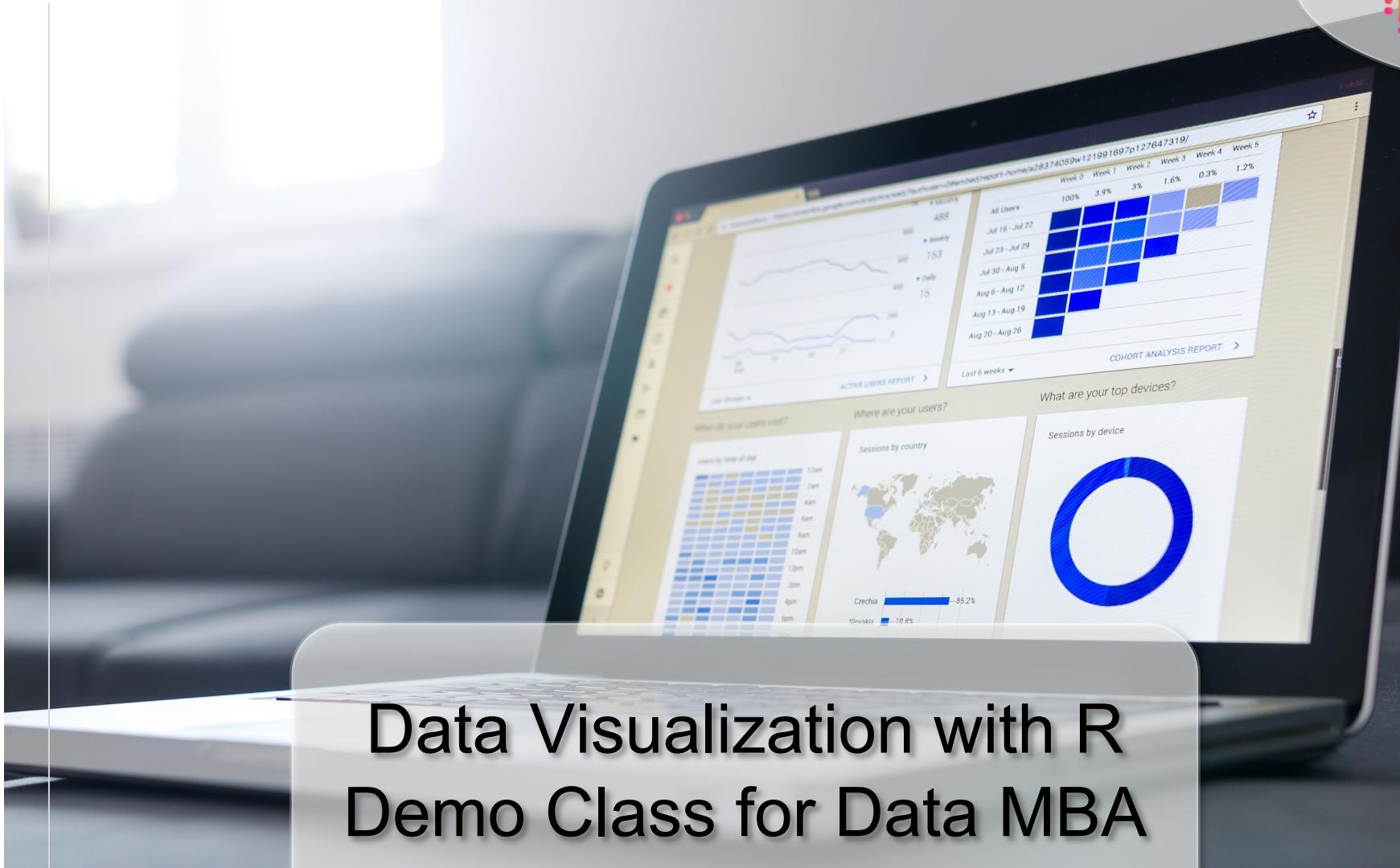


Special Crafted by
Bernardus Ari
Kuncoro

Training Series No.
120.040.12.111.27



Trainer Bio





Bernardus Ari Kuncoro (Ari),
Head of Analytics COE at IYKRA.

In recent 5 years, Ari had experienced as a Data scientist in consultancy, ecommerce, and telecommunication companies. His background is Electrical Engineering (Telecommunication) and Computer Science. He is absolutely and utterly passionate about Data Science and teaching, thus he is looking forward to sharing his passion and knowledge with you! Please connect to one of the online channels below (blog, Instagram or LinkedIn).



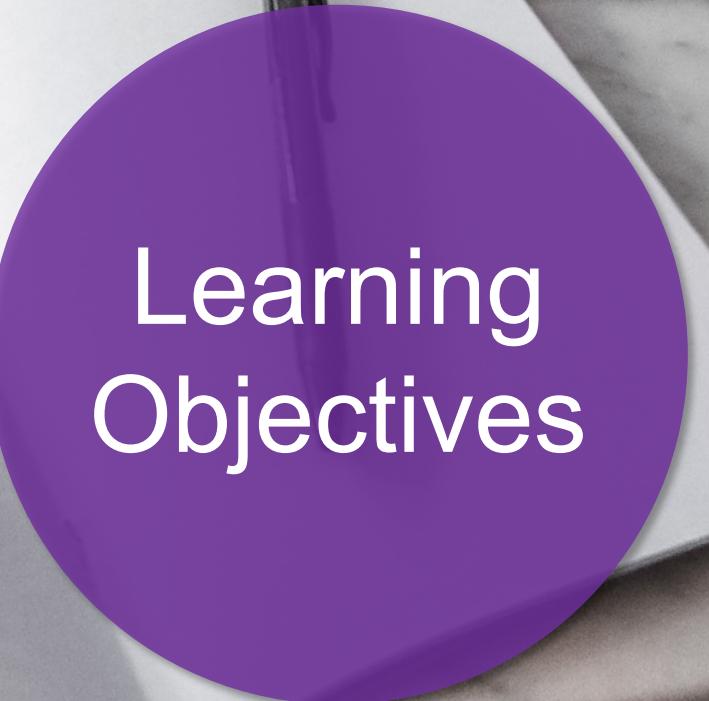
<http://blog.arikuncoro.xyz>



[@arikunc0r0](#)



[Bernardus Ari Kuncoro](#)



Learning Objectives

Learning Objectives

- To understand the objectives of visualization.
- To understand what suitable graphs for different purposes.
- To understand and hands on in creating basic graphs with R package: ggplot2

Today's Agenda

1. Pre-requisites
2. Exploratory vs Explanatory
3. Basic Visualization Types
4. Grammar of graphics
5. Seven most frequent used graphs
6. Summary

The background image shows a panoramic view of a dense urban area, likely Jakarta, Indonesia. In the foreground, there are numerous smaller buildings with red-tiled roofs, interspersed with green trees. Behind them, a massive cluster of modern skyscrapers rises against a clear blue sky with a few wispy clouds. Notable landmarks include the Citi Tower and the FWD Tower. A large white circle is centered over the middle ground, containing the text 'Pre-requisites'. A small purple square at the top left corner of the circle contains the number '1'.

1

Pre-requisites

Please prepare the followings:

- **Install R** from CRAN on your laptop (<https://cran.r-project.org>)
- **Install RStudio Desktop IDE*** on your laptop
(<https://www.rstudio.com/products/rstudio/download>)
- Clone this github: <https://github.com/arikunco/visualization.git>
 - Install packages in R:
`install.packages("tidyverse")`
`install.packages("corrplot")`
`install.packages("corrgram")`

*IDE = Integrated Development Environment

John Tukey

American mathematician



John Wilder Tukey was an American mathematician best known for development of the Fast Fourier Transform algorithm and box plot. The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name. He is also credited with coining the term 'bit'. [Wikipedia](#)

“The **simple graph** has brought **more** information to the data analyst’s mind than any other device.” —
John Tukey

What is Data Visualization?

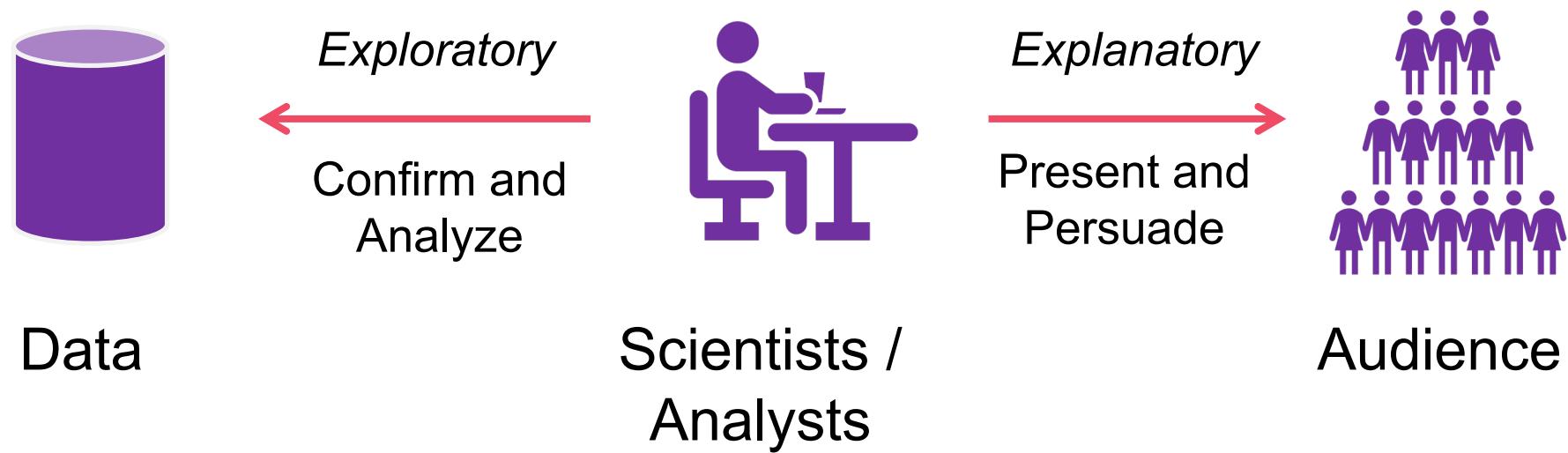
Data visualization refers to the techniques used to **communicate data** or **information** by encoding it as visual objects (e.g., points, lines or bars) contained in graphics.

