REVISION GUIDE: QUESTION & ANSWER SHEET

DECODED

BIG DATA, ANALYTICS & DATA SCIENTIST ROLE

1. BIG DATA

1. Describe the main characteristics of Big Data.	*Volume:* large volumes of data, grows exponentially each year *Velocity:* data is acquired very frequently, and it is aggregated and dispersed very fast *Variety:* the variety of Big Data ranges from traditional, highly structured data to unstructured sources like texts, images, video, audio, etc. *Veracity:* Big Data needs to be as trustworthy or reliable as possible to be able to use your business.
2. What are the 4 parts of a Data Ecosystem, and what is unique to each part?	*Data Devices*: Data devices are anything that are constantly *gathering information*. *Data Collectors*: Anything that *collects* any sort of *data from devices/* *Data Aggregators*: Data aggregators *compile* data from devices and collectors, and then *transform* and package the information to *sell* to others. *Data Users & Buyers*: The users and buyers are those who *directly benefit* from the data collected and aggregated by others by *buying* this information

2. INTRODUCTION TO DATA

1. What are the four types of data structures? Provide an example for each type.	*Structured:* databases *Semi-structured:* spreadsheets, xml files *Quasi-structured:* clickstream data *Unstructured:* image, audio, & video files; tweet;, newspapers

2. Describe three differences between data science and business intelligence roles.	1. Data science deal with the how & why, while business intelligence deals with the when & where 2. Data science work with a range of structured and unstructured data, while business intelligence works with structured data only 3. Data science provides insight & foresight (predictive modelling, optimisation), while business intelligence provides hindsight & insight (alerts, ad hoc reports)

DATA ANALYTICS LIFECYCLE

INTRODUCTION TO DATA

1. Describe the 6 phases of the data lifecycle.	1. *Discovery*: frame the business problem as an analytics challenge to be addressed in subsequent phases; formulate initial hypotheses, begin learning from the data 2. *Data preparation*: Use an analytic sandbox to work with data and perform analytics throughout the project; execute ELT / ETL process for analysis; get familiar with the data and take steps to condition it 3. *Model planning*: determine methods, techniques, and workflow for the subsequent model building phase; data exploration to see the relationships between different variables; choose key variables and appropriate models 4. *Model building*: develop datasets for testing, training, and production purposes; build and executes models based on work in the model planning phase; consider whether its existing tools will suffice for running the models or if a more robust environment is needed 5. *Communicate results*: determine if the results of the project are a success or failure based on criteria from Phase 1; identify key findings, quantify business values, develop a narrative to summarise findings for stakeholders 6. *Optimisation*: delivery of final reports, briefings, code, and technical documents; team may run a pilot project to implement the models in a production environment
What is a data sandbox? In what phase is required?	It is required for phase 2 (data preparation). It is a scalable and developmental platform to explore information sets of an organisation (via interaction and collaboration). It allows a company to realise its investment value in big data.

INITIAL ANALYSIS OF THE DATA

INTRODUCTION TO STATISTICS

1. How do the following differ from each other: *a. z-test and Student's t-test?* *b. z-distributions and t-distributions?*	a. z-tests test a random sample against a known population to see if the sample comes from (is the same) or doesn't come from (different) from population. Student's t-tests test a random sample against another random sample to see if they are the same or different. b. z-distributions have means and standard deviations from a known population; t-distributions have means and standard deviations from a random sample (but when random sample is large, can resemble z-distribution)
2. How do Wilcoxon tests differ from: *a. Student's t-test?* *b. Welch's t-test?*	a. Wilcoxon tests if two samples from two different populations are the same or not; whereas student's t-tests tests two random samples against each other to see if they are similar or different (and thus if they are from the same or different populations). b. Welch's t-tests test if two random samples have the same or different sample means.
3. Aside from ANOVAs, what other statistical tests require normal distribution assumptions be met?	● z-test ● Student's t-test ● Welch's t-test
4. In what situations would ANOVAs be more useful than t-tests or Wilcoxon tests?	If you are testing for similarities/differences of three or more groups, and when groups have the same or similar sample sizes.

