Homerwork 2

YOUR NAME HERE

5/21/23

## Obtain the data

Rows: 144  
Columns: 24  
$ case <chr> "New Mexico neighborhood shooting", "…  
$ location...2 <chr> "Farmington, New Mexico", "Allen, Tex…  
$ date <chr> "5/15/23", "5/6/23", "4/10/23", "3/27…  
$ summary <chr> "Beau Wilson, 18, opened fire on the …  
$ fatalities <dbl> 3, 8, 5, 6, 3, 7, 11, 6, 5, 3, 5, 3, …  
$ injured <dbl> 6, 7, 8, 6, 5, 1, 10, 6, 25, 2, 2, 2,…  
$ total\_victims <dbl> 9, 15, 13, 12, 8, 8, 21, 12, 30, 5, 7…  
$ location...8 <chr> "Other", "Other", "workplace", "Schoo…  
$ age\_of\_shooter <chr> "18", "33", "25", "28", "43", "67", "…  
$ prior\_signs\_mental\_health\_issues <chr> "-", "yes", "yes", "-", "-", "-", "ye…  
$ mental\_health\_details <chr> "-", "Reportedly had a history of men…  
$ weapons\_obtained\_legally <chr> "yes", "yes", "yes", "yes", "yes", "-…  
$ where\_obtained <chr> "-", "-", "gun dealership in Louisvil…  
$ weapon\_type <chr> "semiautiomatic rifle; semiautomatic …  
$ weapon\_details <chr> "AR-15-style rifle", "AR-15-style rif…  
$ race <chr> "-", "Latino", "White", "White", "Bla…  
$ gender <chr> "M", "M", "M", "F (\"identifies as tr…  
$ sources <chr> "https://www.cbsnews.com/news/farming…  
$ mental\_health\_sources <chr> "-", "-", "-", "-", "-", "-", "https:…  
$ sources\_additional\_age <chr> "-", "-", "-", "-", "-", "-", "-", "-…  
$ latitude <chr> "-", "-", "-", "-", "-", "-", "-", "-…  
$ longitude <chr> "-", "-", "-", "-", "-", "-", "-", "-…  
$ type <chr> "mass", "Mass", "Mass", "Mass", "Mass…  
$ year <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2…

## Explore the data

### Specific questions

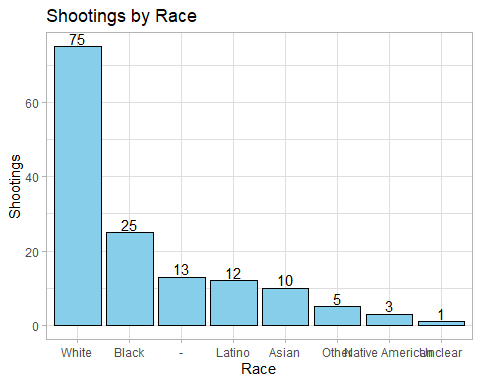
* Generate a data frame that summarizes the number of mass shootings per year.

mass\_shootings %>%   
 group\_by(year) %>%   
 summarise(shootings = n())

# A tibble: 39 × 2  
 year shootings  
 <dbl> <int>  
 1 1982 1  
 2 1984 2  
 3 1986 1  
 4 1987 1  
 5 1988 1  
 6 1989 2  
 7 1990 1  
 8 1991 3  
 9 1992 2  
10 1993 4  
# ℹ 29 more rows

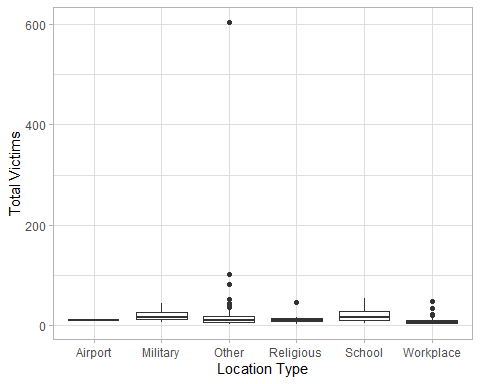
* Generate a bar chart that identifies the number of mass shooters associated with each race category. The bars should be sorted from highest to lowest and each bar should show its number.

#store to table  
race\_shootings <-  
  
 # create data frame using group by and summarise  
 mass\_shootings %>%  
   
 # use mutate to remove case sensitivity  
 mutate(race = str\_to\_title(race)) %>%  
   
 # group by and summarise to get counts by race  
 group\_by(race) %>%   
 summarise(shootings = n())  
   
# create plot  
ggplot(  
 race\_shootings  
 ,aes(  
 # sort race by shootings high to low on the x axis  
 x = fct\_rev(fct\_reorder(race, shootings))  
 ,y = shootings  
 )  
) +  
 # add cols with blue fill and black outline  
 geom\_col(color = "black", fill = "skyblue") +  
  
 # add data labels  
 geom\_text(  
  
 # set label value  
 aes(label = shootings)  
  
 # set label color  
 ,color = "black"  
  
 # adjust vertical position of label  
 ,vjust = -0.2) +  
  
 # apply theme  
 theme\_light() +  
   
 # add labels  
 labs(  
 x = "Race"  
 ,y = "Shootings"  
 ,title = "Shootings by Race"  
 )



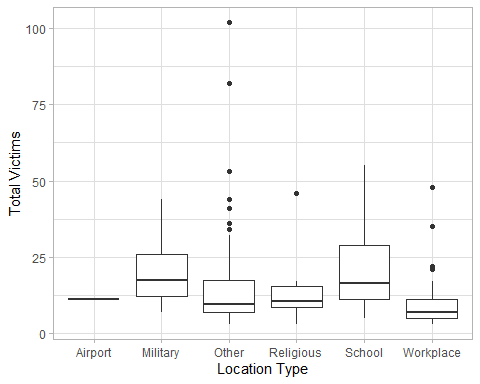
* Generate a boxplot visualizing the number of total victims, by type of location.

