Team 042 Final Report

CSE6242 – Spring 2021

Project Title: The Economic Impact Caused by COVID-19 on People's Life
Team Member Names: Yinghai Yu, Lifei Xu, Ke Wang, Qichen Yu, Hang Yang

Introduction and Problem Definition

The goal of this project is to demonstrate how hard the pandemic hurts people's well-being, in terms of the loss of lives and job opportunities. Most current research works are focused on separate factors, such as the transmission condition of COVID 19 or the change of house market in different areas. This project explores approaches to evaluate the overall impact by considering the impact of the pandemic as well the economic environment together. Specifically, we would like to explore the impact of COVID R0[1] value (basic reproduction number) on house price, CPI and unemployment rate. For this proposal, all team members have contributed a similar amount of effort.

Literature Survey

After the outbreak of COVID-19, researchers have done tons of exploration on how to measure COVID-19 infection rate and human transmission [1]. A basic measurement R0 is defined to evaluate the average number of new infections created by an infectious individual in an entirely susceptible population [2]. Therefore, in this project, we will use R0 value to assess the infection rate of COVID-19 in the US.

The economic consequence of the pandemic on people's life has been evaluated in several research. **House market, household income and unemployment rates** are used to evaluate the impact of COVID-19 in this project. 1). In research for the house market, in [4], study demonstrates that the COVID-19 pandemic has disparately impacted low-resource areas and racial and ethnic minorities are experiencing the worst outcomes of this pandemic. In [3] [5], researcher uses a panel regression model and non-parametric model to evaluate the 'Fear of Missing Out' house purchase and Other COVID-induced fundamental changes in household behavior, and a two-step VAR model is used to forecast the effect of the COVID-19 outbreak in New York. The paper [6] shows that renters are facing difficulty to pay the rent since COVID-19, and even many households are still facing difficulties to pay their mortgage, especially if they have lost jobs or seen their incomes drop dramatically. Another paper [7] illustrates that the demand for homes close to city centers and neighborhoods with large populations decreased. 2). CPI (Consumer Price Index) is one of the indicators showing the impact on household spending. In[8] , researches show that the traditional CPI calculation methodology provided by BLS (Bureau of Labor Statistics) may not be appropriately applicable to 2020 due to COVID-19, because the spending patterns could be changed during the pandemic, like from to 'Recreation, Travel, and Entertainments' to 'Food and Beverages'. Researches proposed alternative methodologies in [9][10][11], which shows the implied CPI is higher than officially reported from BLS. 3). Household income and unemployment rates are usually considered together. In [14][15], research highlights various government policies that offset income losses from unemployment. According to [12], unemployment

[1] See Wiki, Ro : https://en.wikipedia.org/wiki/Basic_reproduction_number

rates increases in worldwide is associated with obvious increases in suicide both in high and low scenario. In [13], research shows that young workers, women, workers with low educational attainment, part-time workers, and racial and ethnic minorities have experienced the largest increases in unemployment rates in pandemic time, and the unemployment rate of each state in April is also greater than their highest unemployment rates during the Great Recession.

Proposed method

Innovation:

- Interactive choropleth corresponding to covid-19;
- Built more comprehensive machine learning model to predict economy;
- Visualize geographic and economy index synchronously via timeline filter ;
- Developed an approach to evaluate the aggregate effects of several factors and is building up a dashboard to visualize the model outputs with several different kinds of graphs, like line charts and maps.

Data Storage and Management

In this project, we analyzed data from various sources:

- The COVID-19 spread condition provided by CDC (~2.4GB)
- The average house price around US at county level (~2MB)
- The CPI data (61KB)
- The unemployment data (4KB)

SQLite was used to store the data. Please check the appendix for the data schema of all tables.
Benefits of using SQLite:

- Easy to manage data. We build a data pipeline to pull the up-to-date data from various sources and refresh the data tables. If we use csv files to store data, it would be quite complicated to track the versions of data files.
- Easy to combine data. Timestamp or location ids are used as primary keys in these tables, which makes it quite convenient to merge the tables.
- Easy to scale up. As time goes by, the volume of data would increase a lot. It is very easy to expand each table by the timestamp of data records.
- Easy to integrate with visualization. SQLite supports seamless data transformation with Pandas and D3 JS, which we use to build up models and visualize the model outputs.