ADVANCED ANALYTICS: THEORY & METHODS

1. CLUSTERING

1. What does the 'k' in k-means represent?	*'k'* represents the number of clusters in the algorithm; it is a user-defined value when testing or calculating the optimal number of clusters
2. What is a centroid and why is it important in k-means?	Centroids are the center-points of a given cluster; each unique cluster has its own centroid. They are important in k-means because the optimal centroid is the mean of all the data points in

	a cluster where the distance between each data point and the centroid is minimized.
3. How do you determine the optimal/best value for k?	WSS is used to determine the optimal value of *k.* When data points are all close to their cluster's centroid, WSS is small; if data points are far from a cluster's centroid, the WSS is big. The Elbow Curve is a visual way to determine the optimal value of *k*, especially when trying to choose between 2 values of *k* that both minimise WSS.

2. ASSOCIATION RULES

1. Evaluate the following from the sample dataset used above:	Supp(Milk) = ⅘ = 0.80 Supp(Milk & Cereal) = ⅖ = 0.4 Supp(Milk & Cheese) = ⅕ = 0.2 Supp(Milk, Cheese, & Bread) = ⅕ = 0.2
2. Evaluate the following from the sample dataset used above: a. *What can you infer from these values?* b. *Does the order of the items matter when calculating confidence?*	Confidence(Milk ⇒ Cheese) = $\frac{Supp\ (Milk\ \&\ Cheese)}{Supp\ (Milk)} = \frac{1/5}{4/5} = 1/4$ Confidence(Cheese ⇒ Milk) = $\frac{Supp\ (Milk\ \&\ Cheese)}{Supp\ (Cheese)} = \frac{1/5}{2/5} = 1/2$ Confidence(Milk ⇒ Cereal) = $\frac{Supp\ (Milk\ \&\ Cereal)}{Supp\ (Milk)} = \frac{2/5}{4/5} = 1/2$ a. We can infer that we are most confident that if someone buys cheese that they'll also buy milk. Interestingly, if someone buys milk, we have less confidence that they will also buy cheese. b. Yes, order does matter. Clear from above as we got different confidence values for {Milk ⇒ Cheese} and {Cheese ⇒ Milk}.
3. Evaluate the following from the sample dataset used above: a. *What can you infer from these values?* b. *Does the order of the items matter when calculating lift?*	Lift(Milk ⇒ Cheese) = $\frac{Supp\ (Milk\ \&\ Cheese)}{Supp\ (Milk) \times Supp\ (Cheese)} = \frac{1/5}{4/5 \times 2/5} = 0.625$ Lift(Milk ⇒ Cereal) = $\frac{Supp\ (Milk\ \&\ Cereal)}{Supp\ (Milk) \times Supp\ (Cereal)} = \frac{2/5}{4/5 \times 2/5} = 1.25$ Lift(Cereal ⇒ Milk) = $\frac{Supp\ (Cereal\ \&\ Milk)}{Supp\ (Milk) \times Supp\ (Cereal)} = \frac{2/5}{4/5 \times 2/5} = 1.25$ a. For milk and cheese, the LHS (milk) seems to negatively affect the presence of the RHS (cheese) in the basket. For the next two examples, the LHS (milk or cereal) seems to positively affect - or 'lift' - the presence of RHS (milk or

