

What's Data analysis?

- => - The process of interpreting data to extract useful information.
- If involves cleaning, transforming & modelling data to gain insights & support decision making.

① What's difference between data mining & data profiling?

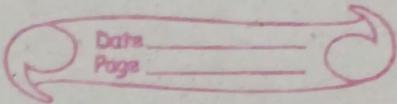
=>

Data mining

- 1) process of finding resultant information which hasn't found before
- 2) It's way in which raw data turn into valuable information

Data profiling

- 1) Done to access dataset for uniqueness, consistency & logic
- 2) cannot identify incorrect / inaccurate data values.



Q.2 Define data wrangling in data analytics?
⇒ (Data munging)

- The process of cleaning, structuring? enriching the raw data into desirable usable formate for better decision making
- main purpose to making raw data usable.

steps: ① Discovering ③ Cleaning ⑤ Validating
 ② Structuring ④ Enriching ⑥ Publishing

1) Discovering

- Data need to understand more deeply.
- Wrangling need to be done in some manner, based on criteria which divide data accordingly.

2) Structuring

- In most case data isn't in any form.
- Data need to be restructured in a proper manner.

3) Cleaning

- Inappropriate data is cleaned here.
- Null values are changed removed.

4) Enriching:

- What's in the data & strategies about how other additional data might augment it.

5) Validation

- To verify data quality, consistency & security.

6) Publishing

- Analysts prepare reusable data that uses further down the line that's is its.

Q.3 Key difference between data analyst & data mining?

- ⇒ - Data analysis involves process of cleaning, organizing & using data to produce meaningful insights.
- Data mining is used to search for hidden pattern in data.
- Data Analysis is used to produce results while data mining for something unique.

Q. What's data validation?

⇒ - The process involves determining accuracy of data & quality of source as well.

(1) Data screening - use models to ensure accuracy of data & no redundancy.

(2) Data verification - If redundancy available, evaluate results based on multiple steps & then calls taken to insure presence of data items.

Q. How do you know if data model is performing well or not?

⇒

- A data model can be evaluate using several metrics such as accuracy, precision, recall & F1 score for classification model & Root.

- Root mean square error & R-squared for regression model.

- Additionally, it's important to compare performance of modal against a human performance (baseline).

- Also need to evaluate data models on unseen data, check if it's generalizing well.

Q. Explain Data cleaning in brief.

- ⇒ • The process of identifying & correcting inaccuracies, inconsistency & missing data in dataset,
• Handling missing values, outliers, duplicate data & data formatting to improve quality of data & make it ready for analysis.

Q. what are some problems that working data analysis might encounter?

⇒

When working with data analysis, some common problems may be encountered includes.

① missing/incomplete data

: make difficult to draw accurate conclusion from the data

② inconsistent data formatting

: can make it difficult to effectively

organize & analyze data.

③ outliers: These can be skew results of analysis & lead to inaccurate conclusion.

④ non-representative data: If sample data of analysis doesn't represent of population, it can lead to inaccurate data.

⑤ Data bias: If data selected in bias way, lead to inaccurate conclusion.

⑥ scalability & computational power

⑦ privacy & security

⑧ Interpreting results

Q. what is data prefilling

\Rightarrow - To analyse all entities present in data to greater depth.

- provide higher accuracy based on data & its attributes such as datatype, occurrence & more

Q. what're scenarios that could cause a model to be retained?

⇒ It depends on several factors.

① High accuracy: if model has high level of accuracy compare to other model, it's likely to be retained.

② Low error rate: If model has low error rate, then likely to retained.

③ Good performance on unseen data: when model generalised with new data, it's likely to be retained.

④ Business value: if a model provide valuable insights, then likely to retained.

⑤ Robustness: If able to robust to small variation of data, then retained.

⑥ Easy to Interpret & explain: likely to retained.

A model that's retained consistently monitored & update to make sure its performing well.

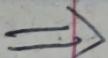
Q. what're prerequisites to become data analyst?



To become proficient in data analyst one should have strong analytics, problem solving skills.

- mathematics & programming language like R, Python / SQL
- knowledge of ML techniques
- strong communication skills & deep understanding of business domain.

Q. what're top tools used to perform data analysis?



- 1) R: open source programming language for statistics & data visualization
- 2) Python: Another open source programming language that widely used for data analysis & visualization as well as machine learning.

- 3) SQL - programming language used for managing & manipulating databases. widely used for data extraction & cleaning.
- 4) Excel: most widely used tool for data analysis, excel is most powerful tool for data manipulation & visualization.
- 5) Tableau / Power BI : Data visualization tool that allows to create interactive & visually appealing charts & dashboard.
- 6) Rapid miner : platform for data science & ml, allows perform data preparation, visualization, modelling & deployment.

Q. What's an outliers?

⇒ - An outlier data point which is significantly different from other data points in a dataset.

- Outliers can have significant impact on result of data analysis & should be examine carefully.

Ex, when we are analyzing height of a group of people, we find the one person's height is much taller than others, that's an outliers.

- Using boxplot, we can find outliers.

- univariate - A data point that's significantly different from other data points in one variable.

- multivariate - A data point is significantly different from other data points in multivariable.

Q. How can we deal with problems that arises when data flows in from variety of sources?



Dealing with data from multiple source can be challenging but it can be done by following these steps

① Data integration - combining data into a single source.

② Data cleaning - Identifying & cleaning incorrect data, inconsistent & missing data in dataset.

③ Data transformation

- Converting data into a format for analysis

④ Data Governance - Implementing policies & consistency & protected from

unauthorized access.

