# X-Files Directors

David Peach

Purely as an exercise for my own learning, I thought I would analyse the top directors of episodes of The X-Files, and see which ones tended to have the most popular episodes by viewership.

### Parsing the episode data from Wikipedia

I wrote the page scraper in python, using the Beautiful Soup library. This just parses the episode data out of the tables from the x files episodes wikipedia page.

```
from bs4 import BeautifulSoup
import requests
import csv


class Episode:
    def __init__(self, series):
        self.series = series

    def getseries(self):
        return self.series

    def settitle(self, title):
        self.title = title

    def gettitle(self):
        return self.title

    def setnumoverall(self, numoverall):
        self.numoverall = numoverall

    def getnumoverall(self):
```

```
        return self.numoverall

    def setnuminseries(self, numinseries):
        self.numinseries = numinseries

    def getnuminseries(self):
        return self.numinseries

    def setdirector(self, director):
        self.director = director

    def getdirector(self):
        return self.director

    def setwriter(self, writer):
        self.writer = writer

    def getwriter(self):
        return self.writer

    def setairdate(self, airdate):
        self.airdate = airdate

    def getairdate(self):
        return self.airdate

    def setprodcode(self, prodcode):
        self.prodcode = prodcode

    def getprodcode(self):
        return self.prodcode

    def setviewersus(self, viewersus):
        self.viewersus = viewersus

    def getviewersus(self):
        return self.viewersus


episodes_url = "https://en.wikipedia.org/wiki/List_of_The_X-Files_episodes"

page = requests.get(episodes_url)
```

```
soup = BeautifulSoup(page.text, features="html.parser")
series_tables = soup.find_all("table", class_="wikiepisodetable")

# Remove the tables for the two films.
del series_tables[5]
del series_tables[9]

all_episodes = []
series_counter = 0
for series in series_tables:
    series_counter += 1

    for episode in series.find_all("tr", class_="vevent"):
        ep = Episode(series_counter)
        ep.setnumoverall(episode.find("th").text)

        info_columns = episode.findAll("td")

        ep.setnuminseries(info_columns[0].text)
        # TODO: handle some odd table cells
        if ep.getnumoverall() == 201202:
            continue

        ep.settitle(info_columns[1].text.replace('"', ""))
        ep.setdirector(info_columns[2].text)
        ep.setwriter(info_columns[3].text)
        ep.setairdate(info_columns[4].text)
        ep.setprodcode(info_columns[5].text)
        ep.setviewersus(info_columns[6].text)

        all_episodes.append(ep)

filename = "x-files-episodes.csv"

with open(filename, "w", newline="") as f:
    writer = csv.writer(f)
    writer.writerow(
        [
            "title",
            "numoverall",
            "series",
            "numinseries",
```

```
            "director",
            "writer",
            "airdate",
            "viewers",
        ]
    )
    for ep in all_episodes:
        writer.writerow(
            [
                ep.gettitle(),
                ep.getnumoverall(),
                ep.getseries(),
                ep.getnuminseries(),
                ep.getdirector(),
                ep.getwriter(),
                ep.getairdate(),
                ep.getviewersus(),
            ]
        )
```

### Getting the top directors in order

I am using `R Lang` for analysing the data and generating the visuals.

First things first, I find the people who have directed the most episodes from The X-Files TV series. I have chosen to use the top eight most active people, but this number could easily but ammended.

```
data <- read.csv("x-files-episodes.csv")

most_active_directors <- head(
  sort(table(data$director), decreasing = TRUE),
  n = 8L
)

as.data.frame(most_active_directors) %>%
  kable(
    col.names = c("Name of Director", "Number of Episodes directed"),
    format = "html"
  )
```

Table 1: Most active directors

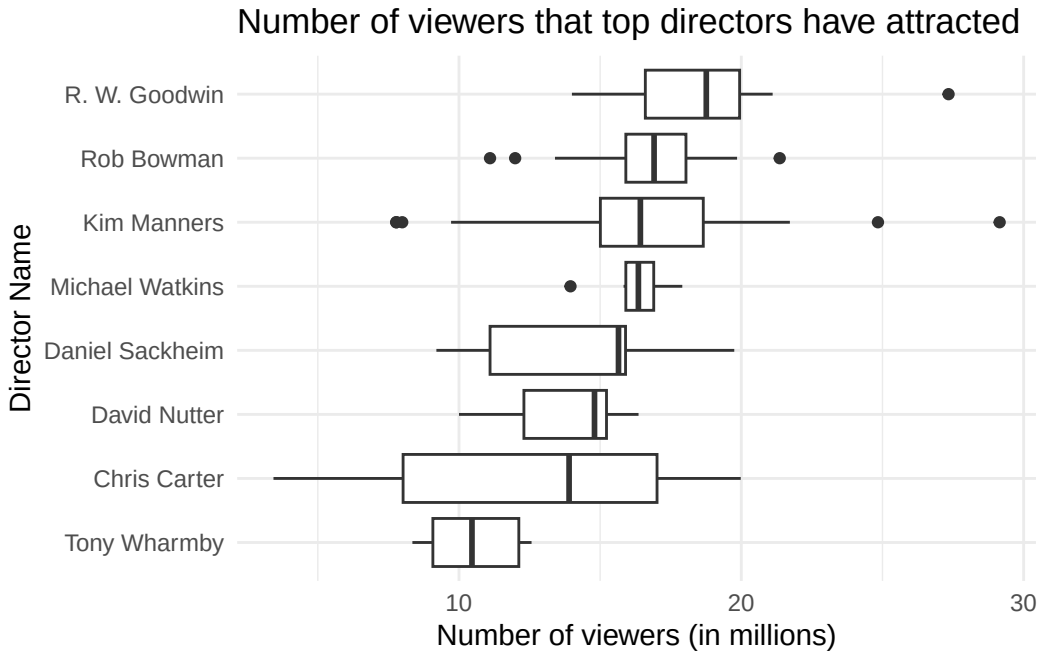
| Name of Director | Number of Episodes directed |
|---|---|
| Kim Manners | 52 |
| Rob Bowman | 33 |
| Chris Carter | 15 |
| David Nutter | 15 |
| R. W. Goodwin | 9 |
| Tony Wharmby | 7 |
| Michael Watkins | 6 |
| Daniel Sackheim | 5 |

## Displaying the data

Next I create a boxplot, ordered by the median (middle value), to show range of viewer numbers, for each of the directors.

```
directors_wanted <- names(most_active_directors)

data %>%
  select(title, viewers, director) %>%
  filter(director %in% directors_wanted) %>%
  mutate(viewers = as.double(
    str_replace(viewers, "\\[[0-9]+\\]", ""))
  ) %>%
  mutate(director = as.factor(director)) %>%
  mutate(g_id = group_indices(., director)) %>%
  ggplot(aes(
    x = fct_reorder(director, viewers, .fun = median),
    y = viewers,
  )) +
  geom_boxplot() +
  coord_flip() +
  # geom_smooth(method = lm, se = FALSE) +
  theme_minimal() +
  labs(
    title = "Number of viewers that top directors have attracted",
    x = "Director Name",
    y = "Number of viewers (in millions)",
  )
```


Number of viewers that top directors have attracted
Director Name
R. W. Goodwin
Rob Bowman
Kim Manners
Michael Watkins
Daniel Sackheim
David Nutter
Chris Carter
Tony Wharmby
10
20
30
Number of viewers (in millions)