

MODULE - 9

Data Visualization

The best way to learn is by doing it

Activity - 1

Take a look at

[Worldwide Covid 19 data visualization here](#)

OR

[India Covid-19 Data visualization here.](#)

The best way to learn is by doing it

Activity - 2

Let us do a pulse rate activity. Find your pulse from your hand or neck and measure for 15 seconds sharp. Multiply by 4 and report the number. We shall then analyze the responses. Note, you can repeat and add measures as many times as you want.

Loading...

Let us answer some questions:

1. Are all responses same? Why, Why not? Can you think of factors that lead to *variability*
2. Will the response be same if you measure it again? If not, can we trust this data to make decisions if any?
3. How can we describe this data to someone? Is there a *central tendency*
4. Can you think of applications where you would want to collect such data?

https://docs.google.com/forms/d/e/1FAIpQLSfMd6Dog86lux0OJ1h_xiHiRnNCr0ZLB-Tv4CWVNC1EiK9GaQ/viewform?fbzx=-4866849235103203839



Basic Terms

Variable

Characteristic of an item or individual like gender, height, age, color, intelligence, income, etc.

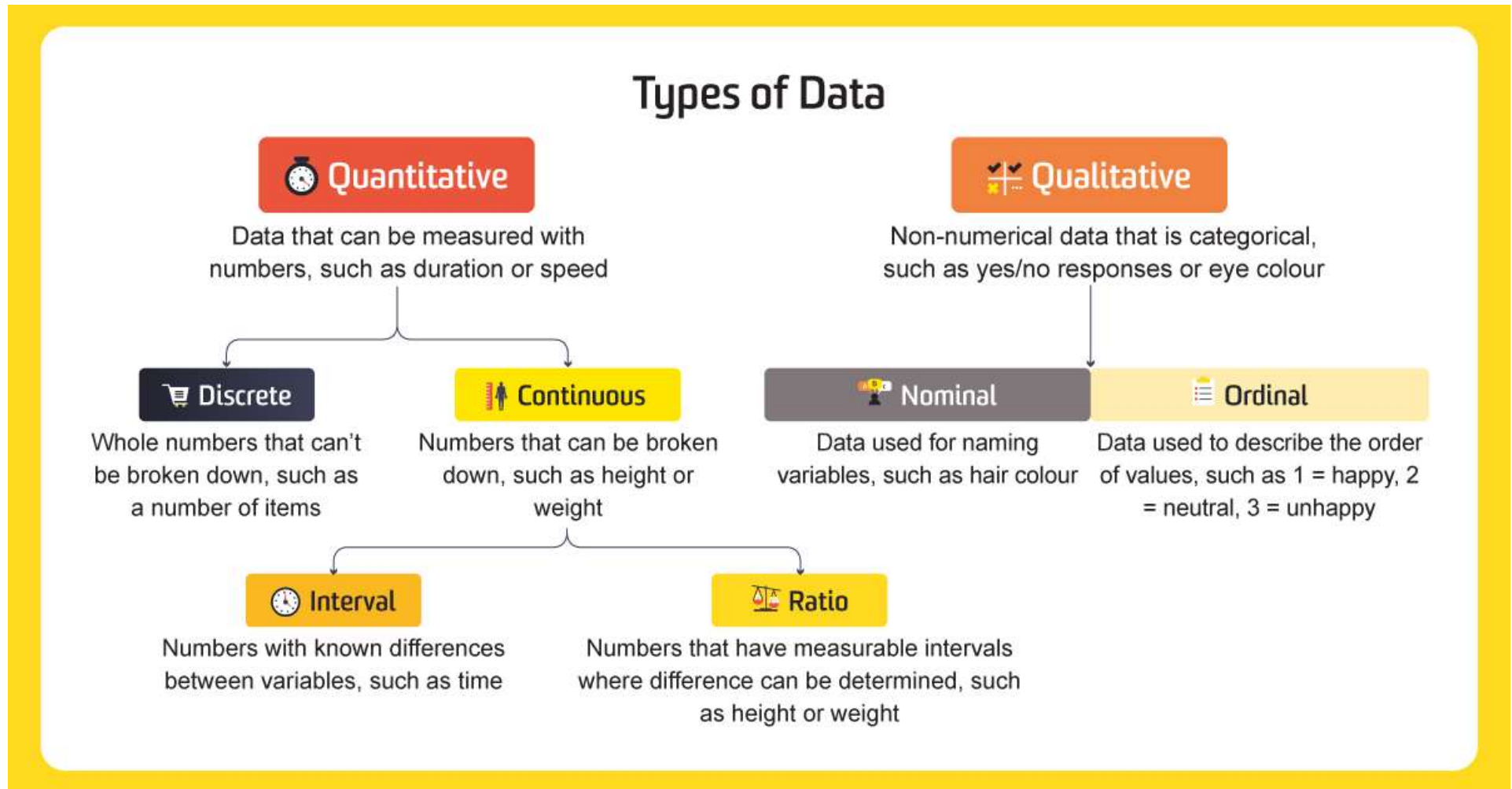
Data

Set of individual values associated with a variable

Statistics

Methods that help us transform data into something useful - e.g. making a decision. When it helps us to describe or summarize data in an easy to understand form, it is called **Descriptive Statistics**. In real world, it is impossible to collect all data for entire population. Often a sample is collected and a hypothesis is formed based upon the sample to describe the population. The ways that help collect data, form and verify hypothesis, is called **Inferential Statistics**.

Types of Variables / Data



Types of Variables / Data

- ❖ Data can be either Quantitative(a number) or Qualitative(not a number). Numeric data itself can be something that can be measured (Continuous) like temperature, weight etc. or something that is counted (Discrete). The counted one is always an integer, while continuous values are floating point numbers. Numbers will always have an order. Money is both measurable and countable. Measured data usually must have units as well.
- ❖ Other type of data is Qualitative or categorical data. This type of data is also of two types - Ordinal, where there is an order e.g. T shirt sizes (S, M, L), Weekdays etc. Nominal data does not have order e.g. gender, country, city etc.
- ❖ This is a brief overview. A detailed guide can be found [here](#).

The best way to learn is by doing it

Activity – 3 and 4

Activity 3

Revisiting our pulse rate experiment, Can you determine what data type is each factor that affect variability of pulse rate in each individual.

Activity 4

Can you revisit Covid Data and determine what are the data types of various features visualized in there.

Importance of Data Types

Data types are important in that they help us:

1. **Describe the data:** Different types of data is described differently
2. **Visualize the data:** We apply different visualization techniques on different types of data
3. **Prepare forms to input data:** Different input elements are used to input different types of data.
4. **Model the data for machine learning:** As we know ML techniques are Computer Algorithms that must work on numbers, so depending on data type we may have to formulate it in a manner that can be fed to appropriate ML model. For example we can use *one hot encoding* for nominal data and *one label encoding* for ordinal data to convert them from labels to numbers. (more on that later)

Why Statistics?

To summarize data to shape how we make decisions.

Decisions about what and why? We usually want to decide on certain features that we can vary in order to predict, or drive, certain outcome. For example, Do costume affect performance? Or Is tumor size a good estimator of malignancy? etc.

Statistics help us place faith in our chosen hypothesis and can also tell us presence of lurking parameters.

The New COKE Mistake

Quite contradictory to the statistical study that Coke did, their new Coke was a BIG Failure. Read more about it [here](#).



Hope you people remember

Population vs Sample

What is sampling?

