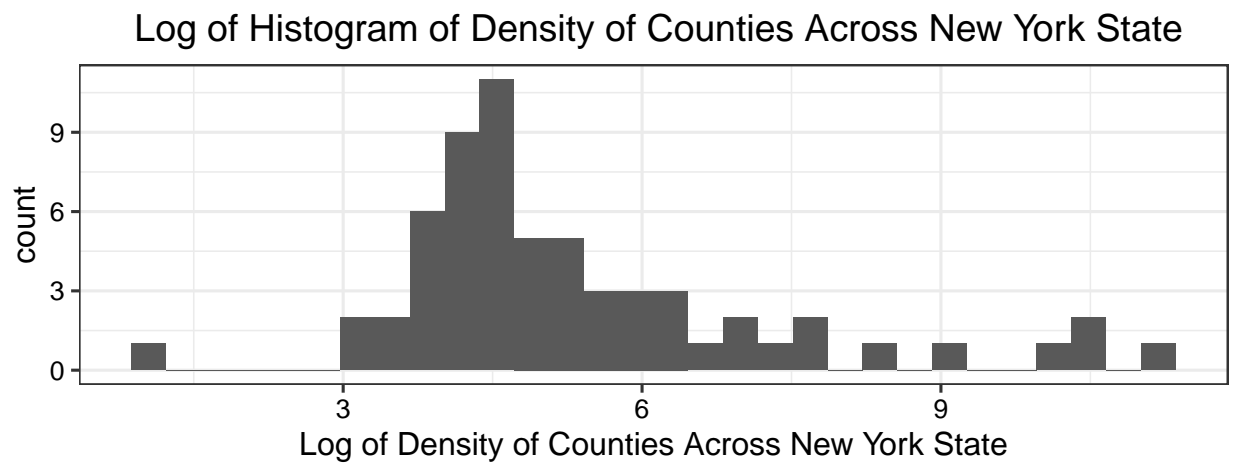


Figure 8: Log of Density of Counties Across New York State

```
## Shapiro-Wilk normality test
##
## data:  log(combined_dat_ny_covid_pop$density_permilessquared)
## W = 0.87808, p-value = 1.73e-05
```

```
#Check normality
```

```
##### Visualize Distribution of Predictor Variables by Settlement Type
```

```
ggplot(data = combined_dat_ny_covid_pop, aes(x= settlement_type,
                                              y= log(density_permilessquared),
                                              fill = settlement_type))+

geom_boxplot() +
ggtitle ("Log of Absolute Population Size of Counties by Settlement Type") +
xlab("Settlement Types") +
ylab("Log of Absolute Population Size 2022")
```

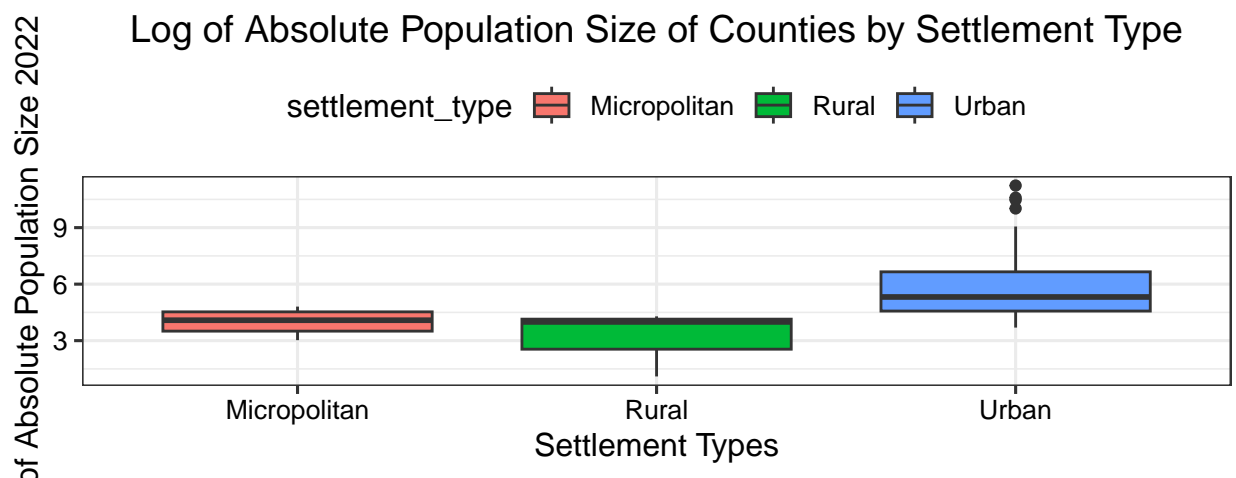


Figure 9: Log of Absolute Population Size of Counties by Settlement Type

```
ggplot(data = combined_dat_ny_covid_pop,
       aes(x= settlement_type, y= log(pop2022), fill = settlement_type))+
geom_boxplot() +
ggtitle ("Log of Population Density of Counties by Settlement Type") +
xlab("Settlement Types") +
ylab("Log of Absolute Population Size 2022")
```

