**DATA 512: Part 2: Extension plan**

1. **Motivation/problem statement**

For more than 3 years we have been experiencing a global pandemic as never before. It has taken millions of lives as it spread throughout the whole world. Given that humans have never experienced such an event before, we had no preparation nor understanding on how to manage such a crisis causing multiple building blocks of our society to fail, leaving millions of households without any income while governments struggled to meet the needs of the people.

Given that the pandemic started in 2020, technology and globalization was widely available, allowing for petabytes of data to be collected during these last 3 years. This data availability is a great resource that we now have at our hands to do scientific research, allowing us to learn the multiple ways in which a global pandemic can hit our society and what were the ultimate consequences around this. That data scientists are taking time to analyze and understand this data will have a huge impact in scientific knowledge, governments and experts need to understand what the consequences of their actions were, which policies worked, and which didn’t, they also need to understand which populations were most vulnerable and which they should have prioritized more attention to. Insights such as these will allow us to learn from our mistakes, grasp what are the top priorities in the event of a pandemic, and what institutions or policies we need to start building to prepare for a next one. Getting prepared is of utter importance if we want to avoid the tragedy, we lived in 2020, optimistically allowing us to save millions of lives and mitigating the pain of the crisis in millions of households that are most vulnerable to these events.

Through this analysis I hope to find insights such as the ones mentioned before, if these insights are useful and novel for Oklahoma County, I will plan to share this in an online article such that this valuable information is shared to the wider community. Sharing has multiple goals, it inspires other data scientists to work on these data sets, it can also drive them to work on top of my work to find additional insights or it can simply inform politicians or advocates about what happened, I would additionally share some suggestions around how this could be prevented in a future pandemic event.

It is important to note that finding insights or contributing in any way to global knowledge around pandemic data is a human-centered problem. We are working with sensible data that represents a tragic story for millions of people, this data for each individual could mean death, pain, suffering, trauma, psychological distress, depression, etc. The pandemic really broke into everyone’s lives in a heavily negative way and as such we need to treat this data with respect, being careful with any conclusions that we draw from it, recognizing any biases, caveats, and generalizations that we might find, always finding good explanations as to why this might be true. Nevertheless, as mentioned previously, it is of utmost importance to encourage thousands of data scientists to dedicate their time to work on these kinds of data problems, if a pandemic occurs soon and politicians or advocates take data-driven decisions with the insights that we might find in the data, we will be able to prevent a lot of harm, deaths, and destruction.

1. **Research questions and/or hypotheses**

As described previously, the pandemic affected and changed almost every aspect of society and human life, hence, there are a vast number of research questions that we could attempt to analyze throughout this project. Given that this is a free research project and that we really want to try to find important insights that might help society. I first searched in the worldwide web for the different data sources that Oklahoma County had to offer and found many areas that could potentially be used for analysis such as diversity, economy, civic, education, housing & living, health, etc. Given that I wanted to focus on areas that had greater impact in our society due to the pandemic, I decided to narrow my search to only economical and health data. After quickly checking and analyzing some of the economic data, I grew interest in it as it is an area that is indirect to COVID but at the same time I’m very positive it got truly hit by it. I would like to know in depth how hard did the pandemic affected the economy inside Oklahoma county, how much relationship there is and if there were any actions taken by the county to counter any changes. To do this, I found two interesting indicators of economy, unemployment insurance claims and employment by industry sector that will be the basis of my analysis to answer the following research questions and hypothesis:

**Research question:** What effect on employment did the pandemic generate inside Oklahoma county? Is there any trend or correlation between employment statistics and confirmed COVID cases in Oklahoma county?

**Hypothesis:** I expect to see multiple effects between employment and the pandemic, specifically I consider:

1. The highest peak of unemployment occurred right at the beginning of the pandemic; this initial peak can be up to 30x the average rate of unemployment for the previous year before the pandemic.
2. The industry sectors that got the heaviest unemployment rates were leisure & hospitality as well as other heavy blue collar job industries.
3. As government announces coming back to reality and pandemic situation is under “control”, the growth rate of employment in each sector mimics pre-pandemic growth.
4. There is a relationship trend between confirmed COVID cases and unemployment, a 1% increase in COVID confirmed cases resulted in 0.3% increase in unemployment.
5. **Data to be used**

I plan to use data from the [datausa.io](https://datausa.io/) web page, they offer extensive US public data that is conveniently at county granularity. Specifically for Oklahoma County, we will use two additional timeseries data:

* **Employment by Industry Sector:** Time series containing monthly employees per sector for Oklahoma county (not-seasonally adjusted) between December 2017 and April 2021.
* **Unemployment Insurance Claims:** Time series containing weekly unemployment insurance claims in Oklahoma county (not-seasonally adjusted) between January 2018 and September 2021.

I believe that these two time-series combined with the data we already have available from part 1 will be enough to answer the research questions previously presented. The source to obtain this dataset is through the following API: <https://datausa.io/about/api/>, specifically the following two links:

* **Employment by Industry Sector:** https://api-ts-vibranium.datausa.io/tesseract/data.jsonrecords?cube=BLS Employment - Supersector Only&drilldowns=Month of Year,Supersector&measures=NSA Employees&State=04000US40
* **Unemployment Insurance Claims:** <https://datausa.io/api/covid19/employment/latest/>

In the following link: <https://datausa.io/about/usage/>, datausa.io specifies all licence/terms of use required to use this data, they allow access to this data for educational purposes, it is presented under a [GNU Affero General Public Licence v3.0 (GPLv3)](https://www.gnu.org/licenses/agpl-3.0.txt). The content can be copied, downloaded for own use provided that suitable acknowledgment of Data USA as source is given.