Measures of Central Tendency

Outlier

Variability – Variance, Standard Deviation

Inter Quartile Range

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

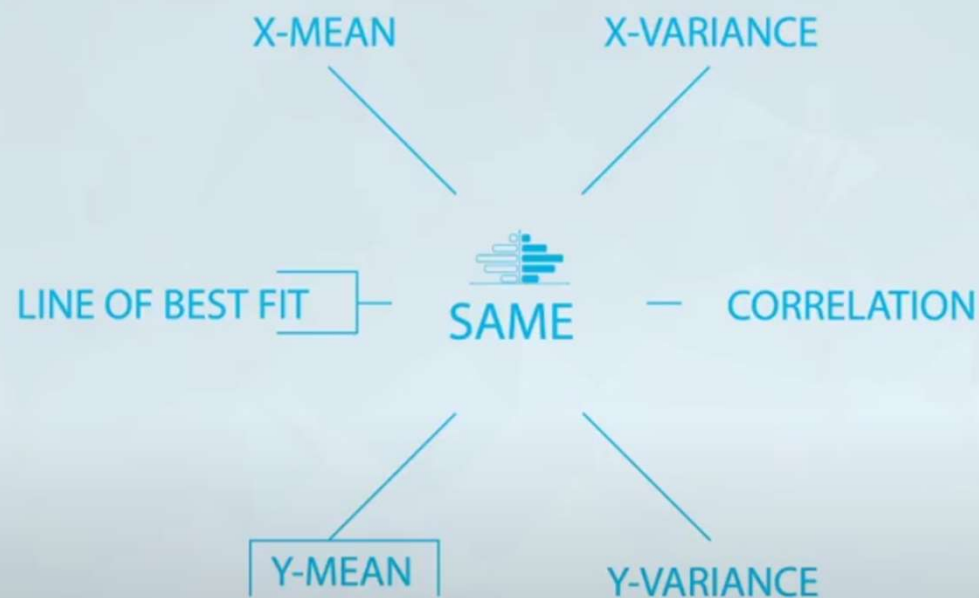
The best way to learn is by doing it

Exercise

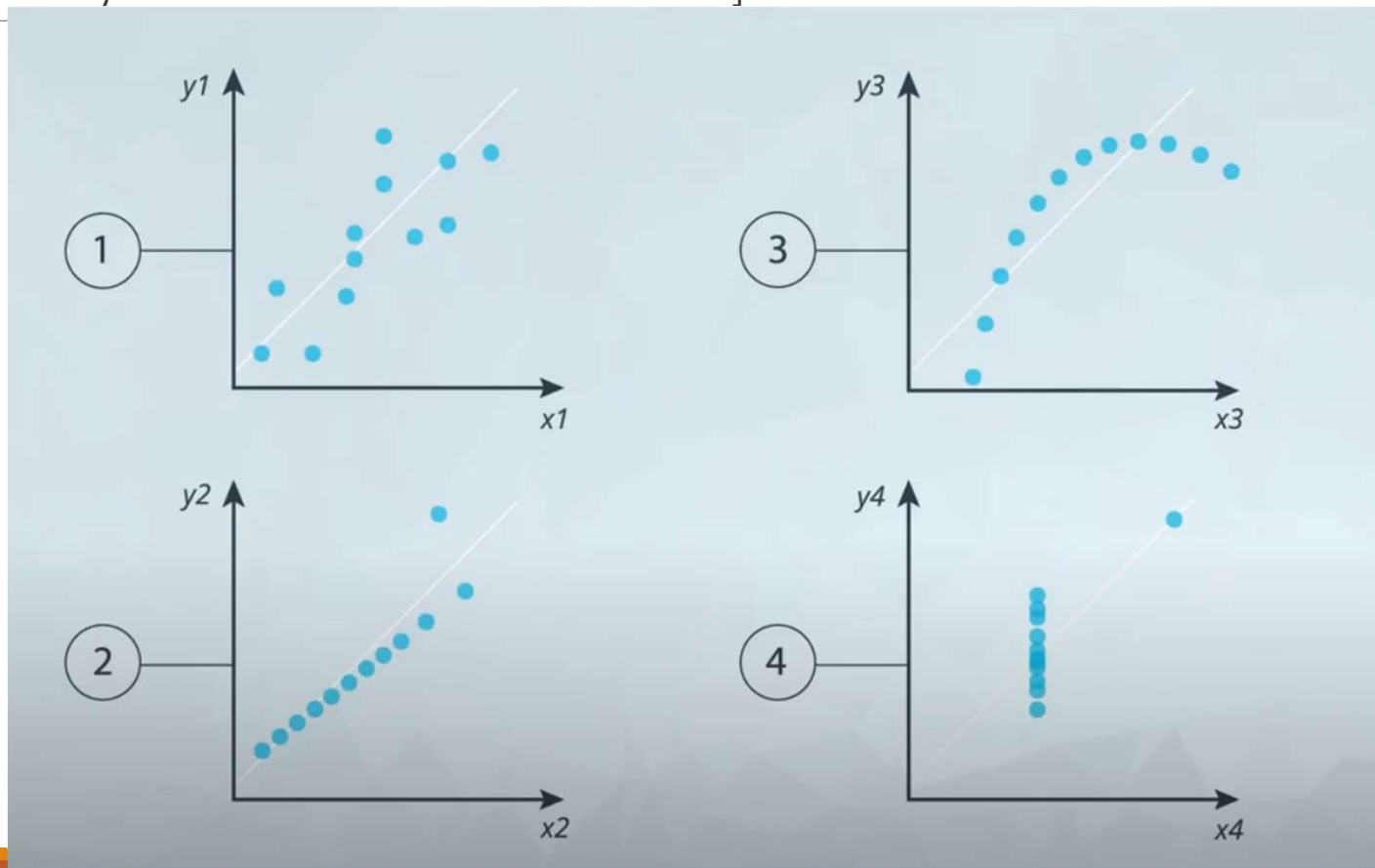
- ✓ Make a copy of this [Google Sheet](#) and answer whether each of the following are the same or different?
 1. means associated with any of the X columns?
 2. means associated with any of the Y columns?
 3. standard deviation associated with any of the X columns?
 4. standard deviation associated with any of the Y columns?
- ✓ NOTE You can use AVERAGE and STDEV functions in excel to do that.

More recently Alberto Cairo created the Datasaurus dataset, which is amazingly insightful and artistic, but is built on the same idea that you just discovered. [Inspired from Udacity Data Vis with Tableau Course]

x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



More recently Alberto Cairo created the Datasaurus dataset, which is amazingly insightful and artistic, but is built on the same idea that you just discovered. [Inspired from Udacity Data Vis with Tableau Course]



Correlation

How change or variability in one variable is related to change or variability in another

Given two variables x and y .

standard deviation of x $\sigma_x = \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{n}$ and standard deviation of y $\sigma_y = \frac{\sqrt{\sum_i (y_i - \bar{y})^2}}{n}$

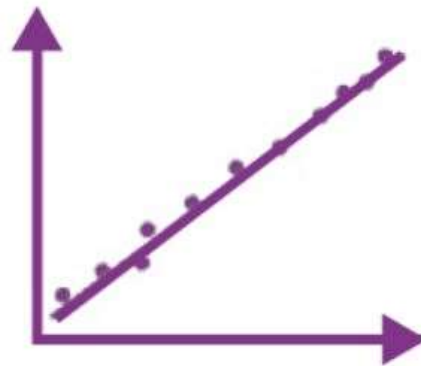
The covariance of x and y measure the joint variability of these two random variables, which is written as

$$\text{cov}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n^2}$$

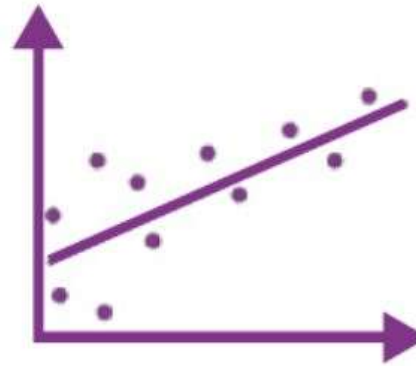
Now the correlation is the ratio of the covariance of x and y to the product of their individual standard deviations (hint - check the units). This correlation coefficient is also called pearson's coefficient.

$$\gamma_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

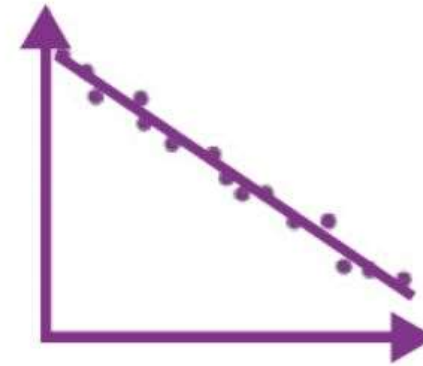
Correlation - Types



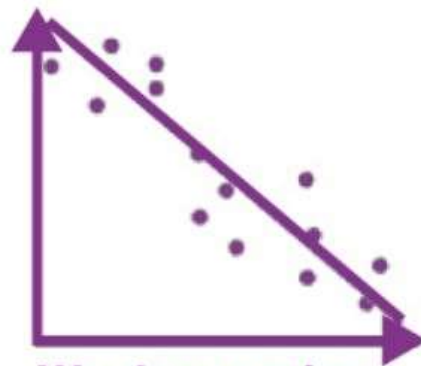
Strong positive correlation



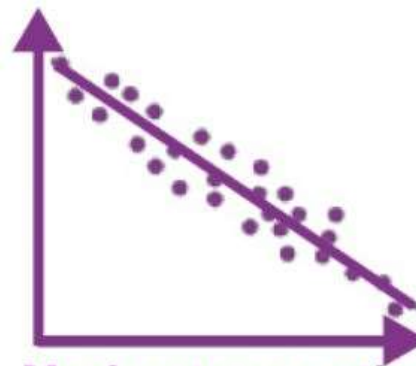
Weak positive correlation



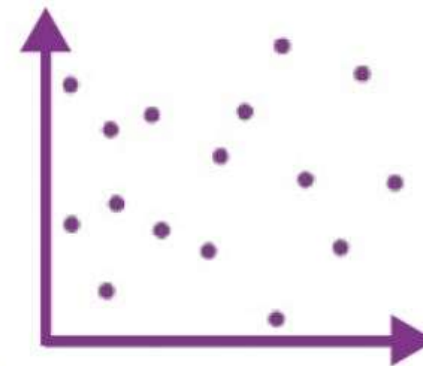
Strong negative correlation



Weak negative correlation



Moderate negative correlation



No correlation

Correlation – contd....

- Correlation coefficient is a value in the range $[-1, 1]$. WHY?
- **Why is correlation important?**
 - Because it helps us establish some sort of dependence or relationship between two random variables. In Machine Learning terms, if an input x is highly correlated with output y , then x is an important feature and can not be neglected. However if another feature z is not correlated with output y at all, we may be able to neglect it as changes in z have no impact on y . For example, your age may impact your income, but the color of your hair may not.
- **Line of Fit**
 - The correlation kind of gives some line of fit for the points. Look at the scatter plots in the figure above. In fact the pearson's coefficient is closely related to the slope of the line of fit.