The goal is to **communicate** information clearly and efficiently to users

The background image shows a panoramic view of a dense urban area, likely Jakarta, Indonesia. In the foreground, there are numerous smaller buildings with red-tiled roofs and green trees. Behind them, a massive cluster of modern skyscrapers rises against a clear blue sky with a few wispy clouds. Some of the buildings have prominent logos, such as 'citi' and 'FWD'.

2

Exploratory vs Explanatory



Exploratory vs Explanatory

If the users are data analysts/data scientists, the objective of visualization is more on **exploratory**.

Exploratory has more freedom and usually only few of the visuals are used to be documented.

If the users are both data analysts, data scientists and general audience, the objective of visualization is more on **explanatory**.

The audience will be explained by your visuals, so they will be easy to understand.

The background image shows a panoramic view of a dense urban area, likely Jakarta, Indonesia. In the foreground, there are numerous smaller buildings with red-tiled roofs and green trees. Behind them, a massive cluster of modern skyscrapers rises against a clear blue sky with a few wispy clouds. The buildings vary in height and design, with some featuring glass facades and others more traditional architectural styles.

3

Basic Visualization Types

Basic Visualization Types

Visualization objectives can be generally derived into 4 things:

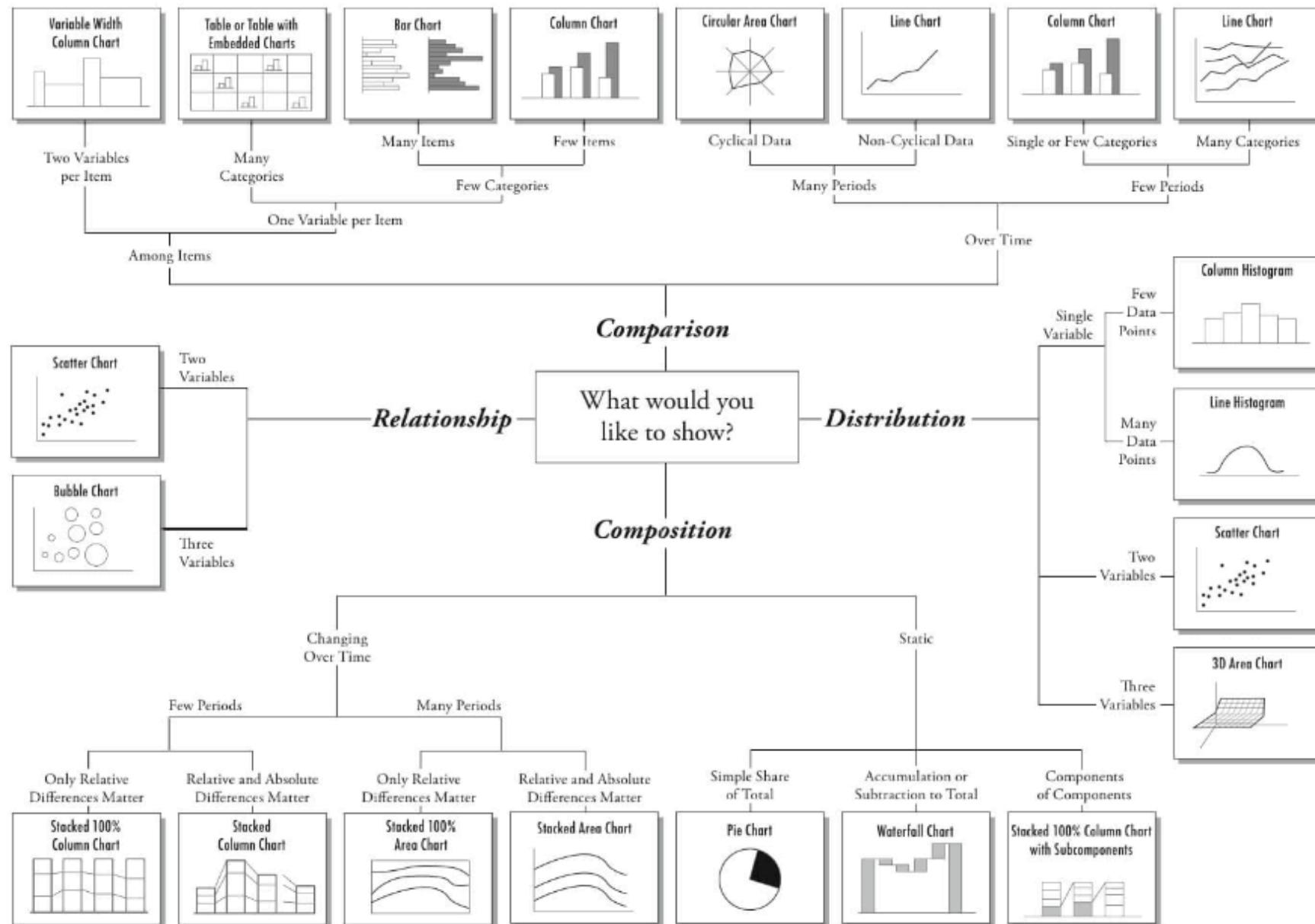
- Comparison
- Composition
- Distribution
- Relationship

3 Minimum Questions to guide you ...

To determine which amongst these is best suited for your data, I suggest you should answer a few questions like:

- How many variables do you want to show in a single chart?
- How many data points will you display for each variable?
- Will you display values over a period of time, or among items or groups?

Chart Suggestions—A Thought-Starter



You can also get inspiration also from

<http://labs.juiceanalytics.com/chartchooser/index.html>

The background image shows a panoramic view of a city skyline, likely Jakarta, featuring numerous modern skyscrapers and office buildings. In the foreground, there is a dense area of lower-rise residential buildings with red-tiled roofs and lush green trees. The sky is clear and blue.

4

Grammar of graphics

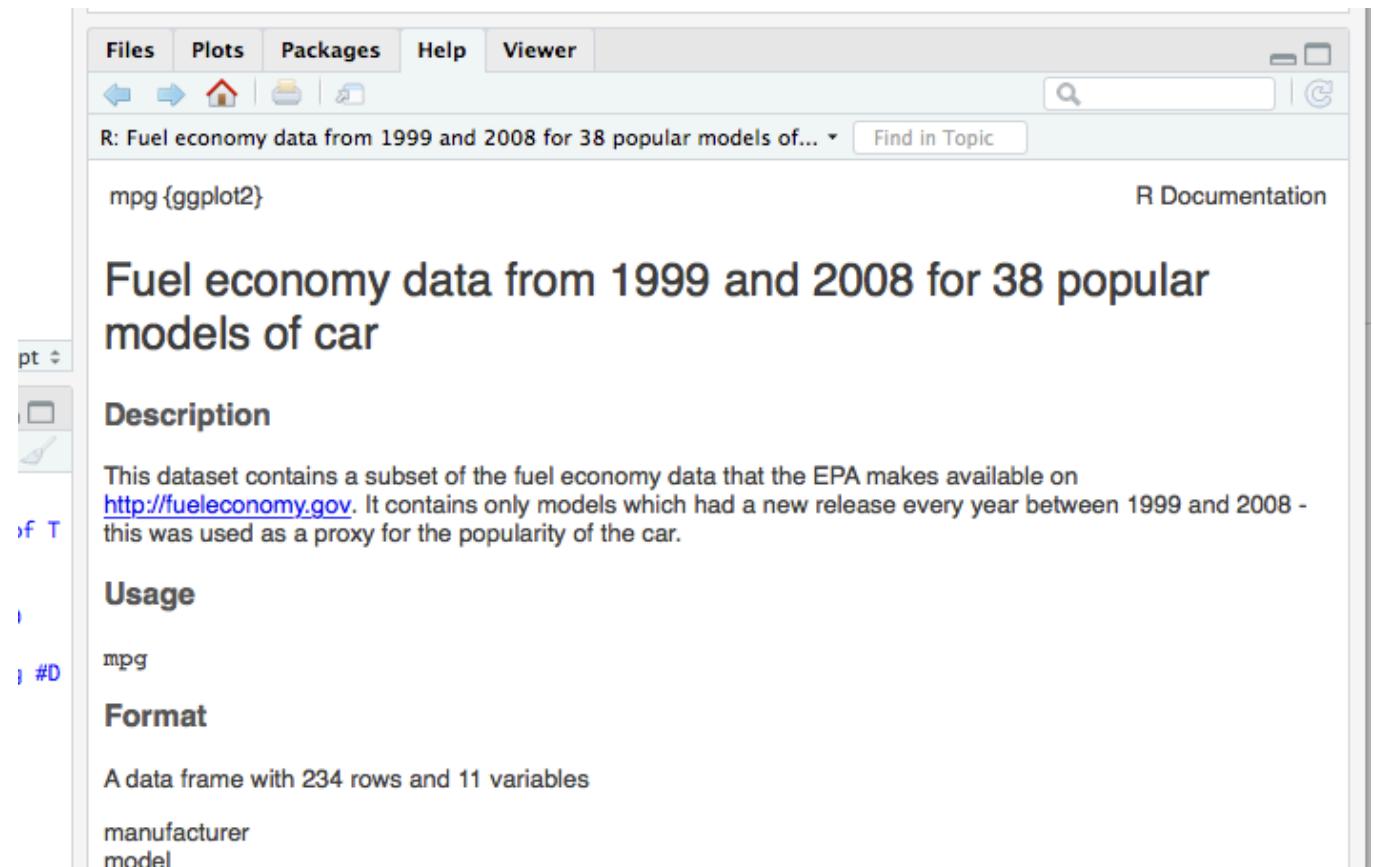
Layer of ggplot (1) - Data

Data

Let's handshake with the data

```
# Perform the  
# following command
```

```
?mpg
```



The screenshot shows the R Documentation interface for the `mpg` dataset. The top navigation bar includes **Files**, **Plots**, **Packages**, **Help**, and **Viewer**. Below the navigation bar, there are icons for back, forward, search, and help. The main title is "R: Fuel economy data from 1999 and 2008 for 38 popular models of car". The title is followed by "mpg {ggplot2}" and "R Documentation". The content area starts with a **Description** section: "This dataset contains a subset of the fuel economy data that the EPA makes available on <http://fueleconomy.gov>. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car." Below that is a **Usage** section with the command `mpg`. The **Format** section states: "A data frame with 234 rows and 11 variables". The variables listed are `manufacturer` and `model`.

