Historical Pricing of World's Fairs

```
library(readr)
url <- "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2024/2024-08-13/worlds
worlds_fairs <- read_csv(url)
```

```
## Rows: 70 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): name_of_exposition, country, city, category, theme, notables
## dbl (8): start_month, start_year, end_month, end_year, visitors, cost, area,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(worlds_fairs)
```

```
## # A tibble: 6 x 14
##   start_month start_year end_month end_year name_of_exposition     country city
##         <dbl>      <dbl>     <dbl>    <dbl> <chr>                 <chr>   <chr>
## 1           4       1851        10     1851 The Great Exhibition  United~ Lond~
## 2           5       1855        11     1855 Exposition Universell~ France  Paris
## 3           5       1862        11     1862 International Exhibit~ United~ Lond~
## 4           4       1867        11     1867 Exposition Universell~ France  Paris
## 5           5       1873        10     1873 Weltausstellung 1873 ~ Austri~ Vien~
## 6           5       1876        11     1876 Centennial Exposition  United~ Phil~
## # i 7 more variables: category <chr>, theme <chr>, notables <chr>,
## #   visitors <dbl>, cost <dbl>, area <dbl>, attending_countries <dbl>
```

Question: How do the themes and categories of World's Fairs relate to their visitor counts and costs over time?

Introduction:

The World's Fairs, also known as the Universal Expositions or Expos, are quinquennial global events in which nations can showcase technological advancements, their cultural heritage, or innovative ideas their country has produce. One famous example of the product of the World's Fairs is the Seattle Space Needle, a major landmark in the United States. These grand exhibitions, held periodically in different locations, draw millions of visitors and often leave behind iconic structures or lasting cultural impacts. The dataset provided by the Tidy Tuesday initiative - a weekly data project which provides open datasets - allows us to analyze information about the World's Fairs. Of interest in this dataset is how the themes or categories of a World's Fair influence the visitor count and associate cost over time, and how clustering methods can help us understand this dataset.

The World's Fairs dataset includes temporal information such as the month or year in which the fair began, as well as the month or year in which it ended. We also know information on the city in which the fair took place, the theme, the number if visitors, the cost, and attending countries, which further eludidate details relevant to the question. We aim to analyze the columns `theme`, `category`, `visitors`, `cost`, and `start_year`, which can help to answer our question. The `theme` column describes the theme of each fair, which allows us to group fairs by their focus (technology, cultural exchange, etc). The `category` column tells us whether the fair which was a World Expo or Specialised Expo, important for separating the data to minimize errors in

our underlying assumptions. The `visitors` column represents the number of visitors in millions. The `cost` records the total cost of the fair, in millions (USD). Finally, the `start_year` column tells us the year each fair began, allowing us to analyze changes over time.

Approach: Our approach is to use clustering methods to group World's Fairs based on their themese, categories, visitor counts, and costs (using `kmeans()`). We use clustering because it identifies patterns and relationships in multidimensional data, which allows us to group fairs which have similar features. Clustering will allow us to analyze how themese and categories are related to the populatity and economics of the fairs over time. This method is most useful for exploring the trends and relationships in the data, which allow us to answer our question most accurately to the data.

To present our findings, we want to shwo the total visitors over time using a line graph (`geom_line()`), as well as visualize the visitors between World Expos and Specialized Expos using a bar graph (`geom_bar()`). Finally, we'll visualize the top ten themes by cost using another bar graph (`geom_bar()`). At the end, we we will use a scatter plot with clustering labels (`geom_point()`) to visualize visitor counts against cost, as well as use shapes (triangle, square) to differentiate between Specialized Expos and Global Expos. These plots will help us visualize the information we wish to analyze from the dataset in an effective way to understand and identify trends in the data. One limitation with clustering is that it doesn't allow us to properly visualize or identify elements in the data, as clusters are not inherently labeled, and thus the data needs to be understood through alternative visualization methods to allow us to properly grasp the information provided through clustering. All together, these visualizations allow us to analyze the interplay between themes, categories, and outcomes, which we can leverage to assess the data and answer the question.

Analysis: First we make a plot of the total visitors over time, to help identify trends in the data.