# store to table  
location\_type\_total\_victims <-  
 mass\_shootings %>%  
   
 # remove \n characters from location\_type  
 mutate(location\_type = str\_remove\_all(location...8, "\n")) %>%  
   
 # make all location types title case  
 mutate(location\_type = str\_to\_title(location\_type))  
  
# create boxplot from data  
ggplot(  
 location\_type\_total\_victims  
 ,aes(  
 x = location\_type  
 ,y = total\_victims  
 )  
) +  
 geom\_boxplot() +  
   
 # add theme  
 theme\_light() +  
   
 # add lables  
 labs(  
 x = "Location Type"  
 ,y = "Total Victims"  
 ,   
 )



* Redraw the same plot, but remove the Las Vegas Strip massacre from the dataset.

# store to table  
location\_type\_total\_victims\_noLV <-  
 location\_type\_total\_victims %>%  
   
 #filter out the las vegas strip massacre  
 filter(case != "Las Vegas Strip massacre")  
  
#create boxplot from data  
ggplot(  
 location\_type\_total\_victims\_noLV  
 ,aes(  
 x = location\_type  
 ,y = total\_victims  
 )  
) +  
 geom\_boxplot() +  
   
 # add theme  
 theme\_light() +  
   
 # add lables  
 labs(  
 x = "Location Type"  
 ,y = "Total Victims"  
 ,   
 )



### More open-ended questions

Address the following questions. Generate appropriate figures/tables to support your conclusions.

* How many white males with prior signs of mental illness initiated a mass shooting after 2000?
* *42*

mass\_shootings %>%  
   
 # use mutate to remove case sensitivity  
 mutate(race = str\_to\_title(race)) %>%  
 mutate(prior\_signs\_mental\_health\_issues = str\_to\_title(prior\_signs\_mental\_health\_issues)) %>%  
   
 # Add filter conditions  
 filter(  
 race == "White"  
 ,prior\_signs\_mental\_health\_issues == "Yes"  
 ,year >- 2000  
 ) %>%   
   
 # Summarise to get count  
 summarise(n())

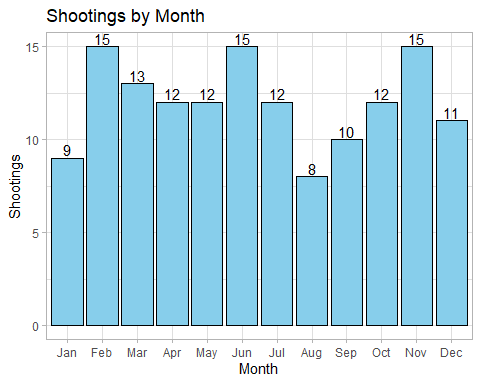
# A tibble: 1 × 1  
 `n()`  
 <int>  
1 42

* Which month of the year has the most mass shootings? Generate a bar chart sorted in chronological (natural) order (Jan-Feb-Mar- etc) to provide evidence of your answer.

#store to table  
month\_shootings <-  
 mass\_shootings %>%   
 # convert strings to date format  
 mutate(date\_formatted = mdy(date) ) %>%  
   
 # extract month from date   
 mutate(date\_month = month(date\_formatted, label = FALSE)) %>%  
   
 # get month name from date  
 mutate(date\_month\_name = month(date\_formatted, label = TRUE)) %>%   
   
 # group by month and summarise to get number of shootings  
 group\_by(date\_month, date\_month\_name) %>%   
 summarise(shootings = n())

`summarise()` has grouped output by 'date\_month'. You can override using the  
`.groups` argument.

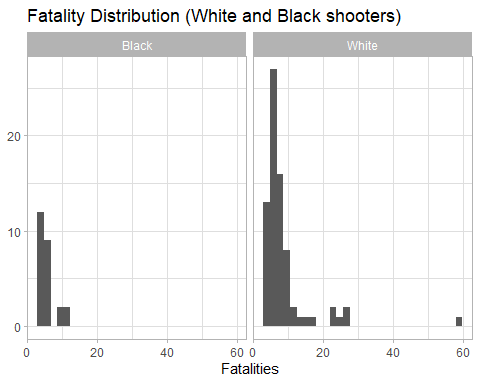
# create plot  
ggplot(  
 month\_shootings  
 ,aes(  
 # sort race by shootings high to low on the x axis  
 x = fct\_reorder(date\_month\_name, date\_month)  
 ,y = shootings  
 )  
) +  
 # add cols with blue fill and black outline  
 geom\_col(color = "black", fill = "skyblue") +  
  
 # Add labels  
 geom\_text(  
  
 # set label value  
 aes(label = shootings)  
  
 # set label color  
 ,color = "black"  
  
 # adjust vertical position of label  
 ,vjust = -0.2) +  
  
 # apply theme  
 theme\_light() +  
   
 # add labels  
 labs(  
 x = "Month"  
 ,y = "Shootings"  
 ,title = "Shootings by Month"  
 )



* How does the distribution of mass shooting fatalities differ between White and Black shooters? What about White and Latino shooters?

