SQL Questions

```
1. SELECT
       AVG(shippedDate) AS AVG_SHIP
  FROM
        orders
  GROUP BY
       orderDate
;
```

```
2. SELECT
      AVG(orderNumber) AS AVG_ORD
  FROM
      orders
  GROUP BY
      OrderDate
  ;
```

```
3. SELECT
      productName
  FROM
      products
  ORDER BY
        MSRP ASC
  LIMIT 1
  ;
```

```
4. SELECT
      productName
  FROM
      products
  ORDER BY
      quantityInStock
  LIMIT 1;
```

```
5. SELECT
      productName
  FROM
      products
  LEFT JOIN
      orderdetails
  ON
      products.productCode = orderdetails.productCode
  ORDER BY
      qunatityOrdered
  LIMIT 1;
```

```
6. SELECT
        customerName
FROM
        customers
LEFT JOIN
        payments
ON
        customers.customerNumber = payments.customerNumber
ORDER BY
        amount
LIMIT 1;
```

```
7. SELECT
        customerNumber,customerName
  FROM
        customers
 WHERE
        city=’Melbourne’
;
```

```
8. SELECT
        customerName
   FROM
        customers
   WHERE
        customerName = “N%”
   ;
```

```
9. SELECT
        customerName
   FROM
        customers
   WHERE
        phone  = “7%” AND city = “Las Vegas”
   ;
```

```
10. SELECT
        customerName
    FROM
        customers
    WHERE
        creditLimit > 1000
    HAVING
        city = “Las Vegas” OR “Nantes” OR “Stavern”
    ;
```

```
11. SELECT
         orderNumber
    FROM
         orderdetails
    WHERE
         quantityOrdered > 10
;
```

```
12. SELECT
         orderNumber
    FROM
         orders
    LEFT JOIN
         customers
    ON
         orders.customerName = customers.customerName
   WHERE
         customerName = “N%”
;
```

```
13. SELECT
         customerName
    FROM
         customers
    LEFT JOIN
         orders
    ON
         customers.customerName = orders.customerName
    WHERE
         status = “Disputed”
;
```

```
14. SELECT
         customerName
    FROM
         customers
    LEFT JOIN
         payments
    ON
         customers.customerNumber = payments.customerNumber
   WHERE
         checkNumber = ‘N%’
  AND
         paymentDate = ‘2004-10-19’
  ;
```

```
15. SELECT
          checkNumber
    FROM
          payments
    WHERE
          amount > 1000
;
```

Statistics Questions

1. The central limit theorem states that if you have a population with mean µ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed..The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.

2. A sampling is defined as a smaller set of data that a researcher chooses or selects from a larger population by using a pre-defined selection method.
Sampling methods :
1) Probability Sampling
   - Simple random sampling
   - Stratified sampling
   - Systematic sampling
   - Cluster sampling
2) Non-Probability Samping:
   - Convenience sampling
   - Judgemental sampling

3. In statistics, a Type I error means rejecting the null hypothesis when it's actually true, while a Type II error means failing to reject the null hypothesis when it's actually false.

4. Normal distribution means that the mean, median and mode are all equal and the plot forms a bell-curve.

5. Correlation is a statistic that measures the degree to which two variables move in relation to each other.
Corvariance is the measure of the relationship between two random variables. The metric evaluates how much the variables change together.

6. Univariate analysis is the analysis of one (“uni”) variable. Bivariate analysis is the analysis of exactly two variables. Multivariate analysis is the analysis of more than two variables.

7.Sensitivity (True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).

8 Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

9. **Quantitative data** is the type of data whose value is measured in the form of numbers or counts, with a unique numerical value associated with each data set. Also known as numerical data, quantitative data further describes numeric variables.

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods.

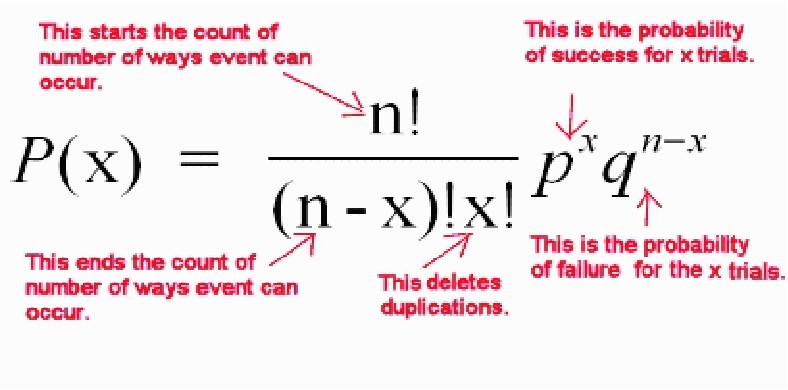
10. Range = Max. Value - Min. Value
 Interquartile range = Quartile3(Q3) - Quartile1(Q1)

11. A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.

12. Z-score method

13. The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

14.


This starts the count of number of ways event can occur.
This is the probability of success for x trials.
P(x) = n! / (n - x)!x! p^x q^(n−x)
This ends the count of number of ways event can occur.
This deletes duplications.
This is the probability of failure for the x trials.

15 **ANOVA** checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not. Another measure to compare the samples is called a t-test. When we have only two samples, t-test and ANOVA give the same results.

Machine Learning Questions

1. C
2. C
3. C
4. A
5. C
6. B
7. A
8. B,C
9. A,C,D
10. A,B,D
11. Outliers are the data points which different significantly from the rest of the data points.

    IQR = Q3 - Q1

    IQR includes the middles 50% data and removes all other data hence removing the outliers.

12. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

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. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.