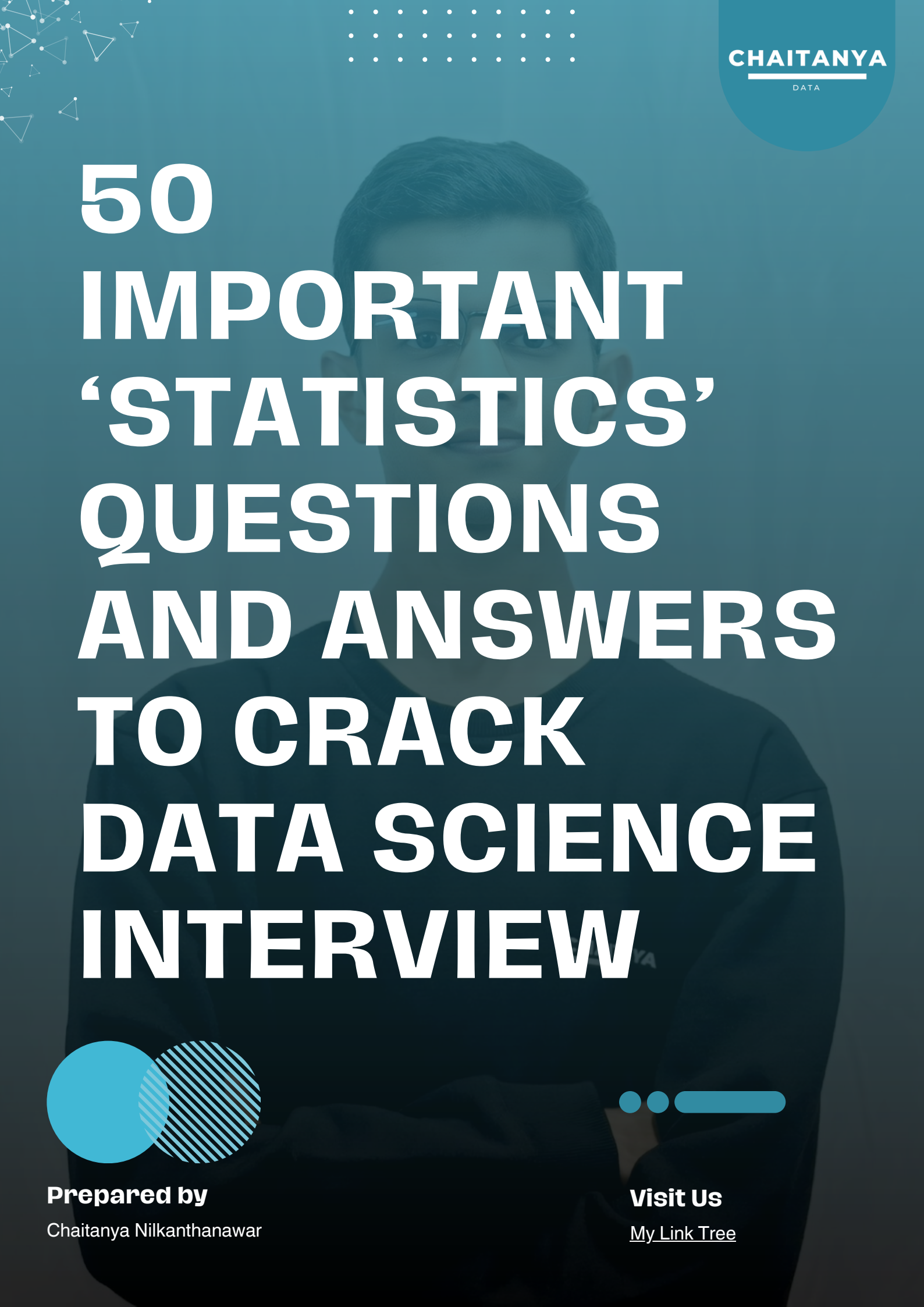
CHAITANYA
DATA

50 IMPORTANT 'STATISTICS' QUESTIONS AND ANSWERS TO CRACK DATA SCIENCE INTERVIEW



Prepared by
Chaitanya Nilkanthanawar

Visit Us
My Link Tree

CHAITANYA
DATA

General Questions

1. What is Statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

2. What is the difference between Descriptive and Inferential Statistics?

Descriptive statistics summarize and describe the features of a dataset, while inferential statistics make predictions or inferences about a population based on a sample.

