A 250-500-word paper summarizing the following: Statistical/Hypothetical Question

* Outcome of your EDA

The Exploratory Data Analysis (EDA) is the crucial aspect of doing preliminary analyses on COVI19 datasets in order to find patterns, identify anomalies, test hypotheses, and verify presumptions with the use of summary statistics and graphical representations. I think understanding the COVID 19 pandemic on weekly data starting from year 2019 till 2023 was an attempt to get as many insights as possible to form a better analysis. The datasets have few important variables like city population, case\_count, hosp\_count, death\_count, pcr\_test\_count and pcr\_test\_pos\_count all these variables can be filtered by date fields like week\_start and week\_end. By using the Probability Mass Functions (PMR) , Cumulative Distributed Functions (CDR) and plotting scatter graph on the above numeric variables, it is evident that during the start week of pandemic the Hospilalization, case counts, death counts and PCR test counts were low and during late 2021 all these variables provides a clear indication of rise in the cases and support the theory virus being very contagious in nature and spread was exponentially very high till it reaches peek during the starting month of year 2022 and then we see a fall in the reported cases. In this exercise we also observe the data is highly distributed and most of the time doesn’t indicates a strong relationship with other variables for example the PCR test count variable doesn’t show any patters with the rise and fall of the pandemic. Additionally, the

case\_ counts and Hospitalization\_count maintains some association during the initial weeks after following weeks the data is just uniformly distributed and indicates fall of the COVID 19 cases during the early weeks of 2023.

The dataset size is approx. 8000 rows, which is massive and good samples to analyze and make some conclusions by the analysts. The main objective here is to observe and analyze the information which includes scrutinizing the data and spotting patterns in data sets behaviors using various statistics, scenarios and plotting graphs. It also ascertains the fact whether or not a predictive model is a workable analytical tool to solve a critical business problems and I think in our case it shows the pattern to comprehend the predictions.

* What do you feel was missing during the analysis?

The dataset was enormous (it had over 8000 rows), and as I was analyzing it, I felt exhaustive because of how big it is. To obtain deeper insights into the data and draw a conclusion more quickly, I may have lowered the sample size to no more than 1500 rows and included a few more variables, such as "Patient Discharge Count," "Negative PCR Test Result," and "Positive PCR Test Result." Data scientists may review the data before drawing any conclusions thanks to good data quality. Additionally, it ensures that the analysis is sound and that the conclusions are precise and pertinent to the goals. In addition, I used the CDC portal to choose the metrics and fill in the King County COVID 19 dataset. I have also taken help with automation scripts to populate large datasets. Over a good and tested data would have helped plotting the graph and more representable.

* Were there any variables you felt could have helped in the analysis?

As I previously stated, a few other variables, such as "Patient Discharge Count," "Negative PCR Test Result," and "Positive PCR Test Result," may have aided in gaining a deeper understanding of the data and hastening the conclusion-making process. These extra metrics would have shown the true efficacy of the PCR testing as well as how many of the total case counts (COVID case recorded) actually turned out to be positive and negative. However, I was unable to input these variables owing to a lack of time and information.

* Were there any assumptions made you felt were incorrect?

As we all understand, COVID-19 has evolved over time, and scientific knowledge continues to be updated as new research and data become available. To create the datasets, I have referenced the CDC portal to collect the dataset for King County. I had impression initially that all the metrics I have used for this project has direct dependencies on each other, however after analyzing and plotting various types of graph, I understood the data is actually distributed and doesn’t show direct dependencies on other variables. Below are a few examples of assumptions that have changed while analyzing the dataset.

* There is no strict relationship between the variables like case count, hospitalization counts, death counts and PCR Test counts. Before data analysis I was assuming that there is a strict relationship between these variables.
* There is also not a regular pattern in the rise and fall of the cases between year 2020 till 2021, however we see definite fall in the cases during the early months of year 2023 which also indicates an end of the pandemic.
* The data for total\_hosp\_incidental and pcr\_test\_count doesn’t provide a clear indication of health agencies taking advantage to control the spread of COVID 19 virus.
* What challenges did you face, what did you not fully understand?

One of the main obstacles was getting ready the vast gathering of term end project data. Given the short time frame of just a few weeks and the requirement for massive datasets, it becomes necessary to use some automated scripts to populate such large datasets. A better study to understand the link between the collection of variables I have generated would have been possible with a better selection of variables and high-quality data. As an illustration, let's look at how improved PCR testing has allowed the health sector to curb the COVID 19 virus's spread. Such eventualities are not represented in the dataset I currently possess. At first, I considered using city population factors and spent a significant amount of time gathering data for each city, However, I don't have a solid scenario to use the population data in relation to be evaluating the COVID 19 dataset while I'm working on the project.