Giulianna Farina

Professor Eleish

Data Analytics

October 22, 2024

Assignment #3

Link to full code and all plots:

For question 1 I used data from Vermont specifically. I had to do this because my computer was not able to run the full dataset due to its large size.

**1A.**

|  |  |
| --- | --- |
| Summary statistics for 2021: | Summary statistics for 2023: |
| A close-up of numbers  Description automatically generated | A close-up of numbers  Description automatically generated |

Plots:

|  |  |
| --- | --- |
| A graph of cases in vermont  Description automatically generatedA graph of cases in vermont  Description automatically generated | A graph with a number of squares and a line  Description automatically generated with medium confidenceA graph with black and white lines  Description automatically generated |
|  |  |

When comparing and analyzing the summary statistics for 2021 and 2023, there were some similarities but also some significant differences. I find it interesting that the minimum number of cases in 2023 was larger than 2021, since precautions in 2021 were much stricter compared to 2023. Although there was a larger number of cases in 2021, there were more deaths in 2023. There is a significant jump in deaths from 2021 to 2023, which is reflected in all the summary statistics.

**1B.**

Plots:

|  |  |
| --- | --- |
| A graph of cases with numbers and lines  Description automatically generatedA graph of cases with numbers and lines  Description automatically generated | A graph of a number of cases  Description automatically generatedA graph of pink and blue bars  Description automatically generated |
|  |  |

When creating and plotting the histograms, applying the normal distribution was the best fit. This is because it made it easier to visualize the changes in numbers of cases and deaths. Finally, in most of the plots the normal distribution line reflected the variables positively, with the top of the curve aligning with the largest data points.

**1C.**

Plots:

|  |  |
| --- | --- |
| A graph with a line  Description automatically generatedA graph with a purple line  Description automatically generated | A graph with blue dots  Description automatically generatedA graph with purple dots  Description automatically generated |
|  |  |

|  |  |
| --- | --- |
| A graph of a graph with a blue line  Description automatically generatedA graph with a purple line  Description automatically generated | A graph with a blue line  Description automatically generatedA graph with a purple line  Description automatically generated |
|  |  |

In 2021, we see the steepness in lines shows there is a significant number of cases/deaths. This is related to the ECDF and QQ plots because of the number of cases/deaths in 2021 were comparable day by day, providing a more even distribution, unlike 2023. 2023 presented a variance of numbers, resulting in the ECDF and QQ plots to have breaks and jumps presented on the graph. The distribution lines on the QQ plots all began closely reflecting the data but loses consistency at the 0 mark on the x-axis.

**2.**

Summary statistics:

|  |  |
| --- | --- |
|  |  |

Plots:

|  |  |
| --- | --- |
| A graph of cases by county  Description automatically generated | A graph of death by county  Description automatically generated |

|  |  |
| --- | --- |
| A graph of cases with numbers and lines  Description automatically generated | A graph of a normal distribution  Description automatically generated |

|  |  |
| --- | --- |
| A graph of cases in a number of cases  Description automatically generated | A graph of death  Description automatically generated |
| A graph of cases and a number of cases  Description automatically generated | A graph of death and normal distribution  Description automatically generated |

Since I was only able to perform analysis one state for the first part of the assignment, for this portion I did a subset of data from one county. I chose to filter the distributions by selecting the county where I am from, Monmouth County, NJ in the year 2021. When analyzing these graphs, it seems that the data is more evenly distributed. The distribution lines, especially in the QQ plots greatly reflect the data with minimal skew when compared to the larger datasets.

**3A.**

Plots:

|  |  |
| --- | --- |
| A graph showing a plot of property  Description automatically generated | A graph showing a black dot  Description automatically generated with medium confidence |

The variable most significantly influencing the house price is the square footage of property. When producing a scatterplot comparing them, we see that the distribution line is not a great fit. Since the square footage has such a large influence, we see in the plots above large amounts of data condensed into one spot, a large significance effecting price and all other variables.

**3B.**

|  |  |
| --- | --- |
| A graph of a number of bathrooms  Description automatically generated | A graph showing a number of small dots  Description automatically generated |

There is great significance when changing the input variables. The first scatterplot is based off data of houses that are over $1 million and have at least 3 bathrooms. These parameters caused a significant change from previous scatterplots and is not as visually appealing due to the strict parameters chosen. It is interesting the residuals of the subset LM gave a similar result to 3A but had presented more outliers.