Introduction to Plotting

R Team

Data Visualization

- Widely used concept in business
- Communication between analysts and stakeholders
- We will be using ggplot2 for graphing, other options are possible in R
- For further readings: http://www.cookbook-r.com/Graphs/
- We will cover different type of charts:
    1. Scatter Plot
    2. Line Charts
    3. Box plots
    4. Histograms

Note: Use View(data_frame) to choose which columns to use as x and y.

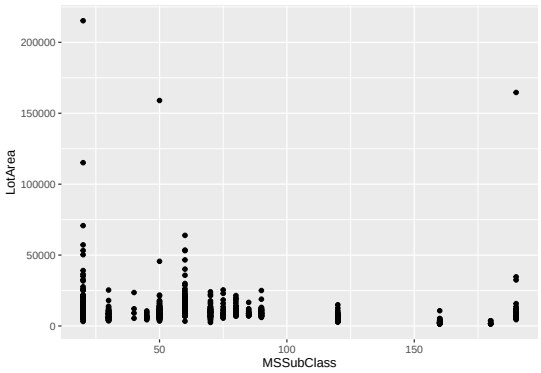
```
# Importing libraries
library(palmerpenguins)
library(ggplot2)
library(tidyverse)

# Importing datasets
spotify_songs <- read.csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
movie_profit <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/201
housing_df <- read_csv("https://www.dropbox.com/s/tvvtf9dwjufo7os/housing_train.csv?dl=1")
penguins <- palmerpenguins::penguins
```

1. Scatter Plot

ggplot(data_frame, aes(x,y)) + geom_point():

- First argument is the dataframe we are graphing
- Second argument identifies the x and y coordinates using the aes function
- Then different type of properties are adding by adding to the initial definition

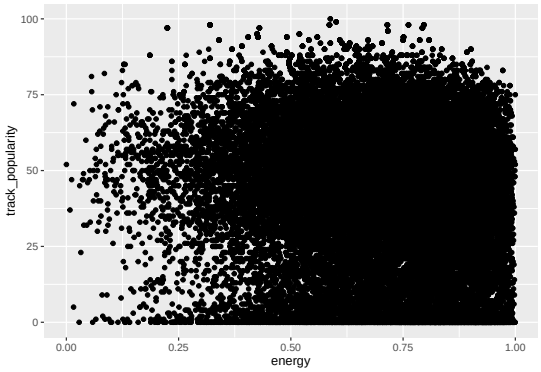
```
ggplot(housing_df, aes(x=MSSubClass, y=LotArea)) + geom_point()
```



Can you see a relation between energy and track_popularity?

- Plots can be helpful in identifying certain correlations.
- One could visualize data and extract insights just by looking at the graphs.

```
ggplot(spotify_songs, aes(x=energy, y=track_popularity)) + geom_point()
```



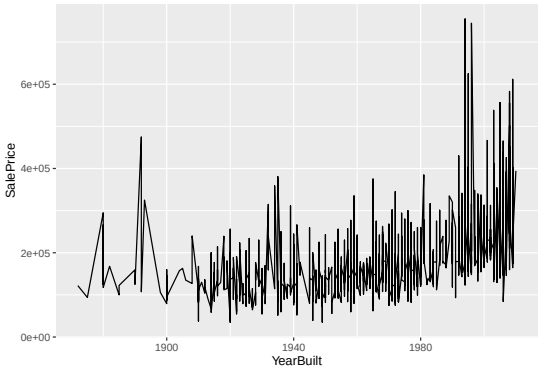
Exercise:

1. Scatter plot of domestic_gross (x) vs worldwide_gross (y)
2. Scatter plot of MasVnrArea (x) vs LotArea (y)

2. Line Chart

- When the lines of the scatter plot are connected

```
ggplot(housing_df, aes(x=YearBuilt, y=SalePrice)) + geom_line()
```



- The geom_line() is the part which identifies the chart as a line chart.

- The rest of the function is the same, we specify the dataframe and x,y columns.
- Plotting the data YearBuilt as x and SalePrice as y.

