Variables: are called variables because they vary between individuals. You classify them according to their type:

		Categorical[1]			**Numerical (measured on a number scale)[2]**	
		{Dummy variables}	{Polytomous variables: more than two categories}			
		Binary/Dichotomus	**Nominal**	**Ordinal[3]**	**Interval / Discrete**	**Ratio/ Continuous**
	Categorical or Numerical? *(Are values divided into distinctive categories or are there distinct scores per person that can take any value in the scale we are using)*	Categorical in simplest form. Only two categories	Categorical. Numbers are only to differentiate between categories	Categorical. Numbers are only to differentiate between categories, but the categories have hierarchy	Also known as discrete where there is a distinct number of values. Technically categorical but usually treated as continuous**	Continuous**, variable can have any numerical value.
		Yes/No (binary category)	*Country of birth, gender, hair colour, voted in last election, ethnicity.*	*Likert scale (agree, like, etc), Freedom House Rankings (free, partially free; pass, fail/ exceed), scales measuring subjective feelings like happiness. Wong-Baker Faces Pain Rating Scale*	*Year of birth, years in full-time education, temperature (Celsius), scores in standardized tests, results in a psychological inventory. Number of siblings*	*Anything you count: number of hours worked, income, DGP per capital, %, weight, height, population, density, temperature (Kelvin)*
			Do you have a university degree? Identify your ethnicity	*What is your highest level of schooling? (primary, secondary, some higher, diploma)*	*What year of school are you in? UG1, UG2, UG 3, UG4, PG 1, PG2*	*How many years have you spent in higher education? 7 years, 9 year, 0 years*
	Sequence, Order of variables *(gaps in the categories are not constant)*		NO	YES	YES	YES
				Strongly agree is more than just agree. Once a week v/s once a month	*Born in Jan 1981 is older than Jan 1982*	*Salary is 25k for one and 35.5k for another*
	Uniform Distance? (can calculate the		NO	NO	YES	YES
					Distance between Jan 1981 and Jan 1982 is one year	*Numerical distance*

[1] Also known as: Qualitative variables (because they describe attributed or qualities), Attribute variables (describe characteristics that classify data into distinct, non-numeric groups) or Factors (such as when using R), Labeled data (Refers to the categories that act as labels for grouping), Non-metric data (variables not measured on a numeric scale)

[2] Also known as: Quantitative variables, measurement variables (because continuous data is generally measured rather than counted), metric data (referring to numerical data that can be measured) or Field/Surface data (for specific software like ArcGIS)

3 Many times self-report data is ordinal data, which has limitations, even if researchers want to treat it as if it's not.

	difference between variables)					
	Add and Subtract Variables?		NO	NO	YES	YES
					1982 is one year younger than 1981	
	Zero is meaningful? (Number can take on full mathematical properties)		NO	NO	NO	YES
	Multiply and Divide Variables?			NO[4]	NO[5]	YES
		Univariate Analysis				
Central Tendency	**Mode (common) incl. bimodal and multimodal distribution**		YES *e.g. most participants are religious*	YES	YES	YES
	Median (middle)			YES[6] *e.g. median price for a 2 bed flat is 200k*	YES	YES
	Mean (average)			NO	YES* e.*g. the average score in the test was 78 points*	YES*

[4] It is possible to calculate an average of multiple ordinal measures to create a continuous/scale variable (calculating the sum or average of multiple ordinal measures). This will refer to the average od ordinal variables rather than categories themselves.

[5] Not possible because there is no absolute zero (eg. 40 Celsius is not twice as hot as 20 Celsius). Relatedly, comparing across different scales needs to be done with caution, as the size of the intervals is the same within the scale but not across the scale)

[6] In an odd-numbered data set, the median is the value at the middle of your data set when it is ranked. In an even-numbered data set, the median is the mean of the two values at the middle of your data set. If you have two categories in the middle (e.g. agree and strongly agree) you can't find the median, even if you codded them numerically because addition and subtraction are not possible with these variables.

* Attention with outliers as they can distort the overall picture.

** Can have discrete or continuous values. Discrete refers to something you can count cand can't subdivide (e.g. number of cats, can't have ½ a cat) but a group of discrete values can have meaning in fraction or decimal, Variable refers to something you can measure and can have any value, with meaningful fraction and decimal values.

