Data Challenge 05

Jake Moore

Objective

In this challenge, you will analyze Olympic medal results across regions and years using tidy summaries and faceted plots. You will:

1. join event data with NOC regions

2. compute group counts and identify "top" groups

3. create a multi-series line plot for the top 6 regions since 1990

4. create faceted trend plots for the top 3 regions by medal type

5. add brief interpretation notes where prompted

Data (same folder as this Rmd):

- `olympic_history_athlete_events.csv`

- `olympic_history_noc_regions.csv`

Load & Tidy Data

1) Load the Olympic data

Keep the exact code. Do **not** modify file names or object names. Ensure your data is located in the corresponding location

```
olympic_events_raw <- read_csv("olympic_history_athlete_events.csv")
noc_regions <- read_csv("olympic_history_noc_regions.csv")
```

2) Explore and describe data (4 pts)

Explore the `olympic_events_raw` and `noc_regions` dataframes. Describe the contents of each dataframe in one-sentence summaries. Your description should answer: "What is this data?"

```
olympic_events_raw_desc <- 'Contains 271,116 Olympic records detailing each athletes participation, dem
noc_regions_desc <- 'Contains 206 NOC codes to standardize and categorize countries and regions'
```

3) Create olympic_events (6 pts)

Create a cleaned dataframe called `olympic_events` using a tidyverse pipe that: 1. starts with `olympic_events_raw` 2. (2 pts) adds `region` by left joining `noc_regions` with `NOC` as by key. 3. (2 pts) ensures all column names are lowercase (hint: use a rename function) 4. (2 pts) drops any rows where `medal` is missing (this is acceptable as we know this means the athlete did not receive a medal which is not information we're interested in for this analysis)

```
olympic_events <- olympic_events_raw %>%
  left_join(noc_regions, by = 'NOC') %>%
  rename_with(tolower) %>%
  filter(!is.na(medal))
```

Success checks:

```
names(olympic_events)
```

```
##  [1] "id"     "name"   "sex"    "age"    "height" "weight" "team"   "noc"
##  [9] "games"  "year"   "season" "city"   "sport"  "event"  "medal"  "region"
## [17] "notes"
```

```
sum(is.na(olympic_events$medal))==0
```

```
## [1] TRUE
```

```
sum(is.na(olympic_events$region))==0
```

```
## [1] TRUE
```

```
ncol(olympic_events)==(ncol(olympic_events_raw)+ncol(noc_regions%>%select(-region)))
```

```
## [1] TRUE
```

Answer Data Questions

In this analysis, we will set out to answer three question: 1. What region has the highest number of Gold, Silver, and Bronze medals in **Summer** Olympics? 2. From the top six countries, who has gotten more medals per year since 1990 in the **Summer** Olympics? 3. From the top three countries, who has gotten more medals segmented by type (gold, silver, and bronze) per year since 1990 in the **Summer** Olympics?

4) DQ1 — Most Summer medals by type and region (15 pts)

1. (10 pts) create a 326 × 3 tibble called `region_medal_count`, mapping the count of each medal in each region, that looks like this:

```r
region_medal_count <- olympic_events %>%
  filter(season == "Summer") %>%
  count(region, medal, sort=T)
```

2. (5 pts) define `p1_reg_most_gold`, `p1_reg_most_silver`, and `p1_reg_most_bronze` as the single character value of the region that had the mode Gold, Silver, and Bronze Medals, respectively.