	cereal) in the basket; there's an equally positive association between milk and cereal, and cereal and milk. b. No, the order does not matter when calculating lift (see second and third calculations), since the equation stays the same if you swap LHS and RHS as they are both calculated in the denominator. This is regardless of whether they are single itemsets, double or more.
4. Evaluate the following from the sample dataset used above: *a. What can you infer from these values?* *b. Does the order of the items matter when calculating leverage?*	Leverage(Milk ⇒ Cheese) = $Supp(Milk \ \& \ Cheese) - (Supp(Milk) \times Supp(Cheese))$ = ⅕ - (⅗*⅖) = -0.12 Leverage(Milk ⇒ Cereal) = $Supp(Milk \ \& \ Cereal) - (Supp(Milk) \times Supp(Cereal))$ = ⅖ - (⅗*⅖) = 0.08 Leverage(Cereal ⇒ Milk) = $Supp(Cereal \ \& \ Milk) - (Supp(Milk) \times Supp(Cereal))$ = ⅖ - (⅗*⅖) = 0.08 a. The first leverage value shows a negative leverage (i.e., an equally negative association between itemsets), which supports the lift value (lift < 1) for the same itemsets. For the second two values, leverage was equally positive, again supporting the lift values (lift > 1) for these different itemsets. b. No, the order of itemsets does not matter when calculating leverage. As with lift, the equation stays the same regardless of the order of LHS and RHS.
5. Evaluate the following from the sample dataset used above: *a. What can you infer from these values?* *b. Does the order of the items matter when calculating conviction?*	Conviction(Milk ⇒ Cheese) $= \frac{1 - Support(Cheese)}{1 - confidence(Milk => Cheese)}$ $= \frac{1 - 2/5}{1 - 1/4}$ = 0.8 Conviction(Milk ⇒ Cereal): $= \frac{1 - Support(Cereal)}{1 - confidence(Milk => Cereal)}$ $= \frac{1 - 2/5}{1 - 1/2}$ = 1.2 Conviction(Cereal ⇒ Milk): $= \frac{1 - Support(Milk)}{1 - confidence(Cereal => Milk)}$ $= \frac{1 - 4/5}{1 - 1}$ = infinity a. The first two conviction values for each of the above combinations doesn't seem to be too high, with {Milk} ⇒ {Cheese} showing the least dependency. With the last conviction value, due to the situation of a zero-value

	denominator, conviction cannot be calculated (and noted as a value of infinity). b. Yes, order of itemsets does matter when calculating conviction because it involves calculating confidence (and the value order with calculating confidence does matter). You can see the difference in {Cereal} ⇒ {Milk} (conviction = infinity) versus {Milk} ⇒ {Cereal} (conviction = 1.2), where the difference is in the order of the itemsets in the confidence value in the denominators.

3. LINEAR REGRESSION

1. In the two graphs above, which appears to have a better fit? Why?	The graph on the left, because the relationship between X and Y is much tighter (data points closer to the line) and so it has a higher r^2
2. If X accounts for all of the variation in Y (i.e., all the data points fall exactly on the regression line), what would the r^2 value be?	r^2 = 1 (X predicts Y 100%)
3. If X accounts for none of the variation in Y (i.e., the regression line is perfectly horizontal), what would the r^2 value be?	r^2 = 0 (X does not predict Y at all)
4. Based on the residuals in the graphs shown, which one appears to be a better-fitting model? Why?	The graph on the left is a better model, since the residual values are much smaller (shorter distance to the predicted line) when compared to the residual values in the graph on the right
5. How does MSE contrast with how MAE residuals are measured?	MSE has residuals squared, summed, and then averaged. MAE takes absolute value of residuals, summed, then takes the average.
6. What is the benefit of using RMSE as opposed to MSE?	Because MSE requires squaring of units, it might be hard to interpret them. By using RMSE (the root mean square error), you can convert the units back to their original form, allowing for interpretation.

4. LOGISTIC REGRESSION (CLASSIFICATION)

1. Identify whether these dependent variables are binary, multinomial, or ordinal.	a. *"Favourite colour":* Multinomial b. *"How do you feel?":* Ordinal c. *"Type of tree":* Binary
2. What is the probability threshold?	The probability threshold is 0.5
3. What are the two categories the threshold is separating?	The threshold is separating groups for tumor type (i.e., malignant group vs. benign group)
4. Which graph has a better division of categories - the left or the right? Why?	The left graph has a better division. Ideally, you want to make the sigmoid the most "s"-shaped as possible, with a clear vertical middle part around the threshold. This indicates a clearer division of groups, wherein the probabilities are closer to 0 or 1 (true positives or true negatives). With the right graph, the sigmoid indicates the division isn't as clear (i.e., probabilities are not as close to 0 or 1 as possible), and there are many more probability values that are closer to the threshold value. When this happens, there is more room for misclassification (i.e., false positives or false negatives).