3. What is a Population in Statistics?

A population is the entire group that you want to draw conclusions about.

4. What is a Sample?

A sample is a subset of the population, selected for analysis to make inferences about the population.

5. What are the different types of Sampling Methods?

Simple random sampling, stratified sampling, cluster sampling, systematic sampling, and convenience sampling.

CHAITANYA
DATA

General Questions

6. What is a P-value?

The p-value is the probability of observing the data, or something more extreme, if the null hypothesis is true.

7. What is Hypothesis Testing?

Hypothesis testing is a statistical method that uses sample data to evaluate a hypothesis about a population parameter.

8. Explain the Central Limit Theorem (CLT).

The CLT states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's distribution.

9. What is a Confidence Interval?

A confidence interval is a range of values, derived from the sample data, that is likely to contain the value of an unknown population parameter.

10. What is the difference between Type I and Type II Errors?

Type I error occurs when the null hypothesis is true, but we reject it. Type II error occurs when the null hypothesis is false, but we fail to reject it.

CHAITANYA
DATA

General Questions

11. What is a t-test?

A t-test is used to determine if there is a significant difference between the means of two groups.

12. What is ANOVA?

ANOVA (Analysis of Variance) is a statistical method used to compare means among three or more groups.

13. What is the difference between a Z-test and a t-test?

A Z-test is used when the sample size is large and population variance is known, while a t-test is used for smaller sample sizes or when population variance is unknown.

14. What is a Normal Distribution?

A normal distribution is a bell-shaped frequency distribution curve where most of the data points are concentrated around the mean.

15. What is Skewness?

Skewness refers to the asymmetry in the distribution of data. Positive skew means a longer tail on the right, negative skew means a longer tail on the left.

CHAITANYA
DATA

General Questions

16. What is Kurtosis?

Kurtosis is a measure of the "tailedness" of the probability distribution. High kurtosis means heavy tails, while low kurtosis means light tails.

17. Explain Variance and Standard Deviation.

Variance measures the spread of the data points around the mean. Standard deviation is the square root of variance and represents the average distance from the mean.

18. What is the Law of Large Numbers?

The law of large numbers states that as the size of a sample increases, the sample mean will get closer to the population mean.

19. What is the difference between Correlation and Causation?

Correlation indicates a relationship between two variables, while causation indicates that one variable causes a change in another.

20. What is a Chi-Square Test?

A Chi-Square test is used to determine if there is a significant association between two categorical variables.

CHAITANYA
DATA

General Questions

21. What is a Regression Analysis?

Regression analysis is a statistical technique for Modeling and analyzing the relationship between a dependent variable and one or more independent variables.

22. What is Multicollinearity?

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine their individual effects.

23. What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of variation explained by the independent variables in the model. Adjusted R-squared adjusts for the number of predictors in the model, providing a more accurate measure.

24. What is the difference between Parametric and Non-Parametric tests?

Parametric tests assume underlying statistical distributions in the data, while non-parametric tests do not assume any specific distribution.

CHAITANYA
DATA

General Questions

25. What is a Bayesian Approach?

The Bayesian approach incorporates prior knowledge along with the current evidence to update the probability of a hypothesis being true.

26. What is a Null Hypothesis (H0)?

The null hypothesis is a statement that there is no effect or no difference, and it is the hypothesis that researchers typically try to disprove.

27. What is an Alternative Hypothesis (H1)?

The alternative hypothesis is a statement that there is an effect or a difference, and it is what researchers typically try to support.

28. What is a One-Tailed Test?

A one-tailed test is used when the direction of the test is specified, such as testing whether a parameter is greater than or less than a certain value.

