Student Name: Jesseca Johnson

Course: DSC 501: Introduction to Data Science

Week 2 Course Project Question

Like many people around the world, I am curious about the origins of the COVID-19 pandemic – caused by the virus *severe acute respiratory syndrome coronavirus 2* (SARS CoV-2) – that has gripped the globe since early 2020, causing unprecedented numbers of death, illness, and changes in daily life. In March 2021, the World Health Organization (WHO) released a joint report it conducted with China about the origins of the virus. The report identified four possible pathways of emergence (WHO, 2021):

1. Direct zoonotic transmission (also termed: spillover)
2. Introduction through an intermediate host followed by zoonotic transmission (the report found this pathway to be the likeliest of the four to have actually occurred)
3. Introduction through the cold/food chain
4. Introduction through a laboratory incident

Although the idea of a possible lab leak is intriguing, the WHO found it to be the least likely of the four pathways of emergence of COVID-19. (It should be noted that that did not stop a group of 18 scientists from writing a letter doubting the methods used for the report and urging further investigation, which was published in the journal *Science* in May 2021, or Dr. Anthony Fauci, the COVID-19 czar of the US, publicly expressing doubts about the report in both his testimony to Congress and when answering a reporter's question at an event, also in May 2021 (Phillips, 2021).)

The first identified cluster of the virus appeared in Wuhan in December 2019 (Sun et al., 2020), but critical questions about the history of the virus remain:

- Where exactly did the virus originate?
- When exactly did the virus begin circulating in the community?

- Was Wuhan the first point of spread, or was spread occurring in another area either before the initial spread or simultaneously?

Because the WHO report found that a bat infected with a virus that was 96% similar to SARS CoV-2 in Yunnan Province, it has been theorized that Yunnan is the likeliest origin location of the virus (WHO, 2021). Disease ecologist Peter Daszak, who was part of a team that completed two weeks of investigation in the origins of the virus in China, has speculated that the virus was perhaps incubated in one of the wildlife farms in or near Yunnan Province, part of a successful 20-year project by the Chinese government that involved breeding wild animals in captivity (Doucleff, 2021).

In 2020, a group of medical researchers in Boston, USA released a study on Harvard's DASH website that made use of alternative data to examine when exactly COVID-19 began spreading in the community in Wuhan. First, they used satellite data to count the number of cars at select Wuhan hospitals between 2018 and 2020, finding that there was increased traffic beginning in August 2019 (Nsoesie et al., 2020). Second, they used search data from the most popular search engine in China, Baidu (Thomala, 2021), which makes public its data, and examined the number of instances of the search terms "cough" and "diarrhea" between 2017 and 2020. They found that although an increase in searches for the term "cough" could be explained by the correlation with flu season, there was a significant increase in searches for "diarrhea" starting in August 2019 that could not be explained by any other known events or factors (Nsoesie et al., 2020). A separate study released in July 2020 found that COVID-19 cases could be predicted using Google searches for symptoms four weeks ahead of time in most states in the US (Ahmad et al., 2020), and a similar study released in December 2020 found that Google searches for symptoms could detect COVID-19 cases up to 11 days in advance in Spain (Jimenez et al., 2020).

For my DSC 501 research project, my research question will be:

Was COVID-19 spreading in Kunming or Guangzhou prior to December 2019?

To answer my research question, I will conduct hypothesis testing. My null hypothesis will be:

There was no statistically significant increase in instances of COVID-19 symptom search terms on Baidu in Kunming/Guangzhou at any point in 2019.

My alternative hypothesis will be:

There was a statistically significant increase in instances of COVID-19 symptom search terms on Baidu in Kunming/Guangzhou at some point in 2019.

I would like to perform a follow-up to the Boston study examining the number of instances of searches for COVID-19 symptoms in two Chinese cities between 2018 and 2020:

1. Kunming, the capital of and largest city in Yunnan, due to the aforementioned hypothesis that Yunnan Province was the origin point of the virus;
2. Guangzhou, because as of 2021, it is the number one flight destination from Wuhan with 127 flights per week ("Wuhan Tianhe International Airport," 2021) and it would perhaps follow that there would also be an increase in searches for symptoms there with some lag time following the increases in Wuhan.

