AI UX & Data Visualisation Design Principles (CA6002)

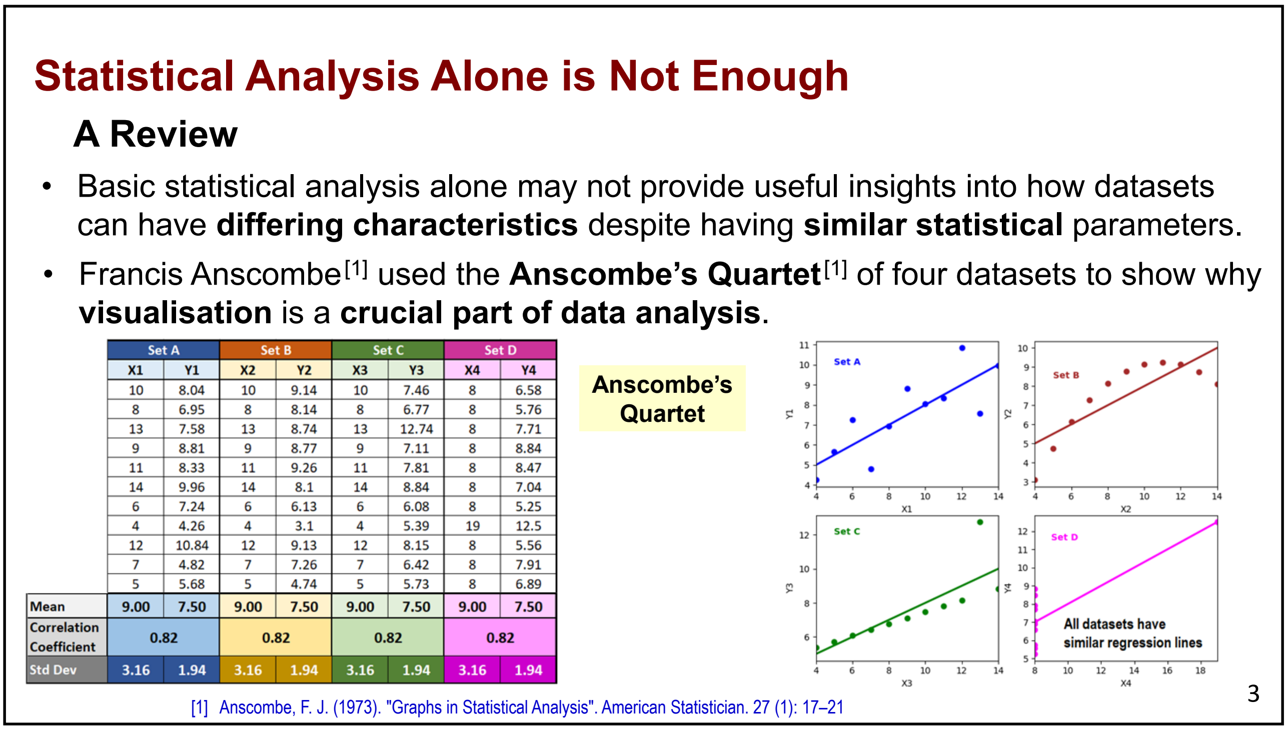
Associate Professor Goh Wooi Boon
College of Computing and Data Science
Nanyang Technological University

email: aswbgoh@ntu.edu.sg

Chapter 2.1 – Correlation Analysis

Contents

- What is Exploratory Data Analysis?
- Correlation Analysis
- Univariate Analysis
- Bivariate Analysis
- Confirmatory Analysis
- Multivariate Analysis



Statistical Analysis Alone is Not Enough

A Review

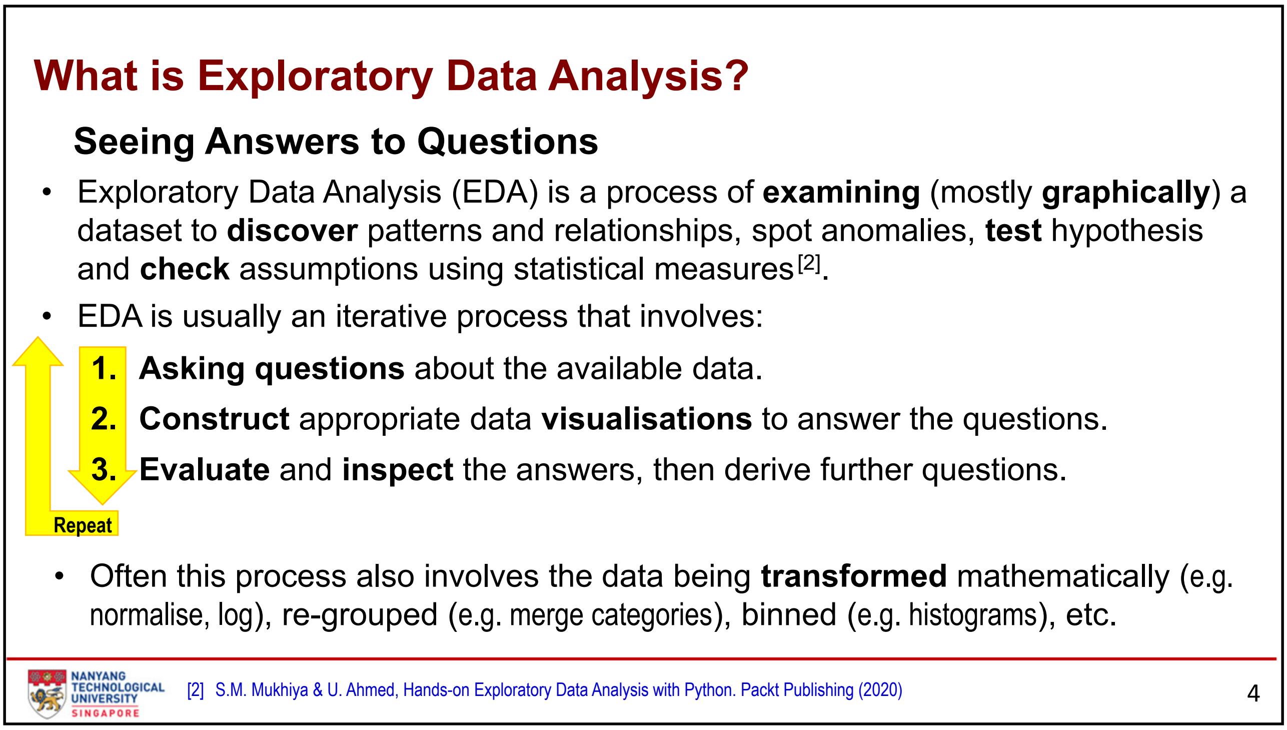
- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.
- Francis Anscombe[1] used the **Anscombe's Quartet**[1] of four datasets to show why **visualisation** is a **crucial part of data analysis**.

