1) Challenge the accuracy of the below statements (agree/disagree) with short justification. [Answer should contain few bullet points of no more than 2 – 4 lines].
a) The roles of a Data Scientist and a Business Analyst are identical.
b) Training set should be larger than Test set. If so, write is the ideal split.

2) A company is planning to expand its business into new markets to increase their business. They currently have five types of products (A, B, C, D and E). The company feels that the new market is similar to their existing market. In the current market, the company has classified their customers into four segments (C1, C2, C3, & C4). The company found that there are approximately 10000 customers in the new market. The company has data from their existing customers and acquired data from the new market. Based on the above case scenario. Answer the following questions. Do not write generic answers. Your answers should revolve around the current business case. That is the ice cream company.
a) What kind of data science problem is this?
b) During our class sessions, we learned about various stages or steps in the data science process/methodology using a couple of case studies. For the case study described above, explain various stages in the data science process.
c) What type of model would be appropriate for this problem?

3) Consider the given data set for carrying out some data science activity.

Customer ID	Name	Age	City	Blood Type	Body Temperature
123	Aravind	25	Bangalore	A	37 deg C
235		22	CHENNAI	A	98.1 deg F
230	Rani Kannan	26	hyderabad	AB	97.5 deg F
222	Rahul	909	Cochin	O	37.5 deg C
123	Aravind	25	Bangalore	A	37 deg C

a) Identify the quality issues with this data set
b) Write any one method to rectify the issues identify in the dataset.
c) Identify the data type of the attributes Blood type and Body Temperature. Also mention whether it is a discrete attribute or a continuous attribute.

4) Using a case scenario in a particular domain, describe Exploratory and Explanatory analysis. Explain the circumstance (for the same scenario) under which you would use these two types of analysis?
a) Clearly describing what is exploratory analysis and explanatory analysis.
b) In the current domain, describe the circumstance where you would use exploratory vs. Explanatory analysis.

5) What is the key difference between the role of Data Scientist and Business Analyst?
a) Explaining role of Data Scientists and Business Analyst
b) Identifying the key difference

6) Raju measures the pressure of all tires coming into his garage and record the values. Unknown to him, his tire gauge is miscalibrated and adds 3 psi to each reading. Using the definition of noise used in the textbook, is this error introduced by the tire gauge considered noise? Answer “yes” or “no” and justify your answer in one line.

7) Describe a scenario where Data Scientists needs to be a bearer of bad news.
a) Just Identification of Scenario

b) Proper Explanation of the scenario

8) List the four outcomes of the Data Science project step where research goal and project charter are defined.

9) Answer the following.
https://ils.unc.edu/courses/2013_fall/inls613_001/INLS_613.midterm.Fall2013.pdf
a) Suppose we train a model to predict whether an email is Spam or Not Spam. After training the model, we apply it to a test set of 200 new email messages (also labelled) and the model produces the contingency table (confusion matrix) as below:
Radha hates seeing spam messages in her inbox! However, she doesn't mind periodically checking the 'junk' folder for messages incorrectlymarkedasspam. Seethadoesn'tevenknowwhere the 'junk' folder is, and she would prefer to see spam messages in her inbox than to miss genuine messages without knowing! What would be the appropriate Performance Measures for Radha and Seetha? Compute them and confirm whether they like the Classifier?

Answer:
Considering Positive -> spam, Radha has NO tolerance for seeing ANY spam messages (highest TPs) in her inbox as she can tolerate non-spam being classified as spam (False Negative) and going to junk folder. The correct Measure for her is: True Positive Rate (or Recall or Sensitivity) = TP/True Spam = 60/(60+0) = 100%. Hence Radha likes this classifier
For Seetha, she doesn't even know about junk folder -> she does not want any non-spam being classified as spam (which ends up in the junk folder). That means, her focus is on TN (True Non-spam) over all Actual Non-spams (FN+TN). The relevant measure for her is: Specificity (True Negative rate) = TN/Actual Non- spams(FP+TN) = 20/(120+20) = 1/7 = 14%
Hence Seetha may not like this classifier (as she expects 100% True Negativity or Specificity)

10) Nephrologists may treat hundreds to thousands of patients with end stage renal disease in their career. A Famous Hospital Chain in India have collected data on over 1 million ESRD patients. For those patients, they have data from every treatment, lab, medication, and assessment resulting in over 4 petabytes of data—that's over 1 million gigabytes of de-identified secured patient data. Imagine that they wanted to find the data from a historical patient who most closely matches the patient currently sitting in the dialysis clinic to inform the best precise personalized treatment. The amount of data is too large for any clinician, nurse, dietitian, social worker, or technician to search. Be a Data Scientist and based on different types of analytics that we discussed in the class, let the hospital management know what kind of analytics can be used in each of the cases below to improve the nephrologist's practice.
https://www.chegg.com/homework-help/questions-and-answers/question--05-subjective-question-hence-write-answer-text-field-given-nephrologists-may-tre-q66663248

a) How many patients were hospitalized last week? Descriptive
b) What percent of patients dropped home therapy in the last month? Descriptive
c) What are the average bone mineral metabolism (BMM) laboratory values for the patient population? Descriptive
d) Why do patients not meet BMM targets? Diagnostic
e) Which patients will have the highest risk of hospitalization next week. Predictive
f) Which patients are likely to switch from home therapy to in centre next month. Predictive
g) Next month's BMM values for each patient. Predictive
h) An extra treatment may help prevent the predicted fluid overload admission for this patient. Prescriptive
i) A home visit from a social worker may help prevent home therapy discharge for this patient. Prescriptive

j) A different BMM therapy may help this patient with medication adherence issues. Prescriptive

11) Netflix has released hundreds of Originals and plans to spend $8 billion over the next year on content. Creators of these stories pour their hearts and souls into turning ideas into joy for our viewers. The sublime art of doing this well is hard to describe, but it necessitates a careful orchestration of creative, business and technical decisions. Here we will focus on the latter two — business & technical decisions like planning budgets, finding locations, building sets, and scheduling guest actors that enable the creative act of connecting with viewers. Explore all the Potential Data Analytics that can be applicable for this scenario. Explain with the type of analytics applicable with suitable examples relevant for the related scenario. [4 marks]

https://netflixtechblog.com/studio-production-data-science-646ee2cc21a1

1) Descriptive Analytics – Statistical Analysis on Budget Spend for Future Plan.
2) Diagnostic Analytics – Historical viewing trends to inform content is consumed across a range of languages and markets. If a piece of content is more popular in a language A than language B, we may sequence our efforts for A before B
3) Predictive Analytics: For upcoming shows, this turns into the following data science problem: Compute Distance metric which include genre, language (both the original language of the content as well as the localized language), and whether the localized content was consumed as dubbed audio or as subtitles.to predict the per-language consumption for each show k months before it is released.
4) Prescriptive Analytics: Using Information from Predictive Analytics to understand and Plan to accommodate Future Trend of Customers in Localizing the Set/scenario of the Story and Model and portray the Guest actors accordingly.

12) Studies show that listening to music while studying can improve your memory. To demonstrate this, a researcher obtains a sample of 36 college students and gives them a standard memory test while they listen to some background music. Under normal circumstances (without music), the mean score obtained was 25 and standard deviation is 6. The mean score for the sample after the experiment (i.e With music) is 28. [3 marks]

a. What is the null hypothesis in this case? Explain Why?
b. After performing the Z-test, what can we conclude ____ ? Explain Why?

https://www.zarantech.com/blog/interview-questions/data-science-statistics-interview-questions-answers/

a) The null hypothesis is a generally assumed statement, that there is no relationship in the measured phenomena. Here the null hypothesis would be that there is no relationship between listening to music and improvement in memory
b) Let's perform the Z test on the given case. We know that the null hypothesis is that listening to music does not improve memory.
Alternate hypothesis is that listening to music does improve memory.
In this case the standard error i.e.
The Z score for a sample mean of 28 from this population is
Z critical value for $\alpha = 0.05$ (one tailed) would be 1.65 as seen from the z table.
Therefore since the Z value observed is greater than the Z critical value, we can reject the null hypothesis and say that listening to music does improve the memory with 95% confidence.

13) In a Dataset of persons in India, for the attribute ‘age’, many of the values are missing as follows. Determine the right Imputation method with justification and fill-in the missing values. [3]
(Gender, Age) [* : missing value]
(M – 42), (M - *), (M – 24), (M - *), (M – 36), (M – 57), (F – 32), (F - *), (F - *), (F – 18), (F - *), (F – 23)

https://www.google.com/url?q=https://www.chegg.com/homework-help/questions-and-answers/question-04-subjective-question-hence-write-answer-text-field-given--accuracy-metrics-s-fo-q62113766&sa=D&source=docs&ust=1640147799176515&usg=AOvVaw2fpJNOWtg5OwY1XtbtmzE5