Let's handshake with the data

```
# Perform the following command
```

```
mpg
```

```
# A tibble: 234 x 11
  manufacturer model   displ  year   cyl trans   drv   cty   hwy fl class
  <chr>        <chr>   <dbl> <int> <int> <chr>   <chr> <int> <int> <chr> <chr>
1 audi         a4      1.8  1999     4 auto(l5) f       18    29 p    compa~
2 audi         a4      1.8  1999     4 manual(~ f      21    29 p    compa~
3 audi         a4      2.0  2008     4 manual(~ f      20    31 p    compa~
4 audi         a4      2.0  2008     4 auto(av) f      21    30 p    compa~
5 audi         a4      2.8  1999     6 auto(l5) f      16    26 p    compa~
6 audi         a4      2.8  1999     6 manual(~ f      18    26 p    compa~
7 audi         a4      3.1  2008     6 auto(av) f      18    27 p    compa~
8 audi         a4 quat~  1.8  1999     4 manual(~ 4    18    26 p    compa~
9 audi         a4 quat~  1.8  1999     4 auto(l5) 4    16    25 p    compa~
10 audi        a4 quat~  2.0  2008     4 manual(~ 4   20    28 p    compa~
# ... with 224 more rows
:  
```

Layer of ggplot (2) - Aesthetic

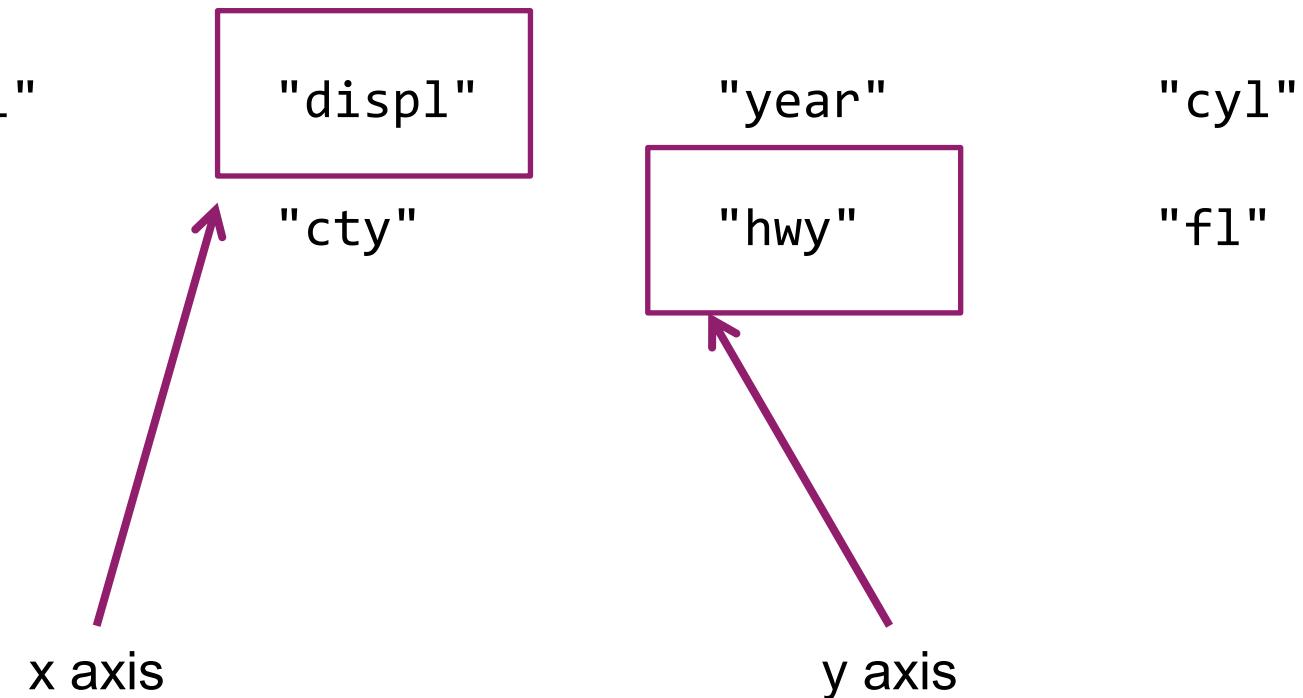
Aesthetic

Data

Layer of ggplot (2) - Aesthetic

```
> colnames(data_mpg)
```

```
[1] "manufacturer" "model"  
[6] "trans"          "drv"  
[11] "class"
```



Layer of ggplot (3) - Geometries

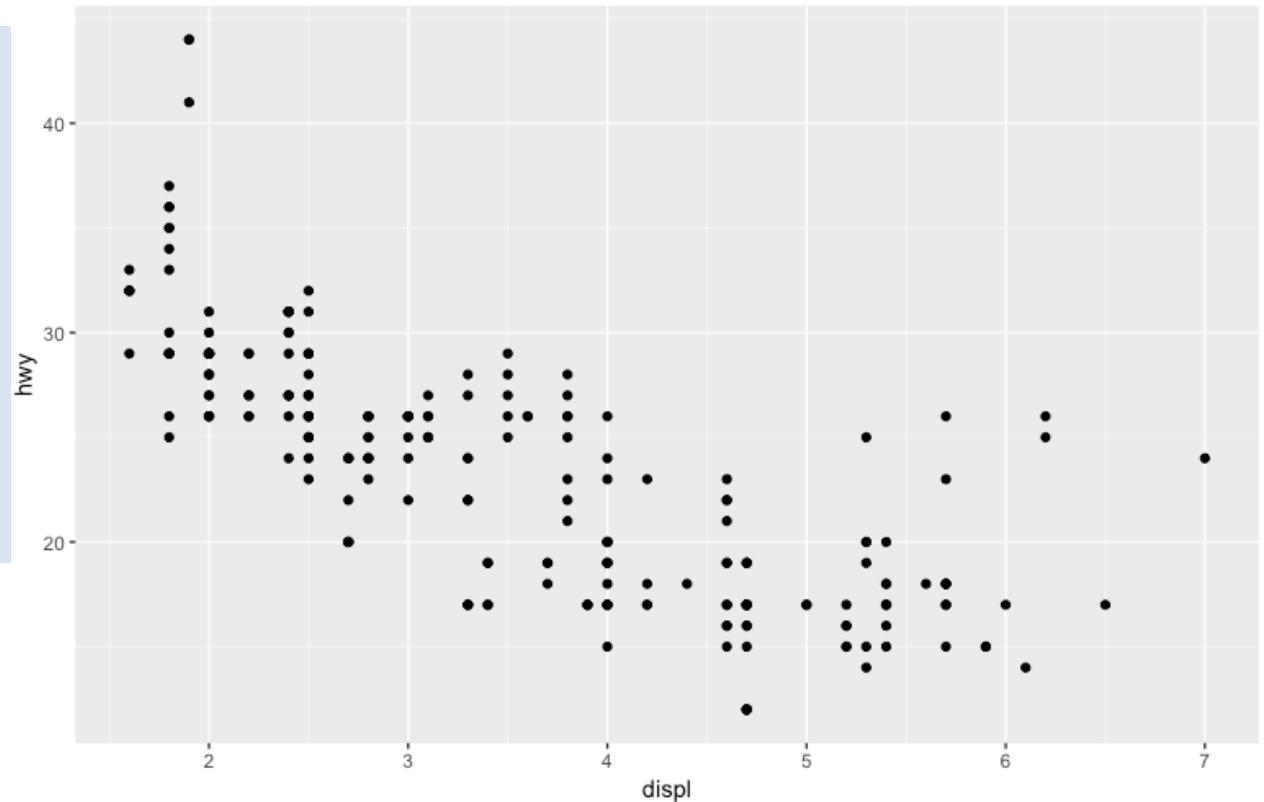
Geometries

Aesthetic

Data

Create basic graph with ggplot

```
# TEMPLATE:  
# ggplot(data = <DATA>) +  
#   <GEOM_FUNCTION>(  
#     mapping = aes(<MAPPINGS>))  
  
ggplot(data = mpg) +  
  geom_point(mapping = aes(x =  
    displ, y = hwy))
```