Description of Methodologies

- Data cleaning strategy and practice (covid-19)

  This project uses covid-19 data from the CDC website. This dataset provides state-level daily covid-19 data, comprising features such as the number of deaths, number of positive cases, and total test cases, etc. The outbreak time of covid varies among different states in the US, therefore this dataset has many missing values. Given that we are going to perform data cleaning to deal with these missing values. For example, missing values in total test cases field should be filled with their immediately previous values. For another example, missing values in the number of daily deaths should be filled with 0. Not only missing values but also some unreasonable values should be cleaned as well, for instance, those extremely large values, and negative total number of daily deaths, etc. Apart from those, some columns that provide useless information will be removed directly from the dataset to improve our model's accuracy. In this project, we will use Python 3.7.x to perform data cleaning procedures. Two packages, pandas and csv, will be used frequently in this project to assist data cleaning.

In addition to covid-19 data, this project also includes unemployment, CPI, and house market data. Similar data cleaning procedures were applied to these datasets as well to make sure that all datasets used by the regression model are reasonable and clean.

In order to make the prediction model reasonable, time intervals of all datasets should be aligned. For example, the time interval of covid-19 data is one day. However, the time interval of house price data is one month. We therefore aggregated all datasets based on the data with the largest time interval, i.e. on a monthly interval, and we parsed the datasets with daily intervals to monthly based on the records from the last day of each month. After all these were done, the cleaned data was passed into our model to make predictions.

- Algorithm & models

  Our algorithm includes 3 steps. The first step is a state level linear regression model that presents the relationship between the R0 value and each of the areas we are interested in, i.e. house price, CPI, and unemployment rates, for 50 states and D.C. area. And our output for each individual state vs. interest area relationship would be shown as:

  $$Y = \alpha X + \beta + \epsilon$$

  Where X is the R0 value at certain period from a certain state, like R0 from NY during Jan'21; Y is the calculated output of one of the area we are interested in from the specific state during the mentioned period, like unemployment rates from NY during Jan'21 generated from the model; α is the model coefficient and β is the Y-intercept for one of our interested area; and ϵ is the error term that follows Gaussian Distribution. Because we are interested in totally three areas across 50 states and the D.C. areas, there are 153 linear regression models generated in the first step.

  The second step is building an Autoregressive integrated moving average (ARIMA) model used to predict a future R0 value for a certain state during a certain period, like R0 from NY during Feb'21. Through observation, we have realized that the R0 values are stationary, and therefore we can run the ARIMA model directly through the R0 values. Our output for X_t, representing R0 difference at time t, would be shown as:

  $$X_t = c + \varphi_1 * X_{t-1} + \varphi_2 * X_{t-2} \dots \varphi_n * X_{t-n} + \theta_1 * \epsilon_{t-1} + \theta_2 * \epsilon_{t-2} \dots \theta_m * \epsilon_{t-m} + \epsilon_t$$

  Where X_t is the calculated R0 difference at time t; X_{t-i} is the R0 difference from the last i period; c is the X_t intercept and φ_i is the autocorrelation for each of the i period (AR term); and ϵ_i is the error term from the last i period that follows Gaussian Distribution (MA term) and θ_i is the error term coefficient for each of the i period. Using this model, we can predict X_{t+1} by using known X_{t-j} and $j>=0$. There are 51 ARIMA models generated in the second steps.

  The third step is a combination of steps one and two, to predict the future performance of the area we determined to be the most relevant to R0 from a certain state. The model could be represented as:

  $$Y_{t+1} = \alpha X_{t+1} + \beta$$

  Where α and β are the coefficients and Y-intercept from step one, X_{t+1} is the predicted R0 value for a certain state during time t+1 from step two, and Y_{t+1} is the predicted future one month performance of the area from a selected state, like predicting unemployment rates from NY during Feb'21. There are totally 204 predictions made, for R0 itself and all 3 areas we are interested in, and across all 51 states and the D.C. area.

