Shubha, Kenna, Rakeb, Suthi - A08 - CSC 285 - 18th Jan 2024

Bike Sharing Data: Visualizing Distributions

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

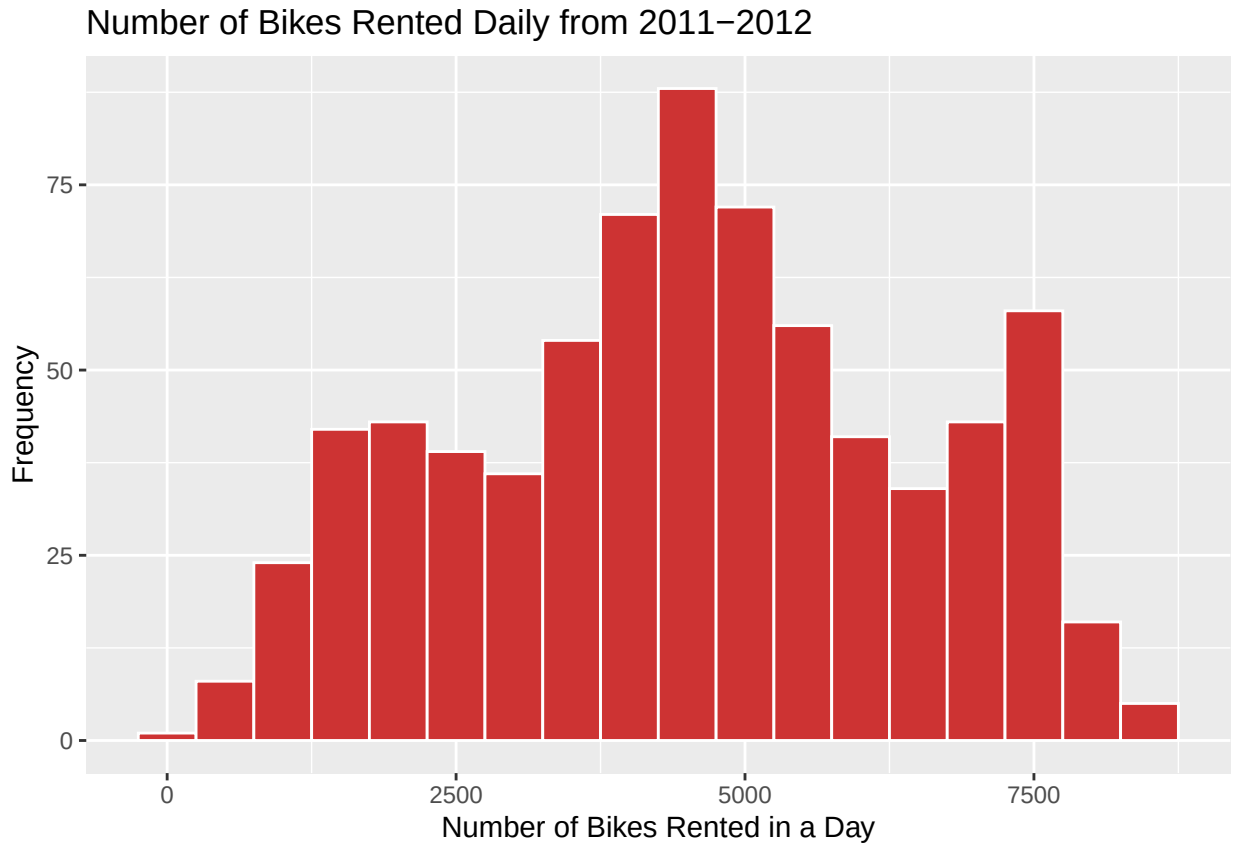
Use R Markdown and ggplot2 to create different visualizations with BikeShare.txt. You will turn in an R

Markdown with each of the required graphs and a brief description of your design choice(s). You'll need to create:

1. A histogram with the number of bike rentals

a. Did you use the default bin width? Why or why not?

```
ggplot(data, aes( x = cnt )) +
  geom_histogram( binwidth = 500 ,
                  fill = "brown3",
                  color = "white" ) +
  labs( title="Number of Bikes Rented Daily from 2011-2012",
        x = "Number of Bikes Rented in a Day",
        y = "Frequency" )
```