29. What is a Two-Tailed Test?

A two-tailed test is used when the direction of the test is not specified, meaning we are testing for any difference from the null hypothesis, either higher or lower.

CHAITANYA
DATA

General Questions

30. Explain the concept of p-hacking.

P-hacking refers to manipulating data or statistical analyses until non-significant results become significant, often leading to false positives.

31. Explain the concept of Overfitting in a statistical model.

Overfitting occurs when a model is too complex and captures noise in the data rather than the underlying trend, leading to poor generalization to new data.

32. Explain the concept of a Confidence Level.

A confidence level represents the proportion of times that the confidence interval will contain the true population parameter if the experiment is repeated multiple times.

33. What is the F-Statistic?

The F-statistic is used in ANOVA and regression analysis to test if the variances between groups are significantly different.

34. What is Heteroscedasticity?

Heteroscedasticity refers to the circumstance in which the variance of the residuals or errors is not constant across all levels of an independent variable.

CHAITANYA
DATA

General Questions

35. What is Homoscedasticity?

Homoscedasticity means that the variance of the residuals is constant across all levels of the independent variable.

36. What is a Log Transformation?

Log transformation is used to stabilize variance, make data more normal distribution-like, and improve the interpretability of a model.

37. What is a Permutation Test?

A permutation test is a non-parametric method that tests the null hypothesis by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

38. Explain the concept of Bootstrapping.

Bootstrapping is a resampling technique used to estimate the distribution of a statistic by sampling with replacement from the original data.

39. What is the significance of the p-value threshold (e.g., 0.05)?

A p-value threshold (e.g., 0.05) is commonly used to determine the statistical significance of a test. If the p-value is below the threshold, the null hypothesis is rejected.

CHAITANYA
DATA

General Questions

40. What is the purpose of the Likelihood Function?

The likelihood function represents the probability of the observed data as a function of the parameters of a statistical model.

41. What is an Outlier?

An outlier is a data point that is significantly different from the other data points in a dataset, potentially indicating an anomaly or error.

42. How can you detect Outliers?

Outliers can be detected using methods like the Z-score, IQR (Interquartile Range), and visualization techniques such as box plots.

43. What is a Quantile?

Quantiles are points in a dataset that divide the data into equal-sized intervals. Common quantiles include quartiles (four parts), percentiles (hundred parts), etc.

44. What is the purpose of a Box Plot?

A box plot is a graphical representation of the distribution of a dataset that shows the median, quartiles, and potential outliers.

CHAITANYA
DATA

General Questions

45. Explain Simpson's Paradox.

Simpson's Paradox occurs when a trend appears in different groups of data but disappears or reverses when the groups are combined.

46. What is the difference between Continuous and Discrete Data?

Continuous data can take any value within a range, while discrete data can only take specific, separate values.

47. What is a Time Series?

A time series is a sequence of data points typically measured at successive times, spaced at uniform time intervals.

48. What is Autocorrelation?

Autocorrelation is the correlation of a time series with a lagged version of itself, indicating how the current value is related to past values.

49. Explain Cross-Validation.

Cross-validation is a technique for assessing how a model generalizes to an independent dataset by partitioning the data into training and validation sets multiple times.

