ISE 201 Math Foundations for Data Science and Decision Science

Descriptive Statistics and EDA

Dr. Shilpa Gupta
ISE @ SJSU
2023-08-31

Questions

In class exercise

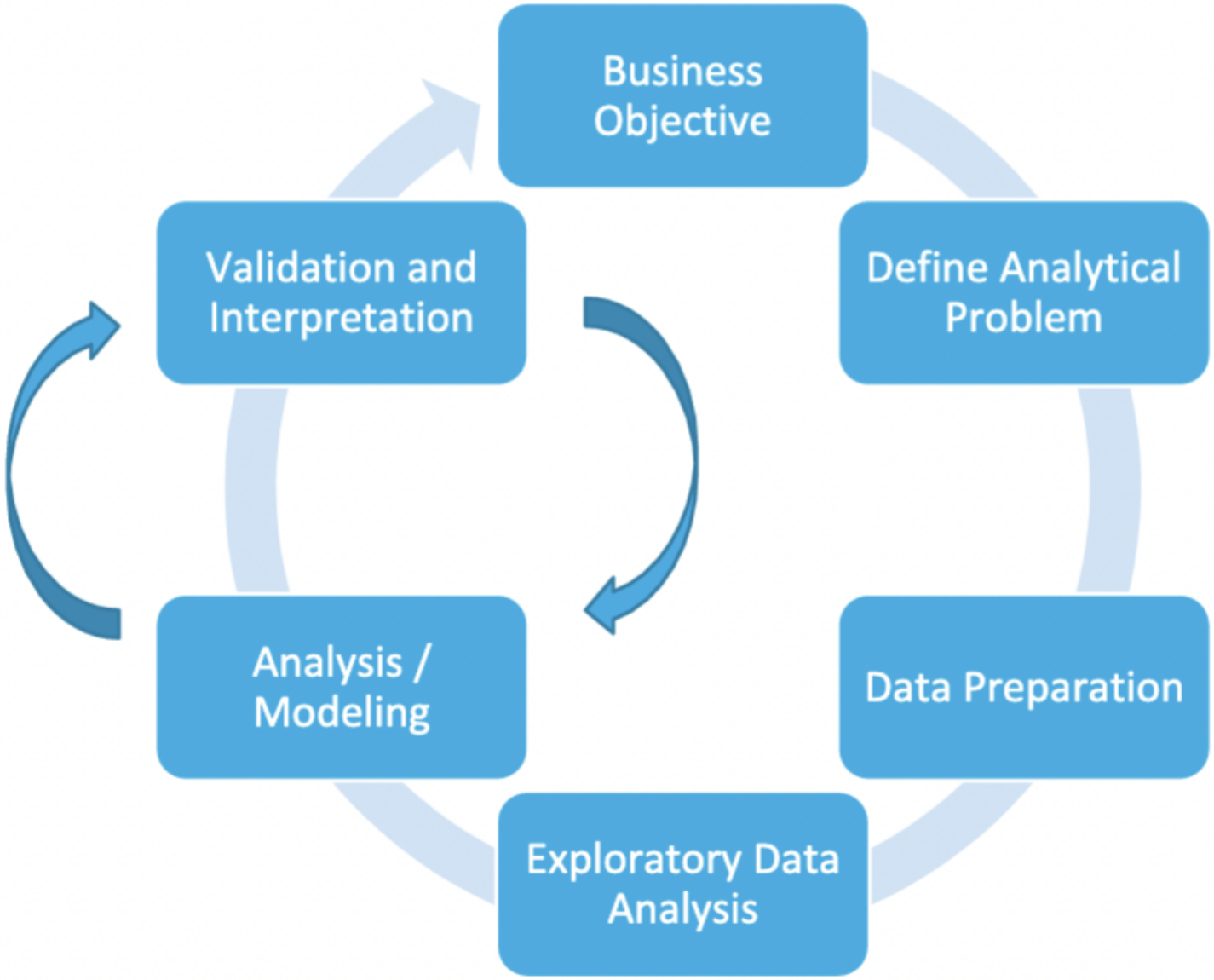
Knit RMarkdown document

[Recap] Science of Data

Collection, analysis, interpretation and presentation of data to:

- Make decisions
- Solve problems
- Design (or improve) product and /or processes

Data Science Process



Population vs Sample

Population: Set of all items or events of interest

Sample: Subset of the population

Statistical Inference

Make valid conclusions about the population by observing the sample

A sample is **representative** and **random**

Free from over-representation or under-representation of a subgroup.

Every entity has equal chance to be selected.

Collecting Data

- Retrospective study

- Historical process data over same period of time
- Helpful in understanding correlation relationships
- Limited by amount and type of data collected

- Observational Study

- Limited in how long the data is collected for
- Include additional quantities not measured in retrospective study

- Designed Experiment

- Deliberate and purposeful changes
- Helpful in understanding causal relationships

