STAT 201: Midterm 1 Practice

```
# load your libraries here for wrangling, plotting, and making pretty tables
library(readr)
library(tidyverse)
library(kableExtra)
```

Dataset 1 (Data retrieved from the R pell package.)

A Pell Grant is a need-based federal grant for undergraduate students paying for college. Students are automatically considered for a Pell Grant award when they submit a FAFSA.

Each case in the dataset `pell` represents the Pell Grant award to a given school in a given year.

- `state`: state/territory shortcode
- `award`: total award amount in USD
- 'recipient": total number of recipients
- `name`: name of college/university
- `session`: meeting ID
- `year`: year

We also have a dataset called `states` that provide information on the states and territories of the US. The variables are:

- `NAME`: name of the state/territory
- `Abbreviation`: state/territory shortcode
- `Type`: political devision ("state" or "territory")

Run the following code chunk to load in the two datasets:

```
pell <- read_csv("https://raw.githubusercontent.com/midd-stat201-spring2025/midd-stat201-s
states <- read_csv("https://raw.githubusercontent.com/midd-stat201-spring2025/midd-stat201
```

Exercise 1

Let's clean and wrangle the data a bit.

- Remove any case where at least one of the following is true:
  - The school is missing information about the award amount
  - The school had less than 1 student receiving a Pell Grant
  - The school receive an award of 0 dollars but had more than 0 students recieve a Pell grant
- Create a new variable called `award_pp` that represents the award amount per person for each school
- Create a new variable called `decade` that takes the value:
  - "1990s" if the award was granted in the 1990s
  - "2000s" if the award was granted in the 2000s
  - "2010s" if the award was granted in the 2010s

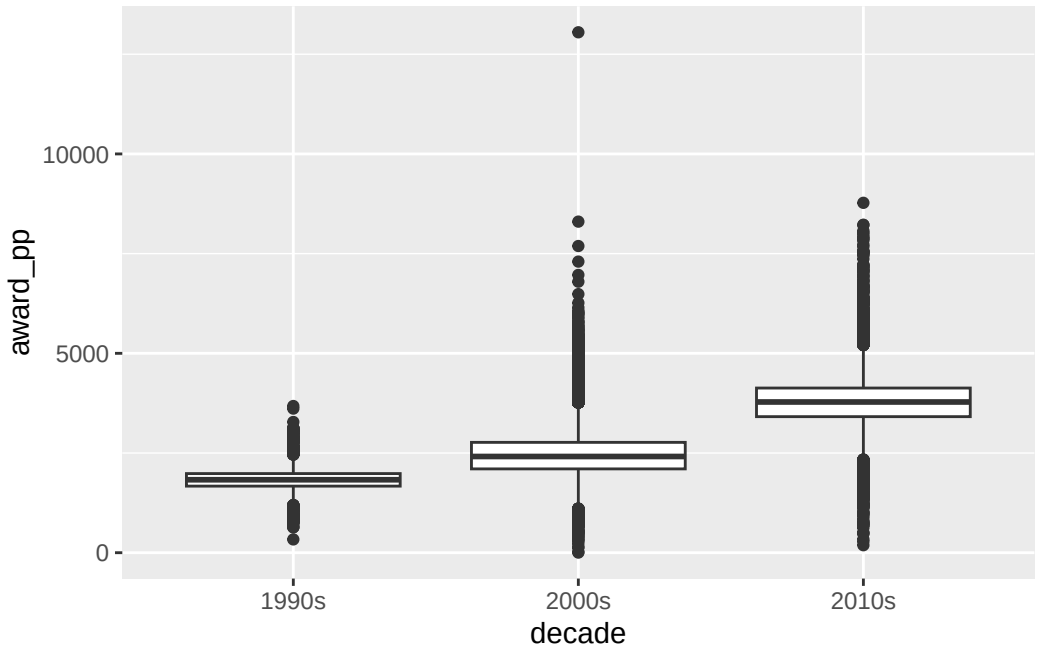
Store your cleaned and wrangled data frame back into `pell`.

```
pell <- pell |>
  filter(!is.na(award)) |>
  filter(!(award == 0 & recipient > 0)) |>
  filter(recipient > 0)
pell <- pell |>
  mutate(decade = case_when(
    year < 2000 ~ "1990s",
    year >= 2000 & year < 2010 ~ "2000s",
    T ~ "2010s"
  )) |>
  mutate(award_pp = award/recipient)
```

Exercise 2

Make a visualization the shows the distribution of the award amount per person for each decade. Have informative labels and titles. Interpret what you see.

```
pell |>
  ggplot(aes(x  = decade, y = award_pp)) +
  geom_boxplot()
```



Answer:

Exercise 3

Now let's bring in the information about states vs territories. Combine the two datasets into a single dataset called `pell2` that retains all observations about pell grants awards for which we also have information about the type of political division (i.e. state or territory).

```
pell2 <- pell |>
  inner_join(states, by = c("state" = "Abbreviation"))
```

Exercise 4

For awards made in 2015, create a beautiful table that displays the mean and standard deviation of the award per person for each of the two political divisions.

```
pell2 |>
  filter(year == 2015) |>
  group_by(Type) |>
  summarise(avg_award_pp = mean(award_pp), sd_award_pp = sd(award_pp)) |>
  kable()
```

Type	avg_award_pp	sd_award_pp
State	3763.108	548.9417
Territory	4133.444	510.0271

Exercise 5

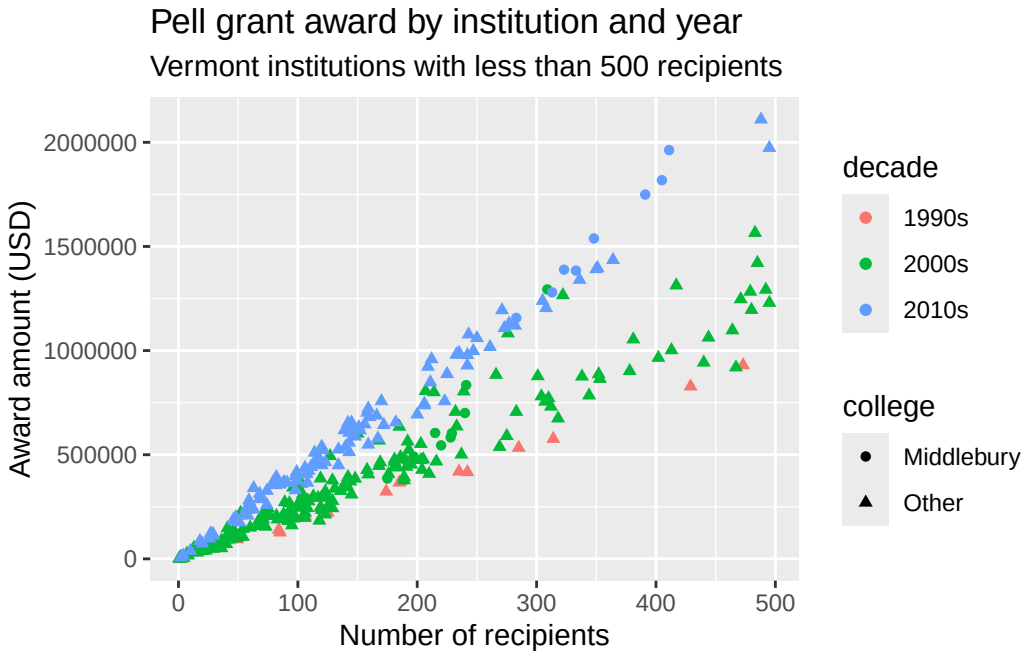
Display a beautiful table of the five U.S. states that received the largest statewide average award per person in the 2010s.

```
pell2 |>
  filter(Type == "State", decade == "2010s") |>
  group_by(Name)  |>
  summarise(state_avg = mean(award_pp)) |>
  ungroup() |>
  arrange(-state_avg) |>
  slice(1:5)
```

```
# A tibble: 5 x 2
  Name          state_avg
  <chr>             <dbl>
1 Mississippi       4062.
2 New York          4030.
3 Idaho             3996.
4 West Virginia     3988.
5 Vermont           3967.
```

Exercise 6

Re-create the following plot:



Dataset 2 (Data retrieved from Kaggle)

We have data from over 1000 wines in the dataset `wine_ratings`. Each case in the dataset represents one bottle of wine. The wines included in the dataset are wines that were tasted and reviewed.

- `country`: country of origin
- `variety`: grape type
- `points`: the number of points WineEnthusiast rated the wine on a scale of 1-100 (100 best)
- `price`: price of the wine (USD)
- `title`: title of the wine review, which often contains the year of the wine
- `taster_name`: name of taster/reviewer
- `description`: flavor and taste profile as written by the reviewer
- `year`: year/vintage of the wine, if available in the title of the review

Run the following code to load in the data and take a look at it before continuing!

```
wine_ratings <- readr::read_csv("https://raw.githubusercontent.com/midd-stat201-spring2025
```

Exercise 1

First, find the three countries that are most represented in the dataset. Once you have identified those countries, create a new data frame called `wine_ratings_top` that only retains observations from those four countries.

```
wine_ratings |>
  count(country) |>
  arrange(-n) |>
  slice(1:3)
```

```
# A tibble: 3 x 2
  country     n
  <chr>   <int>
1 US        534
2 France    221
3 Italy     189
```

```
wine_ratings_top <- wine_ratings |>
  filter(country %in% c("US", "France", "Italy"))
```

Exercise 2

Obtain the mean and standard deviation of the points and prices of the wines in each of the four countries using the observations for which we have information about these variables. Display as a beautiful table .

```
wine_ratings_top |>
  group_by(country) |>
  # na.omit() |>
  filter(!is.na(price)) |>
  summarise(mean_pts = mean(points), mean_price = mean(price), sd_points = sd(points), sd_
```

```
# A tibble: 3 x 5
  country mean_pts mean_price sd_points sd_price
  <chr>      <dbl>      <dbl>     <dbl>    <dbl>
1 France      88.7       48.0      3.23    104.
2 Italy       88.3       44.1      2.74     41.7
3 US          88.4       35.6      3.15     25.7
```

Exercise 3

For each of the three countries identified earlier, what proportion of their wines that were produced before 2010 received over 90 points? Display a beautiful table that only shows the country name and the proportion.

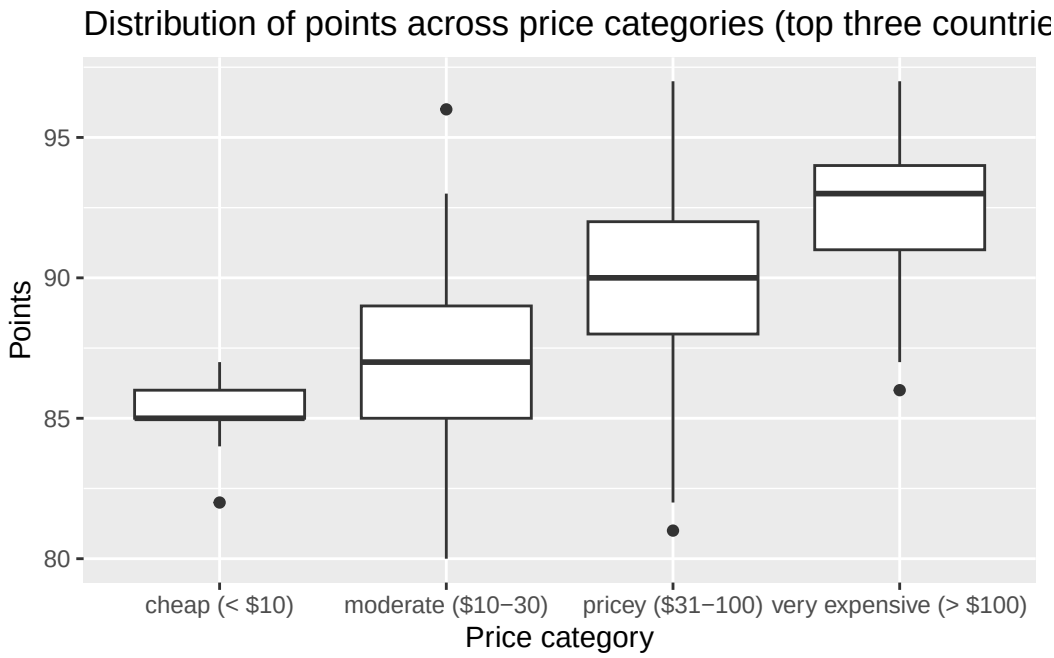
```
wine_ratings_top |>
  filter(year < 2010) |>
  mutate(over90 = if_else(points > 90, T, F)) |>
  count(over90, country) |>
  group_by(country) |>
  mutate(prop = n/sum(n)) |>
  ungroup() |>
  filter(over90 == TRUE) |>
  select(country, prop) |>
  kable()
```

country	prop
France	0.4166667
Italy	0.3000000
US	0.2055556

Exercise 4

Re-create the following plot and interpret it:

```
wine_ratings_top |>
  filter(!is.na(price)) |>
  mutate(price_cat = case_when(
    price < 10 ~ "cheap (< $10)",
    price >= 10 & price <= 30 ~ "moderate ($10-30)",
    price > 30 & price <= 100 ~ "pricey ($31-100)",
    price > 100 ~ "very expensive (> $100)"
  )) |>
  ggplot(aes(x = price_cat, y = points)) +
  geom_boxplot() +
  labs(x = "Price category", y = "Points", title = "Distribution of points across price ca
```


Distribution of points across price categories (top three countrie
Points
95
90
85
80
cheap (< $10)
moderate ($10−30)
pricey ($31−100)
very expensive (> $100)
Price category