	Set A		Set B		Set C		Set D	
	X1	Y1	X2	Y2	X3	Y3	X4	Y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Correlation Coefficient	0.82		0.82		0.82		0.82	
Std Dev	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94

Anscombe's Quartet

All datasets have similar regression lines

[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21



What is Exploratory Data Analysis?

Seeing Answers to Questions

- Exploratory Data Analysis (EDA) is a process of **examining** (mostly **graphically**) a dataset to **discover** patterns and relationships, spot anomalies, **test** hypothesis and **check** assumptions using statistical measures[2].
- EDA is usually an iterative process that involves:

1. **Asking questions** about the available data.
2. **Construct** appropriate data **visualisations** to answer the questions.
3. **Evaluate** and **inspect** the answers, then derive further questions.

Repeat

- Often this process also involves the data being **transformed** mathematically (e.g. normalise, log), re-grouped (e.g. merge categories), binned (e.g. histograms), etc.

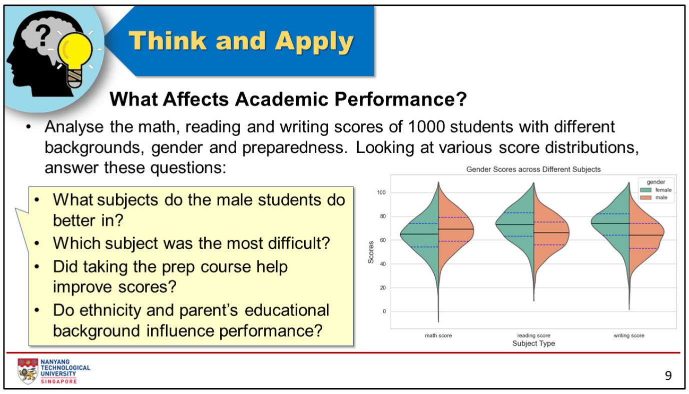
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

What is Exploratory Data Analysis?

Data Visualisation – Our Essential Tool

- Many of the data visualisation techniques that we have learnt so far will come in very handy when we do exploratory data analysis.



Think and Apply

What Affects Academic Performance?

- Analyse the math, reading and writing scores of 1000 students with different backgrounds, gender and preparedness. Looking at various score distributions, answer these questions:

- What subjects do the male students do better in?
- Which subject was the most difficult?
- Did taking the prep course help improve scores?
- Do ethnicity and parent's educational background influence performance?

Example - Analysis of Student Performance using the different Distribution Plots (e.g. Violin plot).

Correlation Analysis

Are There Relationships in the Dataset?

- Datasets often have **different categories** (i.e. columns) containing many **measured values** describing a particular phenomena, event or situation.
- Since these values are collected from the **same event**, there is a possibility that they are **related** to one another[2].
- **Correlation** is the statistical technique that examines these relationships and describes how strongly different categories of measures or values are related.
- Correlation answers questions such as:
  - How does one variable change with respect to another?
  - If it does change, to what degree or strength is this change related?

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Measure Of Correlation

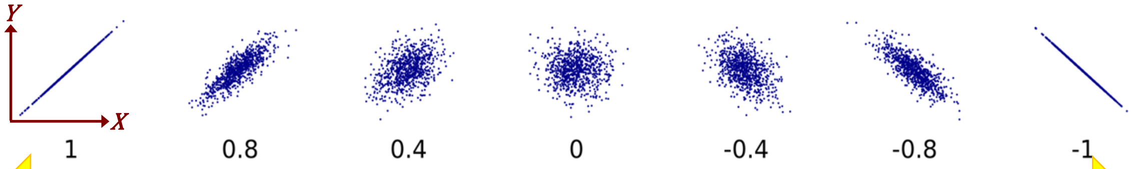
Pearson Correlation Coefficient

- Correlation measures how variables **change in tandem**, either in the same or in opposite directions, and also the magnitude or **strength** of these related changes between the variables.
- The **Pearson correlation coefficient** (ρ_{xy}), with a value between -1 and +1, measures the correlation between two variables X and Y. It is given by:

where: cov is the covariance

σ_X and σ_Y are the standard deviation of X and Y respectively.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$$ Eqn. (1)



Strong Correlation (> 0.7)

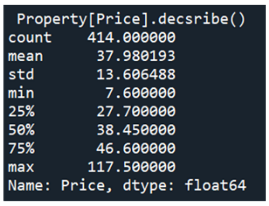
[3] Image from Wikipedia, Pearson Correlation Coefficient - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Strong Correlation (< -0.7)

Univariate Analysis

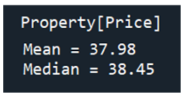
One Variable At A Time

- Univariate analysis involve the analysis of **one variable at a time** in the dataset. As such, it is not meant to find relationships between variables.
- The main purpose of univariate analysis is to:
  - find patterns in the variable such as its **central tendency** (e.g mean, median, mode)
  - understand its dispersion (e.g range, variance, maximum & minimum quartile, etc)
  - detect presence of outliers.



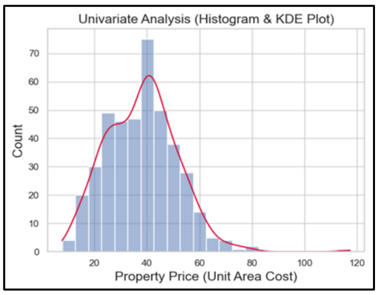
```
 Property[Price].decsribe()
count     414.000000
mean       37.980193
std        13.606488
min         7.600000
25%        27.700000
50%        38.450000
75%        46.600000
max       117.500000
Name: Price, dtype: float64
```

Apply `describe()` on category **Price** in *Property.csv*

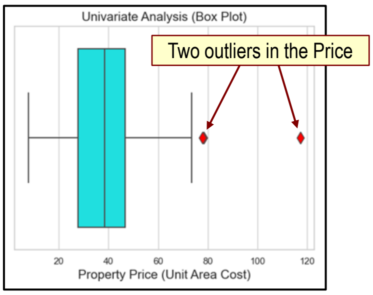


```
 Property[Price]
 Mean = 37.98
 Median = 38.45
```

`mean()` and `median()` values of Price



Histogram and Kernel Density Curve

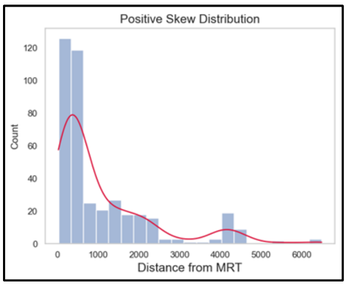


Box Plot (Observing Outliers)

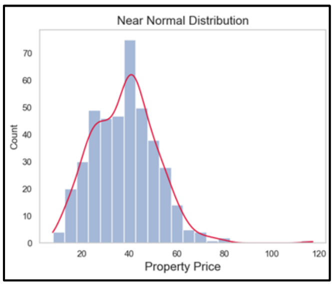
Univariate Analysis

Is My Distribution Normal Enough?