0.3.3 Spatial Context of Data

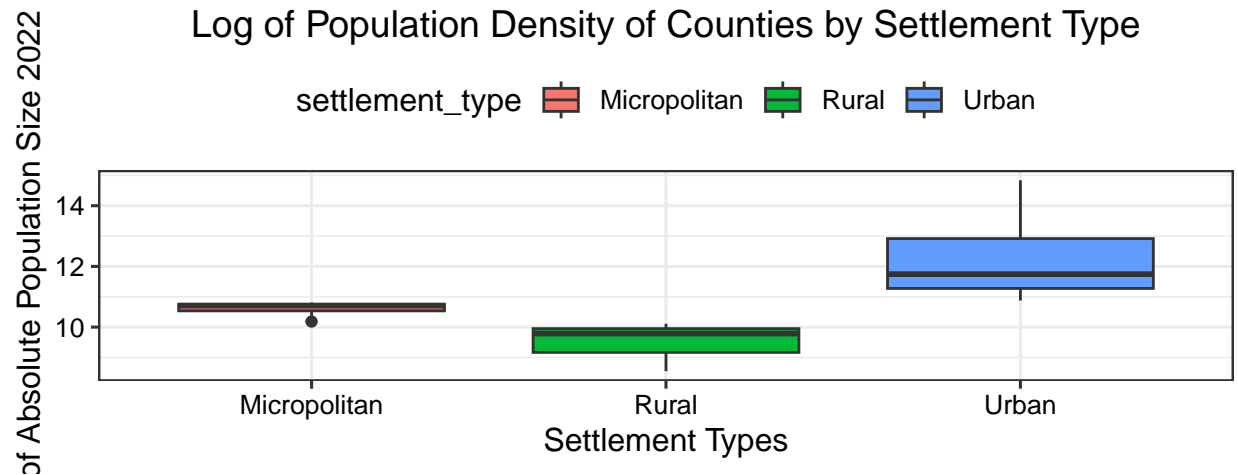


Figure 10: Log of Population Density of Counties by Settlement Type

```
# Map of New York County Population Levels
```

```
# The EPSG code of the New York counties dataset is 4269.
```

```
#This is a geographic coordinate reference system, so it uses
```

```
#angular coordinate units. This CRS (4269) is associated with the datum NAD83.
```

```
uscounty <- st_read('./Data/Spatial_Files/cb_2018_us_county_20m.shp',
                    stringsAsFactors = TRUE) %>%
  filter(STATEFP == 36)
```

```
## Reading layer 'cb_2018_us_county_20m' from data source
```

```
##   '/Users/danleizou/MiltonZouKhalsa_ENV872_EDA_FinalProject3/Data/Spatial_Files/cb_2018_us_county_20m.shp'
```

```
##   using driver 'ESRI Shapefile'
```

```
## Simple feature collection with 3220 features and 9 fields
```

```
## Geometry type: MULTIPOLYGON
```

```
## Dimension:      XY
```

```
## Bounding box:   xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
```

```
## Geodetic CRS:   NAD83
```

```
# Reveal the names of the columns
```

```
colnames(combined_dat_ny_covid_pop)
```

```
##   [1] "County"                "pop2022"
##   [3] "area_milessquared"     "density_permilessquared"
##   [5] "sum_new_positives"     "sum_cumulative_no_of_positive"
##   [7] "sum_total_number_of_tests" "sum_cumulative_no_of_tests"
##   [9] "median_test_positive"   "settlement_type"
```

```
# Join the flow data to our NWIS gage location spatial dataframe.
```

```
combined_dat_join <- merge(x = uscounty,  
                           y = combined_dat_ny_covid_pop,  
                           by.x = "NAME",  
                           by.y = "County")
```

```
# Show the column names of the joined dataset.
```

```
colnames(combined_dat_join)
```

```
## [1] "NAME" "STATEFP"  
## [3] "COUNTYFP" "COUNTYNS"  
## [5] "AFFGEOID" "GEOID"  
## [7] "LSAD" "ALAND"  
## [9] "AWATER" "pop2022"  
## [11] "area_milessquared" "density_permilelessquared"  
## [13] "sum_new_positives" "sum_cumulative_no_of_positive"  
## [15] "sum_total_number_of_tests" "sum_cumulative_no_of_tests"  
## [17] "median_test_positive" "settlement_type"  
## [19] "geometry"
```

```
# Show the dimensions of this joined dataset.
```

```
dim(combined_dat_join)
```

```
## [1] 61 19
```

```
# Map the population of each New York county in 2022.
```

```
ggplot() +  
  geom_sf(data = combined_dat_join, aes(fill = log(pop2022))) +  
  scale_fill_gradientn(colours = terrain.colors(10)) +  
  ggtitle("COVID-19 in New York",  
          subtitle = "Population by County in 2022") +  
  labs(fill = 'Log of Variable')
```

```
# Map the settlement types of each New York county in 2022.
```

```
ggplot() +  
  geom_sf(data = combined_dat_join, aes(fill = settlement_type)) +  
  ggtitle("COVID-19 in New York",  
          subtitle = "Settlement Types by County in 2022") +  
  labs(fill = 'Settlement Types')
```

