HW4_Data-Visualisation.R

Bharathi

2025-03-30

```
##Part 3

#Load the library
library(ggplot2)
```

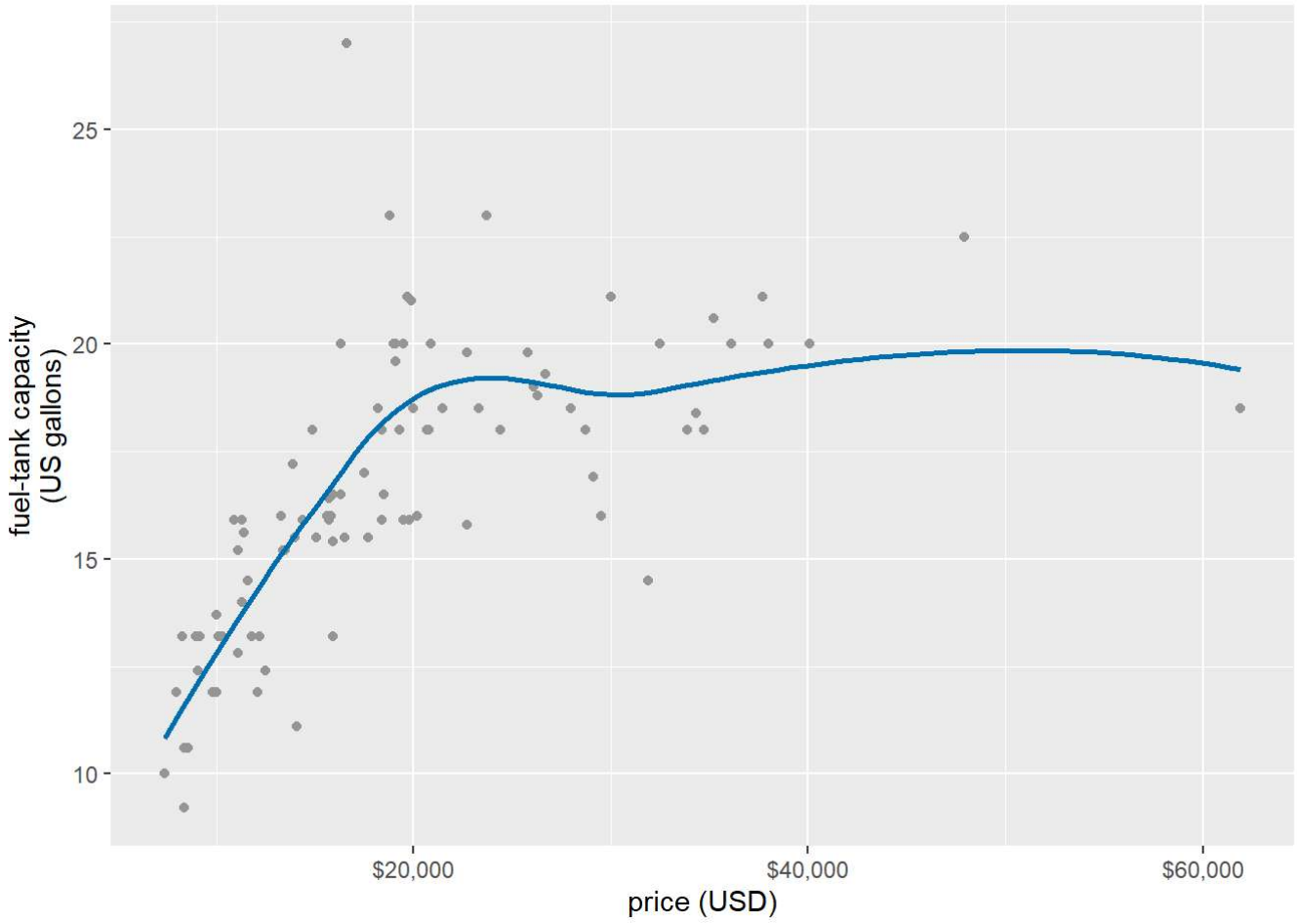
```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(