Past Practice:

1. **True or False:**
   - Your chosen major/specialisation is an example of an activity you participate in daily that creates "data" for Sydney University. False
   - "Analysis of student data" is an example of a problem you (as a data scientist) can solve at Sydney University. False
   - In your role as a Sydney University student, not doing an activity may create "data" for the University. T
   - Because Big Data is expensive to collect, store and analyse, only large for-profit organisations (such as banks, insurance companies, retail organisations) can afford to innovate with Data Science. False
   - The main goal of Data Science in business is to make money for the business/company using big data. F

2. As shown in the TWDI survey companies are at very different stages of Data Science maturity. One of the biggest challenges for companies is:

A. Very high level statistical and computational expertise on any type of data

B. Expertise in helping companies accomplish the cultural and managerial changes of analytic initiatives

C. Developing executive reports and managerial dashboards from raw data.

D. Buying the right technology platforms and up-to-data infrastructure.

3. Distinguishing between Big Data, Data Science and Small Data projects is important because:

A. They have different levels of objectivity, accuracy and predictive capability

B. The distinction is not important as they definitions are contested.

C. For large companies only Data Science is important

D. It depends on who you talk to.

4. Data quality are important in Data Science:

A. When people and/ or machine are used to analyse data.

B. Only when people are looking at data

C. They are not important when the data Volume and Veracity is high

D. When specific types of analytic models are used.

5. In the context of banking fraud and manipulation of financial markets the best way to determine what "the problem" is to:

A. Collect digital data until you have all four V's represented in your data set.

B. Wait until the financial crisis has passed and then collect data to analyse

C. Consult with bankers, lawyers and technical experts to evaluate what data are important

D. Study social media records of all customers with machine learning

6. Sensors, screening techniques, medical tests and other data collection techniques are very likely to contain false positive/negatives. As data scientists these errors can safely be ignored because:

A. The volume of data ensures these errors don't occur

B. This is not relevant because Data science is about numbers

C. It is not the Data Scientists job to make decisions

D. Data Scientists are only concerned with long-term trends

E. It is not safe to ignore any of these errors

7. 'Data' may include behaviours that people perform, things that do not happen and emotions people feel
True or False True

8. It is very important to distinguish between Big Data, Data Science and Small Data projects is because:

A. it is only important so you understand what other people mean by terms they are using

B. you should always seek enough data to use Data Science techniques or your solutions will be inaccurate

C. they have different levels of objectivity, accuracy and predictive capability

D. Big Data is needed to solve the real problems business face

E. The distinction is not important as the definitions are contested

9. Sufficient volumes, variety and velocity of Data will ensure that false positive/ false negatives do not occur. T/F
True **大概看看就好，不要纠结，大概说的就是数据增加，Type I and II Error 都会减少，之前答案给的就是 True**

10. the use of readily available surrogate measures for people behaviour is the best way to collect data for predictive modelling. T/F True (了解就好了，这学期没这句话)

11. in the context of banking fraud and manipulation of financial markets the best way to determine what 'the problem' is to:

A. wait until the business crisis has passed and then collect data to analyse

B. use machine learning to study all data available until the problem is identifies

C. collect digital data until you have all four V's represented in your data set

D. **consult with domain experts such as bankers, lawyers and technical experts to evaluate what data are significant'**

12. data quality dimensions include

A. clarity, detail and order

B. relevance

C. accuracy, completeness and scope

D. a time dimension

E. All of the above (这学期删了这一页 PPT 我也不知道为什么- -)

13. Triple-loop learning or determining what is morally right will be easier as we can collect large values of data and train good algorithms. True or False.
False

14. The most important part of 'Big Data' analysis is trying enough analytic models until you achieve the answer you want. True or False
False

15. Suppose you are given the direct marketing in the pandas DataFrame “Marketing”. You wish to find the total Spending of each age group. Which line of Code will Compute this?

A. Marketing[‘Age’].groupby(sum)[‘AmountSpent’]

B. Marketing[‘Age’].sum()[‘AmountSpent’]

C. Marketing[‘Age’].groupby(‘Age‘).head()[‘AmountSpent’]

D. Marketing[‘Age’].groupby(‘Age’).sum()[‘AmountSpent’]

16. Which of the following examples needs to be addressed by the UNSUPERVISED learning algorithm?

A. Given some house size and price data, learn a house prices prediction model.

B. Provided with some email labelled as spam/not spam, learn a spam filter.

C. Given a dataset of customer classifies as low or high risk customer, predict the risk level of a new customer.

D. Given a data set of customer occupation, age and income data, automatically group the customers into different market segments.

17. Which of the following models is not a linear regression

A. $y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2^2 + \varepsilon$

B. $y = \beta_0 + \beta_1 x_1^2 + \beta_2 e^{x_2} + \beta_3 x_3 + \varepsilon$

C. $y = \beta_0 + \beta_1 e^{x_1} + \beta_2 x_2^2 + \beta_3 x_3 + \beta_4 x_2 x_3 + \varepsilon$

D. $y = \beta_0 + \beta_1 x_1 + \log(\beta_2) x_2 + \beta_3 x_3 + \beta_4 x_3 x_4 + \varepsilon$

18. Suppose you are given the following output where final exam score (out of 40) and attend is the number of lectures attended.

Final = 22.73+0.121Attend, n=680, R^2=0.02

The highest score on the final was 39. In order to predict a score of 39, what would attend have to be?

