ECONOMIC IMPACT OF COVID-19 PREVENTION POLICIES: A COMPARATIVE ANALYSIS OF CHINA AND THE UNITED STATES

By

Xiaohui Ma

A capstone project submitted for Graduation with University Honors

May 8, 2025

University Honors

University of California, Riverside

APPROVED

Dr. Yingzhuo Fu

Department of Statistics

Dr. Begoña Echeverria, Howard H Hays, Jr. Chair

University Honors

ABSTRACT

The COVID-19 pandemic (2019-2023) triggered an unprecedented global health and economic crisis. Governments worldwide implemented various levels of quarantine and lockdown measures to prevent the virus's spread. To understand how different policies might influence their economy, this study makes a comparative analysis of China and the United States. By using regression analysis and spatial autocorrelation diagnostics, we assess the relationship between the strictness of quarantine policy and two economic indicators (unemployment rate and housing price index). The findings show marked differences between the two nations. For the U.S., we found that dummy variable states, workplace closing, international control, and eight other quarantine measures were significant predictors of the unemployment rate of the U.S.. However, only two policy variables were significant in China's case, potentially reflecting the impact of centralized enforcement in mitigating the economic disruptions. These findings can provide policymakers with a better reference for adjusting their policy and facing this kind of infectious disease again in the future.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my advisor, Professor Yingzhuo Fu, who gave me feedback throughout the research process and generously dedicated time during the summer break to guide me through statistical concepts. Your patience and insight have been instrumental to this project. I am also deeply thankful to my boyfriend Ding, who patiently taught me programming and stood by me through countless challenges. My sincere appreciation goes to my parents, whose unwavering support made it possible for me to study abroad. Of course, I want to thank myself - for never giving up. I would also like to thank the Honors Society for offering me guidance and a platform to grow, and for welcoming me into a community of outstanding students and faculty members. Finally, I extend my sincere appreciation to the countless volunteers around the world who helped collect and maintain COVID-19 data during the pandemic. Your efforts have laid a critical foundation for global research and have contributed meaningfully to scientific progress.

TABLE OF CONTENTS

INTRODUCTION

On December 12, 2019, the first case of a new coronavirus, known as COVID-19, was discovered in Wuhan, China. Due to globalization, the virus quickly spread around the world and became a serious public health issue. It took almost four years, from December 12, 2019, to May 11, 2023, to fully control the spread of the virus. During this period, different countries implemented varying quarantine policies to control the spread of the virus. Two distinctive examples include China and the United States. For China, the government adopted a highly centralized and stringent policy, including a national lockdown and travel restrictions.
In contrast, the United States took a more decentralized approach, with individual states having significant autonomy in determining the time, intensity, and type of quarantine measures. However, the same thing between the two countries is that both economies were affected by the pandemic. According to the International Monetary Fund (IMF), the containment measures of China, its economy contracted by 6.8 percent in 2020 Q1. On the other side, "the U.S. economy contracted by 31.4 percent in the second quarter of 2020 but has rebounded strongly since then. The unemployment rate stayed at 5.8 percent in May 2021" (IMF). Nevertheless, it is worth noting that most of the existing literature focuses on immediate economic losses experienced by countries during COVID-19. There is little analysis and comparison of the long-term economic impacts of the policies of the two countries, particularly across countries with differing governance structures. This report aims to address this gap by exploring two central research questions:

(1) To what extent do quarantine policies influence economic outcomes?

(2) How do countries with different government structures, such as the United States and China, experience these effects differently?

This paper aims to address these questions through a comparative analysis of regional unemployment rates and housing price indices, using statistical models and quarantine policy data from both nations. Understanding the economic consequences of quarantine policies is critical for improving policy responses in future crises. This paper's goal is to help governments design more balanced and adaptive frameworks that mitigate both health and financial damage.

DATA

2.1. Data Sources

The data fed to the model are collected from various sources to make a good read on the effect of various quarantine measures. These sources include institutional databases and official government statistics.

2.1.1. Quarantine Policy Index

Data on the stringency of quarantine policy for different countries were obtained from the Oxford COVID-19 Government Response Tracker (OxCGRT), maintained by the Blavatnik School of Government (BSG) at the University of Oxford (Hale et al. 2023). It provides a standardized and comparable record of government response to COVID-19 across 185 countries from January 1, 2020, to December 31, 2022, and the data was recorded daily by a team of trained volunteers. While the dataset consists of 25 indicators measuring government response, including economic policies, vaccination policies, and so on, we only select 8 of them that are related to quarantine policy. These include variables such as school closure, workplace closure, cancellation of public events, restriction on public transport, stay-at-home requirement, restriction on internal movement, and international movement control. These indicators are coded on an ordinal scale ranging from 0 to 5, representing increasing levels of policy stringency. For example, the "Restrictions on gathering " indicator is coded as follows: 0

indicates no restrictions; 1 represents limitations on very large gatherings (over 1000 people); 2 corresponds to restrictions on gatherings between 101-1000 people; 3 corresponds to restrictions on gatherings of 11-100 people and so on.

2.1.2 US Regional Unemployment Data

US regional unemployment rate data was obtained from the National Conference of State Legislatures (NCSL), which likely compiles data originating from the Bureau of Labor Statistics (BLS). This dataset provides a monthly unemployment rate for each state in the United States, covering the period from 1976 to the present. To analyze the changes in the unemployment rate during the pandemic period, data spanning from January 2019 to December 2022 will be selected, and the year 2019 will serve as the pre-pandemic baseline.

2.1.3. China Regional Unemployment Data

Data on the regional unemployment rate in China were obtained from China's National Bureau of Statistics (NBS). In contrast to U.S. unemployment data, China's unemployment data are only reported annually at the provincial level (excluding Hong Kong, Macau, and Taiwan). Moreover, NBS has not released data beyond 2021. As a result, our analysis for China is limited to the years 2019, 2020, and 2021.

2.1.4 US Regional House Price Index

Data on the regional house price index in the U.S. was sourced from the Federal Housing Finance Agency (FHFA). This dataset provides annual house price index (HPI) values for each state (include 50 states and the District of Columbia) from 1975 to 2024, with base years set to both 1990 and 2000. However, for the purpose of analyzing changes of HPI during the pandemic period, we extract data from 2019 to 2022 and re-calculate the data using 2019 as the base year, so that HPI = 100 in 2019 for all states. Therefore, HPIs in each state are expressed relative to

their 2019 levels, allowing us to assess the relative growth or decline in housing prices throughout the pandemic. For example, Alabama's HPI rose to 104 in 2020, indicating a 4% increase in housing prices relative to its 2019 level.

2.2. Data Preparation and Alignment

To combine these disparate datasets and conduct further analysis, careful alignment based on region and time is needed. While the quarantine policy indexes are reported on a daily basis, the economic indicators, such as the unemployment rate and housing price index, are available at monthly or yearly intervals. Therefore, we aggregate the daily policy data to match the frequency of the economic data by calculating their monthly and yearly averages. For example, to obtain a monthly value for the "School closing" index, we calculate the average of its daily scores within that month for each state. The aim of using average values rather than maximum values is to reflect the overall policy strictness experienced over a given period.

METHODOLOGY

The main analytical approach employed in this study is multiple linear regression. To make a meaningful comparison between the U.S. and China, separate regression models are constructed for each country, allowing us to evaluate how different quarantine policies influence their unemployment rate. Moreover, to explore the potential impact of quarantine measures on the housing market in the U.S., an additional regression model is developed with the HPI as the dependent variable. However, given the potential spatial autocorrelation in the housing market, spatial analysis is conducted to evaluate the correlation of HPI among neighboring states. Based on the result, we will determine whether it is appropriate to retain the LM or to adopt a spatial regression framework such as SAR/CAR.