  All models were simulated through the Python scikit-learn and statsmodels packages.

- Visualization interfaces design

We built a dashboard to illustrate the relationship between economy and covid-19, and with a front-end interface that allows users to do interaction with our dataset.

- Interfaces design and goals
  - The choropleth map interface. This interface shows the spread of Covid-19 in the U.S, state level, from Jan. 2020 to Feb. 2021. Each state is filled with a color to present its Covid-19 R0 value in a specific month in 2020, the deeper the color means the worse the situation in this state, based on user's selection in the timeline control panel. Users can also click a specific state in the map to see the detailed information of the state.
  - The chart box interface. This interface includes three line charts to visualize the house market, household spending, household income and unemployment rate change during the Covid-19 period of each state, based on user selection in choropleth map, and it will also highlight the month point based on selection in the timeline control panel.
  - The timeline control panel. This panel is used for users to select specific month time to see the visualization effect in choropleth map and charts.
  - The prediction interfaces. This interface is used to test accuracy of our prediction model and visualize the prediction data for 2021. Users can click the "Prediction" link on the navigation bar to the page. In the choropleth map, users can select CPI, Unemployment rate and House Price from the drop-down list and see the color change in the map, and when users put the mouse over a state, a tooltip shows the predicted and actual data, also the prediction error. The smaller prediction error is, the higher accuracy of the prediction model, or the opposite.
- Tools for build the interface

  HTML, CSS, JavaScript are the main tools used in building the dashboard interfaces, D3.js is applied to develop visualization effects and create interaction functions for users.
- Final effect

  Please see the screenshots of the dashboard in Appendix. We also made a video demo to present the final visualization effect, please check it here.

Experiments:

- **Model selection and parameter tuning**

  We selected a linear regression model to represent the relationship between R0 and each of the following areas: house price, CPI and unemployment rates. The equation of the model is $Y = \alpha X + \beta + \epsilon$. Parameter α(model coefficient) needs tuning while ϵ is the error term that follows Gaussian Distribution , and β is the intercept.

  We selected an ARIMA model used to predict a future R0 value for a certain state during a certain period. The equation of the model is $X_t = c + \varphi_1 * X_{t-1} + \varphi_2 * X_{t-2} \dots \varphi_n * X_{t-n} + \theta_1 * \epsilon_{t-1} + \theta_2 * \epsilon_{t-2} \dots \theta_n * \epsilon_{t-n} + \epsilon_t$. The number of AR orders and MA orders are the most important parameters to tune. Parameter c is the X_t intercept, and φ_i is the autocorrelation, ϵ_i is the error term that follows Gaussian Distribution, and θ_i is the error term coefficient.

  The third model we selected is a combination of the first and the second model. The model, which could be represented as $Y_{t+1} = \alpha X_{t+1} + \beta$, is able to predict a future value for R0. As a result, tuning is not necessary for it.

Evaluation:

1. Evaluation of models

- Regression model - We have developed a linear regression model for each one of the interested areas vs. a state, and we found that the average R-squared for the house price linear regression models is 0.483 with average confidence level of 96.1%. For the unemployment rate models, the average R-square is 0.137 with an average confidence level of 84.7%. For the CPI models, the average R-square is 0.281 with an average confidence level of 42.5%. Therefore, we believe the house price predictions for each state are mostly closed to the actual.
- Autoregression model - We used Partial Autocorrelation plot (PACF) to determine the AR lags p and Autocorrelation plot (ACF) to determine the MA lags q, and we determined the ARIMA (2,0,2) model would be the best approach here. We then used Mean Absolute Error (MAE), Mean Percentage Error (MPE), and Root Mean Squared Error (RMSE) to evaluate the overall model performance.
- Prediction model - We compared our predictions with the actual Feb'21 datasets including monthly average house price, CPI, and unemployment rate across 50 states and D.C. area, and confirmed that our house price predictions are mostly close to the actual, with average prediction error of 4.82%. However, for CPI the average prediction error is as high as 37.1%, and for the unemployment rate the average prediction error is as high as 46.0%.