Start become a Kagglers

- ✓ The [Kaggle](#) is an excellent resource for those who are beginners in data science and machine learning.
- ✓ you will find more details about kaggle [here](#)
 - ✓ <https://youtu.be/AoRSIdLpFqU>
- ✓ For additional information on Kaggle
 - ✓ <https://youtu.be/TPjcRbS9QeQ>
- ✓ **Courses**
 - ✓ Kaggle offers free courses to gain the skills you need to do independent data science projects.
 - ✓ Their courses pare down complex topics to their key practical components, so you gain usable skills in a few hours (instead of weeks or months).
 - ✓ The courses are free, and you can now earn certificates.

Data Visualization

Introduction

- Visualizing data in plots and figures exposes the underlying patterns in the data and provides insights.
- Good visualizations also help you communicate your data to others, and are useful to data analysts and other consumers of the data.
- Visualizing data using charts, graphs, and maps is one of the most impactful ways to communicate complex data.
- choosing the best visualization for your dataset, and how to interpret common plot types like histograms, scatter plots, line plots and bar plots is important for every data scientist.
- Build knowledge on best practices for using colors and shapes in your plots, and how to avoid common pitfalls.
- Matplotlib
 - a powerful Python data visualization library.
 - provides the building blocks to create rich visualizations of many different kinds of datasets.

A plot tells thousand words

Three ways of getting insights

- Calculating summary statistics
 - mean, median, standard deviation etc.
- Running models
 - linear and logistic regression
- Drawing plots
 - scatter, bar, histogram

Motivating visualization

Bitcoin price e

date	price_usd
<input type="text"/>	
2016-01-01	434.463
2016-01-02	433.586
2016-01-03	430.361
2016-01-04	433.493
2016-01-05	432.253
2016-01-06	429.464
2016-01-07	458.28
2016-01-08	453.37
2016-01-09	449.143
2016-01-10	448.964



Look at the Bitcoin prices on January the first each year. Which year began with the highest Bitcoin price?

Watch this excellent video by Hans Rosling at
<https://www.youtube.com/watch?v=jbkSRLYSojo>



Data Visualization

- In the previous discussion you have already gone through COVID Data and tried to get some insights.
- Data Visualization have an artistic side that makes it beautiful and compelling, at the same time it has a math/scientific side that helps deliver the right insights making the visualization both Engaging and Informative
- Data visualization requires some design and logical skills in order to:
 - Identify what plot to use for what data
 - What all data to visualize in a single plot
 - How to plot to make it engaging and informative.

What to plot?

- One can plot
 - A single variable (Univariate)
 - Two variables (Bivariate)
 - More than two variables (Multivariate)

Univariate Analysis

For quantitative data, if we are just looking at one column worth of data, we have four common visuals:

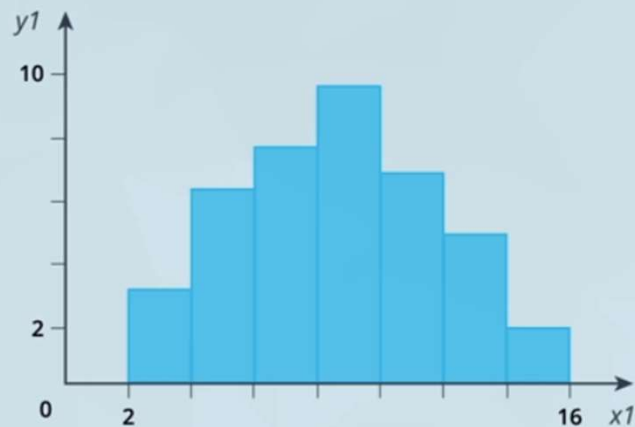
1. Histogram
2. Normal Quantile Plot
3. Stem and Leaf Plot
4. Box and Whisker Plot In most cases, you will want to use a **histogram**.

For categorical data, if we are looking at just one variable (column), we have three common visuals:

1. Bar Chart
2. Pie Chart
3. Pareto Chart In most cases, you will want to use a **bar chart**.

Avoid 3D plots as they only look good but are misleading

UNIVARIATE analysis also shows the Distribution of data and possible determine the outliers in Exploratory Analysis.



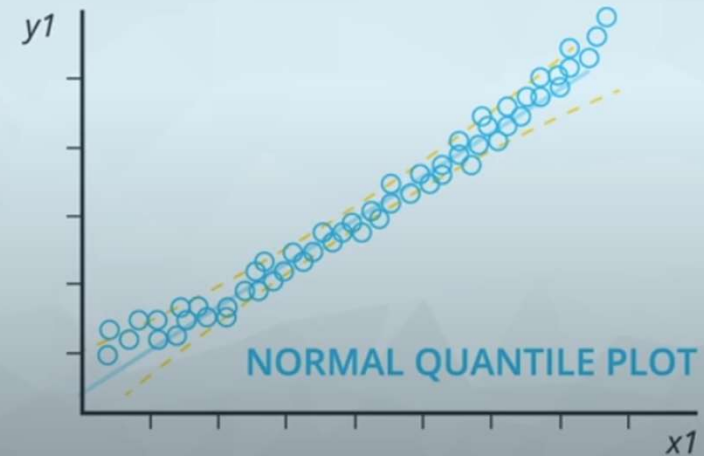
stem	leaf
1	1, 1, 2, 3, 4, 5, 5, 8
2	2, 2, 3, 3, 4, 8, 8, 9
3	2, 2, 2, 3, 4, 5, 7, 8, 8, 9
4	3, 4, 5, 6, 6, 7, 8
5	0, 1, 1, 1, 2, 3, 4, 9

1 | 1 = 11 years-old

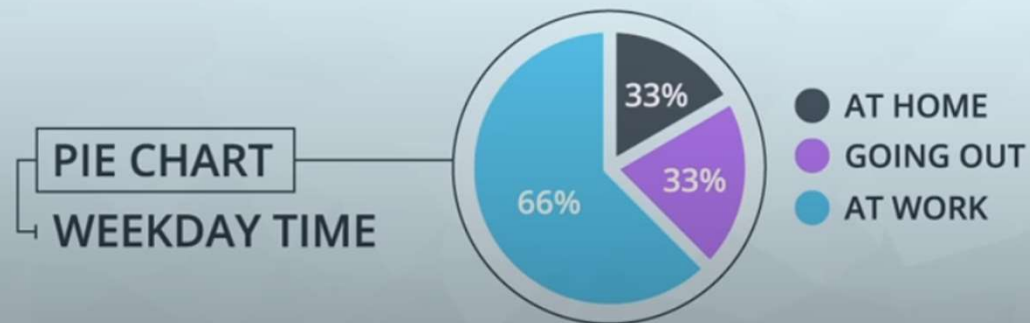
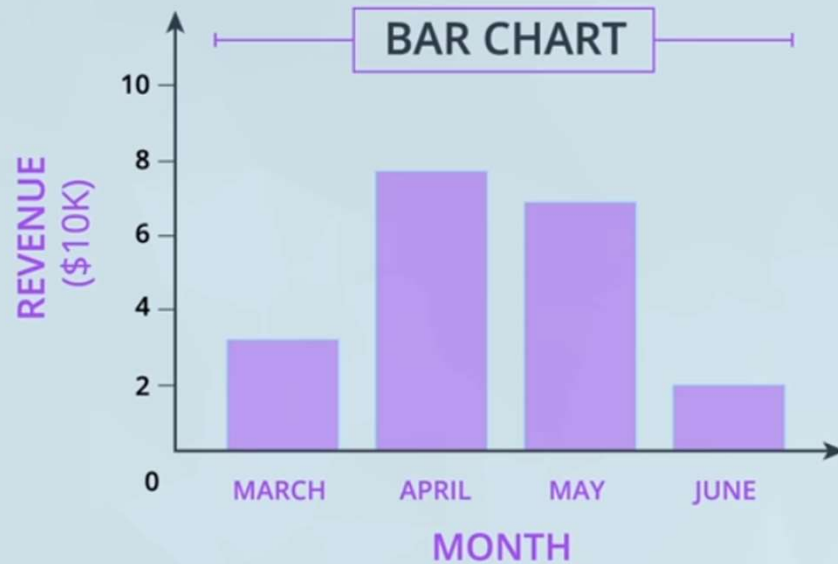
KEY

STEM-AND-LEAF PLOT

BOX-AND-WHISKER PLOT



Univariate Plots for Categorical Data



Histograms

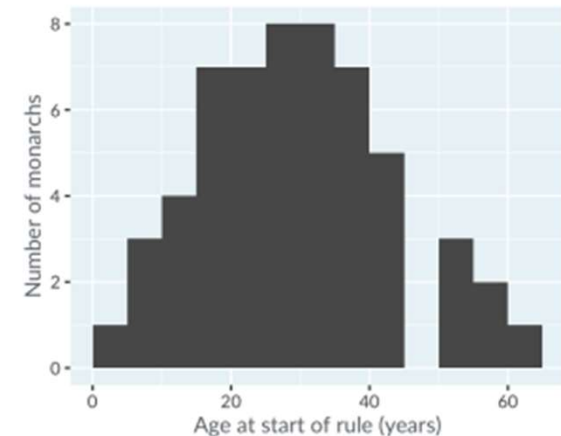
A histogram is a graph showing frequency distributions.