COVID-19 in New York

Population by County in 2022

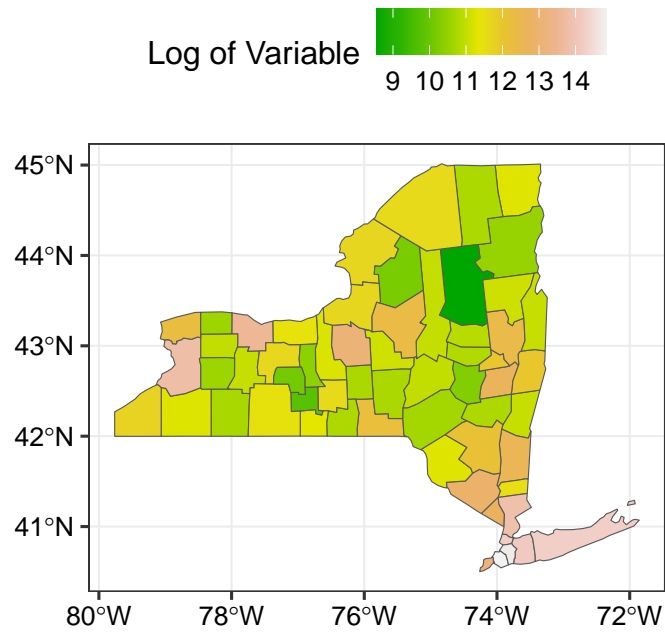


Figure 11: Population of New York Counties

COVID-19 in New York

Settlement Types by County in 2022

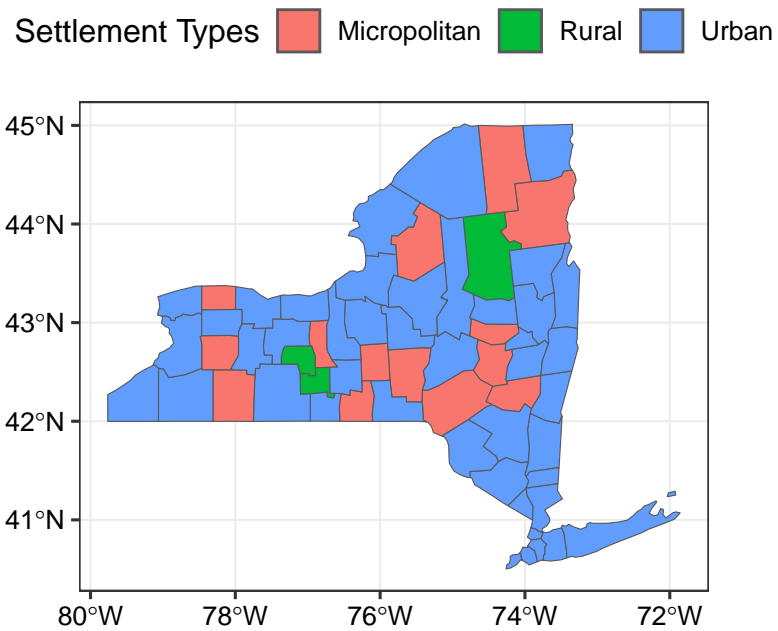


Figure 12: Settlement Types of New York Counties

```
# Map the sum of the cumulative number of tests administered in each  
#New York county in 2022.
```

```
ggplot() +  
  geom_sf(data = combined_dat_join,  
          aes(fill = log(sum_cumulative_no_of_tests))) +  
  scale_fill_gradientn(colours = terrain.colors(10)) +  
  ggtitle("COVID-19 in New York",  
          subtitle = "Cumulative Number of Tests Administered") +  
  labs(fill = 'Log of Variable')
```

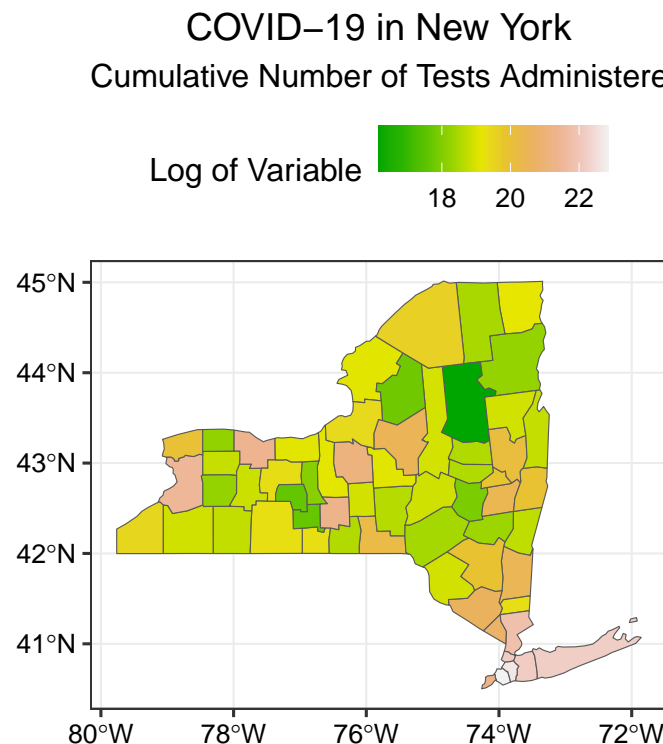


Figure 13: Cumulative Number of Tests Administered in Each New York County

```
# Map the sum of the cumulative number of positive test results in each  
#New York county in 2022.
```

```
ggplot() +  
  geom_sf(data = combined_dat_join,  
          aes(fill = log(sum_cumulative_no_of_positive))) +  
  scale_fill_gradientn(colours = terrain.colors(10)) +  
  ggtitle("COVID-19 in New York",  
          subtitle = "Cumulative Number of Positive Test Results in 2022") +  
  labs(fill = 'Log of Variable')
```

COVID-19 in New York

Cumulative Number of Positive Test Results in 2022

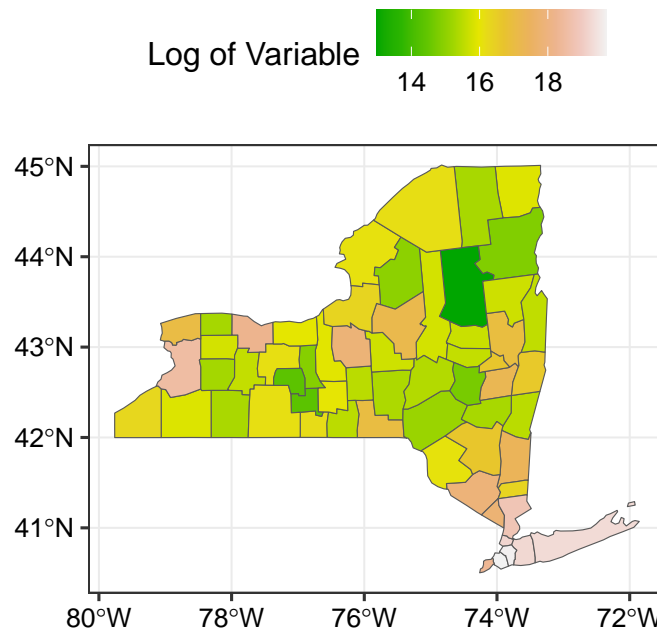


Figure 14: Cumulative Number of Positive Test Results in Each New York County

Export Processed data

```
getwd() #Check working directory
```

```
## [1] "/Users/danleizou/MiltonZouKhalsa__ENV872_EDA_FinalProject3"
```

```
#Full processed data set
```

```
write.csv2(combined_dat_ny_covid_pop,  
            "./Data/Processed_Data/Combined_NY_COVID_Pop.csv")
```

```
#Collapsed data set
```

```
write.csv2(collapsed_dat_settlement_types,  
            "./Data/Processed_Data/collapsed_dat_settlement_types.csv")
```

0.4 Analysis

As a central question of this study is “How well does population density and absolute population predict COVID-19 cases?”, the predictor variables are population density per miles squared and absolute population while the outcome variable is sum cumulative positive number of positive COVID-19 tests which is being used as a proxy for COVID-19 cases. Our unit of analysis is counties and settlement types depending on the primary objective of each sub-question. Settlement types are classified based on our definitions of rural, micropolitan, and urban settlement types. Our hypotheses for each study question is indicated below.

Question 1: What is the correlation between population density and absolute population and the number of COVID-19 cases (cumulative number of positive tests) across New York State?

Prior to the correlation analysis, the normality of absolute population, population density, and COVID-19 cases were examined and were determined to not be normally distributed (see exploratory analysis). The variables were log-transformed but were showed to still not be normally distributed after a Shapiro-Wilk test of normality was performed. As such, a Spearman correlation analysis for non-parametric data was carried out. The classification of the strength of the relationship was determined based on the value of r which ranges from 0 to 1 where $r = 0$ indicates no association and $r = -1$ or $+1$.

0.4.1 Absolute Population and COVID-19 Cases

Hypotheses

Null Hypothesis: As the ranks of absolute population change, the ranks of the number of COVID-19 cases do not change.