Exploratory Data Analysis

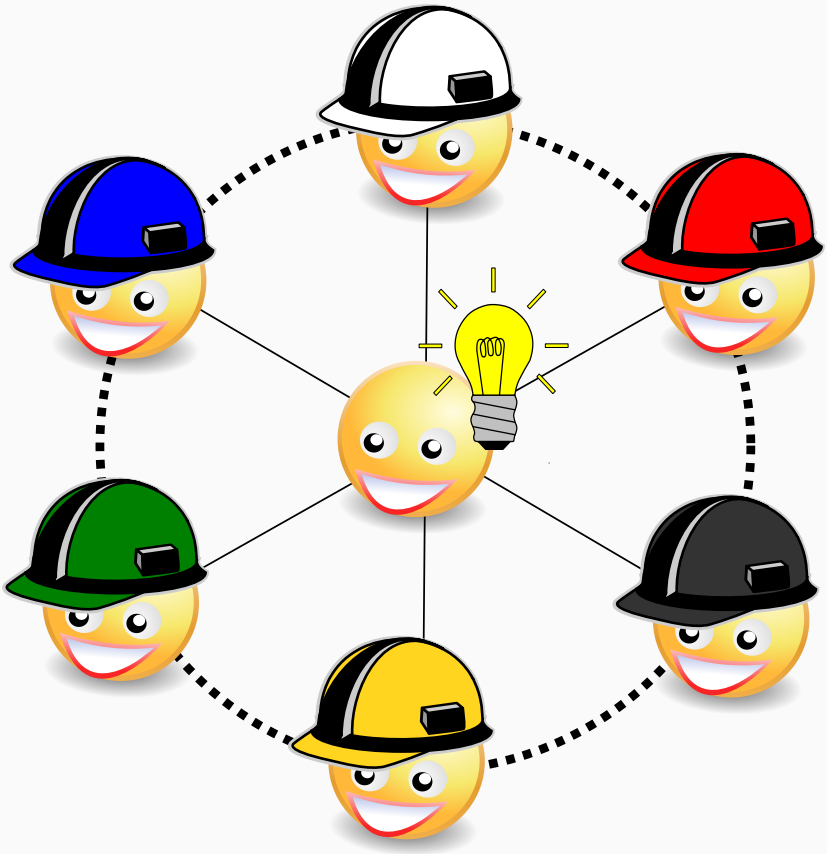
Generate questions about your data.

Search for answers by visualizing, transforming, and modeling your data.

Use what you learn to refine your questions and/or generate new questions.

In class exercise

Ask Questions



.footnote [Source]

Generating Data Questions

Descriptive

What is the summarized statistic of a dataset?

Exploratory

Are there any patterns, trends or relationships within a dataset?

Inferential

What can be inferred about the population from the sample?

Predictive

What would the outcome be for a certain combination of features ?

Causal

How does change in the levels of one factor causes changes in the other factor?

In class exercise

Refine Questions

Descriptive Statistics

Understanding Data

```
head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Random Variable

Data Types: Continuous and Categorical

Frequency Tables

cyl	n
4	11
6	7
8	14

```
mtcars %>%
  count(cyl) %>%
  group_by(cyl) %>%
    knitr::kable(format = "html")
```

Numerical Summaries

Measures of central tendency

Mean, Median, Mode, ...

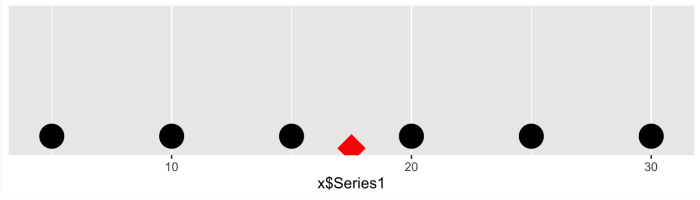
Measure of spread

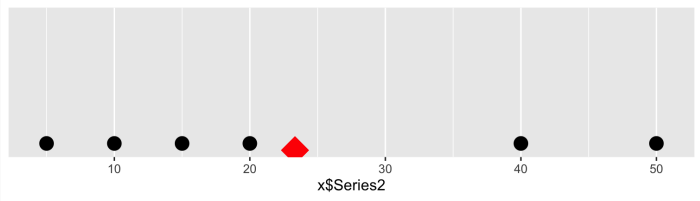
Range, Variance, Standard Deviation...

Central Tendency: Mean, Median, Mode

Arthmetic mean

$$\bar{x} = \frac{1}{n}\sum_{i=i}^{n} x_i$$





Series1	5	10	15	20	25	30
Series2	5	10	15	20	40	50

```
mean(x$Series1)
## [1] 17.5

mean(x$Series2)
## [1] 23.33333
```

Missing data

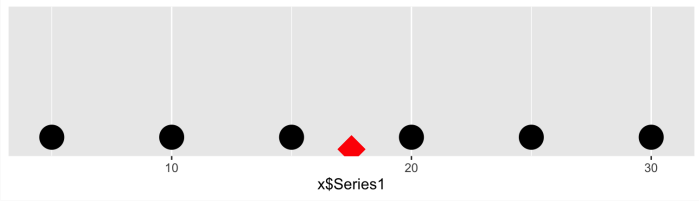
If you have missing data

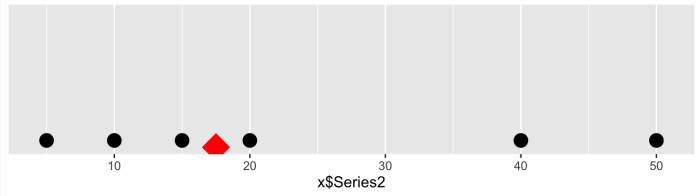
```
x ← c(2, 3, 4, 5, NA)
mean(x)
## [1] NA
```

```
  mean(x, na.rm=TRUE)
## [1] 3.5
```

Central Tendency: Median

Median is the middle score or average of two middle scores when the data is **rank ordered**





Series1	5	10	15	20	25	30
Series2	5	10	15	20	40	50

```
median(x$Series1)
## [1] 17.5

median(x$Series2)
## [1] 17.5
```

Quartile / Quantile / Percentile

Ordered data is divided into five equal parts

Lower Quartile or First Quartile (Min - 25th percentile)

Upper Quartile or Third Quartile (75th percentile - Max)

```
help(quantile)
```

Reference

Central Tendency: Mode

Most frequently occuring data value

cyl	average	median	mode	total
4	26.66364	26.0	21	11
6	19.74286	19.7	21	7
8	15.10000	15.2	15	14

