kable table..

2024-04-21

code so plots work

data cleaning

data combination from jorge

```
"C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa/not combined csv files"
```

```
## [1] "C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa/not combined csv files"
```

```
getwd()
```

```
## [1] "C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa"
```

```
setwd("C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa/not combined csv files")

data1 <- read.csv("Criminal_Offenses_On_campus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x,"_all_campus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_all_campus, unique_id = unique_id_all_campus)

data2 <- read.csv("Criminal_Offenses_On_campus_Student_Housing_Facilities.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x,"_student_housing"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_student_housing, unique_id = unique_id_student_housing)

data3 <- read.csv("Criminal_Offenses_Noncampus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_crim_offense_noncampus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_crim_offense_noncampus, unique_id = unique_id_crim_offense_noncampus

data4 <- read.csv("Criminal_Offenses_Public_property.csv") |>
   mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_crim_offense_public"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_crim_offense_public, unique_id = unique_id_crim_offense_public)

data5 <- read.csv("Arrests_On_campus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_arrests_campus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_arrests_campus, unique_id = unique_id_arrests_campus)
```

```r
data6 <- read.csv("Arrests_On_campus_Student_Housing_Facilities.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_arrests_stuhousing"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_arrests_stuhousing, unique_id = unique_id_arrests_stuhousing)

data7 <- read.csv("Arrests_Noncampus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_arrests_noncampus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_arrests_noncampus, unique_id = unique_id_arrests_noncampus)

data8 <- read.csv("Arrests_Public_Property.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_arrests_public"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_arrests_public, unique_id = unique_id_arrests_public)

data9 <- read.csv("Disciplinary_Actions_On_campus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_disciplinary_campus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_disciplinary_campus, unique_id = unique_id_disciplinary_campus)

setwd("C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa")

data10 <- read.csv("Disciplinary_Actions_Student_Housing_Facilities.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_disciplinary_housing"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_disciplinary_housing, unique_id = unique_id_disciplinary_housing)

setwd("C:/Users/paige/OneDrive/Documents/STAT 472/Team-Koopa/not combined csv files")

data11 <- read.csv("Disciplinary_Actions_Noncampus.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_disciplinary_noncampus"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_disciplinary_noncampus, unique_id = unique_id_disciplinary_noncampus

data12 <- read.csv("Disciplinary_Actions_Public_Property.csv") |>
  mutate(unique_id = paste0(OPEID, "_", Campus.ID)) |>
  rename_with(~ paste0(.x, "_disciplinary_public"), recycle0 = TRUE) |>
  rename(Survey.year = Survey.year_disciplinary_public, unique_id = unique_id_disciplinary_public)

# This is our datasets being joined into one
dataset <- data1 |> left_join(data2) |>
  left_join(data3) |>
  left_join(data4) |>
  left_join(data5) |>
  left_join(data6) |>
  left_join(data7) |>
  left_join(data8) |>
  left_join(data9) |>
  left_join(data10) |>
  left_join(data11) |>
  left_join(data12)
```

```
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
## Joining with ‘by = join_by(Survey.year, unique_id)‘
```

remove useless cols

removing NA values, removing useless columns

```r
#remove NAs
dataset[is.na(dataset)] <- 0

#remove repeated columns (like unitid repeating for each xcel file)
#(3/4/24) just fixed some problems w this

cols_to_remove <- c("Unitid_student_housing", "Institution.name_student_housing", "OPEID_student_housin


## had to change this dataset name before removing the campses ##

cleaned <- dataset[, !names(dataset) %in% cols_to_remove]
```