3.1. Regression Model

The general form of Multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_k x_k + \ \epsilon$$

Where:

- y is the unemployment rate / HPI in different region
- x_i are predictors, including states, month/year, and 8 quarantine policy index (School Closing, Workplace closing, Cancel public event, Restriction on gathering, Close public transport, State at home requirement, Restrictions on internal movement, International travel control).
- β_o is an intercept term, which represents the baseline unemployment rate/HPI for a given state and time period (month/year) under the situation in which no quarantine policies are implemented.
- $\beta_1 \ - \ \beta_k$ are coefficient of each x_i , which represent the estimated effect of that specific quarantine policy on the unemployment rate/HPI, holding all other variables constant.
- ϵ: Error term, which captures the difference between the actual observed value and the value predicted by the model.

3.1.1. Variable Selection and Diagnostics

Before building the model, an exploratory data analysis (EDA) was conducted to gain a comprehensive understanding of the dataset. This step helps identify the distribution of each variable, potential outliers, and the relationship between each x and y, thereby informing suitable modeling strategies and variable selection. For instance, a scatter plot of U.S. unemployment rate against the workplace closing index (colored by different levels of school closing index) (Fig.2)

was generated, and it was observed that the relationship between them varies depending on the intensity of school closures, which suggests the potential interaction term in the model.

The base model, including only the main variables, was first developed. To ensure the validity of the analysis, multicollinearity was checked by calculating the variance inflation factor (VIF). For variables with multiple degrees of freedom, the adjusted GVIF1/(2*DF) was used instead. The variable with high GVIF1/(2*DF) (>5) was iteratively removed, one at a time, until all remaining variables exhibited acceptable levels of collinearity. This step is very crucial because high multicollinearity can inflate the variance of the coefficients of estimates, making them unstable and difficult to interpret. By removing the predictor with high VIF/GVIF1/(2*DF), each predictor's effect on the dependent variable is estimated reliably and independently.

Building upon the reduced model, all possible interactions and quadratic terms for the remaining variables were added to capture non-linear relationships and interaction effects. Subsequently, non-significant terms were removed one by one following the order of quadratic terms, interaction terms, and finally main effects. The non-significant main variables will be retained in the model if they are involved in significant quadratic or interaction terms, in order to ensure proper interpretation of higher-order effects.

Finally, three assumptions must be satisfied by the final model, which includes homoscedasticity (constant variance), independence of residuals, and normality of residuals. If any of the assumptions are violated, corrective action such as removing outliers or applying appropriate data transformation may be necessary.

3.2. Spatial Analysis

For the unemployment rate model, no further analysis is required. However, for the HPI model of the U.S., spatial analysis is necessary for checking the potential spatial autocorrelation across regions.

To examine the spatial correlation in the residual of the HPI regression model, the geographic boundary data for each state in the U.S. were obtained, excluding Puerto Rico. State names are standardized to ensure consistency with the main dataset. Then, a spatial neighborhood structure was created based on the contiguity of state borders, identifying which state shares a border. Using the neighborhood structure, a spatial weight matrix was constructed to measure the spatial relationships between states. To test the degree of spatial dependence, Moran's I test was used on both the average residuals across the entire study period and on residuals for each year (2020, 2021, 2022). If no correlation is detected in the residuals, it would suggest that the linear regression model sufficiently captures the structure of the data, and further spatial analysis is not necessary. Conversely, suppose spatial autocorrelation is present in the residual. In that case, it indicates a violation of the independence assumption for linear regression is violated, suggesting that alternative models such as Spatial Autoregressive (SAR) and Conditional Autoregressive (CAR) are necessary.

SAR Model:

$$Y = X\beta + \lambda W(Y - X\beta) + \epsilon$$

Where:

- Y: vector of the dependent variable (HPI)
- X: the matrix of independent variables (quarantine policy index)
- β : the vector of coefficient for predictor

- λ: spatial autocorrelation parameter
- W: spatial weights matrix indicating neighboring structure
- ϵ: error term

CAR Model:

$$Y_i|Y_{j\sim i} \sim N(X\beta + \lambda W(Y - X\beta) + \ldots + \epsilon)$$

Where:

- Y_i: the dependent variable (HPI) in region i
- $Y_{j\sim i}$: the dependent variable (HPI) in neighbors of unit i
- X: the matrix of independent variable (quarantine policy index)
- β: the vector of coefficient associated with the independent variable (quarantine policy index)
- λ: spatial autoregressive coefficient that capture the spatial dependence
- W: spatial weight matrix
- ϵ: error term

RESULTS

4.1. EDA Result

Based on the distribution plots of dependent variables, it is clear that the U.S. experienced higher and more volatile unemployment rates compared to China during the pandemic period. While China's unemployment rate remained relatively stable between 2% and 4%, the U.S. data exhibited a wide spread, with values ranging from 2% to 30%. It suggests that the economic impact of COVID-19 was more severe and uneven across US states than across CHinese provinces, which shows the same result with the IMF.

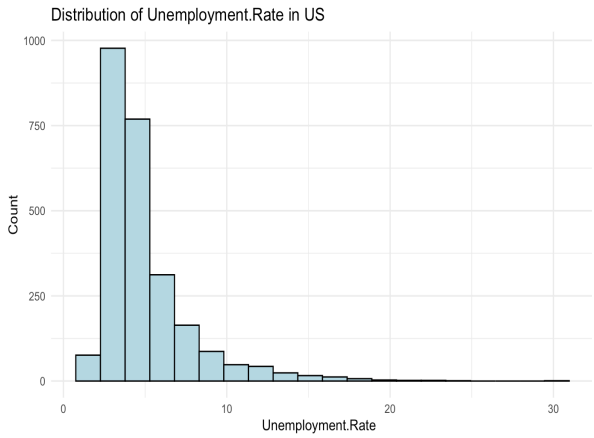

Fig.1 Distribution of monthly unemployment rate across U.S. states.

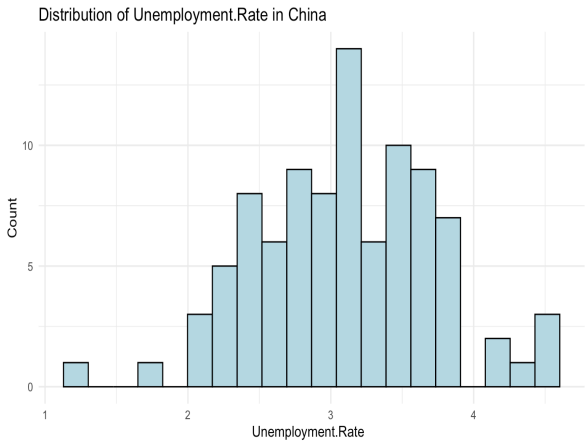

Fig.2 Distribution of yearly unemployment rate across China provinces.