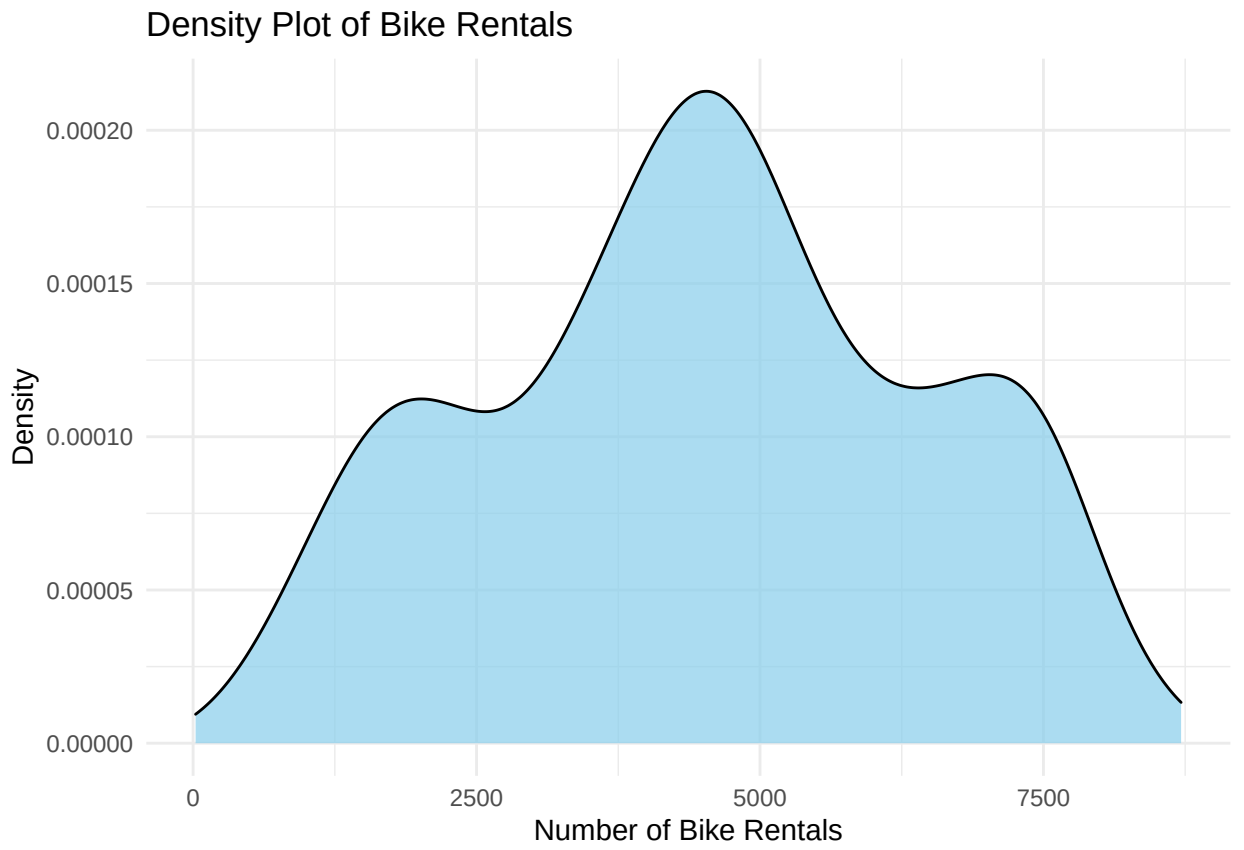


This graph is a simple histogram showcasing the frequency with which different daily rental values of bikes occured during the 2011-2012 time frame. I chose to make the bins' width be 500 bikes because the range of bike rental values is so great that this makes the number of bins to analyze feasible. I also made the lines a reddish color because I associate red with my mental image of a bike and I made the outlines of the bars white so they lined up well and were easy to differentiate between.