Both cities' data should be at least mostly representative of the population as a whole since, as of 2017, 77% of citizens in Yunnan had an Internet connection, and there was more than one Internet connection per person in Guangdong, of which Guangzhou is the capital ("How Web-connected is China?," 2019).

To be clear, the WHO report suggests that an infected animal was likely transported from Yunnan to the Huanan Seafood Market in Wuhan where it began to spread, rather than first spreading in Kunming/Yunnan and then to and in Wuhan (or both simultaneously), but because of the proximity and the WHO report's origin hypothesis, I find the number of COVID-19 symptom search terms in Kunming to be worthy of investigation.

I would like to look for any increases in the number of searches for the following COVID-19 symptoms (CDC, 2021), with each acting as a dependent variable:

1. "diarrhea"
2. "nausea"
3. "loss of taste"
4. "ageusia" (loss of taste)
5. "loss of smell"
6. "anosmia" (loss of smell)
7. "shortness of breath"
8. "breathlessness"

Thus, the independent variable would be the search engine data distribution. As in the Boston study, I will create visualizations using WebPlotDigitizer. I have intentionally excluded terms that are also symptoms of influenza-like illness (CDC, n. d.) in order to avoid any correlation with flu season that might produce misleading findings. The search terms will need to be translated into Mandarin (the most spoken language in Kunming; widely spoken in Guangzhou) and Cantonese (widely spoken in Guangzhou) (Tirosh, 2019).

The source of my data will be Baidu's Open Access Dataset, which is publicly available and the same source that the Boston team used for their research in 2020. I have not yet been able to examine the data, but I imagine the metadata will be able to be broken down at least by city, if not by neighborhood. Before conducting my actual research, I will get a feel for the data using exploratory data analysis as described by O'Neil and Schutt (2014) to see what the trends have been regarding those search terms over the last five or so years for the purpose of comparison. I will then perform a time series analysis on the above-listed search terms between 2018 and 2020 using regression diagnostics and a LOWESS line, perhaps along with other logistic regression techniques, as described in *Practical Statistics for Data Scientists* (2020).

Although multiple studies have concluded that instances of searches for COVID-19 symptoms can predict actual COVID-19 outbreaks (as discussed on page 2), the novelty of search engine data must be noted as a possible weakness, as it is not a time-tested approach. It is also worth noting that, with China's history as a secretive state ("Why Chinese Science Seems So Secretive," 2019), it is not outside the realm of possibility that their Baidu data may not be completely reliable. Regarding logistic regression, the advantage of LOWESS is that, as a non-parametric test, it assumes nothing about the data (Glen, 2013). One potential disadvantage is that it is sensitive to outliers and may thus be unrepresentative of the actual data (Bruce et al., 2020), but since we are using data on a massive scale, that should not be a significant problem.

In conclusion, I would like to contribute to the body of knowledge surrounding the origins of COVID-19. This investigation is ongoing and rapidly changing as new studies and information are released daily. It is critically important that we learn as much as possible about where SARS CoV-2 came from, how it spread, and when, so that we can prevent future outbreaks, learn more about early community transmission and how to monitor it, and how we can better manage future outbreaks in general. I feel that my study could be helpful in this pursuit.

Works Cited

Ahmad, I., Flanagan, R., and Staller, K. (2020). *Increased Internet Search Interest for GI Symptoms May Predict COVID-19 Cases in US Hotspots. Clinical Gastroenterology and Hepatology, 19*(2). https://doi.org/10.1016/j.cgh.2020.06.058

Bloom, J. D., Chan, Y. A., Baric, R. S., Bjorkman, P. J., Cobey, S., Beverman, B. E., Fisman, D. N., Gupta, R., Iwasaki, A., Lipsitch, M., Medzhitov, R., Neher, R. A., Neilsen, R., Patterson, N., Stearns, T., van Nimwegen, E., Worobey, M., and Relman, D.A. (2021). Investigate the origins of COVID-19. *Science*, *372*(6543), 694. https://doi.org/10.1126/science.abj0016

Bruce, A., Bruce, P., and Gedeck, P. (2020). *Practical Statistics for Data Science: 50+ Essential Concepts using R and Python* (2nd ed.). O'Reilly Media, Inc.

