Homework 2

Nikita Reid

5/22/23

# Data Visualisation

# Mass shootings in the US

In July 2012, in the aftermath of a mass shooting in a movie theater in Aurora, Colorado, [Mother Jones](https://www.motherjones.com/politics/2012/07/mass-shootings-map/) published a report on mass shootings in the United States since 1982. They provided the underlying data set as [an open-source database](https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/) for anyone interested in studying criminal behavior.

Rows: 125  
Columns: 14  
$ case <chr> "Oxford High School shooting", "San Jose VTA shoo…  
$ year <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2020, 2020, 2…  
$ month <chr> "Nov", "May", "Apr", "Mar", "Mar", "Mar", "Mar", …  
$ day <dbl> 30, 26, 15, 31, 22, 16, 16, 26, 10, 6, 31, 4, 3, …  
$ location <chr> "Oxford, Michigan", "San Jose, California", "Indi…  
$ summary <chr> "Ethan Crumbley, a 15-year-old student at Oxford …  
$ fatalities <dbl> 4, 9, 8, 4, 10, 8, 4, 5, 4, 3, 7, 9, 22, 3, 12, 5…  
$ injured <dbl> 7, 0, 7, 1, 0, 1, 0, 0, 3, 8, 25, 27, 26, 12, 4, …  
$ total\_victims <dbl> 11, 9, 15, 5, 10, 9, 4, 5, 7, 11, 32, 36, 48, 15,…  
$ location\_type <chr> "School", "Workplace", "Workplace", "Workplace", …  
$ male <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T…  
$ age\_of\_shooter <dbl> 15, 57, 19, NA, 21, 21, 31, 51, NA, NA, 36, 24, 2…  
$ race <chr> NA, NA, "White", NA, NA, "White", NA, "Black", "B…  
$ prior\_mental\_illness <chr> NA, "Yes", "Yes", NA, "Yes", NA, NA, NA, NA, NA, …

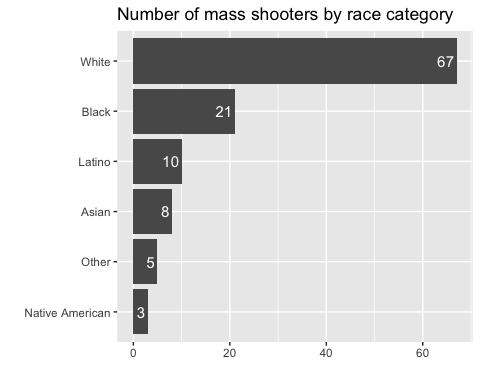
* Generate a data frame that summarizes the number of mass shootings per year.

# create new dataframe   
  
number\_shootings <- mass\_shootings %>%   
   
# group by year and count the number of shooitngs in each   
 group\_by(year) %>%   
 count()  
  
number\_shootings

# A tibble: 37 × 2  
# Groups: year [37]  
 year n  
 <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 27 more rows

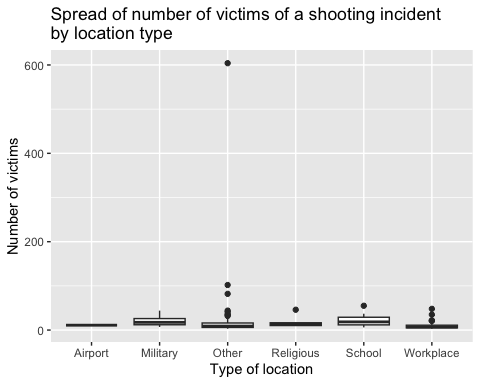
* Generate a bar chart that identifies the number of mass shooters associated with each race category. The bars should be sorted from highest to lowest and each bar should show its number.