Hint: all three are “USA”!

```r
p1_reg_most_gold <- region_medal_count %>%
  filter(medal == "Gold") %>%
  top_n(1) %>%
  .$region
p1_reg_most_silver <- region_medal_count %>%
  filter(medal == "Silver") %>%
  top_n(1) %>%
  .$region
p1_reg_most_bronze <- region_medal_count %>%
  filter(medal == "Bronze") %>%
  top_n(1) %>%
  .$region
```

Inline winners:

- Gold: USA
- Silver: USA
- Bronze: USA

5) DQ2 — Top 6 regions: Summer medals per year since 1990 (30 pts)

1. (5 pts) identify the top 6 regions by all-time total Summer medals, stored as a sorted 6 × 2 tibble `top_6`, that looks like:

```r
top_6 <- olympic_events %>%
  filter(season == "Summer") %>%
  count(region, sort=T)  %>%
  top_n(6)
```

2. (5 pts) filter the `olympic_events` to Summer events for countries in the `top_6` after 1990 (inclusive) and it store as `olympic_top6`.

```r
olympic_top6 <- olympic_events %>%
  filter(season == "Summer",
         year >= 1990,
         region %in% top_6$region)
```

Success Checks

```r
min(olympic_top6$year) == 1992
```

```
## [1] TRUE
```

```r
unique(olympic_top6$season) == "Summer"
```

```
## [1] TRUE
```

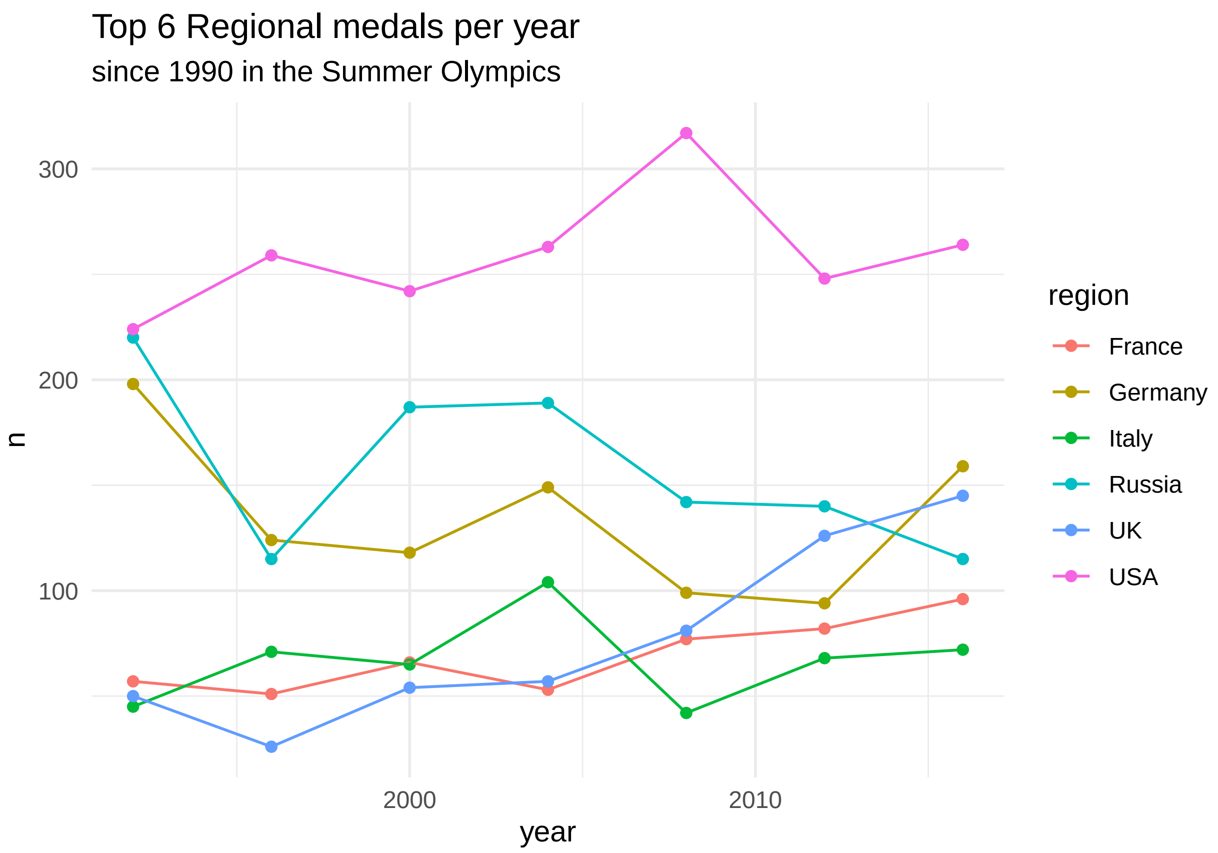
```r
all(unique(olympic_top6$region) == c("Italy","Russia","France","USA","UK","Germany"))
```

```
## [1] TRUE
```