4.2. US Unemployment Rate Model

Model Result:

Characteristic	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9086951	0.0274945	33.05	< 2e-16***
x1 Alaska	0.4828231	0.0342719	14.088	< 2e-16***
x1 Arizona	0.4065343	0.0341345	11.91	< 2e-16***
x1 Arkansas	0.1743081	0.0341237	5.108	3.50e-07***
x1 California	0.414999	0.034448	12.047	< 2e-16***
x1 Colorado	0.1417805	0.0344636	4.114	4.02e-05***
x1 Connecticut	0.360778	0.0342899	10.521	< 2e-16***
x1 Delaware	0.2746458	0.0342201	8.026	1.54e-15***
x1 District of Columbia	0.6651409	0.0361504	18.399	< 2e-16***
x1 Florida	0.1684867	0.034413	4.896	1.04e-06***
x1 Georgia	0.1639869	0.0340947	4.81	1.60e-06***
x1 Hawaii	0.1438556	0.0351059	4.098	4.31e-05***
x1 Idaho	0.0108904	0.0341804	0.319	0.750045
x1 Illinois	0.4305883	0.0345188	12.474	< 2e-16***

x1 Indiana	0.1274678	0.0340047	3.749	0.000182***
x1 Iowa	0.0213118	0.0340142	0.627	0.531008
x1 Kansas	0.0462088	0.0342192	1.35	0.177019
x1 Kentucky	0.291811	0.034424	8.477	< 2e-16***
x1 los Angeles County	0.7814606	0.0361504	21.617	< 2e-16***
x1 Louisiana	0.3506315	0.0347467	10.091	< 2e-16***
x1 Maine	0.0571707	0.0344842	1.658	0.097469.
x1 Maryland	0.1636312	0.0341412	4.793	1.74e-06***
x1 Massachusetts	0.2543606	0.0346077	7.35	2.68e-13***
x1 Michigan	0.4119691	0.0340986	12.082	< 2e-16***
x1 Minnesota	0.0576245	0.0340775	1.691	0.090966.
x1 Mississippi	0.4445322	0.0339999	13.075	< 2e-16***
x1 Missouri	0.071029	0.034203	2.077	0.037933*
x1 Montana	0.0311407	0.0342277	0.91	0.363011
x1 Nebraska	-0.1296065	0.0342513	-3.784	0.000158***
x1 Nevada	0.5982778	0.0340553	17.568	< 2e-16***
x1New Hampshire	-0.0533172	0.0342183	-1.558	0.119326
x1New Jersey	0.3975925	0.034061	11.673	< 2e-16***
x1New Mexico	0.5108082	0.0350483	14.574	< 2e-16***
x1New York	0.4543356	0.0345529	13.149	< 2e-16***
x1New York city	0.8085079	0.0361504	22.365	< 2e-16***
x1 North Carolina	0.2647656	0.0345843	7.656	2.74e-14***
x1 North Dakota	-0.1998517	0.0340636	-5.867	5.03e-09***
x1 Ohio	0.3462657	0.0339881	10.188	< 2e-16***
x1 Oklahoma	0.0848511	0.0343299	2.472	0.013516*
x1 Oregon	0.2668982	0.0342333	7.796	9.30e-15***
x1 Pennsylvania	0.4096364	0.0340445	12.032	< 2e-16***
x1 Rhode Island	0.2419577	0.0345321	7.007	3.13e-12***
x1 South Carolina	0.1183272	0.034171	3.463	0.000544***
x1 South Dakota	-0.1844778	0.0339988	-5.426	6.32e-08***
x1 Tennessee	0.2256587	0.0341725	6.604	4.90e-11***
x1 Texas	0.3083861	0.034178	9.023	< 2e-16***
x1 Utah	-0.1452651	0.0340732	-4.263	2.09e-05***
x1 Vermont	-0.1945936	0.0345274	-5.636	1.94e-08***

x1 Virginia	0.0179843	0.0340692	0.528	0.597634
x1 Washington	0.3007971	0.0341801	8.8	< 2e-16***
x1 West Virginia	0.4118937	0.034122	12.071	< 2e-16***
x1 Wisconsin	0.0905632	0.0340174	2.662	0.007812**
x1 Wyoming	0.2430394	0.0364846	6.661	3.33e-11***
x2 2020	0.2740059	0.0199113	13.761	< 2e-16***
x2 2021	0.2367743	0.0223858	10.577	< 2e-16***
x2 2022	-0.0438445	0.0125217	-3.501	0.000471***
x3 Month	0.0482243	0.0044773	10.771	< 2e-16***
x4 School closing	0.0773384	0.0095034	8.138	6.30e-16***
x5 Workplace closing	-0.1492	0.0172733	-8.638	< 2e-16***
x6 Cancel public event	0.0503323	0.0125773	4.002	6.47e-05***
x8 Close public transport	0.0885373	0.0202724	4.367	1.31e-05***
x9 State home requirement	-0.0271293	0.027504	-0.986	0.324045
x10 Restriction on internal movement	-0.013185	0.0204466	-0.645	0.519083
x11 International travel control	-0.0583943	0.0091698	-6.368	2.27e-10***
(Month)^2	-0.0036407	0.0003332	-10.926	< 2e-16***
(State home requirement)^2	0.0869331	0.0139942	6.212	6.11e-10***
(Restriction on internal movement)^2	0.0606383	0.0112646	5.383	8.01e-08***
(School closing)*(Workplace closing)	0.1197359	0.0074299	16.115	< 2e-16***
(Close public transport)*(Restriction on internal movement)	-0.1051624	0.0158852	-6.62	4.39e-11***

The final model for US unemployment rate:

$$log(y) = 0.91 + \beta_1 x_1 + 0.27x_{21} + 0.24x_{22} - 0.04x_{23} + 0.05x_3 + 0.08x_4 - 0.15x_5 + 0.05x_6 -$$

$$+ 0.09x_8 - 0.03x_9 - 0.01x_{10} - 0.06x_{11} + 0.12(x_4 * x_5) - 0.11(x_8 * x_{10}) - 0.003x_3^2$$

$$+ 0.09x_9^2 + 0.06x_{10}^2$$

Where:

- y = Unemployment rate for U.S.
- x_1 = dummy variable for States
- x_{21} = dummy variable for Year 2020
- x_{22}= dummy variable for Year 2021

- x_{23}= dummy variable for Year 2022
- x_3 = Month
- x_4 = School closing
- x_5 = Workplace closing
- x_6 = Cancel public event
- x_8 = Close public transport
- x_9 = State home requirement
- x_{10} = Restriction on internal movement
- x_{11} = International travel control

When checking the assumption, all three key conditions - homoscedasticity, normality, and independence of residuals - were initially violated. The residual vs. predicted values plot revealed a clear pattern of amplification, which means the variance of residuals increases with the fitted values, indicating homoscedasticity. For testing normality, a Shapiro-Wilk test was conducted. The result p-value (<2.2e-16) is significantly below the conventional threshold of 0.05, violating the normality assumption. Furthermore, the result p-value (<2.2e-16) from the Durbin-Watson test is also smaller than 0.05, indicating possible autocorrelation and violation of the independence assumption. Fortunately, after applying a log transformation to the dependent variable y, the assumption of constant variance was satisfied. The normality assumption is also improved, and most of the points closely followed the reference line in the QQ-plot with only slight deviations at the tails, suggesting a near-normal distribution. However, the independence assumption remained problematic since the data was collected over time for the same states. While this violation may affect the precision of standard error estimates, the primary goal of this study is on the relative effect size. Nonetheless, the result should be interpreted cautiously.

Overall, this model demonstrates strong explanatory power, with an adjusted R-squared of 86.07%, which means that 86.07% of unemployment data in the U.S. can be explained by this

model. The following paragraphs provide a detailed interpretation of each important term included in the model.

In the model, $\beta_1 x_1$ represents a set of dummy variables for the U.S. States, capturing regional effects. However, given the large number of categories (50 states plus the District of Columbia), the term is summarized in compact notation for clarity. According to the results, most of the p-values associated with the state dummy variables are very significant ($p < 0.05$), which represents that unemployment rates vary significantly across different states. For most of the state, the estimated coefficients are positive, indicating a generally increasing trend in unemployment rates compared to the reference state. Among them, New York City exhibits the highest coefficient (0.81), suggesting that it experienced the most serious unemployment impact during the pandemic period.

One of the significant main effects in the model is the international reveal control index, which has a coefficient of -0.06. Given the log-linear specification of the model, this implies that each one-unit increase in the international travel control index is associated with an approximate 0.058% decrease in the unemployment rate, holding other factors constant. This is derived from the transformation $\exp(-0.06) - 1 \approx -0.058$, which converts the change in the log of the unemployment rate into a change in the actual unemployment rate.

The model also includes several interaction and quadratic terms. The quadratic term for State Home Requirement (x_9^2) has a positive coefficient (0.09), suggesting a non-linear U-shaped relationship between “Stay at home restriction” and the unemployment rate. This indicates that the economic impact of this policy is not constant across levels of restriction. Specifically, mild stay-at-home orders may have a limited or neutral effect on unemployment, while more severe or prolonged restrictions are associated with a sharper increase in unemployment. This finding

highlights the importance of considering policy intensity and duration when evaluating the economic consequences of pandemic responses.

Additionally, the model includes both a main effect of school closures and an interaction term between school and workplace closures. The coefficient of workplace closing is 0.08, suggesting that, in the absence of workplace closures (i.e., when x5 = 0), each one unit increase in the school closing index will lead to a 0.08 percentage increase in the log of unemployment rate. However, the interaction terms between school closing and workplace closings indicate that the combined impact of these two policies on unemployment is greater than their individual effects. As shown in Figure 3, when the school closing index increases, higher levels of workplace closures tend to be associated with higher unemployment rates. One possible explanation for this is that school closures can indirectly affect labor force participation because working parents may be forced to stay home to care for caring children. When school closing and workplace closing are implemented simultaneously, the ability for individuals to adapt to this new situation is further constrained, amplifying the overall impact on employment.

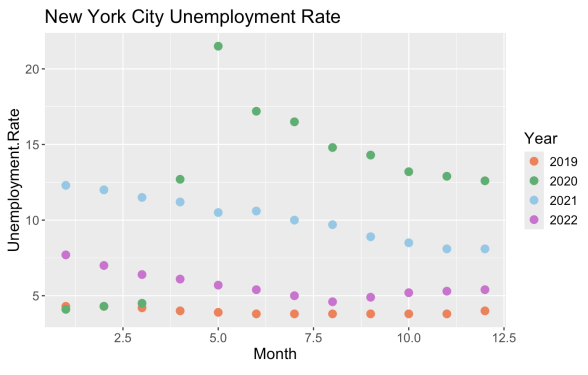

Fig.3 New York City's unemployment rate experienced a marked increase during the beginning of 2020, peaking at 22% in May. Following this apex, the rate steadily

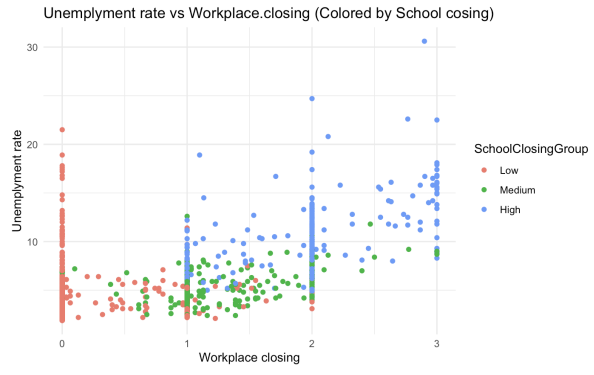

Fig.4 Scatter plot showing the relationship between workplace closings and unemployment rate, categorized by school closing intensity.

declined, eventually close to pre-pandemic levels by 2022.

4.3. China Unemployment Rate Mode:

Model Results:

Characteristic	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.43321	0.21125	11.518	< 2e-16 ***
t1 Beijing	-0.18519	0.28608	-0.647	0.519844
t1 Chongqing	0.73178	0.28508	2.567	0.012730 *
t1 Fujian	0.91111	0.28493	3.198	0.002197 **
t1 Gansu	0.70998	0.28704	2.473	0.016181 *
t1 Guangdong	-0.18306	0.28495	-0.642	0.523008
t1 Guangxi	0.11094	0.28708	0.386	0.700507
t1 Guizhou	1.25338	0.28623	4.379	4.76e-05 ***
t1 Hainan	0.25568	0.28917	0.884	0.380052
t1 Hebei	0.69692	0.28656	2.432	0.017966 *
t1 Heilongjiang	0.73129	0.2849	2.567	0.012734 *
t1 Henan	0.6541	0.28498	2.295	0.025171 *
t1 Hubei	0.31951	0.28497	1.121	0.266598
t1 Hunan	0.09009	0.28922	0.311	0.756488
t1 Inner Mongolia	1.23883	0.28687	4.318	5.88e-05 ***
t1 Jiangsu	0.33936	0.28584	1.187	0.239735
t1 Jiangxi	0.38622	0.2854	1.353	0.180963

t1 Jilin	0.65116	0.28496	2.285	0.025797 *
t1 Liaoning	1.85707	0.2876	6.457	1.97e-08 ***
t1 Ningxia	1.28057	0.28494	4.494	3.18e-05 ***
t1 Qinghai	-0.49796	0.28674	-1.737	0.087506 .
t1 Shaanxi	0.79285	0.28491	2.783	0.007163 **
t1 Shandong	0.50184	0.28512	1.76	0.083405 .
t1 Shanghai	0.71833	0.28496	2.521	0.014341 *
t1 Shanxi	0.09965	0.2851	0.35	0.727899
t1 Sichuan	0.96006	0.28644	3.352	0.001383 **
t1 Tianjin	1.10793	0.28842	3.841	0.000294 ***
t1 Tibet	0.1664	0.2849	0.584	0.561346
t1 Xinjiang	-0.34264	0.28762	-1.191	0.238149
t1 Yunnan	1.04271	0.28492	3.66	0.000529 ***
t1 Zhejiang	0.03972	0.28518	0.139	0.889678
t6 Restriction on gathering	0.11627	0.03694	3.148	0.002547 **

The final model for China unemployment rate:

$$z = 2.43 + \beta_1 t_1 + 0.12 t_6$$

Where:

- z = Unemployment rate for U.S.
- t_1 = Provinces
- t_6 = Restriction on gathering

For China's model, the adjusted R-squared is 68.24%, which is lower than the model for the US. Nevertheless, this model includes only two predictors. Even with this simple structure, these two variables still explain a substantial portion of the variation in the unemployment rate.

One of the predictors in this model is "Restriction on gathering," which has a coefficient of 0.12. It means that holding other variables constant, when the restriction on gathering index increases by 1, the unemployment rate will increase by 0.12. The other predictors in the model are "Provinces". Similar to the US case, the notation 1t1 denotes the set of dummy variables for all provinces. According to the result, Liaoning (coefficient: 1.86) experienced the highest unemployment relative to the reference province. As shown in the figure.3, an interesting observation is that the unemployment rate in some provinces, such as Heilongjiang, Qinghai, and Shandong, continued to decline during the pandemic period. It may be attributed to regional differences because these provinces have greater reliance on agriculture, mining, and state-owned enterprises, which may have experienced less disruption during the pandemic.