It is a graph showing the number of observations within each given interval

Kings and Queens of England & Britain

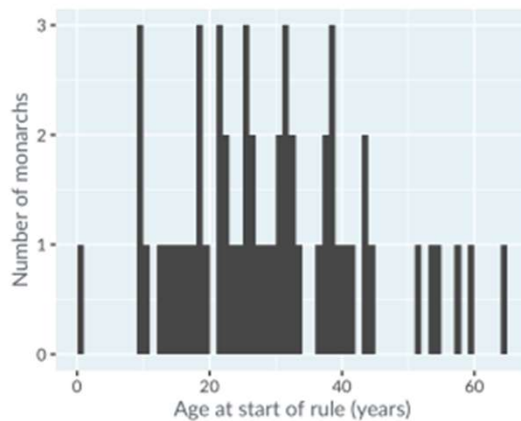
official_name	house	birth_date	start_of_rule	age_at_start_of_rule
Charles III	Windsor	1948-11-14	2022-09-08	73.86575
Elizabeth II	Windsor	1926-04-21	1952-02-06	25.79603
George VI	Windsor	1895-12-14	1936-12-11	40.99110
Edward VIII	Windsor	1894-06-23	1936-01-20	41.57426
...
Eadred	Wessex	0923-07-01	0946-05-26	22.90212
Edmund I	Wessex	0921-07-01	0939-10-27	18.32170
Aethelstan	Wessex	0894-07-01	0924-07-01	29.99863

Histogram of age at start of rule

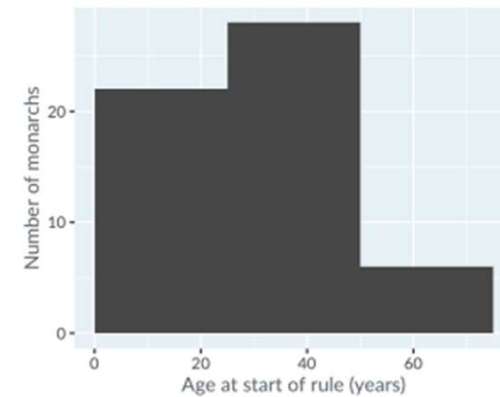


Histograms – Choosing bin width

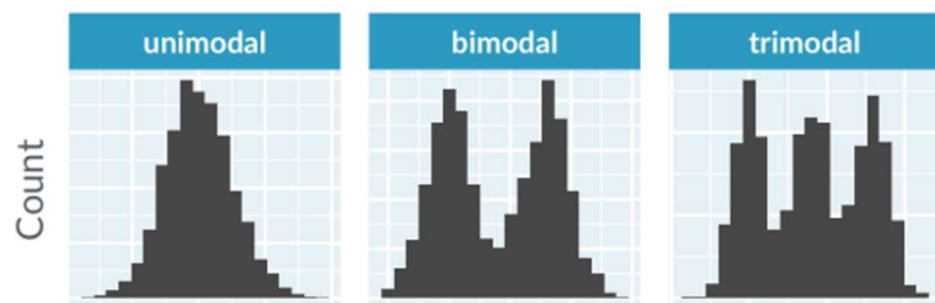
Choosing binwidth: 1 year



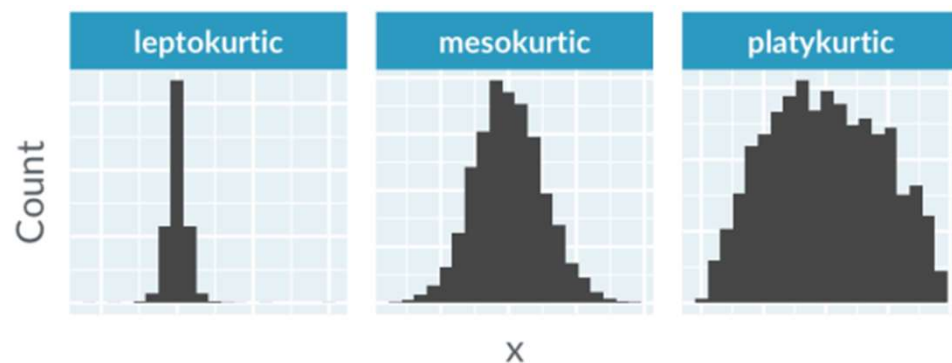
Choosing binwidth: 25 years



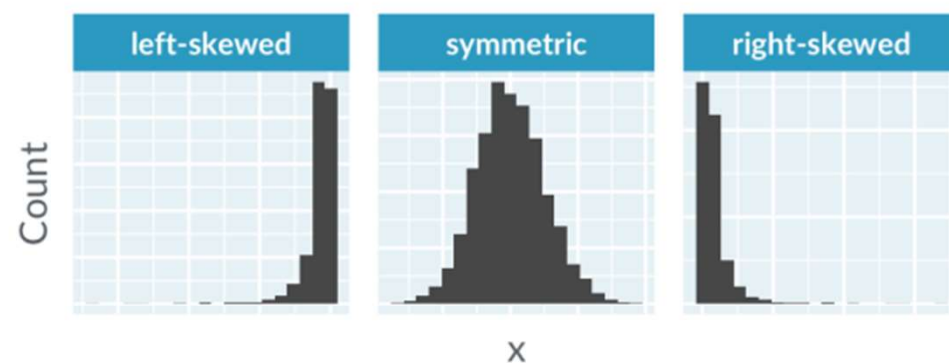
Modality: how many peaks?



Kurtosis: how many extreme values?



Skewness: is it symmetric?

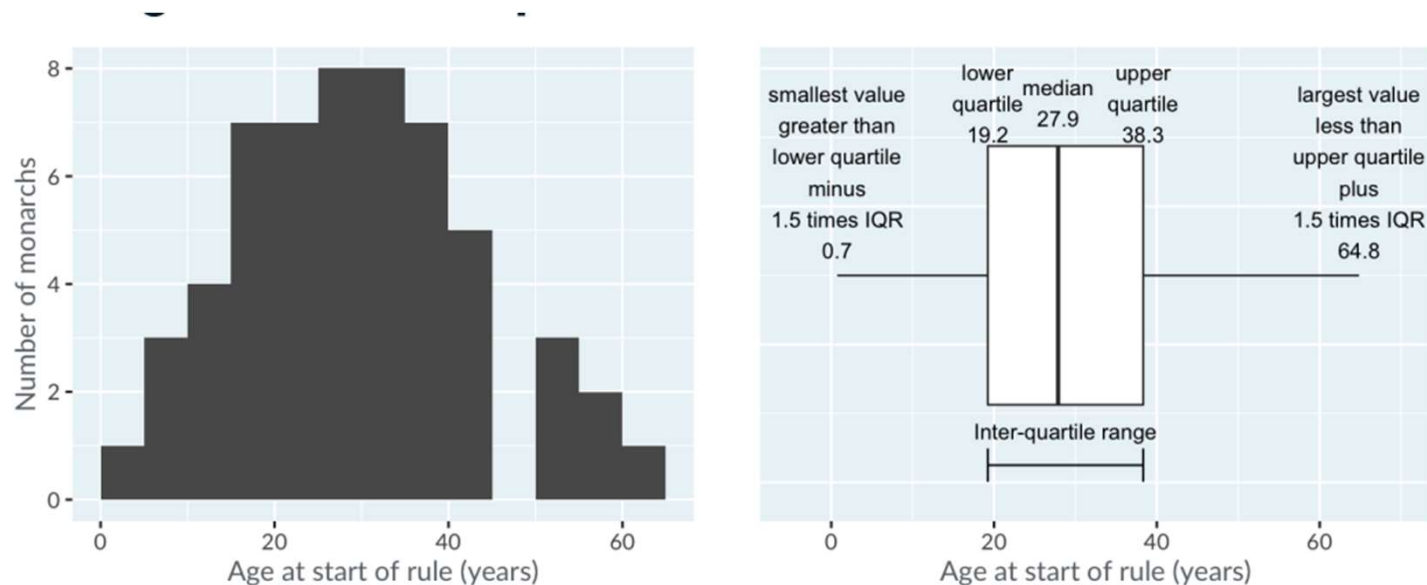


Box Plot

- A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum.
- In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.
- Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.
- Creating Box Plot
 - The matplotlib.pyplot module of matplotlib library provides boxplot() function with the help of which we can create box plots.

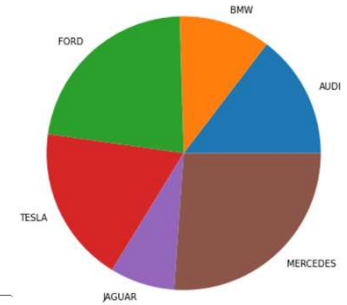
When should you use a box plot?

1. When you have a continuous variable, split by a categorical variable.
2. When you want to compare the distributions of the continuous variable for each category.



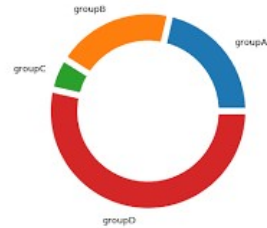
Histogram vs. Box Plot

PIE Chart



- A Pie Chart is a circular statistical plot that can display only one series of data.
- The area of the chart is the total percentage of the given data. The area of slices of the pie represents the percentage of the parts of the data.
- The slices of pie are called wedges. The area of the wedge is determined by the length of the arc of the wedge. The area of a wedge represents the relative percentage of that part with respect to whole data.
- Pie charts are commonly used in business presentations like sales, operations, survey results, resources, etc as they provide a quick summary.
- Creating Pie Chart
 - Matplotlib API has `pie()` function in its `pyplot` module which create a pie chart representing the data in an array.