Variability (measures of dispersion)	**Absolute Measures (Range, Interquartile Range, Variance, Mean Absolute Deviation, Standard Deviation)**		Number of categories with at least one response, percent distribution	Min and Max, Range, IQR, percentiles, percent distribution	Min and Max, Range, IQR, percentiles, percent distribution, Standard Deviation (description including skewness and kurtosis)	Min and Max, Range, IQR, percentiles, percent distribution, Standard Deviation, description including skewness and kurtosis
	Relative Measures (Coefficient of Variation, Coefficient of Quartile Deviation, Coefficient of Mean Deviation)					Relative Standard Deviation or Coefficient of Variation
	Tables	Frequency Tables (number and percent)	Frequency Tables (number and percent)	Frequency Tables (number and percent)	Frequency table of ranges (groups[♦])	Frequency table of ranges (groups[♦])
	Visualization	Usually you only need the text as a visual representation won't enhance the message or understanding of the results.	Bar chart, pie chart	Bar chart, respecting the order of the categories	Histograms, Plot frequency polygon, with groups (X) against frequencies, box plots (when embedded on histogram, or for categorical v/s continuous). Scatterplot (2 continuous variables)	
	Tests			Non-parametric		

Changing the form of the data (and losing detail as you go from continuous to binary, because the statistical analysis of continuous data is more powerful and often simpler):

Original: A baby's weight at birth in kilograms, grams and milligrams [Continuous numeric]

♦ Groups, only when the variable has been categorized. Attention when you categorize not using the same proportion between categories.

♦ Groups, only when the variable has been categorized. Attention when you categorize not using the same proportion between categories.

[Discrete numeric: Birthweight to the nearest 10 grams]
[Ordinal: Birthweight as category between less than 1 kg, 1kg to 2kg, 2 to 3kg, 3kg or more]
[Binary: Low birthweight, Normal birthweight]

Displaying Data:
Consider:

- Number and type of variable you need to present
- The number of observations you have in the sample
- The message you are trying to convey

Variables		
1 Variable categorical (Categorical data or numeric data that has been categorized)	Frequency Table: Displays the number-frequency/percentage (in relation to sample size) of participants within a certain category	
	Bar Chart: Displays the number-frequency/percentage (in relation to sample size) of participants within a certain category. Length of bar represents the frequency or percent in the category. Bars can be ordered and frequencies can be added to strengthen the message. Helps visualize differences better than frequency tables	
	Pie Chart: each slice represents the proportion of the sample within a given category. It can only be used for NOMINAL variables because you can't display hierarchy.	
1 Variable numerical	Histogram, Dot Plot	
	Histogram: an alternative to bar charts. The area of the bar represents the proportion of the sample lying within the range covered. Used for investigating the distribution/shape of a numeric value. Only issue is that you are somewhat categorizing the numeric value and therefore losing some of its detail	
	Dot Plot: each plot represents an observation. It shows each participant, therefore is useful for identifying outliers or	

	spotting patterns in the sample. If you don't use transparency (or use jitter) you miss information if the points overlap.	
Paired data: Any pairing -same individuals- or matching needs to be retained during visualization (and analysis) of the data. If the same person is measured twice, or there are distinct pairs, it needs to be clear in the visualization	Scatterplots 2.0 or line diagram. Scatterplot needs to have a line of equality to emphasize the paired nature of the data. Line diagram: each line represents an individual and gives a stronger message than separate points in a scatterplot, but when you have a large sample, it can get messy.	
2 Categorical	Frequency Table (or Contingency Table) are expanded to include more than one categorical variable. Ordinal and binary variables can be included in the same table. You can add frequency and percentage. Side by side Bar Chart. Charts are added to account for another categorical variable. Stacked bar chart: Here the second category is added on top of the bar instead of side by side. Each bar shows the frequency per group but split by another grouping variable to show multiple relationships within the same graph.	
2 Continuous	Scatter plot: Show their relationship. Each point represents an individual. It can help identify associations and outliers. Helpful for checking assumptions for certain methods (eg)	
1 Categorical and 1 Continuous	Bar chart, box plot or dot plot Dot plot can display numerical data grouped by categories that are exclusive (you belong to just one category). It can help show the differences between groups and highlight outliers.	Test for significance (t test or anova)
	Pictures and diagrams: can help convey the message in a clearer way, for example map of a region with color variations to depict differences between areas within the region. Helpful for location for example (like frequency of pain by parts of the body where it hurts)	