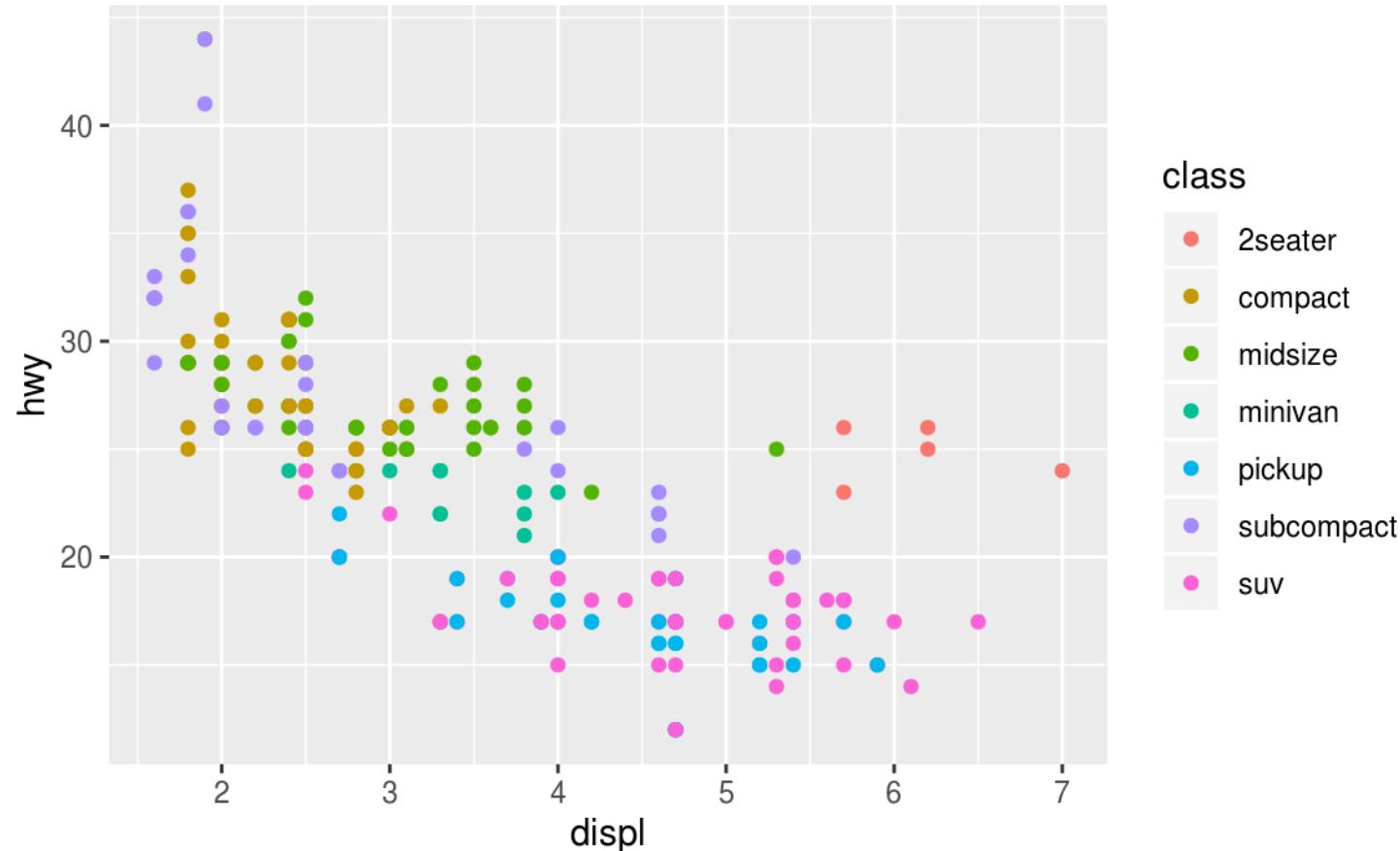


Exercise #1

1. Run `ggplot(data = mpg)`. What do you see?
2. How many rows are in `mpg`? How many columns?
3. What does the `drv` variable describe? Read the help for `?mpg` to find out.
4. Make a scatterplot of `hwy` vs `cyl`.
5. What happens if you make a scatterplot of `class` vs `drv`? Why is the plot not useful?

Aesthetic Mapping

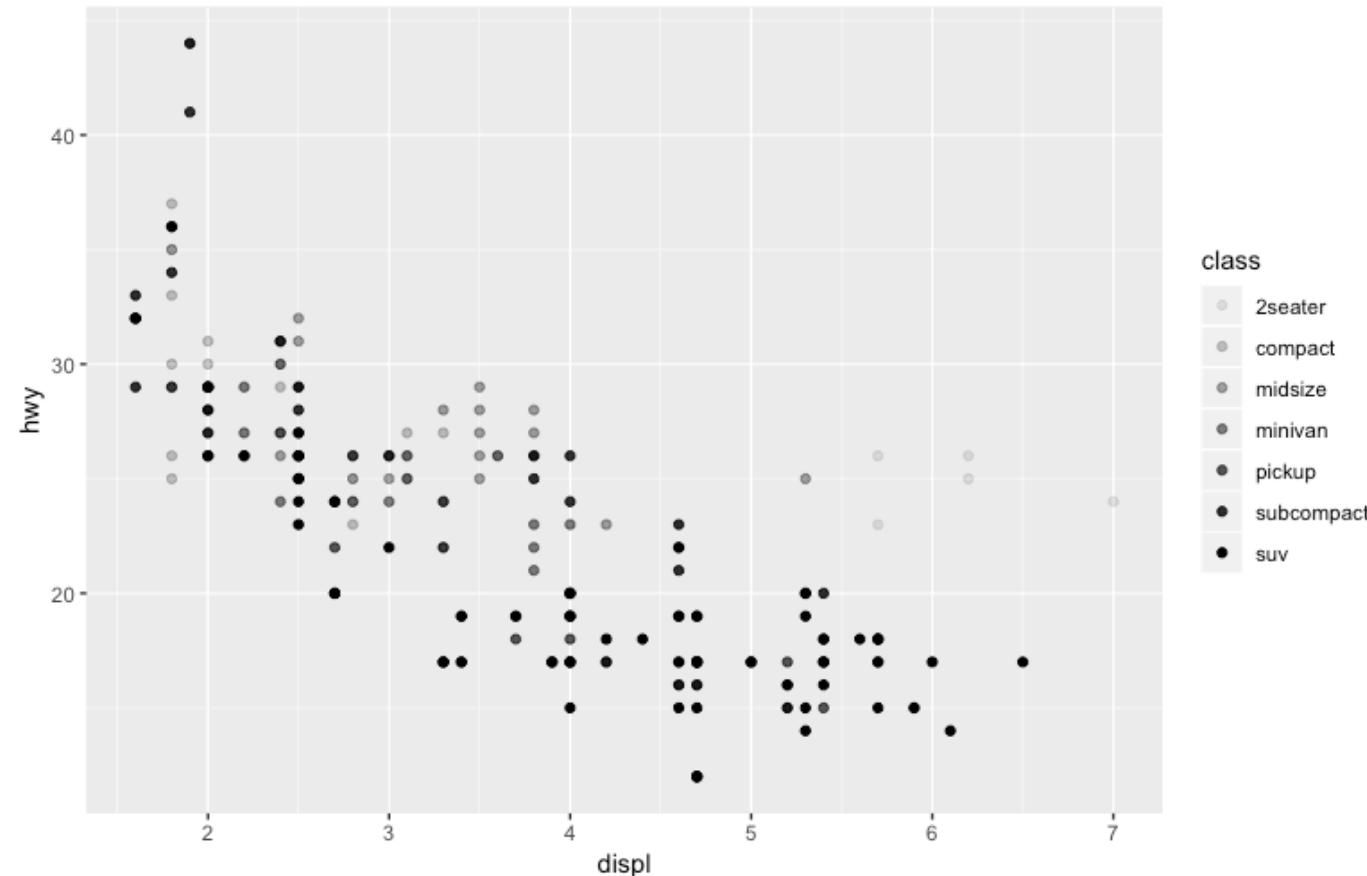
```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = class))
```



You can change with
other variables:
shape,
size,
alpha

Aesthetic Mapping

```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, alpha = class))
```