```
mode ← function(x){
  which.max(tabulate(x))
}

mtcars %>%
  group_by(cyl) %>%
  summarise(average = mean(mpg),
            median = median(mpg),
            mode = mode(mpg),
            count = n())
```

Spread

Range

$$range = \max(x) - \min(x)$$

IQR

$$IQR = UpperQuartile - LowerQuartile$$

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Population versus Sample

Population Variance

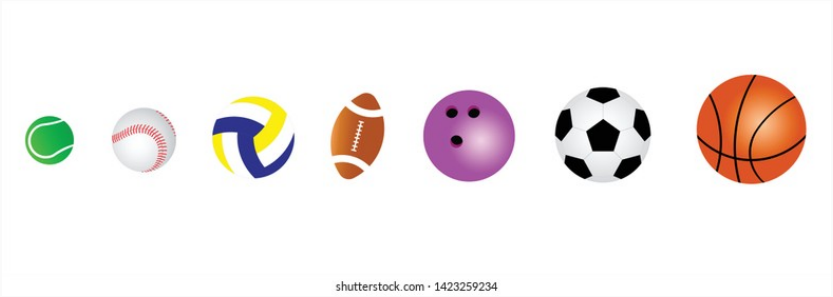
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Degrees of Freedom ?

Degrees of Freedom



How many choices do they have...

Team 1:

Team 2:

Team 3:

Team 4:

Team 5:

Team 6:

Team 7:

Simplifying sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}x_i)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

```
help("sd")

