Homework I - Solution

Gul Inan

10/29/2021

Answer

- **Answer 1:**
- Load the `tidyverse` library first. We will need it throughout the homework.

```
library(tidyverse)
```

- Read in the data into R and check it:

```
#i could not import csv file, and i
#got help from here: https://stackoverflow.com/a/45977065
library(readxl)
music_revenue <- read_excel("data/Revenue_Chart_Full_Data.xlsx", skip = 1)
```

- See a portion of the data:

```
music_revenue %>%
  head(5)
```

```
## # A tibble: 5 x 16
##   `Year of Year Da~ `Adjusted for Infl~ `Adjusted for Inf~ Format   Metric  Year
##               <dbl> <lgl>               <lgl>              <chr>    <chr>  <dbl>
## 1              2005 NA                  NA                 Cassette Value   2005
## 2              2015 NA                  NA                 CD Sing~ Value   2015
## 3              2015 NA                  NA                 Paid Su~ Value   2015
## 4              2017 NA                  NA                 Downloa~ Value   2017
## 5              1986 NA                  NA                 Vinyl S~ Value   1986
## # ... with 10 more variables: Value (For Charting) <chr>,
## #   Adjusted for Inflation Flag <lgl>, Year Date <dbl>,
## #   Format Value # (Billion) <chr>, Format Value # (Million) <chr>,
## #   Total Value # (Billion) <chr>, Total Value # (Million) <lgl>,
## #   Total Value For Year <chr>, Value (Actual) <chr>, Year (copy) <dbl>
```

- Check the structure of the data:

```
str(music_revenue)
```

```
## tibble [453 x 16] (S3: tbl_df/tbl/data.frame)
##  $ Year of Year Date           : num [1:453] 2005 2015 2015 2017 1986 ...
##  $ Adjusted for Inflation Notes: logi [1:453] NA NA NA NA NA NA ...
##  $ Adjusted for Inflation Title: logi [1:453] NA NA NA NA NA NA ...
##  $ Format                      : chr [1:453] "Cassette" "CD Single" "Paid Subscription" "Download Si
##  $ Metric                      : chr [1:453] "Value" "Value" "Value" "Value" ...
##  $ Year                        : num [1:453] 2005 2015 2015 2017 1986 ...
##  $ Value (For Charting)        : chr [1:453] "13.1" "1.196946610" "1156.708513551" "667.875936447" .
```

```
##  $ Adjusted for Inflation Flag : logi [1:453] NA NA NA NA NA NA ...
##  $ Year Date                   : num [1:453] 2005 2015 2015 2017 1986 ...
##  $ Format Value # (Billion)    : chr [1:453] NA NA "$1.2B" NA ...
##  $ Format Value # (Million)    : chr [1:453] "$13.1M" "$1.2M" NA "$667.9M" ...
##  $ Total Value # (Billion)     : chr [1:453] "$12.3B" "$6.7B" "$6.7B" "$8.5B" ...
##  $ Total Value # (Million)     : logi [1:453] NA NA NA NA NA NA ...
##  $ Total Value For Year        : chr [1:453] "$12289.9B" "$6710.8B" "$6710.8B" "$8503.2B" ...
##  $ Value (Actual)              : chr [1:453] "13.1" "1.196946610" "1156.708513551" "667.875936447" .
##  $ Year (copy)                 : num [1:453] 2005 2015 2015 2017 1986 ...
```

- **Answer 2:**

```
revenue_18_20 <- music_revenue %>%
                 select("Year"="Year (copy)", "Format", "Value"="Value (Actual)") %>%
                 filter(Year >= 2018) %>%
                 arrange(Year) %>%
                 mutate(Value = as.numeric(Value))
```

- A portion of the data:

```
revenue_18_20 %>%
              head(10)
```

```
## # A tibble: 10 x 3
##     Year Format                         Value
##    <dbl> <chr>                          <dbl>
##  1  2018 Vinyl Single                   5.72
##  2  2018 SoundExchange Distributions  953.
##  3  2018 Ringtones & Ringbacks         25.0
##  4  2018 SACD                           0.860
##  5  2018 Download Music Video           2.22
##  6  2018 Paid Subscription           4614.
##  7  2018 Kiosk                          1.97
##  8  2018 DVD Audio                      0.329
##  9  2018 CD                           696.
## 10  2018 Music Video (Physical)        28.4
```