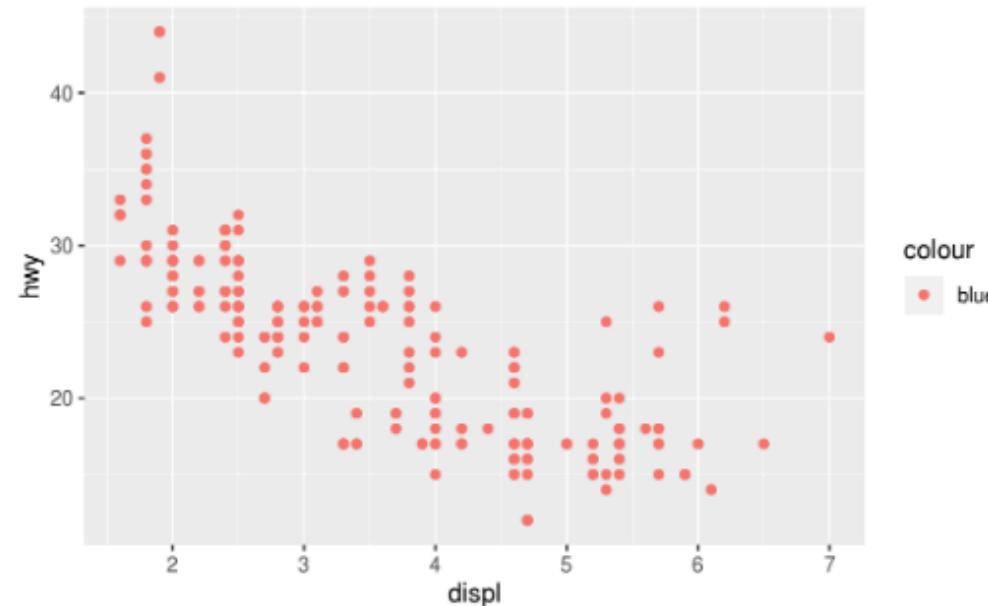


You can change with
other variables:
shape,
size,
alpha

Exercise #2

1. What's gone wrong with this code? Why are the points not blue?

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```



Layer of ggplot (4) - Facets

Facets

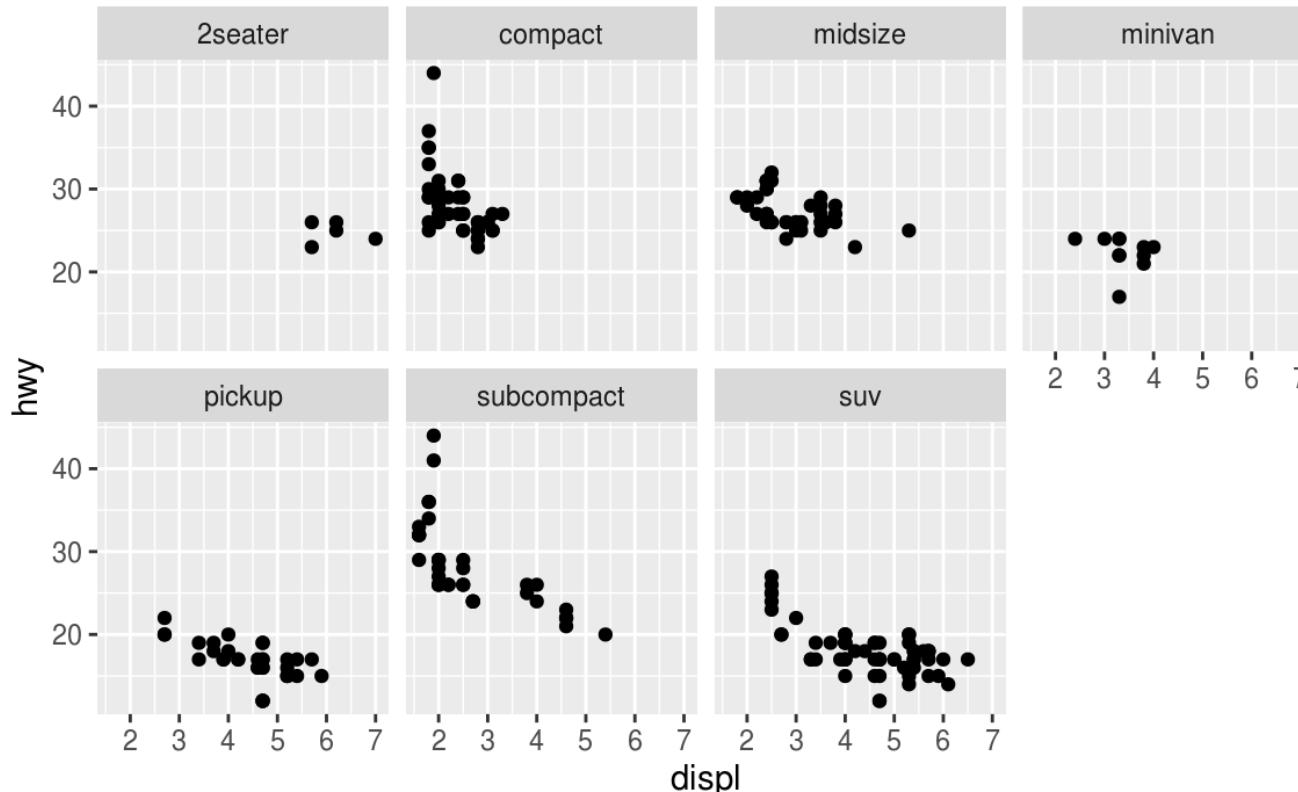
Geometries

Aesthetic

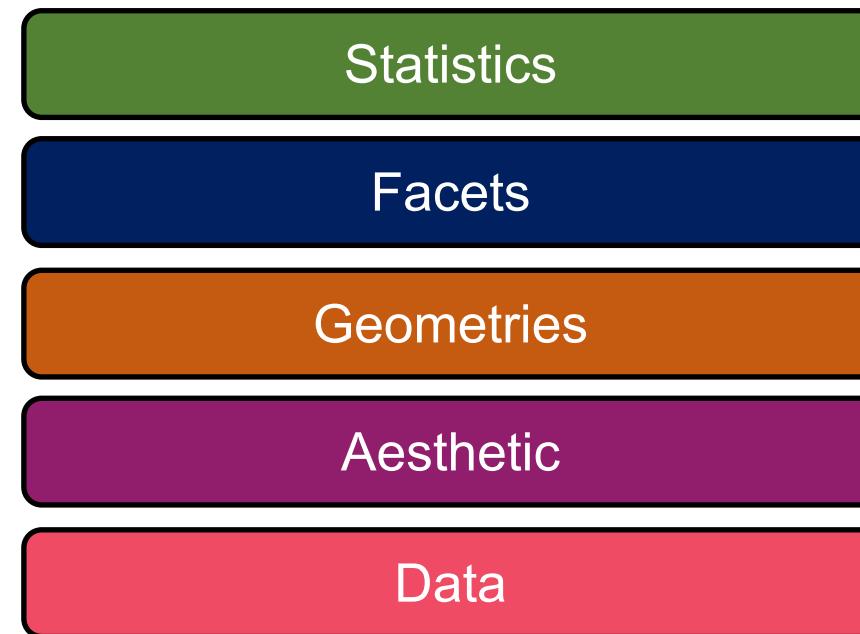
Data

Facets

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) +  
  facet_wrap(~ class, nrow = 2)
```

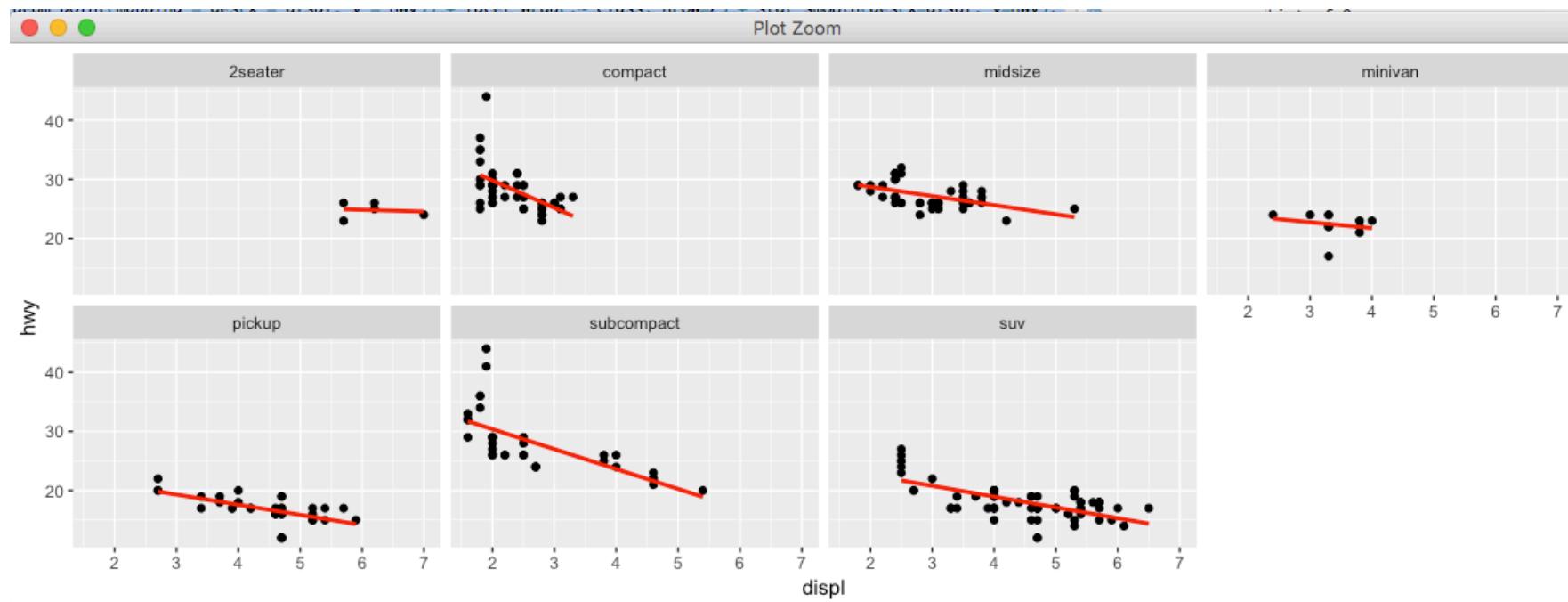


Layer of ggplot (5) - Statistics

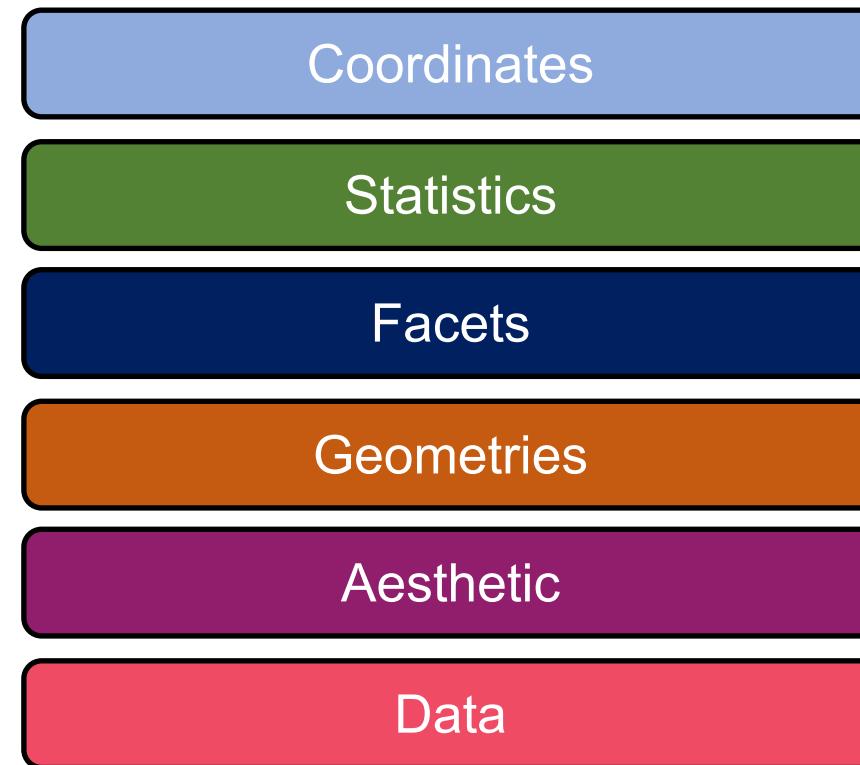


Statistical Transformation

```
ggplot(data = data_mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) +  
  facet_wrap(.~ class, nrow=2) +  
  stat_smooth(aes(x=displ, y=hwy), method = 'lm', se = F, col="red")
```

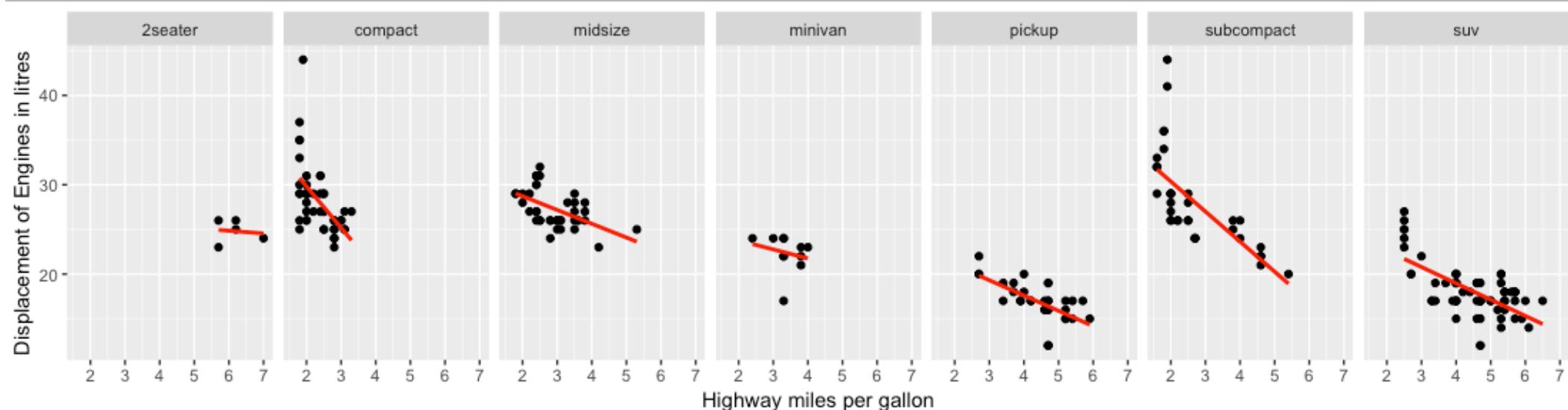


Layer of ggplot (6) - Coordinates



Coordinates

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(.~ class, nrow = 1) +
  stat_smooth(aes(x=displ, y=hwy), method = 'lm', se = F, col="red") +
  scale_y_continuous("Displacement of Engines in litres") +
  scale_x_continuous("Highway miles per gallon") + coord_cartesian()
```

