SQL ANSWERS

```
1 - SELECT date(shipped_date)
     , COUNT(id) AS orderNumber
     , AVG(orderNumber) AS daily_total
  FROM Orders
 GROUP BY date(shipped_date)

2 - SELECT date(orderdate)
     , COUNT(id) AS orderNumber
     , AVG(orderNumber) AS daily_total
  FROM Orders
 GROUP BY date(orderdate)

3 - SELECT productName
FROM products
WHERE MSRP = (SELECT MIN(MSRP) FROM products);

4 - SELECT productName
FROM products
WHERE quantityinStock = (SELECT MAX(quantityinStock) FROM products);

5 - SELECT productName
FROM products
WHERE productName = (SELECT MAX(productName) FROM products);
6 - Select name
```

from Customer

where pay= (select max(pay) from Customer)

7 – select customername , customernumber from Customer where city = ‘Melbourne’

8 – Select CustomerName from Customer where CustomerName Like ‘N%’

9 - Select CustomerName from Customer where Phone Like ‘N%’ AND City IN (‘Las Vegas’)

10 - Select CustomerName from Customer where creditLimit < 1000 AND city in (‘Las Vegas’, ‘Nantes’, ‘Stavern’)

11 – Select ordernumber from orderdetails where Quantityordered<10

12 - SELECT ordernumber from orderdetails where where customername like ‘%N’

13 – Select Customer Name from customers Where Status=”Disputed”

15 – Select Checknumber from Payment where amount > 1000

MACHINE LEARNING ANSWERS

1 - C) between -1 and 1

2 - C) Recursive feature elimination

3 - C) hyperplane

4 - D) Support Vector Classifier

5 - A) 2.205 × old coefficient of 'X'

6 - C) decreases

7 - D) Random Forests provide a reliable feature importance estimate

8 – B and C

9 – A,C and D

10 – A And D

11 - IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12 - In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates

13 - Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines

14 - Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

15 - An advantage of using this method is that we make use of all data points and hence it is low bias. The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point

STATISTICS ANSWERS

1 - The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section

2 - There are two types of sampling methods: Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data

3 - Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true. Type I error is equivalent to false positive. Type II error is equivalent to a false negative

4 - Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve

5 - Covariance is a measure of how much two random variables vary together. Correlation is a statistical measure that indicates how strongly two variables are related

6 - Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables

7 - A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty

8 - The hypothesis actually to be tested is usually given the symbol H0, and is

commonly referred to as the null hypothesis. An alternative hypothesis that specified that the parameter can lie on either side of

the value specified by H0 is called a two-sided (or two-tailed) test.

9 - Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language

10 - To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1

11 - The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape

12 - The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR

13 - The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H 0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested

14 - Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is nCx·px·(1−p)n−x

15- ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables