Box Office Mojo Graph’s Code

Oleg Cherkasskikh

9/8/2018

Libs = c(tidyverse, readxl, treemapify, RColorBrewer)

boxoffice <- read\_csv(boxoffice.csv)

Graphs (in order of appearance in the report)

[1] **Rise of the US Movie Industry**

boxoffice %>%

group\_by(year) %>%

summarise(total\_yearly\_revenue = sum(as.numeric(lifetime\_gross))) %>%

arrange(desc(year)) %>%

filter(year < 2018) %>%

ggplot(aes(year, total\_yearly\_revenue, group = 1)) +

geom\_area(size = 1, fill = "tomato2", alpha = 0.3) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

scale\_x\_discrete(breaks = seq(1920, 2020, 5)) +

annotate("text", x = factor(1977), y = 2e+09, label = "The Star Wars", size = 3, color = "black", angle = 90) +

annotate("text", x = factor(2009), y = 7e+09, label = "$10 bln.", size = 4, color = "black") +

annotate("text", x = factor(1993), y = 7e+09, label = "$5 bln.", size = 4, color = "black") +

annotate("text", x = factor(1977), y = 7e+09, label = "$1 bln.", size = 4, color = "black") +

theme(panel.grid.minor = element\_blank(), panel.grid.major.x = element\_blank(), panel.border = element\_blank()) +

labs(title = "Rise of US Movie Industry",

subtitle = "Based on Total Yearly Box Office",

caption = "Data: BoxOfficeMojo",

y = "Total Yearly Revenue")

[2] **Distribution of Box Office Revenue based on Movie Title Length, Movie Gross > $ 10 mln**

boxoffice %>%

mutate(character\_length = nchar(title)) %>%

mutate(length = ifelse((character\_length <= 14), yes = "short", no = "long")) %>%

group\_by(length) %>%

filter(lifetime\_gross > 10000000) %>%

ggplot(aes(x = length, y = lifetime\_gross, col = length)) +

geom\_jitter(alpha = 0.4) +

geom\_boxplot(color = "black") +

ylab("Box Office") +

xlab("Title Length") +

labs(title = "Distribution of Box Office Revenue based on Movie Title Length",

subtitle = "With Movie Gross > $10 mln") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5),

legend.position = "none")

[3] **Distribution of Box Office Revenue based on Movie Title Length, Movie Gross < $ 10 mln**

boxoffice %>%

mutate(character\_length = nchar(title)) %>%

mutate(length = ifelse((character\_length <= 14), yes = "short", no = "long")) %>%

group\_by(length) %>%

filter(lifetime\_gross < 10000000) %>%

ggplot(aes(x = length, y = lifetime\_gross, col = length)) +

geom\_jitter(alpha = 0.4) +

geom\_boxplot(color = "black") +

ylab("Box Office") +

xlab("Title Length") +

labs(title = "Distribution of Box Office Revenue based on Movie Title Length",

subtitle = "With Movie Gross < $10 mln") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5),

legend.position = "none")

[4] **Distribution of Box Office Revenue based on Movie Title Length, Movie Gross > $ 10 mln (density graph)**

boxoffice %>%

mutate(character\_length = nchar(title)) %>%

mutate(length = ifelse((character\_length <= 14), yes = "short", no = "long")) %>%

group\_by(length) %>%

filter(lifetime\_gross > 100000000) %>%

ggplot(aes(x = lifetime\_gross, fill = length, color = length)) +

geom\_density(alpha = 0.1) +

ylab("Density") +

xlab("Box Office") +

labs(title = "Distribution of Box Office Revenue based on Movie Title Length",

subtitle = "With Movie Gross > $10 mln") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5),

legend.position = "none", axis.ticks.y = element\_blank(), axis.text.y = element\_blank())

[5] **Distribution of Box Office Revenue based on Movie Title Length, Movie Gross > $ 10 mln (density graph)**

boxoffice %>%

mutate(character\_length = nchar(title)) %>%

mutate(length = ifelse((character\_length <= 14), yes = "short", no = "long")) %>%

group\_by(length) %>%

filter(lifetime\_gross < 100000000) %>%

ggplot(aes(x = lifetime\_gross, fill = length, color = length)) +

geom\_density(alpha = 0.1) +

ylab("Density") +

xlab("Box Office") +

labs(title = "Distribution of Box Office Revenue based on Movie Title Length",

subtitle = "With Movie Gross < $10 mln") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5),

legend.position = "none", axis.ticks.y = element\_blank(), axis.text.y = element\_blank())