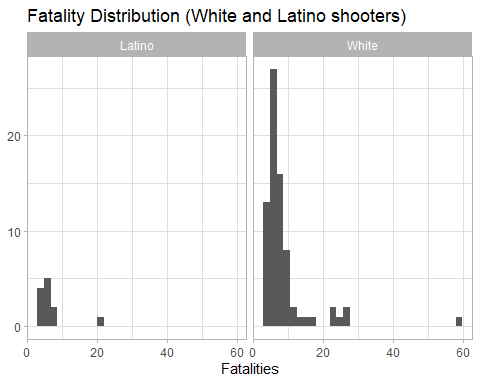
# plt to compare black and white shooters  
ggplot(  
 # use mutate to remove case sensitivity  
 mutate(mass\_shootings, race = str\_to\_title(race)) %>%   
 filter(race %in% c("White", "Black"))  
 ,aes(x = fatalities)  
) +  
  
 # create histogram+  
 geom\_histogram() +  
   
 # add facet wrap on race  
 facet\_wrap(~race) +  
   
 # apply theme and labels  
 theme\_light() +  
 labs(  
 title = "Fatality Distribution (White and Black shooters)"  
 ,x = "Fatalities"  
 ,y = NULL  
 )

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# plt to compare white and Latino shooters  
ggplot(  
 # use mutate to remove case sensitivity  
 mutate(mass\_shootings, race = str\_to\_title(race)) %>%   
 filter(race %in% c("White", "Latino"))  
 ,aes(x = fatalities)  
) +  
  
 # create histogram+  
 geom\_histogram() +  
   
 # add facet wrap on race  
 facet\_wrap(~race) +  
   
 # apply theme and labels  
 theme\_light() +  
 labs(  
 title = "Fatality Distribution (White and Latino shooters)"  
 ,x = "Fatalities"  
 ,y = NULL  
 )

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Very open-ended

* Are mass shootings with shooters suffering from mental illness different from mass shootings with no signs of mental illness in the shooter?

mass\_shootings\_mental\_health <-  
 mass\_shootings %>%   
 mutate(mental\_health\_issue = ifelse(prior\_signs\_mental\_health\_issues %in% c("Yes", "yes"), "Yes", "No / Not Clear")) %>%   
 mutate(age\_of\_shooter = strtoi(age\_of\_shooter))  
  
  
mass\_shootings\_mental\_health %>%   
 split(.$mental\_health\_issue) %>%   
 map(summary)

$`No / Not Clear`  
 case location...2 date summary   
 Length:76 Length:76 Length:76 Length:76   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 fatalities injured total\_victims location...8   
 Min. : 3.000 Min. : 0.00 Min. : 3.00 Length:76   
 1st Qu.: 3.750 1st Qu.: 1.00 1st Qu.: 5.00 Class :character   
 Median : 5.000 Median : 2.00 Median : 8.00 Mode :character   
 Mean : 7.289 Mean : 13.33 Mean : 20.62   
 3rd Qu.: 7.000 3rd Qu.: 8.00 3rd Qu.: 15.00   
 Max. :58.000 Max. :546.00 Max. :604.00   
   
 age\_of\_shooter prior\_signs\_mental\_health\_issues mental\_health\_details  
 Min. :11.00 Length:76 Length:76   
 1st Qu.:22.00 Class :character Class :character   
 Median :31.00 Mode :character Mode :character   
 Mean :33.86   
 3rd Qu.:43.75   
 Max. :70.00   
 NA's :2   
 weapons\_obtained\_legally where\_obtained weapon\_type   
 Length:76 Length:76 Length:76   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
   
 weapon\_details race gender sources   
 Length:76 Length:76 Length:76 Length:76   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 mental\_health\_sources sources\_additional\_age latitude   
 Length:76 Length:76 Length:76   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
   
 longitude type year mental\_health\_issue  
 Length:76 Length:76 Min. :1986 Length:76   
 Class :character Class :character 1st Qu.:2009 Class :character   
 Mode :character Mode :character Median :2017 Mode :character   
 Mean :2013   
 3rd Qu.:2019   
 Max. :2023   
   
  
$Yes  
 case location...2 date summary   
 Length:68 Length:68 Length:68 Length:68   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 fatalities injured total\_victims location...8   
 Min. : 3.000 Min. : 0.000 Min. : 3.00 Length:68   
 1st Qu.: 5.000 1st Qu.: 1.000 1st Qu.: 7.75 Class :character   
 Median : 7.000 Median : 4.000 Median :11.50 Mode :character   
 Mean : 8.279 Mean : 8.721 Mean :17.00   
 3rd Qu.: 9.000 3rd Qu.:13.000 3rd Qu.:21.25   
 Max. :32.000 Max. :70.000 Max. :82.00   
 age\_of\_shooter prior\_signs\_mental\_health\_issues mental\_health\_details  
 Min. :15.00 Length:68 Length:68   
 1st Qu.:23.00 Class :character Class :character   
 Median :34.50 Mode :character Mode :character   
 Mean :33.79   
 3rd Qu.:42.00   
 Max. :72.00   
 weapons\_obtained\_legally where\_obtained weapon\_type   
 Length:68 Length:68 Length:68   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
 weapon\_details race gender sources   
 Length:68 Length:68 Length:68 Length:68   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 mental\_health\_sources sources\_additional\_age latitude   
 Length:68 Length:68 Length:68   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
 longitude type year mental\_health\_issue  
 Length:68 Length:68 Min. :1982 Length:68   
 Class :character Class :character 1st Qu.:1999 Class :character   
 Mode :character Mode :character Median :2012 Mode :character   
 Mean :2008   
 3rd Qu.:2017   
 Max. :2023

*For shooters with confirmed higher mental health issues we see higher mean fatalities but lower total victims. This indicates that these shooters are more lethal.*

*The ages of the shooters do not vary much*

* Assess the relationship between mental illness and total victims, mental illness and location type, and the intersection of all three variables.