- In order to use **parametric statistical methods** (e.g. Student's t-test), our data must have a **normal distribution**, otherwise nonparametric statistical methods must be used.
- The **normality** of the data or how close they conform to a Gaussian distribution can be tested using visual methods (e.g. **statsmodels's** `qqplot()`)[4].
- Numerical evaluations can be done using **statistical methods**[5] such as:
  - Shapiro-Wilk test
  - D'Agostino's K^2 test
  - Anderson-Darling test



Positive Skew – **Property['MRT']**



Near Normal – **Property['Price']**

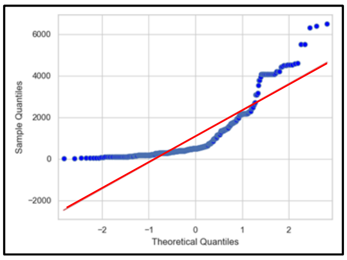
[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

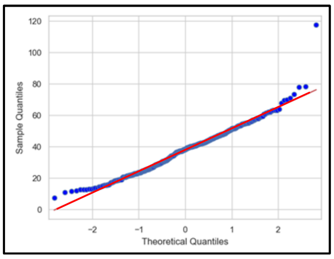
Univariate Analysis

Visualising Normality

- A popular plot for checking normality is the **quantile-quantile plot** or Q-Q plot[6].
- The plot generates its own sample of the ideal Gaussian distribution and divides these samples into groups (e.g. 5) called **quantiles**. Each data point in our sample is **paired** with a similar member in the idealised distribution and plotted on the same cumulative distribution[5].
- **Perfect normality** will see all the data points fall along the 45° **red** line (i.e. matching ideal distribution).
- Dots seen deviating from the line shows deviation from the expected normal distribution.



Positive Skew – **Property['MRT']**



Near Normal – **Property['Price']**

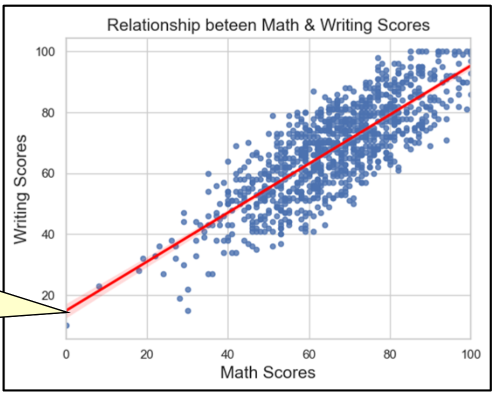
[6] Paras Varshney, Q-Q Plots Explained (2020) - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Bivariate Analysis

How Are These Two Variables Related?

- Bivariate analysis is used to find out if there is a **relationship** between two different variables in the dataset.
- A common plot used for bivariate analysis is the two-axes **scatter plot**.
- A relationship exist if the data points seem to **fit around** a **line** or a **curve**. The **tighter** the clustering, the **stronger** the relationship.



Writing vs Math Scores in *Student performance.csv*

Did students getting high Math scores also got high Writing scores?

Yes, there is a reasonably strong positive correlation between Math and Writing scores

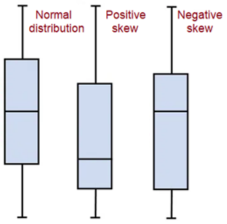
Bivariate Analysis

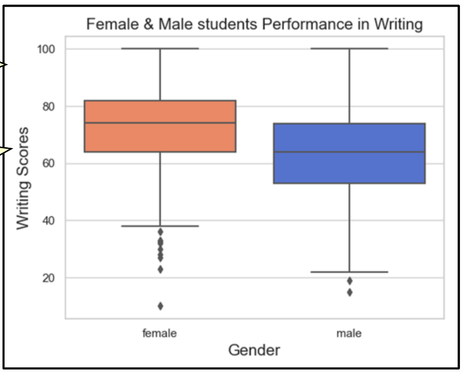
What If One Variable is Categorical?

- Bivariate analysis using a **box plot** allows us to examine the **statistical measures** of one quantitative variable along with the relationship between multiple values of a **categorical variable** in the dataset.
- The **central lines** in the box plot allow us to **compare** the **median** values within the categories (e.g. female & male).
- The nature of the data distribution (i.e. normal or skewed) can be inferred from the **relative position** of the **central line** within the box.

Do female students do better in writing?

They seem to have a higher median score.





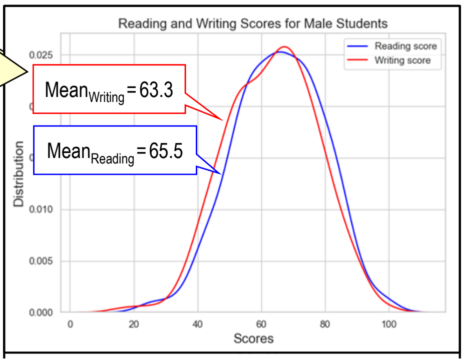
Box plots of Writing scores of Male & Female students in *Student performance.csv*

Bivariate Analysis

Confirmatory Analysis

- Visual analysis is useful for spotting relationships or patterns in data. However, we should check our assumption using appropriate statistical measures and **confirm** if an observation is **statistically significant**.
- The following confirmatory analysis steps could be taken:
  - Formulate a hypothesis.
  - Check normality of data distributions.
  - Select appropriate statistical test (e.g parametric or non-parametric).
  - Compute the test statistics to determine the ***p*** value (typically ***p*** < 0.05 is considered significant)

Are the differences between the reading & writing scores of the boys significant?



KDE plots of Reading & Writing scores of Male students in *Student performance.csv*

Confirmatory Analysis

Formulating a Hypothesis

- A hypothesis is a statement about the value of a population parameter. In order to test a hypothesis, we formulate a **null** hypothesis and the **alternate** hypothesis.
- **Null hypothesis (H_0)** - a default position statement that there is no relationship between two measured phenomena or no association among groups[7].
- **Alternate hypothesis (H_a)** - a statement that is contrary to the null hypothesis. It is usually the hypothesis being tested since we see a pattern, think it is true and want to find evidence to reject the null hypothesis and replace it with the alternate.