More than 2 variables (challenging, ideally the addition of more variables should not add confusion to the message)	Scatterplot, Contingency Tables and Bar Charts. Scatterplot: two show the association between two continuous variables in more than one group. You would use different symbols and colour to distinguish between groups. Contingency table that is extended to display 3 or more categories. Here you need to be careful with the percentages, so that it's clear what you mean by them. Bar chart: Stacked and side by side are combined to show 3 categorical variables. You should only use it if it's helpful for the reader.	

Variable Type	Significance Test	Strength Test
Two categorical	Chi^2 Result: P-value	Cramer's V Result: between 0 and 1 0.3 weak 0.5 moderate 0.6 strong
Two Continuous	Pearson's R Result: P-value	Pearson's R Result: Coefficient Results between -1 to 1 -0.3 and 0.3 weak -0.3 to -0.5 AND 0.3 to 0.5 moderate -0.6 to -1 AND 0.6 and 1 Strong
One Continuous and One Categorical	T-Test and Anova	None, just look at the difference in the means

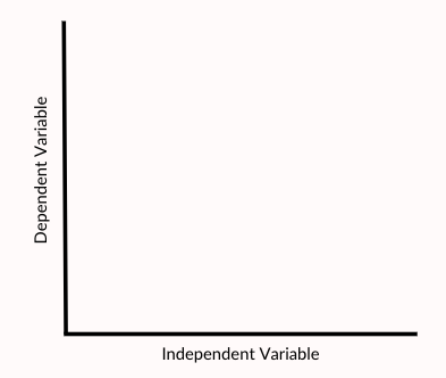
Parametric v/s Non-parametric

Parametric: Analysis us based on a fitted distribution, such as Normal distribution
Preferred when the distribution adequately fits the data and there are no outliers
Have more power and need fewer cases to find significance.
They can estimate the value of data where there are no data (interpolate) and extrapolate beyond the range of the data

Non-parametric: Analysis does not assume a distribution for the data, the analysis is typically based on ranks
Used when assumptions based on the distribution can't be met, or when the outliers can't be removed.
Used when the median is chosen as the measure of the centre of the distribution instead of the mean

Comparison	**Parametric**	**Non-parametric**
Mean/Median to a target value	One sample T-Test	Wilcoxon Signed-Rank Test
Two mean/median from independent samples	Unpaired T-Test	Mann-Whitney Test
Two mean/median from paired samples	Paired T-Test	Wilcoxon Matched-Pairs Test
More than two mean/medians from independent samples	One-way ANOVA	Kruskal-Wallis Test
More than two mean/medians from repeated measure samples	One-way ANOVA with RM	Friedman Test
Two factor DOE	Two-way ANOVA	-
Three factor DOE	Three-way ANOVA	-
Relationship between X and Y	Pearson Correlation	Spearman Correlation
Predict Y as a function of X	Linear and Non-Linear Regression	-

VISUALIZATION



How many variables?	What type of variable?	Appropriate visualization
One	Continuous	Histogram
		Boxplot
	Categorical	Pie Chart

		Bar/Column Chart
Two	Continuous / Continuous	Scatter Plot
	Continuous / Categorical	(Clustered) Bar Chart – could have box plot with error bars and see if they confidence intervals coincide at some point
		Multiple Boxplots
		Multiple Histograms
	Categorical / Categorical	Heat Map
Three or more	All Continuous	Bubble Plot
		Heat Map
		Spider Graphs
		Parallel Coordinates Plot
		Scatter Plot Matrix
	Two Continuous and one categorical	Scatter Plot with multiple series
	Two Continuous and multiple categorical	Multiple Boxplots

Pie Charts:

- Use to compare parts to whole
- All percentages need to sum to 100%
- Max. 6 categories (avoid if some are too small)
- No negative values

Error Bar Charts:

- Mean score
- The error bar sticks out from the section (whisker)
- The error bars can display
    - Confidence interval
    - Standard deviation
    - Standard error of the mean

Attention: Software with automatically generated charts will pick scales to emphasise differences. Scales are important.

