

Part 1: Intro to Data Science

- What are the roles of data scientists
- What are the steps of a collaborative project in data science?
- How would you compare the importance of technical skills vs the domain knowledge
- Explain the “velocity” in Big data
- Is there some overlap between big data and data science? Explain your reasoning
- Why Python and R are the mostly used languages for data science?
- What is the difference between a data scientist and a statistician?

Part 2: Data Science Overview

- What are the differences between supervised, unsupervised, and reinforcement learning? Can you guess what kind of learning might be used for chatGPT ?
- What is A/B testing, and provide examples where it can be used
- What are the differences between linear regression and decision tree?
- Explain in brief: Checking for quality issues in data science– completeness, fidelity and consistency

Part 3: Machine Learning

- What are different examples of regression and classification? What are the different algorithms to use for these methods?
- Examples of supervised, unsupervised and reinforcement learning. What are different algorithms in each of them
- What is cross-validation, and why is it important
- What is hyperparameter optimization
- Explain the 5 steps of approaching an application in machine learning
- What is an objective function in machine learning, and what is its role
- What are the differences between eager and lazy learning algorithms? State with examples
- What are the differences between batch and online learning algorithms? State with examples
- What are the differences between parametric and nonparametric learning algorithms? State with examples
- What are the differences between discriminative and generative learning algorithms? State with examples
- Draw a ven diagram showing how Artificial intelligence, Machine Learning, Deep Learning and Data science interact

Part 4: Statistical Learning and Linear Regression

- What are they key differences between a population line and the least squares line ?
- How to interpret the R^2 value? What does it mean when 1 and 0 ?

- Solve by hand. You have the following data, where the size of an apartment in square foot (“sizeSQFT”) is the independent variable and the rent of the apartment is the dependent variable. You want to fit a line with the following equation:

$$rent = b_0 + b_1 * sizeSQFT$$

sizeSQFT	rent
810.5	1300
512.6	981
1120.8	1400
782.2	1210

Find b_0 and b_1

- Find the R^2 for the above table
- Assume, you fit a line, and the corresponding point in your lines are shown in the “rent_predicted” column. Compute the SSE

sizeSQFT	rent	rent_prediction
810.5	1300	1200
512.6	981	1000
1120.8	1400	1500
782.2	1210	1100

- Imagine, you find a regression line, which is flat (parallel to x axis). How would you interpret the result ?
- What is the interpretation of p-value in following table?

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	12.3514	0.6214	19.8761	0.0000
Newspaper	0.0547	0.0166	3.2996	0.0011

- What is one-way ANOVA test, and what is it used for ?
- What are the assumptions for linear regression? What is Multicollinearity
- When would you find multiple lines in a linear regression problem ? What is the interpretation when these lines are not parallel?

Part 5: Data Processing and Cleaning

- What is data and attribute?
 - What are different types of attributes? What is the difference between ordinal and nominal attributes?
 - What is graph data? Provide an example
 - What is noise and outlier in the data ? Provide examples
 - Explain different techniques in handling missing values for different types of attributes
 - What is the relation between correlation and covariance? Provide examples
-
- John Doe is recruiting a data scientist, and below is a table describing the age of the applicant, and whether or not the person is called for an interview

Age	Status
16	rejected
19	called
27	rejected
28	rejected
31	called
40	rejected
44	called
47	rejected

- Calculate the information gain if we split the age at 23
- Calculate the information gain if we split the age at 27.5
- Calculate the information gain if we split the age at 35.5
- Calculate the information gain if we split the age at 42



Which split is the best and why?

- The following table describes if a person is born in Houston, if the person supports Houston Astros, and if the person goes to UH. Compute the mutual information between
 - Born in Houston and supports Astro
 - Goes to UH and supports Astro

Born in Houston	Support Astros	Goes to UH
yes	no	yes
yes	yes	no
no	yes	no
yes	yes	no

- Explain curse of dimensionality. What are different techniques to reduce dimensionality. Why is it important to reduce dimensionality
- Explain PCA
- What are binarization and discretization?
- Why normalization is done on the data?
- Explain standardization normalization, and min-max normalization
- How would you determine if a categorical or numerical attribute is redundant?
- Compute the Chi-squared value for the following table and provide interpretation of your result

Observed frequencies

	 Female	 Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	43	49

- What is entropy-based binning

Part 6: Data Exploration

- Compute Percentile in a data
- Compute variance and standard deviation

- How is AAD measured, and when it is used
- Compute interquantile range in a data
- How histograms are drawn? What happens to the width of the beans if you want to increase or decrease the number of beans
- Draw a box-plot based on the given data
- What each of the bars mean in the box plot
- Why contour plots are needed
- Given the results on different models in different validation set, how would you choose the best model?
- Practice on EDA from different datasets from kaggle, UCI machine learning repository, zenodo etc.
- Find data from <https://datasetsearch.research.google.com/>

Part 7: Data Similarities and Distances and KNN

- What is similarity, dissimilarity and proximity
- How data matrix and dissimilarity matrix are built?
- What is qualitative and quantitative data
- Why mean absolute deviation is a better approach to handle outliers in a data than standard deviation
- Compute dissimilarity matrix from a provided dataframe using Euclidean and Manhattan distance
- Why logarithmic transformation is useful
- Compute Jaccard coefficient from a given dataframe
- How dissimilarity is computed from binary variables
- How dissimilarity is computed from nominal variables
- How dissimilarity is computed from ordinal variables
- Verify the dissimilarity matrix
- Compute the Minkowski distance from the given data
- Compute the norm
- Compute the Mahalanobis distance between two coordinates
- What is cosine distance? Provide some use cases
- What is KNN and what kind of learning happens in KNN
- State the basic requirements for KNN
- What are the pros and cons of using a KNN classifier
- What is the strategy to find “k” in KNN
- How to handle the following issues in KNN? → Attributes with large range, correlated attributes, symbols, expensive in testing, storage requirements, curse of dimensionality
- Apply KNN to predict some unseen variable, given a dataframe
- Try to perform KNN from 2-d representation of the data

Part 8: Numpy and Pandas

- Study Lab 1 to Lab 4 rigorously
- Perform EDA and data analysis from the following resources: Kaggle, zenodo, UCI machine learning repository, dataverse, dataset search in google
- Learn basic numpy operation: inverse, transpose a matrix, resize a matrix, find determinant of a matrix, matrix operations like add, summation, multiplication, finding eigen vectors and eigen values
- Learn important pandas operation: creating dataframe in different ways (from list, dictionary, series, tuple etc), how to present summary data with info, describe methods, how to perform basic plots, how to do value_counts, merging, grouping, slicing
- Learn plot operations rigorously from matplotlib and Seaborn