help("var")
```

Relationships

Covariance

$$cov_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation

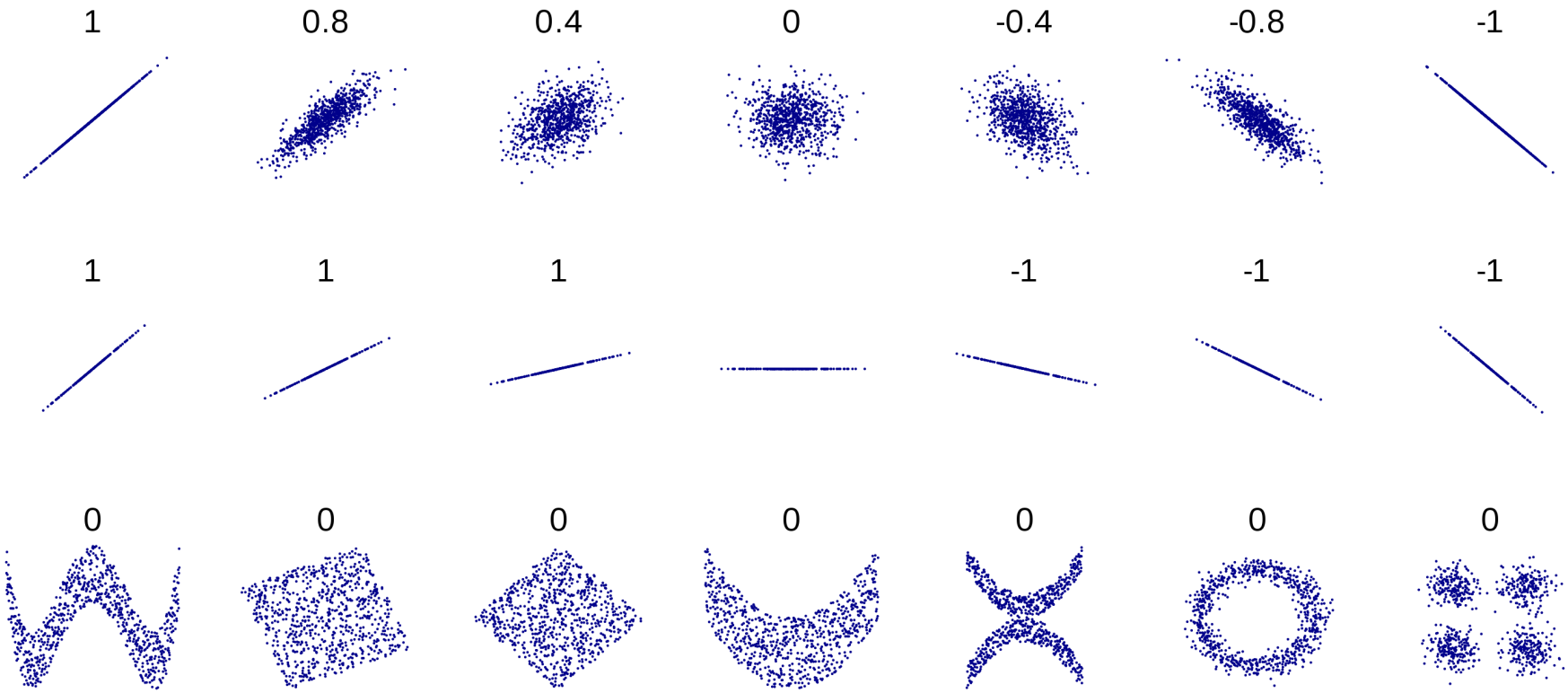
Pearson product moment coefficient

Ranges from -1 to 1

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

$$= \frac{cov_{x,y}}{s_x s_y}$$

Pearson Correlation Coefficient



Source: Wikipedia

Distance Measures

Euclidean Distance

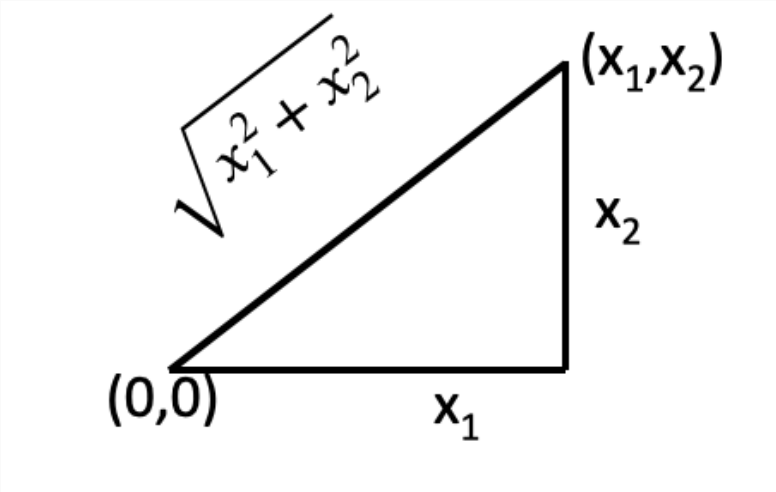
Euclidean distance of point (x_1, x_2) from origin

$$\sqrt{x_1^2 + x_2^2}$$

*each point contributed equally to the calculation of the euclidean distance

Equation of a circle

$$x_1^2 + x_2^2 = c^2$$



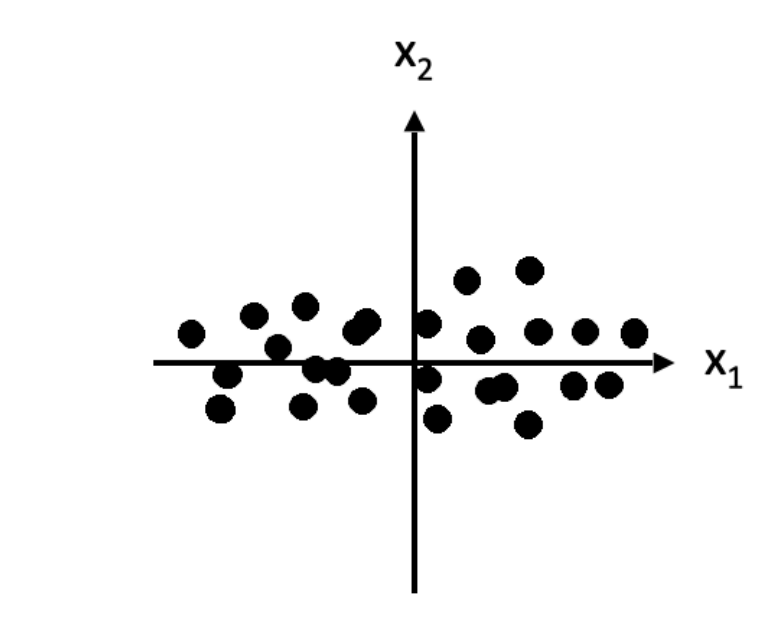
Statistical Distance

Suppose, n pairs of measurements on two variables x_1 and x_2

Mean of each variable is zero

x_1 and x_2 are **independent** (not correlated)

Variability of x_1 > Variability of x_2



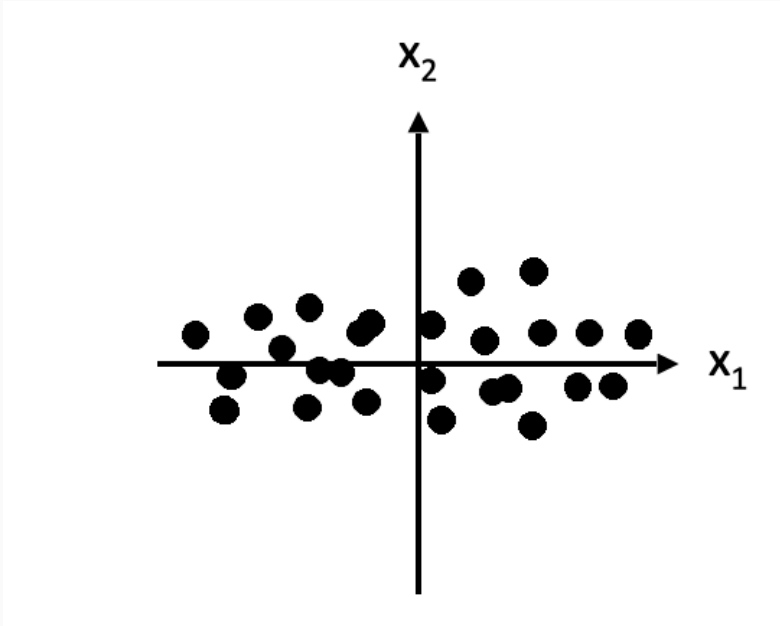
Statistical Distance

Suppose, n pairs of measurements on two variables x_1 and x_2

Mean of each variable is zero

x_1 and x_2 are **independent** (not correlated)

Variability of x_1 > Variability of x_2



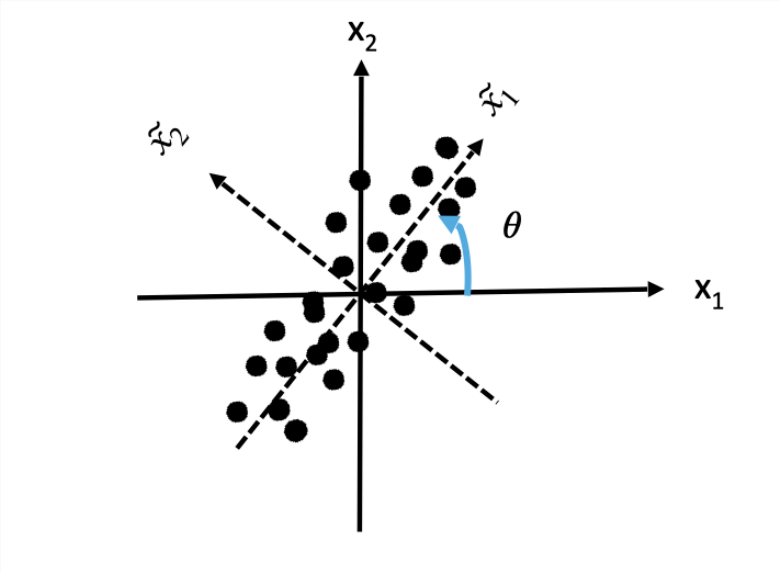
Divide each coordinate by sample standard deviation

$$x_1{}^* = \frac{x_1}{s_{11}}$$

$$x_2{}^* = \frac{x_2}{s_{22}}$$

$$Dist(0, P) = \sqrt{(x_1^*)^2 + (x_2^*)^2}$$

Statistical Distance



$$Dist(0,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}^2} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}^2}}$$

$$\tilde{x}_1 = x_1 cos(\theta) + x_2 sin(\theta)$$

$$\tilde{x}_2 = -x_1 sin(\theta) + x_2 cos(\theta)$$

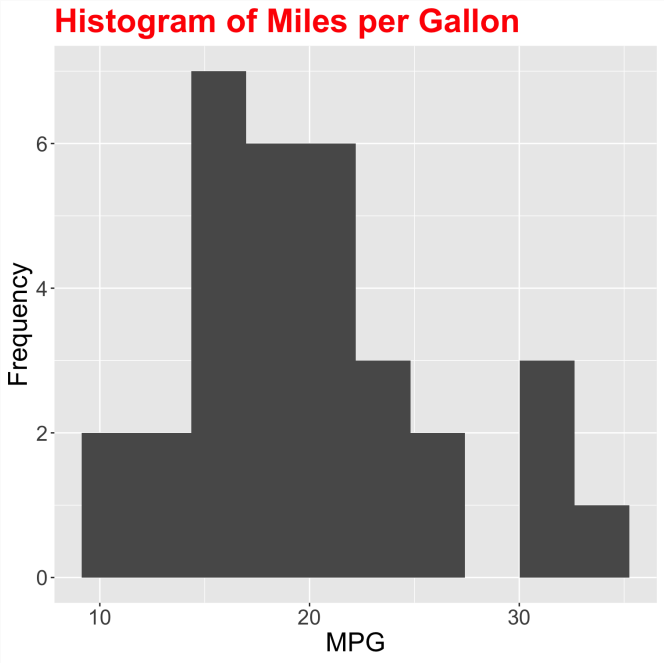
$$D(0,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2}$$

Visualization

Types of Visualization

- Distributions
- Trends
- Relationships
- Groups

Distributions : Histogram

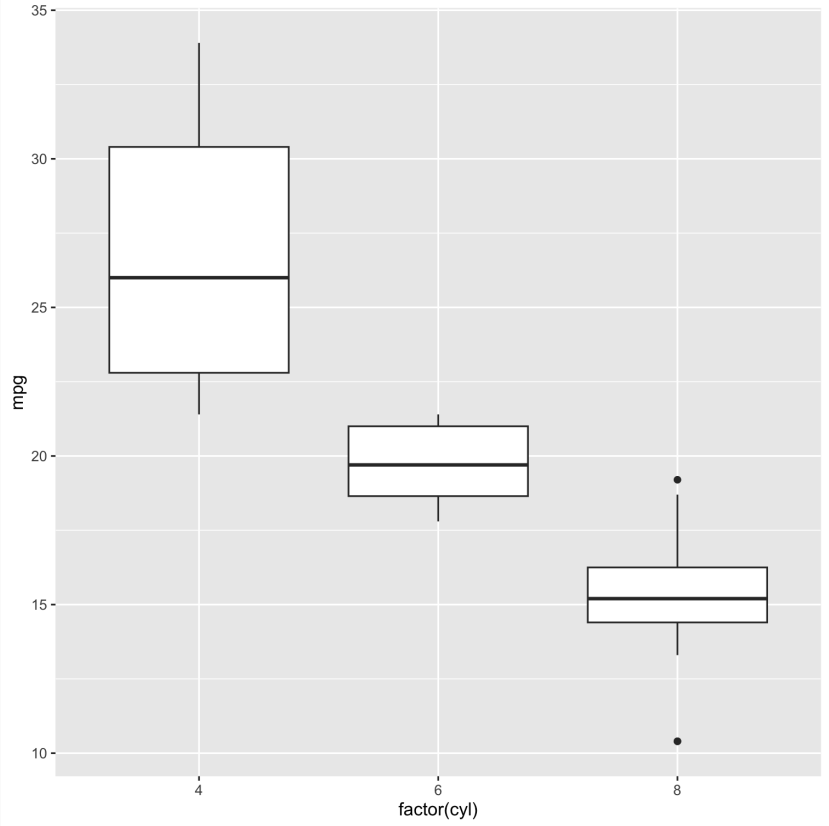