Donut Chart



- Donut charts are the modified version of Pie Charts with the area of center cut out.
- A donut is more concerned about the use of area of arcs to represent the information in the most effective manner instead of Pie chart which is more focused on comparing the proportion area between the slices.
- Donut charts are more efficient in terms of space because the blank space inside the donut charts can be used to display some additional information about the donut chart.
- For being a Donut chart it must be necessarily a Pie chart.
- Creating a Simple Donut Chart
- Creating a Donut Chart involves three simple steps which are as follows :
 - Create a Pie Chart
 - Draw a circle of suitable dimensions.
 - Add circle at the Center of Pie chart

Bivariate Analysis

Quantitative Data

- ✓ Scatter plots are a common visual for comparing two quantitative variables. A common summary statistic that relates to a scatter plot is the correlation coefficient commonly denoted by r .
- ✓ Though there are a few different ways to measure correlation between two variables, the most common way is with Pearson's correlation coefficient. Pearson's correlation coefficient provides the:
 1. Strength - magnitude (absolute value)
 2. Direction - positive or negativeof a linear relationship.

Bivariate Analysis

Correlation coefficients

- Correlation coefficients provide a measure of the strength and direction of a linear relationship.
- We can tell the direction based on whether the correlation is positive or negative.
- A rule of thumb for judging the strength:

Strong $0.7 \leq |r| \leq 1.0$

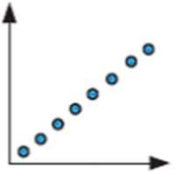
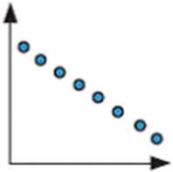
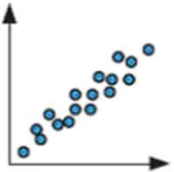
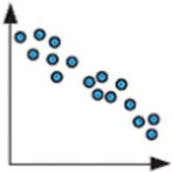
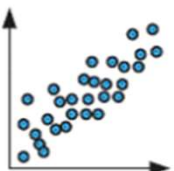
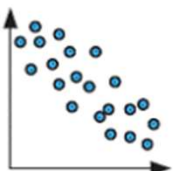
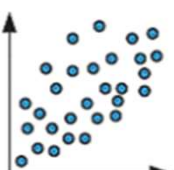
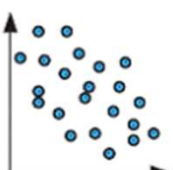
Moderate $0.3 \leq |r| < 0.7$

Weak $0.0 \leq |r| < 0.3$

- It can also be calculated in Excel and other spreadsheet applications using `CORREL(col1, col2)`, where col1 and col2 are the two columns you are looking to compare to one another.

Another interpretation is shown in below figure.

Relation between scatter plots, correlation and pearson's coefficient

r	Description	r	Description
1	perfect positive correlation 	-1	perfect negative correlation 
0.75 to 1	strong positive correlation 	-1 to -0.75	strong negative correlation 
0.50 to 0.75	moderate positive correlation 	-0.75 to -0.50	moderate negative correlation 
0.25 to 0.50	weak positive correlation 	-0.50 to -0.25	weak negative correlation 

Scatter Plots

With Pyplot, you can use the `scatter()` function to draw a scatter plot.

The `scatter()` function plots one dot for each observation. It needs two arrays of the same length, one for the values of the x-axis, and one for values on the y-axis:

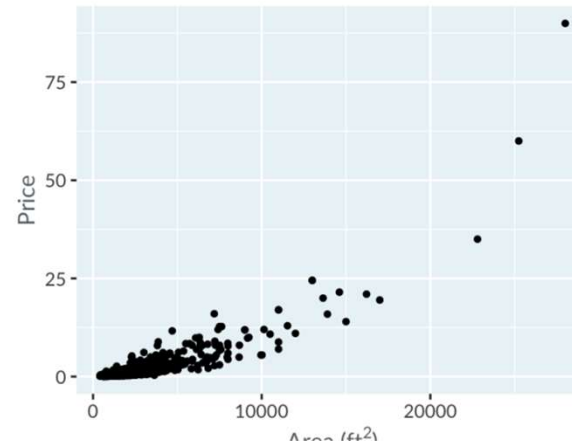
When should you use a scatter plot?

1. You have two continuous variables.
2. You want to answer questions about the relationship between the two variables.

Los Angeles County home prices

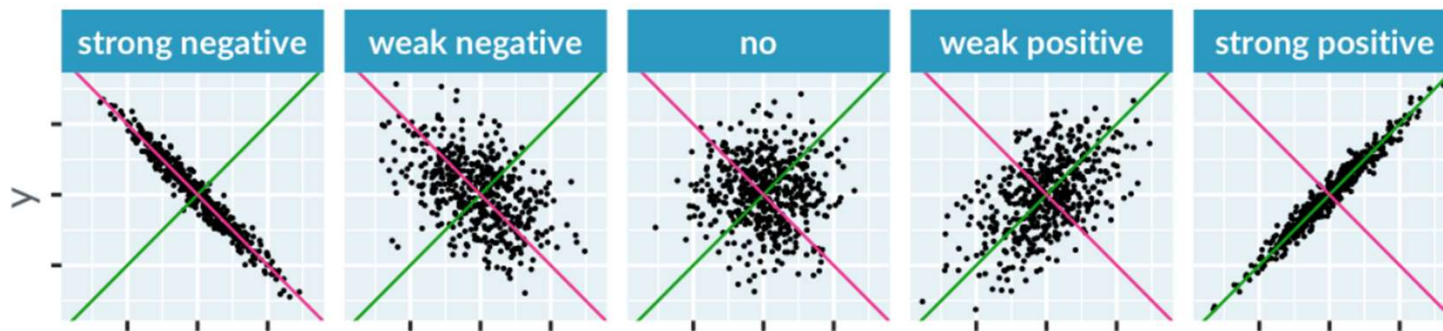
city	n_beds	price_musd	area_sqft
Long Beach	1	0.3250	846
Beverly Hills	3	2.1950	2930
Santa Monica	2	0.5740	1037
Santa Monica	1	0.5990	576
Beverly Hills	5	3.9500	5600
Long Beach	4	0.2999	1571
Westwood	3	0.6950	1913

Prices vs. area



Correlation

How close are you to being able to fit a straight line through the points?



Correlation – Exercise

In each of the following tell what type of correlation it is. There will be two answers for each

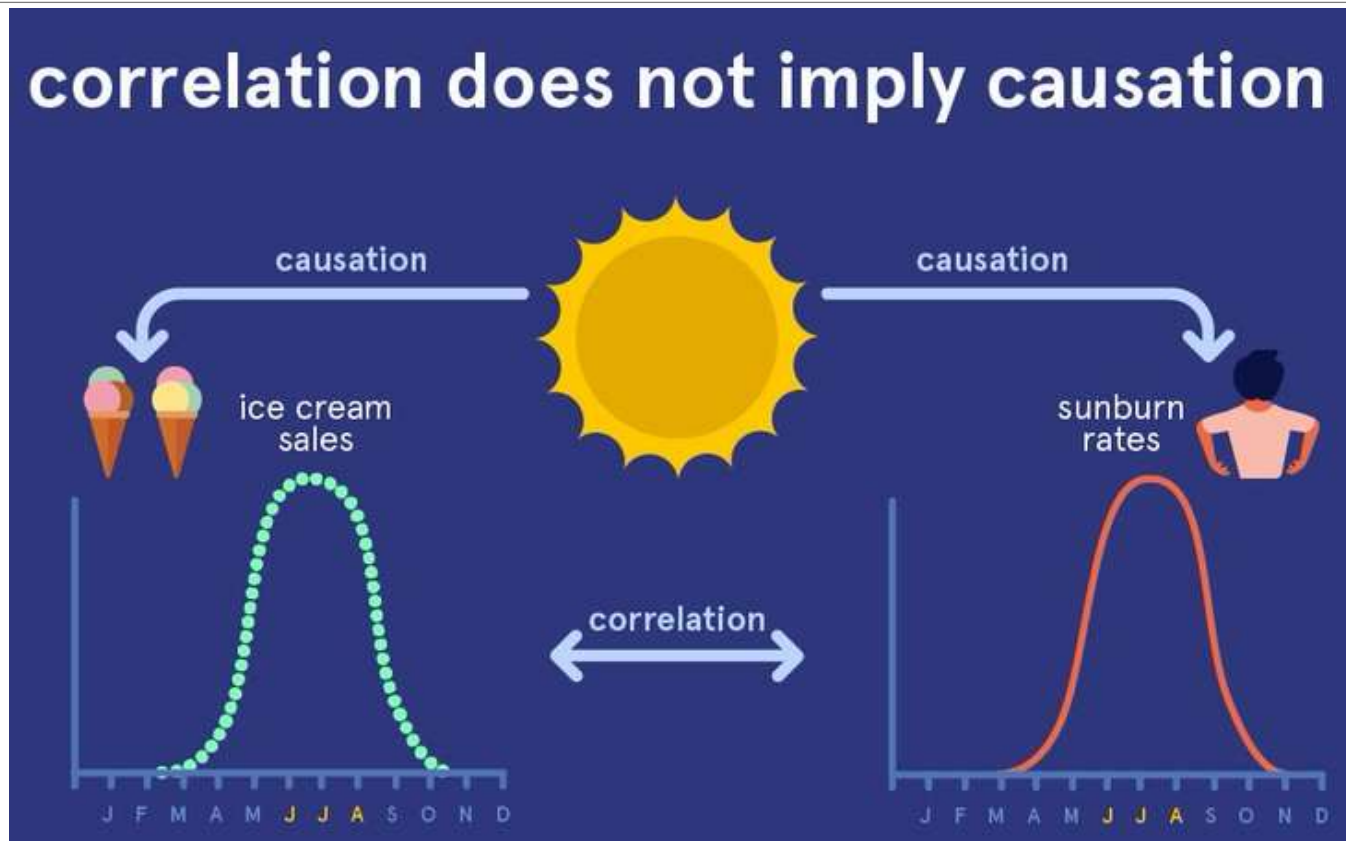
1. Whether there is no, weak, moderate, strong or perfect correlation _____

