Shubha, Kenna, Rakeb, Suthi - A07 - CSC 285 - 17th Jan 2024

Bike Sharing Data: Visualizing amounts

Loading the data

```
data <- read.csv(file="/srv/R/CSC285_public/Shubha-Rakeb-Suthi-Kenna/BikeShare.txt",
                 sep="\t", quote="", comment.char="")
```

Scenario

Data description from Kaggle (https://www.kaggle.com/marklvl/bike-sharing-dataset).Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back to another position. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system in Washington, DC with the corresponding weather and seasonal information.

BikeShare.txt has the following fields

- **instant**: record index
- **dteday** : date
- **season** : season (1:springer, 2:summer, 3:fall, 4:winter)
- **yr** : year (0: 2011, 1:2012)
- **mnth** : month (1 to 12)
- **holiday** : weather day is holiday or not
- **weekday** : day of the week
- **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit**
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- **temp** : Normalized temperature in Celsius. The values are divided to 41 (max)
- **atemp*** : Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- **hum** : Normalized humidity. The values are divided to 100 (max)
- **windspeed** : Normalized wind speed. The values are divided to 67 (max)
- **casual** : count of casual users
- **registered** : count of registered users
- **cnt** : count of total rental bikes including both casual and registered

Create an RMarkdown where you use ggplot2 to create the visualizations specified below with "BikeShare.txt". For each graph listed below, include,