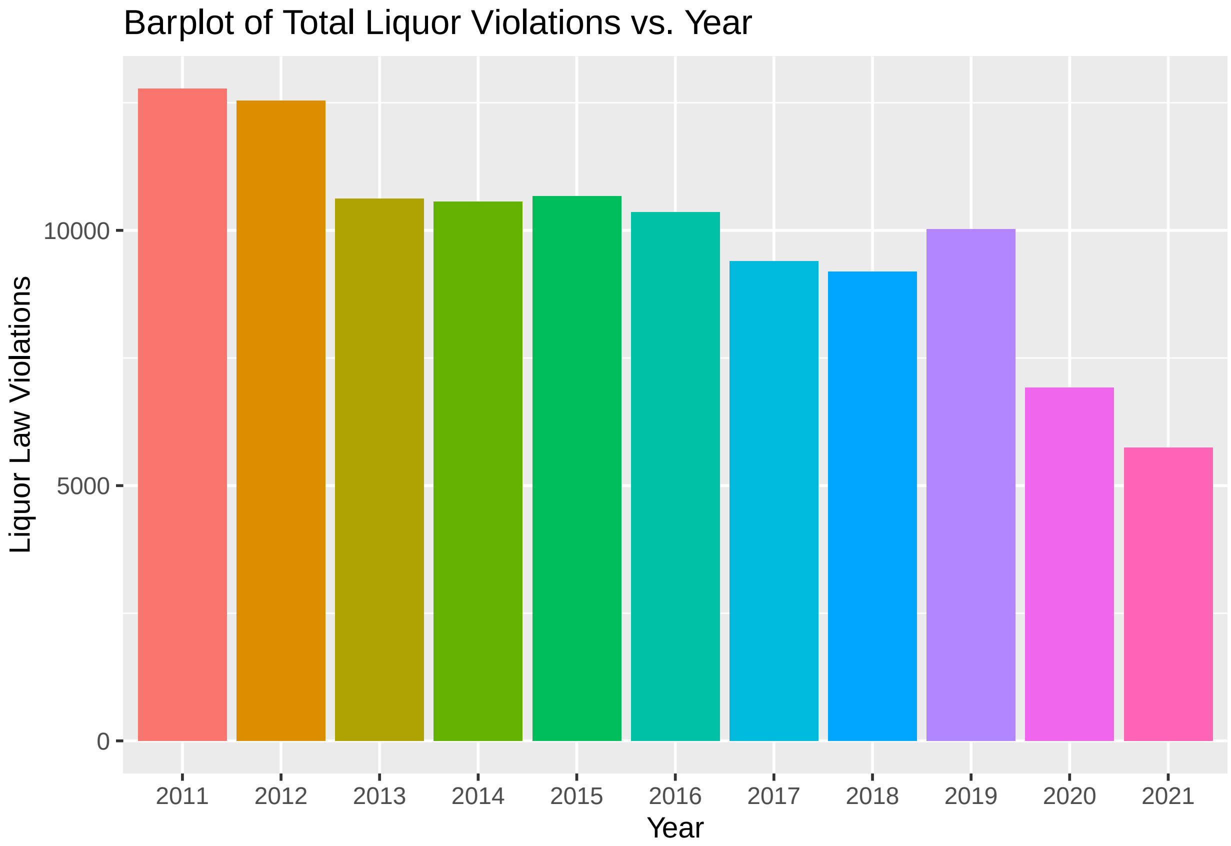
remove campuses

Removes campuses outside of Colorado.

```r
### note!! i had to change the full dataset name from cleaned_data to cleaned (see last line in the chu

to_remove1 <- c("Jacksonville", "San Diego", "Memphis", "Dunnam", "Ft. Drum", "San Luis Obispo", "Syracu

#check vector length
#length(to_remove1)

matches <- unique(grep(paste(to_remove1,collapse="|"),
                        cleaned$Campus.Name_all_campus, value=TRUE))
cleaned_1 <- cleaned |> filter(!Campus.Name_all_campus %in% matches)

to_remove2 <- c("Albuquerque", "Wiesbaden", "Beale", "Gateway", "Ocala Metropolitan Campus", "Baton Rou

#length(to_remove2)

matches <- unique(grep(paste(to_remove2,collapse="|"),
                        cleaned_1$Campus.Name_all_campus, value=TRUE))
cleaned_2 <- cleaned_1 |> filter(!Campus.Name_all_campus %in% matches)

to_remove3 <- c("Webster University St. Louis-Main Campus", "Space Coast", "Fort Worth", "San Francisco
#length(to_remove3)

matches <- unique(grep(paste(to_remove3,collapse="|"),
                        cleaned_2$Campus.Name_all_campus, value=TRUE))
cleaned_data <- cleaned_2 |> filter(!Campus.Name_all_campus %in% matches)


# take a look
#head(cleaned_data)

#new column combining liquor law violations across disciplinary, arrests and location (public, stuhousi
cleaned_data$all_liquor_violations <- cleaned_data$Liquor.law.violations_arrests_campus + cleaned_data$
```