2. A density plot with the number of bike rentals

```
ggplot(data, aes( x = cnt )) +
  geom_density( fill = "skyblue",
                color = "black",
                alpha = 0.7) +
  labs( title = "Density Plot of Bike Rentals",
        x = "Number of Bike Rentals",
```

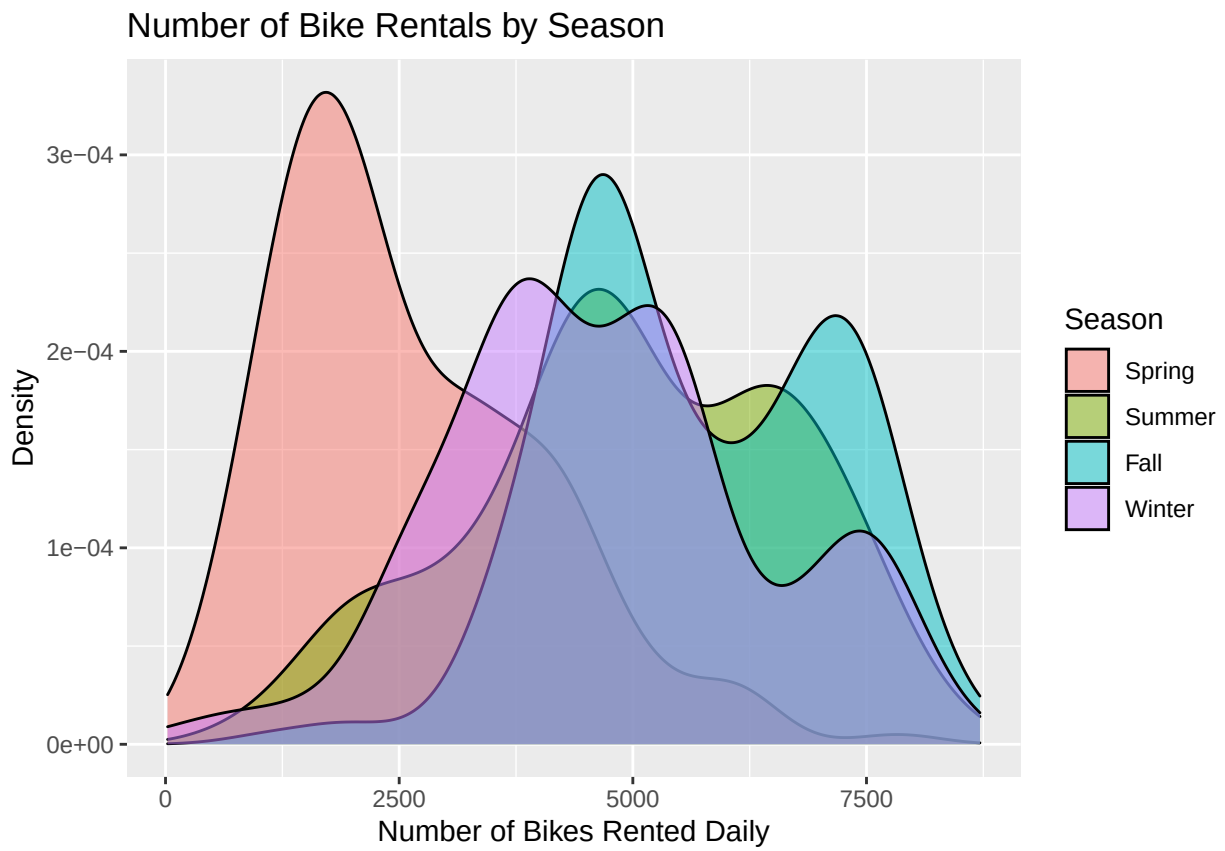
```
      y = "Density" ) +
  theme_minimal()
```



The graph is a density plot that shows the number of bike rentals which showed that it's peak total number of rental is around a 45000. The density plot can smooth out the distribution of values and reduce the noise. It visualizes the distribution of data over a given period, and the peaks show where values are concentrated.

