STAT 201: Midterm 1 Practice 3

Possible solutions

In the first code chunk, load your libraries for wrangling, plotting, and making pretty tables. Run the code chunk to also load in the data (described below) and take a look at it before continuing!

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

Exercise 1

First, modify the `wine_ratings` data to remove observations for which we don't have information about price.

Then 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 three countries.

```
wine_ratings <- wine_ratings |>
  filter(!is.na(price))
# see which ones
wine_ratings |>
  count(country) |>
  arrange(-n) |>
  slice(1:3)
```

```
# A tibble: 3 x 2
  country     n
  <chr>   <int>
1 US        532
2 France    178
3 Italy     152
```

```
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 for each of the three countries. Display as an informative, beautiful table. Briefly interpret what you see.

```
wine_ratings_top |>
  group_by(country) |>
  summarise(mean_pts = mean(points), mean_price = mean(price),
            sd_points = sd(points), sd_price = sd(price)) |>
  kable()
```

country	mean_pts	mean_price	sd_points	sd_price
France	88.69663	47.95506	3.227371	103.75807
Italy	88.31579	44.09868	2.743858	41.70327
US	88.37030	35.64098	3.150957	25.72696

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.

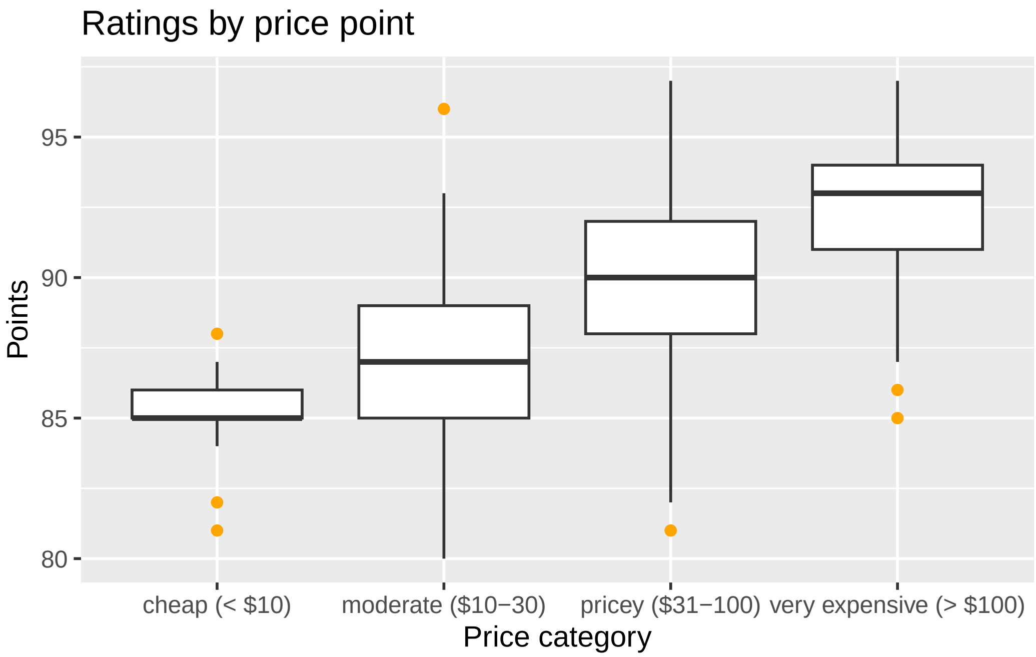
```r
wine_ratings_top |>
  filter(year < 2010) |>
  mutate(over90 = if_else(points > 90, T, F)) |>
  count(over90, country) |>
  group_by(country) |>
  mutate(prop = n/sum(n)) |>
  ungroup() |>
  filter(over90 == TRUE) |>
  select(country, prop) |>
  kable()
```

country	prop
France	0.3947368
Italy	0.2909091
US	0.2055556

Exercise 4

Re-create the following plot and interpret it:

```r
wine_ratings |>
  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(outlier.color = "orange") +
  labs(x = "Price category", y = "Points", title = "Ratings by price point")
```



Exercise 5

This one is a bit difficult!

Among all wines in the original dataset that have at least 25 reviews and information about prices, determine which type of wine grape seems to be the worst value. In your answer, briefly describe your methods/reasoning and identify the wine variety.

```
wine_ratings |>
  mutate(point_price = points/price) |>
  group_by(variety) |>
  summarise(mean = mean(point_price), n = n()) |>
  arrange(-n) |>
  filter(n >= 25) |>
  arrange(mean) |>
  slice(1)
```

```
# A tibble: 1 x 3
  variety      mean     n
  <chr>       <dbl> <int>
1 Pinot Noir   2.53   117
```