These two time-series data sources summarize unemployment, given that it is at a county granularity, there is no personal information from any of the citizens. Given this, it is not sensible data that we would require to handle with greater care. We still need to be conscious that this data represents the suffering and pain of thousands of citizens that passed through harsh moments during the pandemic, as such, this data must be treated fairly and conclusions must be thoroughly inspected and proved as these could be eventually used as data-driven decisions in a next pandemic, affecting thousands to millions of lives.

1. **Unknowns and dependencies**

There are clearly many assumptions that we will be required to do to be able to answer these research questions in the time given. With regards to the pandemic COVID data, we will only work with COVID confirmed cases, this is a limited view of the whole pandemic, we know tests are faulty, have an evident lag, they are biased and seasonal (holidays peak # of tests done), among other. Additionally, we will not be modelling deaths, recoveries nor vaccination rates and their effect.

Regarding answering the research questions, we know that our data sources might not be the true numbers for unemployment but just estimates, there can be many faults such as lags in reporting, not every household report unemployment insurance claims, there can be informal jobs that we are not considering as an industry sector even though this is a very important sector. It might be tough to determine which sectors are more blue-collared than others. Furthermore, we don’t know if behind the pandemic, there was a depression in the economy, confounding our conclusions as we could be concluding that everything is related to the pandemic when it can be other factors as well. Lastly, given that I do not live in this county, I have no true context of how the pandemic was like there, I don’t know if there was special news, special interventions or any special policies enacted during the time of the pandemic that might affect the way these statistics are registered. We will have to make big assumptions in each of these parts, assuming that there are no other confounding factors that might affect these results, and that everything we find in the data is directly related to the pandemic, this simplifying assumption will allow us to draw conclusions in an easier manner, without having to dive much deeper into literature, other data sources, news, etc.

1. **Methodology**

I plan to first clean the original dataset from part 1. This includes transforming the cumulative data into daily confirmed COVID cases. Additionally, given that we know that this data has a weekly seasonality, I plan to use moving averages with a 7-day window to reduce this seasonality. On the other hand, I will also check for the new datasets if they have any seasonality or unit problems that I need to solve, for the moment, I note that the Employment by Industry Sector data is in terms of monthly employees, which makes industry non-comparable, I will transform this to units such as monthly growth (year-over-year) to be able to compare between these different sectors.

After cleaning and processing the data, I will attempt to answer the research questions.

1. To analyze unemployment peaks, I will use a timeseries plot, this will allow me to compare different peaks and conclude if the highest occurred right at the start of the pandemic or if there was a lag, the lag will be found by plotting this curve alongside the confirmed cases timeseries, I will use a lag to see in which moment the two curves meet, this will allow me to know how much time it took unemployment petitions to peak given COVID. I will additionally take the average of the previous year unemployment, I will compare this figure with the highest peak, to understand how many times did unemployment petitions rose, I will also take an average over the pandemic and compare these two using a t-test, the idea is to understand if the difference in means is statistically significant and that we can conclude that the pandemic increased unemployment.
2. To understand which were the industry sectors that were most heavily hit, I will first visualize the difference using timeseries plots. I will also do a couple of additional statistical tests, I will first do an ANOVA test to check if at least one of the industry sectors has a different growth (was hit more than the rest), then to go deeper I will be performing multiple mean t-tests, I will compare the different means of monthly growth across the different sectors during the pandemic, this will allow me to find which were most heavily hit.
3. To check if employment growth rate is equal before and after the worst part of the pandemic, I will first have to determine what is a date that corresponds to a “back to normal” situation, then I will perform for each industry sector two ordinary least squares regressions, one for the 3 months previous to pandemic and one for the 3 months after the pandemic is “normalized”, here the dependent variable will be the number of monthly employees and the independent variable will be time in days (3 months), I will then compare these two slopes using an analysis of covariance (ancova), statistically concluding if there is a change in the slope (employment growth) or not.
4. To find a relationship trend between COVID confirmed cases and unemployment, I will perform a ordinarily least squares regression with unemployment insurance claims as the dependent variable and confirmed COVID cases as the independent variable. I will first check using an F-test or t-test if the relationship is significant, if it is, then I will check what is the value, this will tell us how these two variables are related.
5. **Timeline to completion**

**Monday Nov 14th:** Have a python notebook that obtains (API call and local file readers) pre-processes/cleans all the data sources, having them ready for analysis.

**Monday Nov. 21st:** Have a separate python notebook for analysis that already contains answers to the first two hypotheses.

**Monday Nov. 28th:** Have a finalized analysis python notebook, have answers for the research questions and hypotheses.

**Friday Dec 2nd:** Have a first version of the PowerPoint presentation with all research questions answered, graphs and visualizations.

**Monday Dec. 5th: Course Project, Part 3 - Presentation due**. Have the presentation finalized and submitted to canvas.

Wednesday Dec. 7th: Practice for presentation

**Thursday Dec. 8th: Presentation day.** Present

**Friday Dec 9th:** Make sure repository is clean, organized, results are clear, and process is thoroughly documented such that replication of the whole process is possible.

**Monday Dec. 12th: Course Project, Part 4 – Project Repository due**: Submit repository link.