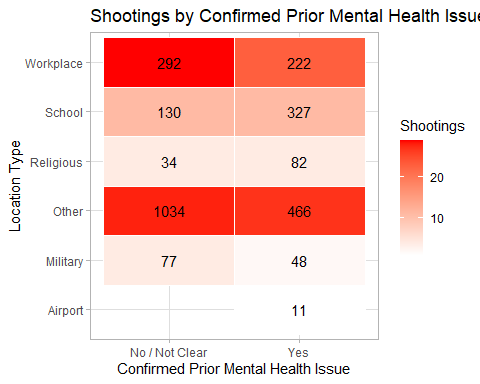
mass\_shootings\_mental\_health %>%   
 group\_by(mental\_health\_issue) %>%   
 summarise(  
 shootings = n()  
 ,total\_victims = sum(total\_victims)  
 ,total\_victims\_per\_shooting = sum(total\_victims) / n()  
 )

# A tibble: 2 × 4  
 mental\_health\_issue shootings total\_victims total\_victims\_per\_shooting  
 <chr> <int> <dbl> <dbl>  
1 No / Not Clear 76 1567 20.6  
2 Yes 68 1156 17

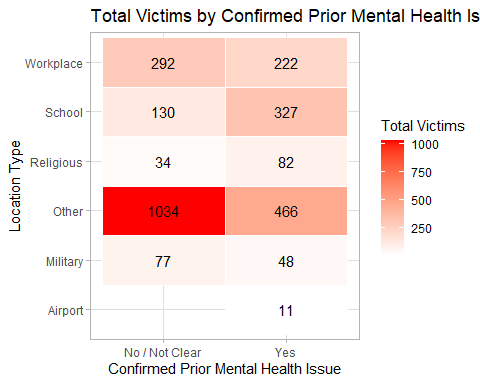
# store to new table  
location\_type\_mental\_health <-  
 mass\_shootings\_mental\_health %>%   
   
 # remove \n characters from location\_type  
 mutate(location\_type = str\_remove\_all(location...8, "\n")) %>%  
   
 # make all location types title case  
 mutate(location\_type = str\_to\_title(location\_type)) %>%   
   
 # group by categories  
 group\_by(mental\_health\_issue, location\_type) %>%   
   
 # summarise to get statistics  
 summarise(  
 shootings = n()  
 ,total\_victims = sum(total\_victims)  
 )

`summarise()` has grouped output by 'mental\_health\_issue'. You can override  
using the `.groups` argument.

# create heatmap for shootings  
ggplot(  
 location\_type\_mental\_health  
 ,aes(  
 x = mental\_health\_issue  
 ,y = location\_type  
 ,fill = shootings  
 )  
) +  
   
 # add tile chart  
 geom\_tile(color = "white") +  
   
 # add value labels  
 geom\_text(  
 aes(label = total\_victims)  
 ,color = "black"  
 ) +  
   
 # add colors to the gradient  
 scale\_fill\_gradient(low = "white", high = "red") +  
   
 # Add labels and theme  
 labs(  
 x = "Confirmed Prior Mental Health Issue"  
 ,y = "Location Type"  
 ,fill = "Shootings"  
 ,title = "Shootings by Confirmed Prior Mental Health Issue and Location Type"  
 ) +  
 theme\_light()



# create heatmap for total victims  
ggplot(  
 location\_type\_mental\_health  
 ,aes(  
 x = mental\_health\_issue  
 ,y = location\_type  
 ,fill = total\_victims  
 )  
) +  
   
 # add tile chart  
 geom\_tile(color = "white") +  
   
 # add value labels  
 geom\_text(  
 aes(label = total\_victims)  
 ,color = "black"  
 ) +  
   
 # add colors to the gradient  
 scale\_fill\_gradient(low = "white", high = "red") +  
   
 # Add labels and theme  
 labs(  
 x = "Confirmed Prior Mental Health Issue"  
 ,y = "Location Type"  
 ,fill = "Total Victims"  
 ,title = "Total Victims by Confirmed Prior Mental Health Issue and Location Type"  
 ) +  
 theme\_light()



Make sure to provide a couple of sentences of written interpretation of your tables/figures. Graphs and tables alone will not be sufficient to answer this question.

*The number of shootings by those with confirmed prior mental health issues and those without are similar but those without confirmed prior mental health issues have a higher number of total victims per shooting.*

*The distribution of shootings by location type are similar between those with confirmed prior mental health issues and those without although there are slightly more workplace shootings by those without.*

*The distribution of Total Victims is heavily skewed by the Los Vegas shooting but otherwise follows a similar distribution as the number of shootings.*

# Exploring credit card fraud

## Obtain the data

Rows: 671,028  
Columns: 14  
$ trans\_date\_trans\_time <dttm> 2019-02-22 07:32:58, 2019-02-16 15:07:20, 2019-…  
$ trans\_year <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2020, …  
$ category <chr> "entertainment", "kids\_pets", "personal\_care", "…  
$ amt <dbl> 7.79, 3.89, 8.43, 40.00, 54.04, 95.61, 64.95, 3.…  
$ city <chr> "Veedersburg", "Holloway", "Arnold", "Apison", "…  
$ state <chr> "IN", "OH", "MO", "TN", "CO", "GA", "MN", "AL", …  
$ lat <dbl> 40.1186, 40.0113, 38.4305, 35.0149, 39.4584, 32.…  
$ long <dbl> -87.2602, -80.9701, -90.3870, -85.0164, -106.385…  
$ city\_pop <dbl> 4049, 128, 35439, 3730, 277, 1841, 136, 190178, …  
$ job <chr> "Development worker, community", "Child psychoth…  
$ dob <date> 1959-10-19, 1946-04-03, 1985-03-31, 1991-01-28,…  
$ merch\_lat <dbl> 39.41679, 39.74585, 37.73078, 34.53277, 39.95244…  
$ merch\_long <dbl> -87.52619, -81.52477, -91.36875, -84.10676, -106…  
$ is\_fraud <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