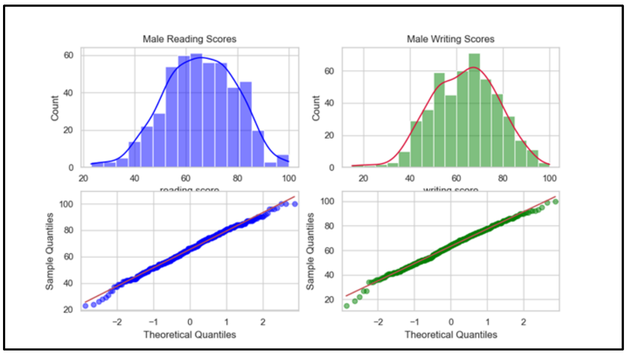
Null hypothesis ($\mathbf{H_0}$)	$\mu_{\text{Reading}} = \mu_{\text{Writing}}$	(No difference in performance)
Alternate hypothesis ($\mathbf{H_a}$)	$\mu_{\text{Reading}} > \mu_{\text{Writing}}$	(Reading scores better than writing)

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://medium.com/data-science/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

Confirmatory Analysis

Check For Normality

- In order to decide if we should use a parametric or non-parametric statistical test, we should check the **normality** of the two distributions we are comparing.
- We can do a normality check using an appropriate statistical tests[5] or use the **Q-Q plot**[4].
- Visually, both the Reading & Writing score distributions of male students seem to conform to a Gaussian distribution.
- We can do further numerical statistical test to verify this.



Histogram/KDE plots of Reading & Writing scores and their Q-Q plots

[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

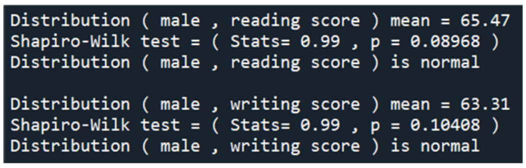
Confirmatory Analysis

Statistical Test for Normality

- The **Shapiro-Wilk** test evaluates a sample of data and quantifies how likely the data sample is drawn from a Gaussian distribution[5]. Python **scipy.stats** library provides several functions for normality statistical tests, including the Shapiro-Wilks.

```
from scipy.stats import shapiro              # import the Shapiro-Wilks test
S,p = shapiro(Dist)                          # run the Shapiro-Wilks test
if (p > 0.05):                               # check p value
  print('Distribution is normal')            # distribution is normal if p > 0.05
```

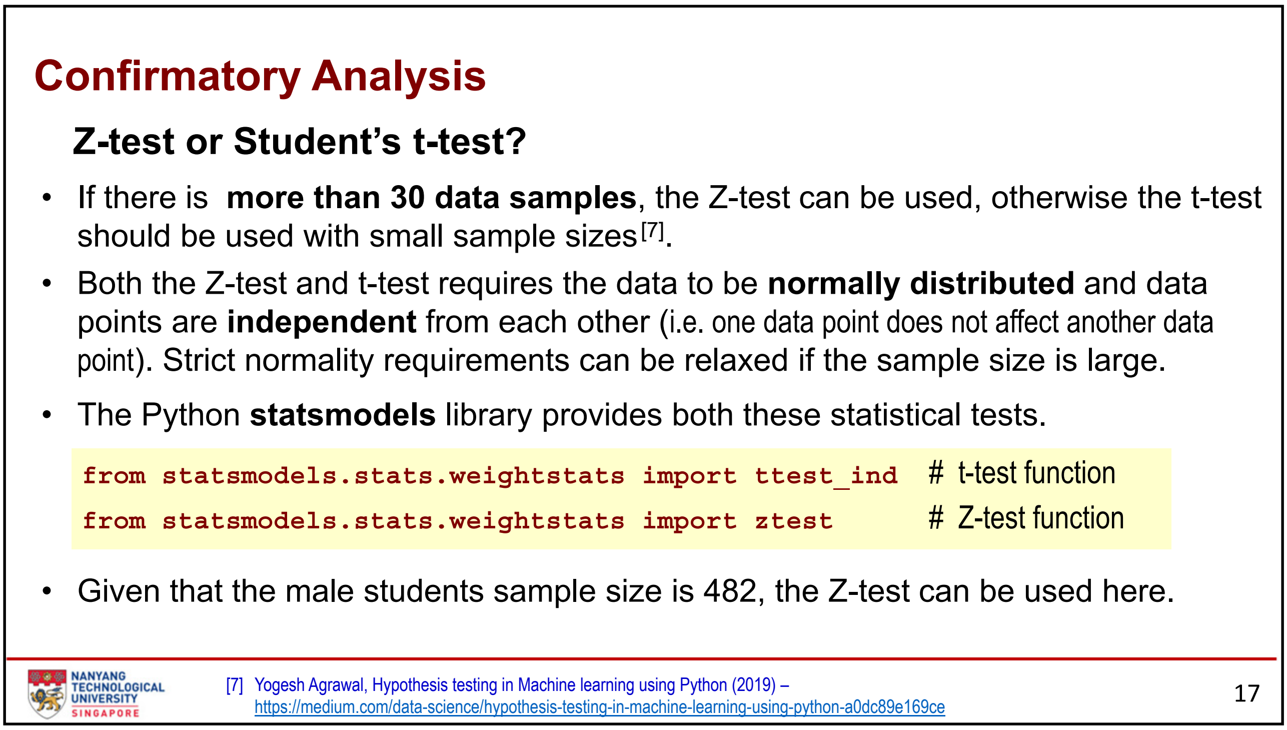
- Both Reading & Writing distributions meet the normality confidence level of 0.05.
- We can thus look at using a parametric test (e.g. the t-test or Z-test)



```
Distribution ( male , reading score ) mean = 65.47
Shapiro-Wilk test = ( Stats= 0.99 , p = 0.08968 )
Distribution ( male , reading score ) is normal

Distribution ( male , writing score ) mean = 63.31
Shapiro-Wilk test = ( Stats= 0.99 , p = 0.10408 )
Distribution ( male , writing score ) is normal
```