Alternative Hypothesis: As the ranks of absolute population change, the ranks of the number of COVID-19 cases change.

Assumptions:

1. The relationship between both variables is monotonic.

```
# Spearman rank correlation test as distributions are skewed
```

```
correlation_results_absol_pop <-  
  cor.test(x=log(combined_dat_ny_covid_pop$pop2022),  
           y=log(combined_dat_ny_covid_pop$sum_cumulative_no_of_positive),  
           method = "spearman")  
correlation_results_absol_pop
```

```
##
## Spearman's rank correlation rho
##
## data: log(combined_dat_ny_covid_pop$pop2022) and log(combined_dat_ny_covid_pop$sum_c
## S = 428, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.9892221
```

```
# Visualize plot (Log-transformed for better visualization)

ggplot(combined_dat_ny_covid_pop, aes(x = log(pop2022),
                                     y = log(sum_cumulative_no_of_positive)))+
  geom_point() +
  ylab ("Log of COVID-19 Cases") +
  xlab ("Log of Absolute Population Across Counties") +
  ggtitle ("The Correlation Between Absolute Population
           Across Counties and COVID-19 Cases") +
  stat_smooth (method = 'lm') +
  annotate("text",x=9,y=20, label=(paste0("R=0.99")))
```

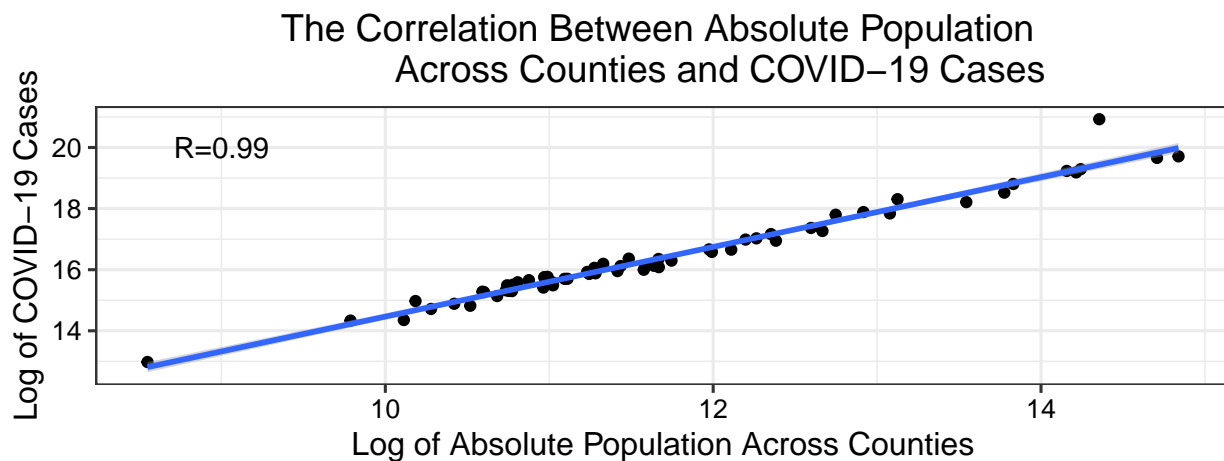


Figure 15: Correlation Between Absolute Population Across Counties and COVID-19 Cases

0.4.2 Population Density and COVID-19 Cases

Hypotheses:

Null Hypothesis: As the ranks of population density change, the ranks of the number of COVID-19 cases do not change.

Alternative Hypothesis: As the ranks of population density change, the ranks of the number of COVID-19 cases change.

Assumptions:

1. The relationship between both variables is monotonic.

```
# Spearman rank correlation test as distributions are skewed
```

```
correlation_results_pop_density <- cor.test(x=log(combined_dat_ny_covid_pop$density_per_mile_squared),
      y=log(combined_dat_ny_covid_pop$sum_cumulative_no_of_positive),
      method = "spearman")
correlation_results_pop_density
```

```
##
## Spearman's rank correlation rho
##
## data: log(combined_dat_ny_covid_pop$density_per_mile_squared) and log(combined_dat_ny_covid_pop$sum_cumulative_no_of_positive)
## S = 4522, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.8861273
```

```
# Visualize plot (Log-transformed for better visualization)
```

```
ggplot(combined_dat_ny_covid_pop, aes(x = log(density_per_mile_squared),
      y = log(sum_cumulative_no_of_positive)))+
  geom_point() +
  xlab (" Log of Population Density (per miles squared)") +
  ylab ("Log of COVID-19 Cases") +
  ggtitle ("The Correlation Between Population Densities
    Across Counties and COVID-19 Cases") +
  geom_smooth (method = 'lm')+
  annotate("text",x=3,y=20, label=(paste0("R=0.89")))
```

Question 2: What is the relationship between the predictor variable: settlement types and the response variable: cumulative COVID-19 cases?

0.4.2.1 Negative Binomial Regression

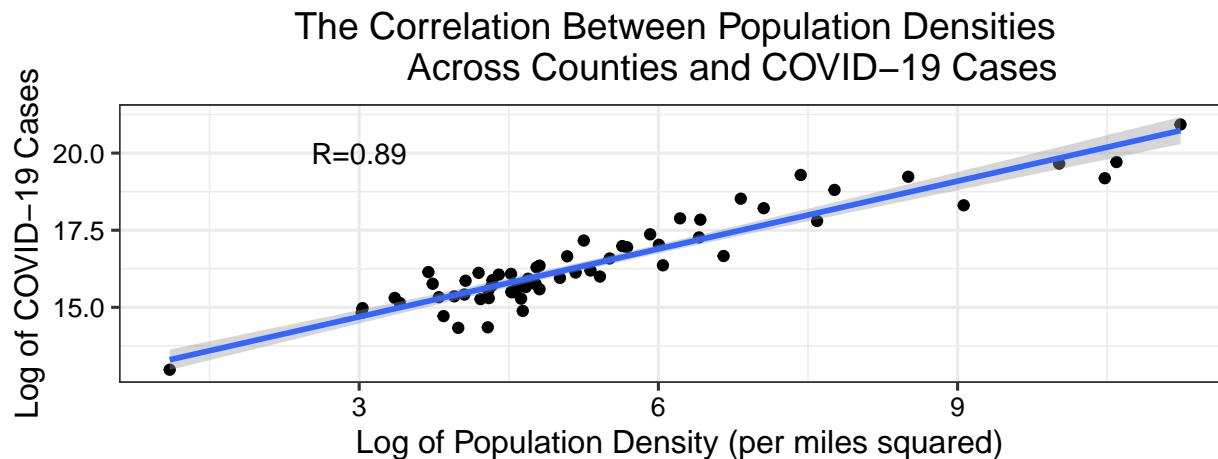


Figure 16: Correlation Between Population Densities Across Counties and COVID-19 Cases

