

STAT 201: Midterm 1 Extra Coding Practice 1

Possible Solutions

Load your packages necessary for wrangling code, plotting, and creating beautiful tables:

The data containing the records of gun violence incidents are found in `gun_violence`. The data are obtained from Kaggle, and have been slightly modified. The variable definitions are as follows:

- `incident_id`: ID number
- `date`: date of incident
- `state`: state in which incident occurred
- `city_or_county`: city or county in which incident occurred
- `address`: address where incident occurred
- `n_killed`: number of people killed
- `n_injured`: number of people injured, but not killed
- `n_guns_involved`: number of guns involved
- `doy`: day of year (month/day) of incident
- `year`: year
- `month`: month

Question 1

In which five cities or counties did the most incidents occur? What are the probabilities that a randomly selected incident in our dataset occurred in each of these cities/counties? Answer this by creating a beautiful table that contains this information.

```
gun_violence |>
  count(city_or_county) |>
  arrange(-n) |>
  mutate(prop = n / sum(n)) |>
  slice(1:5) |>
  kable()
```

city_or_county	n	prop
Chicago	10298	0.0462661
Baltimore	3703	0.0166366
New Orleans	2939	0.0132041
Philadelphia	2824	0.0126875
Jacksonville	2372	0.0106567

Question 2

While there are a variety of definitions of mass shootings, one definition is “any incidents in which four or more people were shot, whether injured or killed”.

Based on this definition, create a beautiful table that displays the top six incidents that meet this definition of mass shooting in descending order of number of people shot. Only display the date, state, city or county, and number of people.

```
gun_violence |>
  mutate(n_shot = n_killed + n_injured) |>
  arrange(-n_shot) |>
  slice(1:6) |>
  select(date, state, city_or_county, n_shot) |>
  kable()
```

date	state	city_or_county	n_shot
2016-06-12	Florida	Orlando	103
2017-11-05	Texas	Sutherland Springs	47
2015-12-02	California	San Bernardino	35
2015-05-17	Texas	Waco	27
2017-07-01	Arkansas	Little Rock	25
2016-07-25	Florida	Fort Myers	21

Question 3

Create a data frame called **state_incidents** where each row is an observation of a state and its total number of incidents in a given year.

Then using your new data frame, answer the following using a single table: For the combined years 2014-2017, which three states had the lowest average number of incidents? Which three states had the highest average number of incidents?

```
state_incidents <- gun_violence |>
  count(state, year)
```

```
state_incidents |>
  group_by(state) |>
  summarise(avg_n = mean(n)) |>
  arrange(-avg_n) |>
  slice(c(1:3, 48:50)) |>
  kable()
```

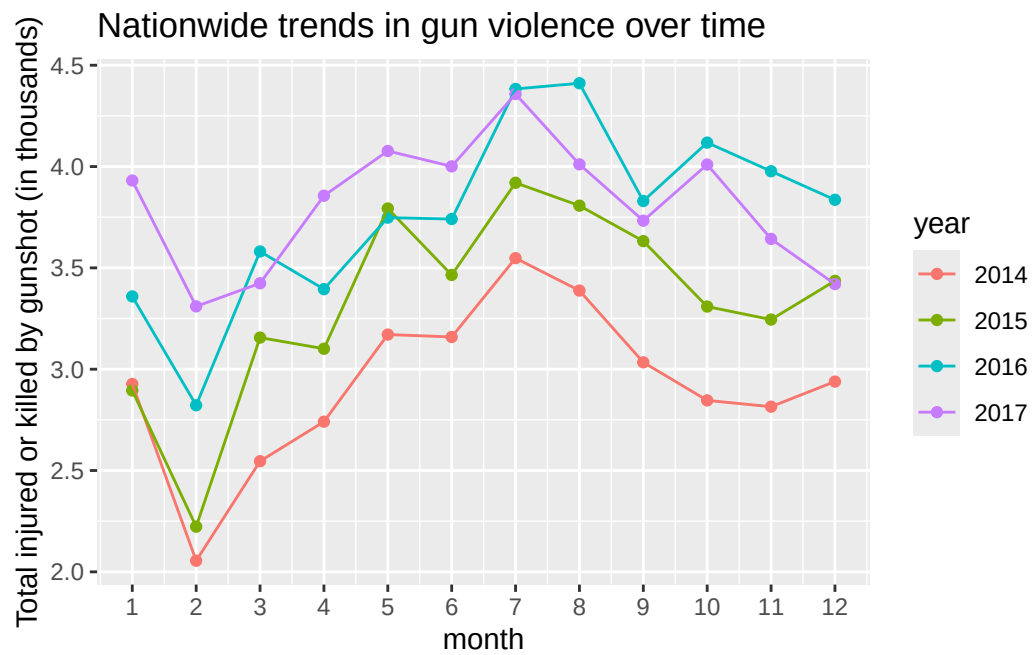
state	avg_n
Illinois	4173.75
California	3792.75
Florida	3549.25
Wyoming	120.00
Vermont	113.00
Hawaii	67.50

Question 4

Re-create the following plot to the best of your abilities. The following code will be helpful to you:

- If you'd like to turn a numeric variable into a character variable, you can use the `as.character(<variable>)` function. If you'd like to turn a character variable into a numeric variable, you can use the `as.numeric(<variable>)` function.
- `scale_x_continuous(breaks = 1:12)` makes a numeric x-axis have tick marks at the values 1-12.

```
gun_violence |>
  mutate(month = as.numeric(month)) |>
  mutate(shot = n_injured+n_killed) |>
  group_by(month, year) |>
  summarise(shot = sum(shot)/1000) |>
  ggplot(aes(x = month, y = shot, col = year)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = 1:12) +
  labs(title = "Nationwide trends in gun violence over time",
       y = "Total injured or killed by gunshot (in thousands)")
```



Then interpret what you see.