# Python Interview Questions

Turn in a document with the following 10 questions answered:

1. What is the difference between a list and a tuple in Python?
2. Is indentation required in Python?
3. What is a comment in Python and how would you write one?
4. What is a dictionary in Python?
5. How would you convert a string to all uppercase letters?
6. What are negative indices and why are they used?
7. How would you reference a specific item within a Python dictionary?
8. What does the .pop() method do to a list?
9. Given a list, how would you access the last 3 items in that list?
10. What is the enumerate function and why is it used?

# Pandas Interview Questions

Optional - turn in a document with the following 10 questions answered:

1. What is a Pandas DataFrame?
2. What is a Pandas Series?
3. How would you make an empty DataFrame in Python?
4. In Pandas, how would you find the number of unique values in a column?
5. What are some benefits of using Pandas?
6. How would you add a column to a pandas DataFrame?
7. How would you rename a column in a Pandas DataFrame?
8. How would you convert a Pandas DataFrame to a numpy array?
9. How would you save a Pandas DataFrame to a csv file?
10. Define GroupBy in Pandas.

# Data Visualization Interview Questions

Optional - turn in a document with the following 10 questions answered:

1. What makes data visualization good?
2. Why might you want to use data visualization?
3. What are some basic elements of a graph?
4. How would you visualize the relationship between two quantitative variables?
5. What is a heat map?
6. Name different types of graphs to visualize data.
7. What are data visualization tools that you are familiar with?
8. Name several data visualization libraries in Python.
9. Name several data visualization tools outside of Python.
10. What is the difference between a line graph and a scatter plot?

# Graphs

1. What are some ethical considerations when creating graphs?
2. What are some considerations for making graphs accessible for a diverse audience?
3. Why are pie graphs not the best option for sharing information, and what would be a better choice?
   When it comes to making graphs for explaining information to others, what are several design features you should avoid?
4. The type of visuals that you use depends on your audience. Give an example of some visuals that would be appropriate for a technical audience, but not appropriate for a non-technical audience?
5. Describe the pros and cons of using a boxplot vs a histogram for distribution.
6. Define correlation and explain the difference between a positive and a negative correlation.
7. What is the difference between a histogram and a bar chart and when should you use each?
8. How does the bin size affect the output of a histogram?
9. What is a donut chart and when is it used?

## Intro to Machine Learning Interview Questions

10. What is the difference between supervised and unsupervised learning?
11. What is the difference between regression and classification?
12. What is a target in supervised machine learning?
13. Explain the purpose of a train-test split?
14. What is machine learning?
15. What are some common applications of machine learning?
16. What are the benefits of using a pipeline?
17. What is data leakage and why is it a problem?
18. What is meant by "fitting" a model?
19. Describe the essential steps to pre-processing data for machine learning.

## Tree-Based Models Interview Questions

1. How do decision trees work?
2. What is the difference between a decision tree and a random forest?
3. What is bootstrapping?
4. What is bagging?
5. Why might a random forest be better than a decision tree?
6. What is training data and what is it used for?
7. What is a test set and why use one?
8. What is cross validation and why is it useful?
9. What are the main hyperparameters that you can tune for decision trees?
10. What are some ways to reduce overfitting with decision trees?

## Logistic Regression Interview Questions

1. What is the difference between a model parameter and a learning hyperparameter?
2. Name and briefly explain several evaluation metrics that are useful for classification problems.
3. Is logistic regression a regressor or a classifier?
4. What is an ROC curve and how is it interpreted?
5. How would you interpret a coefficient from a logistic regression model?
6. Given a confusion matrix, calculate precision.
7. Given a confusion matrix, calculate sensitivity.
8. What is the difference between precision and recall?
9. What is the difference between sensitivity and specificity?
10. Which is better - false positives or false negatives?

## Dimensionality Reduction Interview Questions

1. What is boosting?
2. What is the difference between bagging and boosting?
3. How does gradient boosting differ from traditional decision tree algorithms?
4. What hyperparameters can be tuned in gradient boosting that are in addition to each individual tree's hyperparameters?
5. How can you reduce overfitting when doing gradient boosting?
6. Why might multiple random weak trees be better than one long tree?
7. What is gradient boosting?
8. What is stacking?
9. What are the pros and cons of bagging?
10. What are the pros and cons of boosting?

## Clustering Interview Questions

1. Compare and contrast two different clustering algorithms. When would you use one vs the other?
2. Explain the difference between k-means and hierarchical clustering
3. Briefly explain how the k-means clustering algorithm works.
4. What is one common use case for k-mean clustering?
5. Why is it difficult to identify the "ideal" number of clusters in a dataset using k-means clustering?
6. What is one heuristic to select "k" for k-means clustering?
7. What is the difference between unsupervised and supervised learning?
8. What is the difference between k-means and k-nearest neighbors?
9. In a level of technical detail that's most comfortable to you, describe how DBSCAN works and in what cases would you choose it over K-means.
10. Suppose you're trying to cluster your observations into groups. However, you have too many features in your data. How would you decide which ones to use in the clustering model?

## Dimensionality Reduction Interview Questions

1. Why would you want to use dimensionality reduction techniques to transform your data before training?
2. Why would you want to avoid dimensionality reduction techniques to transform your data before training?
3. Name a popular dimensionality reduction algorithm and briefly describe it.
4. After doing dimensionality reduction, can you transform the data back into the original feature space? How?
5. How do you select the number of principal components needed for PCA?
6. When would you use PCA?
7. Why would you perform PCA even if you don't have a lot of features?
8. In what cases would you NOT use PCA?
9. What is the geometric interpretation of an eigenvector and eigenvalue?
10. What is the algebraic interpretation of an eigenvector and eigenvalue?

## Deep Learning Interview Questions

1. When would you use a recurrent neural network?
2. How do recurrent neural networks work?
3. What hyperparameters are available in LSTMs?
4. Explain neural networks to a non-technical coworker.
5. Imagine you are using a neural network to classify images, but your image dataset is very small. You need additional data to build a decent model but you do not have access to any other images. What would you do?
6. How does backpropagation work?
7. What types of neural networks have you worked with?
8. What types of datasets are neural networks best for?
9. Why are CNNs good for image data?
10. What are the drawbacks of data augmentation?

## SQL Interview Questions

1. Describe the different types of SQL joins.
2. What is the difference between WHERE and HAVING in a SQL query?
3. What is DBMS?
4. What is RDBMS?
5. What is a primary key?
6. What is a foreign key?
7. What is the difference between SQL and PostgreSQL?
8. Given a table with columns ‘a’, ‘b’, ‘c’, and ‘d’, write a query that selects the average of column ‘b’ greater than a certain number *n*.
9. Given a table with columns ‘a’, ‘b’, ‘c’, and ‘d’, write a query that selects all of the columns and orders them by column ‘a’ from highest value to lowest value.