# remove missing values  
mass\_shootings %>%  
 drop\_na(race) %>%   
  
# count the frequency of different races   
 count(race) %>%  
   
# arrange races in order  
 mutate(race = fct\_reorder(race,n)) %>%  
  
# plot bar chart  
 ggplot(aes(x = n, y = race)) +   
 geom\_bar(stat = "identity") +  
  
# add text for count to each bar  
 geom\_text(aes(label = n, x = n - 0.5),   
 colour = "white", size = 4 , hjust = 1) +  
 labs(title = "Number of mass shooters by race category"  
 , x ="", y ="")



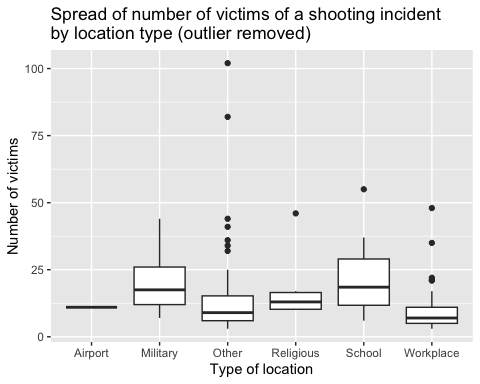
* Generate a boxplot visualizing the number of total victims, by type of location.

# boxplot of total shooting victims by location  
mass\_shootings %>%  
 ggplot(aes(x = location\_type, y = total\_victims)) +   
 geom\_boxplot() +  
  
# some labels we probably need   
 labs(title = "Spread of number of victims of a shooting incident \nby location type",   
 x = "Type of location", y= "Number of victims")



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

mass\_shootings %>%  
   
# filter to use all data except the Las Vegas strip massacre  
 filter(case != "Las Vegas Strip massacre") %>%   
   
# boxplot of total shooting victims by location  
 ggplot(aes(x = location\_type, y = total\_victims)) +   
 geom\_boxplot() +  
  
# some labels we probably need   
 labs(title = "Spread of number of victims of a shooting incident \nby location type (outlier removed)",   
 x = "Type of location", y= "Number of victims")



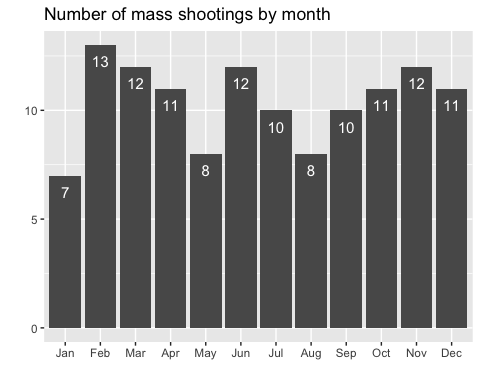
* How many white males with prior signs of mental illness initiated a mass shooting after 2000? 22

mass\_shootings %>%   
   
# filter for incident happening after 2000  
 filter(year>2000) %>%   
# filter for male and white and prior mental illness   
 filter(male==TRUE &   
 race=="White" &   
 prior\_mental\_illness=="Yes") %>%   
# count them   
 count()

# A tibble: 1 × 1  
 n  
 <int>  
1 22

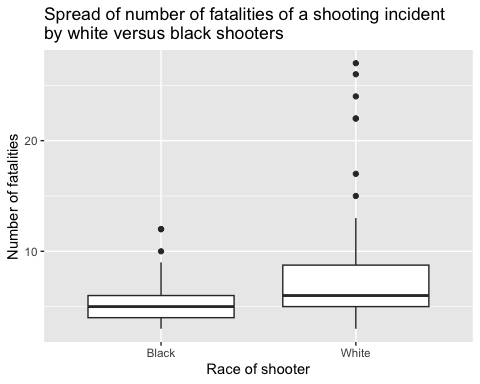
* Which month of the year has the most mass shootings? Generate a bar chart sorted in chronological (natural) order (Jan-Feb-Mar- etc) to provide evidence of your answer. From the bar chart, February has the most mass shootings, with 13.

mass\_shootings %>%   
   
# count mass shootings for each month   
 group\_by(month) %>%   
 count() %>%   
   
# plot bar chart   
 ggplot(aes(x = month, y = n)) +   
 geom\_bar(stat = "identity") +  
   
# fix x scale to put months in natural order   
 scale\_x\_discrete(limits= month.abb) +  
   
# add text to bars to show number of shootings for each month  
 geom\_text(aes(label = n, y= n - 0.5),   
 colour = "white", size = 4 , vjust = 1) +  
   
