HW4

JAYNA CLARK

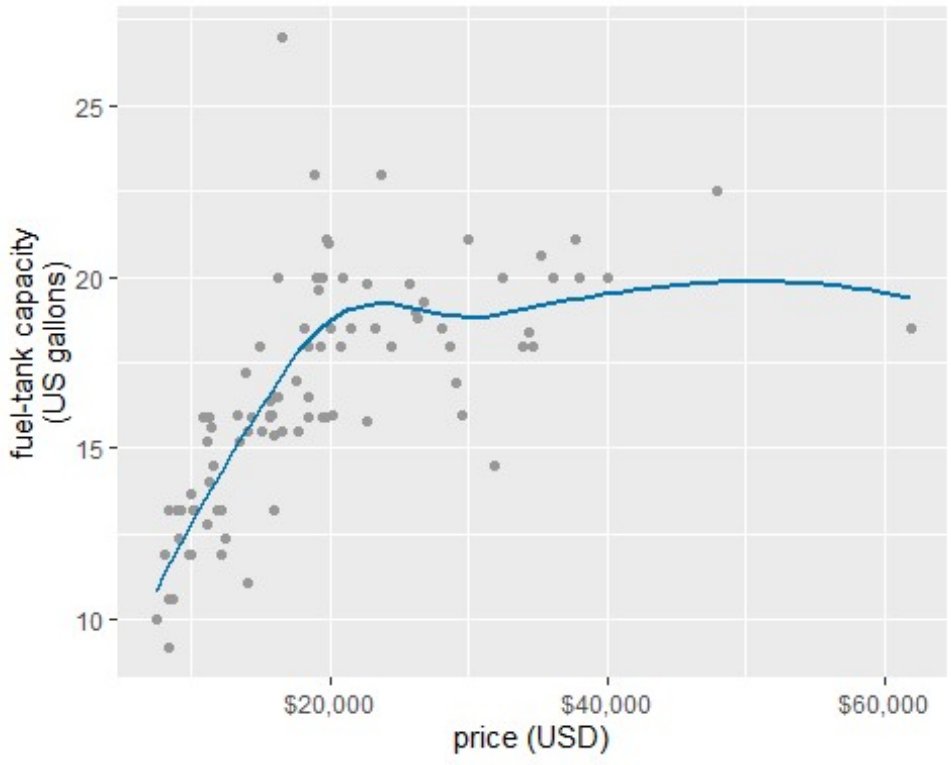
2025-03-30

```
library(ggplot2)
library(tidyr)
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(lubridate)

## Warning: package 'lubridate' was built under R version 4.4.2

##
## Attaching package: 'lubridate'

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

Number 3

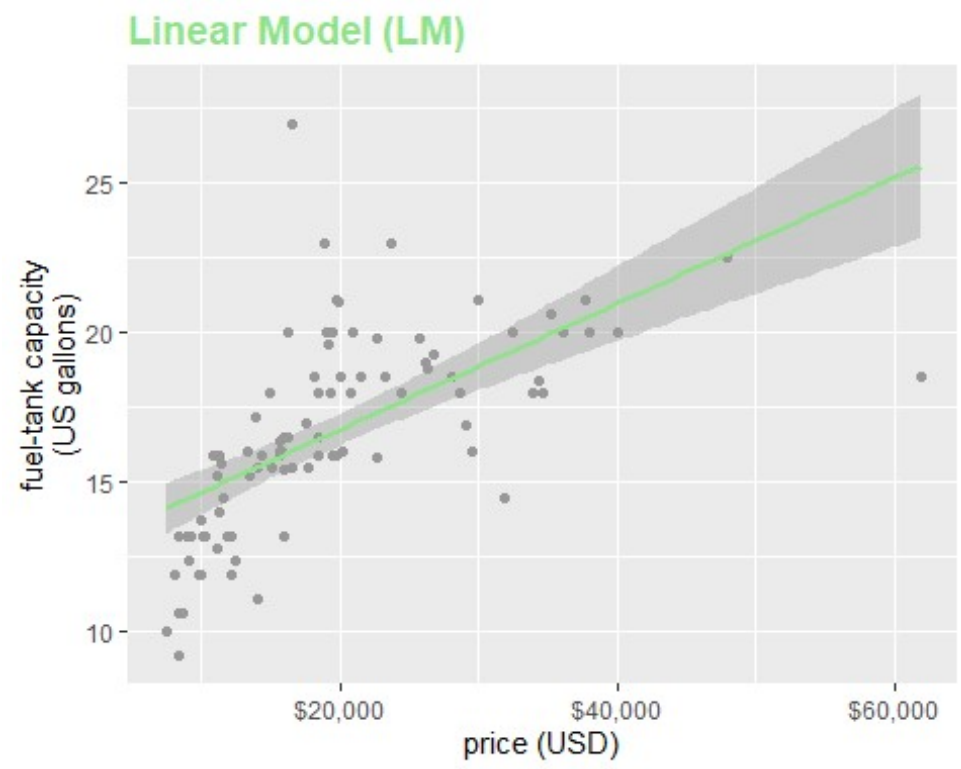
given code:

```
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)")
```

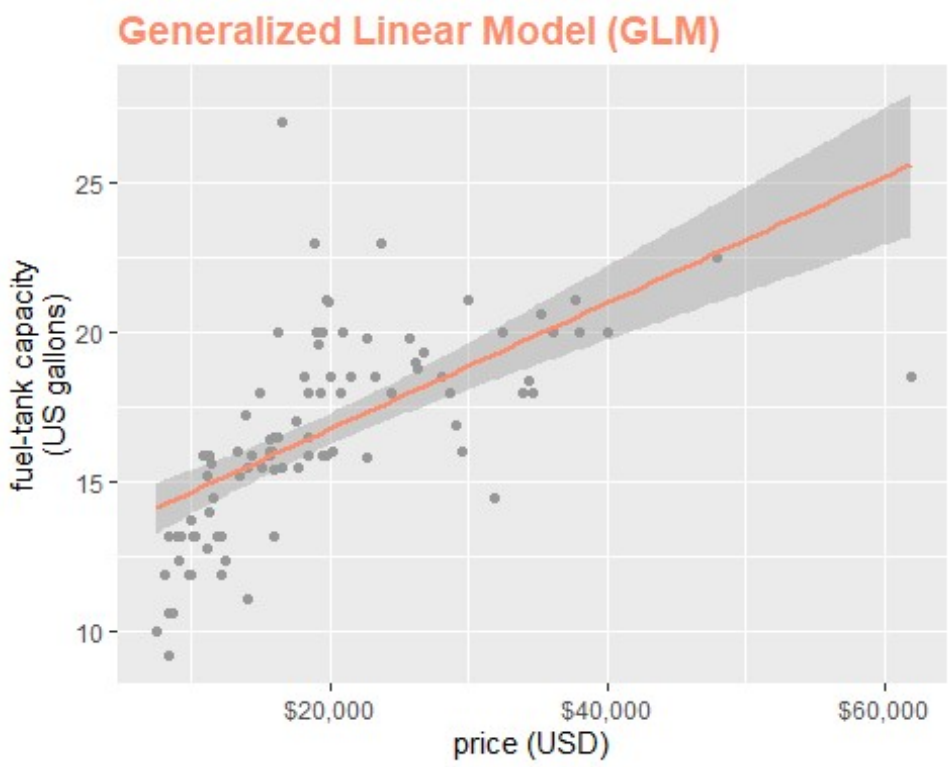


(a) Use ”lm”, ”glm”, ”gam” methods in the geom_smooth() function to create three figures. (b) Set the se parameter to TRUE to show the standard error (shaded area around the fitted line) (c) For every method above change the color of the line with the following color codes: #8fe388, #fe8d6d, #7c6bea (d) Please search for the method to add a title to your ggplot figure and add titles for each figure to indicate the method that you used for smoothing. (e) Please search for the theme() function for ggplot