2. Whether it is positive or negative

- strength vs income – moderate to strong negative correlation
- age vs income
- age vs strength
- weight of car vs fuel efficiency
- number of siblings vs intelligence quotient
- hunger vs satisfaction after eating mess food
- weight of car vs price
- attendance in class vs marks obtained

Correlation does not mean Causation

For example in a hospital we will find more sick people. So person in hospital is correlated to person being sick. But it does not mean that hospital causes you to be sick.



Bar Chart

A barplot (or barchart) is one of the most common types of graphic. It shows the relationship between a numeric and a categoric variable

With Pyplot, you can use the `bar()` function to draw bar graphs:

The `bar()` function takes arguments that describes the layout of the bars.

The categories and their values represented by the first and second argument as arrays.

When should you use a bar plot?

Most common cases:

1. You have a categorical variable.
2. You want counts or percentages for each category.

Occasionally:

1. You want another numeric score for each category, and need to include zero in the plot.

Stacked Bar Chart

A Stacked Percentage Bar Chart is a simple bar chart in the stacked form with a percentage of each subgroup in a group.

Stacked bar plots represent different groups on the top of one another. The height of the bar depends on the resulting height of the combination of the results of the groups.

It goes from the bottom to the value instead of going from zero to value. A percent stacked bar chart is almost the same as a stacked bar chart.

Subgroups are displayed on top of each other, but data are normalized to make in a sort that the sum of every subgroup is the same as the total for each one.

Relative Stacked Bar Chart

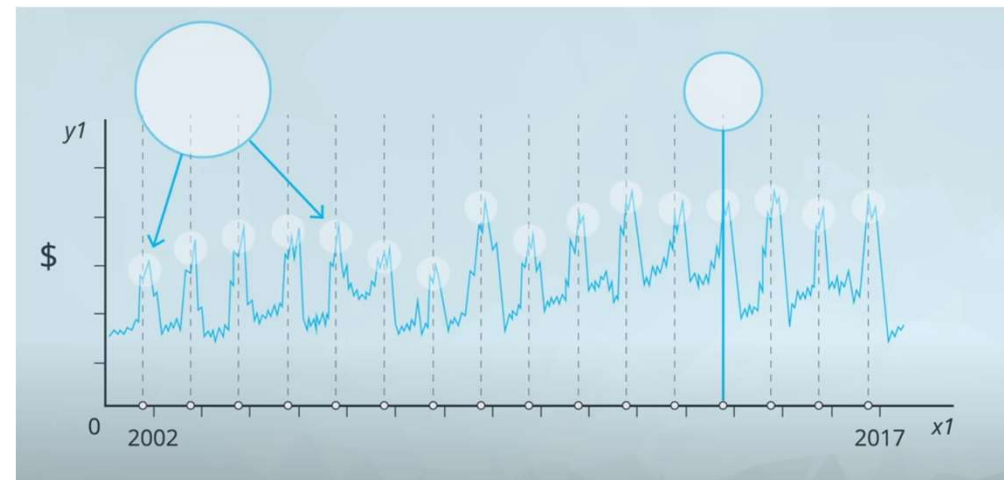
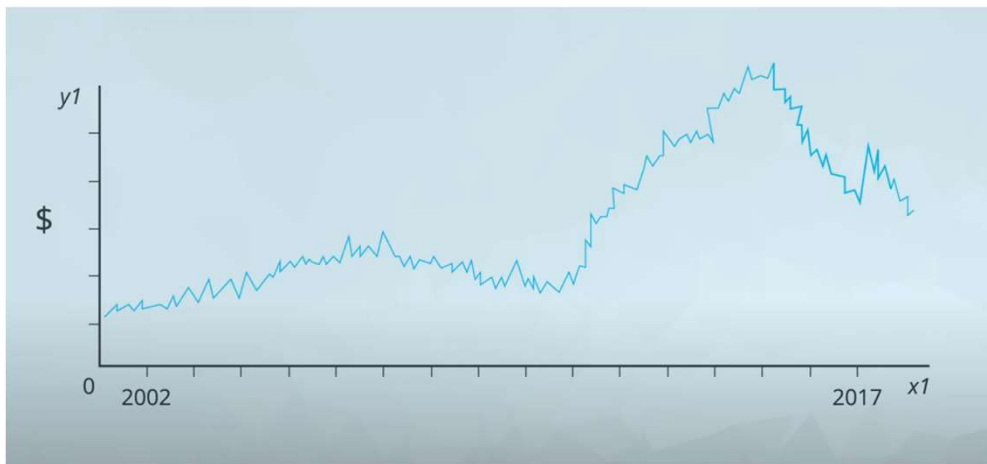
Children's fruit and veg consumption

n_portions	year	pct_children
n = 0	2001	10.921779
n < 1	2001	3.843093
1 <= n < 2	2001	23.659102
...
4 <= n < 5	2018	12.28728
5 <= n	2018	17.87497



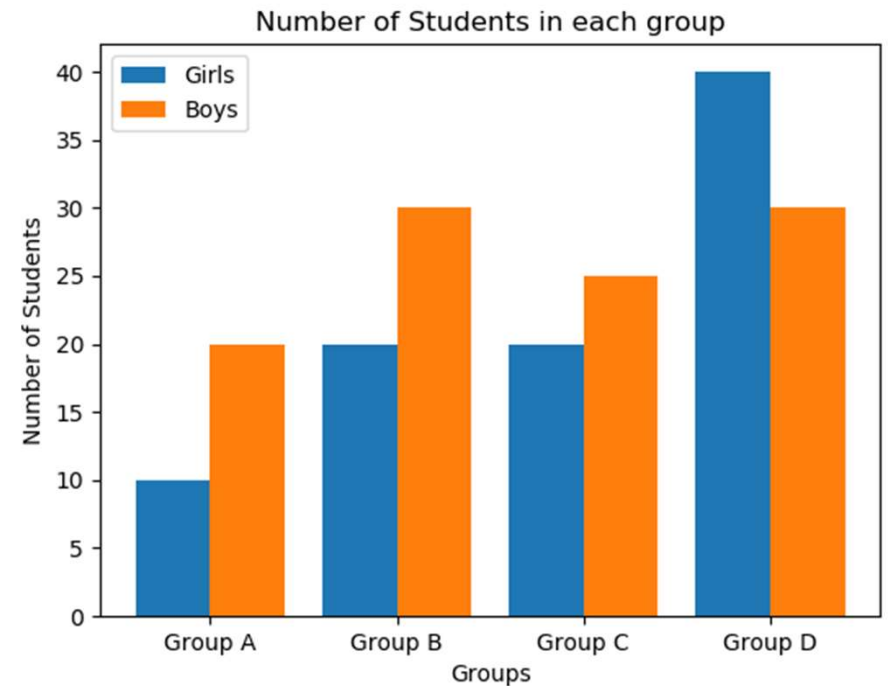
Line Plots

Line plots are a common plot for viewing data over time. These plots allow us to quickly identify overall trends, seasonal occurrences, peaks, and valleys in the data. You will commonly see these used in looking at stock prices over time, but really tracking anything over time can be easily viewed using these plots.