In addition to the correlation analysis, we carried out a negative binomial regression to further understand the relationship between settlements types (determined by levels of absolute population) and cumulative COVID-19 cases. We used a negative binomial regression as based on the exploratory plots, we believed that the outcome variable can be classified as over-dispersed count data. We check our assumption that the outcome variable, “sum_cumulative_no_of_positive”, was overdispersed by carrying out a Poisson regression and found that the negative binomial regression was a better fit for the data. Further, we carried out a Chi-Square Goodness of Fit test for our negative binomial regression model with residual deviance of 75.355 on 59 degrees of freedom and the p-value obtained was 0.07. This means that the evidence is more supportive of the fact that the model is a good fit.

Hypotheses:

Null Hypothesis: There is no trend between the predictor variable (settlement types) and outcome variable (cumulative COVID-19 cases).

Alternative Hypothesis: There is a trend between the response (settlement types) and outcome variables (cumulative COVID-19 cases).

Assumptions:

1. The response variable (COVID cases) is a count variable.
2. Each observation of data is independent of each other.
3. The mean and variance are not equal.

Negative Binomial Regression

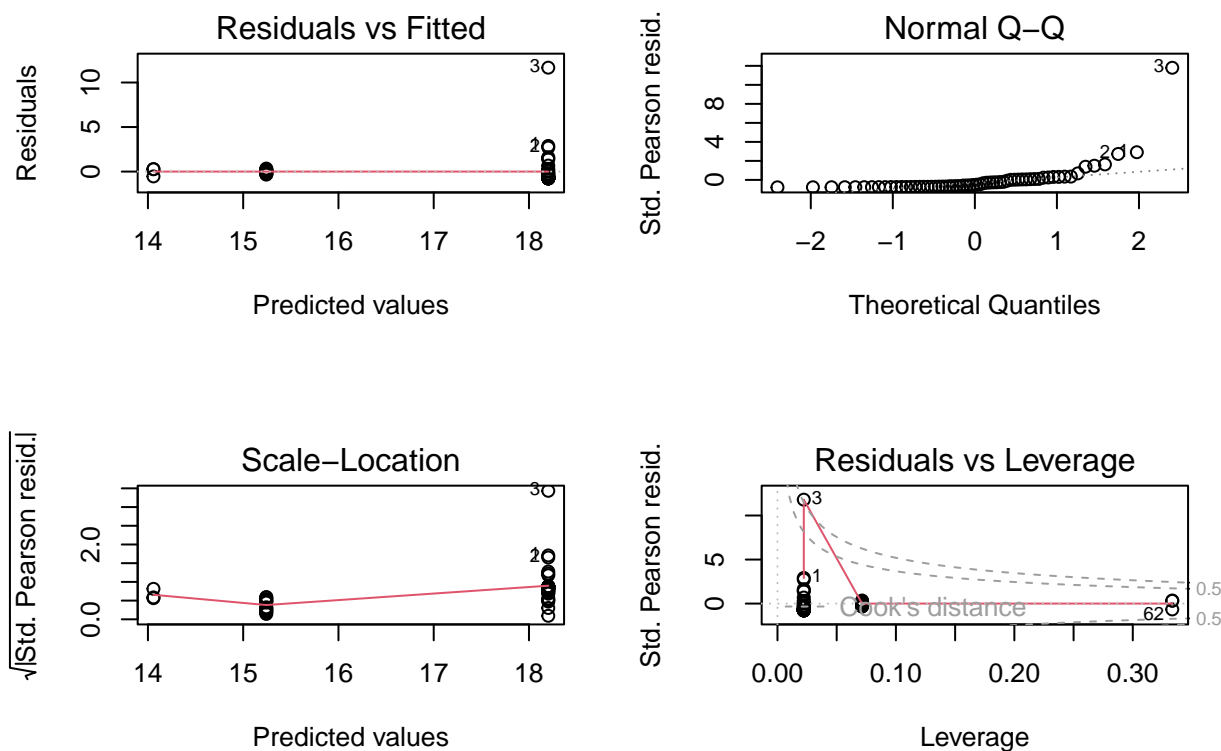
```

summary (modelnb <-
  glm.nb(sum_cumulative_no_of_positive ~ factor(settlement_type),
    data = combined_dat_ny_covid_pop))

##
## Call:
## glm.nb(formula = sum_cumulative_no_of_positive ~ factor(settlement_type),
## data = combined_dat_ny_covid_pop, init.theta = 0.6722916249,
## link = log)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5768  -1.2869  -0.6950   0.0667   3.9330
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      15.2425     0.3260  46.763 < 2e-16 ***
## factor(settlement_type)Rural  -1.1844     0.7759  -1.526   0.127
## factor(settlement_type)Urban   2.9602     0.3732   7.931 2.17e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6723) family taken to be 1)
##
##      Null deviance: 122.511  on 61  degrees of freedom
## Residual deviance:  75.366  on 59  degrees of freedom
## AIC: 2273.8
##
## Number of Fisher Scoring iterations: 3
##
##              Theta:  0.672
##            Std. Err.:  0.102
##
## 2 x log-likelihood:  -2265.783

par (mfrow = c(2,2))
plot(modelnb) #Check fit

```



```
# Test of Fitness
```

```
pchisq(q=modelnb$deviance, df=modelnb$df.residual,
lower.tail=FALSE) #Check model fit
```

```
## [1] 0.0740414
```

```
## Check our assumption that a negative binomial is a better model than the
#poisson model
```

```
summary (modelpoisson <-
  glm(sum_cumulative_no_of_positive ~ factor(settlement_type),
      family = "poisson", data = combined_dat_ny_covid_pop))
```

```
##
```

```
## Call:
```

```
## glm(formula = sum_cumulative_no_of_positive ~ factor(settlement_type),
##      family = "poisson", data = combined_dat_ny_covid_pop)
```

```
##
```

```
## Deviance Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -11105  -10010   -5977    168   66186
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      15.2424841  0.0001309  116408  <2e-16 ***
## factor(settlement_type)Rural -1.1844175  0.0005279   -2244  <2e-16 ***
## factor(settlement_type)Urban  2.9601778  0.0001320   22427  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 1.0813e+10  on 61  degrees of freedom
## Residual deviance: 8.9559e+09  on 59  degrees of freedom
## AIC: 8955869537
##
## Number of Fisher Scoring iterations: 6
```

Summary inference: Residual deviance is much larger than degrees of freedom (DF). Negative binomial plot better.