* In this dataset, how likely are fraudulent transactions? Generate a table that summarizes the number and frequency of fraudulent transactions per year.

card\_fraud %>%  
   
 # group by transaction year  
 group\_by(trans\_year) %>%   
   
 # use the is\_fraud field to get counts of (non) fraud transactions and frequency  
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 )

# A tibble: 2 × 4  
 trans\_year fraud\_count non\_fraud\_count fraud\_freq  
 <dbl> <int> <int> <dbl>  
1 2019 2721 475925 0.00572  
2 2020 1215 191167 0.00636

* How much money (in US$ terms) are fraudulent transactions costing the company? Generate a table that summarizes the total amount of legitimate and fraudulent transactions per year and calculate the % of fraudulent transactions, in US$ terms.

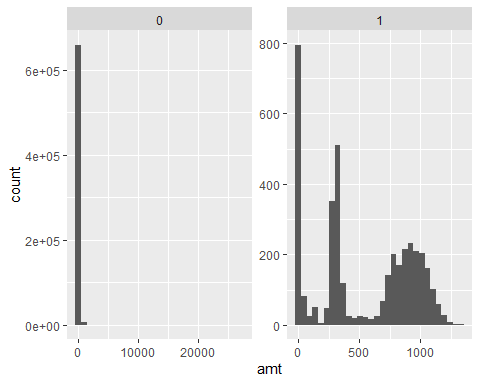
card\_fraud %>%   
   
 # group by year  
 group\_by(trans\_year) %>%   
   
 # get fraud and non fraud amounts  
 summarise(  
 fraud\_amt = sum(amt, is\_fraud == 1)  
 ,non\_fraud\_amt = sum(amt, is\_fraud == 0)  
 ) %>%   
   
 # add the pct  
 mutate(fraud\_amt\_pct = fraud\_amt / (fraud\_amt + non\_fraud\_amt))

# A tibble: 2 × 4  
 trans\_year fraud\_amt non\_fraud\_amt fraud\_amt\_pct  
 <dbl> <dbl> <dbl> <dbl>  
1 2019 33608762. 34081966. 0.497  
2 2020 13579078. 13769030. 0.497

* Generate a histogram that shows the distribution of amounts charged to credit card, both for legitimate and fraudulent accounts. Also, for both types of transactions, calculate some quick summary statistics.

# generate histogram  
ggplot(  
 card\_fraud  
 ,aes(x = amt)  
) +  
 geom\_histogram() +  
 facet\_wrap(~is\_fraud, scales = "free")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



summary(filter(card\_fraud, is\_fraud == 1))

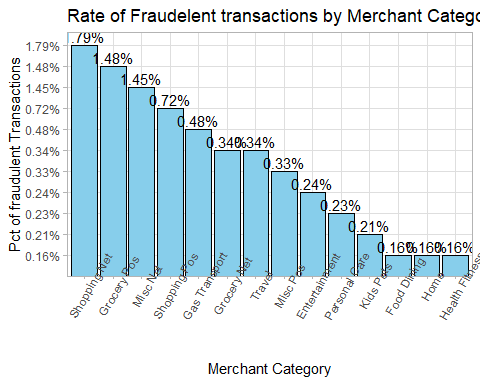
trans\_date\_trans\_time trans\_year category   
 Min. :2019-01-02 01:06:37.00 Min. :2019 Length:3936   
 1st Qu.:2019-05-04 23:31:33.00 1st Qu.:2019 Class :character   
 Median :2019-09-29 05:54:09.50 Median :2019 Mode :character   
 Mean :2019-09-25 17:23:00.88 Mean :2019   
 3rd Qu.:2020-02-08 00:08:40.50 3rd Qu.:2020   
 Max. :2020-06-21 03:59:46.00 Max. :2020   
 amt city state lat   
 Min. : 1.06 Length:3936 Length:3936 Min. :20.03   
 1st Qu.: 240.49 Class :character Class :character 1st Qu.:35.06   
 Median : 368.83 Mode :character Mode :character Median :39.43   
 Mean : 527.21 Mean :38.65   
 3rd Qu.: 900.95 3rd Qu.:41.84   
 Max. :1334.07 Max. :66.69   
 long city\_pop job dob   
 Min. :-165.67 Min. : 23 Length:3936 Min. :1925-08-29   
 1st Qu.: -96.70 1st Qu.: 741 Class :character 1st Qu.:1958-03-18   
 Median : -86.69 Median : 2526 Mode :character Median :1971-08-20   
 Mean : -89.91 Mean : 94096 Mean :1970-09-21   
 3rd Qu.: -79.99 3rd Qu.: 19803 3rd Qu.:1986-11-24   
 Max. : -68.56 Max. :2906700 Max. :2005-01-29   
 merch\_lat merch\_long is\_fraud  
 Min. :19.53 Min. :-166.40 Min. :1   
 1st Qu.:35.12 1st Qu.: -96.72 1st Qu.:1   
 Median :39.42 Median : -86.88 Median :1   
 Mean :38.64 Mean : -89.91 Mean :1   
 3rd Qu.:41.92 3rd Qu.: -79.91 3rd Qu.:1   
 Max. :67.44 Max. : -67.57 Max. :1

summary(filter(card\_fraud, is\_fraud == 0))