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/



Confirmatory Analysis

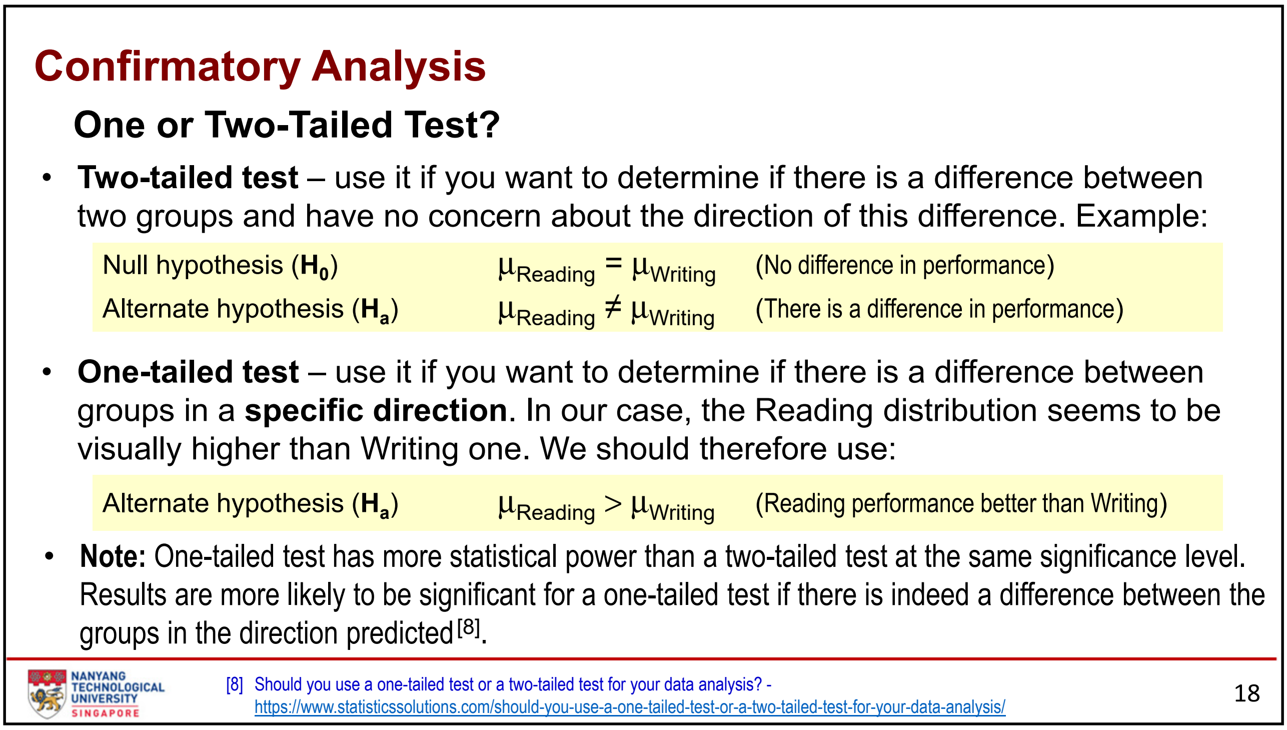
Z-test or Student’s t-test?

- If there is **more than 30 data samples**, the Z-test can be used, otherwise the t-test should be used with small sample sizes[7].
- Both the Z-test and t-test requires the data to be **normally distributed** and data points are **independent** from each other (i.e. one data point does not affect another data point). Strict normality requirements can be relaxed if the sample size is large.
- The Python **statsmodels** library provides both these statistical tests.

```
from statsmodels.stats.weightstats import ttest_ind   # t-test function
from statsmodels.stats.weightstats import ztest       # Z-test function
```

- Given that the male students sample size is 482, the Z-test can be used here.

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://medium.com/data-science/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce



Confirmatory Analysis

One or Two-Tailed Test?

- **Two-tailed test** – use it if you want to determine if there is a difference between two groups and have no concern about the direction of this difference. Example:

Null hypothesis (**H_0**)	$\mu_{Reading} = \mu_{Writing}$	(No difference in performance)
Alternate hypothesis (**H_a**)	$\mu_{Reading} \neq \mu_{Writing}$	(There is a difference in performance)

- **One-tailed test** – use it if you want to determine if there is a difference between groups in a **specific direction**. In our case, the Reading distribution seems to be visually higher than Writing one. We should therefore use:

Alternate hypothesis (**H_a**)	$\mu_{Reading} > \mu_{Writing}$	(Reading performance better than Writing)

- **Note:** One-tailed test has more statistical power than a two-tailed test at the same significance level. Results are more likely to be significant for a one-tailed test if there is indeed a difference between the groups in the direction predicted[8].

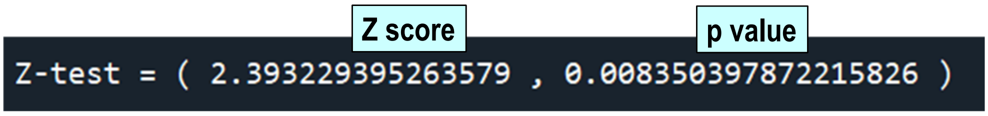
[8] Should you use a one-tailed test or a two-tailed test for your data analysis? - https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

Confirmatory Analysis

Statistical Significance

- **Z-test results** – Applying one-tailed ('larger') Z-test[9] on the **Reading** & **Writing** data:

```
Z,p = ztest(Reading, Writing, alternative='larger')   # call Z-test function
```



```
Z-test = ( 2.393229395263579 , 0.008350397872215826 )
```

- **Rejecting null hypothesis** – the threshold at which is generally considered safe to reject the null hypothesis is a ***p*** value of (***p*** < 0.05). This means there is a less than 5% chance that the observed data is due to chance.
- **Statistically Significant** – since the ***p*** value computed by the Z-test is way smaller than 0.05, we can safely say that the observed mean Reading score being higher than the mean Writing score for Male students is statistically significant.

[9] Statsmodels ztest fucntion - https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

Confirmatory Analysis

Selecting The Statistical Test

- Selecting an appropriate statistical test for your confirmatory analysis requires consideration of various factors such:
  - Continuous or discrete data.
  - Number of samples to compare.
  - Sample size.
  - Parametric or non-parametric test.
- A helpful flowchart in[10] shows some statistical tests that can be use based on the nature of the data and test requirements.