# don't need axis labels here   
 labs(title = "Number of mass shootings by month"  
 , x ="", y ="")

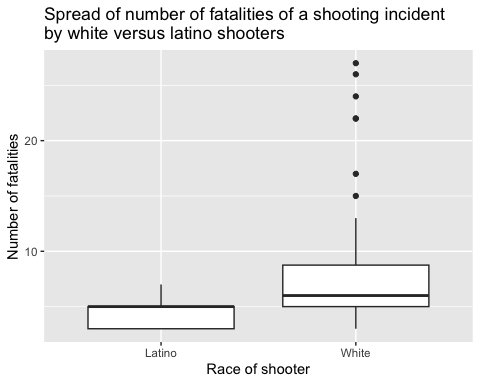


* How does the distribution of mass shooting fatalities differ between White and Black shooters? What about White and Latino shooters?
* On average (from box plots) the number of fatalities in incidents committed by white shooters is higher than those comitted by both black and latino shooters as well as there being a non-neglible number of extreme cases where white shooters claimed many victims. The trend seems to suggest that fatalities are higher in incidents when white shooters are involved.

# boxplot of total shooting fatalities by race white vs black   
mass\_shootings %>%  
   
# filter for race white or black first then plot   
 filter(race=="White"| race=="Black") %>%   
   
# filter out the Las Vegas massacre again because it ruins plots!  
 filter(case != "Las Vegas Strip massacre") %>%   
   
 ggplot(aes(x = race, y = fatalities)) +   
 geom\_boxplot() +  
  
# labels  
 labs(title = "Spread of number of fatalities of a shooting incident \nby white versus black shooters",   
 x = "Race of shooter", y= "Number of fatalities")

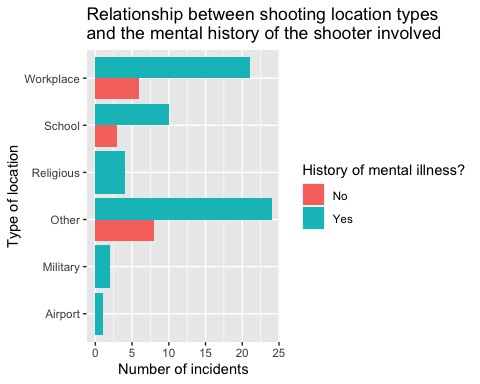


# boxplot of total shooting fatalities by race white vs latino   
mass\_shootings %>%  
   
# filter for race white or latino first then plot   
 filter(race=="White"| race=="Latino") %>%   
   
# filter out the Las Vegas massacre again because it ruins plots!  
 filter(case != "Las Vegas Strip massacre") %>%   
   
 ggplot(aes(x = race, y = fatalities)) +   
 geom\_boxplot() +  
  
# labels  
 labs(title = "Spread of number of fatalities of a shooting incident \nby white versus latino shooters",   
 x = "Race of shooter", y= "Number of fatalities")

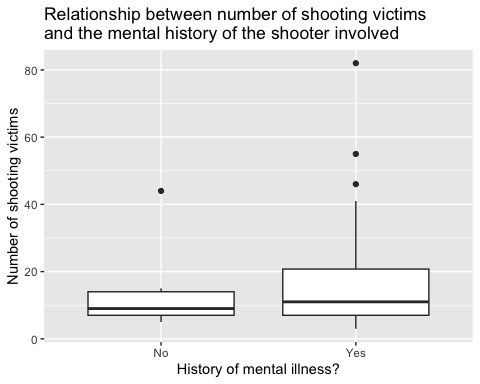


* Are mass shootings with shooters suffering from mental illness different from mass shootings with no signs of mental illness in the shooter?
* In general, from box plot, shooters with a prior mental illness documented tend to harm more victims in a shooting, on average and in extreme outlying cases. Furthermore, from bar chart, shooters with a history of mental illness tend to engage in shootings at all types of locations, far more frequently than shooters without a history of mental illness. This probably makes sense, as these people tend to be erratic, and their motivations for violence are often not rational or true to any type of pattern. Shooters with no mental history only shot at schools, workplaces or other locations. This seems to suggest more “explainable” motives such as a vendetta against former coworkers or schoolmates.