3. Overlapping density plots using the number of bike rentals and a categorical variable of your choice

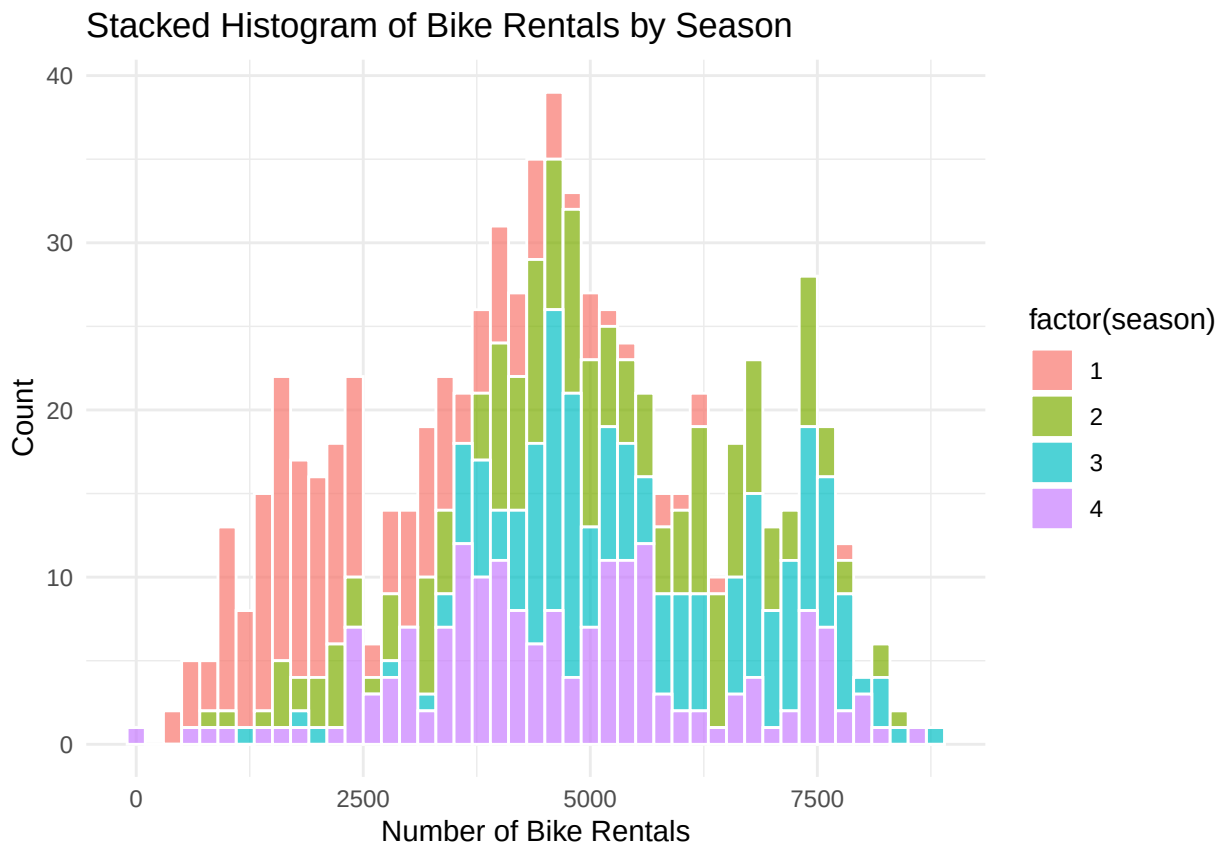
```
ggplot(data, aes( x = cnt , fill = factor (season))) +
  geom_density( alpha = 0.5 ) +
  labs( title = "Number of Bike Rentals by Season",
       x = "Number of Bikes Rented Daily",
       y = "Density",
       fill = "Season" ) +
  scale_fill_discrete ( labels = c( 'Spring' , 'Summer' , 'Fall' , 'Winter' ))
```



This overlayed density plot showcases the number of bike rentals daily and divides the occurrences of each count by season. Thus, we can see the different trends amongst the seasons: people don't tend to rent many bikes in spring, but they do during the fall. I chose to leave the colors the default R colors for this graph because they provide enough contrast to see the different seasons but are still not that jarring against each other.

4. A stacked histogram of Bike Rentals by Season

```
# Count of Bike Rentals for Each Season in terms of Registered or Casual Rentals
ggplot( data, aes( x = cnt, fill = factor (season))) +
  geom_histogram( binwidth = 200, position = "stack", color = "white", alpha = 0.7 ) +
  labs(title = "Stacked Histogram of Bike Rentals by Season",
       x = "Number of Bike Rentals",
       y = "Count" ) + theme_minimal()
```