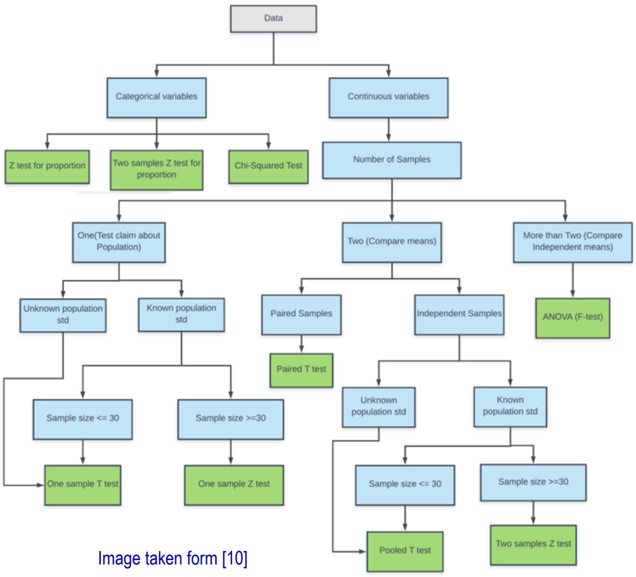
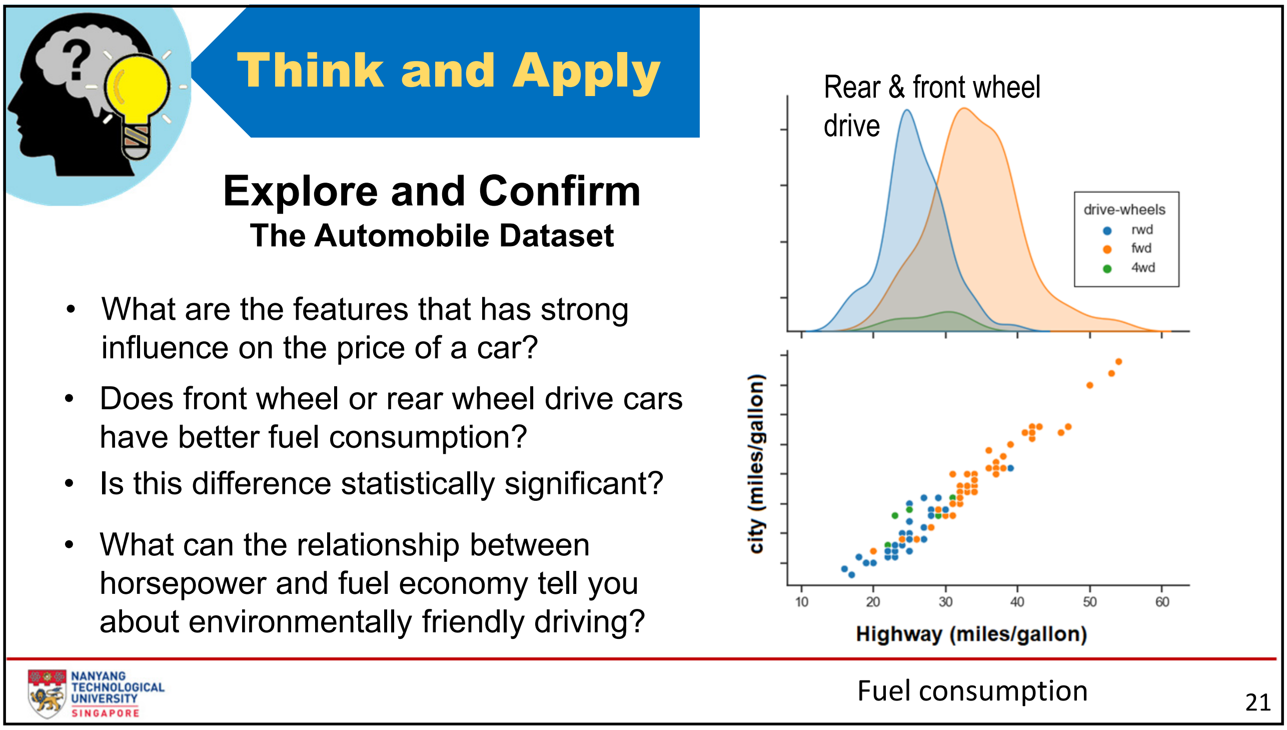


Image taken form [10]

[10] Jagandeep Singh, Statistical Tests with Python (2020) - https://python.plainenglish.io/statistical-tests-with-python-880251e9b572



Think and Apply

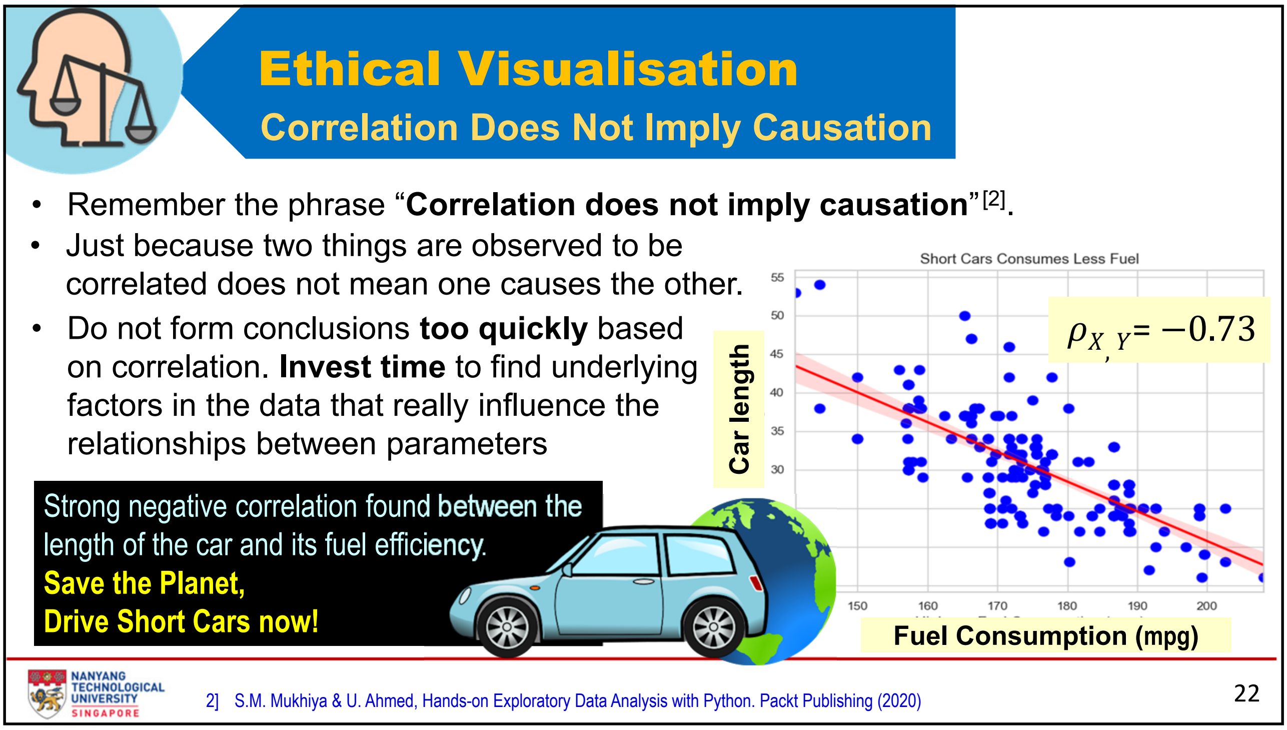
Explore and Confirm

The Automobile Dataset

- What are the features that has strong influence on the price of a car?
- Does front wheel or rear wheel drive cars have better fuel consumption?
- Is this difference statistically significant?
- What can the relationship between horsepower and fuel economy tell you about environmentally friendly driving?

