IST 687: Introduction to Data Science

Group E Final Project

LaMiah Tysinger, Katherine Rossi, Nick Videtti, Anthony Klein

**Table of Contents:**

Page 1: Title Page

Page 2: Table of Contents

Page 3: Introduction

Page 4-15: Analysis

Page 16: Conclusions

Page 17: References

**Introduction:**

This dataset is from Nielsen Media Research, the standardized metric for television ratings. There were 14k rows and 26 columns prior to cleaning the dataset. Below is an in depth analysis of the CSV.

* Section number includes the line item when running this information in the database. In this case the line item is by network and program.
* Section label is the type of report that was run.
* Network indicates the network the program ran on.
* Period includes the time period of 2021 and is consistent across the dataset.
* Program indicates the name of the show.
* Time displays the daypart/time of day that these programs aired. Similar to period, this is the same across the board since the dataset was inclusive of total day (6AM-6AM).
* Days show which days the program ran in a format that needed to be adjusted.
* Median age of the program, based on people 2+.
* Demographic (IE: “P18-34 AA (000s)”) impressions, or the number of the average audience that viewed the program and the (000s) denote the number is in thousands.
* Impressions mean any given person within the demo that had watched.
* P2+ means total audience since Nielsen can only measure people aged 2-99.
* T/C gives the number of times the program ran over the period.
* Duration is the number of minutes the program was on air.

The dataset has changed over the past year due to COVID with the scheduling changes of sports as well as ratings increases and decreases. Live sports scheduling was adjusted once mandates were lifted or loosened, and some games were shifted to different nights of the week or different dayparts, which then impacted ratings. Additionally, towards summer of 2021, ratings shifted and saw declines due to loosened mandates. On the other hand, during early 2021, certain genres experienced ratings increases due to government mandated lockdowns, giving people additional time to view programs on TV, which eventually was reversed.

Through quantitative and prescriptive analysis, our team hoped to uncover the top rated programs among certain demographics, highest rated awards shows, and which media properties had the most viewing share across the cable and broadcast landscape.

**Analysis:**

First and foremost, our data needed to be cleaned in order to be usable for our analysis. Going through all the details would be quite time consuming, so we will summarize the steps taken and then provide the full code for more detail.

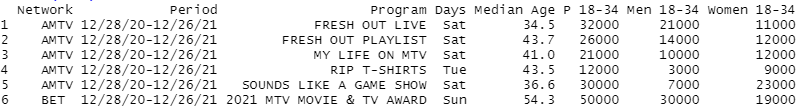
-Read in data using read\_csv()

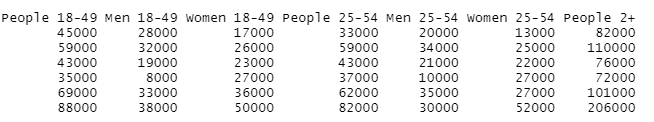
-Delete all extra rows and columns that do not contain data we are interested in

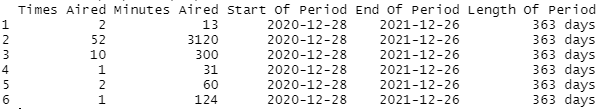
-Change column names and row names

-Transform fields for numerical and/or formatting reasons

Here are the first few rows of our resulting data, followed by the code that we used.