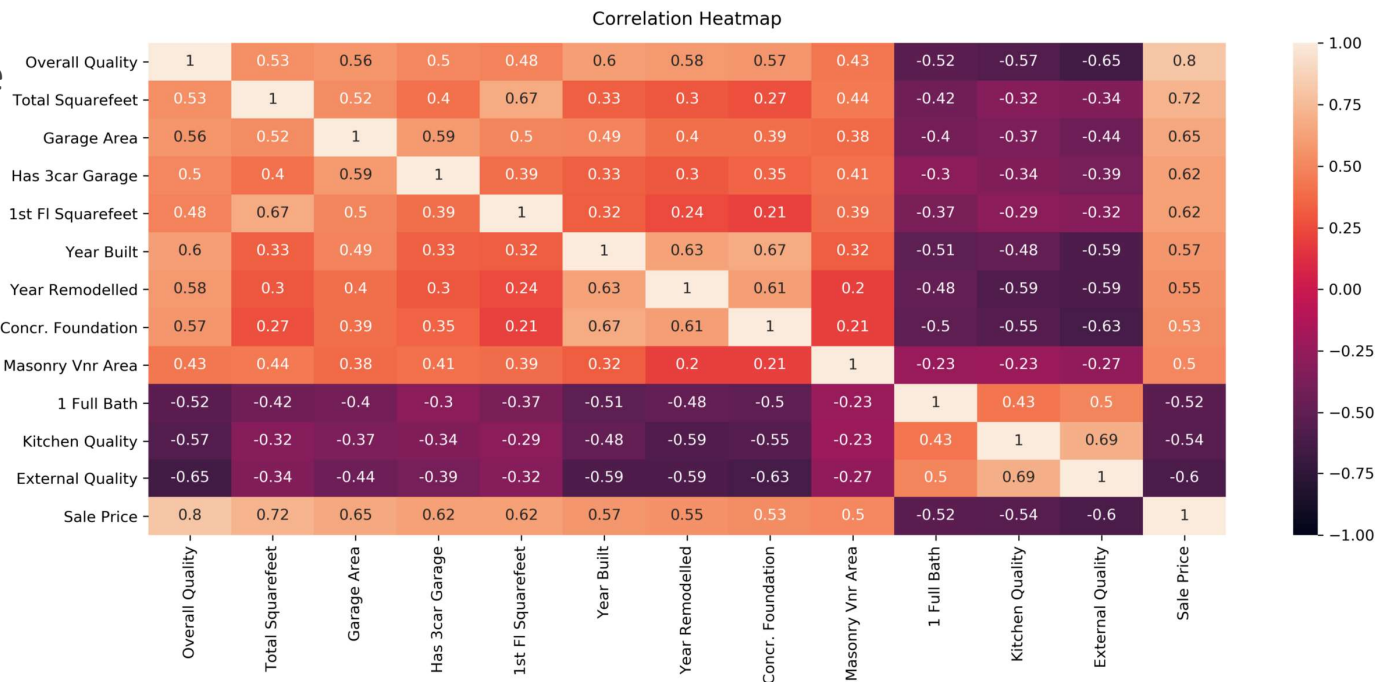
Multivariate Analysis

- In general we avoid 3D plots as they are difficult to interpret. However we can add more categorical variables to our plot by adding more elements like shapes, positions or colors to our plots.
- Adding Categorical Data on an axis demands other elements to be added



Heatmaps

Suppose we give a color hue to the person's coefficient's absolute value from 0 to 1. If data has 10 features, we can plot a 10x10 matrix where each rectangle is colored based upon the pearson's coeff value between corresponding variables. These are called Correlation Heatmaps. You also see heatmaps on github that indicate over time on which week or month you had made contributions etc.



Exercise

What chart will you use for visualizing the following?

- Want to determine how two variables relate to one another
- Want to show how a quantity is distributed in various parts
- Want to illustrate total scores of teams in Olympics
- Find the distribution of data values
- Show trends of certain value(s) over time
- Show change of a value over time
- Determine in one shot how all the variables relate to one another in a data set.

Six lessons of communicating with data

1



Understand the context

4



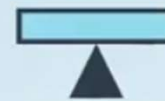
Focus attention where you want

2



Choose an appropriate
visual display

5



Think like a designer

3



Eliminate clutter

6



Tell a story

Exploratory Vs Explanatory Analysis

There are two main reasons for creating visuals using data:

Exploratory analysis is done when you are searching for insights. These visualizations don't need to be perfect. You are using plots to find insights, but they don't need to be aesthetically appealing. You are the consumer of these plots, and you need to be able to find the answer to your questions from these plots.

Explanatory analysis is done when you are providing your results for others. These visualizations need to provide you the emphasis necessary to convey your message. They should be accurate, insightful, and visually appealing.

The five steps of the data analysis process



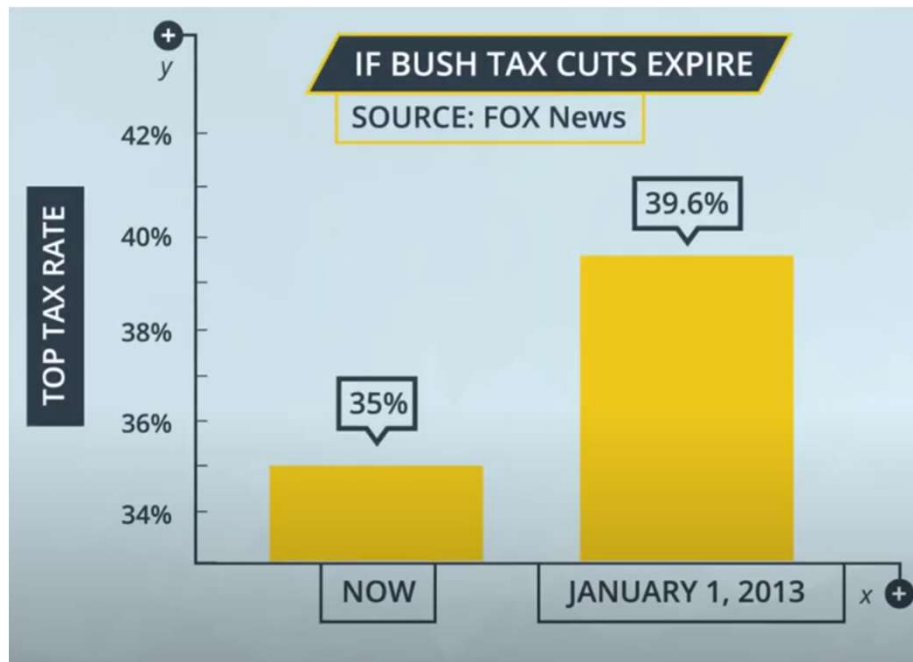
1. **Extract** - Obtain the data from a spreadsheet, SQL, the web, etc.
2. **Clean** - Here we could use exploratory visuals.
3. **Explore** - Here we use exploratory visuals.
4. **Analyze** - Here we might use either exploratory or explanatory visuals.
5. **Share** - Here is where explanatory visuals live.

Bad Data Visualization

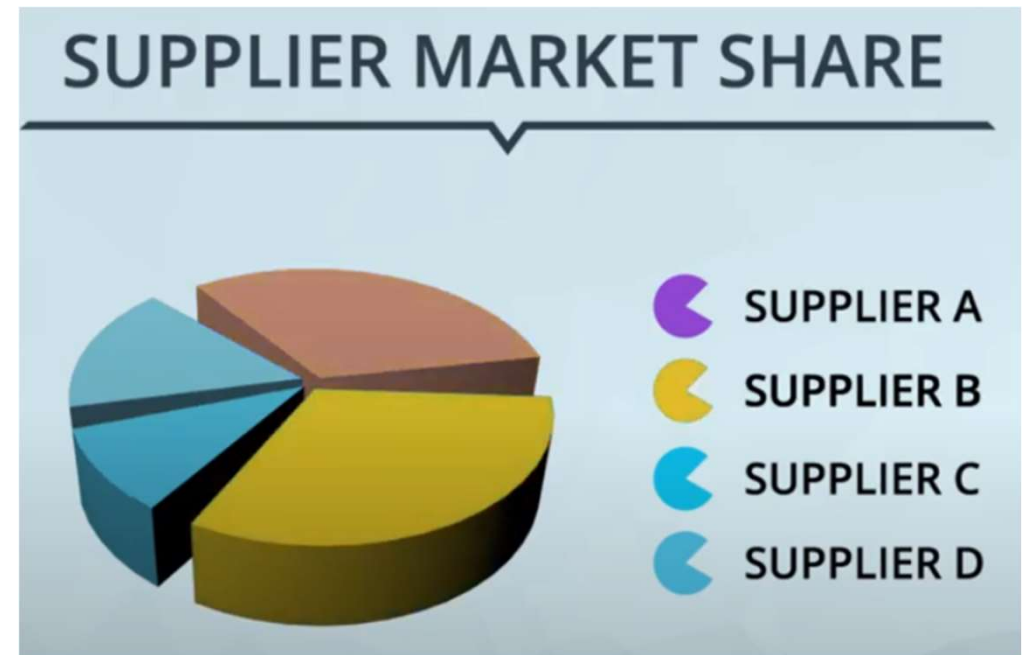
Visuals can be bad if they:

- Don't convey the message.
- Are misleading (hiding, distracting or biasing)

This seems straightforward, but often visuals are created that do one or both of these unintentionally.