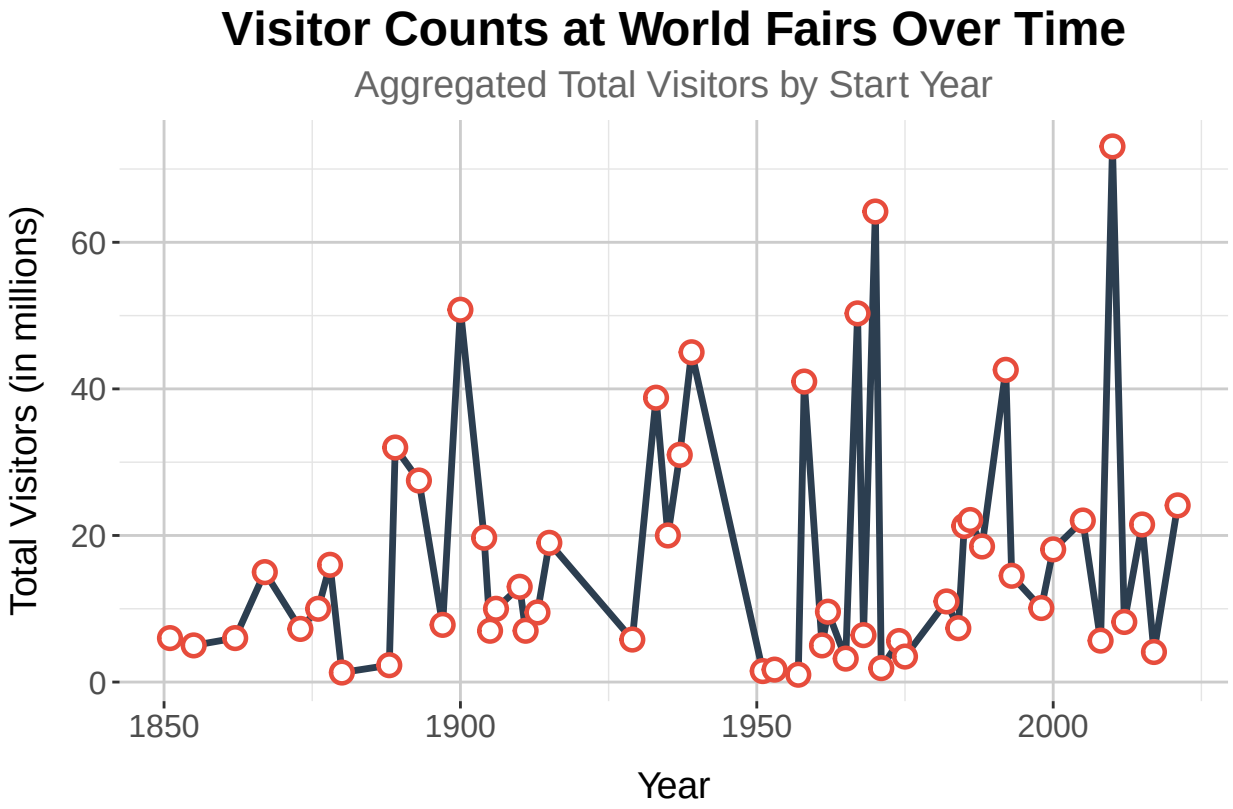
```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v stringr   1.5.1
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
worlds_fairs %>%
  filter(!is.na(visitors)) %>%
  group_by(start_year) %>%
  summarize(total_visitors = sum(visitors, na.rm = TRUE)) %>%
  ggplot(aes(x = start_year, y = total_visitors)) +
  geom_line(color = "#2C3E50", size = 1.2) +
  geom_point(color = "#E74C3C", size = 3, shape = 21, fill = "white", stroke = 1.2) +
  labs(
    title = "Visitor Counts at World Fairs Over Time",
    subtitle = "Aggregated Total Visitors by Start Year",
    x = "Year",
    y = "Total Visitors (in millions)",
    caption = "Data Source: Worlds Fairs Dataset"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, color = "gray40"),
    axis.title.x = element_text(size = 14, margin = margin(t = 10)),
    axis.title.y = element_text(size = 14, margin = margin(r = 10)),
```

```
    axis.text = element_text(size = 12),
    panel.grid.major = element_line(color = "gray80", linewidth = 0.5),
    panel.grid.minor = element_line(color = "gray90", linewidth = 0.25),
    panel.background = element_rect(fill = "white"),
    plot.caption = element_text(size = 10, hjust = 1, color = "gray50")
  )
```