MASS)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

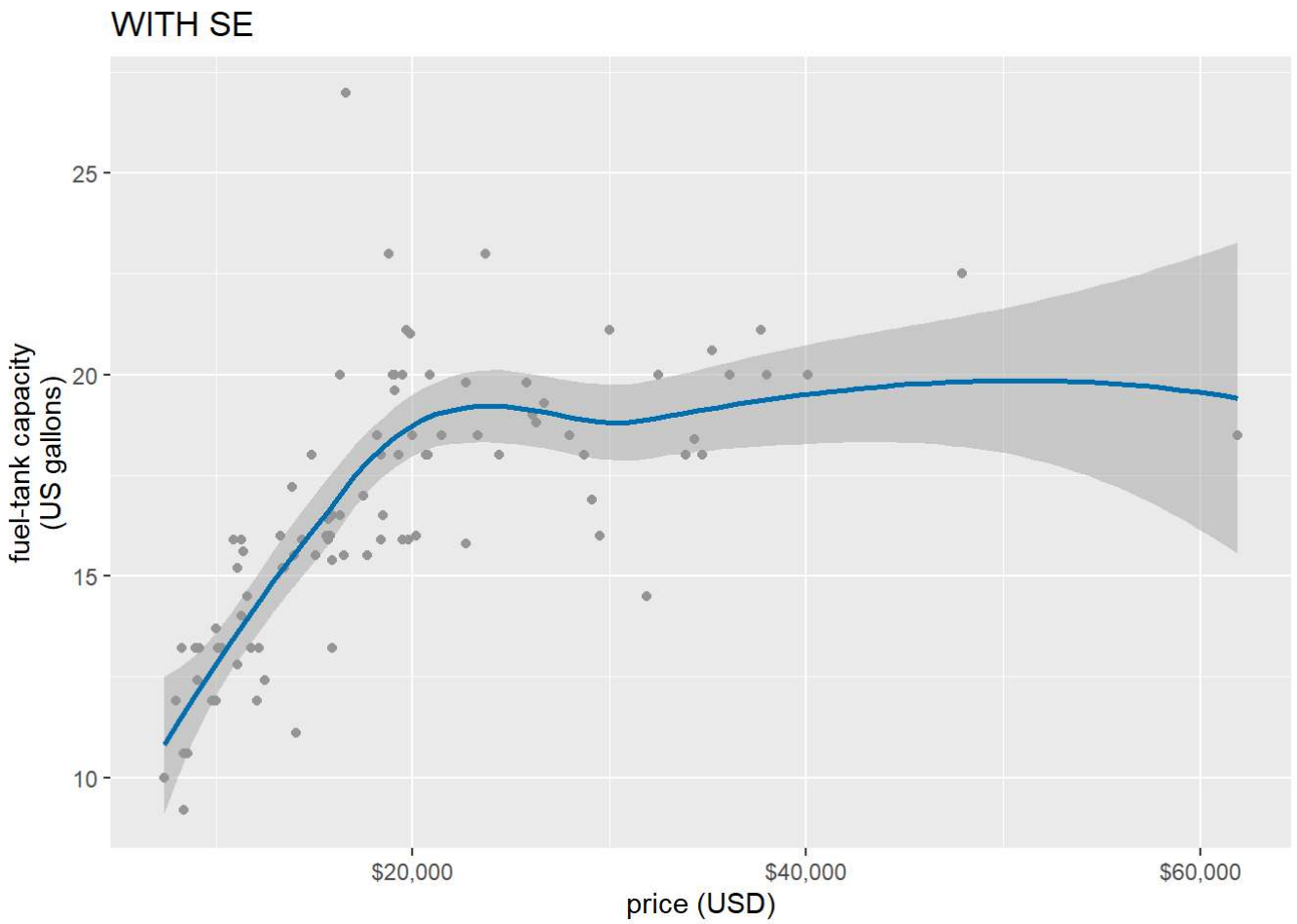
```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
#Load the dataset
cars93 = MASS::Cars93

ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = FALSE, method = "loess", formula = y ~ x, color = "#0072B2") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)")
```


fuel-tank capacity
(US gallons)
25
20
15
10
$20,000
$40,000
$60,000
price (USD)

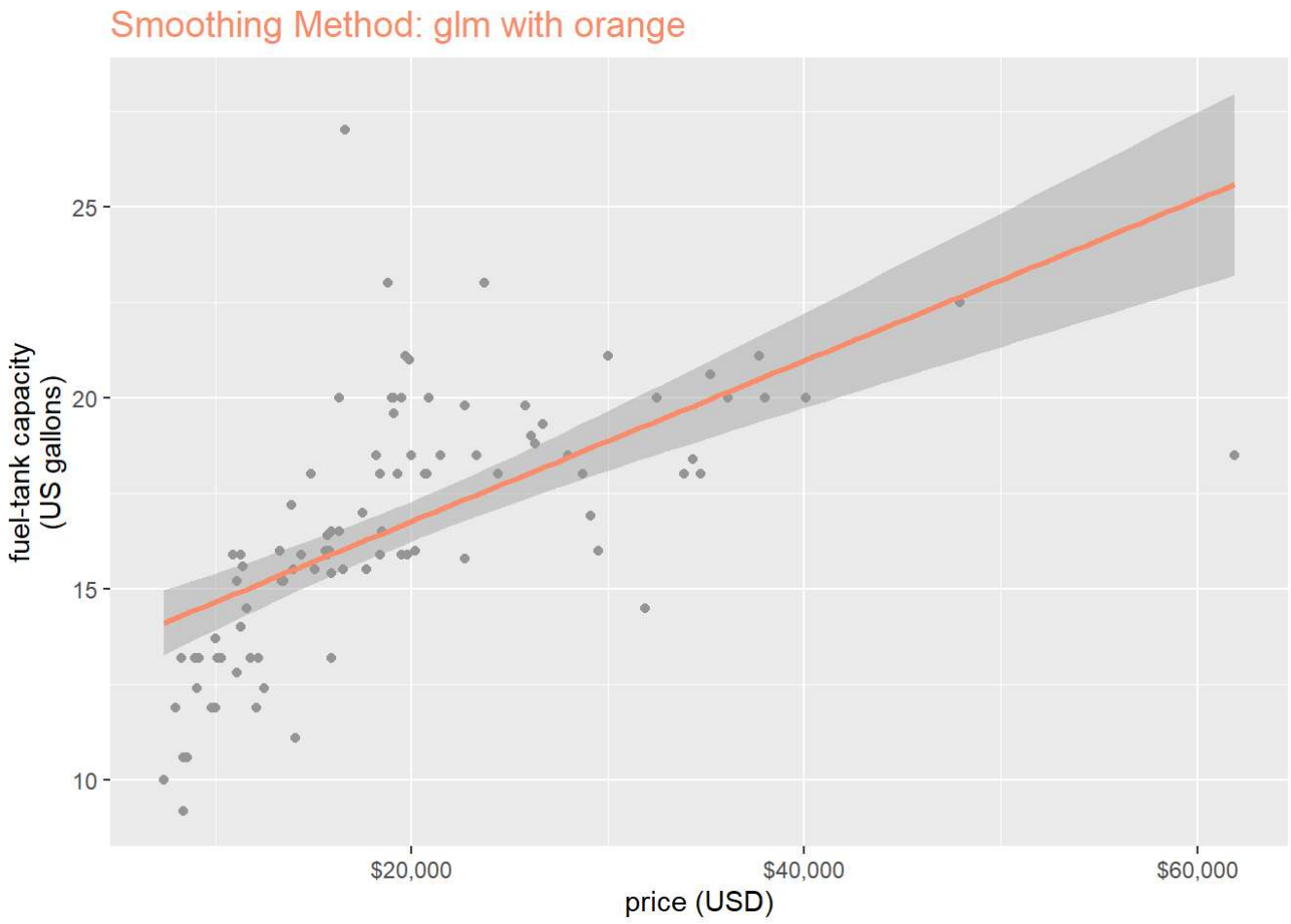
```
#(a) lm