barplot

```
year_factor <- as.factor(cleaned_data$Survey.year)

ggplot(cleaned_data, aes(x = year_factor, y = all_liquor_violations, fill = year_factor)) +
  geom_bar(stat = "identity") +
  labs(x = "Year", y = "Liquor Law Violations", fill = "Year") +
  ggtitle("Barplot of Total Liquor Violations vs. Year") +
  theme(legend.position = "none")
```



split data

```
set.seed(4242)

## split cleaned data into 25/75
smp_size <- floor(0.75 * nrow(cleaned_data))

train_split <- sample(seq_len(nrow(cleaned_data)), size = smp_size)

# create train = 75% and test = 25% set
train <- cleaned_data[train_split,] |> as_tibble() |> mutate(train = TRUE)
test <- cleaned_data[-train_split,] |> as_tibble() |> mutate(train = FALSE)
```

lasso coef table

```
set.seed(4242)

#for lasso
#install.packages("glmnet")
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```
train_num <- dplyr::select_if(train, is.numeric)

#specify y
y <- train_num$all_liquor_violations

#train$Liquor

exclude_columns <- c("Unitid_all_campus", "OPEID_all_campus",
                     "Campus.ID_all_campus", "all_liquor_violations",
                     "Liquor.law.violations_arrests_campus",
                     "Liquor.law.violations_arrests_public",
                     "Liquor.law.violations_arrests_noncampus",
                     "Liquor.law.violations_arrests_stuhousing",
                     "Liquor.law.violations_disciplinary_campus",
                     "Liquor.law.violations_disciplinary_noncampus",
                    "Liquor.law.violations_disciplinary_public",
                    "Liquor.law.violations_disciplinary_housing",
                    "new_column")

train_finalset <- train_num[, !names(train_num) %in% exclude_columns]

#specify x
x <- data.matrix(train_finalset)


# k fold cv for lambda
cv_model <- cv.glmnet(x,y,alpha = 1)
best_lambda <- cv_model$lambda.min
#best_lambda

#plot(cv_model)

#find optimal lasso model
best_lasso <- glmnet(x, y, alpha = 1, lambda = best_lambda)

#coefficients from lasso model
lasso_coef <- coef(best_lasso)

#lasso_coef

#make coefficients matrix
lc_mat <- as.matrix(lasso_coef)

#make coefficients dataframe
lc_df <- as.data.frame(lc_mat)

#filter out coefficients that are 0
rows_to_keep <- apply(lc_mat, 1, function(row) any(row > 0, row < 0))

lc_df_filtered <- lc_df[rows_to_keep,]

#lc_df_filtered

#remove intercept
lc_df_clean <- lc_df_filtered[-1]

#lc_df_clean

lc_table_df <- data.frame(
  Variable = c("Institution Size", "Sex Offenses (all campus)", "Arson (all campus)", "Rape (student hou
  Coefficients = lc_df_clean)

#table of lasso coefficients
knitr::kable(lc_table_df, caption = "LASSO Coefficients", digits = 3)
```

Table 1: LASSO Coefficients

Variable	Coefficients
Institution Size	0.001
Sex Offenses (all campus)	4.213
Arson (all campus)	7.350
Rape (student housing)	13.193
Fondling (student housing)	14.171
Robbery (student housing)	67.500
Assault (student housing)	35.636
Burglary (student housing)	15.433
Vehicle Theft (student housing)	-19.912
Arson (student housing)	82.575
Assault (criminal offense, noncampus)	32.531
Vehicle Theft (criminal offense, noncampus)	-6.897
Arson (criminal offense, noncampus)	80.363
Sex Offenses (criminal offense, public)	3.729
Fondling (criminal offense, public)	64.643
Drug Law Violations (arrest, student housing)	4.979
Drug Law Violations (arrest, noncampus)	12.475
Drug Law Violations (disciplinary, campus)	1.109
Drug Law Violations (disciplinary, housing)	1.474