| **Data Cleaning Code:**  Import = function(data){  CSV <- read.csv(data,header = FALSE)  Data <- CSV[-(1:21),-c(1,2,5,6,7,8,10)]  colnames(Data) <- CSV[21,-c(1,2,5,6,7,8,10)]  rownames(Data) <- rownames(CSV)[1:nrow(Data)]  colnames(Data)[6:15] <- CSV[19,13:22]  for (x in 6:17) {Data[,x] <- as.numeric(Data[,x])\* (if (x >= 6 & x <= 15) 1000 else 1)}  Data[,"MedAge"] <- as.numeric(Data[,"MedAge"])  for (x in 1:nrow(Data)) {Data[x,"MedAge"] <- if (is.na(Data[x,"MedAge"])) 0 else Data[x,"MedAge"]}  for (x in 1:nrow(Data)) {Data[x,"DaysFixed"] <- paste(  if (substr(Data[x,"Days"],1,1) == "M") "Mon",  if (substr(Data[x,"Days"],2,2) == "T") "Tue",  if (substr(Data[x,"Days"],3,3) == "W") "Wed",  if (substr(Data[x,"Days"],4,4) == "T") "Thu",  if (substr(Data[x,"Days"],5,5) == "F") "Fri",  if (substr(Data[x,"Days"],6,6) == "S") "Sat",  if (substr(Data[x,"Days"],7,7) == "S") "Sun")}  for (x in 1:nrow(Data)) {Data[x,"Days"] <- gsub(" ","",Data[x,"DaysFixed"])}  Data <- Data[,colnames(Data) != "DaysFixed"]  Data$Period <- gsub("!","",Data$Period)  Data$StartOfPeriod <- as.Date(substr(Data$Period,1,8),"%m/%d/%y")  Data$EndOfPeriod <- as.Date(substr(Data$Period,10,17),"%m/%d/%y")  Data$LengthOfPeriod <- Data$EndOfPeriod - Data$StartOfPeriod  colnames(Data)[colnames(Data) == "T/C"] <- "Times Aired"  colnames(Data)[colnames(Data) == "Duration"] <- "Minutes Aired"  colnames(Data)[colnames(Data) == "MedAge"] <- "Median Age"  colnames(Data)[colnames(Data) == "StartOfPeriod"] <- "Start Of Period"  colnames(Data)[colnames(Data) == "EndOfPeriod"] <- "End Of Period"  colnames(Data)[colnames(Data) == "LengthOfPeriod"] <- "Length Of Period"  for (x in 7:16) {  if (substr(colnames(Data)[x],1,1) == "P") colnames(Data)[x] <- gsub("P","People",colnames(Data)[x])  colnames(Data)[x] <- gsub("M","Men",colnames(Data)[x])  colnames(Data)[x] <- gsub("W","Women",colnames(Data)[x])}  CSV <<- CSV  Data <<- Data}  Import("C:/Users/nvidetti/Downloads/2021 L3 PROGRAM RANKS.L3D.csv") |
| --- |

**Question 1: How many programs within the Top 150 are ViacomCBS originals?**

In order to find the number of programs within the Top 150 that are ViacomCBS originals, we needed to determine the age demographic that would best suit the analysis. From doing this, we determined that people ages 25-54 properly represented a wide range of viewers who would be interested in watching cable television.   
  
 To begin the analysis, we had to clean the Network column and remove any spaces that were typed before or after the actual network name. After this, we created a new dataframe to avoid messing up the original dataset. We then ordered this dataset by the “People Ages 25–54” column (Highest viewership to lowest viewership), and sorted the new dataset for the first 150 rows.

| **Sorting for the Top 150 Rows:**  Data$Network <- gsub(" ","",Data$Network)  sorteddf <- data.frame(Data)  sorteddf <- Data[order(-Data$`People 25-54`),]  Top150 <- sorteddf[1:150,] |
| --- |

This information got us to the baseline of having the Top 150 television shows that people aged 25-54 enjoy the most. For the second half of the analysis, we dove deeper into a quantitative analysis. We took the sum of the amount of rows within the Network column containing a ViacomCBS original Network. These specific networks are: AMTV, CMDY, CMT, LOGO, MTV, MTV2, NAN, NICK, NKJR, NKTN, POP, PAR, SMTH, TNNK, TVLC, TVLD, VH1, BET, BHER, CBS, and CW. We received a result that out of the top 150 shows, only 41 are ViacomCBS Originals. After learning this, we went a step further to calculate the overall percentage of ViacomCBS-specific viewership via the ‘stats’ vector, and received a result of 27.3%.

| **Quantitative Analysis:**  sum(Top150$Network == "AMTV"| Top150$Network == "CMDY"| Top150$Network == "CMT"| Top150$Network == "LOGO"| Top150$Network == "MTV"  | Top150$Network == "MTV2"| Top150$Network == "NAN"| Top150$Network == "NICK"| Top150$Network == "NKJR"| Top150$Network == "NKTN"  | Top150$Network == "POP" | Top150$Network == "PAR"| Top150$Network == "SMTH"| Top150$Network == "TNNK"| Top150$Network == "TVLC"  | Top150$Network == "TVLD"| Top150$Network == "VH1"| Top150$Network == "BET"| Top150$Network == "BHER"| Top150$Network == "CBS"  | Top150$Network == "CW")  #Of the top 150 shows within the 25-54 age range, only 41 are ViacomCBS originals  stats <- (41/150)\*100  stats |
| --- |

**Question 2: Which media companies are the most and least popular for total viewers?**

We were interested in seeing which media group owned the most viewing share across the cable and broadcast television landscape on the program level. This type of analysis can be used to determine which media groups should be paid a premium for ad dollars and where the largest amount of viewership lies.