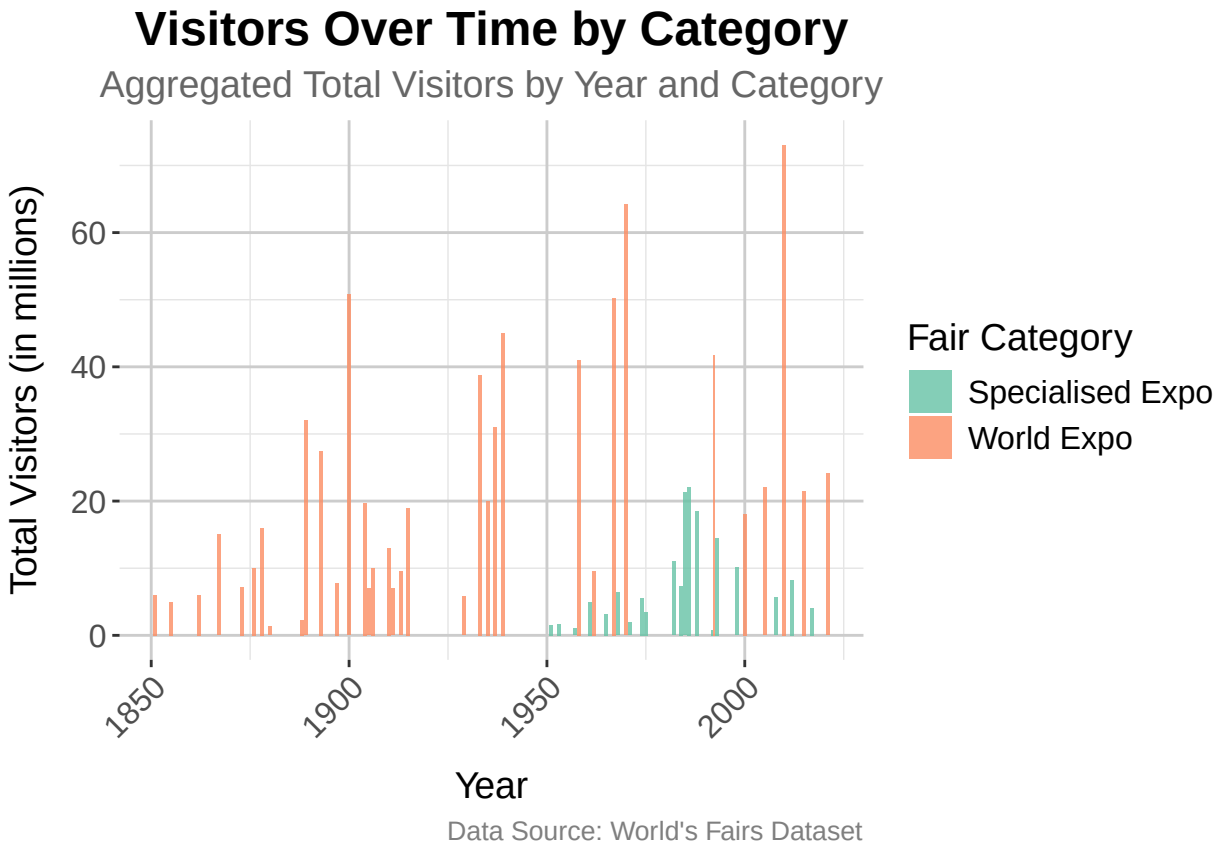
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Next, we make a plot showing how many visitors there are and how they compare between the World's Fair and World's Expo.

```
library(ggplot2)
library(tidyverse)

worlds_fairs <- worlds_fairs %>%
  filter(!is.na(start_year), !is.na(visitors), !is.na(category)) %>%
  mutate(
    start_year = as.numeric(start_year),
    visitors = as.numeric(visitors)
  )

aggregated_data <- worlds_fairs %>%
  group_by(start_year, category) %>%
```