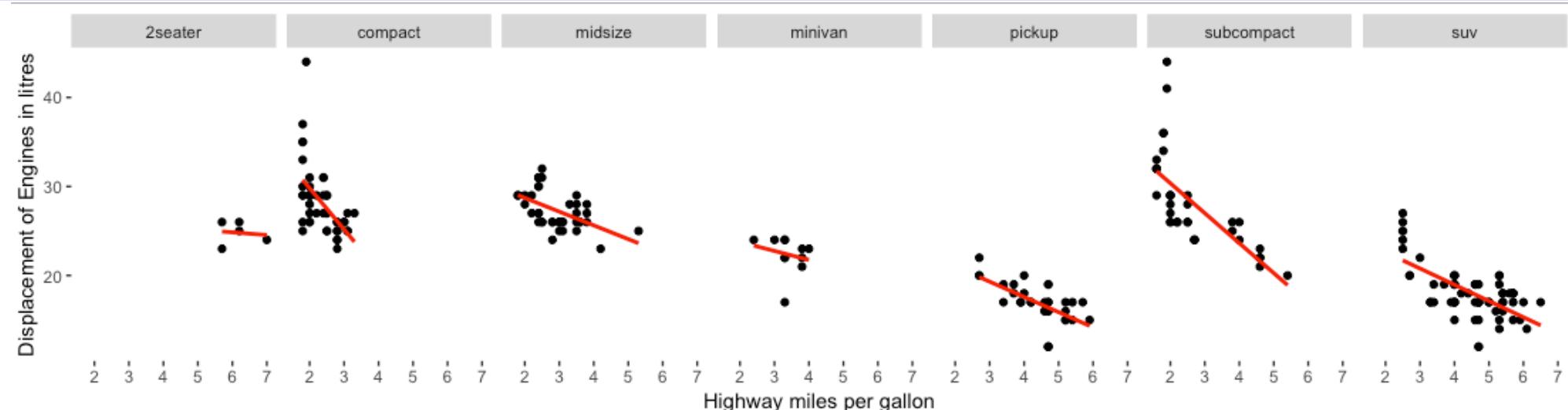


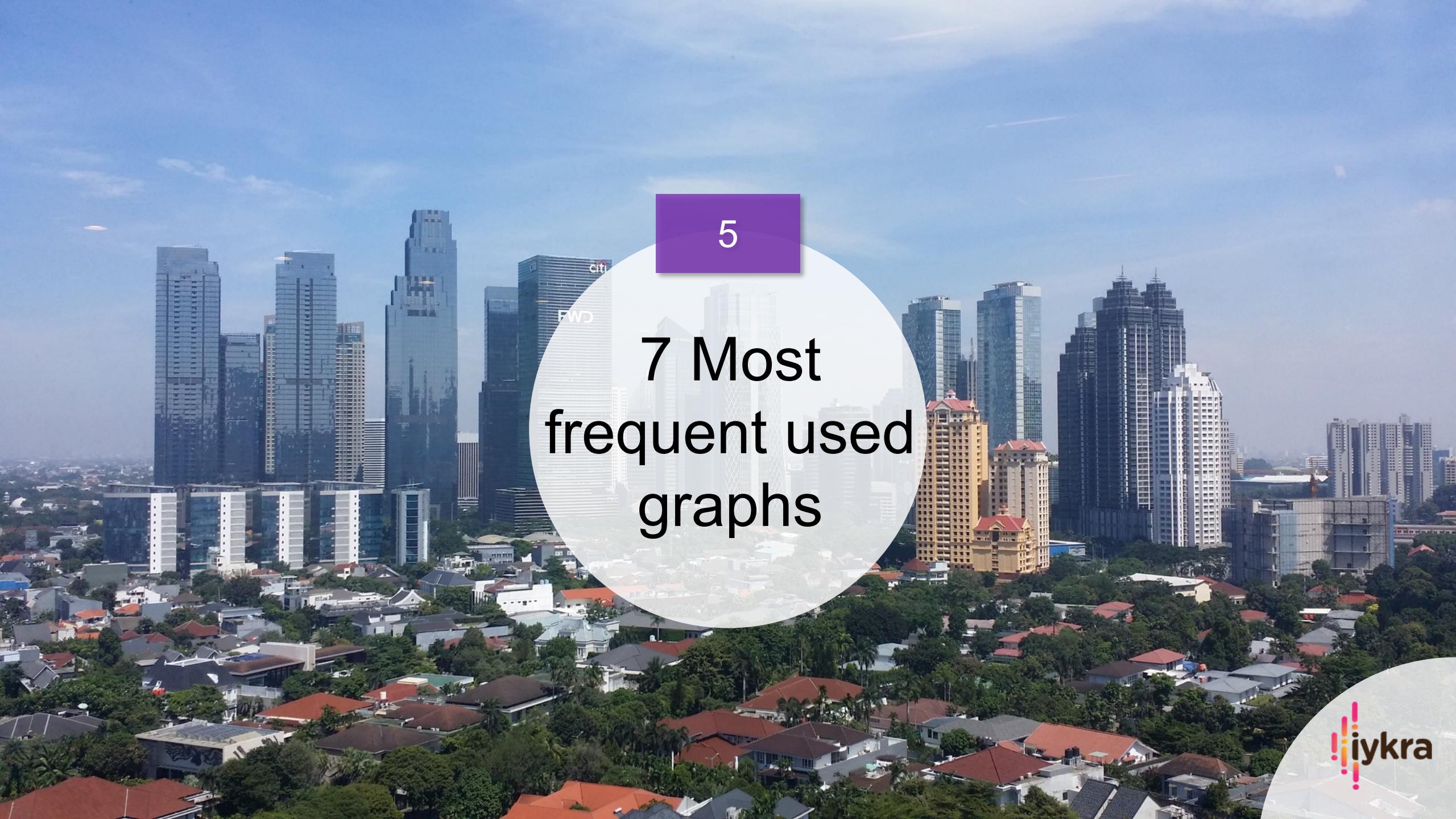
Layer of ggplot (6) - Themes



Themes

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(.~ class, nrow = 1) +
  stat_smooth(aes(x=displ, y=hwy), method = 'lm', se = F, col="red") +
  scale_y_continuous("Displacement of Engines in litres") +
  scale_x_continuous("Highway miles per gallon") + coord_cartesian() +
  theme(panel.background = element_blank())
```



The background image shows a panoramic view of a dense urban area, likely Jakarta, Indonesia. In the foreground, there's a mix of lower residential buildings with red roofs and taller office buildings. The skyline is dominated by several modern skyscrapers, including one with a prominent 'FWD' logo and another with a 'Citi' logo. The sky is clear and blue.

5

7 Most frequent used graphs

Discrete vs Continuous

- **Discrete Data can only take certain values.**

Example: Number of people in a certain room, model of the cars, number of cars in a park area.

- **Continuous Data can take any value (within a range)**

Example: weight of body in kilograms, length of the routes in km, the capacity of a car fuel tank in litres, and so on.

Quiz: Which data is discrete?

- A person's height.
- Time in a race.
- A dog's weight.
- The length of a leaf.
- The number of students in a class.
- The results of rolling 2 dice.
- Gender that contains female and male.

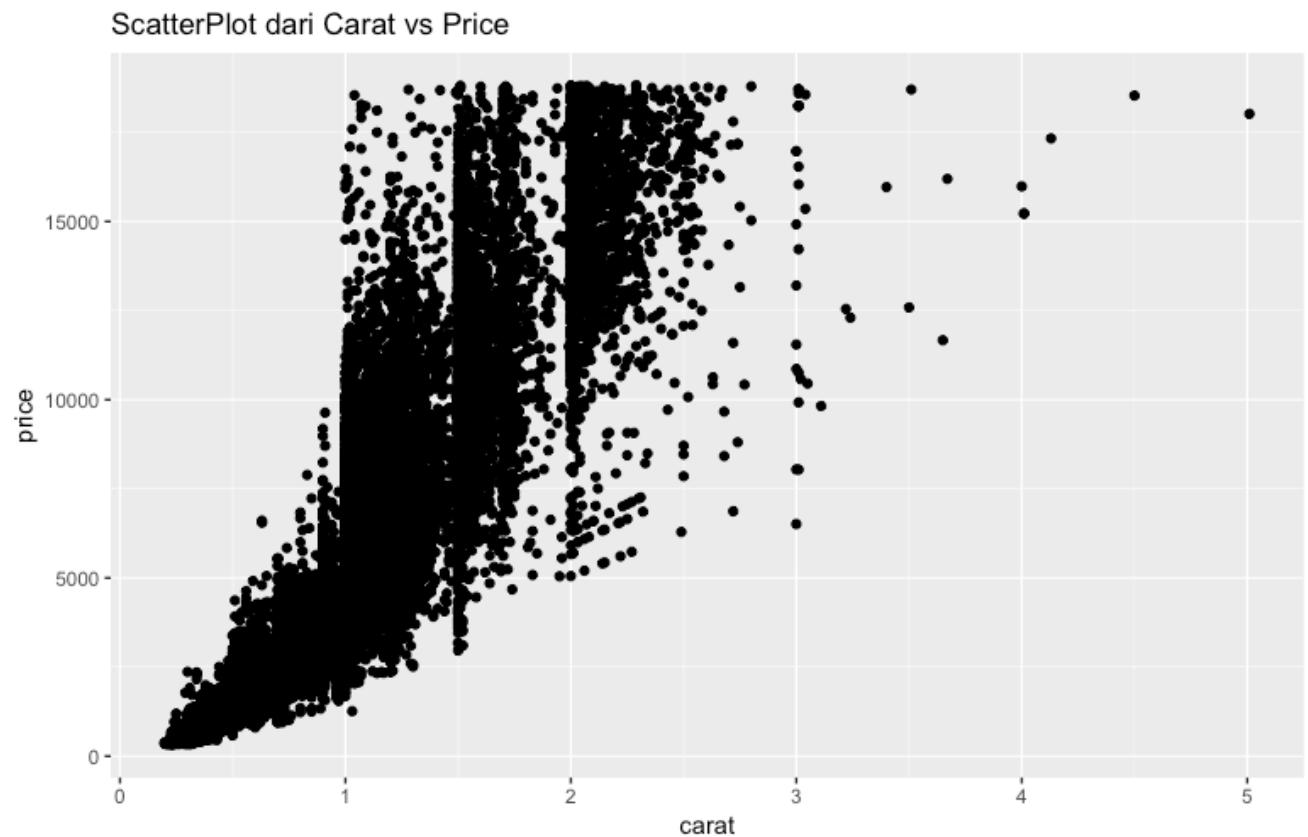
Answer: Which data is discrete?

- A person's height.
- Time in a race.
- A dog's weight.
- The length of a leaf.
- The number of students in a class.
- The results of rolling 2 dice.
- Gender that contains female and male.

1. Scatter Plot

When to use: Scatter Plot is used to see the relationship between two continuous variables.

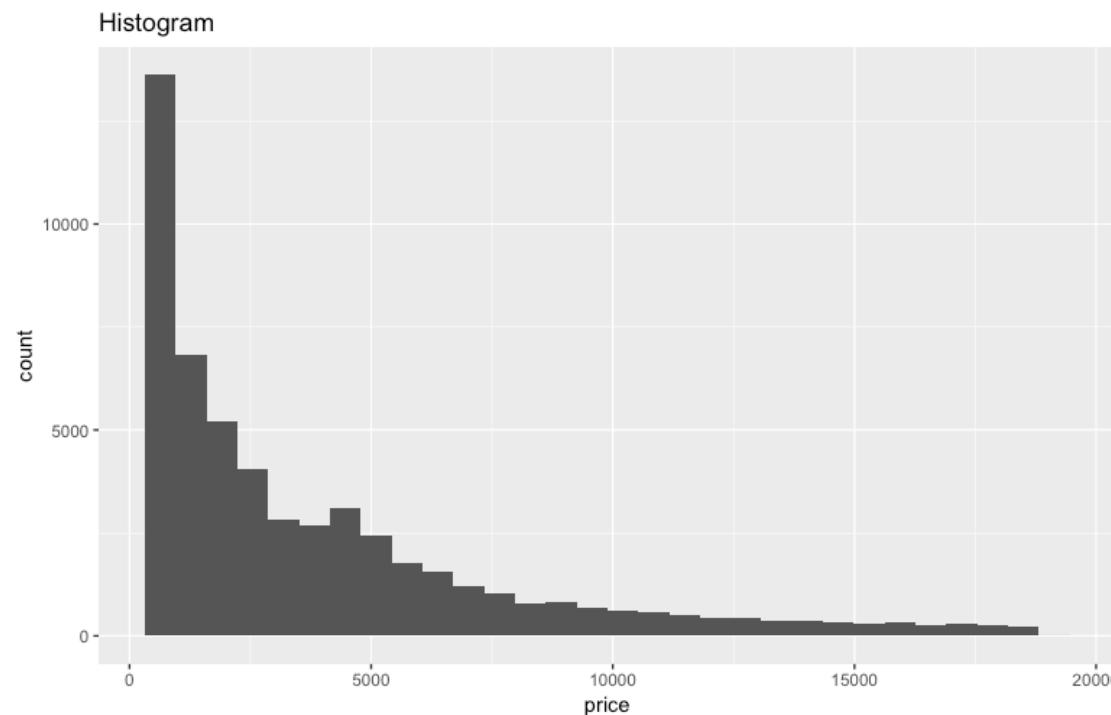
```
ggplot(diamonds) +  
  geom_point(mapping = aes(x = carat,  
                           y = price)) +  
  labs(title="ScatterPlot Carat vs  
Price")
```



2. Histogram

When to use: Histogram is used to plot continuous variable. It breaks the data into bins and shows frequency distribution of these bins. We can always change the bin size and see the effect it has on visualization.

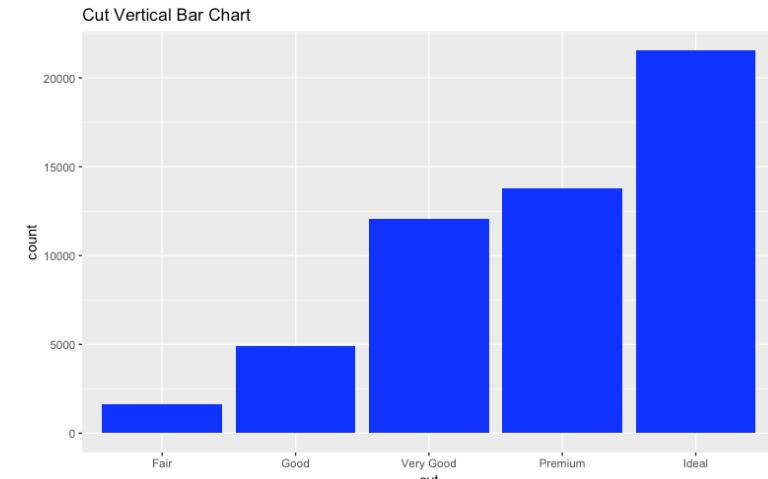
```
ggplot(diamonds) + geom_histogram(mapping= aes(x=price), bins=30) + labs(title="Histogram")
```



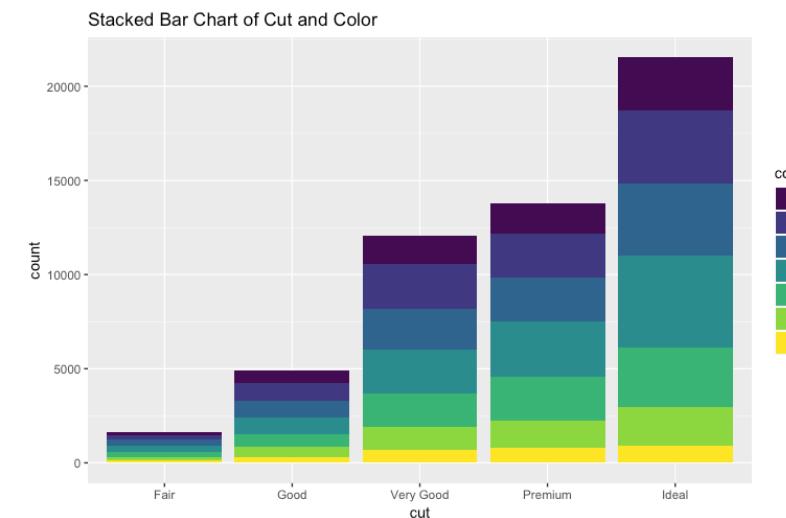
3. Bar and Stack Bar Chart

When to use: Bar charts are recommended when you want to plot a categorical variable or a combination of continuous and categorical variable.

```
ggplot(diamonds) + geom_bar(aes(x=cut)) +
  labs(title="Cut Vertical Bar Chart")
```



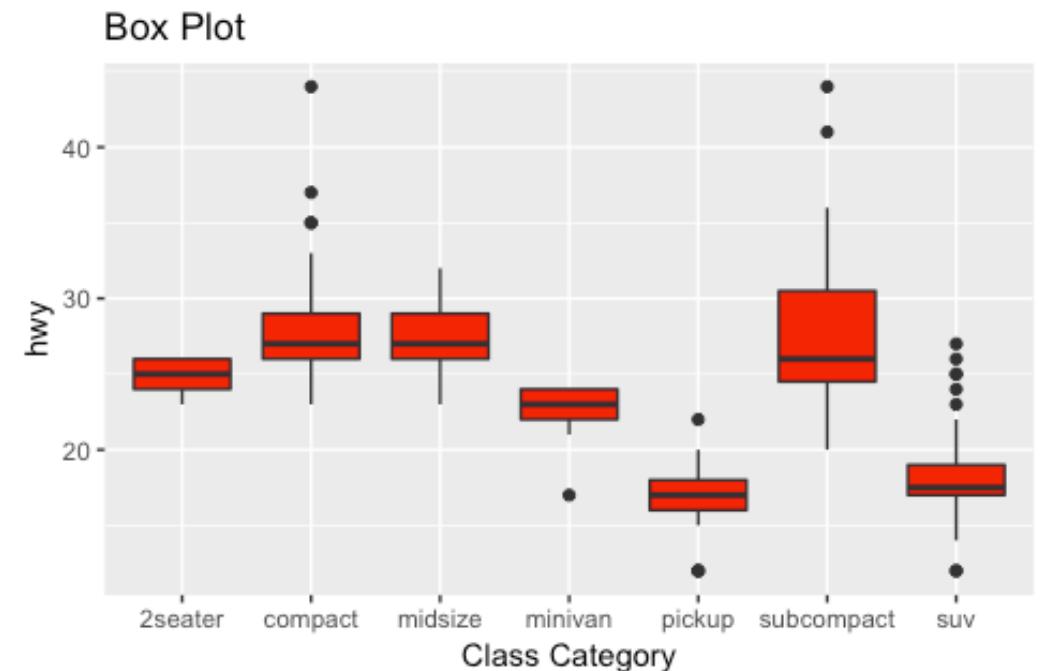
```
ggplot(diamonds) + geom_bar(aes(x=cut,
  fill=color)) + labs(title="Stacked Bar
Chart of Cut and Color")
```



4. Boxplot

When to use: Box Plots are used to plot a combination of categorical and continuous variables.