# first let's investigate where these shootings take place   
 mass\_shootings %>%   
   
# remove missing values to investigate prior mental illness   
 drop\_na(prior\_mental\_illness) %>%   
   
# collect values by location type and history of mental illness   
 group\_by(location\_type, prior\_mental\_illness) %>%   
 count() %>%   
   
# plot separate side by side bars by location type   
# for shooters with and without a history of mental illness   
 ggplot(aes(x = n, y = location\_type,   
 fill= prior\_mental\_illness)) +   
 geom\_bar(stat="identity",position ="dodge") +  
   
 labs(title = "Relationship between shooting location types \nand the mental history of the shooter involved",  
 x = "Number of incidents", y = "Type of location",  
 fill = "History of mental illness?")

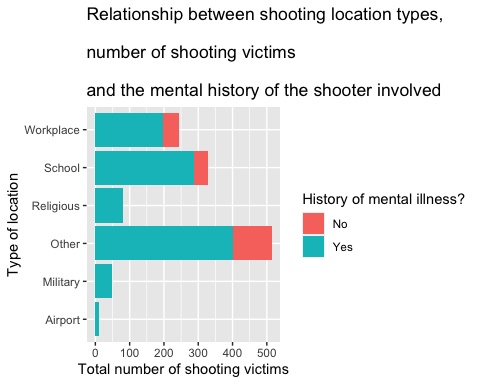


# next let's investigate how number of victims relate to mental illness  
mass\_shootings %>%   
   
# remove missing values to investigate prior mental illness   
 drop\_na(prior\_mental\_illness) %>%   
   
# plot boxplot to see distribution of victims by mental history   
 ggplot(aes(x = prior\_mental\_illness, y = total\_victims)) +  
 geom\_boxplot() +  
   
 labs(title = "Relationship between number of shooting victims \nand the mental history of the shooter involved",  
 x = "History of mental illness?",   
 y = "Number of shooting victims")



* Assess the relationship between mental illness and total victims, mental illness and location type, and the intersection of all three variables.
* I have already investigated the relationship between mental illness and total victims, mental illness and location type above as these seemed to be the best variables to analyse the difference in shootings based on history of mental illness. No other variables were particularly interesting. So, see above.
* Now, let’s examine the intersection between all three variables:
* Along with the above, it is clear that, across all locations, shooters with a history of mental illness harm far more victims.

# investigate this with bar chart with some colour for third dimension   
 mass\_shootings %>%   
   
# remove missing values to investigate prior mental illness   
 drop\_na(prior\_mental\_illness) %>%   
   
# plot stacked bars by location type against number of victims  
# for shooters with and without a history of mental illness   
 ggplot(aes(x = total\_victims, y = location\_type,   
 fill= prior\_mental\_illness)) +   
 geom\_bar(stat="identity") +  
   
 labs(title = "Relationship between shooting location types,  
 \nnumber of shooting victims  
 \nand the mental history of the shooter involved",  
 x = "Total number of shooting victims",   
 y = "Type of location",  
 fill = "History of mental illness?")



# Exploring credit card fraud

The data set we will use consists of credit card transactions and it includes information about each transaction including customer details, the merchant and category of purchase, and whether or not the transaction was a fraud.

* In this data set, how likely are fraudulent transactions? Generate a table that summarizes the number and frequency of fraudulent transactions per year. Approximately 0.6% of transactions were fraudulent in both 2019 and 2020.

# group transactions by year   
card\_fraud %>%   
 group\_by(trans\_year) %>%   
# count number of fraudulent and non- fraudulent transactions   
 count(is\_fraud) %>%   
# calculate variable for frequency of fraud   
 mutate(frequency\_percentage = n/sum(n) \*100) %>%   
   
# filter and display on data for fraudulent transactions in table  
 filter(is\_fraud==1)

# A tibble: 2 × 4  
# Groups: trans\_year [2]  
 trans\_year is\_fraud n frequency\_percentage  
 <dbl> <dbl> <int> <dbl>  
1 2019 1 2721 0.568  
2 2020 1 1215 0.632

* 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.