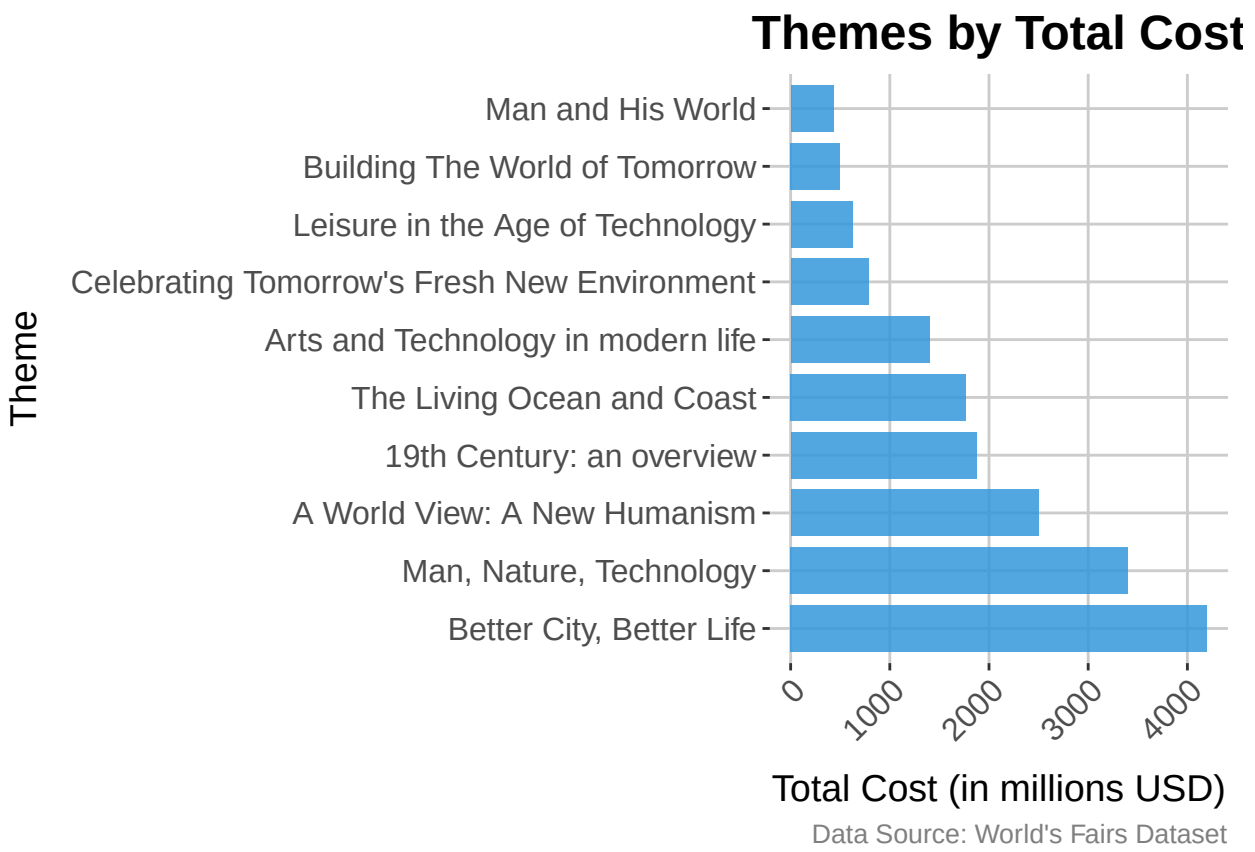
```r
  summarize(total_visitors = sum(visitors, na.rm = TRUE), .groups = "drop")

ggplot(aggregated_data, aes(x = start_year, y = total_visitors, fill = category)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.8) +
  scale_fill_brewer(palette = "Set2") +  # Professional color palette
  labs(
    title = "Visitors Over Time by Category",
    subtitle = "Aggregated Total Visitors by Year and Category",
    x = "Year",
    y = "Total Visitors (in millions)",
    fill = "Fair Category",
    caption = "Data Source: World's Fairs Dataset"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, color = "gray40"),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 12),
    axis.title.x = element_text(size = 14, margin = margin(t = 10)),
    axis.title.y = element_text(size = 14, margin = margin(r = 10)),
    axis.text.y = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12),
    panel.grid.major = element_line(color = "gray80", linewidth = 0.5),
    panel.grid.minor = element_line(color = "gray90", linewidth = 0.25),
    panel.background = element_rect(fill = "white"),
    plot.caption = element_text(size = 10, hjust = 1, color = "gray50")
  )
```



Fi-

nally, we want to make a bar plot which shows the top ten themes which made the most money.