trans\_date\_trans\_time trans\_year category   
 Min. :2019-01-01 00:00:51.0 Min. :2019 Length:667092   
 1st Qu.:2019-06-04 06:41:49.0 1st Qu.:2019 Class :character   
 Median :2019-10-03 16:40:52.0 Median :2019 Mode :character   
 Mean :2019-10-03 16:30:11.0 Mean :2019   
 3rd Qu.:2020-01-28 16:17:57.5 3rd Qu.:2020   
 Max. :2020-06-21 12:12:32.0 Max. :2020   
 amt city state lat   
 Min. : 1.00 Length:667092 Length:667092 Min. :20.03   
 1st Qu.: 9.60 Class :character Class :character 1st Qu.:34.62   
 Median : 47.17 Mode :character Mode :character Median :39.35   
 Mean : 67.62 Mean :38.54   
 3rd Qu.: 82.41 3rd Qu.:41.89   
 Max. :27119.77 Max. :65.69   
 long city\_pop job dob   
 Min. :-165.67 Min. : 23 Length:667092 Min. :1924-10-30   
 1st Qu.: -96.80 1st Qu.: 741 Class :character 1st Qu.:1962-08-13   
 Median : -87.48 Median : 2456 Mode :character Median :1975-11-30   
 Mean : -90.23 Mean : 88876 Mean :1973-10-15   
 3rd Qu.: -80.16 3rd Qu.: 20328 3rd Qu.:1987-04-23   
 Max. : -67.95 Max. :2906700 Max. :2005-01-29   
 merch\_lat merch\_long is\_fraud  
 Min. :19.03 Min. :-166.67 Min. :0   
 1st Qu.:34.73 1st Qu.: -96.90 1st Qu.:0   
 Median :39.37 Median : -87.44 Median :0   
 Mean :38.53 Mean : -90.23 Mean :0   
 3rd Qu.:41.95 3rd Qu.: -80.23 3rd Qu.:0   
 Max. :66.68 Max. : -66.95 Max. :0

* What types of purchases are most likely to be instances of fraud? Consider category of merchants and produce a bar chart that shows % of total fraudulent transactions sorted in order.

card\_fraud %>%  
   
 # group by merchant category  
 group\_by(category) %>%   
   
 # use the is\_fraud field to get counts of (non) fraud transactions and frequency  
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 ) %>%   
   
 mutate(category = str\_replace\_all(category, "\_", " ")) %>%  
 mutate(category = str\_to\_title(category)) %>%   
   
 # create plot frame  
 ggplot(  
 aes(  
 x = fct\_rev(fct\_reorder(category, fraud\_freq))  
 ,y = percent(fraud\_freq, 0.01)  
 )  
 ) +  
  
 # add bar chart  
 geom\_col(color = "black", fill = "skyblue") +  
  
 # Add labels  
 geom\_text(  
  
 # set label value  
 aes(label = percent(fraud\_freq, 0.01))  
  
 # set label color  
 ,color = "black"  
  
 # adjust vertical position of label  
 ,vjust = -0.2) +  
  
 # apply theme  
 theme\_light() +  
 theme(axis.text.x = element\_text(angle = 60)) +  
   
 # add labels  
 labs(  
 x = "Merchant Category"  
 ,y = "Pct of fraudulent Transactions"  
 ,title = "Rate of Fraudelent transactions by Merchant Category"  
 )



* When is fraud more prevalent? Which days, months, hours? To create new variables to help you in your analysis, we use the lubridate package and the following code

mutate(  
 date\_only = lubridate::date(trans\_date\_trans\_time),  
 month\_name = lubridate::month(trans\_date\_trans\_time, label=TRUE),  
 hour = lubridate::hour(trans\_date\_trans\_time),  
 weekday = lubridate::wday(trans\_date\_trans\_time, label = TRUE)  
 )

* Are older customers significantly more likely to be victims of credit card fraud? To calculate a customer’s age, we use the lubridate package and the following code

mutate(  
 age = interval(dob, trans\_date\_trans\_time) / years(1),  
 )

card\_fraud\_dates <-  
 card\_fraud %>%   
 mutate(  
 date\_only = lubridate::date(trans\_date\_trans\_time)  
 ,month\_name = lubridate::month(trans\_date\_trans\_time, label=TRUE)  
 ,hour = lubridate::hour(trans\_date\_trans\_time)  
 ,weekday = lubridate::wday(trans\_date\_trans\_time, label = TRUE)  
 ,age = interval(dob, trans\_date\_trans\_time) / years(1)  
 )  
  
card\_fraud\_dates %>%   
 group\_by(weekday) %>%   
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 ) %>%   
 arrange(desc(fraud\_freq))

# A tibble: 7 × 4  
 weekday fraud\_count non\_fraud\_count fraud\_freq  
 <ord> <int> <int> <dbl>  
1 Thu 542 75658 0.00716  
2 Fri 557 78394 0.00711  
3 Wed 468 67471 0.00694  
4 Sat 626 103413 0.00605  
5 Tue 496 82434 0.00602  
6 Mon 639 130780 0.00489  
7 Sun 608 128942 0.00472

card\_fraud\_dates %>%   
 group\_by(month\_name) %>%   
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 ) %>%   
 arrange(desc(fraud\_freq))