-If you create a new dataframe where averages are calculated, have RMarkdown display the averages you calculated. Regular R output will suffice!

-The graph

-A description of the graph that would serve as the caption. Describe the graph to the reader and tell them what you want them to notice when looking at the graph (essentially a caption).**

-Put this directly below the graph in your RMarkdown

-A brief description of your design choices, or if you think a different graph would be more appropriate.

Questions

1. Create a bar chart that shows the average number of bikes rented per day by season.

a. Note that you will need to create a new data frame with the averages. Write a loop in R to create this new data frame, then create the plot.

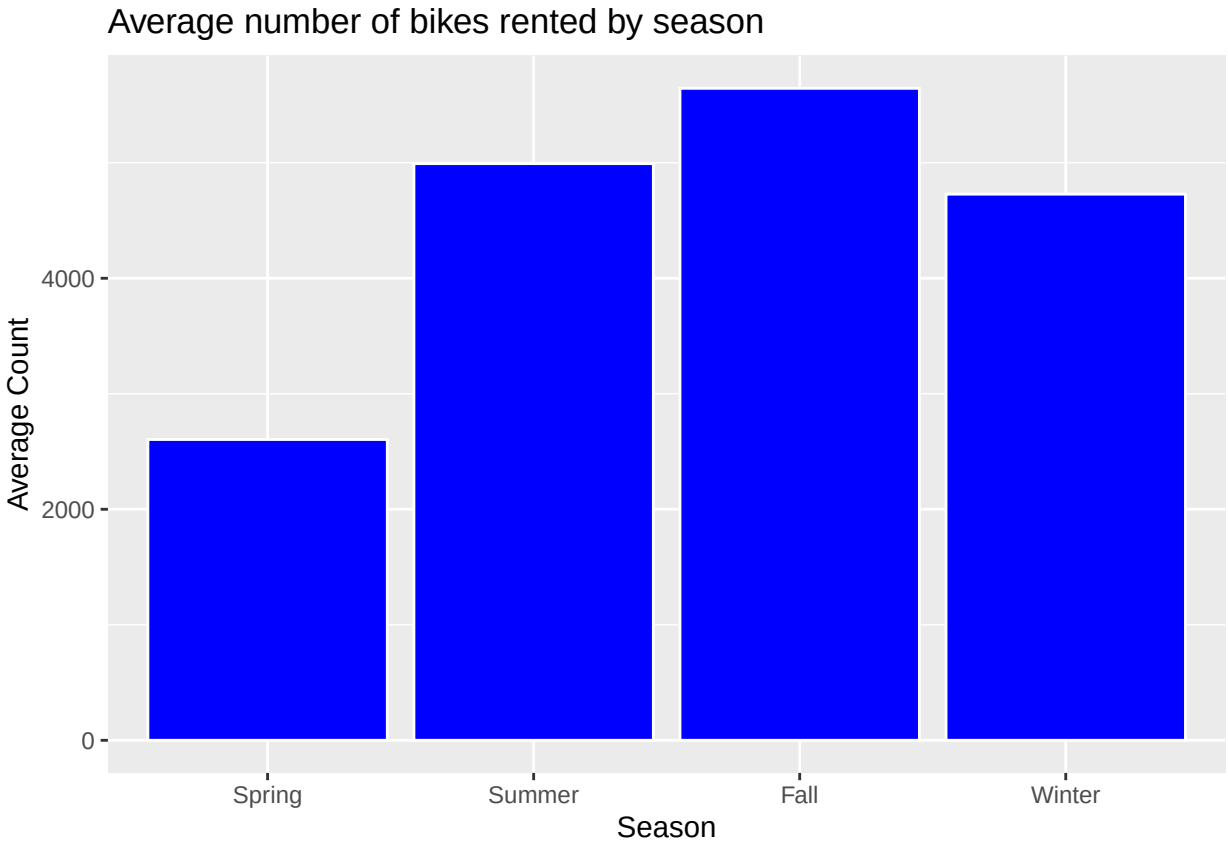
b. Print the output that shows your averages in the RMarkdown

```
#converting number strings into integers
data$cnt <- as.numeric(data$cnt)
data$season <- as.numeric(data$season)

#calculating average number of bikes per season
avg_bikes_by_season <- aggregate(cnt ~ season, data, mean)
avg_bikes_by_season
```

```
##   season      cnt
## 1      1 2604.133
## 2      2 4992.332
## 3      3 5644.303
## 4      4 4728.163
```

```
#creating a bar graph
ggplot(avg_bikes_by_season, aes(x = factor(season), y = cnt)) +
  geom_bar(stat = "identity", fill = "blue", color = "white") +
  labs(x = "Season", y = "Average Count",
       title = "Average number of bikes rented by season") +
  scale_x_discrete(labels=c('Spring', 'Summer', 'Fall', 'Winter'))
```



