SQL

1) result = cursor.execute('''SELECT * from movie''')
2) result = cursor.execute('''SELECT runtime FROM movie WHERE runtime = (SELECT MAX(runtime) FROM movie)''')
3) result = cursor.execute('SELECT title FROM movie WHERE revenue = (SELECT MAX(revenue) FROM movie)')
4) result = cursor.execute('''SELECT title FROM movie WHERE revenue = (SELECT MAX(revenue) FROM movie)''')
5) result = cursor.execute('''SELECT person.person_name, gender.gender, movie_cast.charactor_name, movie_cast.cast_order, movie.title FROM movie
   JOIN movie_cast ON movie.movie_id = movie_cast.movie_id
   JOIN gender ON movie_cast.gender_id = gender.gender_id
   JOIN person ON person.person_id = movie_cast.person_id''')

6)- result = cursor.execute('SELECT country.country_name, COUNT(production_country.movie_id) AS num FROM production_country JOIN country ON production.country_id = country.country_name GROUP BY production_country.country_id ORDER BY num DESC LIMIT 1')

7) result = cursor.execute('''SELECT genre.genre_name, movie_genre.genre_id

FROM genre JOIN movie_genre ON genre.genre_id = movie_genre.genre_id''')

8) –result = cursor.execute('SELECT language.language_name, movie_language.movie_id FROM movie_language JOIN language ON movie_language.language_id = language.language_id')

9)-

10 result = cursor.execute('''SELECT title FROM movie ORDER BY popularity DESC LIMIT 10''')

11) result = cursor.execute('''SELECT title, revenue FROM movie ORDER BY revenue DESC LIMIT 1 OFFSET 2''')

12) result = cursor.execute('SELECT title FROM movie WHERE movie_status LIKE "%rumored%"')

13) result = cursor.execute('SELECT MAX(movie.revenue), movie.title, country.country_name FROM movie JOIN production_country ON movie.movie_id = production_country JOIN country ON country.country_name = production_country WHERE country_name LIKE "%United_States_of_America%"')

14) result = cursor.execute('SELECT movie_company.movie_id, production_company.company_name FROM movie_company JOIN production_company ON production_company.company_id = movie_company.company_id')

15) result = cursor.execute('SELECT title FROM movie ORDER BY budget LIMIT 20')

MACHINE LEARNING

1) - R-squared is a method which is used to check the goodness of model which result in percentage
   - Residual Sum of Squares are the error which is resulting figure from actual value minus predicted value and then squared in order to Ignore the negative value
   - R-squared is used to check the goodness of model to know how well it is performing

2) Regularization method is used to overcome the problem of overfitting of linear regression model
   - Problem of overfitting arise when model fails to understand the pattern inside the data
   - There are two method lasso and ridge
   - Lasso also used for feature selection
3) Hyper parameter tunning is a method which helps in finding the best parameter in order to build the best model
4) In gradient decent if we choose large learning rate then model will take big step in order to reach lowest point or cost point and because of that it may miss the optimal point
5) No logistic regression model deals with only linear dataset if we apply logistic regression model on non linear dataset it will not perform better and prediction will be inaccurate and accuracy level will be low
6) K fold cross validation is method which is used to check the how well model can perform with different dataset it show maximum score or accuracy model can give on given dataset which is performed In conjunction with model
7) Adaboost assigns weights to each sample in the training set, Gradient boosting, on the other hand, doesn't assign weights to samples. Adaboost uses a range of weak learners, including decision tree, Gradient boosting typically uses decision trees as weak learners
8) Bias variance trade off is method which is used to find the right balance between bias and variance in order to build the generalize model for new data
   - Bias occur when model make overly simplistic assumption about data resulting being underfit or overfit
   - Variance occur when model is overly sensitive to noise or outliers, which results in overfitting
9) Out-of-bags a technique used in random forests for assessing the performance of the model without separate validation set. The data points that are not included in a particular tree's bootstrap sample are called the out-of-bag data for that tree.
   - OOB error is the prediction error of the model on the data points that were not used to train it, but are still included in the overall dataset
10) Bagging stands for Bootstrap Aggregating, and it is a technique where multiple models are trained on different subsets of the training data, where each subset is selected
    - Boosting technique where multiple models are trained sequentially, with each model trying to improve on the errors of the previous model

11) Ensemble method is method where multiple models are combined to improve the overall performance of a prediction task. ensemble can capture the strengths of different models and compensate for their weaknesses

12) The tree is built by splitting the data into smaller subsets based on the values of the input features, with the goal of minimizing the impurity
    - If a decision tree is unregularized, it will continue to split the data until each leaf node contains only one data point.
    - which can result in a very complex and overfit model that performs poorly on new data

13) Gini impurity is a measure of impurity used in decision trees and random forests to determine the quality of a split in the data.
    - The Gini impurity index ranges from 0 to 1
    - 0 indicates perfect purity
    - 1 indicates perfect impurity

14) 1) Linear kernel - The linear kernel is the simplest kernel, which assumes that the data is linearly separable. The linear kernel can be used when the data can be separated by a single hyperplane
    2) Radial basis function - RBF kernel can handle non-linearly separable data
    3) Polynomial kernel - polynomial kernel is another can handle non-linearly separable data

15)

TSS - TSS is a measure of the total variation in the dependent variable, It is sum of the squared differences between the actual y values and the mean of y

RSS – RSS is a measure of the variation in the dependent variable, It is calculated as the sum of the squared differences between the actual y values and the predicted y values

ESS - It is calculated as the difference between TSS and RSS.

Stats –

1) A) Mean
2) C) Frequency
3) C) 6
4) A) Normal Distribution
5) C) F Distribution
6) B) hypothesis
7) A) Null hypothesis
8) D) zero tailed
9) B) research Hypothesis
10) A) np