# group transactions by year and fraud status  
card\_fraud %>%   
 group\_by(trans\_year, is\_fraud) %>%  
   
# calculate total dollar amount for each fraud group and year   
 summarize(total\_amount\_dollars = sum(amt)) %>%   
   
# calculate percentage of total amount for each fraud group   
 mutate(percentage\_dollars =   
 total\_amount\_dollars/sum(total\_amount\_dollars)\*100)

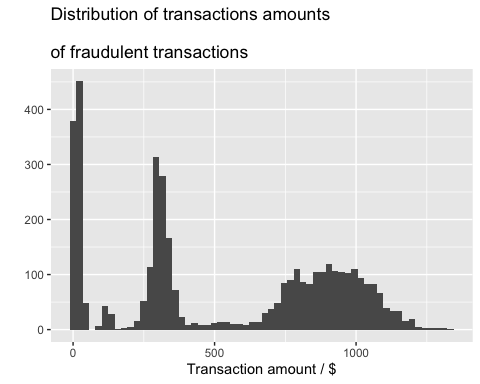
# A tibble: 4 × 4  
# Groups: trans\_year [2]  
 trans\_year is\_fraud total\_amount\_dollars percentage\_dollars  
 <dbl> <dbl> <dbl> <dbl>  
1 2019 0 32182901. 95.8   
2 2019 1 1423140. 4.23  
3 2020 0 12925914. 95.2   
4 2020 1 651949. 4.80

* 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.

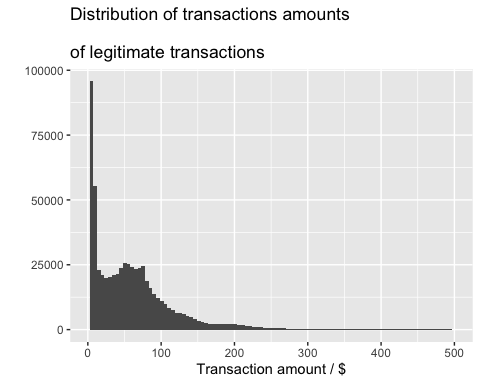
# some quick summary stats using summarize   
# group by fraud versus legitimate   
card\_fraud %>%   
 group\_by(is\_fraud) %>%   
# calculate summary stats of amounts for each  
 summarize(mean\_amount= mean(amt),   
 median\_amount = median(amt),   
 minimum\_amount = min(amt),   
 maximum\_amount = max(amt))

# A tibble: 2 × 5  
 is\_fraud mean\_amount median\_amount minimum\_amount maximum\_amount  
 <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0 67.6 47.2 1 27120.  
2 1 527. 369. 1.06 1334.

# now plot histograms   
card\_fraud %>%  
# first only fraudulent transactions  
 filter(is\_fraud==1) %>%   
# plot distribution - make bins look as good as possible   
 ggplot(aes(x = amt)) +  
 geom\_histogram(bins=60) +  
  
 labs(title = "Distribution of transactions amounts  
 \nof fraudulent transactions",   
 x = "Transaction amount / $", y="")

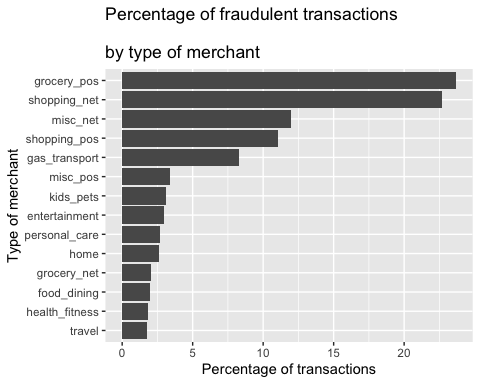


# now plot histograms   
card\_fraud %>%  
# next only legitimate transactions  
 filter(is\_fraud==0) %>%   
# plot distribution - make bins look as good as possible   
 ggplot(aes(x = amt)) +  
 geom\_histogram(bins=100) +   
# take out large amount outliers for this purpose so the graph looks decent   
 scale\_x\_continuous(limits= c(1,500)) +  
   
 labs(title = "Distribution of transactions amounts  
 \nof legitimate transactions",   
 x = "Transaction amount / $", y="")



