**MEMORANDUM**

From: Thomas Greve

To: Dr. Craig

Date: May 10, 2021

Subject: Analysis of COVID-19 Statistics for Local Counties

**INTRODUCTION AND OVERVIEW**

The global coronavirus pandemic has been going on for well over a year. With the vaccine getting widely distributed as cases slowly fall, it appears as though there is light at the end of the tunnel. A wide variety of government agencies collected a multitude of data points throughout the past year to analyze the spread of the virus. By evaluating this data, data researchers can determine what methods of containing the virus worked best and where improvements needed to be made. Personally, I care about this information because the virus has drastically affected my family and friends. In addition, we should all care about this information for the welfare of society. I want to find out what could have been done better to limit the spread, and what type of information we can find out from the data that has been collected.

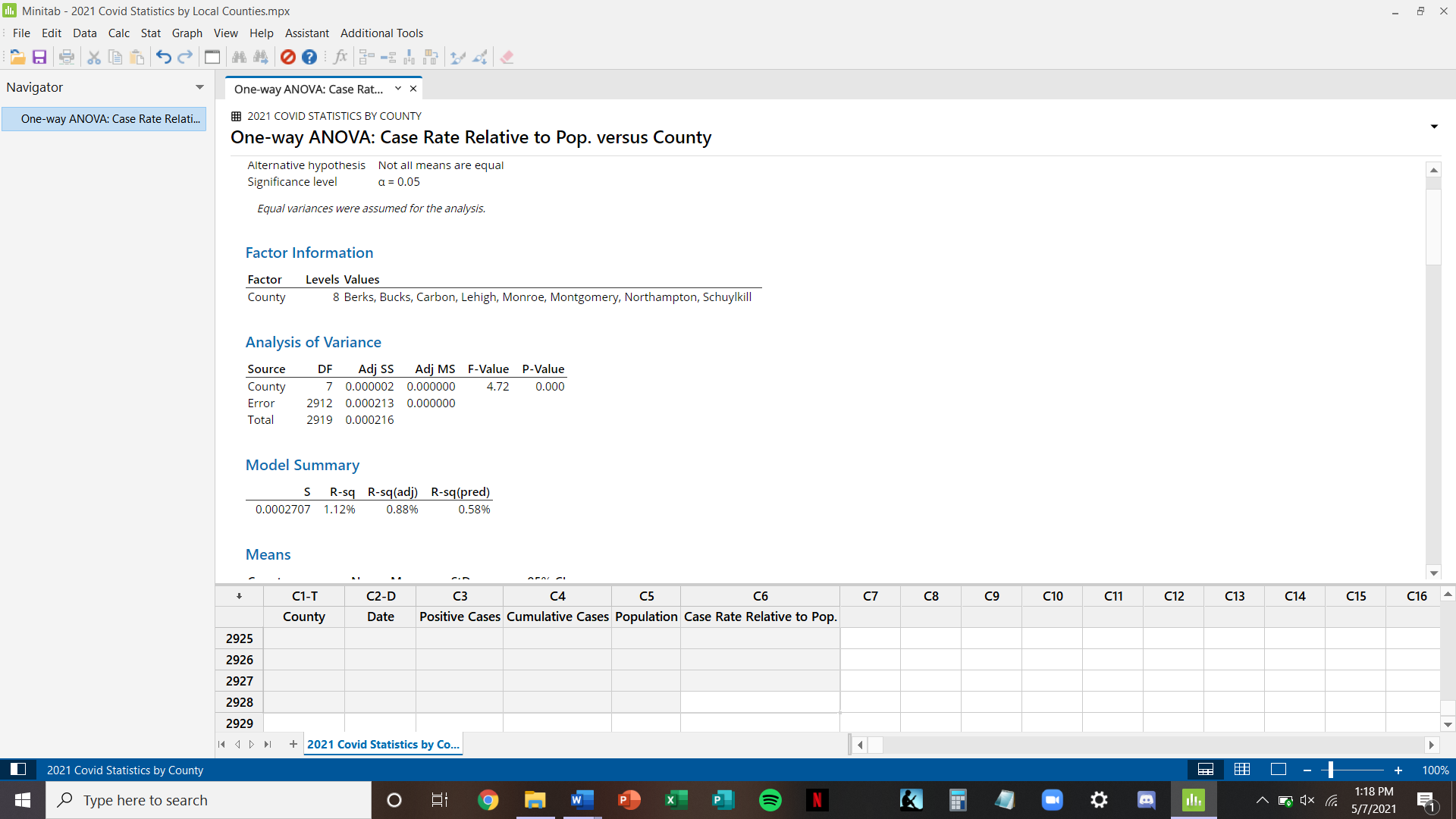
**Section 1: STATEMENT OF HYPOTHESIS**

One of the most widely reported statistics from the pandemic are daily statistics of new COVID-19 cases. I want to explore how the daily COVID statistics vary based on regional variable. My hypothesis is that cases will vary depending on what county is reporting them. The null hypothesis (H0) in this experiment is that these results are not different, and that these counties experience roughly the same number of daily cases. The alternative hypothesis (H1) is that these results are different, and different counties experience a significantly different number of cases daily. This can be proven at a 95% level of confidence by performing a hypothesis test.

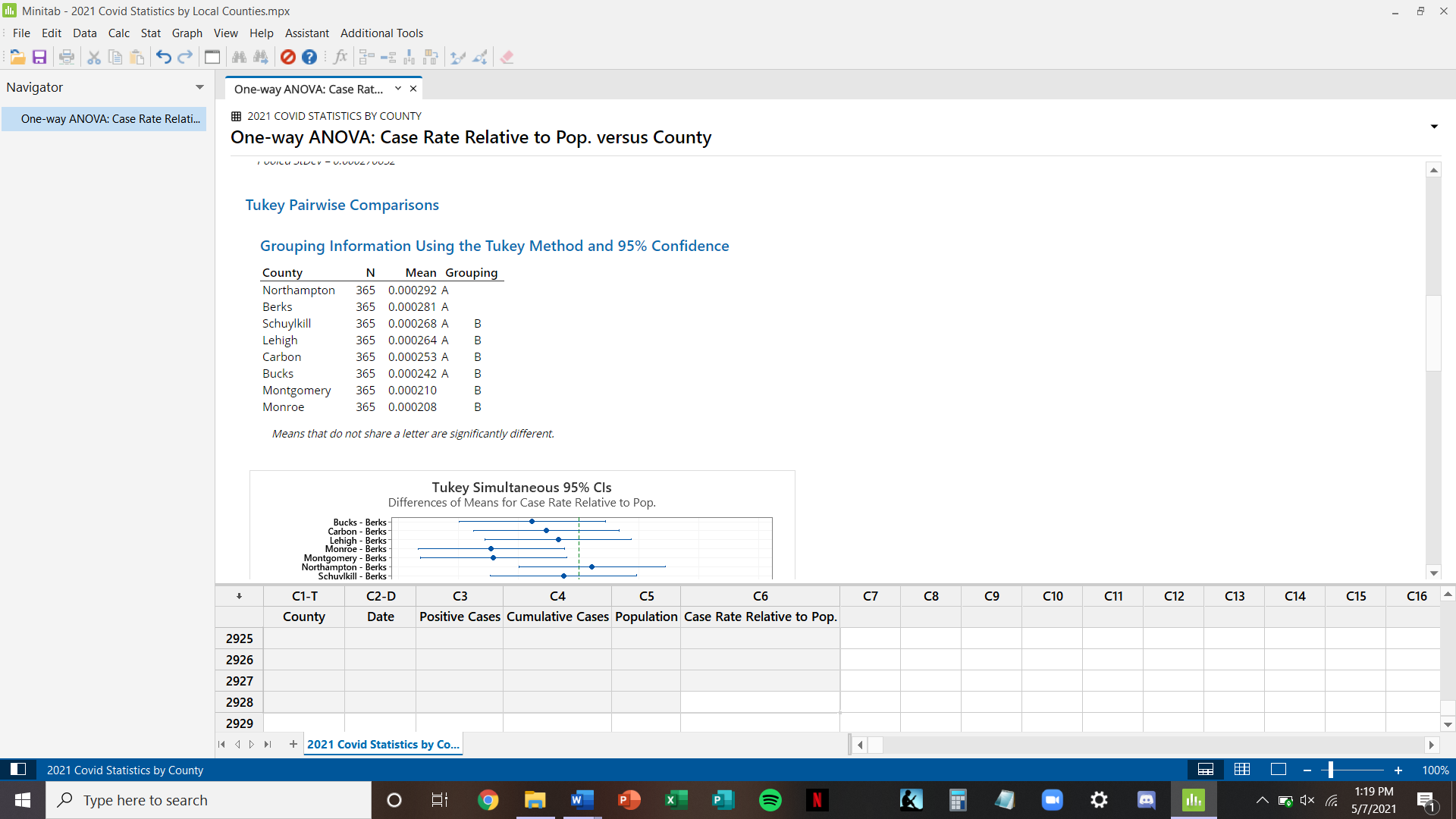
**Section 2: DATA AND ANALYSIS METHODOLOGY**