- The structure of the data:

```
revenue_18_20 %>%
  str()
```

```
## tibble [57 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Year  : num [1:57] 2018 2018 2018 2018 2018 ...
##  $ Format: chr [1:57] "Vinyl Single" "SoundExchange Distributions" "Ringtones & Ringbacks" "SACD" ..
##  $ Value : num [1:57] 5.72 952.8 24.96 0.86 2.22 ...
```

- Get the summary statistics:

```
annual_retail <- revenue_18_20 %>%
                 group_by(Year) %>%
                  summarize(tol_revenue = sum(Value))
```

```
annual_retail %>%
              print()
```

```
## # A tibble: 3 x 2
##    Year tol_revenue
##   <dbl>       <dbl>
```

```
## 1  2018       9738.
## 2  2019      11130.
## 3  2020      12153.
```

- **Answer 3:**

```
streaming <- c("Paid Subscription", "On-Demand Streaming (Ad-Supported)", "Other Ad-Supported Streaming
"SoundExchange Distributions", "Limited Tier Paid Subscription")

stream_20 <- revenue_18_20 %>%
             filter(Year == 2020) %>%
             filter(Format %in% streaming) %>%
             select(Value) %>%
             summarize(total_stream = sum(Value))
```

```
stream_20  %>%
           View()
```

- **Answer 4:**

```
#define some character vectors to increase code readability.
stream <- c("Paid Subscription", "On-Demand Streaming (Ad-Supported)", "Other Ad-Supported Streaming",
physical <- c("LP/EP", "CD")
download <- c("Download Album", "Download Single")
```

```
revenue_categorized_20 <- revenue_18_20 %>%
                          filter(Year == 2020) %>%
                          mutate(category = case_when( Format %in% stream ~ "Streaming",
                               Format == "Synchronization" ~ "Synchronization",
                               Format %in% physical ~ "Physical",
                               Format %in% download ~ "Digital Downloads",
                               TRUE ~ "The Others"

  ))
```

- View the new data

```
View(revenue_categorized_20)
```

- Calculate the percentage of each category share:

```
share_table   <- revenue_categorized_20 %>%
                 group_by(category) %>%
                 summarize(cat_total=sum(Value)) %>%
                 mutate(prop = round(cat_total /sum(cat_total) * 100,2))
```

- View the data.

```
share_table %>%
  View()
```

- Get the table.

```
share_table %>%
  print()
```

```
## # A tibble: 5 x 3
##   category          cat_total  prop
##   <chr>                 <dbl> <dbl>
## 1 Digital Downloads     632.   5.2
```

```
## 2 Physical             1103.   9.08
## 3 Streaming           10075.  82.9
## 4 Synchronization       265.   2.18
## 5 The Others             78.3  0.64
```

- **Answer 5:**
- Reproduce Figure 5. It is a stacked bar chart. We need to convert Year and Format into character variables.

```
fig_5_data <- revenue_18_20 %>%
              filter(Format %in% c("SoundExchange Distributions", "Other Ad-Supported Streaming")) %>%
mutate_at(vars("Year", "Format"), as.character)
```

```
fig_5_data %>%
  View()
```

```
str(fig_5_data)
```