CHAITANYA
DATA

General Questions

50. What is the A/B Testing?

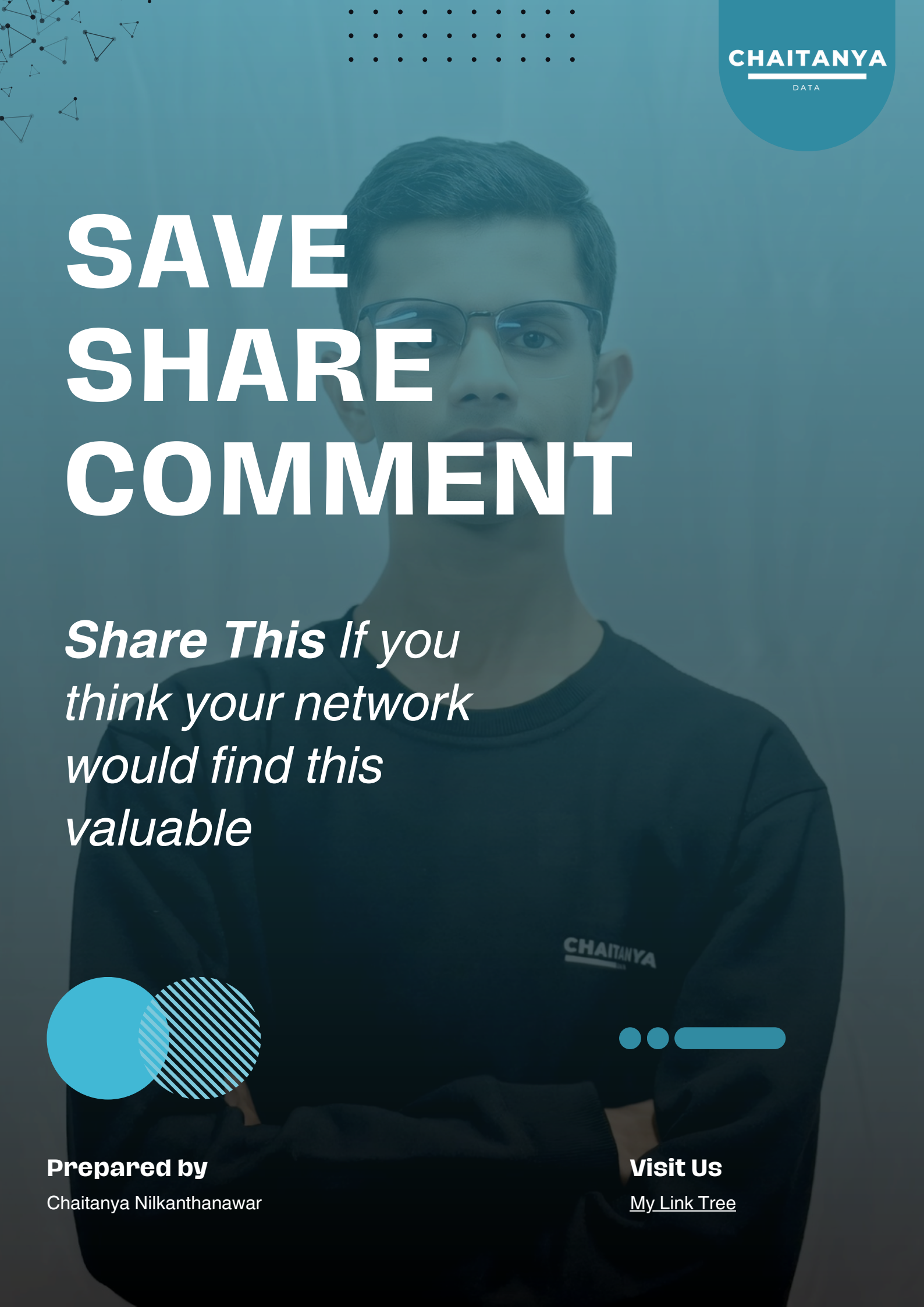
A/B testing is a statistical method used to compare two versions of a webpage, app, or feature to determine which one performs better.

CHAITANYA
DATA

Important Note

I hope you like my "50 Important 'Statistics' Questions And answers to crack Data Science Interview" document. I honestly tell you, it took me 6 months to collect these types of questions and answers from the 'FAANG' Companies (Facebook, Amazon, Apple, Netflix, and Google) and many other MNC companies. Do save this document and also share it with your friends.

These questions and answers cover a broad range of topics and scenarios that a Data Scientist / Data Analyst might encounter. Preparing thoroughly will help you demonstrate your knowledge, skills, and experience during your interviews.
Good luck!



Prepared by
Chaitanya Nilkanthanawar

Visit Us
My Link Tree