Untitled

Name: CHUYANG CHEN

data cleaning

```
#install.packages("blscrapeR")
library(blscrapeR)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------- tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# get the lastest nemployment data.
df <- get_bls_county()
WIunemployment = df %>% filter(fips_state == 55)

# get the unemployment data from November 2019 (the data from previous month)
df2 = get_bls_county("November 2019")
WIunemployment_Nov = df2 %>% filter(fips_state == 55)

# rename the column names to aovid confusion
colnames(WIunemployment_Nov)[colnames(WIunemployment_Nov) == "unemployed"] = "unemployed_Nov"
colnames(WIunemployment_Nov)[colnames(WIunemployment_Nov) == "unemployed_rate"] = "unemployed_rate_Nov"

# get the bridges data in WI.
bridges = read_csv("https://www.fhwa.dot.gov/bridge/nbi/2018/delimited/WI18.txt")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   STRUCTURE_NUMBER_008 = col_character(),
##   ROUTE_NUMBER_005D = col_character(),
##   HIGHWAY_DISTRICT_002 = col_character(),
##   COUNTY_CODE_003 = col_character(),
##   FEATURES_DESC_006A = col_character(),
##   CRITICAL_FACILITY_006B = col_logical(),
##   FACILITY_CARRIED_007 = col_character(),
##   LOCATION_009 = col_character(),
##   LRS_INV_ROUTE_013A = col_character(),
```

```
##   LAT_016 = col_character(),
##   LONG_017 = col_character(),
##   MAINTENANCE_021 = col_character(),
##   OWNER_022 = col_character(),
##   FUNCTIONAL_CLASS_026 = col_character(),
##   DESIGN_LOAD_031 = col_character(),
##   RAILINGS_036A = col_character(),
##   TRANSITIONS_036B = col_character(),
##   APPR_RAIL_036C = col_character(),
##   APPR_RAIL_END_036D = col_character(),
##   NAVIGATION_038 = col_character()
##   # ... with 41 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 3 parsing failures.
##   row                    col               expected actual
##  5739 OTHR_STATE_STRUC_NO_099 no trailing characters   B010 'https://www.fhwa.dot.gov/bridge/nbi/2018
## 11175 OPR_RATING_METH_063     a double                 F    'https://www.fhwa.dot.gov/bridge/nbi/2018
## 11175 INV_RATING_METH_065     a double                 F    'https://www.fhwa.dot.gov/bridge/nbi/2018
```

```
# selecting varibles which are relevent to the unemployment number / rate.
table2 = bridges %>%
  group_by(COUNTY_CODE_003) %>%
  left_join(WIunemployment,by = c("COUNTY_CODE_003" = "fips_county")) %>%
  select(area_title,unemployed,unemployed_rate,APPR_WIDTH_MT_032,PERCENT_ADT_TRUCK_109,APPR_ROAD_EVAL_07
         FUNCTIONAL_CLASS_026)
```

```
## Adding missing grouping variables: `COUNTY_CODE_003`
```

```
# Add the unemployed data from November 2019 into table2
table3 = table2 %>%
  group_by(COUNTY_CODE_003) %>%
  left_join(WIunemployment_Nov,by = c("COUNTY_CODE_003" = "fips_county"))
```

bulding linear model with relevent variables

This linear model has 4 independent variables.

Continous variables: - APPR_WIDTH_MT_03 represents the width of usable roadway approaching the bridge. - PERCENT_ADT_TRUCK_109 represents the percentage of average daily traffic that is trunk. - ADT_029 represents the overall average daily traffic

Dummy variables (categorical variables): - FUNCTIONAL_CLASS_026 represents the location and function of the bridge. 01-09 indicates the bridge is in a rural area, and 11 - 19 means the bridge is in an urban area.

```
# Build a linear model to predict unempolyment rate
lm_unemployed_rate = lm(data = table2, unemployed_rate ~ APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 + AD
summary(lm_unemployed_rate)
```

```
##
## Call:
## lm(formula = unemployed_rate ~ APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 +
##     ADT_029 + as.character(FUNCTIONAL_CLASS_026), data = table2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2051 -0.6476 -0.1664  0.4666  3.3527
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           3.212e+00  5.301e-02  60.590  < 2e-16 ***
## APPR_WIDTH_MT_032                    -1.573e-02  2.507e-03  -6.276 3.56e-10 ***
## PERCENT_ADT_TRUCK_109                 1.641e-02  1.588e-03  10.331  < 2e-16 ***
## ADT_029                              -4.415e-06  1.030e-06  -4.285 1.84e-05 ***
## as.character(FUNCTIONAL_CLASS_026)02  4.639e-01  4.515e-02  10.275  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)06  4.911e-01  4.618e-02  10.636  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)07  4.883e-01  4.765e-02  10.248  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)08  3.676e-01  5.425e-02   6.777 1.28e-11 ***
## as.character(FUNCTIONAL_CLASS_026)09  4.705e-01  4.566e-02  10.305  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)11  3.248e-01  5.069e-02   6.407 1.53e-10 ***
## as.character(FUNCTIONAL_CLASS_026)12  4.649e-02  6.521e-02   0.713   0.4759
## as.character(FUNCTIONAL_CLASS_026)14  2.002e-01  5.118e-02   3.912 9.19e-05 ***
## as.character(FUNCTIONAL_CLASS_026)16  2.752e-01  5.281e-02   5.212 1.89e-07 ***
## as.character(FUNCTIONAL_CLASS_026)17  7.783e-02  7.213e-02   1.079   0.2806
## as.character(FUNCTIONAL_CLASS_026)19  9.853e-02  5.528e-02   1.783   0.0747 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8437 on 14196 degrees of freedom
##   (64 observations deleted due to missingness)
## Multiple R-squared:  0.05906,    Adjusted R-squared:  0.05813
## F-statistic: 63.64 on 14 and 14196 DF,  p-value: < 2.2e-16
```