This bar chart depicts the average number of bikes written during the seasons of the year, showing the most bikes are rented in Fall and the least in Spring. There weren't that many design choices for this graph, but we made the bars blue and labelled the seasons so the graph would be easier to analyze.

2. Additionally, the city might be interested in registered vs casual users by season. Create a stacked bar chart that shows the average number of bike rentals by user type and season.

a. Note that you will need to create a new data frame with the averages. Write a loop in R to create this new data frame, then create the plot.

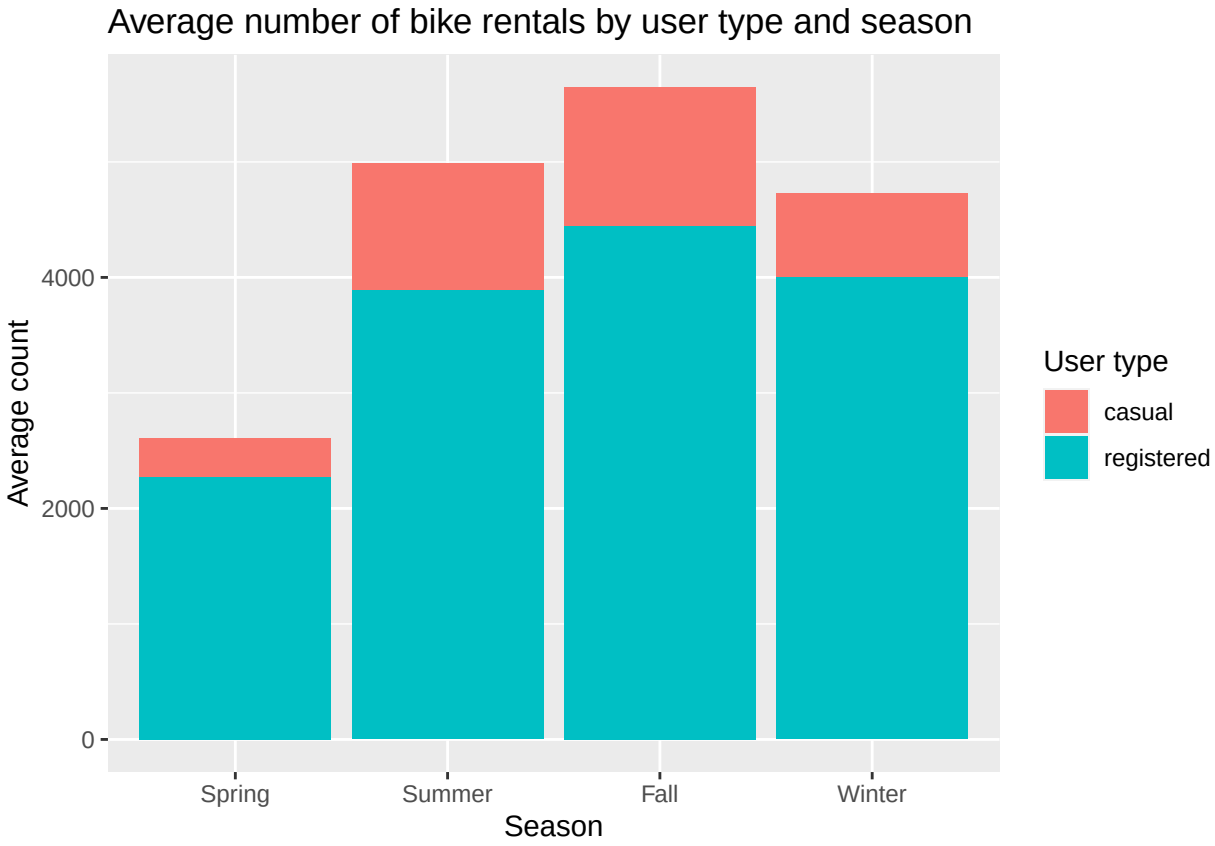
```
#calculating average number of bikes by user and season
avg_bikes_by_user_season <- aggregate(cbind(casual, registered) ~ season, data, mean)

#reshape for plotting
avg_bikes_by_user_season_long <- reshape2::melt(avg_bikes_by_user_season, id.vars = "season")

avg_bikes_by_user_season_long
```

```
##   season   variable     value
## 1      1     casual  334.9282
## 2      2     casual 1106.0978
## 3      3     casual 1202.6117
## 4      4     casual  729.1124
## 5      1 registered 2269.2044
## 6      2 registered 3886.2337
## 7      3 registered 4441.6915
## 8      4 registered 3999.0506
```