```
ggplot(data = mtcars) +
  geom_histogram(mapping = aes(x = mpg), bins = 10) +
  labs(title = "Histogram of Miles per Gallon", x = "MPG",y = "Frequency") +
  theme(plot.title = element_text(color = "red", size = 25, face = "bold")) +
  theme(text = element_text(size = 20))
```

*Number of bins ~ $\sqrt{n}$

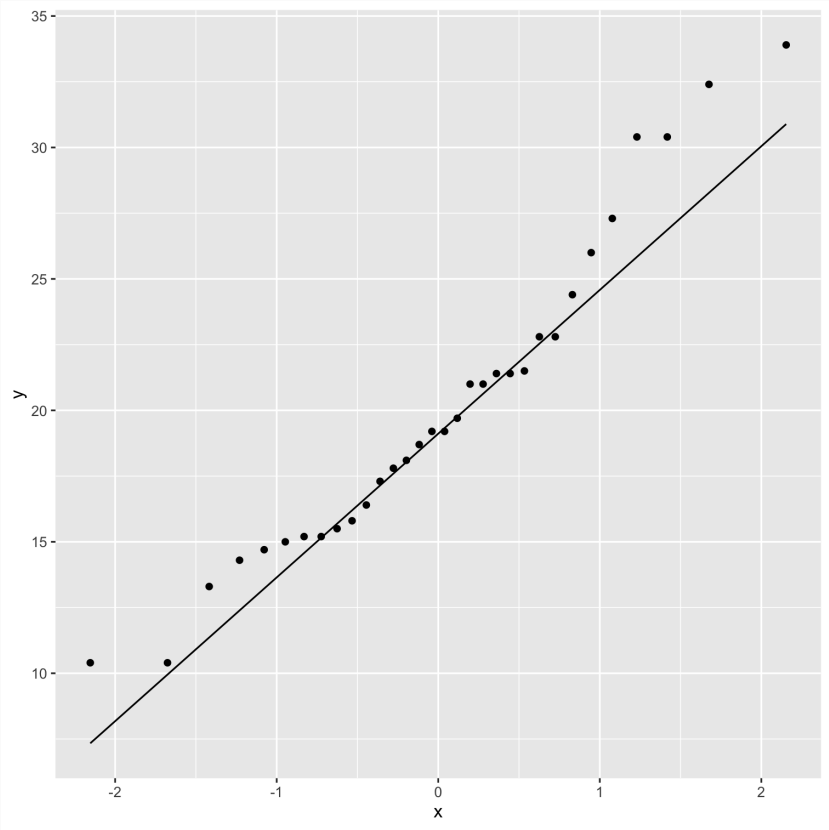
Distributions : Boxplot

box plot example



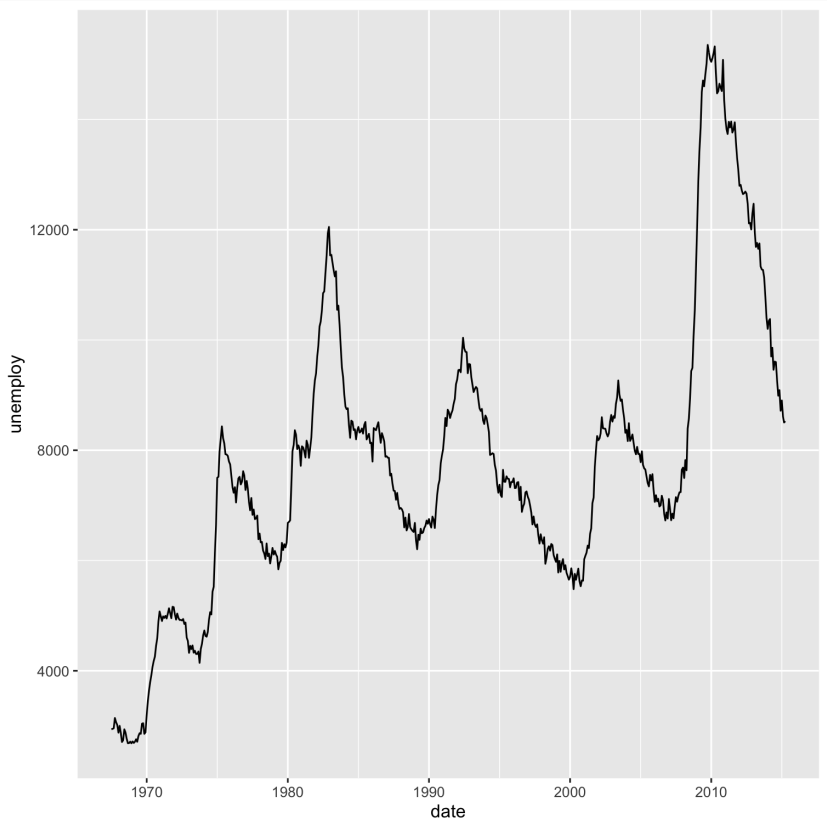