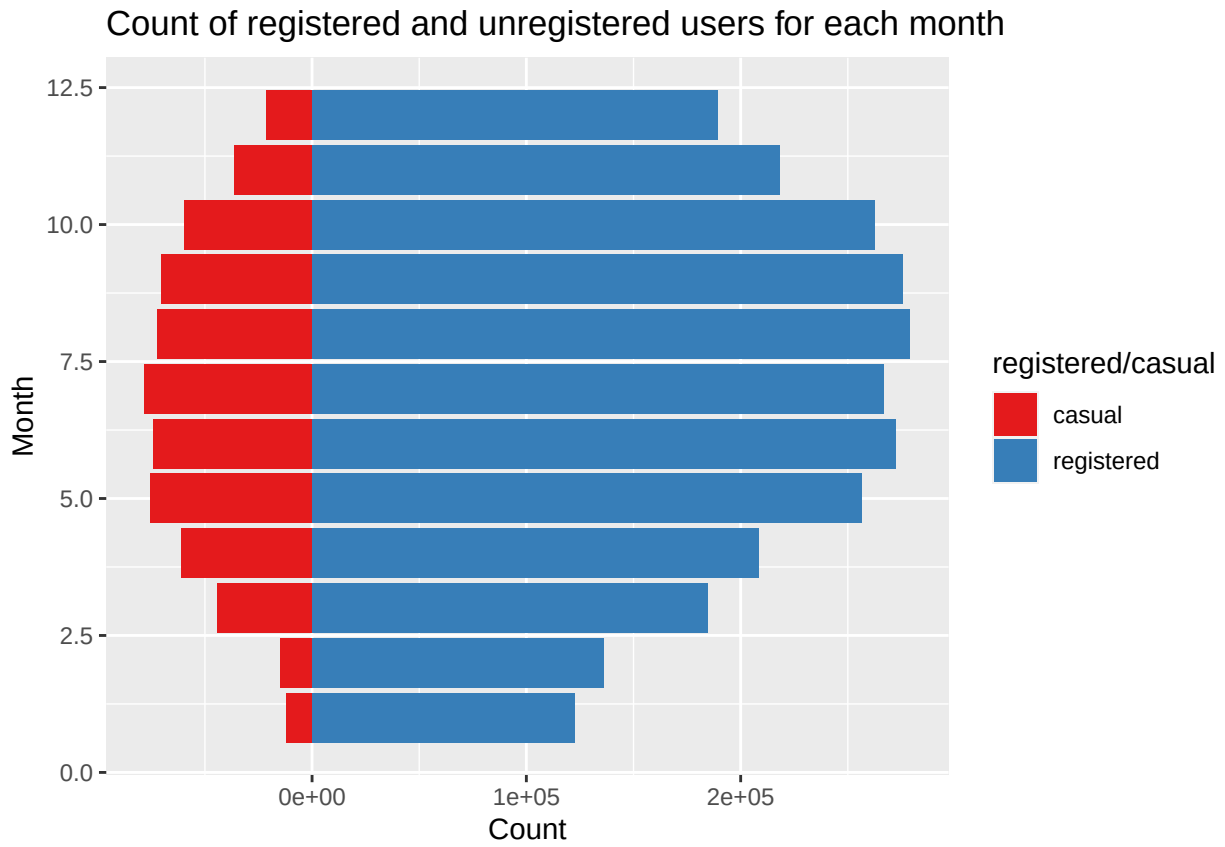
The stack histogram shows the distribution of bike rental categorized by working days(1) and non working days(2). The x axis represents the number of bike rentals whereas the y axis represents the count, while the bars are stacked based on the working day status. This graphs reveal that a high number of rentals happen on working days compared to non working days.

5. An 'age pyramid' with registered and unregistered users

a. Make sure to think carefully about what variables should be visualized here

```
# making registered and unregistered users columns to a longer format
data2 <- data %>% pivot_longer(cols=c('casual', 'registered'),
                               names_to='user_type',
                               values_to='users')

# making age pyramid
data2 %>%mutate(
    users = ifelse(user_type=="casual", users*(-1),
                           users*1))%>%
    ggplot(aes(x = mnth ,y = users , fill= user_type)) +
    geom_bar(stat = "identity") +
    coord_flip()+
  labs(title = "Count of registered and unregistered users for each month",
       x = "Month",
       y = "Count",
       fill = "registered/casual") +
    scale_fill_brewer(type = "qual",palette = 6)
```