The result of this linear model shows that larger width of usable roadway approaching the bridge is associated with lower unemployment rate. As the percentage of trunk traffic increases, the unemployment rate will also rise. The average daily traffic tells a different story: more traffic in gerenal leads to lower unemployment rate. The location of bridges shows that rural area is associated with higher unemployment rate, which makes sense. Even though the coefficients of urban area are also postive, we cannot say urabn area is associated with higher unemployment rate since their p-values are very large, which means that the coefficients do not have significant different with zero.

```r
# Build a linear model to predict unempolyed number
lm_unemployed = lm(data = table2, unemployed ~ APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 + ADT_029 + as
summary(lm_unemployed)
```

```
##
## Call:
## lm(formula = unemployed ~ APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 +
##     ADT_029 + as.character(FUNCTIONAL_CLASS_026), data = table2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11980.6   -953.0   -397.9    375.9  15534.7
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.712e+03  1.882e+02   9.093  < 2e-16 ***
## APPR_WIDTH_MT_032                     4.367e+01  8.902e+00   4.905 9.44e-07 ***
## PERCENT_ADT_TRUCK_109                -9.187e+01  5.640e+00 -16.290  < 2e-16 ***
```

```
## ADT_029                              6.869e-02  3.658e-03  18.777  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)02 -6.033e+02  1.603e+02  -3.763 0.000169 ***
## as.character(FUNCTIONAL_CLASS_026)06 -6.027e+02  1.640e+02  -3.676 0.000238 ***
## as.character(FUNCTIONAL_CLASS_026)07 -7.784e+02  1.692e+02  -4.601 4.25e-06 ***
## as.character(FUNCTIONAL_CLASS_026)08 -6.510e+02  1.926e+02  -3.379 0.000729 ***
## as.character(FUNCTIONAL_CLASS_026)09 -6.697e+02  1.621e+02  -4.131 3.64e-05 ***
## as.character(FUNCTIONAL_CLASS_026)11  5.616e+03  1.800e+02  31.201  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)12  1.205e+03  2.316e+02   5.206 1.96e-07 ***
## as.character(FUNCTIONAL_CLASS_026)14  2.545e+03  1.817e+02  14.004  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)16  4.958e+03  1.875e+02  26.442  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)17  2.462e+03  2.561e+02   9.612  < 2e-16 ***
## as.character(FUNCTIONAL_CLASS_026)19  3.948e+03  1.963e+02  20.111  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2996 on 14196 degrees of freedom
##   (64 observations deleted due to missingness)
## Multiple R-squared:  0.4251, Adjusted R-squared:  0.4246
## F-statistic: 749.9 on 14 and 14196 DF,  p-value: < 2.2e-16
```