Histograms (or frequency distributions):

- They help visualize the general distribution or the shape of the data
- You can also create density distributions the same way (has to do with skew?)
- In Stata you can combine them with density distribution to see if your data is normally distributed or not
- Is difficult to understand them if you have different categories, they work best for just one variable
- No gaps between bars -continuous data-(you can specify the beam size when visualizing the data) the intervals should be equal size

Box Plots

- They help visualize the distribution or the shape of the data (symmetrical or skewed), the outliers and the quintiles

- They show 5 main parts of the data, the IQR (2 and 3 quartile rectangle, if long, it means that there is significant variation, if one side is longer than the other, the distribution is skewed) with a line in the middle that represents the median and the first ¼ are out, like whiskers.
- Outliers are represented with circles and the extreme values are represented with stars.
- Not all boxplots have outliers
- You can find out more about the data than with histograms
- They can also be helpful with categorical variables (side by side boxplots)

Scatterplots

- Useful for 2 continuous variables
- Allow us to visualize the relationship between variables
- Y-axis is used for the dependant (y) variable, the X axis is used for the independent variable
- You look at the patter of the distribution of the data
- You can have a grouped scatter plot, where you can examine two independent variables (like make and female)

Scatterplot matrix

- (like a correlation matrix but with visualization, you have different combinations.
- Used to visualize bivariate relationships between combinations of variables
- It can hep you see which variables you should be including in your analysis

Line Graphs

- Useful to show trends over time.
- It could be used with things that are not time, but, to connect with a line things that don't connect is misleading. In that case is better to use a bar chart.
- Can get confusing if there are too many categories in gone chart
- Suggestion: hide grids and reducing the scale y-axis will let you see trends more easily.

**Deciding the scale to use for the plot should be theoretically driven.

Heat Maps:

- Useful for two categorical variables
- Is like a frequency table but it uses…?

Best practices charts:

- Title: clear and informative
- Provide a caption (explain clearly why you are using it)
- Scale should be relevant to the issue at hand
- Label the X and Y axis
- Use color effectively (consider color blindness and using lines instead)
- Avoid chart junk (e.g. 3-d graphs)

Best practices tables:

- Title clear and informative
- Data source should be included in the caption

- Present the percents, not just the frequencies
- Always report the missing data
- Export the table into word (ugly option) or the tabout command (pretty option) in Stata
- Spend some time editing
    - Use labels and make sure that tables are as self-explanatory as possible (variable labels and values labels)

Outcome:
Continuous: linear regression
Categorical (Binary): Logistic regression

When we are doing regression, we care about three aspects:

1. the slope of the line: Measured by the slope of the line, also known as Beta (you move it and doesn't change how close the values are and where the line crosses the Y.

This is the effect size, or the magnitude of the effect.

2. where the line crosses the y-axis: the intercept or the alpha. The value we would expect if X is zero. We would report it and put it in the regression table but you don't really comment on it or discuss it.
3. how close to the line the points are.

```
. regress vax white
```

Source	SS	df	MS	Number of obs	=	451
				F(1, 449)	=	795.21
Model	56764.0048	1	56764.0048	Prob > F	=	0.0000
Residual	32050.6465	449	71.3822861	R-squared	=	0.6391
				Adj R-squared	=	0.6383
Total	88814.6513	450	197.365892	Root MSE	=	8.4488

vax	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
white	.5778113	.0204901	28.20	0.000	.5375429	.6180798
_cons	10.31004	1.800428	5.73	0.000	6.771728	13.84835

Interpret:
*Regression coefficient(.577)
Is the change in the depended variable for every one unit change in the dependent variable
The increase is 0.5, so for every unit, it goes up about half (0.5%)
For a 1 unit increase is 0.5
For a 10 unit increase is 0.5 times 10 =

*Intercept (where does the line crosses the y) (10.3)
If the x value was zero, in this case no white people, the vaccination rate would be 10%

*Check p value of the independent variable to see if it;'s significant
*Check R2