For this analysis, we created a separate dataframe inclusive of networks, programs, and P2+ impressions to look at total viewers on the program level only. From there, we grouped each media company by network including ViacomCBS, AMC, A+E, Disney, NBC, ABC, Fox and Crown networks. We coded each media group to include programs from networks that were owned by those groups and once each group was made, we summed the total viewing impressions for each group. Following that, we summed the total media groups together in order to determine the share percentage for each property to see which media group earned the most amount of viewers by dividing each media company’s P2+ impressions sum into the total number of P2+ impressions across each media group. Our analysis showed that the three main players in the cable and broadcast industry are ViacomCBS, NBC, and Disney.

Below includes the code for how we did this.

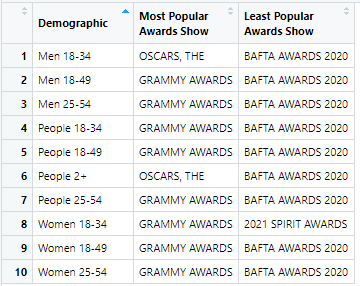
| #adding in media group DFs  totalViewers <- data.frame(Data$Network, Data$Program, Data$`People 2+`)  View(totalViewers)  #grouping total viewership by network group for ViacomCBS  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  VCBSnets = data.frame(c("AMTV", "CMDY", "CMT", "LOGO", "MTV", "MTV2", "NAN",  "NICK", "NKJR", "NKTN", "POP", "PAR", "SMTH", "TNNK", "TVLC", "TVLD",  "VH1", "BET", "BHER", "CBS", "CW"))  names(VCBSnets) = "Network"  totalVCBSnetsViewers <- merge(totalViewers,VCBSnets,by.x = "Data.Network",by.y = "Network")  view(totalVCBSnetsViewers)  sum(totalVCBSnetsViewers$Data..People.2..)  #grouping total viewership by network group for AMC  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  AMCnets = data.frame(c("AMC", "BBCA", "IFC", "SUND", "WETV"))  names(AMCnets) = "Network"  totalAMCnetsViewers <- merge(totalViewers,AMCnets,by.x = "Data.Network",by.y = "Network")  view(totalAMCnetsViewers)  sum(totalAMCnetsViewers$Data..People.2..)  #grouping total viewership by network group for NBCUniversal  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  NBCnets = data.frame(c("BRAV", "CNBC", "COZI", "EN", "GOLF", "MSNBC", "NBCS", "OXYG", "UKID", "SYFY", "USA", "NBC"))  names(NBCnets) = "Network"  totalNBCnetsViewers <- merge(totalViewers,NBCnets,by.x = "Data.Network",by.y = "Network")  view(totalNBCnetsViewers)  sum(totalNBCnetsViewers$Data..People.2..)  #grouping total viewership by network group for Warner Media  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  WMnets = data.frame(c("ADSM", "BOOM", "CNN", "HLN", "TBSC", "TNT", "TOON", "TRU"))  names(WMnets) = "Network"  totalWMnetsViewers <- merge(totalViewers,WMnets,by.x = "Data.Network",by.y = "Network")  view(totalWMnetsViewers)  sum(totalWMnetsViewers$Data..People.2..)  #grouping total viewership by network group for Discovery/Scripps  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  DSnets = data.frame(c("AHC", "APL", "DAM", "DFC", "DISC", "DLIF", "ID", "OWN", "SCI", "TLC", "MT", "CC", "MAG", "FOOD", "GACF", "HGTV", "TRAV"))  names(DSnets) = "Network"  totalDSnetsViewers <- merge(totalViewers,DSnets,by.x = "Data.Network",by.y = "Network")  view(totalDSnetsViewers)  sum(totalDSnetsViewers$Data..People.2..)  #grouping total viewership by network group for Fox  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  FOXnets = data.frame(c("BTN", "FBN", "FOXN", "FS1", "FS2", "FOX"))  names(FOXnets) = "Network"  totalFOXnetsViewers <- merge(totalViewers,FOXnets,by.x = "Data.Network",by.y = "Network")  view(totalFOXnetsViewers)  sum(totalFOXnetsViewers$Data..People.2..)  #grouping total viewership by network group for Disney  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  DISnets = data.frame(c("DSJR", "DSNY", "DXD", "ESNU", "ESP2", "ESPN", "FRFM", "FX", "FXM", "FXX", "NGC", "NGWD", "ABC"))  names(DISnets) = "Network"  totalDISnetsViewers <- merge(totalViewers,DISnets,by.x = "Data.Network",by.y = "Network")  view(totalDISnetsViewers)  sum(totalDISnetsViewers$Data..People.2..)  #grouping total viewership by network group for A+E  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  AEnets = data.frame(c("AEN", "FYI", "HIST", "LIFE", "LMN", "VICE"))  names(AEnets) = "Network"  totalAEnetsViewers <- merge(totalViewers,AEnets,by.x = "Data.Network",by.y = "Network")  view(totalAEnetsViewers)  sum(totalAEnetsViewers$Data..People.2..)  #grouping total viewership by network group for Crown Media  totalViewers$Data.Network <- gsub(" ","",totalViewers$Data.Network)  CMnets = data.frame(c("HALL", "HMM", "HDRM"))  names(CMnets) = "Network"  totalCMnetsViewers <- merge(totalViewers,CMnets,by.x = "Data.Network",by.y = "Network")  view(totalCMnetsViewers)  sum(totalCMnetsViewers$Data..People.2..)  #summing across all media groups for easier viewing  sumVCBS <- sum(totalVCBSnetsViewers$Data..People.2..)  sumAMC <- sum(totalAMCnetsViewers$Data..People.2..)  sumNBC <- sum(totalNBCnetsViewers$Data..People.2..)  sumWM <- sum(totalWMnetsViewers$Data..People.2..)  sumDS <- sum(totalDSnetsViewers$Data..People.2..)  sumFOX <- sum(totalFOXnetsViewers$Data..People.2..)  sumDIS <- sum(totalDISnetsViewers$Data..People.2..)  sumAE <- sum(totalAEnetsViewers$Data..People.2..)  sumCM <- sum(totalCMnetsViewers$Data..People.2..)  sumVCBS  sumAMC  sumNBC  sumWM  sumDS  sumFOX  sumDIS  sumAE  sumCM  #Determining share of impressions across TV  netSum <- sum(sumVCBS, sumAMC, sumNBC, sumWM, sumDS, sumFOX, sumDIS, sumAE, sumCM)  VCBSshare <- sumVCBS/netSum  AMCshare <- sumAMC/netSum  NBCshare <- sumNBC/netSum  WMshare <- sumWM/netSum  DSshare <- sumDS/netSum  FOXshare <- sumFOX/netSum  DISshare <- sumDIS/netSum  AEshare <- sumAE/netSum  CMshare <- sumCM/netSum  #Running percentages for ease of use  VCBSshare  AMCshare  NBCshare  WMshare  DSshare  FOXshare  DISshare  AEshare  CMshare  #just checking to make sure all adds up to 100%  sum(VCBSshare, AMCshare, NBCshare, WMshare, DSshare, FOXshare, DISshare, AEshare, CMshare) |
| --- |

**Question 3: What are the most popular and least popular awards shows for each demographic, as well as overall?**

For this part of our analysis, we first needed to identify which programs were awards shows. This required us to use the merge() function to combine our data with an Excel file that contains the names of the awards shows. This allowed us to filter our data down to only awards shows data. Due to the awards show names Excel file being an Excel file, we needed to load the readxl library to use the read\_excel() function.

Next, we needed to load the tidyverse library so we could pipe our data into the group\_by() and summarize() functions. This allowed us to create a data frame for the number of impressions by awards show, per demographic. Using that data frame, we were able to identify the most popular awards show and the least popular awards show for each demographic.