Using the unemployed number and rate from the previous month as additional predictors to the linear model

```
# linear model predicting unempolyment rate
lm_unemployed_rate2 = lm(data = table3, unemployed_rate ~ unemployed_rate_Nov + unemployed_Nov + APPR_WI
summary(lm_unemployed_rate2)
```

```
##
## Call:
## lm(formula = unemployed_rate ~ unemployed_rate_Nov + unemployed_Nov +
##     APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 + ADT_029 + as.character(FUNCTIONAL_CLASS_026),
##     data = table3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71299 -0.22962 -0.04881  0.18638  1.04813
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -1.683e-01  2.712e-02  -6.208 5.50e-10 ***
## unemployed_rate_Nov                   1.277e+00  5.227e-03 244.256  < 2e-16 ***
## unemployed_Nov                       -5.780e-05  1.002e-06 -57.674  < 2e-16 ***
## APPR_WIDTH_MT_032                    -5.946e-03  1.099e-03  -5.408 6.49e-08 ***
## PERCENT_ADT_TRUCK_109                -1.075e-03  7.051e-04  -1.525  0.12731
## ADT_029                              -1.417e-07  4.570e-07  -0.310  0.75653
## as.character(FUNCTIONAL_CLASS_026)02  9.495e-02  1.984e-02   4.785 1.72e-06 ***
## as.character(FUNCTIONAL_CLASS_026)06  1.371e-01  2.028e-02   6.760 1.44e-11 ***
## as.character(FUNCTIONAL_CLASS_026)07  1.130e-01  2.094e-02   5.397 6.87e-08 ***
## as.character(FUNCTIONAL_CLASS_026)08  7.595e-02  2.380e-02   3.191  0.00142 **
## as.character(FUNCTIONAL_CLASS_026)09  1.120e-01  2.006e-02   5.585 2.38e-08 ***
## as.character(FUNCTIONAL_CLASS_026)11  1.177e-02  2.301e-02   0.512  0.60880
## as.character(FUNCTIONAL_CLASS_026)12  4.946e-02  2.859e-02   1.730  0.08362 .
```

```
## as.character(FUNCTIONAL_CLASS_026)14 -5.091e-03  2.259e-02  -0.225  0.82171
## as.character(FUNCTIONAL_CLASS_026)16  4.798e-03  2.374e-02   0.202  0.83982
## as.character(FUNCTIONAL_CLASS_026)17 -1.045e-02  3.170e-02  -0.330  0.74178
## as.character(FUNCTIONAL_CLASS_026)19 -2.575e-02  2.456e-02  -1.048  0.29453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3695 on 14194 degrees of freedom
##   (64 observations deleted due to missingness)
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8193
## F-statistic:  4028 on 16 and 14194 DF,  p-value: < 2.2e-16
```

The result shows that a higher unemployment rate of previous month are likely to lead to a higher unemployment rate of future month. For some reason, the predictor unemployed number from previous month tells a different story, and I am not sure why this is happening. A larger width of usable roadway approaching the bridge is associated with even lower unemployment rate in this model compared to the model without unemployment data from previous month. A larger percetage of trunk traffic is associated with lower unemployment rate, which is contradict to the first model. Higher average daily traffic in gereanl leads to lower unemployment rate, which totally makes sense, but unfortunately the p-value here is way too large which means we cannot get this conclusion. Rural areas are associated with higher unemployment rate, and the p-value for urban areas are too large so being in urban area might not be an effective predictor.

```r
# linear model predicting unempolyed number
lm_unemployed2 = lm(data = table3, unemployed ~ unemployed_rate_Nov + unemployed_Nov + APPR_WIDTH_MT_03
summary(lm_unemployed_rate2)
```

```
##
## Call:
## lm(formula = unemployed_rate ~ unemployed_rate_Nov + unemployed_Nov +
##     APPR_WIDTH_MT_032 + PERCENT_ADT_TRUCK_109 + ADT_029 + as.character(FUNCTIONAL_CLASS_026),
##     data = table3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71299 -0.22962 -0.04881  0.18638  1.04813
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.683e-01  2.712e-02  -6.208 5.50e-10 ***
## unemployed_rate_Nov                  1.277e+00  5.227e-03 244.256  < 2e-16 ***
## unemployed_Nov                      -5.780e-05  1.002e-06 -57.674  < 2e-16 ***
## APPR_WIDTH_MT_032                   -5.946e-03  1.099e-03  -5.408 6.49e-08 ***
## PERCENT_ADT_TRUCK_109               -1.075e-03  7.051e-04  -1.525  0.12731
## ADT_029                             -1.417e-07  4.570e-07  -0.310  0.75653
## as.character(FUNCTIONAL_CLASS_026)02  9.495e-02  1.984e-02   4.785 1.72e-06 ***
## as.character(FUNCTIONAL_CLASS_026)06  1.371e-01  2.028e-02   6.760 1.44e-11 ***
## as.character(FUNCTIONAL_CLASS_026)07  1.130e-01  2.094e-02   5.397 6.87e-08 ***
## as.character(FUNCTIONAL_CLASS_026)08  7.595e-02  2.380e-02   3.191  0.00142 **
## as.character(FUNCTIONAL_CLASS_026)09  1.120e-01  2.006e-02   5.585 2.38e-08 ***
## as.character(FUNCTIONAL_CLASS_026)11  1.177e-02  2.301e-02   0.512  0.60880
## as.character(FUNCTIONAL_CLASS_026)12  4.946e-02  2.859e-02   1.730  0.08362 .
## as.character(FUNCTIONAL_CLASS_026)14 -5.091e-03  2.259e-02  -0.225  0.82171
## as.character(FUNCTIONAL_CLASS_026)16  4.798e-03  2.374e-02   0.202  0.83982
## as.character(FUNCTIONAL_CLASS_026)17 -1.045e-02  3.170e-02  -0.330  0.74178
## as.character(FUNCTIONAL_CLASS_026)19 -2.575e-02  2.456e-02  -1.048  0.29453
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3695 on 14194 degrees of freedom
##   (64 observations deleted due to missingness)
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8193
## F-statistic:  4028 on 16 and 14194 DF,  p-value: < 2.2e-16
```