3. (20 pts) plot a timeseries of the yearly medal counts since 1990 for those regions with the following elements:

    1. (10 pts) time series as scatter plot with overlayed lines
    2. (5 pts) maps `x = year` and `y = n`, using `region` to color each line
    3. (2 pts) title of `"Top 6 Regional medals per year"`
    4. (2 pts) subtitle of `"since 1990 in the Summer Olympics"`
    5. (1 pt) minimal theme

```r
p2_regional_line <- olympic_top6 %>%
  count(year, region) %>%
  ggplot(aes(x = year, y = n, color = region)) +
  geom_point() +
  geom_line() +
  ggtitle("Top 6 Regional medals per year","since 1990 in the Summer Olympics") +
  theme_minimal()
p2_regional_line
```



Short interpretation (1–2 sentences) → `q2_note`. (not graded, recommended)

```
q2_note <- 'The visualization above shows that the United Statesd has led all regions in total Summer Ol
```

6) DQ3 — Top 3 by medal type, faceted (45 pts)

1. (5 pts) identify the top 3 regions by all-time total Summer medals, stored as a sorted 3 × 2 tibble `top_3`, that looks like:

```
  region      n
  <chr>   <int>
1 USA      5002
2 Russia   3188
3 Germany  3126
```

```
top_3 <- olympic_events %>%
  filter(season == "Summer") %>%
  count(region, sort=T)  %>%
  top_n(3)
```

2. (5 pts) filter the `olympic_events` to Summer events for countries in the `top_3` after 1990 (inclusive) and it store as `olympic_top3`.

Success checks:

- `min(olympic_top3$year) == 1992`
- `unique(olympic_top3$season) == "Summer"`
- `all(unique(olympic_top3$region) == c("Russia","USA","Germany"))`

```
olympic_top3 <- olympic_events %>%
  filter(season == "Summer",
         year >= 1990,
         region %in% top_3$region)
```

Success checks

```
min(olympic_top3$year) == 1992
```

```
## [1] TRUE
```

```
unique(olympic_top3$season) == "Summer"
```

```
## [1] TRUE
```

```
all(unique(olympic_top3$region) == c("Russia","USA","Germany"))
```

```
## [1] TRUE
```

3. (10 pts) convert columns `medal` and `region` into a leveled factor type. We should order medals as `c("Gold","Silver","Bronze")` and regions as `top_3$region`. Save as `olympic_top3_ordered`

```
olympic_top3_ordered <- olympic_top3 %>%
  mutate(medal = factor(medal, levels = c("Gold","Silver","Bronze")),
         region = factor(region, levels = top_3$region))
```

Success checks

```
summary(olympic_top3_ordered$medal)
```

```
##   Gold Silver Bronze
##   1624   1033   1209
```

