```
---
output:
  pdf_document: default
  html_document: default
---
title: "Assignment 1"
author: "2316631"
output: pdf_document


  # Section 1: Warm-up

  ## Question 1
  ### Explanation
  The following code loads the required packages, downloads the 2024 UK General Election data,
and prepares it for analysis.

```{r setup, message=FALSE}
library(stringr)
library(readr)
library(ggplot2)
library(curl)
library(tidyverse)

uk_csv <- "https://electionresults.parliament.uk/general-elections/6/candidacies.csv"
uk_raw <- readr::read_csv(uk_csv, show_col_types = FALSE)
names(uk_raw) <- names(uk_raw) %>% tolower() %>% str_replace_all("\\s+", "_")

uk1 <- uk_raw %>% filter(!`election_is_by-election`)
uk2 <- uk1 %>% select(
  constituency = constituency_name,
  country = country_name,
  valid_votes = candidate_vote_count,
  electorate = electorate
)
uk3 <- uk2 %>%
  group_by(country, constituency) %>%
  summarise(
    valid_votes = sum(valid_votes, na.rm = TRUE),
    electorate = max(electorate, na.rm = TRUE),
    .groups = "drop"
  )

uk <- uk3 %>%
  mutate(turnout_registered = 100 * valid_votes / electorate) %>%
  filter(is.finite(turnout_registered), electorate > 0)

#  This code shows a very simplified version of the data.

glimpse(uk)

#=============================================================================
# Section 1: Data Wrangling and Organisation
#=============================================================================


# This code is finding the 10 constituencies with the highest turnout (top10)
# and the 10 with the lowest turnout (bottom10).

top10 <- uk %>% arrange(desc(turnout_registered)) %>% slice_head(n = 10)
bottom10 <- uk %>% arrange(turnout_registered) %>% slice_head(n = 10)

top10
bottom10

#=============================================================================
#=============================================================================
# Question 2: What patterns do you notice in high- vs low-turnout places?
# ANSWER
```
```

```
# High-turnout constituencies tend to be smaller towns or suburban areas, while
# low-turnout constituencies are mostly large cities. The same pattern is clear
# when comparing just the top and bottom 10.


#Now explore the top and bottom 25 places? Would you draw the same conclusions?
#Why or why not?
#ANSWER
# Yes, the same conclusions would likely hold. Expanding to the top and bottom 25
# would still show that smaller or suburban areas have higher turnout, while
# larger city constituencies tend to have lower turnout, because people in smaller
# communities may feel more directly connected to local politics, while urban areas
# often experience lower engagement and voter fatigue.

#==============================================================================
#==============================================================================
# Aggregating/Making Groups

# What about if we look across England, Wales, Scotland and NI?


# Step 7: Turnout by Constituent Countries of the UK

# 1) Country summary (mean, median, n), ordered by mean turnout

turnout_by_country <- uk %>%
  group_by(country) %>%
  summarise(
    mean_turnout   = mean(turnout_registered, na.rm = TRUE),
    median_turnout = median(turnout_registered, na.rm = TRUE),
    n              = n(),
    .groups = "drop"
  ) %>%
  arrange(desc(mean_turnout))

turnout_by_country
```

Country	Mean Turnout	Median Turnout	n
England	60.1	60.7	543
Scotland	59.2	58.6	57
Northern Ireland	57.1	58.2	18
Wales	56.2	56.8	32

```
#Question 3: What is the n here? Why might that matter?
# ANSWER
# The “n” shows how many constituencies are in each country for instance, 543 in
# England, 57 in Scotland, 18 in Northern Ireland, and 32 in Wales.
# This matters because countries with more constituencies, like England, have a
# larger influence on overall averages and more variation within their data, while
# smaller samples (like Northern Ireland) can make averages less reliable or more
# sensitive to outliers.

#==============================================================================
# Section 1: Data Visualisation
#==============================================================================

ggplot(turnout_by_country, aes(country, mean_turnout)) +
  geom_col(width = 0.7, fill = "#0072B2") +  # bar = mean turnout
  scale_y_continuous(limits = c(0, 100),
                     expand = expansion(mult = c(0, 0.05))) +
  labs(
    title = "UK 2024 General Election: Mean Turnout by Country",
    x = "Country",
    y = "Turnout (% of registered voters)"
  ) +
  theme_classic()
```

```
###ADD BLUE TABLE HERE###

#==============================================================================
#==============================================================================

#Question 4: Which country shows the highest median turnout? What is one plausible reason for
this?
# ANSWER
# It isn't easy to tell but England shows the highest median turnout.
# A plausible reason is that England has many smaller suburban constituencies where
# voters may feel a stronger connection to local issues and candidates, leading to
# slightly higher participation compared to larger or more urban areas in other countries.

#Question 5: Make 2 changes to the visualisation in order to improve it? What are the changes
#and why do they improve the figure? Document 2 approaches that didn't work and explain why
they didn't work also?
#Make sure that you add the visuals to your code so I can see.
# ANSWER
# Two changes that would improve the visualisation are adding data labels on top
# of each bar and reordering the bars by turnout instead of listing countries
# alphabetically. Adding labels would make it easier to read the exact values
# without relying on the y-axis, while ordering the bars by turnout would make
# the differences between countries clearer.
#==============================================================================
#==============================================================================

#==============================================================================
# Section 2, EXERCISE 1: "The Intergenerational Transmission of Advantage"
# Story: How does parental background shape children's opportunities?
#==============================================================================

kidiq <- read_csv("kidiq-2.csv") %>%
  mutate(
    work_status = case_when(
      mom_work == 1 ~ "Not working",
      mom_work == 2 ~ "Part-time",
      mom_work == 3 ~ "Full-time",
      mom_work == 4 ~ "Full-time+"
    ),
    family_advantage = case_when(
      mom_hs == 0 & mom_work <= 2 ~ "Low",
      mom_hs == 0 & mom_work > 2 ~ "Mixed",
      mom_hs == 1 & mom_work <= 2 ~ "Mixed",
      mom_hs == 1 & mom_work > 2 ~ "High"
    ),
    iq_group = case_when(
      mom_iq <= quantile(mom_iq, 0.33, na.rm = TRUE) ~ "Lower IQ",
      mom_iq <= quantile(mom_iq, 0.67, na.rm = TRUE) ~ "Middle IQ",
      TRUE ~ "Higher IQ"
    )
  )

ggplot(kidiq, aes(x = family_advantage, y = kid_score, fill = iq_group)) +
  geom_boxplot(alpha = 0.8) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "How Family Advantage Shapes Children's Test Scores",
    subtitle = "Children of higher-IQ, educated, and working mothers score higher on average",
    x = "Family Advantage",
    y = "Child's Test Score",
    fill = "Mother's IQ Group"
  ) +
  theme_minimal(base_size = 13)

### ADD HISTOGRAM HERE ###


## YOUR TURN: Create your visualization here
kidiq <- kidiq %>%
```

```
  mutate(
    mom_hs = factor(mom_hs, labels = c("Did not complete HS", "Completed HS"))
  )

ggplot(kidiq, aes(x = mom_iq, y = kid_score, color = mom_hs)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Children's Test Scores: Merit vs Structure",
    subtitle = "Mother's IQ predicts performance, but completing high school raises outcomes
further",
    x = "Mother's IQ (proxy for merit)",
    y = "Child's Test Score",
    color = "Mother's Education"
  ) +
  scale_color_brewer(palette = "Set2") +
  theme_minimal(base_size = 13)

###### add merit vs STRUCTURE HERE #######
```