Group Project Check-in

submitted by Radhika Kaul, Odiche Nwabuikwu, & Ruochen Wang

Question of Interest

We will develop different models to predict the number of unemployment claims filed in each US state in the second week of May (when the project is due). Our primary data source for the outcome variable is the Unemployment Insurance Weekly Claims data from the U.S. Department of Labor (DOL)'s Employment & Training Administration.

Project Progress

We have decided on the following variables for our models:

Outcome Variables: either New Initial Claims filed per week OR percentage of the labor force filing for unemployment claims per week. New Initial Claims is one of the most sensitive and most frequently used official statistics in analyzing unemployment in the labor market (see Aaron Soujourner and Paul Goldsmith Pinkham).

Predictors *(all variables are at the state level)*

Structural predictors include:

- Region
- Population
- Working Population
- Per Capita Personal Income
- Percent Population w/ a High School Degree
- Top Industry in terms of GDP Contribution
- Democratic or Republican Governor

We are considering standardizing/normalizing/logging the populations variables due to concerns of high correlations between these variables and the outcome variable.

Real-Time Predictors include:

- Number of COVID Cases
- Number of Days Since the issurance of Stay-at-Home Order
- Number of Days Before the Stay-at-Home Order ends
- Peak of Claims Predicted (we haven't found a specific data source but will keep looking)
- Whether there has been a protest/protests in the state

Predictive Models: We will have three predictive models based off of data from when the first case was reported in the US, the first death was reported in the US, and the modified date when the first death related to COVID-19 was reported (see https://www.cnn.com/2020/04/22/us/california-deaths-earliest-in-us/index.html)

Algorithms: KNN/CART/Random Forest. We will test each algorithm and get the error rates and pick the best one.

We propose a **Random Forest** model given that we have both categorical and continuous inputs and that we have correlated predictors. Also given the fluctuation in the weekly claims data we are going to explore, this algorithm will make more accurate out-of-sample predictions as compared to KNN and CART.

Error Metric: RMSE (OOB Error for Random Forest)

Data Cleaning

```
library(tidyverse)
library(readxl)
```

State GDP by Industry

```
industry <-
  read.csv("data/raw/gdp_bystate_byindustry_2019.csv")

# data cleaning
industry <- industry %>%
  filter(str_detect(string = Description,
                    pattern = "    ")) %>%
  filter(!str_detect(string = Description,
                     pattern = "     ")) %>% # removing sub-industries
  filter(GeoName != "United States") %>% # removing federal data
  filter(X2019 != "(NA)" & X2019 != "(D)") %>% # removing non-numeric values from GDP
  rename(gdp = X2019, state_name = GeoName) # renaming GDP and state variables

# reshape data
industry <- industry %>%
  pivot_wider(names_from = Description, state_name, values_from = gdp) %>%
  filter(!state_name %in% c("New England", "Mideast", "Great Lakes", "Plains", "Southeast", "Southwest"
```

Unemployment Claims

```
# load data
claims <- read_excel("data/raw/unemployment_bystate_weekly.xls")

claims <- claims %>%
  rename(state_name = State)
```

Personal Income by State

```
# load data
income <- read_csv("data/raw/percapitapersonalincome_bystate_2019.csv")

income <- income %>%
  filter(as.numeric(GeoFips) >= 1000 & as.numeric(GeoFips) <= 56000) %>% # remove fed/regional data
  rename(income = "2019", state_name = GeoName) %>%
  select(-GeoFips)
```

Population by State

```
# load data
pop <- read_csv("data/raw/popest_bystate_2019.csv")

pop <- pop %>%
  filter(!NAME %in% c("United States", "Puerto Rico Commonwealth")) %>%
  select(-SUMLEV, -REGION, -DIVISION) %>%
  rename(fips = STATE, state_name = NAME, pop_2019 = POPESTIMATE2019, pop_over18 = POPEST18PLUS2019, pc
```

Employment by Occupation by State

```
# load data
occ <- read_excel("data/raw/employment_byoccupation_bystate_may2019.xlsx")


occ <- occ %>%
  filter(o_group == "major") %>%
  select(area, area_title, occ_code, occ_title, tot_emp, jobs_1000, h_mean, h_median) %>%
  rename(fips = area, state_name = area_title, mean_hourly_wage = h_mean, med_hourly_wage = h_median )
```

NOTE: This dataset will be further formatted and merged with the others after further discussion.

Census Regions

```
# load data
regions <- read_excel("data/raw/censusregion_bystate.xlsx")

regions <- regions %>%
  select(State, Region) %>%
  rename(state_name = State, region_name = Region)
```

Merge Datasets

```
clean_data <- claims %>%
  left_join(income) %>%
  left_join(industry) %>%
  left_join(pop) %>%
  left_join(regions) %>%
  write_csv("data/clean/merged.csv")
```

(Immediate) Next Steps

We are still in the process of cleaning and merging datasets - finals week has been insane! We will meet with the instructors to finalize the variables we are to include in our models as well as the algorithms, and finish up building our final dataset.