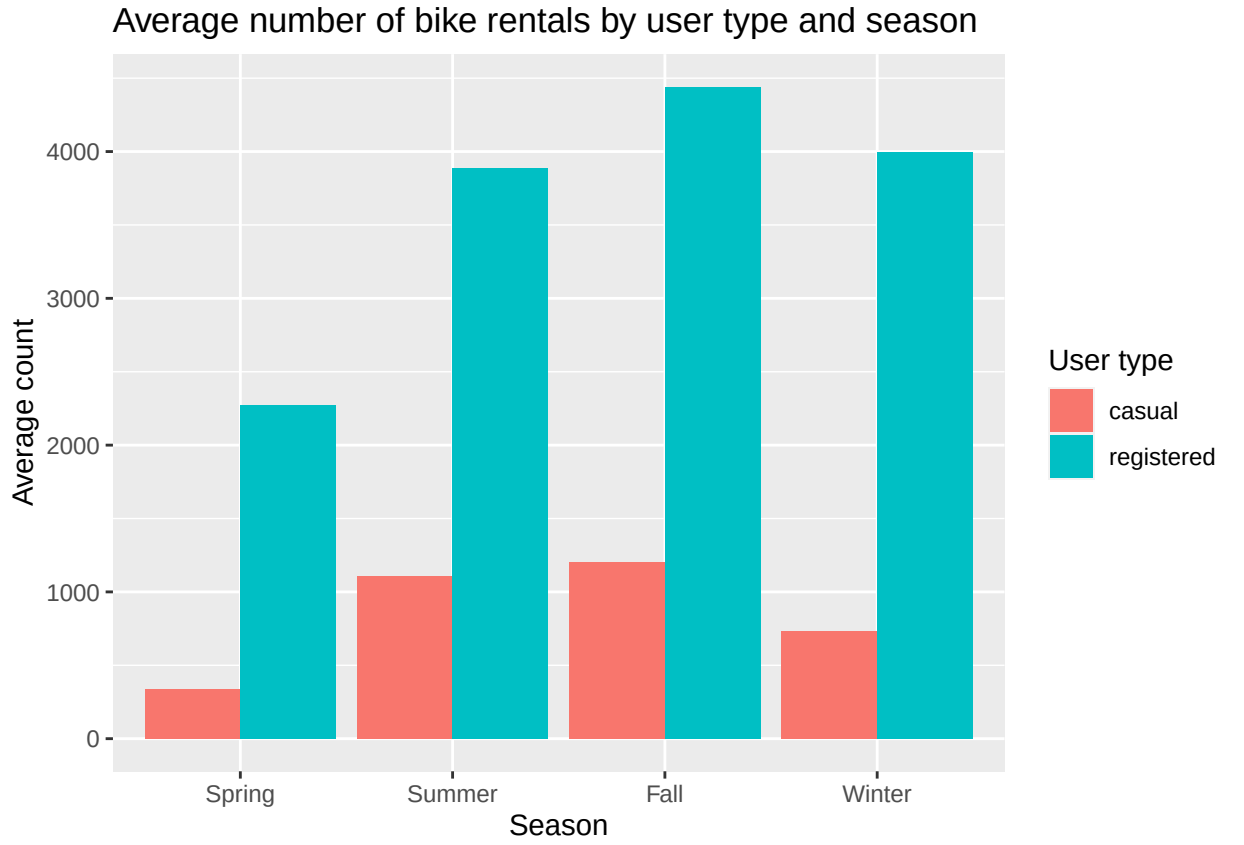
```
#creating a stacked bar graph
ggplot(avg_bikes_by_user_season_long,
       aes(x = factor(season), y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Season", y = "Average count",
       fill = "User type",
       title = "Average number of bike rentals by user type and season") +
  scale_x_discrete(labels=c('Spring', 'Summer', 'Fall', 'Winter'))
```



This graph goes into further detail regarding user trends during the various seasons of the year. It depicts the same overall averages during the seasons as the last graph, but divides the recorded averages into casual users and registered users. In all seasons, the registered users far outnumber the casual. When designing this graph, we used two near complementary colors to really portray the contrast between casual and registered users. This data is easy to deduce using the stacked bar graph format because it gives the total averages on top of the split between user types.

```
#grouped bar chart by changing the position to doge
ggplot(avg_bikes_by_user_season_long, aes(x = factor(season), y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Season", y = "Average count",
       fill = "User type",
       title = "Average number of bike rentals by user type and season") +
  scale_x_discrete(labels=c('Spring', 'Summer', 'Fall', 'Winter'))
```

3. Using the same data from second question above, create a grouped bar chart.



This graph depicts the same variables as the last graph, but in a grouped rather than stacked format. The takeaway is the same, but it may be easier for some to visualize the disparity between casual and registered users when looking at the bars side by side.

