````
---
title: "hw1"
author: "Jiahui Yang"
date: "2024-02-13"
output: html_document

---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
library(ggplot2)
library(dplyr)
library(tidyr)
```

```{r}
df <- readRDS("C:/Users/Admin/Desktop/HCMST_couples.rds")
meet <- distinct(select(df, meeting_type))
options(max.print = 10000)
meet
names(df)
```
```{r}
#Question 1:Dating trends over time

meeting_category <- c("School", "Social", "Work", "Online","Other")
df_simplified <- df %>%
  mutate(meeting_category = case_when(
    meeting_type %in% c('Primary or Secondary School', 'college') ~ 'School',
    meeting_type %in% c('military','Church', 'Volunteer Organization','Bar or Restaurant','Private
Party') ~ 'Social',
    meeting_type %in% c('Customer-Client Relationship','One-time Service Interaction','Business
Trip','Work Neighbors') ~ 'Work',
    meeting_type %in% c('Internet','Internet Dating or Phone App', 'Internet Social Network','Online
Gaming','Internet Chat','Internet Site','Met Online') ~ 'Online',
    TRUE ~ 'Other'
  )) %>%
  group_by(Q21A_Year,Q21A_Month,meeting_category) %>%
  summarise(count = n(),.groups = 'drop')

meeting_counts <- df_simplified %>%
  group_by(Q21A_Year, meeting_category) %>%
  summarise(count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = meeting_category, values_from = count, values_fill = list(count = 0))

# Line Chart
ggplot(meeting_counts, aes(x = Q21A_Year)) +
  geom_line(aes(y = School, colour = "School", group = 1)) +
  geom_line(aes(y = Social, colour = "Social", group = 1)) +
  geom_line(aes(y = Work, colour = "Work", group = 1)) +
  geom_line(aes(y = Online, colour = "online", group = 1)) +
  geom_line(aes(y = Other, colour = "other", group = 1)) +
  labs(title = "Trend of Meeting Types Over Years",
       x = "Year",
       y = "Count",
       colour = "Meeting Type") +
  theme_minimal()

# Stacked Bar Chart
gmeeting_category <- ggplot(df_simplified, aes(x = Q21A_Year, y = count, fill = meeting_category)) +
  geom_bar(stat = "identity") +
  theme_minimal()
````
````

```
gmeeting_category + theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1, size = 8))

names(df_simplified)

```
The first graph is a line chart, which is an excellent choice for displaying data over time, allowing the viewer to observe trends and patterns. Line charts are particularly effective in showing how data points are connected over a period and can illustrate the rise and fall of counts over the years.
The second graph is a stacked bar chart, which is useful for comparing the total count across different categories while also breaking down each category's contribution to the total. This type of visualization supports the principle of comparison, enabling the reader to compare the composition of meeting types for each year at a glance.
```{r}
#Question 2: Age is just a number

ggplot(df, aes(x = ppage, y = Q9, color = ppgender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Age Relationship Between Partners",
       x = "Respondent's Age",
       y = "Partner's Age") +
  theme_minimal()

```
There's a clear positive correlation between the respondent's age and their partner's age. This indicates that generally, partners tend to be close in age.  The plot shows two distinct trend lines, one for male respondents and one for female respondents. This suggest a gender difference in age dynamics within partnerships.
```{r}
#Question3. Politics and Dating

# Plot 1: Distribution of relationship duration by political affiliation
df_summary <- df %>%
  group_by(partyid7) %>%
  summarise(average_duration = mean(duration, na.rm = TRUE))

ggplot(df_summary, aes(x = partyid7, y = average_duration)) +
  geom_col() +
  labs(title = "Average Relationship Duration by Political Affiliation",
       x = "Political Affiliation",
       y = "Average Duration (days)") +
  theme_minimal()