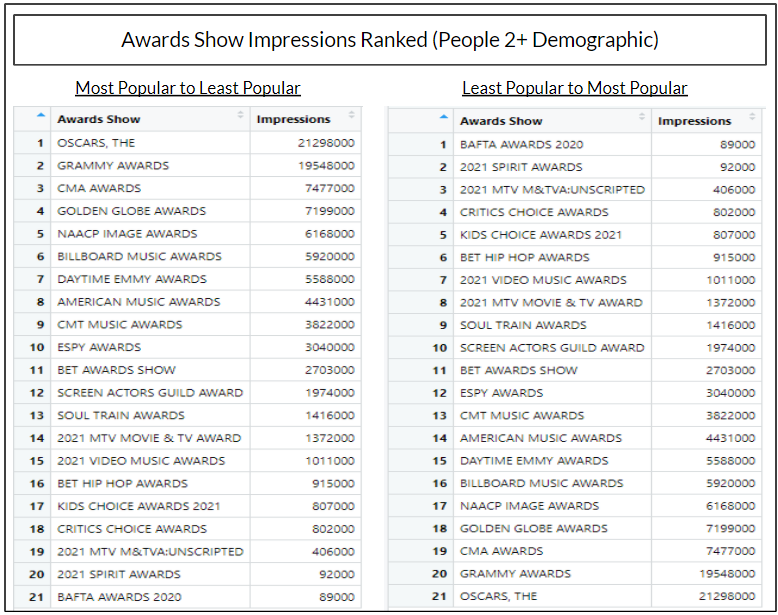
The results we found are below, followed by the code used to generate them.

****

| #What are the most popular and least popular awards shows for each demographic, as well as overall?  library(readxl)  AwardsShows <- read\_excel("C:/Users/nvidetti/Downloads/Awards Shows.xlsx")[,3]  AwardsShows <- merge(AwardsShows,Data,by = "Program")  library(tidyverse)  ImpressionsByAwardShow <- as.data.frame(AwardsShows %>% group\_by(Program) %>% summarize(`People 18-34` = sum(`P 18-34`), `Men 18-34` = sum(`Men 18-34`), `Women 18-34` = sum(`Women 18-34`), `People 18-49` = sum(`People 18-49`), `Men 18-49` = sum(`Men 18-49`), `Women 18-49` = sum(`Women 18-49`), `People 25-54` = sum(`People 25-54`), `Men 25-54` = sum(`Men 25-54`), `Women 25-54` = sum(`Women 25-54`), `People 2+` = sum(`People 2+`)))  Top\_People\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 18-34` == max(ImpressionsByAwardShow$`People 18-34`),1]  Top\_Men\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 18-34` == max(ImpressionsByAwardShow$`Men 18-34`),1]  Top\_Women\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 18-34` == max(ImpressionsByAwardShow$`Women 18-34`),1]  Top\_People\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 18-49` == max(ImpressionsByAwardShow$`People 18-49`),1]  Top\_Men\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 18-49` == max(ImpressionsByAwardShow$`Men 18-49`),1]  Top\_Women\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 18-49` == max(ImpressionsByAwardShow$`Women 18-49`),1]  Top\_People\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 25-54` == max(ImpressionsByAwardShow$`People 25-54`),1]  Top\_Men\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 25-54` == max(ImpressionsByAwardShow$`Men 25-54`),1]  Top\_Women\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 25-54` == max(ImpressionsByAwardShow$`Women 25-54`),1]  Top\_People\_2\_Plus <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 2+` == max(ImpressionsByAwardShow$`People 2+`),1]  TopShows <- data.frame(c("People 18-34","Men 18-34","Women 18-34","People 18-49","Men 18-49","Women 18-49","People 25-54","Men 25-54","Women 25-54","People 2+"),c(Top\_People\_18\_34,Top\_Men\_18\_34,Top\_Women\_18\_34,Top\_People\_18\_49,Top\_Men\_18\_49,Top\_Women\_18\_49,Top\_People\_25\_54,Top\_Men\_25\_54,Top\_Women\_25\_54,Top\_People\_2\_Plus))  names(TopShows) <- c("Demographic","Awards Show")  Bottom\_People\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 18-34` == min(ImpressionsByAwardShow$`People 18-34`),1]  Bottom\_Men\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 18-34` == min(ImpressionsByAwardShow$`Men 18-34`),1]  Bottom\_Women\_18\_34 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 18-34` == min(ImpressionsByAwardShow$`Women 18-34`),1]  Bottom\_People\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 18-49` == min(ImpressionsByAwardShow$`People 18-49`),1]  Bottom\_Men\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 18-49` == min(ImpressionsByAwardShow$`Men 18-49`),1]  Bottom\_Women\_18\_49 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 18-49` == min(ImpressionsByAwardShow$`Women 18-49`),1]  Bottom\_People\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 25-54` == min(ImpressionsByAwardShow$`People 25-54`),1]  Bottom\_Men\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Men 25-54` == min(ImpressionsByAwardShow$`Men 25-54`),1]  Bottom\_Women\_25\_54 <- ImpressionsByAwardShow[ImpressionsByAwardShow$`Women 25-54` == min(ImpressionsByAwardShow$`Women 25-54`),1]  Bottom\_People\_2\_Plus <- ImpressionsByAwardShow[ImpressionsByAwardShow$`People 2+` == min(ImpressionsByAwardShow$`People 2+`),1]  BottomShows <- data.frame(c("People 18-34","Men 18-34","Women 18-34","People 18-49","Men 18-49","Women 18-49","People 25-54","Men 25-54","Women 25-54","People 2+"),c(Bottom\_People\_18\_34,Bottom\_Men\_18\_34,Bottom\_Women\_18\_34,Bottom\_People\_18\_49,Bottom\_Men\_18\_49,Bottom\_Women\_18\_49,Bottom\_People\_25\_54,Bottom\_Men\_25\_54,Bottom\_Women\_25\_54,Bottom\_People\_2\_Plus))  names(BottomShows) <- c("Demographic","Awards Show")  TopBottomByDemo <- merge(TopShows,BottomShows,by = "Demographic")  names(TopBottomByDemo) <- c("Demographic","Most Popular Awards Show","Least Popular Awards Show")  View(TopBottomByDemo) |
| --- |