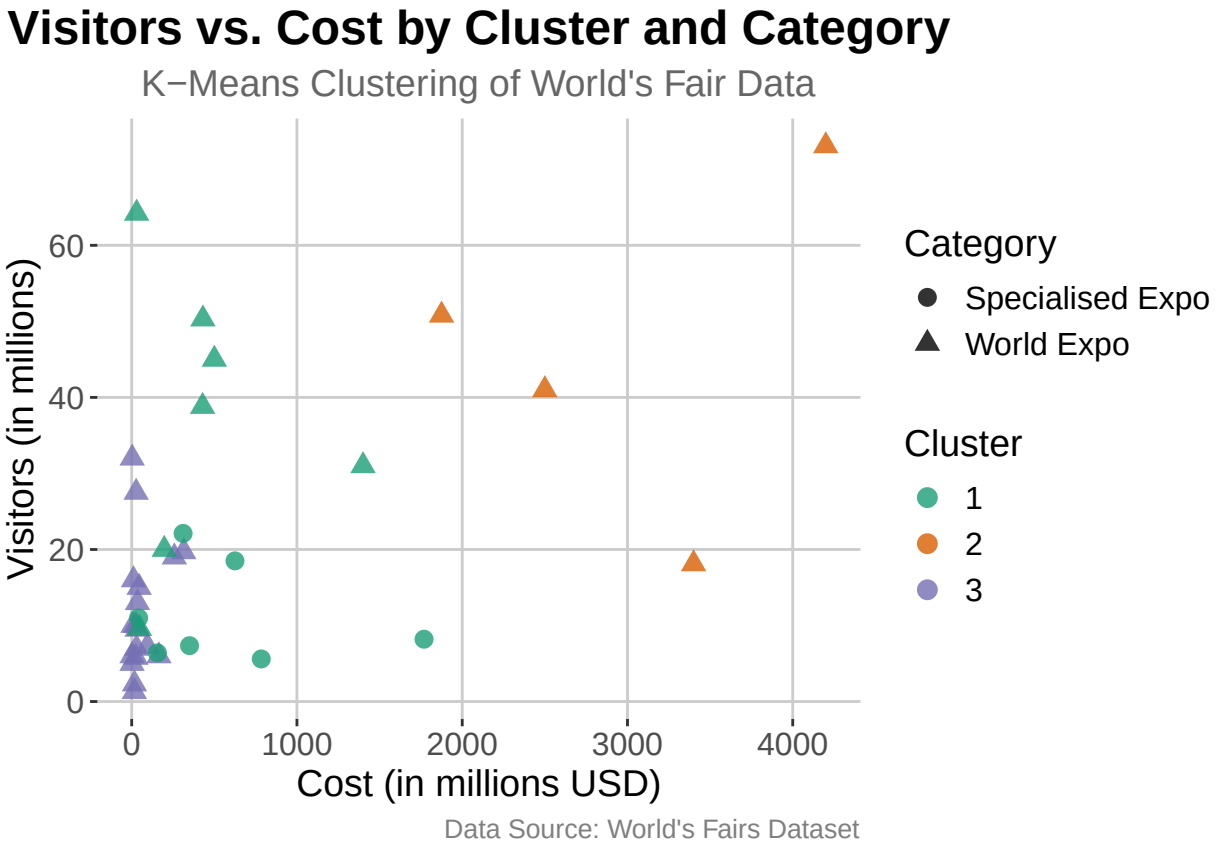
```
library(tidyverse)

worlds_fairs <- worlds_fairs %>%
  filter(!is.na(theme), !is.na(cost)) %>%
  mutate(cost = as.numeric(cost))

aggregated_cost <- worlds_fairs %>%
  group_by(theme) %>%
  summarize(total_cost = sum(cost, na.rm = TRUE), .groups = "drop") %>%
  arrange(desc(total_cost)) %>%
  slice_head(n = 10)  # Select top 10 themes by total cost

ggplot(aggregated_cost, aes(x = reorder(theme, -total_cost), y = total_cost)) +
  geom_bar(stat = "identity", fill = "#3498DB", alpha = 0.85, width = 0.8) +
  labs(
    title = "Themes by Total Cost",
    x = "Theme",
    y = "Total Cost (in millions USD)",
    caption = "Data Source: World's Fairs Dataset"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, color = "gray40"),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 12),
    axis.title.x = element_text(size = 14, margin = margin(t = 10)),
    axis.title.y = element_text(size = 14, margin = margin(r = 10)),
    axis.text.y = element_text(size = 12),
    panel.grid.major = element_line(color = "gray80", linewidth = 0.5),
    panel.grid.minor = element_blank(),  # Remove minor gridlines for clarity
    panel.background = element_rect(fill = "white"),
    plot.caption = element_text(size = 10, hjust = 1, color = "gray50")
  ) +
  coord_flip()
```