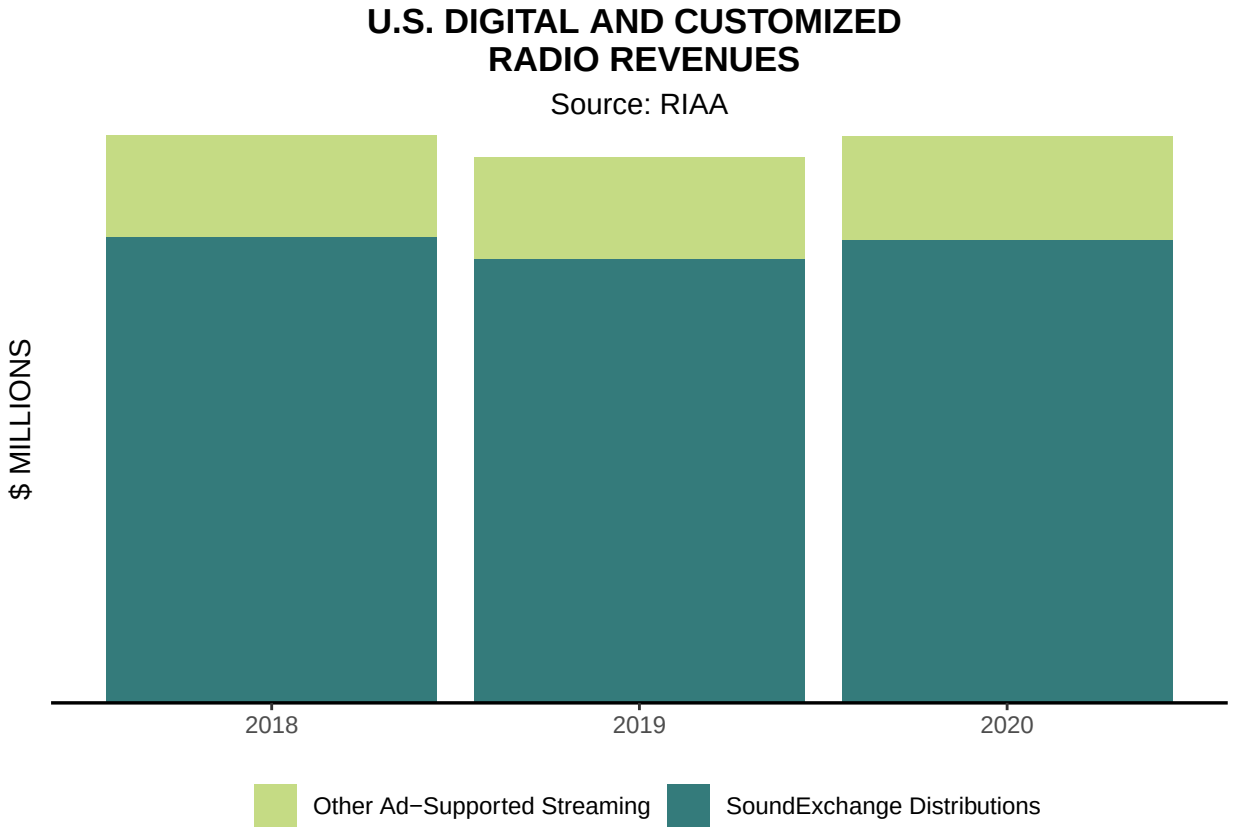
```
## tibble [6 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Year  : chr [1:6] "2018" "2018" "2019" "2019" ...
##  $ Format: chr [1:6] "SoundExchange Distributions" "Other Ad-Supported Streaming" "Other Ad-Supporte
##  $ Value : num [1:6] 953 208 207 908 211 ...
```

- Generate the Figure.

```
library(ggplot2)
```

```
labels <- c("$ 1,161", "$1,115", "$1,1559")
```

```
fig_5_data %>%
ggplot(aes(x=Year, y=Value, fill=Format)) +
  geom_col(position="stack") +
  scale_fill_manual(values = c("#c5db84","#347b7b")) +
  scale_y_continuous(expand = c(0,0)) +
  ggtitle("U.S. DIGITAL AND CUSTOMIZED \n RADIO REVENUES", subtitle = "Source: RIAA" ) +
  labs("y" = "$ MILLIONS") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, face="bold"),
        plot.subtitle =  element_text(hjust = 0.5),
        axis.title.x = element_blank(),
        axis.line.x  = element_line(size = 0.6),
        axis.line.y  = element_blank(),
        axis.text.y  = element_blank(),
        axis.ticks.y = element_blank(),
        legend.title = element_blank(),
        legend.position = "bottom")
```




```
#geom_text(aes(label = labels, y = c(208.17685,208.17685,208.17685)), size = 3)
```