Question 3: What is the correlation between absolute population and population density and the number of COVID-19 tests done?

Similar to question 1, prior to the correlation analysis, the normality of absolute population, population density, and COVID-19 cases were examined and were determined to not be normally distributed (see exploratory analysis). The variables were log-transformed but were showed to still not be normally distributed after a Shapiro-Wilk test of normality was performed. As such, a Spearman correlation analysis for non-parametric data was carried out. The classification of the strength of the relationship was determined based on the value of r which ranges from 0 to 1 where $r = 0$ indicates no association and $r = -1$ or $+1$.

0.4.3 Absolute Population and COVID-19 Cases

Hypotheses

Null Hypothesis: As the ranks of absolute population change, the ranks of the number of COVID-19 tests conducted do not change.

Alternative Hypothesis: As the ranks of absolute population change, the ranks of the number of COVID-19 tests conducted change.

Assumptions

1. The relationship between both variables is monotonic.

```
# Spearman rank correlation test as distributions are skewed
```

```
correlation_results_absol_pop <-  
  cor.test(  
    x = log(combined_dat_ny_covid_pop$pop2022),  
    y = log(combined_dat_ny_covid_pop$sum_cumulative_no_of_tests),  
    method = "spearman")  
correlation_results_absol_pop
```

```
##  
## Spearman's rank correlation rho  
##  
## data: log(combined_dat_ny_covid_pop$pop2022) and log(combined_dat_ny_covid_pop$sum_c  
## S = 996, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho  
## 0.9749188
```

```
# Visualize plot (Log-transformed for better visualization)
```

```
ggplot(combined_dat_ny_covid_pop,  
  aes(x = log(pop2022),  
      y = log(sum_cumulative_no_of_tests)))+  
  geom_point() +  
  ylab("Log of COVID-19 Tests Conducted") +  
  xlab("Log of Absolute Population Across Counties") +  
  ggtitle("The Correlation Between Absolute Population  
          Across Counties and COVID-19 Tests Conducted") +  
  stat_smooth(method = 'lm') +  
  annotate("text", x=9, y=22, label=(paste0("R=0.97")))
```

0.4.4 Population Density and COVID-19 Cases

Hypotheses:

Null Hypothesis: As the ranks of population density change, the ranks of the number of COVID-19 tests conducted do not change.

Alternative Hypothesis: As the ranks of population density change, the ranks of the number of COVID-19 tests conducted change.

Assumptions:

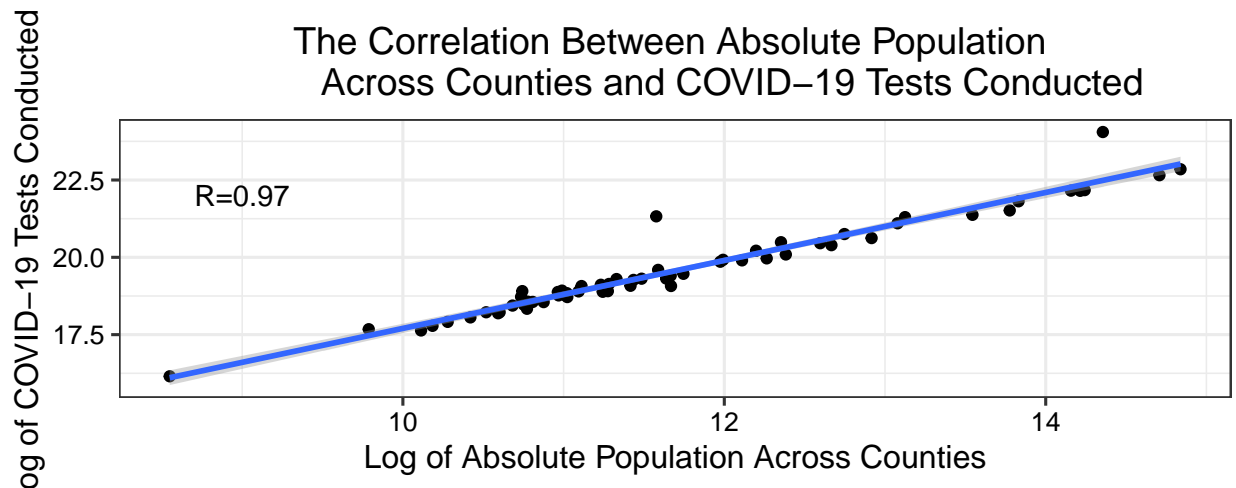


Figure 17: Correlation Between Absolute Population Across Counties and COVID-19 Tests Conducted

1. The relationship between both variables is monotonic.

```
# Spearman rank correlation test as distributions are skewed
```

```
correlation_results_pop_density <- cor.test(
  x= log(combined_dat_ny_covid_pop$density_per_milessquared),
  y = log(combined_dat_ny_covid_pop$sum_cumulative_no_of_tests),
  method = "spearman")
correlation_results_pop_density
```

```
##
## Spearman's rank correlation rho
##
## data: log(combined_dat_ny_covid_pop$density_per_milessquared) and log(combined_dat_ny_covid_pop$sum_cumulative_no_of_tests)
## S = 5602, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.8589308
```

```
# Visualize plot (Log-transformed for better visualization)
```

```
ggplot(combined_dat_ny_covid_pop, aes(x = log(density_per_milessquared), y = log(sum_cumulative_no_of_tests))) +
  geom_point() +
  xlab (" Log of Population Density (per miles squared)") +
```

```
ylab ("Log of COVID-19 Tests Conducted") +
ggtitle ("The Correlation Between Population Densities
        Across Counties and COVID-19 Tests Conducted") +
geom_smooth (method = 'lm')+
annotate("text",x=3,y=23, label=(paste0("R=0.86")))
```