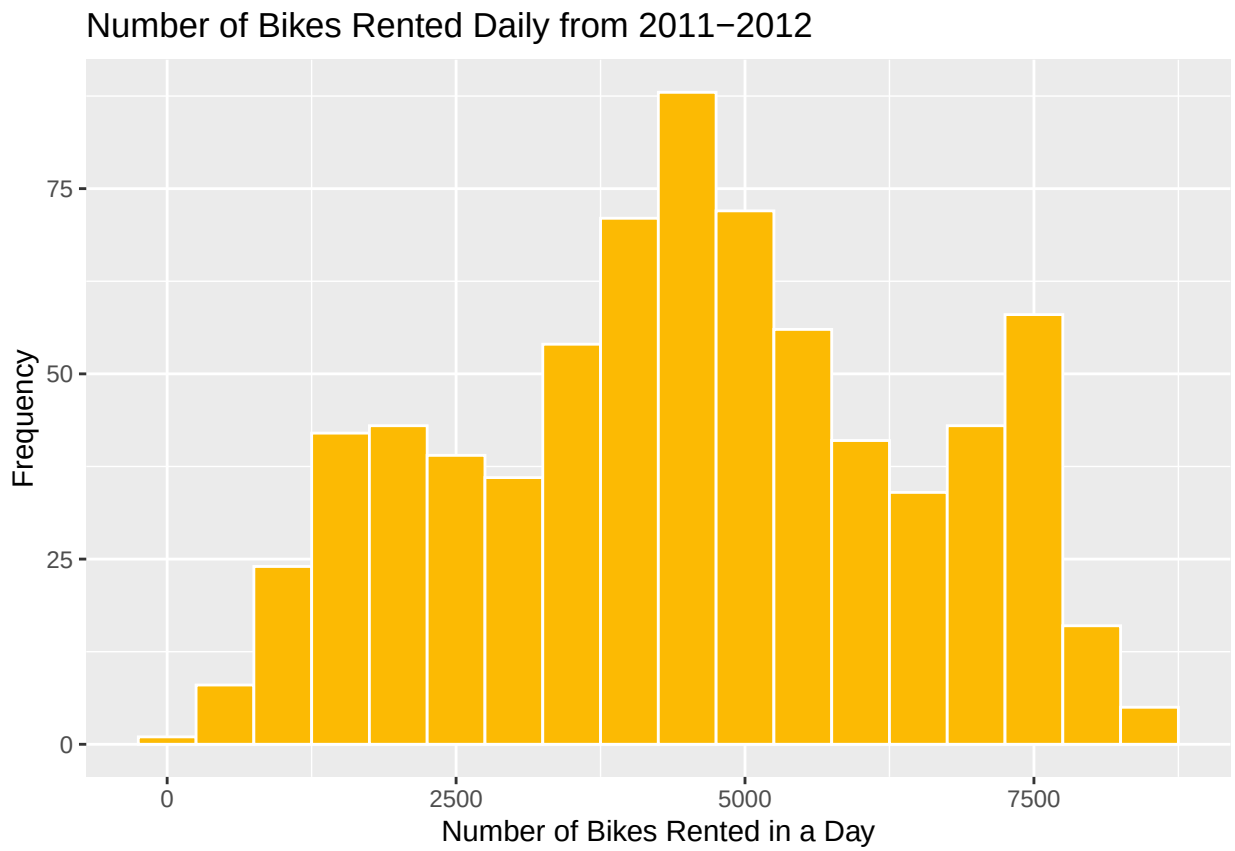
An age pyramid, using a paired bar chart-type graphic, shows the count of registered and unregistered users for each month. This type of graphic provides a very clear picture of a count of registered and unregistered users composition of each month.The shape of the pyramid reflects whether the population, which in this case is "cnt" is growing, stable or declining. We used two contrasting colors to visualize better

Other:

6. Change the colors from default in at least one graph

a. Either do this now on a new graph or tell me which graph had the colors changed

```
ggplot(data, aes( x = cnt )) +
  geom_histogram( binwidth = 500 ,
                  fill = "#fcba03",
                  color = "white" ) +
  labs( title="Number of Bikes Rented Daily from 2011-2012",
        x = "Number of Bikes Rented in a Day",
        y = "Frequency" )
```