and change the font size of the titles to 14 and match their colors with the line colors you used above.

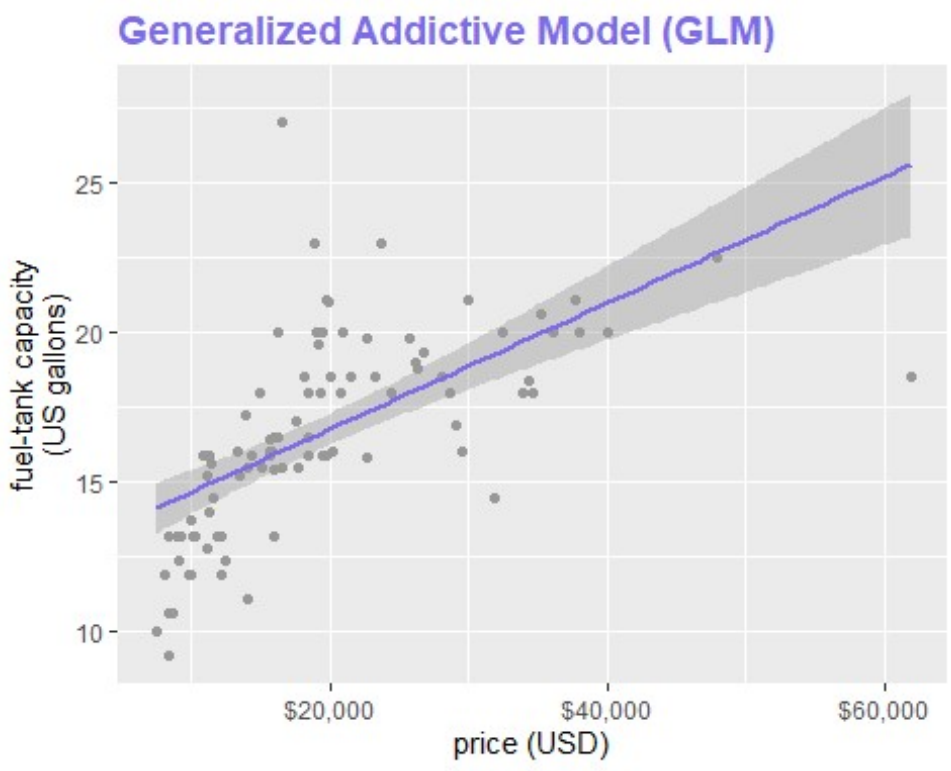
```
#Lm
ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
geom_point(color = "grey60") +
geom_smooth(se = TRUE, method = "lm", formula = y ~ x, color = "#8fe388") +
ggtitle("Linear Model (LM)")+
scale_x_continuous(
name = "price (USD)",
breaks = c(20, 40, 60),
labels = c("$20,000", "$40,000", "$60,000")) +
scale_y_continuous(name = "fuel-tank capacity\n(US gallons)")+
  theme(
    plot.title = element_text(size = 14, color = "#8fe388", face = "bold")
  )
```