plot_lm <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = FALSE, method = "lm", formula = y ~ x, color = "#0072B2") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: lm")


#(a) glm

plot_glm <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = FALSE, method = "glm", formula = y ~ x, color = "#0072B2") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: glm")

#(a) gam

plot_gam <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = FALSE, method = "gam", formula = y ~ x, color = "#0072B2") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: gam")


#(b)

ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = TRUE, method = "loess", formula = y ~ x, color = "#0072B2") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
```

```
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("WITH SE")
```

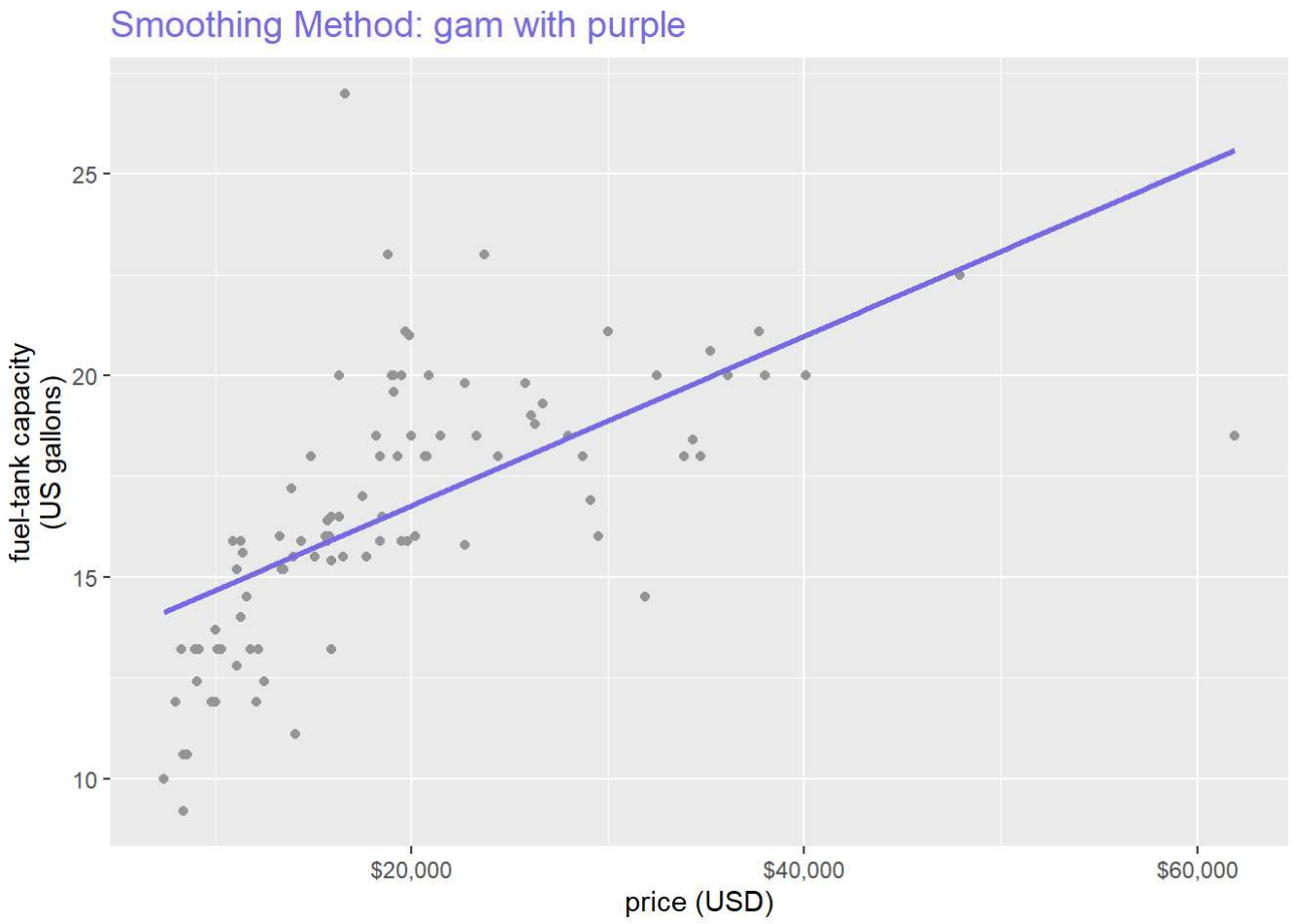



```
#(c)&(d)&(e) lm, using ggtitle

plot_lm_green <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = TRUE, method = "lm", formula = y ~ x, color = "#8fe388") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: lm with green") +
  theme(plot.title = element_text(size = 14, color = "#8fe388"))


print(plot_lm_green)
```


```
#(c)&(d)&(e) glm, using ggtitle

plot_glm_orange <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = TRUE, method = "glm", formula = y ~ x, color = "#fe8d6d") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: glm with orange") +
  theme(plot.title = element_text(size = 14, color = "#fe8d6d"))