```
ggplot(data = mtcars) +
  geom_boxplot(mapping = aes(y = mpg, x = factor(cyl)))
```

Distributions: Quantile & Prob plots



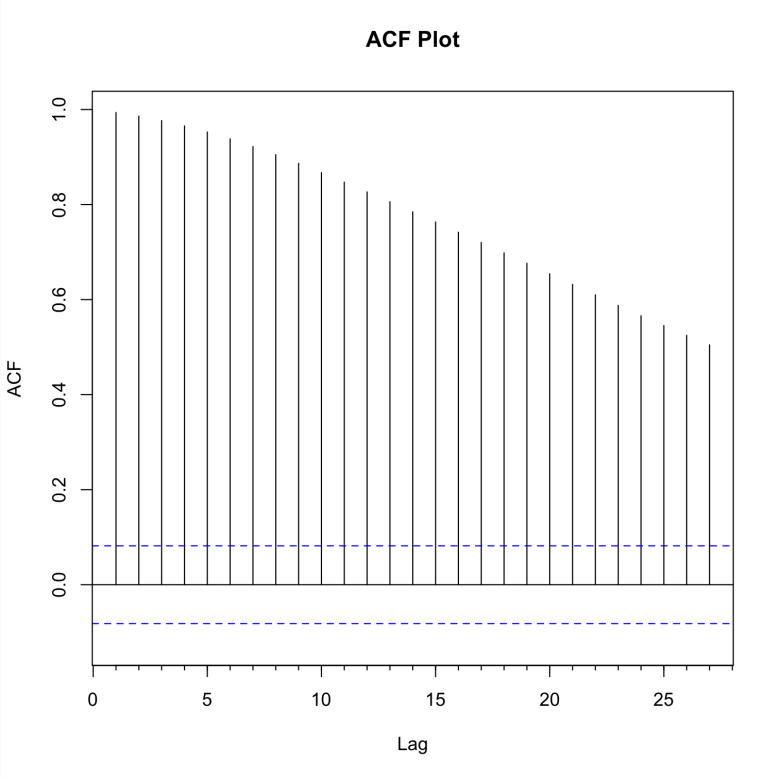