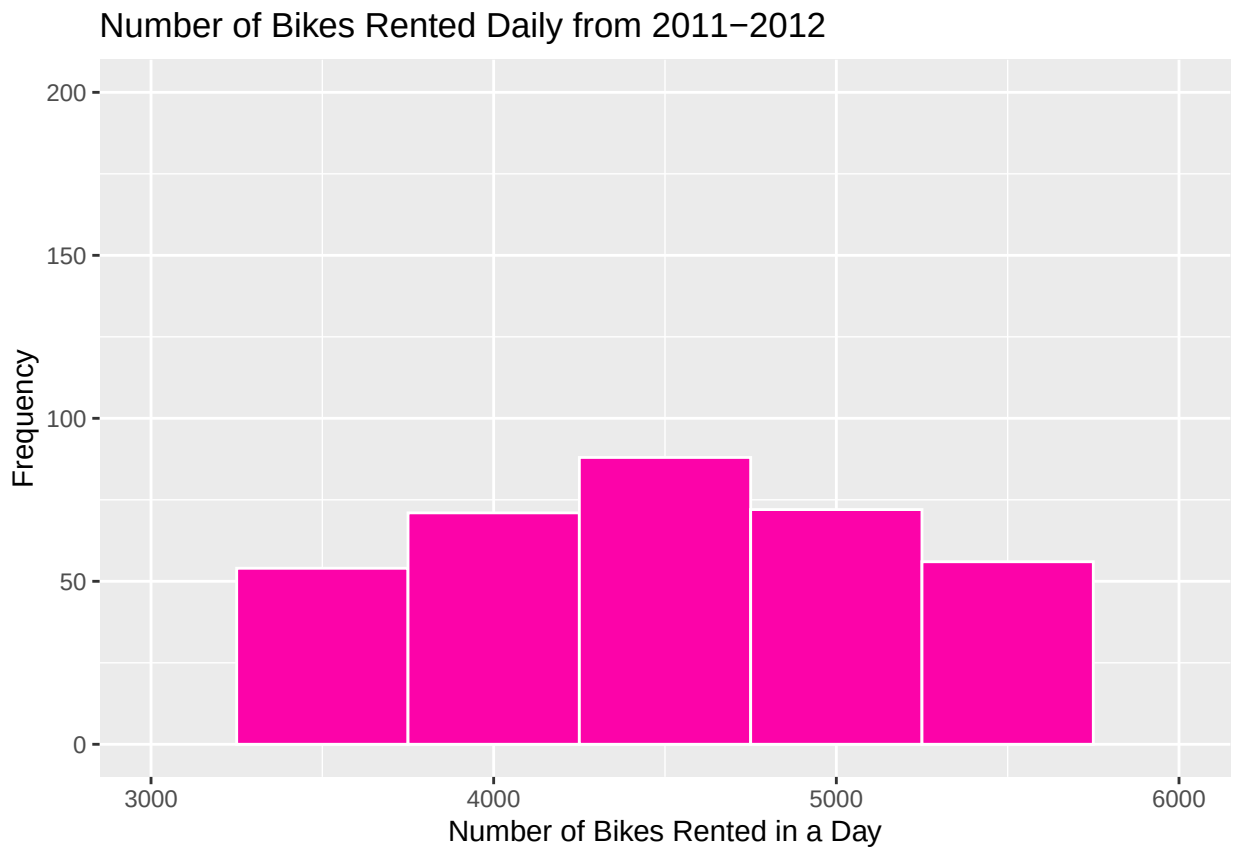
We changed the color of the histogram from question 1. Design choice wise, everything is very much same from the question 1.

