Group Project Check-in

submitted by Radhika Kaul, Odiche Chidinma Nwabuikwu, & Ruochen Wang

Question of Interest: Predicting the number of unemployment claims filed per U.S. state in in the second week of May (When project is due) using data from various sources but the major source of what we want to predict comes from the Unemployment Insurance Weekly Claims data from the U.S. Department of Labor (DOL)'s Employment & Training Administration.

Project Progress: In the wake of the COVID-19 pandemic and responses set up by different states to this crisis, we have finally decided on pursuing the following predicted model for our analysis:

What we are trying to predict: Either New Initial Claims filed per week OR percentage of the labor force filing for unemployment claims per week. Initial claims is one of the most-sensitive, high-frequency official statistics used to detect changes in the labor market (Source: Aaron Soujourner and Paul Goldsmith Pinkham UI trends data.

Predictors: Structural predictors include population, per capita income,top industry contributing to GDP, working population, region, and percentage population with a highschool degree. Some concerns of using these variables is high correlation in population variables and the need to pre-process (standardize/normalize/log) them. Real-time predictors include Number of COVID cases per state, number of days since the stay-at-home orders started, peaks predicted for each state (although we haven't found a particular data source), Democratic or Republic governor, number of days till stay-at-home orders will end, and days since the protests started per state.

- **Error Metric**: RMSE (we'll have to explain why?)
- **Pre-processing data**:
- **Predictive model**: CART/RANDOM FOREST.. explain and tell them we'll have three predictive models on the basis of the following initial information: First case reported, First death reported, New info on first case reported

Data Cleaning

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

Data Cleaning

State GDP by Industry

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

# data cleaning
```

```r
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

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

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

Personal Income by State

```r
# 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

```r
# 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

```r
# 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

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

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

Merge Datasets

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