[6] **Number of Movies Made by Each Studio**

boxoffice %>%

group\_by(studio) %>%

summarise(movies\_made = n()) %>%

arrange(movies\_made) %>%

ungroup() %>%

mutate(studio = factor(studio, studio, ordered = T)) %>%

arrange(desc(movies\_made)) %>%

ggplot(aes(x = studio, y = movies\_made, fill = "tomato2", label = as.character(movies\_made))) +

geom\_bar(stat = 'identity') +

theme(legend.position = "none", axis.text.x = element\_blank(), axis.ticks.x = element\_blank(), axis.title = element\_text(size = 12), plot.title = element\_text(hjust = 0.5)) +

ylab("Movies Made") +

labs(title = "Number of Movies Made by Each Studio")

[7] **Number of Movies Made by Each Studio, movies made > 1**

boxoffice %>%

group\_by(studio) %>%

summarise(movies\_made = n()) %>%

arrange(movies\_made) %>%

ungroup() %>%

mutate(studio = factor(studio, studio, ordered = T)) %>%

arrange(desc(movies\_made)) %>%

filter(movies\_made > 1) %>%

ggplot(aes(x = studio, y = movies\_made, fill = "tomato2", label = as.character(movies\_made))) +

geom\_bar(stat = 'identity', width = 1.05) +

theme(legend.position = "none", axis.text.x = element\_blank(), axis.ticks.x = element\_blank(), axis.title = element\_text(size = 12), plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5)) +

ylab("Movies Made") +

labs(title = "Number of Movies Made by Each Studio",

subtitle = "Movies made > 1")

[8] **Number of Movies Made by Top 10 studios**

boxoffice %>%

group\_by(studio) %>%

summarise(movies\_made = n()) %>%

arrange(movies\_made) %>%

ungroup() %>%

mutate(studio = factor(studio, studio, ordered = T)) %>%

top\_n(10) %>%

arrange(desc(movies\_made)) %>%

ggplot(aes(x = studio, y = movies\_made, fill = "tomato2", label = as.character(movies\_made))) +

geom\_bar(stat = 'identity', width = 0.5) +

geom\_label(label.size = 0.5, size = 5, color = "black", fill = "white") +

theme(legend.position = "none", axis.title = element\_text(size = 12),

plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5)) +

ylab("Movies Made") +

xlab("Studio") +

labs(title = "Number of Movies Made by Top 10 studios")

[9] **Historical Studio Density**

boxoffice %>%

group\_by(year, studio) %>%

summarise(studio\_gross = sum(as.numeric(lifetime\_gross))) %>%

ggplot() + geom\_histogram(aes(x = factor(year), y = studio\_gross, fill = studio), position = "fill", stat = "identity", show.legend = FALSE) +

scale\_x\_discrete(breaks = seq(1920, 2020, 10)) +

xlab("Year") +

ylab("") +

labs(title = "Historical Studio Density",

subtitle = "Based on revenues per year; all studios") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))

[10] **Studio Density since 1977**

spectral2 <- brewer.pal(9, "Spectral")

spectral\_range <- colorRampPalette(spectral2)

boxoffice %>%

group\_by(year, studio) %>%

summarise(studio\_gross = sum(as.numeric(lifetime\_gross))) %>%

top\_n(n = 5, wt = studio\_gross) %>%

filter(year > 1977) %>%

ggplot() + geom\_histogram(aes(x = factor(year), y = studio\_gross, fill = studio), position = "fill", stat = "identity", size = 1.5) +

scale\_fill\_manual(values = spectral\_range(14)) +

scale\_x\_discrete(breaks = seq(1977, 2017, 5)) +

xlab("Year") +

ylab("") +

labs(title = "Studio Density since 1977",

subtitle = "Based on revenues per year; top 5 studios") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))

[11] **Studio Density since the beginning of MCU**

boxoffice %>%

group\_by(year, studio) %>%

summarise(studio\_gross = sum(as.numeric(lifetime\_gross))) %>%

top\_n(n = 5, wt = studio\_gross) %>%

filter(year > 2007) %>%

ggplot() + geom\_histogram(aes(x = factor(year), y = studio\_gross, fill = studio), position = "fill", stat = "identity", size = 1.5) +

scale\_fill\_manual(values = spectral\_range(6)) +

xlab("Year") +

ylab("") +

labs(title = "Studio Density since the beginning of MCU",

subtitle = "Based on revenues per year; top 5 studios") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))