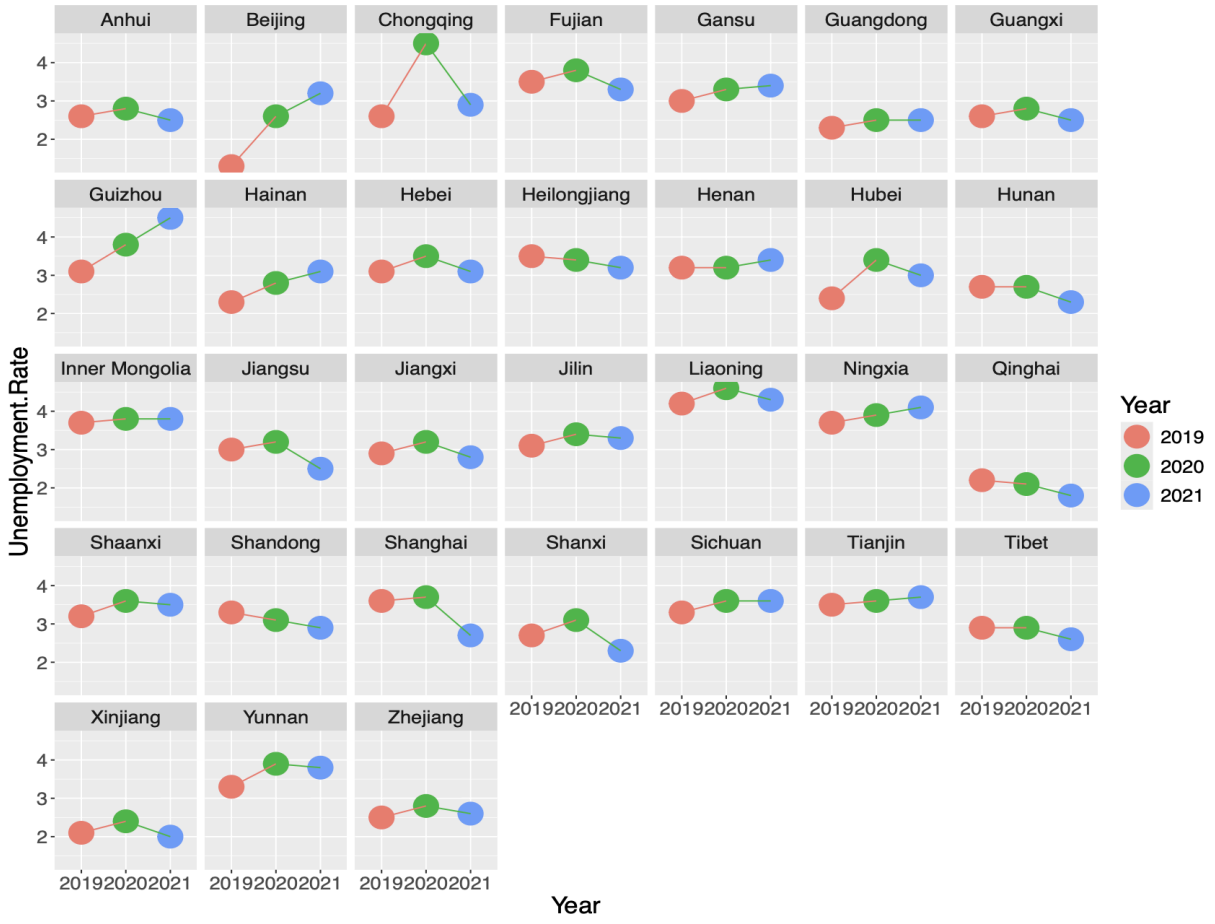

Fig.5 This chart displays the unemployment rates (%) for various provinces in China over three consecutive years: 2019 (red), 2020 (green), and 2021 (blue). The data reflects the impact of the COVID-19 pandemic, with noticeable increases in unemployment in 2020 for many regions, followed by partial recovery in 2021.

4.4. US Housing Price Model

Linear Regression Model Result:

Characteristic	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	101.2534	13.409	7.551	4.00e-12 ***
k1 Alaska	6.6998	18.3719	0.365	0.715869
k1 Arizona	17.1876	17.9407	0.958	0.339603
k1 Arkansas	-46.3586	17.9309	-2.585	0.010685 *
k1 California	173.36	18.5606	9.34	< 2e-16 ***

k1 Colorado	10.1029	18.4481	0.548	0.584759
k1 Connecticut	-4.3169	18.3176	-0.236	0.814012
k1 Delaware	-6.8738	18.2757	-0.376	0.707363
k1 District Of Columbia	123.6573	17.8833	6.915	1.28e-10 ***
k1 Florida	-69.9196	18.1871	-3.844	0.000178 ***
k1 Georgia	8.8871	18.1179	0.491	0.624493
k1 Hawaii	43.3936	19.399	2.237	0.026778 *
k1 Idaho	-18.2501	18.2586	-1	0.319158
k1 Illinois	-36.446	18.3156	-1.99	0.048433 *
k1 Indiana	-15.893	18.3842	-0.864	0.388708
k1 Iowa	-3.1137	17.8818	-0.174	0.862001
k1 Kansas	1.2948	18.1364	0.071	0.943183
k1 Kentucky	5.4734	18.3313	0.299	0.765676
k1 Louisiana	-10.3841	18.5513	-0.56	0.57649
k1 Maine	35.7011	18.217	1.96	0.051889 .
k1 Maryland	4.8081	18.2697	0.263	0.79278
k1 Massachusetts	18.6553	18.3336	1.018	0.310541
k1 Michigan	-27.8396	18.4146	-1.512	0.132697
k1 Minnesota	7.5089	18.2804	0.411	0.681839
k1 Mississippi	-59.2291	18.173	-3.259	0.001384 **
k1 Missouri	7.1354	18.2033	0.392	0.695631
k1 Montana	24.4165	18.2176	1.34	0.182197
k1 Nebraska	-24.047	17.9364	-1.341	0.182065
k1 Nevada	18.8984	18.2818	1.034	0.302939
k1New Hampshire	-26.589	17.8772	-1.487	0.139046
k1New Jersey	7.3982	18.3027	0.404	0.686634
k1New Mexico	-1.7386	18.4843	-0.094	0.925191
k1New York	21.203	18.3429	1.156	0.249561
k1 North Carolina	-34.2305	18.2791	-1.873	0.063075 .
k1 North Dakota	-35.6857	17.8728	-1.997	0.047685 *
k1 Ohio	9.7403	18.4729	0.527	0.598786
k1 Oklahoma	-0.6999	17.9001	-0.039	0.968864
k1 Oregon	61.0997	18.5161	3.3	0.001210 **

k1 Pennsylvania	-44.0626	18.1129	-2.433	0.016173 *
k1 Rhode Island	26.2264	18.3099	1.432	0.154133
k1 South Carolina	4.9517	17.9646	0.276	0.783211
k1 South Dakota	-1.2267	17.8731	-0.069	0.945372
k1 Tennessee	-2.2958	18.2228	-0.126	0.899913
k1 Texas	-9.0901	18.3564	-0.495	0.621188
k1 Utah	21.5207	18.0344	1.193	0.234643
k1 Vermont	8.6034	18.2886	0.47	0.63874
k1 Virginia	4.3757	18.2412	0.24	0.810753
k1 Washington	47.0049	18.5246	2.537	0.012195 *
k1 West Virginia	-64.8578	18.1224	-3.579	0.000466 ***
k1 Wisconsin	41.539	18.3142	2.268	0.024758 *
k1 Wyoming	4.832	18.1069	0.267	0.789946
k2 2020	20.8753	11.5744	1.804	0.073318 .
k2 2021	14.712	7.1623	2.054	0.041716 *
k2 2022	17.6775	5.0057	3.532	0.000550 ***
k6 Restriction on gathering	-9.2295	4.3699	-2.112	0.036351 *

The Linear Regression model for US housing price index:

$$h = 101.25 + \beta_1 k_1 + 20.88k_{21} + 14.71k_{22} + 17.68k_{23} - 9.23k_6$$

Where:

- h = housing price index for different states in different years standardized using data in 2019 as baseline
- k_1: dummy variable for States
- k_{21}: dummy variable for Year 2020
- k_{22}: dummy variable for Year 2021
- k_{23}: dummy variable for 2022
- k_6: Restriction on gathering

Moran's I Test Result:

Moran I statistics	Expectation	Variance	s.d.	p-value
0.005302399	-0.020833333	0.001949384	0.59195	0.2769

To test whether there is a spatial autocorrelation within a dataset, the Moran I statistic is one of the most widely used indicators. A Moran's I statistic close to +1 indicates a strong positive spatial autocorrelation, meaning that similar values tend to cluster together geographically. Conversely, a value close to -1 suggests strong negative spatial autocorrelation, where neighboring regions tend to have different values. A value near 0 implies no significant spatial autocorrelation, indicating a random spatial distribution of the variable. According to the result of Moran's I test in this study, the Moran I statistic was 0.0053, which is very close to 0. As a result, no significant spatial autocorrelation was detected in the residuals of the US HPI model, indicating that the linear regression model is sufficient to capture the relationship between housing prices and the predictors. According to the final LM model, the adjusted R-squared is 70.6%, which represents a strong fit.