2. Evaluation of the Presentation

- Functionality testing.
  Manually checked all functions work, including all links in interfaces, timeline filter, mouse-over and floating-menu are functional, all the displayed data is correct.
- Usability testing.
  We conducted a survey among 20 participants to collect their feedback about experiences and satisfaction in using the dashboard. The survey questions and summary is added in the appendix.

3. Plan of activities

- Progress review meeting.
- Data cleaning collaboration meeting.
- UI design meeting.
- Product test and debug meeting.

4. Roles of team members

- Data engineer: Lifei Xu, Yinghai Yu
- Modeler: Ke Wang, Lifei Xu
- UI design and visualization: Qichen Yu, Ke Wang, Yinghai Yu, Hang Yang
- Project management: Yinghai Yu
- Project report: Qichen Yu, Hang Yang

Conclusions and Discussion

In the page "Visualization For 2020", we observed that the R0 value peaked across all states in Mar'20, but the state Unemployment Rates reached zeniths mostly in May'20, and the CPIs dropped to nadirs even until Jun'20. We believe that the lags in economics metrics were more driven by political policies instead of by the COVID infection itself. Since Jun'20, we have seen that both CPIs and Unemployment Rates are recovering along with the R0 values stabilized. Also, we observed that the house prices were increasing MoM and not being impacted crucially by neither the COVID infection nor the policy controls.

On the "Prediction for 2021" page, our models predicted the housing prices across different states with the highest average accuracy, and this could be driven by the straight MoM increase of housing price as mentioned in the last paragraph. The small prediction errors for housing prices are primarily from underestimation. For CPIs, our models predicted the CPIs with highest average accuracy for the states from the Northeast and West regions, but with higher errors when predicting the CPIs for the states from the Midwest and South regions. Overall, our models tended to underestimate CPIs. For the unemployment rate, our models tend to overestimate by 2% on average across all states. We believe that the data volatility caused our model predictions to be less accurate for CPIs and unemployment rates.

If we have more time for this project, we could develop our own algorithm to calculate R0 value. Also, we may develop the linear regression models by adding other variables, like monthly state level death rate and newly infectious rate as additional dependent variables.

References

1. Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak[J]. International journal of infectious diseases, 2020, 92: 214-217.
2. Bettencourt L M A, Ribeiro R M. Real time bayesian estimation of the epidemic potential of emerging infectious diseases[J]. PloS one, 2008, 3(5): e2185.
3. Zhao Y. US Housing Market during COVID-19: Aggregate and Distributional Evidence[J]. 2020.
4. Jones A, Grigsby-Toussaint D S. Housing stability and the residential context of the COVID-19 pandemic[J]. Cities & Health, 2020: 1-3.
5. Maria N, Zaid A, Catrin S, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review[J]. International Journal of Surgery, 2020, 78: 185-193.
6. Food D G E. Tracking the COVID-19 Recession's Effects on Food, Housing, and Employment Hardships[J].
7. Baker, S., Farrokhnia, R., Meyer, S., Pagel, M. and Yannelis, C., 2020. How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic. *SSRN Electronic Journal*,.
8. Cavallo, A., 2020. Inflation with Covid Consumption Baskets. *SSRN Electronic Journal*,. 9. Baker, S., Farrokhnia, R., Meyer, S., Pagel, M. and Yannelis, C., 2020. How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic. *SSRN Electronic Journal*,.
10. Liu S, Su Y. The impact of the Covid-19 pandemic on the demand for density: Evidence from the US housing market[J]. Available at SSRN 3661052, 2020.
11. Federal Reserve Bank of St. Louis, and Michael McCracken. "How COVID-19 May Be Affecting Inflation." STL FED, 2 Feb. 2021, www.stlouisfed.org/on-the-economy/2021/february/covid19- affecting-inflation.
12. Falk G, Carter J A, Nicchitta I A, et al. Unemployment rates during the COVID-19 pandemic: In brief[J]. Congr Res Serv, 2020: 2-16.
13. "COVID-19 Pandemic's Impact on Household Employment and Income." Congressional Research Service, 9 Nov. 2020, https://crsreports.congress.gov/product/pdf/IN/IN11457
14. "Tracking the COVID-19 Recession's Effects on Food, Housing, and Employment Hardships." Center on Budget and Policy Priorities, 10 Mar. 2020, www.cbpp.org/research/poverty-and inequality/tracking-the-covid-19-recessions-effects-on-food-housing-and.