7. Create a misleading distribution graph and explain why it is misleading

```
# Changing the scale of the x and y axis
ggplot(data, aes( x = cnt )) +
  geom_histogram( binwidth = 500 ,
                  fill = "#fc03a9",
                  color = "white" ) + xlim(3000, 6000 )  + ylim(0, 200 )  +
  labs( title="Number of Bikes Rented Daily from 2011-2012",
        x = "Number of Bikes Rented in a Day",
        y = "Frequency" )
```

```
## Warning: Removed 352 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



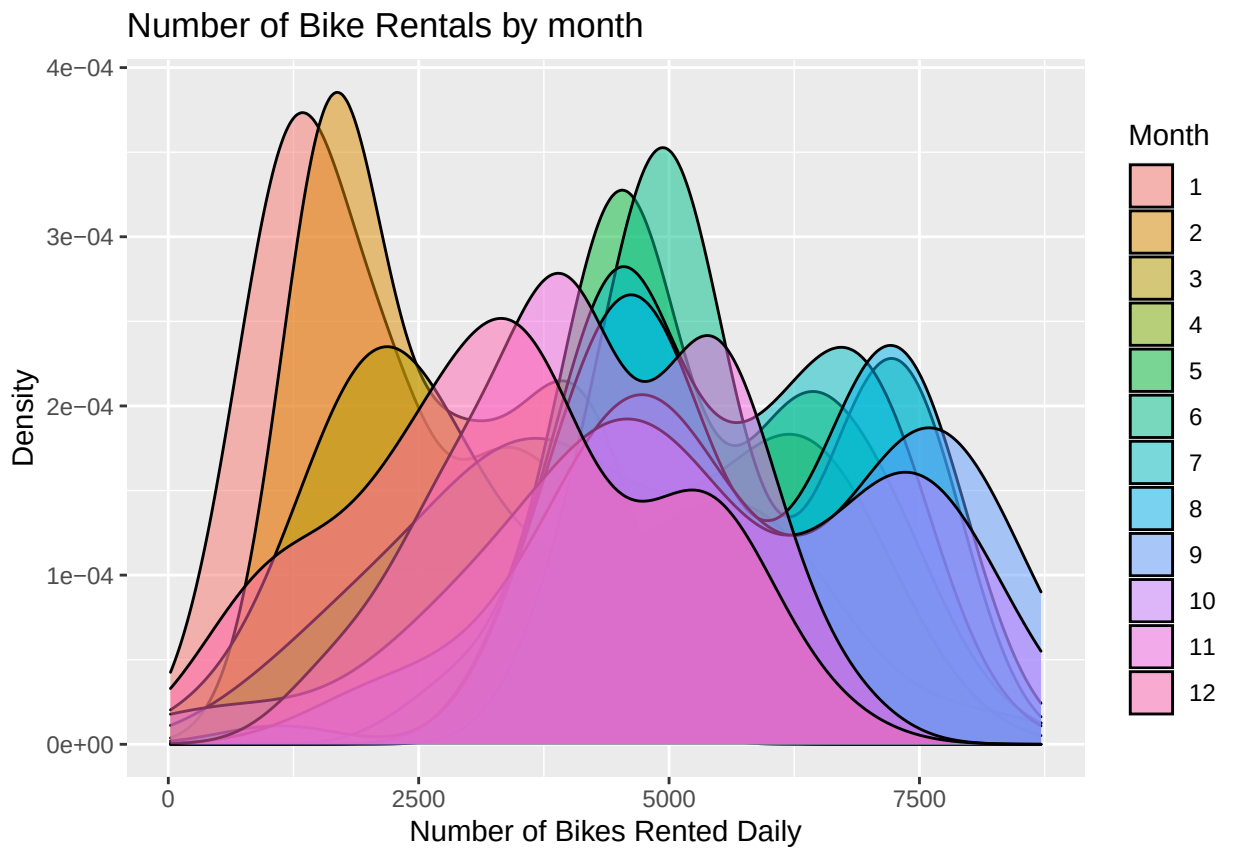
This graph could be misleading because scales of both x (between 3000 and 6000) and y (between 0 and 200) axes are not accurate, which might make the observer feel like output values are smaller in the y axis, from the first glance. We changed the color of the histogram from question 1. Design choice wise, everything is very much same from the question 1.

8. Think of a question about the data that could be addressed with histograms or density plots that has not yet been explored. Create this graph and record the question that this graph addresses.

a. For example, for #5 the question might be 'How does the distribution of unregistered users differ from that of registered users?'

How does the number of bike rentals vary each month?

```
ggplot(data, aes( x = cnt , fill = factor (mnth))) +
  geom_density( alpha = 0.5 ) +
  labs( title = "Number of Bike Rentals by month",
        x = "Number of Bikes Rented Daily",
        y = "Density",
        fill = "Month" ) +
  scale_fill_discrete ( labels = c( '1' , '2' , '3' , '4', '5' , '6' ,
                                    '7', '8' , '9' , '10', '11' , '12' ))
```



We selected "mnth" as a variable as it would show us the density variation of bike rentals for each month between all users and it would be a helpful plot to interpret Also we would have a different color representing each month.