```r
#glm
ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity))+
geom_point(color = "grey60") +
geom_smooth(se = TRUE, method = "glm", formula = y ~ x,  color = "#fe8d6d") +
ggtitle("Generalized Linear Model (GLM)")+
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)")+
  theme(
    plot.title = element_text(size = 14, color = "#fe8d6d", face = "bold")
  )
```




```r
#gam
ggplot(cars93, aes(x = Price, y = Fuel.tank.capacity)) +
geom_point(color = "grey60") +
geom_smooth(se = TRUE, method = "gam", formula = y ~ x, color = "#7c6bea") +
ggtitle("Generalized Addictive Model (GLM)")+
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)")+
    theme(
    plot.title = element_text(size = 14, color = "#7c6bea", face = "bold")
  )
```



Number 4

Please inspect the following code which can be also found in TimeSeries_Trends.R and try to run how it generates three time series in a single plot. Then, perform the steps in the following bullet points. Please recall that %>% is called the pipe operator which passes the output of previous step to the next step.

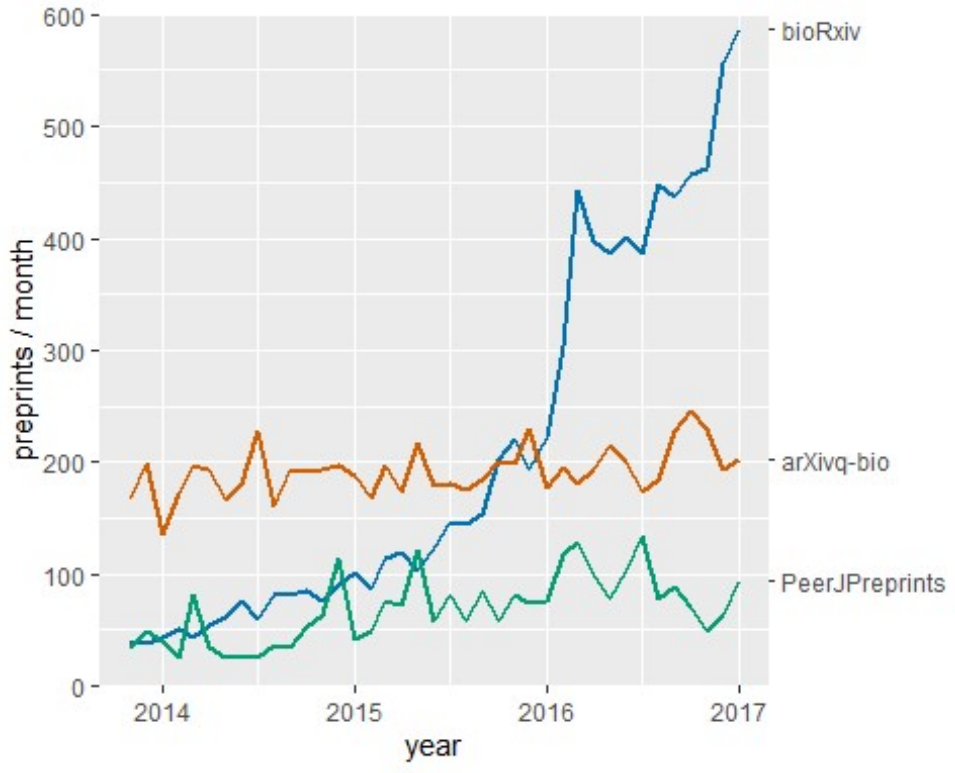
given code

```
load("./preprint_growth.rda") #please change 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")) %>%filter(count > 0) %>%
mutate(archive = factor(archive, levels = c("bioRxiv", "arXiv q-bio", "PeerJ
Preprints")))
```

```r
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()`).
```



(a) By using drop_na( ) and filter( ) on preprint_growth data frame, get the rows which have count greater than 0 and year later than 2004, and output it to another data frame called preprint_full.

(b) Use the filter function again to select the rows that have ”bioRxiv”, ”F1000Research” in it only by looking at the example in the code above.

(c) Draw line graphs for these two time series, ”bioRxiv” and ”F1000Research”, by coloring them with ”#7c6bea” and ”#fe8d6d”.

(d) Put the legend to the right of the figure.

(e) For the x-axis, start the values from Feb 2014.

(f) Add a title ”Preprint Counts” to the figure.

```
#a
preprint_full <- preprint_growth %>%
  drop_na() %>%
  filter(count > 0, year(date) > 2004)

#b
preprints_selected <- preprint_full %>%
  filter(archive %in% c("bioRxiv", "F1000Research"))

#c,d,e and f
ggplot(preprints_selected) +
  aes(x = date, y = count, color = archive) +
  geom_line(size = 1) +
  theme(legend.position = "right") +
  scale_x_date(name = "Year", limits = c(ymd("2014-02-01"),
```

```
max(preprints_selected$date))) +
  scale_color_manual(values = c("bioRxiv" = "#7c6bea", "F1000Research" =
"#fe8d6d"), name = "Archive") +
  ggtitle("Preprint Counts") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"))

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