134.4628

19. You Are given a Pandas DataFrame (df) with some features(Columns) missing all their values and some features missing a few entries. Which line of code would remove the columns that are completely empty or null?

A. df.dropna(axis=0, how = ‘any’)

B. df.dropna(axis=0, how = ‘all’)

C. df.dropna(axis=1, how = ‘any’)

D. df.dropna(axis=1, how = ‘all’)

	Coefficient	Std.Error	t-statistics	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

20. The table contains the regression results for predicting the number of units sold given advertisement budgets on TV, radio and in newspapers. Sales are measured in $1000's. Advertisement budgets are measured in $1000's too. What is the most effective way to advertise?

    A. TV

    B. Radio

    C. Newspaper

21. Suppose that the simple regression of sales on newspaper advertising expenditure reveals a large positive effect. Does it make sense for the multiple regression to suggest no relationship between sales and newspaper while simple linear implies the opposite?

    A. Yes

    B. No

	Coefficient	Std.Error	t-statistics	p-value
Intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
Radio	0.0289	0.009	3.24	0.0014
TV*Radio	0.0011	0.000	20.73	<0.0001

22. Which of the following is most accurate?

    A. The p-value for TV is very small indicating that we cannot reject the null hypothesis: $\beta_{TV} = 0$

    B. None of the listed answer is correct

    C. TV advertising is significantly less effective when combined with radio advertising

    D. All of the p-values are smaller than 0.01 indicating when testing for statistical significance probability of false positives is lower than 1%

23. According to the regression result above:

    A. The p-value for TV*radio is very small indicating that we cannot reject null hypothesis: $\beta_{TV*radio} = 0$

    B. All of the p-values are smaller than 0.01 indicating when testing for statistical significance probability of false positives is lower than 1%

    C. TV advertising is significantly more effective when combined with radio advertising.

    D. AC are correct

    E. BC are correct

24. Which of the flowing steps is NOT part of the Cross Industry Standard Process for Data Science? (CRISP)

    A. Business understanding

    B. Data preparation

    C. Resource allocation

    D. Modelling

	Coef.	Std. err	t	P> |t|	[0.025	0.975]
Const	-4.847e+04	1.16e+04	-4.183	0.000	-7.12e+04	-2.57e+04
SQFT	85.8384	3.821	22.465	0.000	78.341	93.336
Bedrooms	-2.825e+04	4586.877	-6.158	0.000	-3.72e+04	-1.92e+04
Baths	4.745e+04	5547.113	8.554	0.000	3.66e+04	5.83e+04
Pool	-1625.4289	8926.263	-0.182	0.856	-1.91e+04	1.59e+04

25. The above is the regression results for predicting house price based on the size of the house in square feet (SQFT), the number of bedrooms, the number of bathrooms and whether or not the house has a pool. What is the most significant factor contributing to the house price?

    A. SQFT

    B. Baths

    C. Bedrooms

    D. Pool

26. Given the multiple regression result for predicting the house price. Does it make sense for the multiple regression to suggest no relationship between house price and the presence of a pool in the house?

    A. Yes

    B. No

27. Which of the following examples are instances of supervised learning?

    A. The question of whether customers who buy infant milk formulas tent to buy nappies

    B. The question of whether a bank customer with a credit score of 550 should be have his loan application approved

    C. The task of finding spam emails according to where an email is sent from and the content of the email

    D. The question of grouping different customers into high-income group, middle-income group and low-income group

    E. AD

    F. BC

28. Suppose that X is an independent variable and Y is the response variable a linear regression model can be used to answer all of the following questions except for one. Which one is not a valid use of a linear regression?

    A. To determine if a change in X causes a change in Y

    B. To predict the value of Y for an individual, given that individual's X value

    C. To estimate the change in Y for a one-unit change in X

    D. To estimate the average value of Y at a specified value of X

29. Which of the following might NOT lead to a problem inverting X'X when estimating a regression model?

    A. When one regressor is close to being a linear combination of the others

    B. When all observation on one regressor, e.g. age, happen to be twice larger than the corresponding observations on the other regressor, e.g., years of education

    C. When there are more parameters and observations.

    D. When you include a log transformation of a regressor and a log transformation of the square of the same regressor

30. Given two ndarray's A and B, which line of Python code would successfully compute A times B' and store the result in C?

    A. C= np.dot(A, B.transpose())

    B. C=A*B.transpose()

    C. C= np.dot(A, B)

    D. C= A*B

31. Given the following code

Marketing=pd.read_excel ('DirectMarketing. xlsx')

Where marketing has columns labelled 'Salary', 'Age', 'Catalogs', 'AmountSpent'. You wish to find customers in the data who received more than 10 catalogs and are not Old. Which line of Python will compute this?

A. Marketing[(marketing['Catalogs']>10)&(marketing['Age']!= 'Old')]

B. Marketing[marketing['Catalogs']>10& marketing['Age']!= 'Old']

C. Marketing[(marketing['Catalogs']>10)&(marketing['Age']== 'Old')]

D. Marketing[(marketing[Catalogs]>10)&(marketing[Age]==Old)]

32. Given a categorical variable with values of 'Sydney', 'New York', 'London' and 'Paris'. Suppose that we want to convert it into a numerical variable, how many dummy variables do we need?

    A. 3

    B. 2

    C. 4

    D. 5