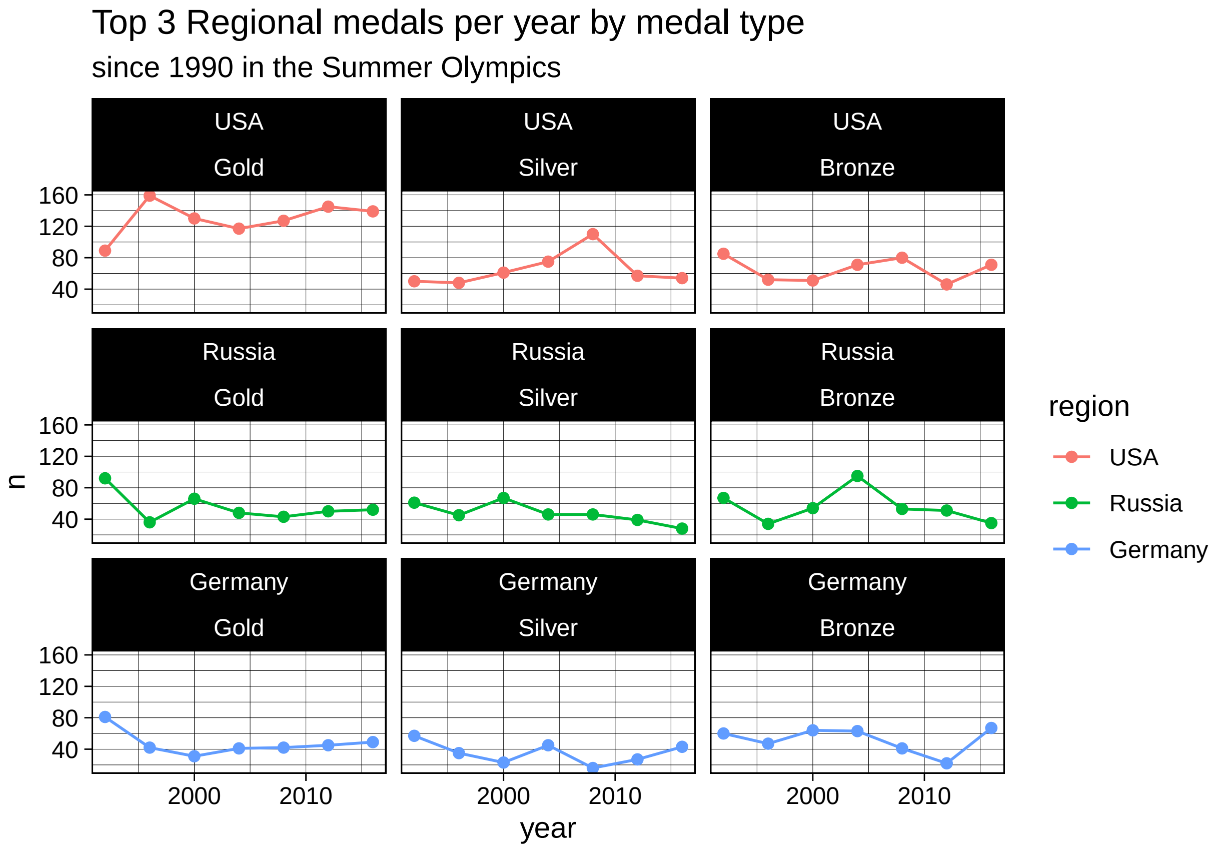
```
summary(olympic_top3_ordered$region)
```

```
##     USA  Russia Germany
##    1817    1108     941
```

4. (25 pts) using `olympic_top3_ordered`, plot timeseries subplots of the yearly medal counts since 1990 with the following elements:

1. (5 pts) scatter plots with overlayed lines
2. (5 pts) maps `x = year` and `y = n`, using `region` to color each line
3. (10 pts) split into subplots of `ncol=3` with axes y-axis as `medal` and x-axis as `region` (`region~medal`)
4. (2 pts) title of `"Top 3 Regional medals per year by medal type"`
5. (2 pts) subtitle of `"since 1990 in the Summer Olympics"`
6. (1 pt) "linedraw" theme

```
p3_line_subplots_ordered <- olympic_top3_ordered %>%
  count(year, region, medal) %>%
ggplot(aes(x = year, y = n, color = region)) +
  geom_point() +
  geom_line() +
ggtitle("Top 3 Regional medals per year by medal type","since 1990 in the Summer Olympics") +
  facet_wrap(region~medal, ncol=3) +
  theme_linedraw()
p3_line_subplots_ordered
```



Short interpretation (1–2 sentences) → `q3_note`. (not graded, recommended)

```
q3_note <- 'The visualization above shows that the United States consistently wins the most medals acro
```