# Plot 2: Frequency of meeting types by political affiliation
df_party <- df %>%
  mutate(meeting_category = case_when(
    meeting_type %in% c('Primary or Secondary School', 'college') ~ 'School',
    meeting_type %in% c('military','Church', 'Volunteer Organization','Bar or Restaurant','Private Party') ~ 'Social',
    meeting_type %in% c('Customer-Client Relationship','One-time Service Interaction','Business Trip','Work Neighbors') ~ 'Work',
    meeting_type %in% c('Internet','Internet Dating or Phone App', 'Internet Social Network','Online Gaming','Internet Chat','Internet Site','Met Online') ~ 'Online',
    TRUE ~ 'Other'
  )) %>%
  group_by(Q21A_Year,Q21A_Month,meeting_category,partyid7) %>%
  summarise(count = n(),.groups = 'drop')
meeting_party <- df_party %>%
  group_by(partyid7, meeting_category) %>%
  summarise(count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = meeting_category, values_from = count, values_fill = list(count = 0))

ggplot(meeting_party, aes(x = partyid7)) +
  geom_col(aes(y = School, fill = "School"), position = "dodge") +
```

```
  geom_col(aes(y = Social, fill = "Social"), position = "dodge") +
  geom_col(aes(y = Work, fill = "Work"), position = "dodge") +
  geom_col(aes(y = Online, fill = "Online"), position = "dodge") +
  geom_col(aes(y = Other, fill = "Other"), position = "dodge") +
  scale_fill_manual(values = c("School" = "blue", "Social" = "red", "Work" = "green", "Online" =
"purple", "Other" = "orange")) +
  labs(title = "Meeting Types by Political Affiliation",
       x = "Political Affiliation",
       y = "Frequency") +
  theme_minimal() +
  guides(fill = guide_legend(title = "Meeting Category"))


```
```{r}
#Question 4. Your turn to choose

#The relationship between the number of people they met by App and the Education level of the people
they met.
ggplot(df, aes(x = w6_how_many_app, fill = Q10)) +
  geom_histogram(stat = "count", position = "dodge") +
  theme_minimal() +
  labs(x = "number of the people they met on app", y = "Partner's Education Level", title = "Education
Level by Using APP")


```

```{r}
install.packages("plotly")
library(plotly)
library(DT)
# Convert a ggplot object to a plotly object
p <- ggplot(meeting_party, aes(x = partyid7)) +
  geom_col(aes(y = School, fill = "School"), position = "dodge") +
  geom_col(aes(y = Social, fill = "Social"), position = "dodge") +
  geom_col(aes(y = Work, fill = "Work"), position = "dodge") +
  geom_col(aes(y = Online, fill = "Online"), position = "dodge") +
  geom_col(aes(y = Other, fill = "Other"), position = "dodge") +
  scale_fill_manual(values = c("School" = "blue", "Social" = "red", "Work" = "green", "Online" =
"purple", "Other" = "orange")) +
  labs(title = "Meeting Types by Political Affiliation",
       x = "Political Affiliation",
       y = "Frequency") +
  theme_minimal() +
  guides(fill = guide_legend(title = "Meeting Category"))

ggplotly(p)
#Without ggplot
datatable(df_simplified, options = list(
  pageLength = 10,
  searchHighlight = TRUE,
  filter = 'top',
  autoWidth = TRUE
))

```
```{r}
#Question 6:

library(dplyr)
library(DT)

# Select a few columns from the dataframe and rename them for clarity
simplified_df <- df %>%
```

```
  select(
    PartnerRce = Q6B,  # Renaming 'pppa1635' to 'GenderIdentity'
    PartnerAge = Q9,             # Renaming 'Q9' to 'PartnerAge'
    EducationLevel = Q10,        # Renaming 'Q10' to 'EducationLevel'
    # Include any other variables you find relevant
  )



# Create an interactive data table with renamed columns
datatable(simplified_df, options = list(
  pageLength = 10,
  searchHighlight = TRUE,
  filter = 'top',  # Enable filtering for each column
  autoWidth = TRUE))
```
Main questions I wanna show based on this data table:
Are age disparities more pronounced within certain racial or educational groups?
Does racial background show specific trends in the choice of partner's education level?
Is education level a key criterion in partner selection, with individuals of different educational attainments tending to match with each other?