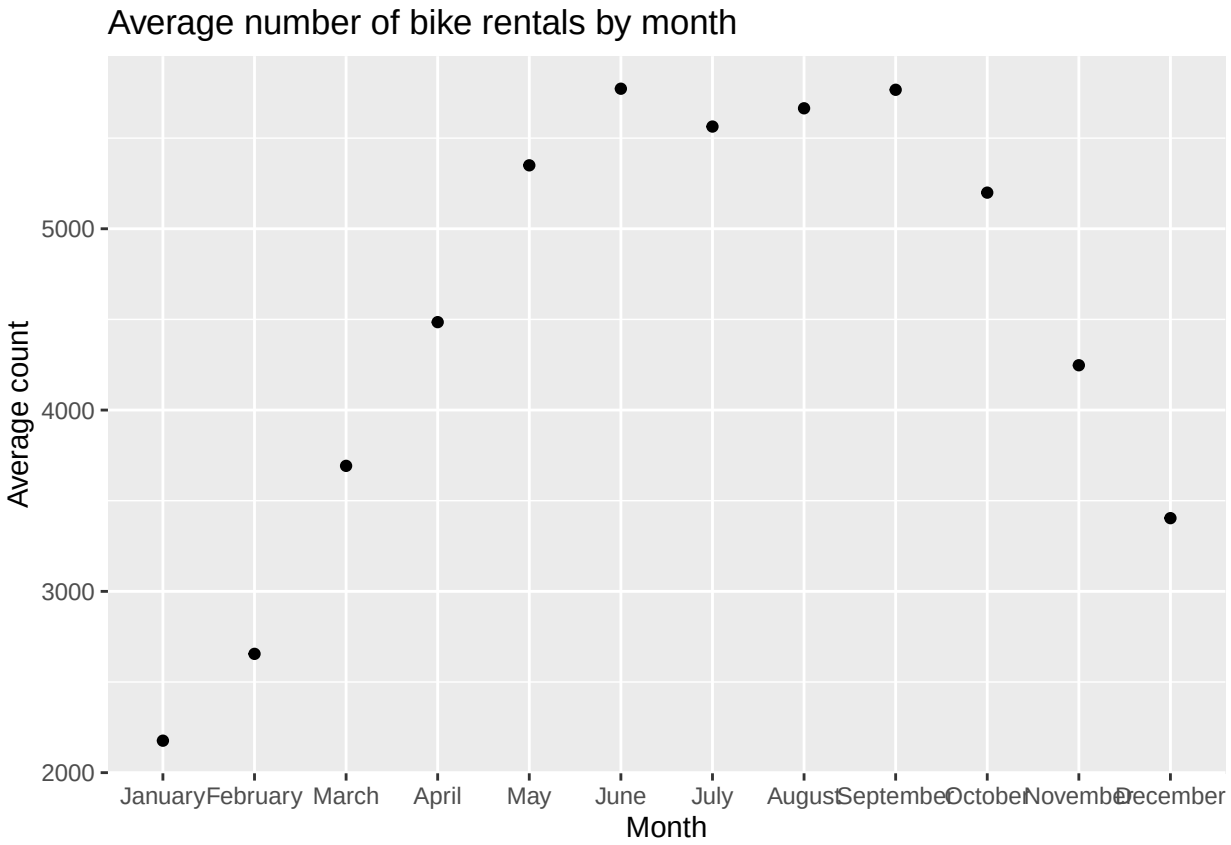
4. Create a dot plot that shows the average number of bike rentals per month.

a. Note that you will need to create a new data frame with the averages. Write a loop in R to create this new data frame, then create the plot.

```
#calculating average number of bikes per moth
avg_bikes_by_month <- aggregate(cnt ~ mnth, data, mean)
avg_bikes_by_month
```

```
##    mnth      cnt
## 1     1 2176.339
## 2     2 2655.298
## 3     3 3692.258
## 4     4 4484.900
## 5     5 5349.774
## 6     6 5772.367
## 7     7 5563.677
## 8     8 5664.419
## 9     9 5766.517
## 10   10 5199.226
## 11   11 4247.183
## 12   12 3403.806
```

```
#creating a dot plot
ggplot(avg_bikes_by_month, aes(x = factor(mnth), y = cnt)) +
  geom_point() +
  labs(x = "Month", y = "Average count",
       title = "Average number of bike rentals by month") +
  scale_x_discrete(labels=c( 'January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
```



This dot plot depicts the average number of bike rentals by month, rather than season. It shows that bike usage dips low during the winter and early spring, but stays pretty high and steady during summer and fall. Depicting this data with a dot plot over a bar chart showcases the linear trends in the data as it moves month to month, but a bar chart may be better to visualize the total averages.