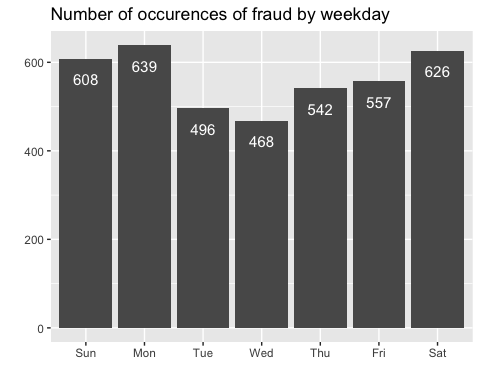
* What types of purchases are most likely to be instances of fraud? Consider category of merchants and produce a bar chart that shows % of total fraudulent transactions sorted in order.

card\_fraud %>%   
# filter for only farud transactions  
 filter(is\_fraud==1) %>%   
   
# count them by category and rearrange bars based on count  
 count(category) %>%   
 mutate(category = fct\_reorder(category,n)) %>%  
   
# calculate percentage of fraud by catgeory of merchant   
 mutate(percentage = n/sum(n)\*100) %>%   
   
# plot categories of fraud percentage frequency  
 ggplot(aes(y=category, x=percentage)) +   
 geom\_bar(stat = "identity") +  
 labs(title="Percentage of fraudulent transactions   
 \nby type of merchant",   
 x = "Percentage of transactions",  
 y = "Type of merchant")

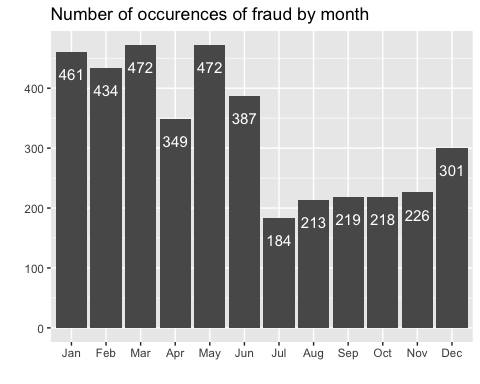


* When is fraud more prevalent? Which days, months, hours?
* Fraud seems to occur far more frequently during night time hours when people are asleep and more frequently on weekend days as these are the times during which people are paying less attention to their bank accounts. Furthermore, fraud is far more common in the first six months of the year than the last.

# use lubridate to create some new variables   
# to investigate which days/months/hours have increased fraud   
# and save them in a new data frame for plotting   
new\_card\_fraud <- 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)  
 )   
  
# plot bar graph to investigate fraud by day of the week   
new\_card\_fraud %>%   
   
# filter only for fraudulent transactions  
 filter(is\_fraud==1) %>%   
   
# count number of transactions of fraud by weekday   
 group\_by(weekday) %>%   
 count() %>%   
   
# plot bars for weekdays   
 ggplot(aes(x = weekday, y = n)) +   
 geom\_bar(stat="identity") +  
   
# add some text to show count of transactions for each day   
 geom\_text(aes(label = n, y= n - 10),   
 colour = "white", size = 4 , vjust = 2) +  
 labs(x='', y='',   
 title = "Number of occurences of fraud by weekday")



# plot bar graph to investigate fraud by month of the year   
new\_card\_fraud %>%   
   
# filter only for fraudulent transactions  
 filter(is\_fraud==1) %>%   
   
# count number of transactions of fraud by month   
 group\_by(month\_name) %>%   
 count() %>%   
   
# plot bars for months   
 ggplot(aes(x = month\_name, y = n)) +   
 geom\_bar(stat="identity") +  
   
# add some text to show count of transactions for each month   
 geom\_text(aes(label = n, y= n - 10),   
 colour = "white", size = 4 , vjust = 2) +  
 labs(x='', y='',   
 title = "Number of occurences of fraud by month")



# too many hours in a day to investigate nicely with a graph   
# make a table instead   
new\_card\_fraud %>%   
   
# filter only for fraudulent transactions  
 filter(is\_fraud==1) %>%   
   
# count number of transactions of fraud by hour of the day   
 group\_by(hour) %>%   
 count() %>%  
# arrange from most fraud occurences to least   
 arrange(desc(n))