Unlike its positive effect on unemployment in China, the "Restriction on gathering" policy negatively influenced the HPI in the US. Specifically, the HPI is estimated to decrease by 9.23 points for every one-unit increase in the restriction index. It could be attributed to factors such as limited mobility for housing viewing, disrupted construction activities, and reduced consumer confidence. Interestingly, although some states showed consistent price decline, others experienced flattening or even growth, indicating that policy impact was not uniform and may have interacted with local economic structures.

DISCUSSION

This study provides a comparative analysis of how quarantine policies affected regional unemployment rates in the United States and China during the COVID-19 pandemic. In addition, it also investigates various containment measures affecting housing price trends in the U.S., with particular attention to spatial dependencies across states.

One of the key findings of this paper is that quarantine policies have a measurable and significant impact on economic indicators. Another valuable finding is that different countries have different sensitivities to specific measures, reflecting differences in governance structures, economic composition, and policy implementation. By comparing the unemployment model for the US and China, we can find that their economies are affected by different quarantine policies. In the US, the unemployment rate is most strongly influenced by measures such as workplace closing, school closing, and international travel controls. In contrast, in China, the unemployment rate is primarily associated with restrictions on the gathering policy. Furthermore, it is revealed that the US unemployment rate is influenced by a broader range of quarantine measures, indicating greater policy sensitivity or regional variability in economic response. The reason behind it might be that the US has a decentralized nature of policy implementation, and each individual state adopted different measures. On the contrary, China's centralized and uniform enforcement of policies might reduce regional variation, resulting in fewer policy variables showing significant economic effects.

However, when economics is influenced by multiple quarantine policy variables simultaneously, policies may produce overlapping effects, interaction effects or nonlinear effects. As a result, it can be complicated to predict and manage economic impacts. Therefore, in a decentralized system like the U.S., better coordination among state and federal governments

could help ensure more consistent and effective policy implementation, reducing regional disparities in economic outcomes. Nonetheless, while China enforces uniform measures efficiently, incorporating more targeted and regional-specific economic support could improve resilience. Although quarantine policies are necessary to prevent the spread of the virus, their design should consider labor market dynamics and housing sector sensitivities in different provinces. Because national/regional shutdowns may be effective in the short term, but can have long-lasting economic effects.

Finally, certain limitations associated with this analysis must be acknowledged. One key issue is the uneven data period of the U.S. and China. While the U.S. dataset includes monthly unemployment data up to 2022, the data for China are only available on an annual basis and end in 2021. This difference may affect the comparability of post-pandemic recovery trends and also limit the precision of cross-national conclusions. Additionally, economic outcomes may also be influenced by concurrent stimulus policies, such as government subsidies or fiscal relief programs, which were not included in the model. The omission of these factors may have confounded the estimated result, as they could have either amplified or attenuated the observed economic impacts. Therefore, future researchers may consider including macroeconomic policy variables into the model, which allows for a more comprehensive assessment of policy interactions. Furthermore, since only two countries are included in this analysis, expanding to additional countries-especially those with hybrid systems or different cultural compliance norms, could also improve the generalizability of the findings.

REFERENCE

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2013). *Applied spatial data analysis with R* (Vol. 2). New York: Springer.

Chen, Y. H., & Fang, C. T. (2020). Mortality from COVID-19: A cross-country comparison of containment versus mitigation strategy. *Journal of the Formosan Medical Association, 119*(11), 1710–1712. https://doi.org/10.1016/j.jfma.2020.05.029

FHFA.gov. (2025). *FHFA House Price Index® Datasets | FHFA*. https://www.fhfa.gov/data/hpi/datasets#qpo

Kira, B., Saptarshi, M., Thayslene, M., Oliveira, M., Nagesh, R., Phillips, T., Pott, A., Sampaio, J., Tatlow, H., Wade, A., Webster, S., Wood, A., Zha, H., Zhang, Y., Torness, W., Vaccaro, A., Laping, S., Kamenkovich, N., Ren, L., & Hale, T. (2023). *BSG Working Paper Series: Variation in government responses to COVID-19*. https://www.bsg.ox.ac.uk/sites/default/files/2023-06/BSG-WP-2020-032-v15.pdf

Lyu, S., et al. (2023). One Pandemic, Two Solutions: Comparing the U.S.-China Response and Health Priorities to COVID-19 from the Perspective of "Two Types of Control". *Healthcare (Basel), 11*(13), 1848. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10341116/

Mendenhall, W., & Sincich, T. (2012). *A second course in statistics: Regression analysis*. Boston: Prentice Hall.

National Bureau of Statistics of China. (2024). *National Data*. https://data.stats.gov.cn/english/easyquery.htm?cn=A01

APPENDIX

This section provides the R code used for the analysis of U.S. unemployment and housing price index (HPI) models. The code includes data processing, model construction, and diagnostic checks. The analytical procedure for China's unemployment model follows a similar structure to U.S. unemployment analysis and is therefore not repeated in full.

Appendix A. R code for US unemployment analysis

#install and library package
```{r}
library(RColorBrewer)
library(ggplot2)
library(gridExtra)
#install.packages("maps")
library(maps)
library(tigris)
library(dplyr)
```

The Data
```{r}
us_unemployment <- read.csv("/Users/maxiaohui/Desktop/Honor Capstone project 2024-25/Honor project data/Unemployment Rate/us_unemployment_combine.csv")
head(us_unemployment)

summary(us_unemployment)

us_unemployment$Year <- as.factor(us_unemployment$Year)
str(us_unemployment)

#Check NA value
colSums(is.na(us_unemployment))
```

##EDA: Distribution of response variables
```{r}
r <- ggplot(us_unemployment, aes(x=Unemployment.Rate  )) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
```

```
  labs(title="Distribution of Unemployment.Rate ", x="Unemployment.Rate  ", y="Count") +
  theme_minimal()
r
```