```
ggplot(mpg) + geom_boxplot(aes(class,  
hwy), fill="red") + labs(title = "Box  
Plot", x = "Class Category")
```

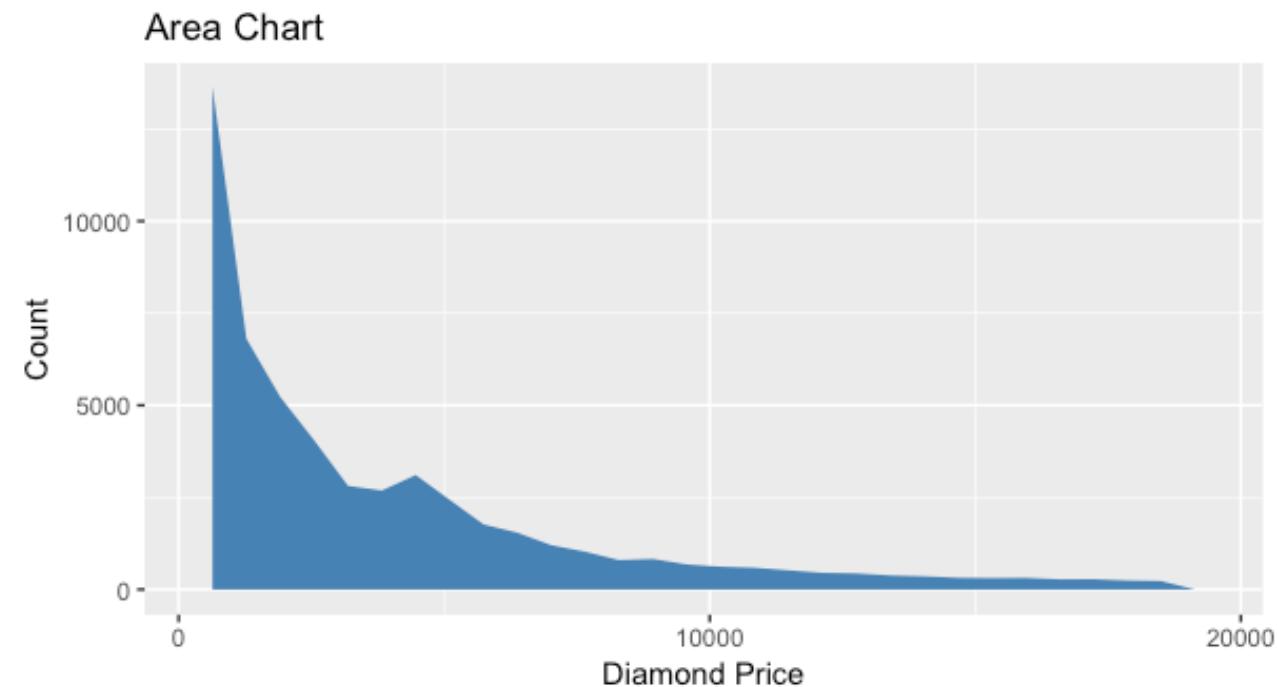


This plot is useful for visualizing the spread of the data and detect outliers. It shows five statistically significant numbers- the minimum, the 25th percentile, the median, the 75th percentile and the maximum.

5. Area Chart

When to use: Area chart is used to show continuity across a variable or data set. It is very much same as line chart and is commonly used for time series plots.

```
ggplot(diamonds) + geom_area(aes(price),  
stat="bin", bins=30, fill="steelblue") +  
scale_x_continuous(breaks=seq(0,20000,1000  
0)) + labs(title = "Area Chart",  
x="Diamond Price", y= "Count")
```

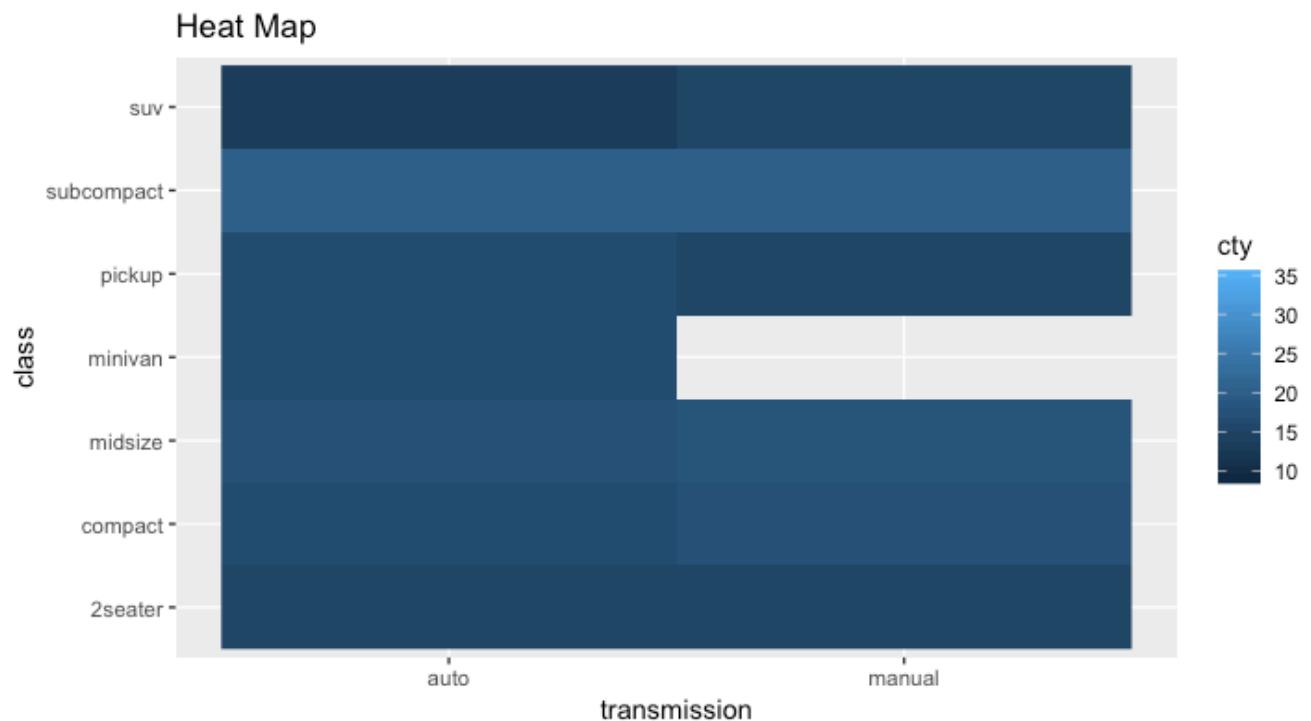


6. Heatmap

When to use: Heat Map uses intensity (density) of colors to display relationship between two or three or many variables in a two dimensional image.

```
mpg$transmission <- ifelse(grepl("auto",
  mpg$trans), "auto", "manual" )
```

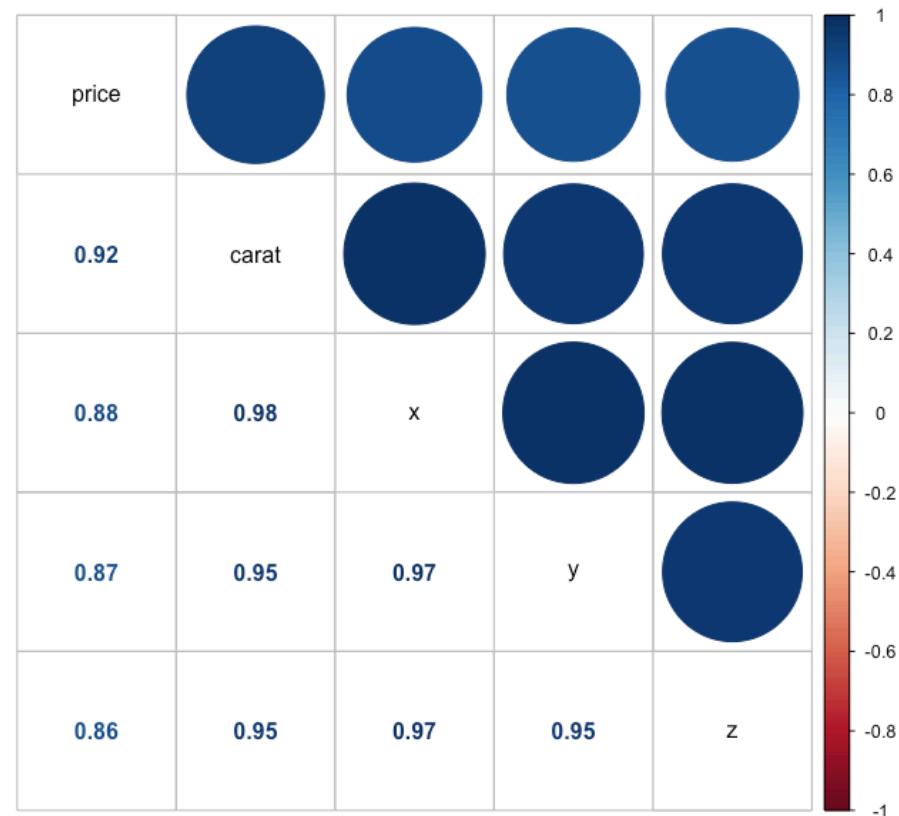
```
ggplot(mpg, aes(transmission, class)) +
  geom_raster(aes(fill=cty)) +
  labs(title="Heat Map", x="transmission",
  y="class") +
  scale_fill_continuous(name="cty")
```



7. Correlogram

When to use: Correlogram is used to test the level of co-relation among the variable available in the data set. The cells of the matrix can be shaded or colored to show the co-relation value.

```
library(corrplot)  
  
corrplot.mixed(cor_numVar, tl.col="black")
```



Let's practice!

Under this github, please open exercise.R

<https://github.com/arikunco/visualization.git>

Summary

- Exploratory vs Explanatory
- Basic presentation types: Comparison, Composition, Distribution, Relationship
- Grammar of graphics
- 7 most frequent used graphs: (scatterplot, histogram, bar & stack bar chart, box plot, area chart, heatmap, correlogram)

References

www.storytellingwithdata.com/book

<https://r4ds.had.co.nz/>

<https://www.mathsisfun.com/data/data-discrete-continuous.html>

<http://labs.juiceanalytics.com/chartchooser/index.html>

<https://www.datacamp.com>

<https://www.r-bloggers.com/7-visualizations-you-should-learn-in-r/>