5. Create a (mini) heat map that shows the median number of bike rentals by year and month.

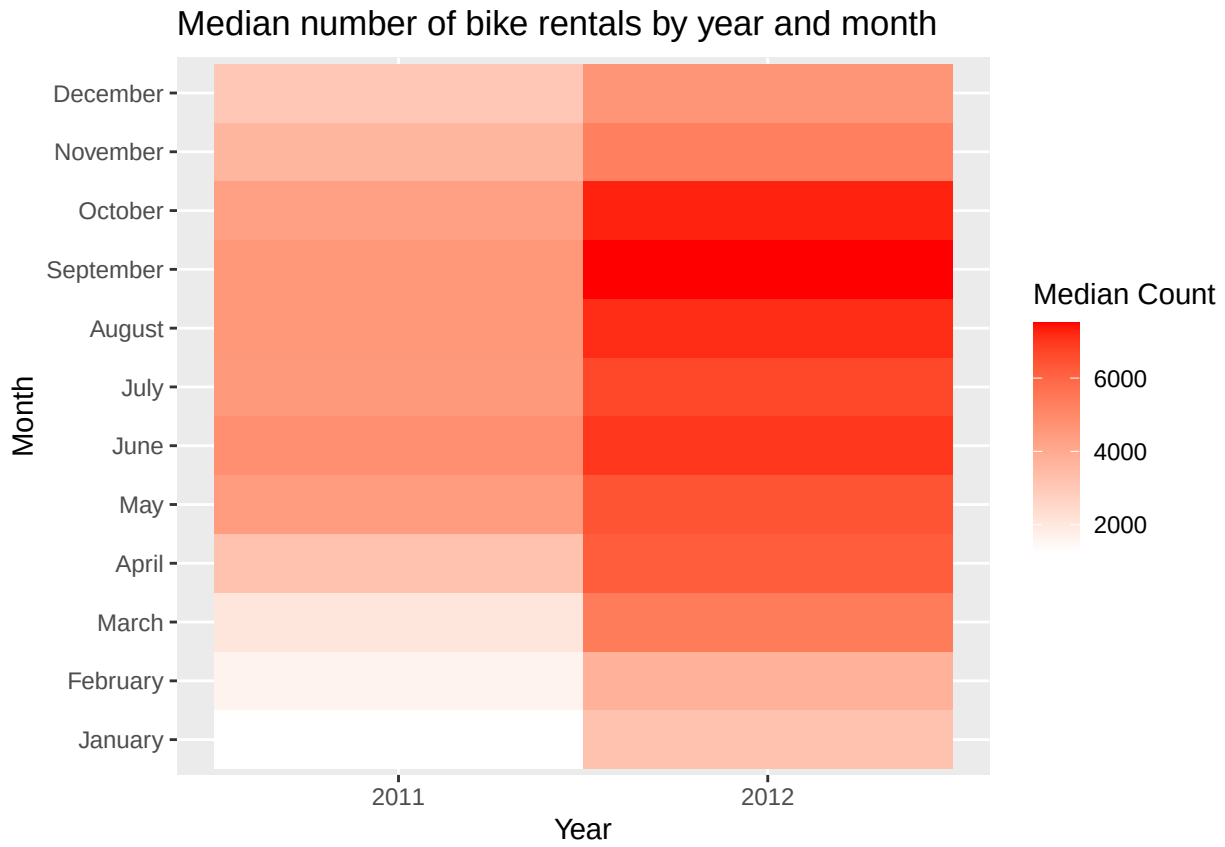
a. Note that you will need to create a new data frame with the medians. Write a loop in R to create this new data frame, then create the plot.

```
#calculating median number of bikes per month and year
median_bikes_by_year_month <- aggregate(cnt ~ yr + mnth, data, median)
median_bikes_by_year_month
```

```
##    yr mnth    cnt
## 1   0    1 1248.0
## 2   1    1 3243.0
## 3   0    2 1629.0
## 4   1    2 3777.0
## 5   0    3 2077.0
## 6   1    3 5382.0
```

```
## 7   0    4 3226.5
## 8   1    4 6214.5
## 9   0    5 4451.0
## 10  1    5 6421.0
## 11  0    6 4839.5
## 12  1    6 6988.0
## 13  0    7 4541.0
## 14  1    7 6685.0
## 15  0    8 4602.0
## 16  1    8 7148.0
## 17  0    9 4584.5
## 18  1    9 7514.5
## 19  0   10 4304.0
## 20  1   10 7282.0
## 21  0   11 3631.0
## 22  1   11 5319.0
## 23  0   12 3068.0
## 24  1   12 4649.0
```

```
#creating a heat map using tile function from ggplot2
ggplot(median_bikes_by_year_month, aes(x = factor(yr), y = factor(mnth),
                                       fill = cnt)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(x = "Year", y = "Month",
       fill = "Median Count",
       title = "Median number of bike rentals by year and month") +
  scale_x_discrete(labels=c('2011', '2012')) +
  scale_y_discrete(labels=c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', '
```