print(plot_glm_orange)
```




```
#(c)&(d)&(e)gam, using ggtitle

plot_gam_purple <- ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
  geom_point(color = "grey60") +
  geom_smooth(se = FALSE, method = "gam", formula = y ~ x, color = "#7c6bea") +
  scale_x_continuous(
    name = "price (USD)",
    breaks = c(20, 40, 60),
    labels = c("$20,000", "$40,000", "$60,000")
  ) +
  scale_y_continuous(name = "fuel-tank capacity\n(US gallons)") +
  ggtitle("Smoothing Method: gam with purple") +
  theme(plot.title = element_text(size = 14, color = "#7c6bea"))

print(plot_gam_purple)
```




```
##Part 4
getwd()
```

```
## [1] "C:/Users/Bharathi/OneDrive/Documents/Data Visualisation/DATA"
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

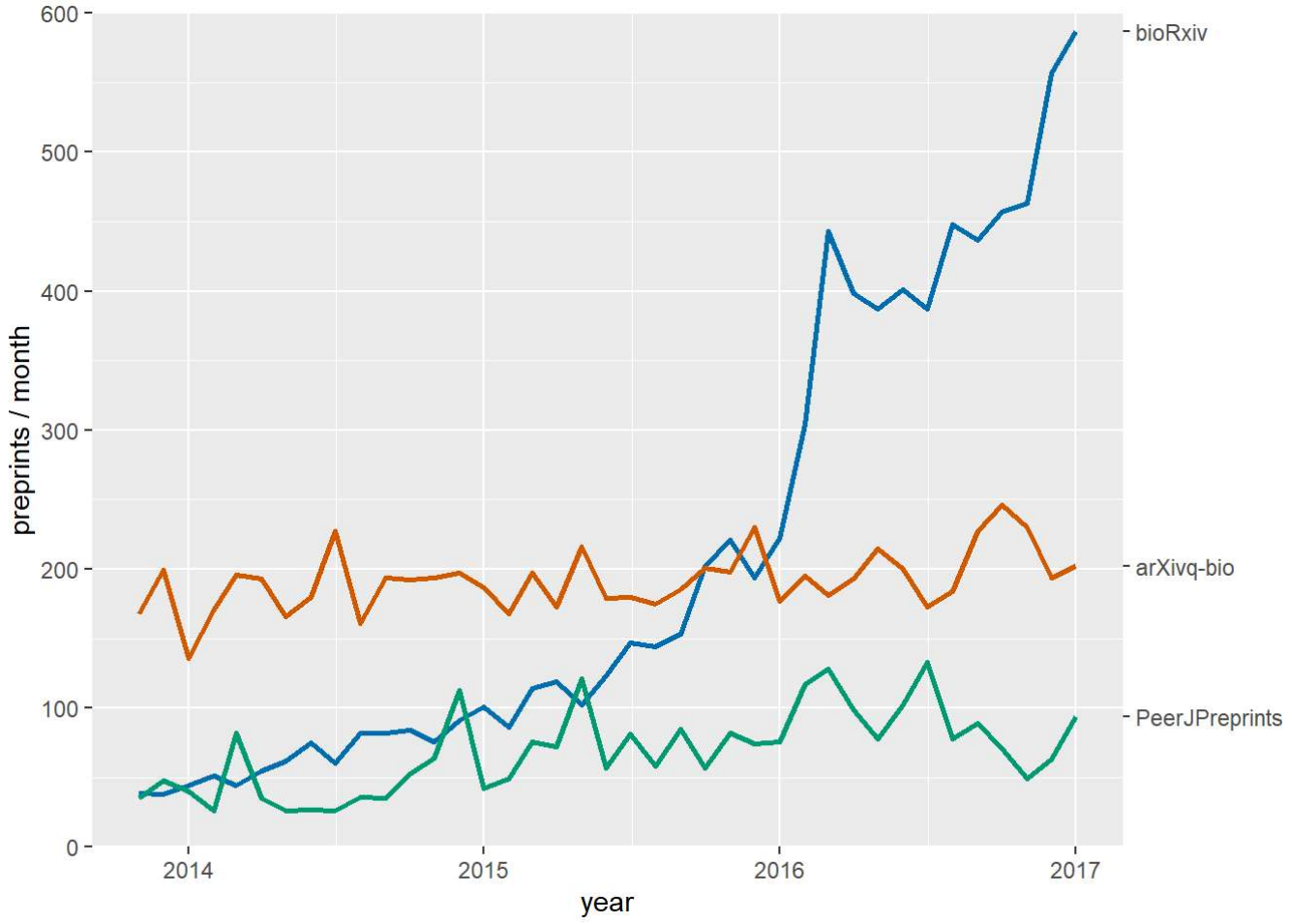
```
load("C:/Users/Bharathi/OneDrive/Documents/Data Visualisation/preprint_growth.rda") #please chan
ge the path if needed
head(preprint_growth)
```

```
## # A tibble: 6 × 3
##   archive           date       count
##   <chr>             <date>     <int>
## 1 arXiv q-bio       2007-01-01    40
## 2 Nature Precedings 2007-01-01     3
## 3 F1000Research     2007-01-01     0
## 4 PeerJ Preprints   2007-01-01     0
## 5 bioRxiv           2007-01-01     0
## 6 Winnower          2007-01-01     0
```

```
preprint_growth %>% filter(archive == "bioRxiv") %>%
  filter(count > 0) -> biorxiv_growth
preprints<-preprint_growth %>% filter(archive %in%