# A tibble: 24 × 2  
# Groups: hour [24]  
 hour n  
 <int> <int>  
 1 23 1012  
 2 22 981  
 3 0 348  
 4 1 332  
 5 3 326  
 6 2 313  
 7 19 52  
 8 18 49  
 9 17 48  
10 13 45  
# ℹ 14 more rows

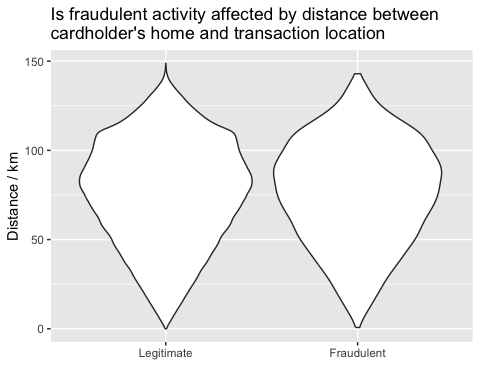
* Are older customers significantly more likely to be victims of credit card fraud? The table suggests that older customers are not more likely to be victims of credit card fraud than younger customers. All ages are affected.

card\_fraud %>%   
# use lubridate package to calculate customer age   
 mutate(age = interval(dob, trans\_date\_trans\_time)   
 / years(1),) %>%   
   
# create variable of age rounded to the nearest year  
 mutate(age\_years = round(age, digits=0)) %>%   
   
# group by rounded age and count from highest to lowest   
 group\_by(age\_years) %>%   
 count() %>%   
 arrange(desc(n))

# A tibble: 83 × 2  
# Groups: age\_years [83]  
 age\_years n  
 <dbl> <int>  
 1 47 20712  
 2 35 20011  
 3 34 18754  
 4 44 18391  
 5 32 18116  
 6 48 18090  
 7 33 16708  
 8 43 16509  
 9 29 15919  
10 46 15668  
# ℹ 73 more rows

* Is fraud related to distance? There seems to be (from violin plot) absolutely no relationship between the distance between the cardholders’ home and the transaction location and whether fraudulent activity occurs. This is not a useful data factor to explain fraud.

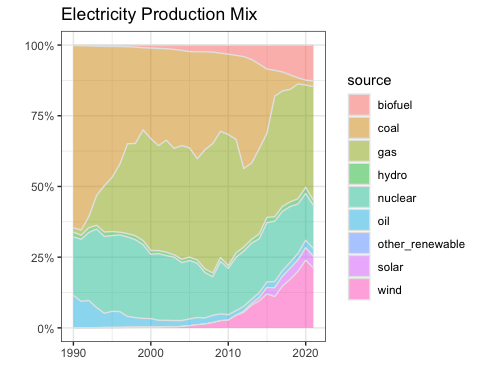
# distance between card holder's home and transaction  
# code adapted from https://www.geeksforgeeks.org/program-distance-two-points-earth/amp/  
  
fraud <- card\_fraud %>%  
 mutate(  
   
 # convert latitude/longitude to radians  
 lat1\_radians = lat / 57.29577951,  
 lat2\_radians = merch\_lat / 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)))   
  
  
# create new column to turn is\_fraud into characters   
fraud %>%   
 mutate(is\_fraud\_new= as.character(is\_fraud)) %>%   
   
# violin plot  
 ggplot(aes(x = is\_fraud\_new, y=distance\_km)) +  
 geom\_violin() +   
# label fraud categories   
 scale\_x\_discrete(labels= c("Legitimate","Fraudulent")) +  
   
 labs(x='', y = "Distance / km" ,   
 title = "Is fraudulent activity affected by distance between \ncardholder's home and transaction location")



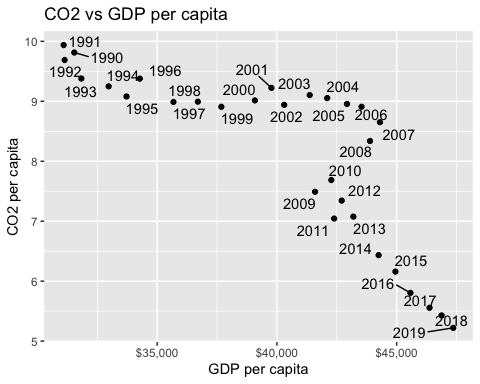
# Exploring sources of electricity production, CO2 emissions, and GDP per capita.