rmse table

```r
## potential libraries

#install.packages("keras")
library(keras)
```

```
## Warning: package 'keras' was built under R version 4.3.3
```

```r
library(tensorflow)
```

```
## Warning: package 'tensorflow' was built under R version 4.3.3
```

```
##
## Attaching package: 'tensorflow'
```

```
## The following object is masked from 'package:caret':
##
##     train
```

```r
library(nnet)

#install.packages("neuralnet")

#compute object is masked from package:dplyr
library(neuralnet)
```

```
## Warning: package ’neuralnet’ was built under R version 4.3.3
```

```
##
## Attaching package: ’neuralnet’
```

```
## The following object is masked from ’package:dplyr’:
##
##     compute
```

```
#get plots side by side, grid.arrange()
#install.packages("gridExtra")
library(gridExtra)
```

```
## Warning: package ’gridExtra’ was built under R version 4.3.3
```

```
##
## Attaching package: ’gridExtra’
```

```
## The following object is masked from ’package:dplyr’:
##
##     combine
```

```
#for dredge()
#install.packages("MuMIn")
library(MuMIn)
```

```
## Warning: package ’MuMIn’ was built under R version 4.3.3
```

```
# set seed for reproducibility
set.seed(4242)

# NN test to see when model breaks
NN_1 <- neuralnet(all_liquor_violations ~ Rape_student_housing + Burglary_student_housing + Arson_studen
                  data = train, hidden = 1, linear.output=TRUE)

NN_2 <- neuralnet(all_liquor_violations ~ Rape_student_housing, hidden = 1, data = train, linear.output

NN_3 <- neuralnet(all_liquor_violations ~ Rape_student_housing + Burglary_student_housing, data = train

NN_4 <- neuralnet(all_liquor_violations ~ Rape_student_housing + Burglary_student_housing, data = train

NN_5 <- neuralnet(all_liquor_violations ~ Rape_student_housing + Burglary_student_housing + Arson_studen

NN_6 <- neuralnet(all_liquor_violations ~ Rape_student_housing + Burglary_student_housing + Drug.law.vio

library(modelr)
```

```
## Warning: package ’modelr’ was built under R version 4.3.3
```

```r
## test rmse

nn_rmse <- data.frame(
  rmse_1 <- rmse(NN_1, data=test),
  rmse_2 <- rmse(NN_2, data=test),
  rmse_3 <- rmse(NN_3, data=test),
  rmse_4 <- rmse(NN_4, data=test),
  rmse_5 <- rmse(NN_5, data=test),
  rmse_6 <- rmse(NN_6, data=test)
)

new_rmse <- t(nn_rmse)

rmse_table <- data.frame(
  Variable = c("1", "2", "3", "4", "5", "6"),
  Coefficients = new_rmse)

rownames(rmse_table) <- NULL

rmse_table
```

```
##   Variable Coefficients
## 1        1     423.2550
## 2        2     436.6905
## 3        3     420.3293
## 4        4     420.3293
## 5        5     417.5463
## 6        6     423.2502
```

```r
kable(rmse_table, col.names = c("Model #", "Test RMSE"), caption = "Neural Network Model Evaluations", 
```

Table 2: Neural Network Model Evaluations

Model #	Test RMSE
1	423.255
2	436.691
3	420.329
4	420.329
5	417.546
6	423.250

```r
#kable(n_rmse, col.names = c("RMSE 1", "RMSE 2", "RMSE 3", "RMSE 4", "RMSE 5", "RMSE 6"), caption = "Ne
```

```r
final_rmse <- data.frame(
  Variable = c("XGBoost", "Neural Net"),
  Coefficients = c("164.725", "417.546"))

kable(final_rmse, col.names = c("Method", "Test RMSE"), caption = "Final Model Evaluations", digits = 3
```

Table 3: Final Model Evaluations

Method	Test RMSE
XGBoost	164.725
Neural Net	417.546