1. Describe the 5V of big data analytics

Volume:huge volume of data(billions \* millios)

Variety: Complexity of data types and structures

Velocity: Speed of new data creation and growth

Value: Data value density is relatively low

Veracity:Accuracy and reliability of data

1. Desrcibe the difference between business intelligence and data science.

Concept: Data Science is a field that uses mathematics,statistics and various other tools to discover the hidden pattern in the data;BI is basically a set of technologies,applications and processes that are used by the enterprises for business data analysis.

Focus:The data science is focus on the future,and BI is focus on the past and present.Data science

Data: DS focus on structured as well as unstructured data. BI deals with structured data.

Method: DS make use of scientific method BI makes use of analytic method.

Complexity: DS has a higher complexity in comparison to bussiness intelligence.

1. Name the six phases of data analytics lifecycle and briefly describe each phases.
2. Explain why visualisation is important in exploratory data analysis.
3. Which R functions can be used to create and load .Rdata files
4. Explain false positive rate(FPR) and true positive rate(TPR) in classification
5. A woolworths store manager would like to study the purchase behaviour of its customers.The following figure shows the density plot of the purchase sizes. What visualisation technique can you recommend to reveal more structures that otherwise might be missed? Sketch the plot after the recommended technique is applied
6. Explain the purpose of hypothesis test. Suppose you are given a task to identify whether a new computational model produces higher accuracy on weather forecast. Define the null hypothesis and alternative hypothesis for this task.
7. Explain the following concepts:significance level,p-value,t-statistic,and confidence interval.
8. Describe the goal and procedure of analysis of variance(ANOVA)
9. Describe the procedure of k-means clustering
10. Describe the “within sum of squares(wss)” and its role in k-means clustering
11. Why we usually run the k-means clustering multiple times for a given k value?

What option in the function kmeans() is used to indicate the times?

1. Name and describe a clustering method that can handle categorical data
2. Following shows the transaction of five customers.Which itemsets satisfy the minimum support of 0.5?

Customer 1:(milk,egg,ham)

Customer 2:(milk,ham)

Customer 3:(egg,ham)

Customer 4:(milk ,egg,ham,butter)

Customer 5:(milk,ham,butter)

1. Describe the following concepts:support ,confidence,lift,and leverage.And use their relationship to explain the “straight line” patterns int the follwing figure
2. Explain the concept of Apriori property and describe the key steps of Apriori algorithm.
3. After training a linear regression model, you plot the residuals as follows. Describe what you can observe from this figure (1)
4. Decribe how logistic regression can be used for a classification task.(1)
5. If b3=-0.5isan estimated coefficient in a logistic regression modely=b0+b1x1+b2x2+b3x3,what is the effect on the odds ratio for each one unit increase in the value of x2 (2)
6. Describe the criterion used by decision tree to pick the most informative attribute (1)
7. Describe the two simplifications that convert Bayes’ theroem to Naive Bayes classifier(1)
8. As a data scientist,you are working on a classification project.It is fount that the dataset contains many correlated variables and most of them are categorical variables.

From logistic regression,decision tree,and naive Bayes,which classifier will you consider using

Justify your choice (2)

A company would like to monitor what is being said about its products in social media.The company is

Interested in (1) whether people mention its products and (2) what is being said,good or bad,Briefly describe your paln as a data scientist for this task (2)