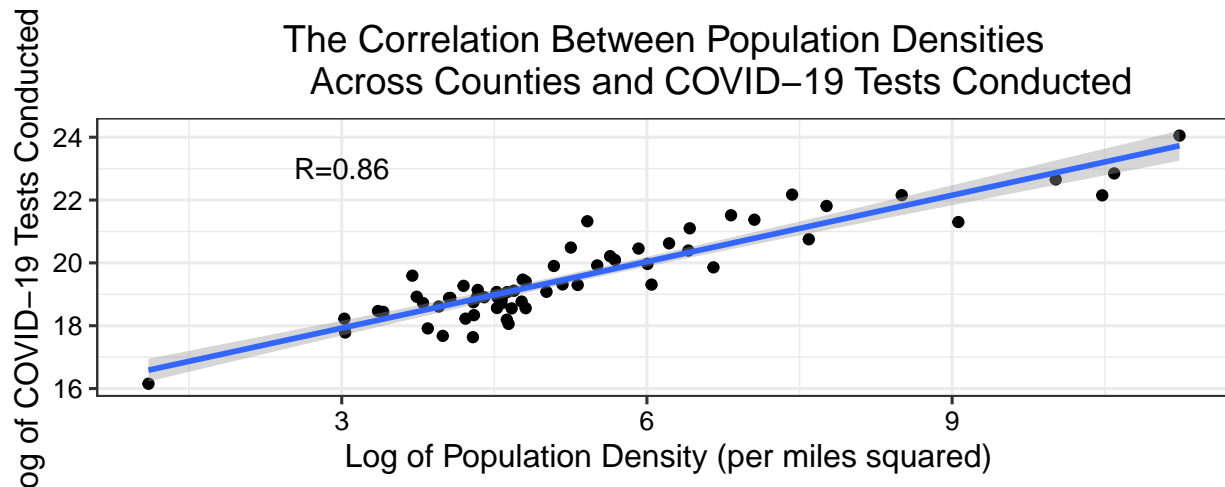


Figure 18: Correlation Between Absolute Population Densities Across Counties and COVID-19 Tests Conducted

Question 3b: Does the stronger correlated relationship vary among settlement types?

```
### Correlation Plots by Settlement Type

##Urban

Urban_dat <- combined_dat_ny_covid_pop %>%
  filter(settlement_type=="Urban")

correlation_results_absolute_pop_urban <-
  cor.test(x= log(Urban_dat$pop2022),
           y = log(Urban_dat$sum_cumulative_no_of_tests), method = "spearman")

correlation_results_absolute_pop_urban
```

0.4.4.1 Absolute Population and Settlement Types

```
##
## Spearman's rank correlation rho
##
## data: log(Urban_dat$pop2022) and log(Urban_dat$sum_cumulative_no_of_tests)
## S = 632, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.9583663
```

Micropolitan

```
Micropolitan_dat <- combined_dat_ny_covid_pop %>%
  filter(settlement_type=="Micropolitan")

correlation_results_absolute_pop_microp <-
  cor.test(x= log(Micropolitan_dat$pop2022),
           y = log(Micropolitan_dat$sum_cumulative_no_of_tests),
           method = "spearman")

correlation_results_absolute_pop_microp
```

```
##
## Spearman's rank correlation rho
##
## data: log(Micropolitan_dat$pop2022) and log(Micropolitan_dat$sum_cumulative_no_of_tests)
## S = 106, p-value = 0.002105
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.767033
```

Rural

```
Rural_dat <- combined_dat_ny_covid_pop %>%
  filter(settlement_type=="Rural")

correlation_results_absolute_pop_rural <-
  cor.test(x= log(Rural_dat$pop2022),
           y = log(Rural_dat$sum_cumulative_no_of_tests), method = "spearman")

correlation_results_absolute_pop_rural
```



```
##
## Spearman's rank correlation rho
##
## data: log(Rural_dat$pop2022) and log(Rural_dat$sum_cumulative_no_of_tests)
## S = 2, p-value = 1
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.5
```

```
ggplot(combined_dat_ny_covid_pop, aes(x = log(pop2022),
                                     y = log(sum_cumulative_no_of_positive)))+
  geom_point() +
  ylab ("Log of Cumulative Number of COVID-19 Cases") +
  xlab ("Absolute Population (2022)") +
  facet_wrap(vars(settlement_type))
```

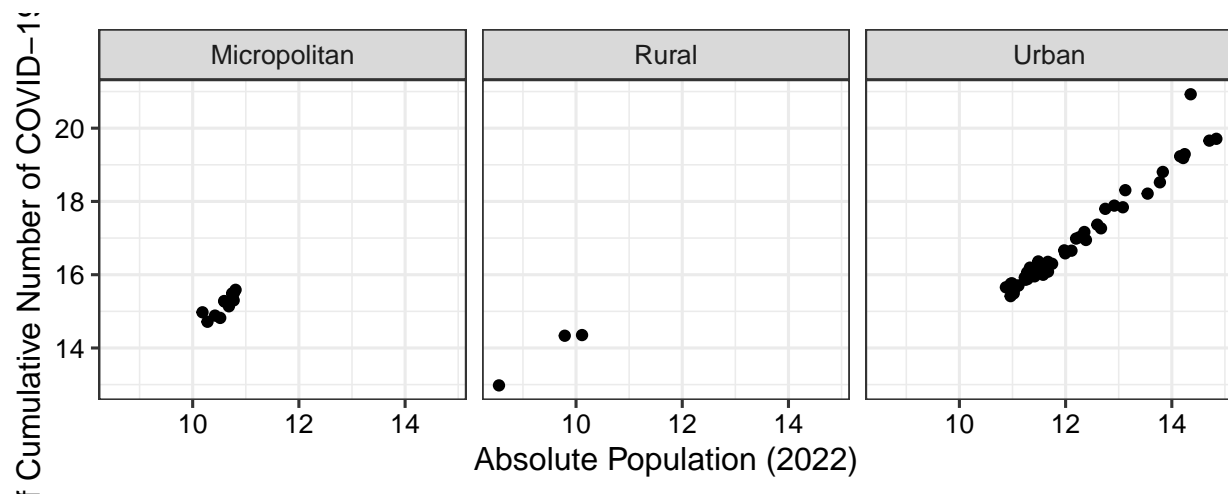


Figure 19: Log of Cumulative COVID-19 Cases Across Settlement Types

0.4.4.2 Visualization of 3b

0.5 Summary and Conclusions

0.5.0.1 General Overview of Data

The most populous county in New York State according to the information we obtained from the US Census Bureau is Kings County which has a population of over 2.7 million people. Meanwhile, the least populous county was Hamilton county with a little over 5000 residents. Despite Kings County have the most residents in the state, New York City was the most densely populated county - with 75,739 residents per squared miles. In addition to being the least populated county, Hamilton was the least dense of the 62 counties in New York State.