Rear & front wheel drive
drive-wheels
● rwd
● fwd
● 4wd
city (miles/gallon)
10
20
30
40
50
60
Highway (miles/gallon)

Fuel consumption



Ethical Visualisation

Correlation Does Not Imply Causation

- Remember the phrase “**Correlation does not imply causation**”[2].
- Just because two things are observed to be correlated does not mean one causes the other.
- Do not form conclusions **too quickly** based on correlation. **Invest time** to find underlying factors in the data that really influence the relationships between parameters

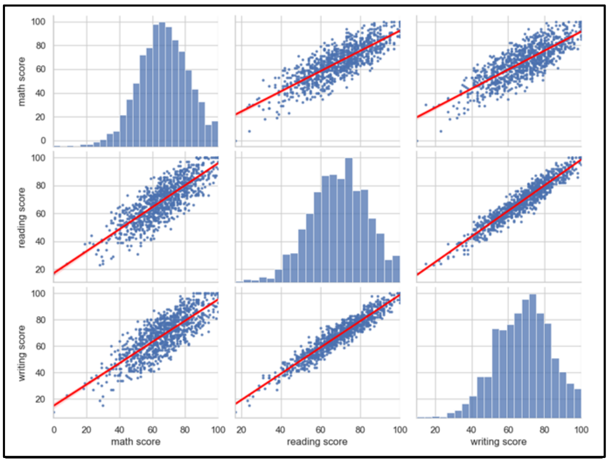
Strong negative correlation found between the length of the car and its fuel efficiency.
Save the Planet,
Drive Short Cars now!

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Multivariate Analysis

Analysis Of Three Or More Variables

- Multivariate analysis allows us to look at the **correlation of three** or **more** variables at a time.
- A common way of visualising multivariate data is to use a **matrix** of multiple **scatter plots**.
- The convenient and powerful Seaborn `pairplot()` function provides this feature[11].
- Setting the `kind` parameter to `kind=‘reg’` will show the linear regression lines.



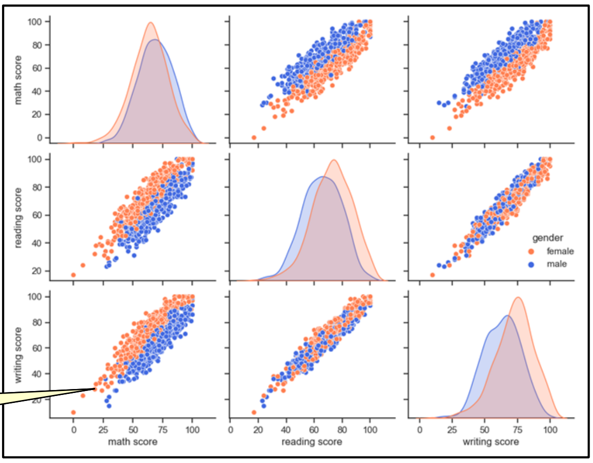
Analysing Math, Reading & Writing scores in *Student performance.csv*

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

Multivariate Analysis

Adding In Categorical Variables

- An additional **categorical variable** can be visualised in **colours** by employing the `hue` parameter in `pairplot()`[11].
- By assigning the hue parameter to a specific categorical variable (e.g. `hue=‘gender’`), each type in the category is assign a different colour from a default or specified palette.
- Such multivariate analysis allows interesting relationships and patterns to be observed in the different categorical types.



Boys do better in math than writing

Math, Reading & Writing scores of Female & Male students

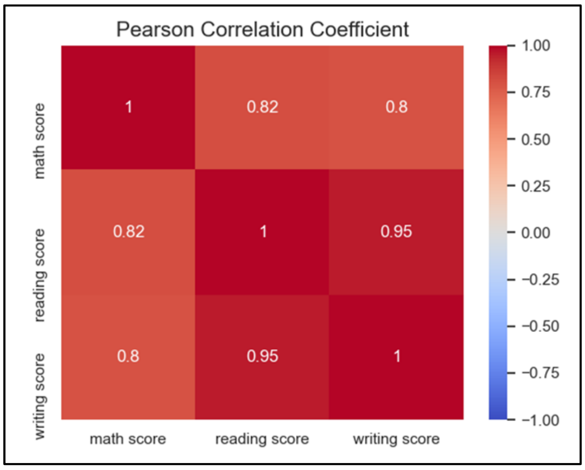
[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

Multivariate Analysis

Getting The Correlation Values

- Instead of just visualising the correlation, it is often useful to compute a numerical measure for each correlation pair.
- **Pandas** `corr()` function and **Seaborn**’s `heatmap()` can compute and display the **pairwise** linear (Pearson) correlation coefficients for all the columns in the dataframe with **numerical** values[12].

```
sns.heatmap(
   Data.corr(method='pearson’),    # Pearson coeff
   annot=True,                     # show coeff values
   vmin=-1, vmax=1,                # min-max scale
   cmap='coolwarm’))               # diverging palette
```



Correlation coefficients of Math, Reading & Writing scores

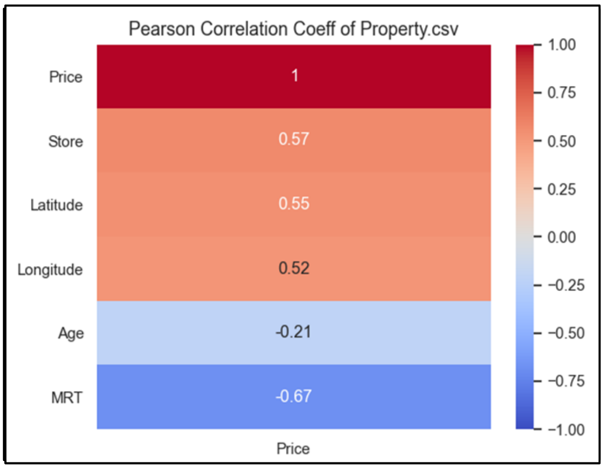
[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