EDA: Distribution of predictors

```{r}
p1 <- ggplot(us_unemployment, aes(x=School.closing)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of School.closing", x="School.closing", y="Count") +
  theme_minimal()
p2 <- ggplot(us_unemployment, aes(x=Workplace.closing)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Workplace.closing", x="Workplace.closing", y="Count") +
  theme_minimal()
p3 <- ggplot(us_unemployment, aes(x=Cancel.public.event)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Cancel.public.event ", x="Cancel.public.event ", y="Count") +
  theme_minimal()
p4 <- ggplot(us_unemployment, aes(x=Restriction.on.gathering)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Restriction.on.gathering", x="Restriction.on.gathering", y="Count")
+
  theme_minimal()
p5 <- ggplot(us_unemployment, aes(x=Close.public.transport)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Close.public.transport", x="Close.public.transport", y="Count") +
  theme_minimal()
p6 <- ggplot(us_unemployment, aes(x=State.at.home.requirement )) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of State.at.home.requirement ", x="State.at.home.requirement ",
y="Count") +
  theme_minimal()
p7 <- ggplot(us_unemployment, aes(x=Restrictions.on.internal.movement)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of internal.movement control", x="Restrictions.on.internal.movement",
y="Count") +
  theme_minimal()
p8 <- ggplot(us_unemployment, aes(x=International.travel.control )) +
```

```
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of International control ", x="International.travel.control", y="Count")
+
  theme_minimal()

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol= 2)
```
EDA: Response Vs. Predictors

```{r}
library("ggplot2")
library("gridExtra")

s1<-ggplot(us_unemployment, aes(x=School.closing, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s2<-ggplot(us_unemployment, aes(x=Year, y=Unemployment.Rate)) + geom_point(alpha = 0.5,
size = 1.5) + theme_minimal()
s3<-ggplot(us_unemployment, aes(x=Month, y=Unemployment.Rate)) + geom_point(alpha =
0.5, size = 1.5) + theme_minimal()
s4<-ggplot(us_unemployment, aes(x=Workplace.closing, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s5<-ggplot(us_unemployment, aes(x=Cancel.public.event, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s6<-ggplot(us_unemployment, aes(x=Restriction.on.gathering, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s7<-ggplot(us_unemployment, aes(x=Close.public.transport, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s8<-ggplot(us_unemployment, aes(x=State.at.home.requirement, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s9<-ggplot(us_unemployment, aes(x=Restrictions.on.internal.movement,
y=Unemployment.Rate)) + geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s10<-ggplot(us_unemployment, aes(x=International.travel.control, y=Unemployment.Rate)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()


grid.arrange(s1, s2,s3,s4, s5, s6, s7, s8,s9, s10, ncol = 3, nrow = 4)
```
```{r}
#Scatterplot for Month vs Unemployment rate for New York City
selected_states2 <- c("New York city")
```

```
filtered_data2 <- us_unemployment %>%
  filter(State %in% selected_states2)

ggplot(data = filtered_data2) +
geom_point(mapping = aes(x = Month, y = Unemployment.Rate, color = Year, line = Year),size
= 3) +
  scale_color_manual(values = c("2019" = "#FF7F50", "2020" = "#3CB371", "2021" =
"#87CEEB", "2022" = "#DA70D6")) +
  labs(title = "New York City Unemployment Rate")+
  theme(text = element_text(size = 14))
```
```{r}
#scatterplot of unemployment rate vs workplace closing (colored by school closing level)
us_unemployment$SchoolClosingGroup <- cut(
  us_unemployment$ School.closing,
  breaks = 3,
  labels = c("Low", "Medium", "High"),
  include.lowest = TRUE
)

ggplot(us_unemployment , aes(x = Workplace.closing, y = Unemployment.Rate, color =
SchoolClosingGroup)) +
  geom_point() +
  labs(title = "Unemplyment rate vs Workplace.closing (Colored by School cosing)",
       x = "Workplace closing",
       y = "Unemplyment rate") +
  theme_minimal()
```
EDA: Correlation check
```{r}
# Install and load the ggcorrplot package
library(ggcorrplot)

reduced_data <- subset(us_unemployment, select = c(School.closing, Workplace.closing,
Cancel.public.event, Restriction.on.gathering, Close.public.transport,State.at.home.requirement,
Restrictions.on.internal.movement, International.travel.control))

# Compute correlation at 2 decimal places
corr_matrix = round(cor(reduced_data), 2)
```

```
# Compute and show the  result
ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower",
         lab = TRUE)
```
Model Selection

model base

```{r}
library("car")
library(MASS)

#defining variables for easier model fitting
x1 <- us_unemployment$State
x2 <- us_unemployment$Year
x3 <- us_unemployment$Month
x4 <- us_unemployment$School.closing
x5 <- us_unemployment$Workplace.closing
x6 <- us_unemployment$Cancel.public.event
x7 <-  us_unemployment$Restriction.on.gathering
x8 <- us_unemployment$Close.public.transport
x9 <- us_unemployment$State.at.home.requirement
x10 <- us_unemployment$Restrictions.on.internal.movement
x11 <- us_unemployment$International.travel.control

model_base <- lm(Unemployment.Rate ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 +
x11, data = us_unemployment)

summary(model_base)

vif(model_base)
```
model 1 (full model with all interaction and quadratic term)

```{r}

model_1 <- lm(Unemployment.Rate ~ (x1 + x2 + x3 + x4 + x5 +x6 + x7 + x8 + x9 + x10 +
x11)^2 + I(x3^2) + I(x4^2) + I(x5^2) + I(x6^2) + I(x7^2) + I(x8^2) + I(x9^2) + I(x10^2) +
I(x11^2), data = us_unemployment)
```

```
summary(model_1)
```
model 2 (reduce quadratic terms)

```{r}
model_2 <- lm(Unemployment.Rate ~ (x1 + x2 + x3 + x4 + x5 +x6 + x7 + x8 + x9 + x10 + x11)^2 + I(x3^2) + I(x9^2) + I(x10^2) + I(x11^2), data = us_unemployment)

summary(model_2)
```
model 3 (reduce interaction terms)
```{r}

model_3 <- lm(Unemployment.Rate ~ x1 + x2 + x3 + x4 + x5 +x6 + x7 + x8 + x9 + x10 + x11 + x3:x4  + x4:x5  + x8:x10 + I(x3^2) + I(x9^2) +I(x10^2) , data = us_unemployment)

summary(model_3)
```
model 4 (reduce predictor terms)
```{r}
model_4 <- lm(Unemployment.Rate ~ x1 + x2 + x3 + x4 + x5 +x6 + x8 + x9 + x10 + x11  + x3:x4  + x4:x5  + x8:x10 + I(x3^2) + I(x9^2) +I(x10^2) , data = us_unemployment)

summary(model_4)

AIC(model_1, model_2, model_3, model_4)
```
Checking Linearity and Constant Variance

```{r}

plot_data2 <- data.frame(fitted = fitted(model_4), residuals = resid(model_4))

library(ggplot2)
ggplot(data = plot_data2, aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")
```

Checking Normality Assumption

```{r}
qqnorm(resid(model_4))
qqline(resid(model_4))

shapiro.test(resid(model_4))
```
Checking Independence

```{r}
library(car)
dwt(model_4)
```
model 5 (take log transformation)
```{r}
model_5 <- lm(log(Unemployment.Rate) ~ x1 + x2 + x3 + x4 + x5 +x6 + x8 + x9 + x10 + x11 +
x4:x5  + x8:x10 + I(x3^2) + I(x9^2) +I(x10^2) , data = us_unemployment)

summary(model_5)

```
###Checking Linearity and Constant Variance
```{r}
plot_data3 <- data.frame(fitted = fitted(model_5), residuals = resid(model_5))

```library(ggplot2)
ggplot(data = plot_data3, aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")
```
#Check independence
```{r}
#install.packages("lmtest")
library(lmtest)
library(car)
dwtest(model_5)
```
#Check normality
```

```{r}
qqnorm(resid(model_5))
qqline(resid(model_5))

shapiro.test(resid(model_5))
```

Appendix B. R code for US HPI analysis

```{r}
#install.packages(c("sp", "sf", "spdep"))
#install.packages("tigris")
#install.packages("spdep")
library(sp)
library(sf)
library(spdep)
library(tigris)
library(dplyr)
```

The data
```{r}
us_hpi <- read.csv("/Users/maxiaohui/Desktop/Honor Capstone project 2024-25/Honor project data/Unemployment Rate/us_hpi_combine.csv")
head(us_hpi)

summary(us_hpi)

us_hpi$Year <- as.factor(us_hpi$Year)
str(us_hpi)

#check NA.
colSums(is.na(us_hpi))
```
##EDA: Distribution of response variables
```{r}
library(ggplot2)
```

```
res <- ggplot(us_hpi, aes(x=HPI.with.base.2019)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of HPI ", x="HPI with 2019 base  ", y="Count") +
  theme_minimal()

res
```