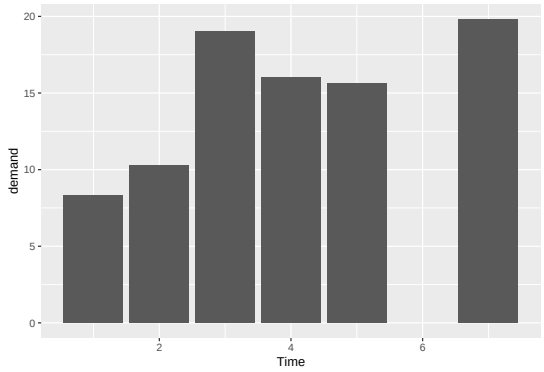
Exercise:

1. Scatter plot of domestic_gross (x) vs worldwide_gross (y)
2. Scatter plot of MasVnrArea (x) vs LotArea (y)

3. Bar Chart

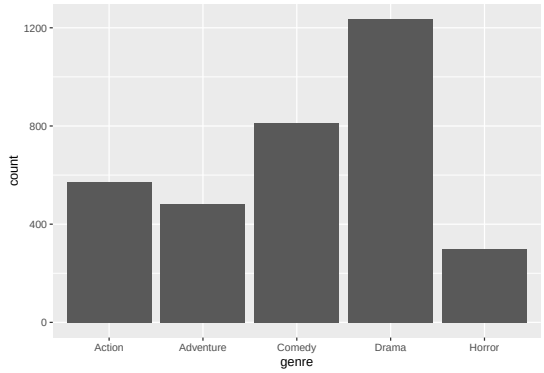
- It has the same structure but instead of geom_point we use geom_bar() with its identity parameter set to identity

```
ggplot(BOD, aes(x=Time, y=demand)) + geom_bar(stat='identity')
```



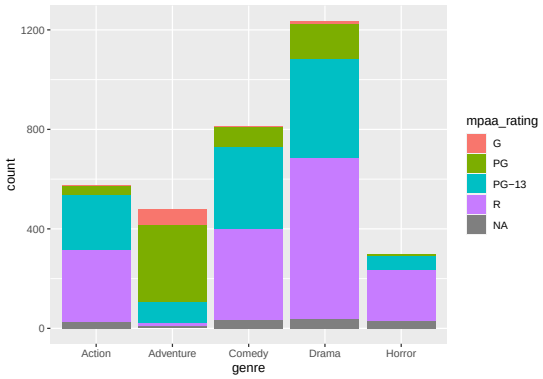
- **Counting repetition**: We are interested to know which movie genre has the most movies (which genre has the highest repetition):
  1. First we pass movie_profit as the dataframe
  2. We choose genre as our x since we are interested in knowing the repetition
  3. We add geom_bar() to identify the graph as a bar chart.

```
ggplot(movie_profit, aes(x = genre )) + geom_bar()
```



- Imagine that we are interested in knowin what are the repetitions of the rating types within each genre. We could do this using an attribute called fill:
  1. We have followed the same process for steps 1 to 3.
  2. ***What does fill do?*** It groups data based on rating. We have already grouped data into genres, fill will group each genre group based on its ratings.

```
ggplot(movie_profit, aes(x = genre, fill = mpaa_rating )) + geom_bar()
```



Note: The position value with geom_bar() should either be dodge or fill

Exercise: (ET: 3/5)

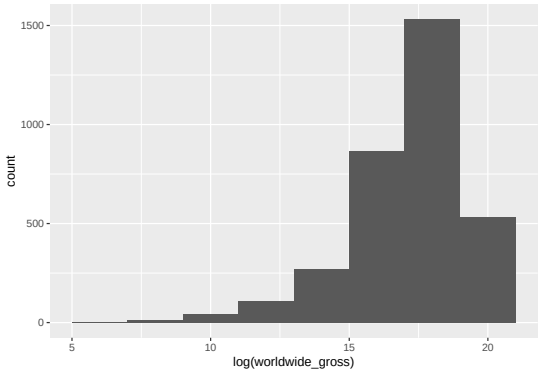
1. Plot clarity column of diamonds dataframe with cut as the fill.
2. Plot clarity column of diamonds dataframe with color as the fill.

4. Histograms

- Bar charts with connected bars

```
ggplot(movie_profit, aes(x=log(worldwide_gross) )) + geom_histogram(binwidth = 2)
```

```
## Warning: Removed 36 rows containing non-finite values (stat_bin).
```



1. dataframe: we specify which dataframe we are interested in
2. aes(x=log(worldwide_gross)): mapping the needed x and y values
3. - geom_histogram(): Generates a histogram