Multivariate Analysis

Putting The Heat On One Variable

- It is helpful to focus on the **correlation strength** between all the **independent** variables and the **dependent** variable.
- Besides using **Seaborn**’s `heatmap()`, the **Pandas** `sort_values()` feature could be used to **sort** the correlation strength for easy visualisation[13].

```
sns.heatmap(
   Data.corr(method='pearson’)    # Pearson coeff
   [[‘Price’]].sort_values(by=‘Price’,
   ascending=False),              # sort descending
   annot=True,                    # show coeff values
   vmin=-1, vmax=1,               # min-max scale
   cmap='coolwarm’))              # diverging palette
```



Pearson correlation coefficients of Property.csv

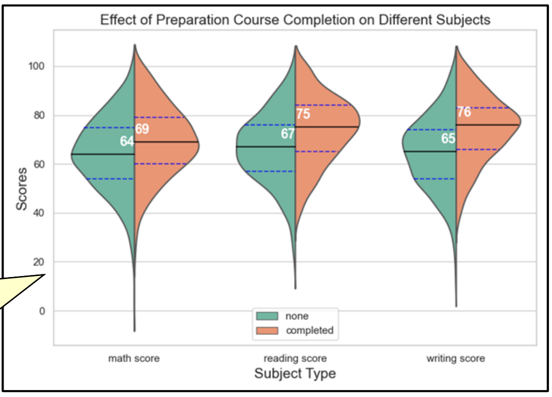
[13] Bibor Szabo, How to Create a Seaborn Correlation Heatmap in Python? (2020) - https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

Multivariate Analysis

Playing With Many Violins

- Multivariate analysis using **violin plots** allows us to compare the **probability density curve** of multiple quantitative variables for a **categorical variable pair**.
- The **central lines** in the violin plot allow us to **compare** the **median** values within the categories (e.g. those who completes test preparation course & those who did not).
- In Seaborn's `violinplot()`, the `hue` parameter can be assigned to a categorical variable with two values. Setting `split=True`, then compares them side-by-side.

Students do better in all subject when completing test preparation.



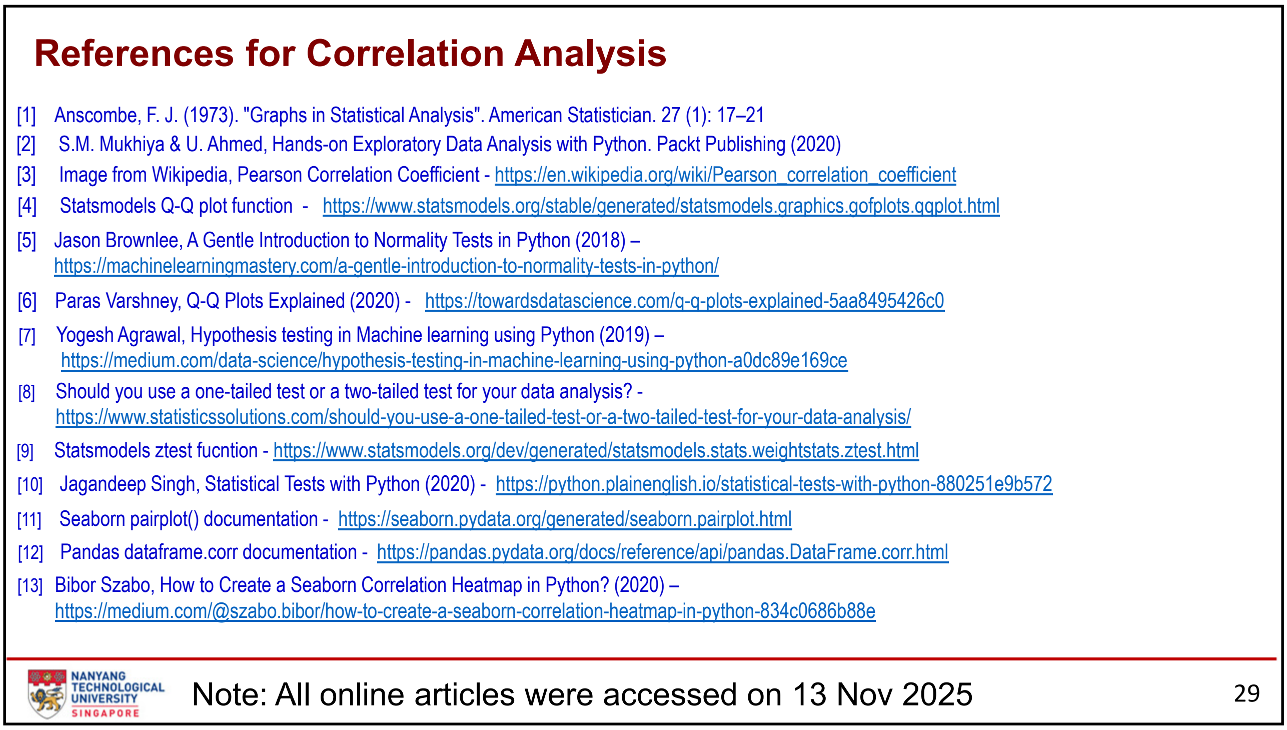
Violin plots showing effects of prep course completion on Math, Reading & Writing scores

Summary

Correlation Analysis

- Correlation analysis allows us to visualise **relationships** within the dataset.
- Data visualisation techniques for this purpose include histogram & kernel density plots (univariate), scatter plot (bivariate), box & violin plots (bivariate & multivariate), paired plot & heatmap (multivariate).
- Correlation between variable pairs can be quantified using the **Pearson correlation coefficient** and this values can be effectively visualised using appropriate **colour palettes** in heatmaps and paired plots.
- Visual observation of relationships and patterns should be confirmed with appropriate **statistical analysis** based on the **nature** and **number** of data distributions being compared

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



References for Correlation Analysis

[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

[3] Image from Wikipedia, Pearson Correlation Coefficient - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

[6] Paras Varshney, Q-Q Plots Explained (2020) - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://medium.com/data-science/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

[8] Should you use a one-tailed test or a two-tailed test for your data analysis? - https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

[9] Statsmodels ztest fucntion - https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

[10] Jagandeep Singh, Statistical Tests with Python (2020) - https://python.plainenglish.io/statistical-tests-with-python-880251e9b572

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

[13] Bibor Szabo, How to Create a Seaborn Correlation Heatmap in Python? (2020) – https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

Note: All online articles were accessed on 13 Nov 2025