I gathered a dataset from the Department of Health from the Commonwealth of Pennsylvania (<https://www.health.pa.gov/topics/disease/coronavirus/Pages/Cases.aspx>). The database reports different statistics counted by every county in the state and is updated daily. The raw data from the database can be downloaded as a CSV file. For the sake of this project, I used Minitab to work with the data and run the tests. Some of the key variables from this dataset include county, date, number of cases reported, population, and percentage of population. For this project, I will be using data collected from the past year (May 6, 2020 to May 5, 2021). I reduced the number of counties being compared to the 7 immediate counties by Lehigh (Northampton, Berks, Bucks, Montgomery, Carbon, Monroe, and Schuylkill). The percentage of population variable is a standardized and calculated statistic that allows us to compare the figures without any bias. From this statistic, we can set up an analysis of variance test to determine if any of the COVID reports from local counties are statistically and significantly different from each other.

**Section 3: RESULTS**



The p-value from the ANOVA test is 0.000, which is well below the alpha level of 0.05. This means that we can reject the null hypothesis with 95% confidence. Essentially, there is sufficient evidence to suggest that the mean percentage of population that tested positive for COVID in these counties over the past year is statistically and significantly different from each other. For an ANOVA test, this means that even just one county can be different from the rest.



The Tukey comparison of this ANOVA test shows 2 different groupings for the mean percentage of population variable. The first group, A, includes counties with a range from .0242% to .0292%. The second group, B, contains points within a range from .0208% to .0268%. There is some overlap between these groupings, as 4 counties belong to both group A and group B. Still, these groups are significantly different enough from each other that the null hypothesis can be rejected. It can be concluded with a 95% level of confidence that the number of cases relative to the population are different across these 8 counties.

**DISCUSSION AND CONCLUSION**

For this project, I wanted to understand the spread of COVID in my area of residence more in depth. My analysis of the PA Department of Health dataset provided some very interesting insights. These data suggest that the percentage of the population that tested positive for COVID daily over the past year was different between all the counties that were tested. My home county (Montgomery) experienced the greatest mean daily number and overall number of COVID cases. However, it experienced one of the lowest percentages of positive cases relative to the entire population since last year. This was one of the most surprising results of this experiment. Generally, I am not surprised that these counties experienced significantly different rates of COVID cases because of a multitude of confounding variables that would potentially make the transmission of the virus greater in different areas. For instance, in areas with a higher population density the rates of COVID would be higher than those with lower population density. In addition, COVID transmission could have been affected by the rate of mask wearing in different counties. It is also necessary to acknowledge some of the limitations of this testing. Due to the time constraints of the data, I only used the daily statistics collected from the past year. This means that there is a lot of room for error regarding the statistics. If one county experienced high rates of spread over a two-week time span but generally did exceptionally for the rest of the year, then their data could have been skewed. For future analyses, it may be necessary to collect a sample taken over a longer period. It also may be necessary to account for these confounding variables when computing the average daily case rates. In conclusion, this dataset has showed us that there is significant variation between daily COVID cases reported by local counties.