# A tibble: 12 × 4  
 month\_name fraud\_count non\_fraud\_count fraud\_freq  
 <ord> <int> <int> <dbl>  
 1 Jan 461 53345 0.00864  
 2 Feb 434 50226 0.00864  
 3 Mar 472 74006 0.00638  
 4 May 472 75329 0.00627  
 5 Nov 226 36107 0.00626  
 6 Oct 218 35869 0.00608  
 7 Sep 219 36314 0.00603  
 8 Jun 387 73827 0.00524  
 9 Apr 349 69527 0.00502  
10 Aug 213 45067 0.00473  
11 Dec 301 72685 0.00414  
12 Jul 184 44790 0.00411

card\_fraud\_dates %>%   
 group\_by(hour) %>%   
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 ) %>%   
 arrange(desc(fraud\_freq))

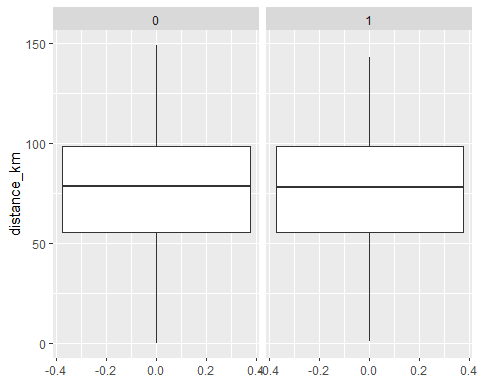
# A tibble: 24 × 4  
 hour fraud\_count non\_fraud\_count fraud\_freq  
 <int> <int> <int> <dbl>  
 1 23 1012 33613 0.0301   
 2 22 981 33693 0.0291   
 3 0 348 21722 0.0160   
 4 1 332 21775 0.0152   
 5 3 326 21866 0.0149   
 6 2 313 21909 0.0143   
 7 7 35 21820 0.00160  
 8 19 52 33935 0.00153  
 9 5 32 21720 0.00147  
10 18 49 34082 0.00144  
# ℹ 14 more rows

card\_fraud\_dates %>%   
 group\_by(round(age, 0)) %>%   
 summarise(  
 fraud\_count = sum(is\_fraud == 1)  
 ,non\_fraud\_count = sum(is\_fraud == 0)  
 ,fraud\_freq = sum(is\_fraud == 1) / sum(is\_fraud == 0)  
 ) %>%   
 arrange(desc(fraud\_freq))

# A tibble: 83 × 4  
 `round(age, 0)` fraud\_count non\_fraud\_count fraud\_freq  
 <dbl> <int> <int> <dbl>  
 1 87 28 1599 0.0175  
 2 92 38 2399 0.0158  
 3 18 33 2103 0.0157  
 4 77 35 2354 0.0149  
 5 79 54 3936 0.0137  
 6 86 22 1719 0.0128  
 7 81 26 2034 0.0128  
 8 94 10 819 0.0122  
 9 71 57 4716 0.0121  
10 76 23 1922 0.0120  
# ℹ 73 more rows

* Is fraud related to distance? The distance between a card holder’s home and the location of the transaction can be a feature that is related to fraud. To calculate distance, we need the latidue/longitude of card holders’s home and the latitude/longitude of the transaction, and we will use the [Haversine formula](https://en.wikipedia.org/wiki/Haversine_formula) to calculate distance. I adapted code to [calculate distance between two points on earth](https://www.geeksforgeeks.org/program-distance-two-points-earth/amp/) which you can find below

# distance between card holder's home and transaction  
# code adapted from https://www.geeksforgeeks.org/program-distance-two-points-earth/amp/  
  
  
card\_fraud %>%  
 mutate(  
   
 # convert latitude/longitude to radians  
 lat1\_radians = lat / 57.29577951,  
 lat2\_radians = merch\_lat / 57.29577951,  
 long1\_radians = long / 57.29577951,  
 long2\_radians = merch\_long / 57.29577951,  
   
 # calculate distance in miles  
 distance\_miles = 3963.0 \* acos((sin(lat1\_radians) \* sin(lat2\_radians)) + cos(lat1\_radians) \* cos(lat2\_radians) \* cos(long2\_radians - long1\_radians)),  
  
 # calculate distance in km  
 distance\_km = 6377.830272 \* acos((sin(lat1\_radians) \* sin(lat2\_radians)) + cos(lat1\_radians) \* cos(lat2\_radians) \* cos(long2\_radians - long1\_radians))  
  
 ) %>%   
   
 # add plot  
 ggplot(  
 aes(  
 y = distance\_km  
 )  
 ) +  
 geom\_boxplot() +  
   
 # use facet wrap to separate by is\_fraud  
 facet\_wrap(~is\_fraud)



Plot a boxplot or a violin plot that looks at the relationship of distance and is\_fraud. Does distance seem to be a useful feature in explaining fraud?

*The boxplots are extremely similar.*

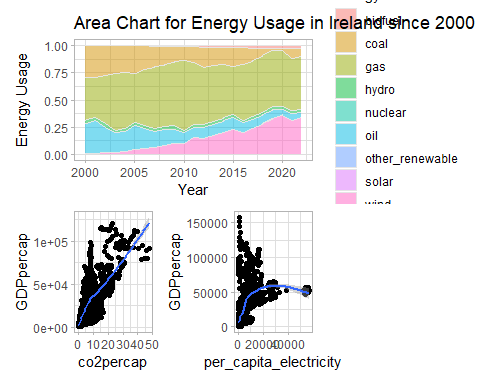
# Exploring sources of electricity production, CO2 emissions, and GDP per capita.