[12] **Top 5 studios with their 5 top movies**

ggplot(top5fromtop5, aes(area = lifetime\_gross, fill = studio, label = title, subgroup = studio)) +

geom\_treemap() +

geom\_treemap\_subgroup\_text(place = "bottom", grow = T, alpha = 0.25, color = "black", fontface = "italic", min.size = 0) +

geom\_treemap\_text(color = "white", place = "center", reflow = T, fontface = "italic") +

theme(legend.position = "none") +

labs(title = "Top 5 studios with their 5 top movies",

subtitle = "Based on all-time revenue") +

theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))

[13] **Studio with Biggest Box Office Each Decade**

boxoffice %>%

group\_by(decade, studio) %>%

summarise(studio\_gross = sum(as.numeric(lifetime\_gross))) %>%

ungroup() %>%

group\_by(decade) %>%

filter(studio\_gross == max(studio\_gross)) %>%

arrange(desc(decade)) %>%

ggplot(aes(x = decade, y = studio\_gross, fill = studio, label = studio)) +

geom\_bar(stat = 'identity') +

geom\_text(label.size = 0.5, size = 5, color = "black", fill = "white", vjust = -0.5) +

theme(legend.position = "none", axis.title = element\_text(size = 12),

plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5)) +

ylab("Box Office") +

xlab("Decade") +

labs(title = "Studio with Biggest Box Office Each Decade",

subtitle = "Based on sum of revenues for each studio")

[14] **Major Studios Revenue Trend**

boxoffice %>%

group\_by(year, studio) %>%

summarise(studio\_gross = sum(as.numeric(lifetime\_gross))) %>%

ungroup() %>%

filter(studio %in% c("BV", "Fox", "Par.", "Sony", "Uni.", "WB")) %>%

filter(year > 1950) %>%

ggplot(aes(x = as.numeric(year), y = studio\_gross)) +

geom\_jitter(aes(col = studio), alpha = 0.3) +

geom\_smooth(aes(col = studio), se = FALSE, size = 1.2) +

ylab("Box Office") +

xlab("year") +

labs(title = "Major Studios Revenue Trend") +

theme(plot.title = element\_text(hjust = 0.5))

[15] **Best Studios per ROI**

boxoffice %>%

group\_by(studio) %>%

summarise(total\_revenue = sum(lifetime\_gross), number\_of\_movies = n()) %>%

mutate(avg\_return\_on\_investment = total\_revenue / number\_of\_movies) %>%

filter(number\_of\_movies >= 10) %>%

arrange(desc(avg\_return\_on\_investment)) %>%

top\_n(10) %>%

ggplot(aes(x = reorder(studio, avg\_return\_on\_investment), y = avg\_return\_on\_investment, fill = "tomato2")) +

geom\_bar(stat = 'identity') +

theme(legend.position = "none", axis.title = element\_text(size = 12),

plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5)) +

ylab("Average ROI") +

xlab("Studio with") +

labs(title = "Best studios per ROI",

subtitle = "Top 10 studios shown; n(movies\_made) > 10")

[16] **Buena Vista’s Output vs. Income**

BV1 <- boxoffice\_BV %>%

ggplot() +

geom\_bar(aes(x = year, y = movies\_made), stat = 'identity', fill = "tomato2") +

theme(axis.text.x = element\_blank(),

axis.ticks.x = element\_blank(),

axis.title.x = element\_blank(),

panel.grid.minor.x = element\_line(),

plot.title = element\_text(hjust = 0.5)) +

ylab("Number of Movies") +

labs(title = "Buena Vista's Output vs Income")

BV2 <- boxoffice\_BV %>%

ggplot() +

geom\_area(aes(x = year, y = lifetime\_gross), group = 1, color = "black", fill = "tomato2", alpha = 0.5) +

scale\_x\_discrete(breaks = seq(1960, 2020, 5)) +

ylab("Box Office") +

theme(panel.grid.minor.x = element\_line())

grid.draw(rbind(ggplotGrob(BV1), ggplotGrob(BV2), size = "last"))

[17] **Buena Vista’s Return per Movie vs. Industry Average**

ggplot(boxoffice2selected, aes(x = title, y = gross\_z, label = gross\_z)) +

geom\_point(stat = "identity", fill = "black", size = 7) +

geom\_segment(aes(y = 0,

x = title,

yend = gross\_z,

xend = title),

color = "black") +

geom\_text(color = "white", size = 3) +

labs(title = "Buena Vista's Return per Movie vs. Industry Average") +

ylim(-1, 21) +

coord\_flip() +

xlab("") +

ylab("Deviation from Industry's Mean") +

theme\_bw()