```
ggplot(mtcars, aes(sample=mpg)) + geom_qq() + stat_qq_line()
```

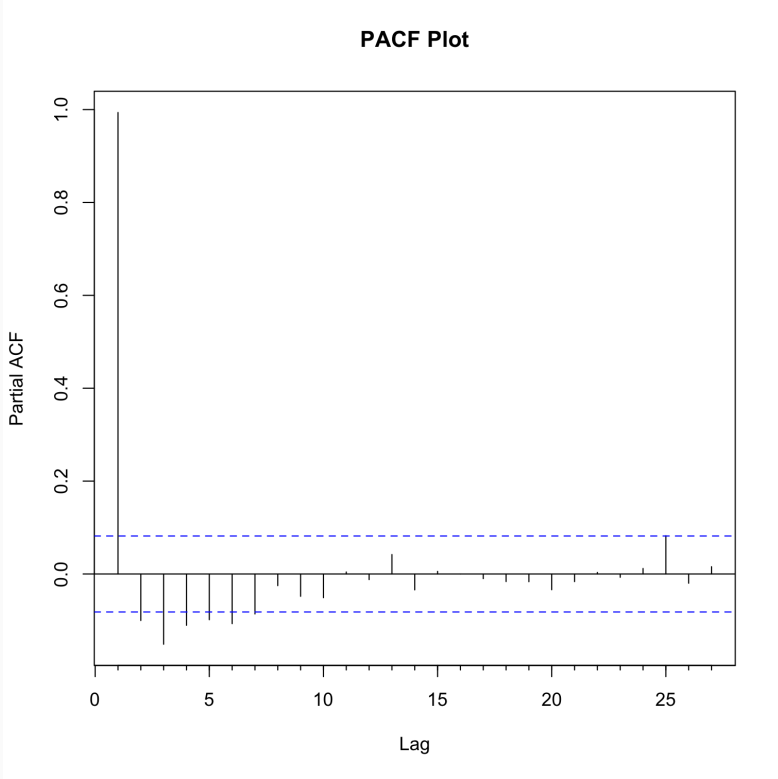
Trend over time




```
ggplot(data = ggplot2::economics) +
  geom_line(mapping = aes(y =  unemploy, x = date))
```

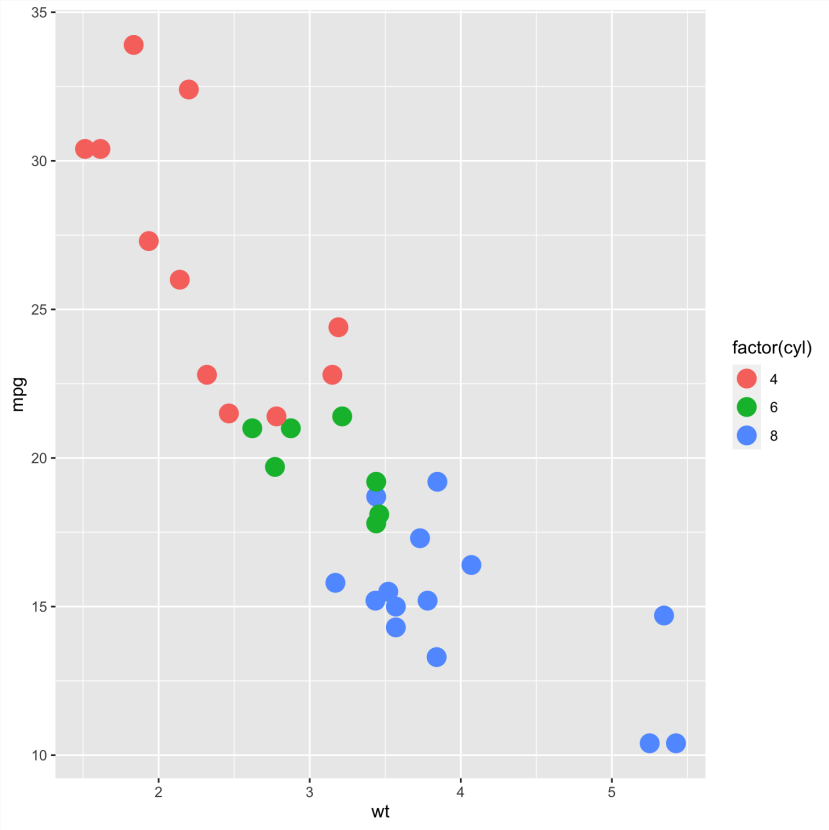
Autocorrelation





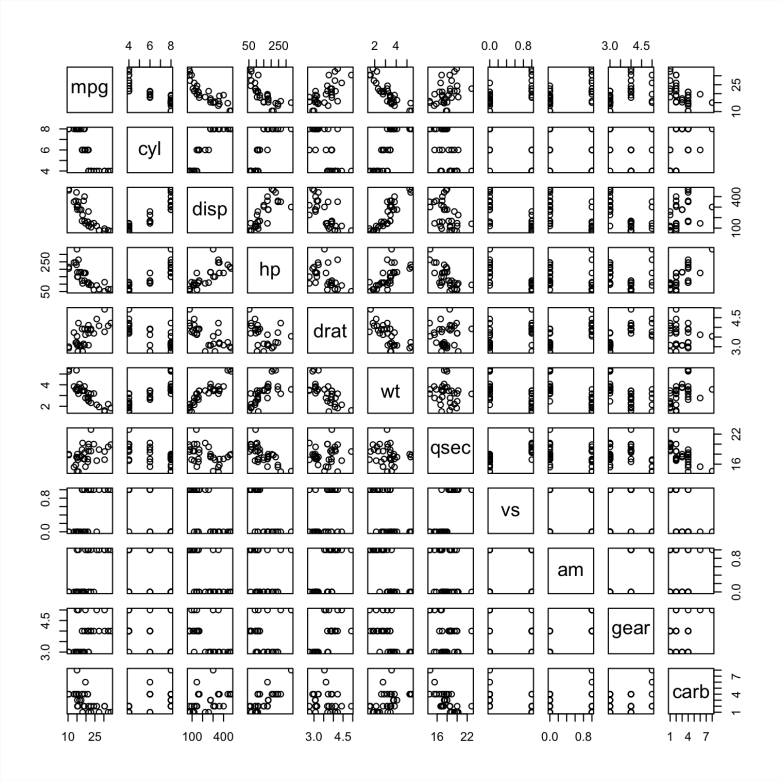
Relationships : Scatter plots

Scatter plots