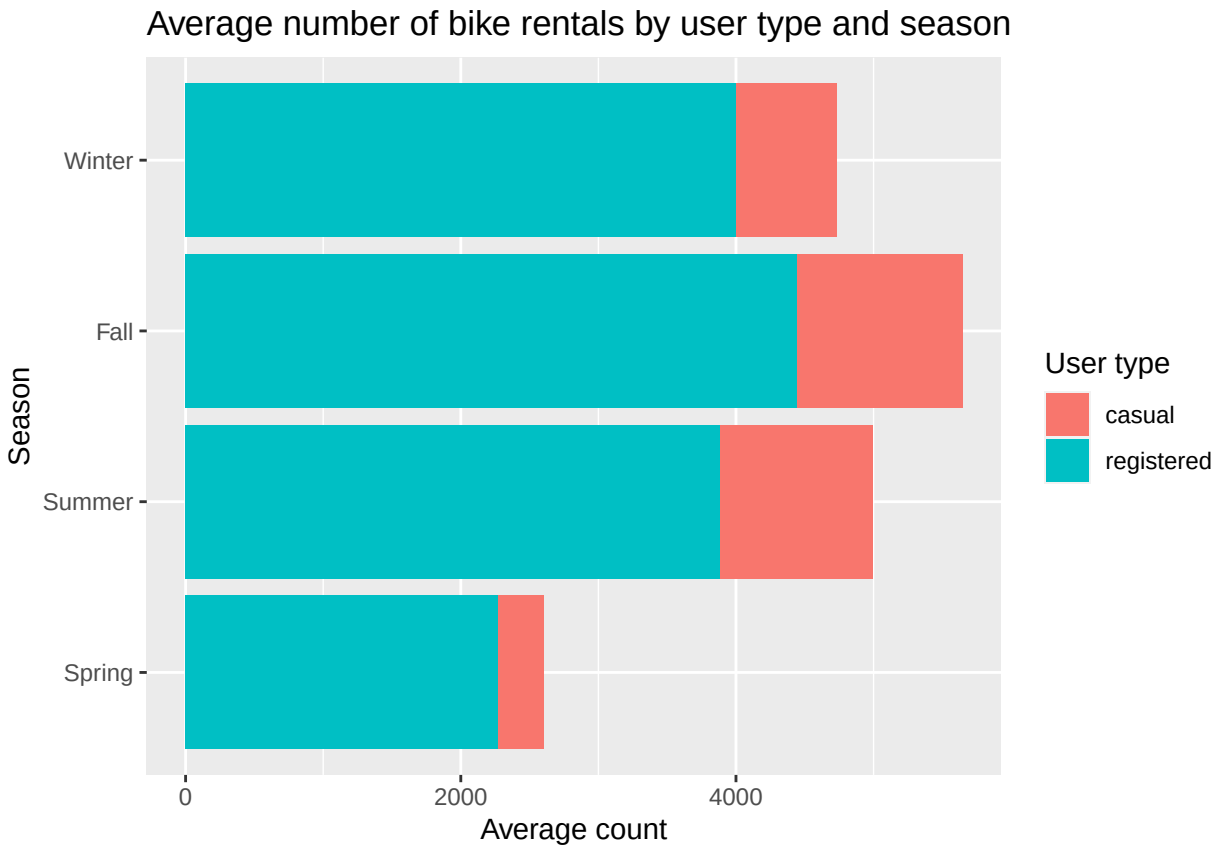


This heat map actually showcases a lot of trends. Not only does it showcase the times of the year when bikes are more popular to rent, from summer to fall, but it also shows that the median number of bikes rented during the year as a whole increased from 2011 to 2012. This data is well-represented with a heat map as it reflects the continuous nature of the count variable. The orange/red color is a classic for a heatmap and really easily presents the concentration of bike usage during the high-use months.