5. NAIVE BAYES CLASSIFIERS (CLASSIFICATION)

1. In your own words, what is the Bayes Theorem testing?	The probability of an event happening given another event has already happened
2. Note down some additional examples of the above applications of Bayes Theorem	Possible examples include: ● The probability of a stock price falling given another (different) stock price falls ● The probability of someone getting a positive medical test result given previous positive test diagnoses
3. Why would you perform Laplace smoothing?	So that in the event you have a value of '0' (a non-event) for a particular measure, the algorithm can run, as mathematically, you cannot divide a number by zero. By adding a very small number, you allow the algorithm to run, but because the value is small, it will not affect the overall probability.
4. If you wanted to analyse text documents, what type of naïve Bayes	Multinomial naive Bayes

distribution would you choose?	

6. DECISION TREES (CLASSIFICATION)

1. Using the Confusion Matrix below, calculate the following values:	a. *Accuracy* = (100+50)/(100+50+10+20) = 150/180 = 15/18 b. *Recall* = 100/(100+10) = 100/110 = 10/11 c. *Precision* = 100/(100+20) = 100/120 = 10/12 d. *False Positive Rate* = 20/(20+50) = 20/70 = 2/7
2. Match the ROC curves to the corresponding classification model where the probability threshold has a value of:	*a. (Threshold = 0.95): Curve A* *b. (Threshold = 0.05): Curve C* *c. (Threshold = 0.50): Curve B*
3. Assume all three curves come from models with thresholds all equal to 0.5. Match the ROC curve to the model where the AUC has a value of:	*a. (AUC = 0.5): Curve B* *b. (AUC = 0): Curve C* *c. (AUC = 1): Curve A*
4. *What are the advantages and disadvantages of using decision trees?*	Advantages: they have high accuracy, stability, are easy to interpret, and can map relationships between non-linear data. Disadvantages: they can be prone to overfitting / class bias, can lose information when using continuous data, and are sensitive to small changes in the data.

7. TIME SERIES ANALYSIS

1. In the model ARIMA(2,1,3), what are the values of q, d, & p?	$q = 3$; $d = 1$; $p = 2$
2. What is the difference between PACF and ACF?	*PACF* - used to determine the value of 'p', and measures the correlation between two data points with a given time lag after removing the effects of other time lags in between. *ACF* - used to determine the value of 'q', and measures how much a certain data point at a given time point is correlated with its previous time points' residuals.

3. What are the three steps of the Box-Jenkins Method?	1. *Model Identification* - Identifies autoregressions and seasonality to determine the differencing value and the time lag. 2. *Parameter Fitting* - Uses model fit procedures to determine the coefficients in the time series model. 3. *Model Evaluation* - Tests the residual errors to determine the temporality (amount, type) not captured by the model.

8. TEXT ANALYSIS

1. What are the differences between tokenisation, stemming, and lemmatisation?	*Tokenisation*: breaking down text / documents into smaller chunks, usually sentences or words *Stemming*: process that reduces words to their main root by removing common beginnings (affixes) or endings (suffixes) *Lemmatisation*: an extension of stemming that accounts for variations in spelling and usage in parts of speech when reducing words to their common roots.
2. What is Sentiment Analysis, and what are its limitations?	Sentiment analysis is analysing and classifying *opinions* (e.g., polarity, emotion, intent) in text data. This can be done at the word-, phrase-, sentence-, or document-level. Limitations include the inability for sentiment analysis to detect or account for the following without the use of custom coding or modelling: subjectivity/tone; context/polarity; sarcasm/irony; comparisons; presence of emojis; and defining neutral polarity
3. Refer to the image below: a. *What term has the lowest document frequency? The highest document frequency?* b. *Based on the information provided, which term is likely to be the most "informative"? Why?*	a. Lowest document frequency: "insurance"; Highest document frequency: "best" b. Term likely to be the most informative is "insurance" because it occurs the least out of many documents (or, has the highest TF-IDF value)