Appendix

Table Schemas:

Table 1: Covid-19 data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
Rt_mean:	float 32

Table 2: Housing market data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
Re_price:	float 32

Table 3: CPI data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
CPI_12mo:	float 32

Table 4: Unemployment data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
Ur_rate:	float 32

Table 5: Combined data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
RegionID	integer
Month	string
Rt_mean:	float 32
Re_price:	float 32

Table 6: Prediction data

Column names	Data type
Date	datetime (primary key)
State	string
State_cd	integer (primary key)
RegionID	integer
Month	string
Rt_mean:	float 32
Re_price:	float 32

CPI_12mo:	float 32
Ur_rate:	float 32

CPI_12mo:	float 32
Ur_rate:	float 32
Rt_predict:	float 32
Re_price_predic:	float 32
CPI_12mo_predict:	float 32
Ur_rate_predict:	float 32

Visualization Dashboard

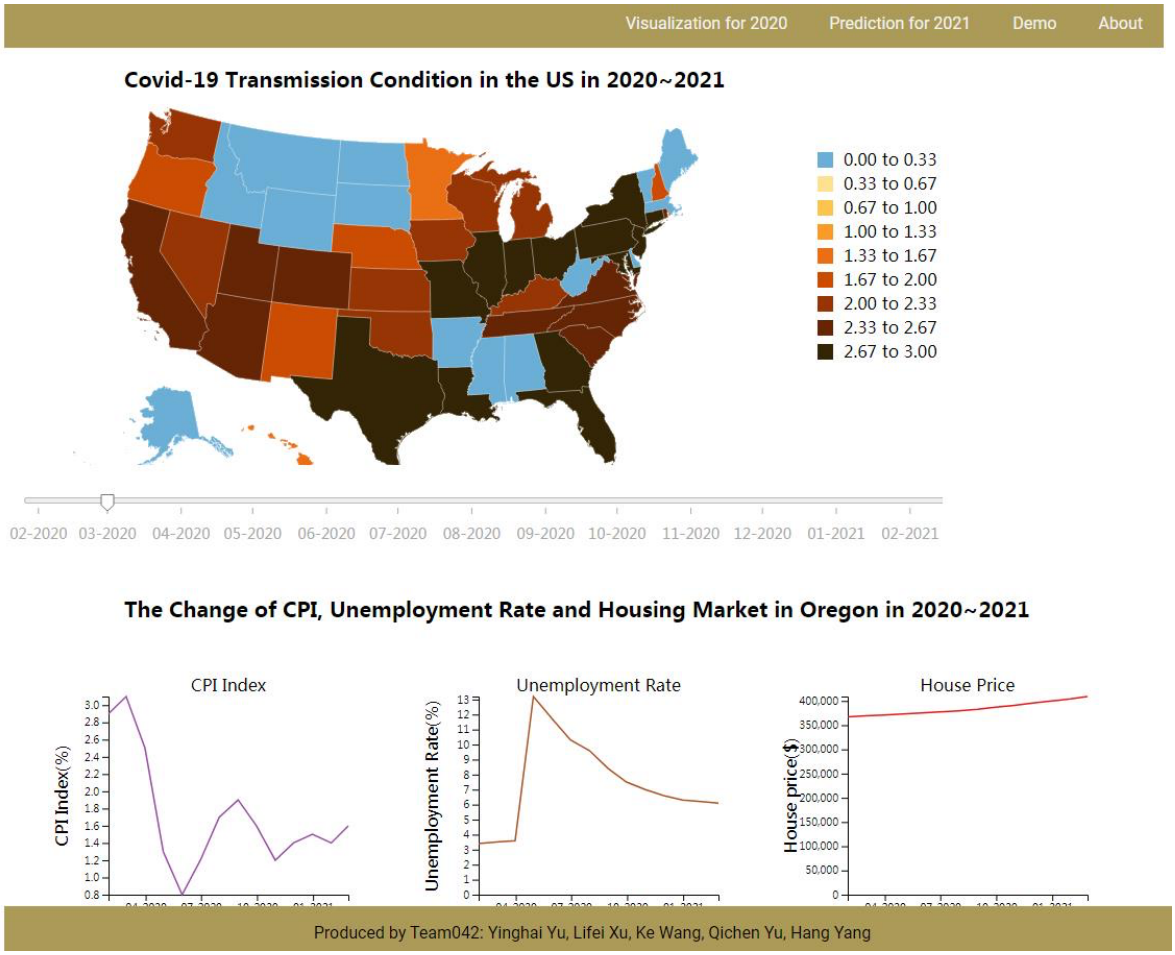

Fig 1. Visualization for 2020

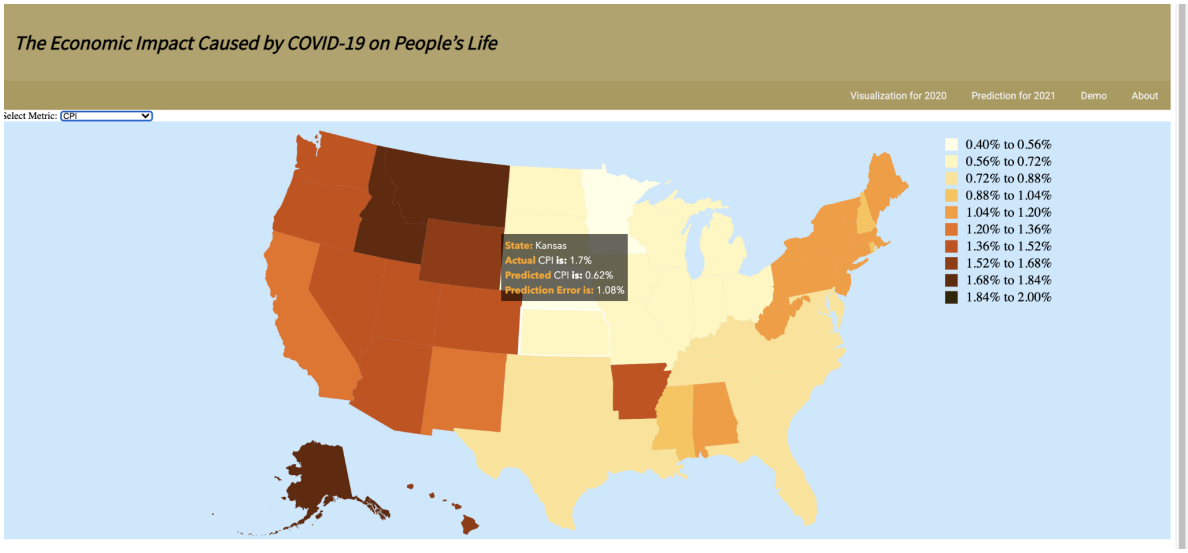

Fig 2. Visualization for 2021

Survey Results

Question	**Choice (1-5)**	**Average Score**
1. Do you think the dashboard is helpful for you to understand the economic impact by COVID-19?	1 for not helpful at all, 5 for extremely helpful	4.65
2. Overall, how easy is it to use the dashboard to get information?	1 for extremely easy, 5 for extremely hard	1.07
3.How easy is it to use selection filters in the dashboard?	1 for extremely easy, 5 for extremely hard	1.32
4.How easy is it to understand the Rt value visualized in the map? (i.e. the gradient scheme in the map)	1 for extremely easy, 5 for extremely hard	2.10
5.How easy is it to see the relation between line charts and the map?	1 for extremely easy, 5 for extremely hard	1.41
6.Overall, from 1 to 5, how do you evaluate the whole COVID-19 impact visualization project?	1-5. 1 for terrible, 5 for excellent	4.89