                                        c("bioRxiv", "arXiv q-bio", "PeerJ Preprints")) %>%filte
r(count > 0) %>%
  mutate(archive = factor(archive, levels = c("bioRxiv", "arXiv q-bio", "PeerJ Preprints")))

preprints_final <- filter(preprints, date == ymd("2017-01-01"))
ggplot(preprints) +
  aes(date, count, color = archive, fill = archive) +
  geom_line(size = 1) +
  scale_y_continuous(
    limits = c(0, 600), expand = c(0, 0),
    name = "preprints / month",
    sec.axis = dup_axis( #this part is for the second y axis
      breaks = preprints_final$count, #and we use the counts to position our labels
      labels = c("arXivq-bio", "PeerJPreprints", "bioRxiv"),
      name = NULL)
  ) +
  scale_x_date(name = "year",
               limits = c(min(biorxiv_growth$date), ymd("2017-01-01"))) +
  scale_color_manual(values = c("#0072b2", "#D55E00", "#009e73"),
                     name = NULL) +
  theme(legend.position = "none")
```

```
## 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.
```

```
## Warning: Removed 131 rows containing missing values or values outside the scale range
## (`geom_line()`).
```




```
# Part 4(a)

library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ forcats 1.0.0     ✓ stringr 1.5.1
## ✓ purrr   1.0.4     ✓ tibble  3.2.1
## ✓ readr   2.1.5     ✓ tidyr   1.3.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::select() masks MASS::select()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
head(preprint_growth)
```