Lastly, we used our impressions by awards show and demographic data frame and looked at only the “People 2+” demographic. For all intents and purposes, we consider this demographic to be everybody, and in this data set, every other demographic is a subset of the “People 2+” demographic. Once we are narrowed down to only “People 2+”, we order the data by number of impressions, descending for most to least and ascending for least to most.

The results and code we used to generate them are below.



| TopShowsOverall <- ImpressionsByAwardShow[order(-ImpressionsByAwardShow$`People 2+`),c("Program","People 2+")]  names(TopShowsOverall) <- c("Awards Show",'Impressions')  rownames(TopShowsOverall) <- c(1:nrow(TopShowsOverall))  BottomShowsOverall <- ImpressionsByAwardShow[order(ImpressionsByAwardShow$`People 2+`),c("Program","People 2+")]  names(BottomShowsOverall) <- c("Awards Show",'Impressions')  rownames(BottomShowsOverall) <- c(1:nrow(BottomShowsOverall))  View(TopShowsOverall)  View(BottomShowsOverall) |
| --- |

**Conclusions:**

We discovered that The Oscars and Grammys were consistently the highest rated awards shows among various demographics, while the BAFTA awards were the lowest. Given that The Oscars and Grammys are both on broadcast networks, they have a better chance of reaching a larger audience. Our analysis shows that cable awards shows are less highly rated, due to the smaller audience within the cable network landscape.

We also found that ViacomCBS has the largest share of viewing among total viewership (P2+). ViacomCBS networks make up 28% of viewership, while NBC closely follows at 23%, and Disney at 20%. All three media companies have broadcast networks and large tentpole events like sports, specials, and awards shows, indicating that they may reach a wider audience than the cable-only media groups.

This information is useful when analyzing television ratings. This data is used to predict awards shows ratings ahead of time in order to sell commercial spots and sponsorships through to advertisers. In terms of analyzing the ratings for top rated programs, this information can be used for competitive analysis purposes and understanding what programs are being watched by certain demographics to determine their viewing interests. Last, understanding which media groups have the highest share of viewership indicates which cable properties should be charging a premium for commercial ad spend.

In a broader sense, this information will be helpful when it comes to scheduling programs, year over year comparisons for long running programs and tentpoles, which demographics to target for advertisers, ad sales research, competitive analysis, and predictive analysis in the future.

**References:**

Saltz, Jeffrey S., and Jeffrey M. Stanton. *Data Science for Business with R*. SAGE Publications, Inc., 2022.