ADVANCED ANALYTICS: TECHNOLOGY & TOOLS

1. DATA WAREHOUSING

1. Define "data warehousing"	A large data repository that collects data usually from many (sometimes unrelated) sources, and sorts and cleans the data
2. Define "data repository"	The multiple ways in which you can collect and store data
3. Name and define two data storage tools in the Hadoop ecosystem	1. *Hadoop Distributed File System (HDFS)* - File system that is able to break down data into smaller blocks, and stores and distributes them across a cluster. When possible, uses MapReduce parallel processing to do so. 2. *Apache HBase* - Real-time read and write access for data stored in the Hadoop environment, and can handle databases with a vast amount of data. Does not rely on MapReduce to access data.
4. Name and define two data processing tools in the Hadoop ecosystem	1. *Hadoop MapReduce* - "heart" of Apache Hadoop, allows for large scalability across hundreds or thousands of servers in a Hadoop cluster. 2. *Yet Another Resource Negotiator (YARN)* - Revamped MapReduce functionality in Hadoop that allows for other tools to run in Hadoop. It separates resource management of clusters from scheduling/monitoring of jobs
5. Name and define three data access tools in the Hadoop ecosystem	1. *Apache Pig* - has a high-level scripting language that allows developers to write high-level code that can be translated to MapReduce program 2. *Apache Hive* - consists of a SQL-like language that allows developers to write and process high-level code that can be translated to MapReduce programs 3. *Mahout* - provides analytical toolset that directs Hadoop to analyse and produce meaningful results; provides executable java code to run algorithms like clustering, classification, and collaborative filtering

2. BASIC & ADVANCED SQL

1. What are System Functions? List 4 examples.	System functions are built in to the SQL system that can handle single-row and group functions. Examples: length(), power(), avg(), and round()

2. What are Window Functions? List 4 examples.	Window functions enable subtotal analyses (similar to ROLLUP and CUBE), but allow you to maintain granular row data as well. They do not modify columns, but create a new column with the output. They are *always* associated with the '*OVER*' clause. Examples: row_number(), rank(), percent_rank(), ntile()
3. What are User-defined Functions?	User-defined functions are custom functions created by the user; generally start with the keywords *'CREATE'* (e.g., 'create function')

OPERATIONALISING AN ANALYTICS PROJECT & DATA VISUALISATION TECHNIQUES

VISUALISATION

1. Describe the 4 main deliverables from an analytics project.	1. *Code*: for technical people in the production environment 2. *Technical specifications*: for implementing code 3. *Presentations for project sponsors*: high-level takeaways for stakeholders and/or executives, delivering the key messages, with clean, simple visualisations 4. *Presentations for analyst audiences*: describes any changes to business processes and reports to data scientists / technical staff, using technical graphs that convey important analytical details
2. What are the main differences in presentations for project sponsors versus technicians/analysts?	Presentations for project sponsors are meant for non-technical audiences, so they should be simple, and without any technical details. Generally, there should not be any discussion about the models used, and include information about how the findings can impact or help the business. Presentations for analysts are more concerned with the details. As such, they should include more technical graphs and information about the models, the variables, findings, and the code/technology used. They are less concerned with business impact, and more concerned with implementation or deployment in the business.
3. What are the key characteristics of visualisations for project sponsors versus technicians/analysts?	Visualisations for project sponsors should be clean, simple, high-level information, and using traits like colour to emphasize certain key points. Visualisations for analysts/technical audiences should have more detail and use more technical graphs that convey information about the statistics or models.