EDA: Distribution of predictors

```{r}
library("gridExtra")

p1 <- ggplot(us_hpi, aes(x=School.closing)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of School.closing", x="School.closing", y="Count") +
  theme_minimal()
p2 <- ggplot(us_hpi, aes(x=Workplace.closing)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Workplace.closing", x="Workplace.closing", y="Count") +
  theme_minimal()
p3 <- ggplot(us_hpi, aes(x=Cancel.public.event)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Cancel.public.event ", x="Cancel.public.event ", y="Count") +
  theme_minimal()
p4 <- ggplot(us_hpi, aes(x=Restriction.on.gathering)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Restriction.on.gathering", x="Restriction.on.gathering", y="Count")
+
  theme_minimal()
p5 <- ggplot(us_hpi, aes(x=Close.public.transport)) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of Close.public.transport", x="Close.public.transport", y="Count") +
  theme_minimal()
p6 <- ggplot(us_hpi, aes(x=State.at.home.requirement )) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of State.at.home.requirement ", x="State.at.home.requirement ",
y="Count") +
  theme_minimal()
p7 <- ggplot(us_hpi, aes(x=Restrictions.on.internal.movement)) +
```

```
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of internal.movement control", x="Restrictions.on.internal.movement",
y="Count") +
  theme_minimal()
p8 <- ggplot(us_hpi, aes(x=International.travel.control )) +
  geom_histogram(fill="lightblue",color="black",bins=20) +
  labs(title="Distribution of International control ", x="International.travel.control", y="Count")
+
  theme_minimal()

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol= 2)

```
EDA: Response Vs. Predictors

```{r}
library("ggplot2")

s1<-ggplot(us_hpi, aes(x=School.closing, y=HPI.with.base.2019)) + geom_point(alpha = 0.5,
size = 1.5) + theme_minimal()
s2<-ggplot(us_hpi, aes(x=Workplace.closing, y=HPI.with.base.2019)) + geom_point(alpha =
0.5, size = 1.5) + theme_minimal()
s3<-ggplot(us_hpi, aes(x=Cancel.public.event, y=HPI.with.base.2019)) + geom_point(alpha =
0.5, size = 1.5) + theme_minimal()
s4<-ggplot(us_hpi, aes(x=Restriction.on.gathering, y=HPI.with.base.2019)) + geom_point(alpha
= 0.5, size = 1.5) + theme_minimal()
s5<-ggplot(us_hpi, aes(x=Close.public.transport, y=HPI.with.base.2019)) + geom_point(alpha =
0.5, size = 1.5) + theme_minimal()
s6<-ggplot(us_hpi, aes(x=State.at.home.requirement, y=HPI.with.base.2019)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s7<-ggplot(us_hpi, aes(x=Restrictions.on.internal.movement, y=HPI.with.base.2019)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()
s8<-ggplot(us_hpi, aes(x=International.travel.control, y=HPI.with.base.2019)) +
geom_point(alpha = 0.5, size = 1.5) + theme_minimal()


grid.arrange(s1, s2,s3,s4, s5, s6, s7, s8, ncol = 4, nrow = 2)
```
EDA: Correlation check

```{r}
# Install and load the ggcorrplot package
library(ggcorrplot)


reduced_data <- subset(us_hpi, select = c(School.closing, Workplace.closing,
Cancel.public.event, Restriction.on.gathering, Close.public.transport,State.at.home.requirement,
Restrictions.on.internal.movement, International.travel.control))

# Compute correlation at 2 decimal places
corr_matrix = round(cor(reduced_data), 2)

# Compute and show the  result
ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower",
        lab = TRUE)
```
Model Selection

model base
```{r}
library("car")
library(MASS)

#defining variables for easier model fitting
z1 <- us_hpi$State
z2 <- us_hpi$Year
z3 <- us_hpi$School.closing
z4 <- us_hpi$Workplace.closing
z5 <- us_hpi$Cancel.public.event
z6 <-  us_hpi$Restriction.on.gathering
z7 <- us_hpi$Close.public.transport
z8 <- us_hpi$State.at.home.requirement
z9 <- us_hpi$Restrictions.on.internal.movement
z10 <- us_hpi$International.travel.control

model_base <- lm( HPI.with.base.2019 ~ z1 + z2 + z3 + z4 + z5 + z6 + z7 + z8 + z9 + z10, data
= us_hpi)

summary(model_base)
```

check multicollinearity
```{r}
vif(model_base)
```
model 1 (reduce main predictors with high multicollinearity)
```{r}
model_1 <- lm( HPI.with.base.2019 ~ z1 + z2 + z4 + z6 + z7 + z8 + z9, data = us_hpi)

summary(model_1)

vif(model_1)
```
model 2 (add quadratic terms)

```{r}
model_2 <- lm( HPI.with.base.2019 ~ z1 + z2 + z4 + z6 + z7 + z8 + z9 + I(z4^2) + I(z6^2) + I(z7^2) + I(z8^2) + I(z9^2), data = us_hpi)

#model_2 <- lm( HPI.with.base.2019 ~ z1 + z2 + z4 + z6 + z7 + z8 + z9 + z1:z2 + z1:z4 + z1:z6 + z1:z7 + z1:z8 + z1:z9 + z2:z4 + z2:z6 + z2:z7 + z2:z8 + z2:z9 + z4:z6 + z4:z7 + z4:z8 + z4:z9 + z6:z7 + z6:z8 + z6:z9 +z7:z8 +z7:z9 +z8:z9 + I(z4^2) + I(z6^2) + I(z7^2) + I(z8^2) + I(z9^2), data = us_hpi)

summary(model_2)
```
model 3 (reduce quadratic terms)

```{r}
model_3<- lm( HPI.with.base.2019 ~ z1 + z2 + z4 + z6 + z7 + z8 + z9 + I(z4^2), data = us_hpi)

summary(model_3)
```
model 4 (add interaction terms)


```{r}
model_4<- lm( HPI.with.base.2019 ~ z1 + z2 + z4 + z6 + z7 + z8 + z9 + I(z4^2), data = us_hpi)

summary(model_4)
```

#model 5 (reduce non-significant main predictors)
```{r}
model_5<- lm( HPI.with.base.2019 ~  z1+ z2 + z6 , data = us_hpi)

summary(model_5)
```
#checking assumptions
```{r}
pre_res <- data.frame(fitted = fitted(model_5), residuals = resid(model_5))

library(ggplot2)
ggplot(data = pre_res, aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")

qqnorm(resid(model_5))
qqline(resid(model_5))

shapiro.test(resid(model_5))

library(car)
dwt(model_5)
```
#Spatial analysis
```{r}
us_hpi$residuals <- residuals(model_5)

us_hpi
# average residual by states
average_residual_by_state <- aggregate(residuals ~ State, data = us_hpi, FUN = mean)

average_residual_by_state
```

```{r}
#Download US state borders with polygonal geometry
states_sf2 <- states(cb = TRUE, resolution = "20m", year = 2024)

glimpse(states_sf2)
```

```
head(states_sf2)

#find the different data
setdiff(states_sf2$NAME, us_hpi$State)
setdiff(us_hpi$State, states_sf2$NAME)

#delete extra states
states_sf_clean2 <- states_sf2 %>%
  filter(!(NAME %in% c("Puerto Rico")))

#change "District of Columbia" to "District Of Columbia"
states_sf_clean2 <- states_sf_clean2 %>%
  mutate(NAME = ifelse(NAME == "District of Columbia", "District Of Columbia", NAME))

# Create the neighbor list
neighbor_list <- poly2nb(states_sf_clean2)

# Create a spatial weights matrix
weights <- nb2listw(neighbor_list, style = "W", zero.policy = TRUE)  # W for row-standardized
weights

# left join US unemployment_combine data with state_sf

us_hpi_nb <- left_join(average_residual_by_state, states_sf_clean2, by=c("State" = "NAME"))
head(us_hpi_nb)

# try Moran's test
moran.test(average_residual_by_state$residuals, weights)
moran.test(us_hpi[us_hpi$Year== 2020, ]$residuals, weights)
moran.test(us_hpi[us_hpi$Year== 2021, ]$residuals, weights)
moran.test(us_hpi[us_hpi$Year== 2022, ]$residuals, weights)
```