Doucleff, M. (2021, March 15). *WHO Points To Wildlife Farms In Southern China As Likely Source Of Pandemic*. NPR. Retrieved May 23, 2020 from https://www.npr.org/sections/goatsandsoda/2021/03/15/977527808/who-points-to-wildlife-farms-in-southwest-china-as-likely-source-of-pandemic

Flu Symptoms & Complications. (n.d.). Centers for Disease Control and Prevention (CDC). Retrieved May 23, 2021 from https://www.cdc.gov/flu/symptoms/symptoms.htm

Glen, S. (2013, October 6). *Lowess Smoothing in Statistics: What is it?* Statistics How To. Retrieved May 23, 2021 from https://www.statisticshowto.com/lowess-smoothing/#:~:text=LOWESS%20(Locally%20Weighted%20Scatterplot%20Smoothing,between%20variables%20and%20foresee%20trends

How Web-connected is China? (2019, April 18). ChinaPower. Retrieved May 22, 2021 from https://chinapower.csis.org/web-connectedness/

Jimenez, A. J., Estevez-Reboredo, R. M., Santed, M. A., and Ramos, V. (2020). COVID-19 Symptom-Related Google Searches and Local COVID-19 Incidence in Spain: Correlational Study. *Journal of Medical Internet Research 22(*12). https://doi.org/10.2196/23518

Nsoesie, E., Rader, B., Barnoon, Y., Goodwin, L., and Brownstein, J. (2020). Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019. *Harvard Medical School (HMS) Scholarly Articles*. https://dash.harvard.edu/bitstream/handle/1/42669767/Satellite_Images_Baidu_COVID19_manuscript_DASH.pdf?sequence=3&isAllowed=y

O'Neil, C. and Schutt, R. (2014). *Doing Data Science*. O'Reilly Media, Inc.

Phillips, M. (2021, May 23). *Fauci 'not convinced' COVID-19 developed naturally*. Yahoo! News. Retrieved May 23, 2021 from https://news.yahoo.com/fauci-apos-not-convinced-apos-120653229.html

Sun, J., He, W., Wang, L., Lai, A., Ji, X., Zhai, X., Li, G., Suchard, M. A., Tian, J. Zhou, J., Veit, M., and Su, S. (2020). COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends in Molecular Medicine*. https://doi.org/10.1016/j.molmed.2020.02.008

Symptoms of COVID-19. (2021, February 22). Centers for Disease Control and Prevention (CDC). Retrieved May 22, 2021 from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

Thomala, L. L. (2021, April 9). *Market share of search engines in China 2021, by pageview*. Statista. Retrieved May 23, 2020 from https://www.statista.com/statistics/253340/market-share-of-search-engines-in-china-pageviews/

Tirosh, O. (2019, September 28.) *What are the Commonly Used Chinese Dialects?* Retrieved May 23, 2021 from https://www.chineasy.com/what-are-the-commonly-used-chinese-dialects/

Why Chinese science seems so secretive – and how it may be about to change. (2019, January 24). The Conversation. Retrieved May 23, 2020 from https://theconversation.com/why-chinese-science-seems-so-secretive-and-how-it-may-be-about-to-change-110326

World Health Organization (WHO). (2021). *WHO-convened Global Study of Origins of SARS-CoV-2: China Part (Joint WHO-China Study)*. WHO. https://www.who.int/docs/default-source/coronaviruse/final-joint-report_origins-studies-6-april-201.pdf?sfvrsn=4f5e5196_1&download=true

Wuhan Tianhe International Airport. (2021, May 22). Flightradar24. Retrieved May 22, 2021 from https://www.flightradar24.com/data/airports/wuh