Other changes to your plots:

```
# Converting the stacked bar graph by flipping the axes
ggplot(avg_bikes_by_user_season_long,
       aes(x = factor(season), y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Season", y = "Average count",
       fill = "User type",
       title = "Average number of bike rentals by user type and season") +
  scale_x_discrete(labels=c('Spring', 'Summer', 'Fall', 'Winter')) +
  coord_flip()
```

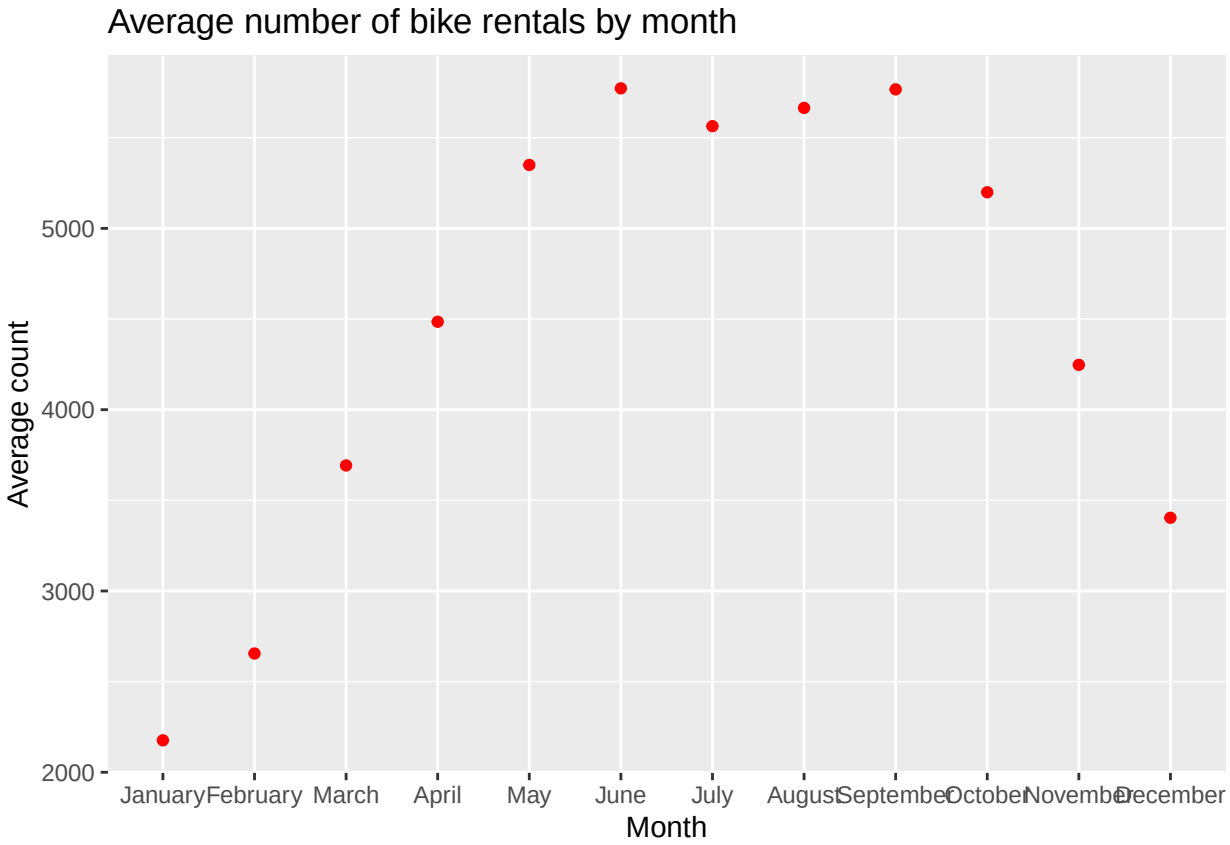
6. Create at least one bar chart where the x and y axis are flipped. If you already did this, simply tell me which graph. If you did not flip the axes in any of the original plots, create a new plot



now.

```
# Changing colors of dots in dot plot
ggplot(avg_bikes_by_month, aes(x = factor(mnth), y = cnt)) +
  geom_point(color="red") +
  labs(x = "Month", y = "Average count",
       title = "Average number of bike rentals by month") +
  scale_x_discrete(labels=c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', '
```

7. Change the colors from default in at least one graph. If you already did this, simply tell me which graph. If you did not flip the axes in any of the original plots, create a new plot now.


Average number of bike rentals by month
Average count
5000
4000
3000
2000
January
February
March
April
May
June
July
August
September
October
November
December
Month