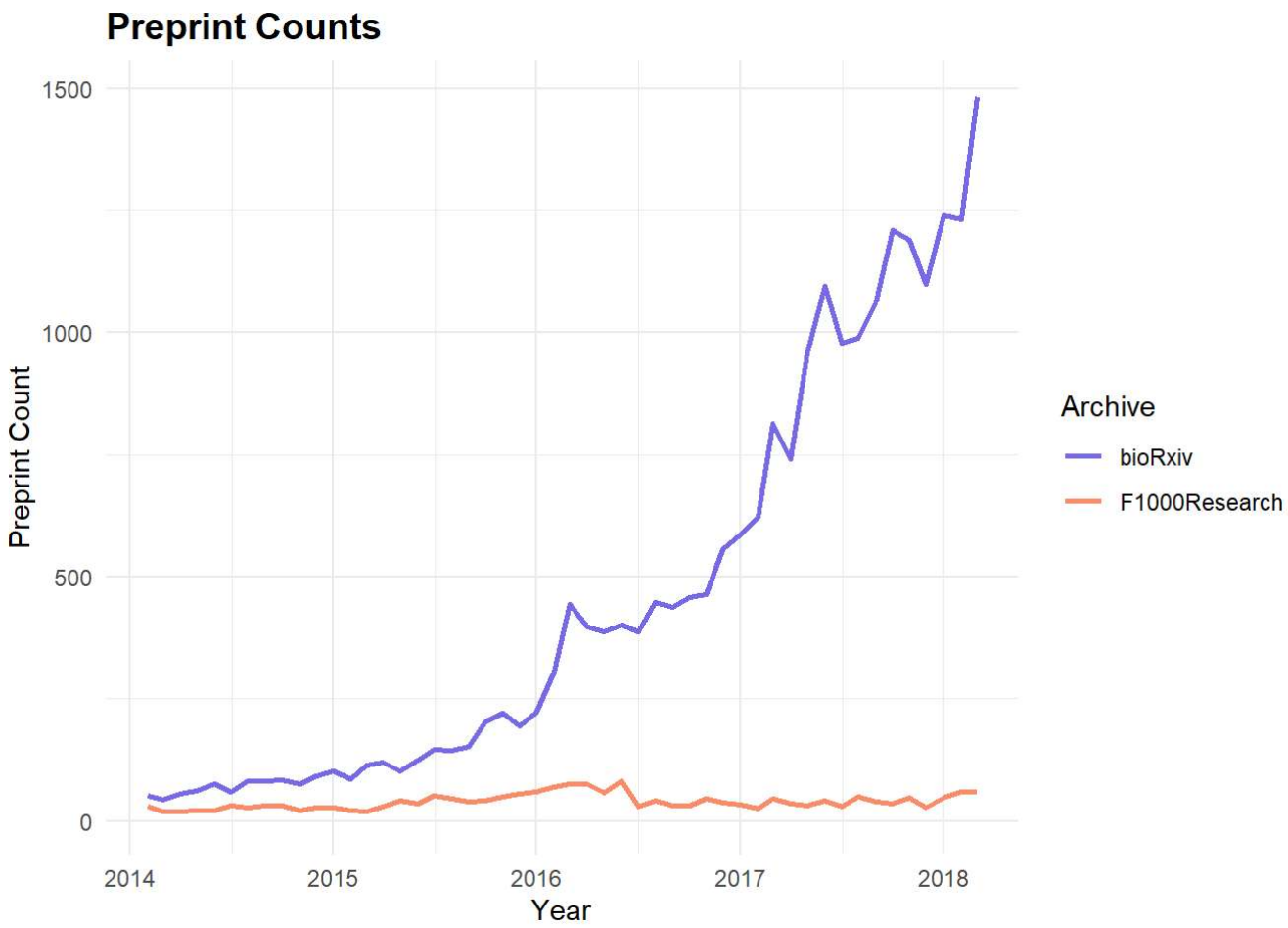
```
## # A tibble: 6 × 3
##   archive           date       count
##   <chr>             <date>     <int>
## 1 arXiv q-bio       2007-01-01    40
## 2 Nature Precedings 2007-01-01     3
## 3 F1000Research     2007-01-01     0
## 4 PeerJ Preprints   2007-01-01     0
## 5 bioRxiv           2007-01-01     0
## 6 Winnower          2007-01-01     0
```

```
preprint_full = preprint_growth %>%
  drop_na() %>%
  filter(count > 0, year(date) > 2004)

head(preprint_full)
```

```
## # A tibble: 6 × 3
##   archive           date       count
##   <chr>             <date>     <int>
## 1 arXiv q-bio       2007-01-01    40
## 2 Nature Precedings 2007-01-01     3
## 3 arXiv q-bio       2007-02-01    44
## 4 arXiv q-bio       2007-03-01    55
## 5 arXiv q-bio       2007-04-01    41
## 6 arXiv q-bio       2007-05-01    59
```

```
#Part 4(b)

preprint_filtered = preprint_full %>%
  filter(archive %in% c("bioRxiv","F1000Research"))

#Part 4(c) & (d) & (e) & (f)

# Create line graph
ggplot(preprint_filtered, aes(x = date, y = count, color = archive)) +
  geom_line(size = 1) +  # Line thickness
  scale_color_manual(values = c("bioRxiv" = "#7c6bea", "F1000Research" = "#fe8d6d")) +
  labs(title = "Preprint Counts",  # Updated title
       x = "Year",
       y = "Preprint Count",
       color = "Archive") +
  scale_x_date(limits = c(as.Date("2014-02-01"), max(preprint_filtered$date))) +  # X-axis start
s Feb 2014
  theme_minimal() +
  theme(plot.title = element_text(size = 14, face = "bold"),  # Title formatting
        legend.position = "right")  # Move legend to the right
```

```
## Warning: Removed 22 rows containing missing values or values outside the scale range
## (`geom_line()`).
```


Preprint Counts
1500
1000
500
0
Preprint Count
2014
2015
2016
2017
2018
Year
Archive
bioRxiv
F1000Research