We will get energy data from the Our World in Data website, and CO2 and GDP per capita emissions from the World Bank.

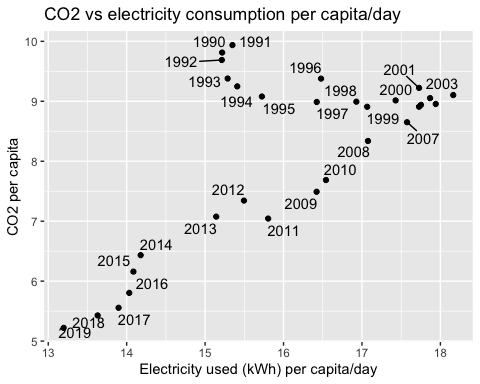
# Download electricity data  
url <- "https://nyc3.digitaloceanspaces.com/owid-public/data/energy/owid-energy-data.csv"  
  
energy <- read\_csv(url) %>%   
 filter(year >= 1990) %>%   
 drop\_na(iso\_code) %>%   
 select(1:3,  
 biofuel = biofuel\_electricity,  
 coal = coal\_electricity,  
 gas = gas\_electricity,  
 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  
 )   
  
   
# pivot longer to get the different categories of electricity source into one column and their %s into another  
new\_energy <- energy %>%  
 pivot\_longer(cols=4:12,  
 names\_to = "electricity\_source"  
 , values\_to = "percentage\_of\_source") %>%   
   
# filter for my country and remove missing values   
 filter(country == "United Kingdom") %>%   
 drop\_na(percentage\_of\_source) %>%   
   
# group by source of electricity   
 group\_by(electricity\_source) %>%   
  
# plot graph number 1 with year on x axis, % of source on y   
# use are to fill the different electricity sources   
 ggplot(aes(x =year, y= percentage\_of\_source,   
 fill=electricity\_source)) +   
 geom\_area(colour="grey90",  
 alpha = 0.5, position = "fill") +  
   
# fix labels according to example image   
# and get y axis in a % scale as in example image   
 labs(x='', y ='', title= "Electricity Production Mix", fill="source") + scale\_y\_continuous(labels = scales::percent) +  
 theme\_bw()  
   
new\_energy



#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)  
  
# left join the two tables by 3 digit iso code, year   
graph\_2\_data <- left\_join(x = gdp\_percap, y = co2\_percap, by = c("iso3c", "year")) %>%   
   
# select my country and remove missing values   
 filter(iso3c=="GBR") %>%   
 drop\_na(GDPpercap) %>%   
   
# scatterplot of gdppercap vs co2percap   
 ggplot(aes(x = GDPpercap, y= co2percap)) +  
 geom\_point() +  
# fix labels as in example figure   
 labs(title="CO2 vs GDP per capita",   
 x = "GDP per capita", y= "CO2 per capita") +   
# label the data points with year   
 geom\_text\_repel(aes(label= year)) +   
# fix x axis scale to dollars as in example figure   
 scale\_x\_continuous(labels = scales::dollar)  
   
graph\_2\_data



# left join the two tables by country, year   
graph\_3\_data <- left\_join(x = co2\_percap, y = energy,  
 by = c("country", "year")) %>%   
   
# select my country and remove missing values   
 filter(country=="United Kingdom") %>%   
 drop\_na(co2percap) %>%   
   
# mutate new column for elec used / capita / day   
 mutate(elec\_cap\_day = per\_capita\_electricity/365) %>%   
   
# scatterplot of elec used / capita / day vs co2percap   
 ggplot(aes(x = elec\_cap\_day, y= co2percap)) +  
 geom\_point() +  
   
# fix labels as in example figure   
 labs(title="CO2 vs electricity consumption per capita/day",   
 x = "Electricity used (kWh) per capita/day",   
 y= "CO2 per capita") +   
# label the data points with year   
 geom\_text\_repel(aes(label= year))   
   
graph\_3\_data



# import patchwork   
library(patchwork)  
  
# stitch plots together   
new\_energy / (graph\_2\_data | graph\_3\_data)