## 1. A stacked area chart that shows how your own country generated its electricity since 2000.

## 2. A scatter plot that looks at how CO2 per capita and GDP per capita are related

## 3. A scatter plot that looks at how electricity usage (kWh) per capita/day GDP per capita are related.

# Download electricity data  
url <- "https://nyc3.digitaloceanspaces.com/owid-public/data/energy/owid-energy-data.csv"  
  
energy <- read\_csv(url) %>%   
 filter(year >= 1990) %>%   
 drop\_na(iso\_code) %>%   
 select(1:3,  
 biofuel = biofuel\_electricity,  
 coal = coal\_electricity,  
 gas = gas\_electricity,  
 hydro = hydro\_electricity,  
 nuclear = nuclear\_electricity,  
 oil = oil\_electricity,  
 other\_renewable = other\_renewable\_exc\_biofuel\_electricity,  
 solar = solar\_electricity,  
 wind = wind\_electricity,   
 electricity\_demand,  
 electricity\_generation,  
 net\_elec\_imports, # Net electricity imports, measured in terawatt-hours  
 energy\_per\_capita, # Primary energy consumption per capita, measured in kilowatt-hours Calculated by Our World in Data based on BP Statistical Review of World Energy and EIA International Energy Data  
 energy\_per\_gdp, # Energy consumption per unit of GDP. This is measured in kilowatt-hours per 2011 international-$.  
 per\_capita\_electricity, # Electricity generation per capita, measured in kilowatt-hours  
 )   
  
# Download data for C02 emissions per capita https://data.worldbank.org/indicator/EN.ATM.CO2E.PC  
co2\_percap <- wb\_data(country = "countries\_only",   
 indicator = "EN.ATM.CO2E.PC",   
 start\_date = 1990,   
 end\_date = 2022,  
 return\_wide=FALSE) %>%   
 filter(!is.na(value)) %>%   
 #drop unwanted variables  
 select(-c(unit, obs\_status, footnote, last\_updated)) %>%   
 rename(year = date,  
 co2percap = value)  
  
  
# Download data for GDP per capita https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD  
gdp\_percap <- wb\_data(country = "countries\_only",   
 indicator = "NY.GDP.PCAP.PP.KD",   
 start\_date = 1990,   
 end\_date = 2022,  
 return\_wide=FALSE) %>%   
 filter(!is.na(value)) %>%   
 #drop unwanted variables  
 select(-c(unit, obs\_status, footnote, last\_updated)) %>%   
 rename(year = date,  
 GDPpercap = value)  
  
# save to p1  
p1 <-  
energy %>%   
   
 # select only cols needed  
 select(1:12) %>%   
   
 # filter to relevant data  
 filter(  
 country == "Ireland"  
 ,year >= 2000  
 ) %>%   
   
 # pivot to long format  
 pivot\_longer(  
 cols = 4:12  
 ,names\_to = "energy\_source"  
 ,values\_to = "energy\_usage"  
 ) %>%   
   
 # create plot  
 ggplot(  
 aes(  
 x = year  
 ,y = energy\_usage  
 ,fill = energy\_source  
 )  
 ) +  
   
 # add area chart  
 geom\_area(colour="grey90", alpha = 0.5, position = "fill") +  
  
 # add theme labels  
 theme\_light() +  
 labs(  
 x = "Year"  
 ,y = "Energy Usage"  
 ,fill = "Energy Source"  
 ,title = "Area Chart for Energy Usage in Ireland since 2000"  
 )  
  
# save to p2  
p2 <-  
  
# join tables on country and year  
inner\_join(  
 x = co2\_percap  
 ,y = gdp\_percap  
 ,by = c("iso3c", "year")  
 ,keep = TRUE  
) %>%  
   
 # set up plot  
 ggplot(  
 aes(  
 x = co2percap  
 ,y = GDPpercap  
 )  
 ) +  
   
 # add scatter  
 geom\_point() +  
   
 # add line of best fit  
 geom\_smooth() +  
   
 # add theme  
 theme\_light()  
  
# save to p3  
p3 <-  
  
# left join eregy to gdp  
left\_join(  
 x = select(energy, c("iso\_code", "year", "per\_capita\_electricity"))  
 ,y = select(gdp\_percap, c("iso3c", "year", "GDPpercap"))  
 ,by = c("iso\_code" = "iso3c", "year")  
) %>%   
 ggplot(  
 aes(  
 x = per\_capita\_electricity  
 ,y = GDPpercap  
 )  
 ) +  
   
 # add scatter  
 geom\_point() +  
   
 # add line of best fit  
 geom\_smooth() +  
   
 # add theme  
 theme\_light()  
  
# use patchwork to display the charts  
p1 / (p2 + p3)



Specific questions:

1. How would you turn energy to long, tidy format?

* *See pivot in code block above*

1. You may need to join these data frames
   * Use left\_join from dplyr to [join the tables](http://r4ds.had.co.nz/relational-data.html)
   * To complete the merge, you need a unique *key* to match observations between the data frames. Country names may not be consistent among the three dataframes, so please use the 3-digit ISO code for each country
   * An aside: There is a great package called [countrycode](https://github.com/vincentarelbundock/countrycode) that helps solve the problem of inconsistent country names (Is it UK? United Kingdon? Great Britain?). countrycode() takes as an input a country’s name in a specific format and outputs it using whatever format you specify.
2. Write a function that takes as input any country’s name and returns all three graphs. You can use the patchwork package to arrange the three graphs as shown below

* *I cannot get the image to load!!*

# Deliverables

There is a lot of explanatory text, comments, etc. You do not need these, so delete them and produce a stand-alone document that you could share with someone. Knit the edited and completed Quarto Markdown (qmd) file as a Word document (use the “Render” button at the top of the script editor window) and upload it to Canvas. You must be commiting and pushing tour changes to your own Github repo as you go along.

# Details

* Who did you collaborate with: *n/a*
* Approximately how much time did you spend on this problem set: *4 hours*
* What, if anything, gave you the most trouble: *The number of questions!*