```
ggplot(data = mtcars) +
  geom_point(mapping = aes(x = wt, y = mpg, color = factor(cyl)), size = 5)
```

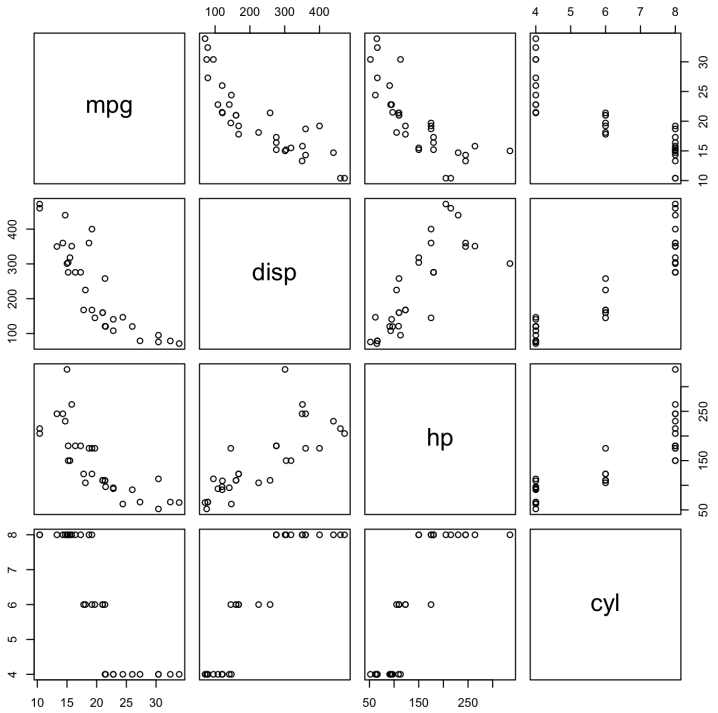
Relationships : Multiple Scatter plots




```
pairs(mtcars)
```

Reference

Relationships : Multiple Scatter plots




```
pairs(mtcars[,c("mpg", "disp", "hp",
                "cyl")])
```

Reference

Tidy Data

Tidy Data

Each variable must have its own column.

Each observation must have its own row.

Each value must have its own cell.

Making Data Tidy : Pivot

country	indicator	2000	2001	2002	2003	2004	
USA	SP.URB.TOTL	2.230691e+08	2.257923e+08	2.284003e+08	2.308766e+08	2.335327e+08	2.
USA	SP.URB.GROW	1.512011e+00	1.213380e+00	1.148419e+00	1.078360e+00	1.143885e+00	1

country	Year	SP.URB.TOTL	SP.URB.GROW	SP.POP.TOTL	SP.POP.GROW
USA	2000	223069137	1.512011	282162411	1.1127690
USA	2001	225792302	1.213380	284968955	0.9897414
USA	2002	228400290	1.148419	287625193	0.9277975

Making Data Tidy : Pivot (Code)

country	indicator	2000	2001	2002	2003	2004
USA	SP.URB.TOTL	2.230691e+08	2.257923e+08	2.284003e+08	2.308766e+08	2.335327e+08
USA	SP.URB.GROW	1.512011e+00	1.213380e+00	1.148419e+00	1.078360e+00	1.143885e+00
USA	SP.POP.TOTL	2.821624e+08	2.849690e+08	2.876252e+08	2.901079e+08	2.928053e+08
USA	SP.POP.GROW	1.112769e+00	9.897414e-01	9.277975e-01	8.594817e-01	9.254840e-01

```
world_bank_pop %>%
  filter(country == "USA" ) %>%
  pivot_longer(str_c(2000:2017), names_to = "Year") %>%
  pivot_wider(names_from = indicator, values_from = value)
```

Combine columns (unite)

	mpg	cyl	disp	hp	drat	wt	qsec	vs_am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0_1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0_1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1_1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1_0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0_0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1_0	3	1

```
unite(mtcars, "vs_am", c("vs","am")) %>% head()
```

Go to slide 1 for original data view

Splitting columns (separate)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

```
mtcars %>%
  unite("vs_am", vs, am) %>%
  separate(vs_am, c("vs", "am")) %>%
  head() %>%
```

Go to slide 1 for original data view

Making Data Tidy: Missing values

```
# Checking for missing values
sum(is.na(world_bank_pop))
```

```
help("complete")
```

```
help("fill")
```

Questions?

3 phases of visualization

1. Exploration for self

2. Exploratory Data Analysis for generating consensus among team members

3. Storytelling with data for decision makers

As we go from $1 \rightarrow 3$, the **impact** of the visualization is increasing and so is the number of people seeing it.

R Demo

Sample EDA