0.5.0.2 Correlation between Absolute Population and Population Density with COVID-19 Cases

We set out to answer our central question, “How well absolute population and population density predict COVID-19 case numbers in the diversely sized New York State?” by first examining the correlation between population density and absolute population and the number of COVID-19 cases across New York. In both cases of population density and absolute population, after performing a Spearman correlation for both variables we had a p-value of $2.2e-16$, which suggests that the correlated relationship was statistically significant. This means we have sufficient evidence to reject our null hypothesis that as the ranks of absolute population and population density change, the ranks of the number of COVID-19 cases do not change and instead accept our alternative hypothesis that as the ranks of absolute population and population density change, the ranks of the number of COVID-19 cases does change. It is also demonstrated in the visualized plots that there is indeed a positive correlation between the variables as the R-squared value for the population density plot was 0.89 and 0.99 for the absolute population plot. Both show a strong correlation, but it is stronger with absolute population.

0.5.0.3 Association between COVID-19 Cases and Settlement Types

In the second question, we examined the relationship between our predictor variable (settlement types) and response variable (cumulative COVID-19 cases). We did not control for potential confounding factors which limits the strength of the conclusions we are able to draw from this analysis. After performing a negative binomial regression, we observed that the variable level, rural, has a coefficient of -1.1844 which is associated with a p-value of 0.127. This means that when compared to the reference settlement level, micropolitan, the difference in the logs of the expected counts is expected to be 1.1844 lower for rural when compared to micropolitan. However, with a p-value of 0.127, the evidence is more supportive of the fact that this relationship is not significant. Meanwhile, we observed that when compared to the reference level, micropolitan, the difference in the logs of

expected counts is expected to be 7.931 higher for urban settlements. The p-value obtained for this association is less than 0.05 indicating that the evidence is more supportive of the fact that this association is statistically significant.

0.5.0.4 Correlation between Absolute Population and Population Density with Number of COVID-19 Tests Conducted

In our last question, we examined the correlation between absolute population and population density and the number of COVID-19 tests conducted in New York. Similar to the results of our first question, after performing a Spearman rank correlation test with both absolute population and population density we found that both had a p-value of $2.2e-16$, indicating that the correlated relationship was statistically significant. Therefore, we have sufficient evidence to reject both null hypotheses that as ranks of absolute population and population density change, the ranks of the number of COVID-19 tests conducted do not change. Instead, we accept our alternative hypotheses that as ranks of absolute population and population density change, the ranks of the number of COVID-19 tests conducted does change. This is also shown in our visualized plots, with the absolute population plot having an R-squared value of 0.97 and the population density plot having an R-squared value of 0.86. Both show a strong correlation, but it is stronger with absolute population.

After determining that the correlation between absolute population and number of COVID-19 tests conducted is stronger, we examined within those parameters to see if this correlation varies among settlement types. We performed the Spearman rank correlation test on all three settlement types. Urban settlements had a p-value of $2.2e-16$ and an R-squared value of 0.96, micropolitan settlements had a p-value of 0.002 and an R-squared value of 0.77, and rural settlements had a p-value of 1 and an R-squared value of 0.5. Two of the p-values show that the correlated relationships are significant, and rural didn't have a significant correlated relationship. The correlation is strongest with urban settlements. However, we do acknowledge that the rural settlements only had 3 data points, which is much fewer than the amount of data of micropolitan and urban settlements.

0.6 Conclusion

In conclusion, we do see a strong correlation between both absolute population and population density and the number of COVID-19 cases in New York, as well as the relationship between both absolute population and population density and the number of COVID-19 tests conducted in New York. The absolute population correlation to the number of COVID-19 tests conducted was the stronger relationship. It is important for us to know the correlation between COVID-19 cases and settlement type so we can be better informed about what factors influence the spread of infectious diseases for the future.

0.7 References

1. CDC. (2020, February 11). COVID-19 and Your Health. Centers for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/your-health/reinfection.html>
2. CDCMMWR. (2020). Geographic Differences in COVID-19 Cases, Deaths, and Incidence—United States, February 12–April 7, 2020. MMWR. Morbidity and Mortality Weekly Report, 69. <https://doi.org/10.15585/mmwr.mm6915e4>
3. Coronavirus. (n.d.). Retrieved March 14, 2022, from <https://www.who.int/westernpacific/health-topics/coronavirus>
4. Hamidi, S., Sabouri, S., & Ewing, R. (2020). Does Density Aggravate the COVID-19 Pandemic? *Journal of the American Planning Association*, 86(4), 495–509. <https://doi.org/10.1080/01944363.2020.1777891>
5. New York Population 2022 (Demographics, Maps, Graphs). (n.d.). Retrieved December 10, 2022, from <https://worldpopulationreview.com/states/new-york-population>
6. New York State Tracking Program | Tracking | NCEH | CDC. (2019, February 20). https://www.cdc.gov/nceh/tracking/profiles/New_York_State_Profile.htm
7. Settlements Overview & Types | What are Settlements? - Video & Lesson Transcript. (n.d.). Study.Com. Retrieved December 10, 2022, from <https://study.com/learn/lesson/settlements-overview-types.html>
8. Times, T. N. Y. (2020, April 1). New York Coronavirus Map and Case Count. *The New York Times*. <https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html>
9. Truelove, S., Smith, C. P., Qin, M., Mullany, L. C., Borchering, R. K., Lessler, J., Shea, K., Howerton, E., Contamin, L., Levander, J., Kerr, J., Hochheiser, H., Kinsey, M., Tallaksen, K., Wilson, S., Shin, L., Rainwater-Lovett, K., Lemaitre, J.
10. C., Dent, J., ... Viboud, C. (n.d.). Projected resurgence of COVID-19 in the United States in July—December 2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. *ELife*, 11, e73584. <https://doi.org/10.7554/eLife.73584>
11. USDA ERS - What is Rural? (n.d.). Retrieved November 27, 2022, from <https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/what-is-rural.aspx>
12. What Is Coronavirus? (2022, July 29). <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus>
13. Wong, D. W. S., & Li, Y. (2020). Spreading of COVID-19: Density matters. *PLOS ONE*, 15(12), e0242398. <https://doi.org/10.1371/journal.pone.0242398>