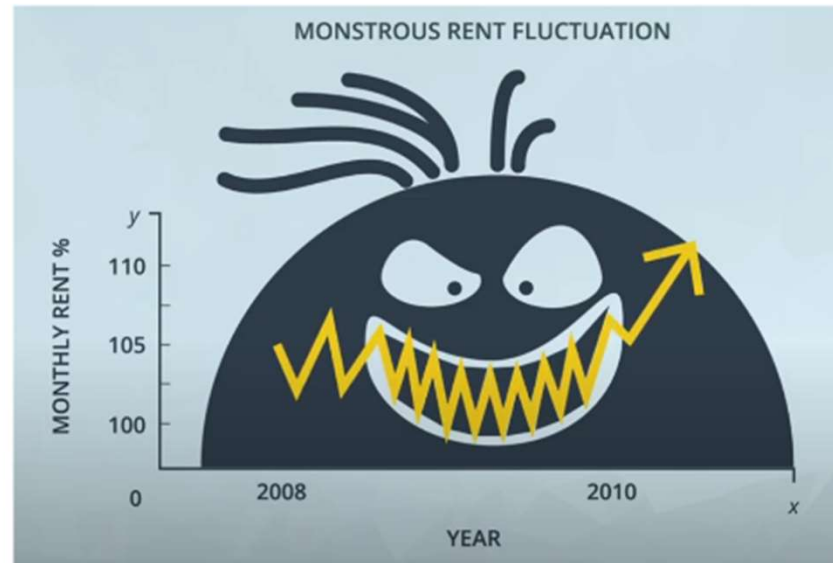
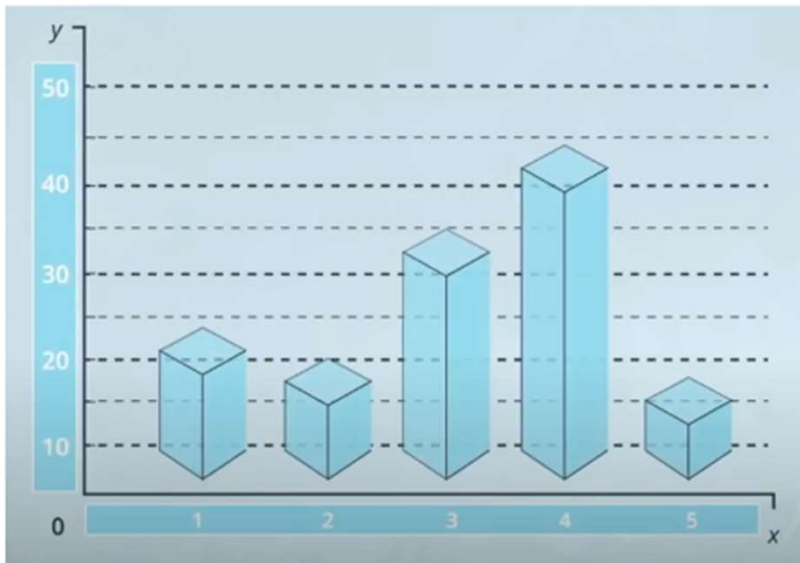


Misleads by suggesting too big a change

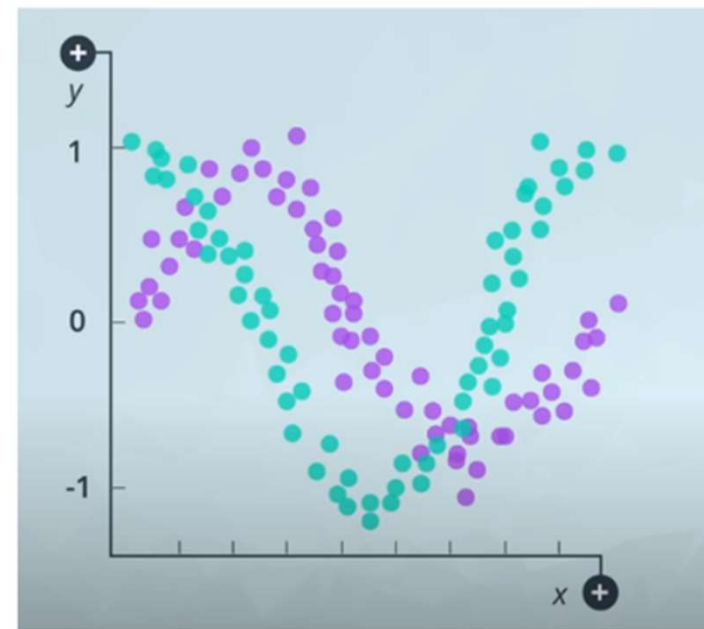
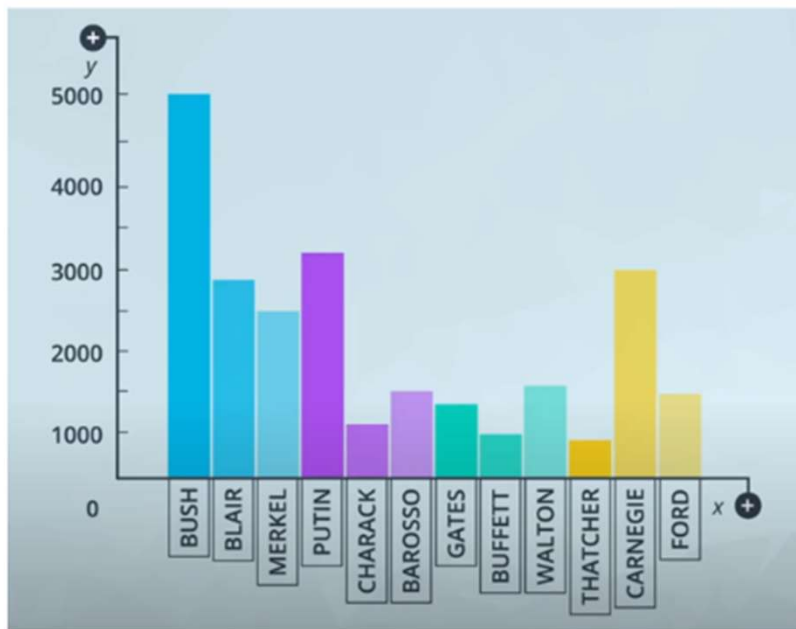


3D is misleading

Bad Visualization



Exercise: In the plots below identify which plot uses color effectively and why?

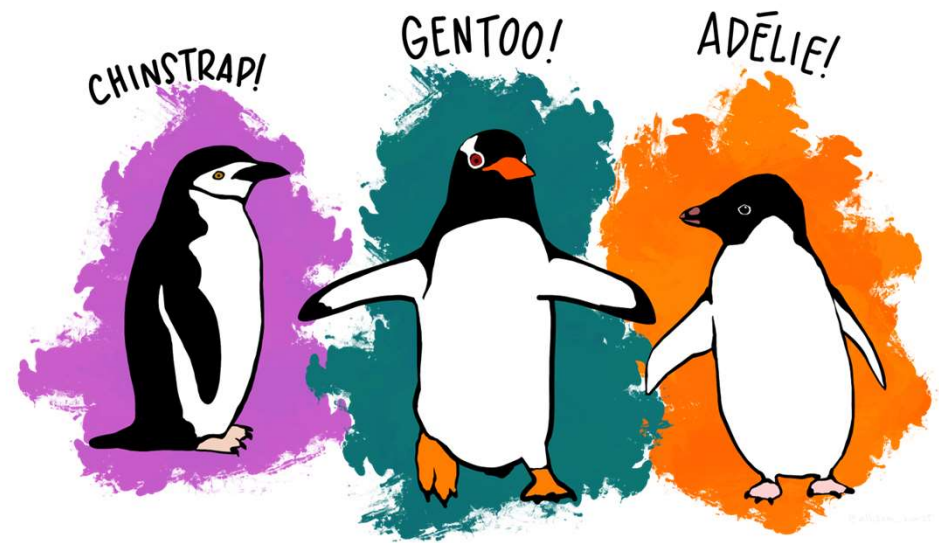
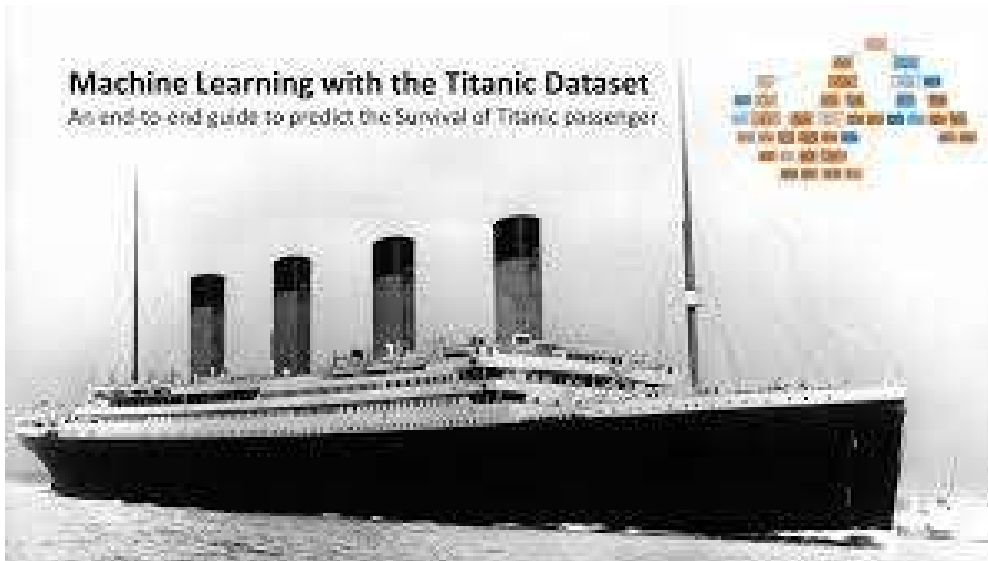


Do Kaggle Data Visualization Course

<https://www.kaggle.com/learn/data-visualization>

A solid orange horizontal bar spanning the width of the slide at the bottom.

Try to download **titanic** and **penguin** datasets and perform some visualizations to understand the patterns in the dataset.



Plotting of Covid Data

Explore the following resources and plot some graphs on covid data

<https://www.geeksforgeeks.org/covid-19-data-visualization-using-matplotlib-in-python/>

<https://towardsdatascience.com/covid-19-data-visualization-using-python-3c8bcfaeff5f>

Self Learn

Plotting all the above in seaborn library

Plotting heapmaps

Plotting wordclouds etc.