Now, we want to cluster the data, separating expos and world's fairs to see if any differences emerge, and plot the visitors vs. cost to determine similar features.

```
library(tidyverse)

worlds_fairs_clean <- worlds_fairs %>%
  filter(!is.na(visitors), !is.na(cost), !is.na(start_year), !is.na(category)) %>%
  mutate(
    visitors = as.numeric(visitors),
    cost = as.numeric(cost),
    start_year = as.numeric(start_year)
  ) %>%
  select(theme, category, visitors, cost, start_year)

clustering_data <- worlds_fairs_clean %>%
  select(visitors, cost, start_year) %>%
  scale()

set.seed(123)
km_fit <- kmeans(clustering_data, centers = 3, nstart = 10)

worlds_fairs_clustered <- worlds_fairs_clean %>%
  mutate(cluster = factor(km_fit$cluster))

ggplot(worlds_fairs_clustered, aes(x = cost, y = visitors, color = cluster, shape = category)) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_brewer(palette = "Dark2") +
  labs(
```

```
    title = "Visitors vs. Cost by Cluster and Category",
    subtitle = "K-Means Clustering of World's Fair Data",
    x = "Cost (in millions USD)",
    y = "Visitors (in millions)",
    color = "Cluster",
    shape = "Category",
    caption = "Data Source: World's Fairs Dataset"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, color = "gray40"),
    axis.title = element_text(size = 14, margin = margin(t = 10, r = 10)),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12),
    panel.grid.major = element_line(color = "gray80", linewidth = 0.5),
    panel.grid.minor = element_blank(),  # minimizes visual clutter
    panel.background = element_rect(fill = "white"),
    plot.caption = element_text(size = 10, hjust = 1, color = "gray50")
  )
```



Discussion:

Using the information from the first three graphs, we can make some preliminary remarks on features of the data. From the first graph of visitor counts over time, there are several peaks corresponding to events when the World's Fair had a significantly higher attendance. These correspond to major historical events or well-publicized fair, such as the post-WW2 expositions, or the turn of the millenium. Troughs in attendance correspond to global conflicts or economic downturns, such as seen during the Great Depression or the world

wars. This information is not necessarily reflected in the data and produces bias in our results. The second visualization shows visitor counts between World Expos and Specialzed Expos. What we observe is that specialized expos only started after 1950, and represent a bump data for later dates but are consistently lower than World Expos visitor counts. We can assume that World Expos are much higher-profile and Specializd Expos represent an insignificant event, which generally does not overlap with World Expos, and doesn't influence the results of analysis as much as other factors such as world events or global economic conditions. Finally, the graph of the top 10 themes shows that themes such as `Better City, Better Life` and `Man, Nature, Technology` represent those themes with the highest costs compared to other events. The clustering graph shows the visitors vs. cost by cluster and category, which breaks down into three distinct clusters: 1) low-cost, low-visitor fairs, predominantly Specialized Expos; 2) medium-cost, medium-visitor fairs, a mix of Specialized and smaller World Expos; 3) and high-cost, high-visitor fairs, which are exclusively World Expos. These clusters show that higher costs correlate with higher visitor counts for World Expos. We can see that visitors and cost are highest for World Expos, which correspond to events which represented the higher cost events. The information shows that themes and categories of World's Fairs strongly influence their visitor counts and costs. World Expos consistently attract higher visitors and incur higher costs compared to Specialized Expos, indicating their larger scale and significance. Themes such as "Better City, Better Life" are tied to high-cost, high-visitor events, reflecting their broad appeal and economic impact.