Q. What're some popular tools used in big data?

⇒ There are many tools that're used to handle a Big data. Some of the most popular ones are follows.

- ① Hadoop
- ③ Scala
- ⑤ Flume
- ② PySpark
- ④ Hive
- ⑥ Mahout

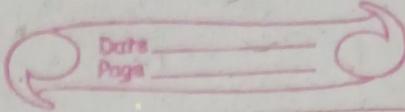
Q. What is the use of pivot table.

⇒ An pivot table is a tool in data analysis to summarize & organize large amount of data by grouping it into categories & calculating the total, average or other summary statistics of the group.

Q. Explain KNN imputation method, in brief.

⇒ (i) KNN is an supervised learning model
(ii) Required selection of no. of nearest neighbours & distance metrices at same time.

(iii) Ability to predict both discrete &



Continuous attribute of dataset.

iv) A distance function is used to find similarity between two or more attributes, which will help in clustering analysis.

v) a) Classification - common class among K-nearest neighbours is chosen.

b) Regression - average value among K-nearest neighbours is chosen as prediction.

Q. what are top Apache framework used in distributed computing?

⇒ ① Hadoop - open-source framework for distributed storage & processing of large dataset using Map-Reduce programming model.

② Spark: An open-source, general purpose clustering computing framework for big data processing

Q. What's Hierarchical clustering?

- ⇒ - A method of clustering in which data objects are organized into hierarchy of clusters.
- Starts by treating each data point as separate cluster & then iteratively merge clusters.
- Final results representation of clusters, called dendogram.

Q. What're steps involve when working with data analysis?

⇒ ① Data collection & preparation

: Gathering & cleaning of data

② Data exploration visualization

: Analyzing the data to understand its characteristics.

③ Data modelling: Building models to extract insight from data.

④ Evaluation: The performance of models & refining them is necessary.

⑤ Deployment: The models in production environment & monitoring their performance over time.

Q. Can you share some of statistical methodologies used by data analysts?



① Regression Analysis

: To identify relationship between a dependent variable & one or more independent variables

② Time series analysis (Hidden Markov model)

: used to study the trends & patterns of data over time

③ Principal Component Analysis

: used to reduce the dimensionality of data by identifying the underlying structure

④ Cluster analysis

: used to group similar data points together in order to identify patterns & trends.

⑤ Decision tree analysis

: used to analyse decision-making process by breaking down a problem into smaller & smaller sub-problems

⑥ Survival Analysis

: used to analyse time-to-event data & estimate probability of an event occurring over time.

⑦ Bayesian Analysis

: statistics method that used prior beliefs & data to update & make inferences about hypothesis.

⑧ Discriminant Analysis

: used to identify which variables discriminate between two or more groups.

⑨ non-parametric Statistical

: used to when underlying data distribution is not known.



Q. what is Time Series Analysis (TSA) ?

\Rightarrow - Time Series Analysis (TSA) is statistical method used to study the trends & patterns of data over time.

- involves the presence of the data at particular intervals of time.

Q. where is Time Series Analysis used?

\Rightarrow Since Time Series analysis (TSA) have wide scope of usage, can multiple domain

- ① Statistics ② signal process
- ③ Economics ④ weather forecasting
- ⑤ Astronomy ⑥ Astrology

Q. what're some of properties of clustering algorithms?

- ⇒
- ① Flat or hierarchy
 - ② Iterative
 - ③ Disjunction

Q. what's collaborative Filtering?

⇒ A method of making recommendations by looking at preferences of users similar to you.

- If User post behaviour & preferences of users to predict what you might like.

Ex, when browser through e-commerce site,
a section called "Recommendation for you" present.

- This is done using browsing history alongside analysis the previous purchase & collaborative filtering.

Q. what're types of Hypothesis testing used today's?

=> ① A-B testing: Used to compare two version of product or campaign to determine which one perform better.

② T-test: used to determine if there is a significant difference between two groups.

③ Chi-square: used to determine if there is significant associate between two (categorical) variables.

④ Anova: Analysis is conducted between the mean values of multiple groups

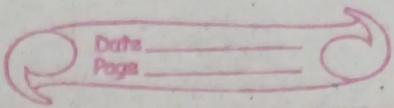
⑤ Z-test: used to determine if mean of population is different from a specific value.

⑥ paired t-test: used to compare mean related or paired groups.

Q. what're some Data validation methodologies use in data analysis?

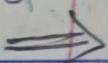
=> ① Field-level validation

: validation is done across each of field to ensure there are no error in data entered by user.



- ② Form-level validation: Here, validation is done when user completes working with form but before information being saved.
- ③ Data saving validation: Form of validation takes place when file or database record is being saved.
- ④ Search criteria validation: Kind validation is used to check whether valid results are returned when user is looking for something.

Q. what is K-means algorithm?



- K-means algorithm is a unsupervised algorithm.
- K-means, Algorithm clustering data into different sets based on how close data points to each other.
- no. of clusters 'K' in 'K'-means algorithm
- tries to make good amount of separation between each of clusters.
- The clusters doesn't have any label.

Q. what's difference between recall & true positive rate?

=>

Recall & true positive rate, Both are totally identical.

$$\text{Recall} = (\text{true positive}) / (\text{true + false pos})$$

(The proportion of actual positive observation that're correctly identified).

Q. what are ideal situation in which z-test & t-test can be used?

=>

- A z-test should be used when sample size is large & population standard deviation is known (greater than 30)

- A t-test should be used when sample size is small & population standard deviation is unknown.

Q. why is Naive Bayes called Baye's Naive?

=> - It makes strong & unrealistic assumption that features in dataset are independent of each other, but in reality they may be correlated which can affect accuracy of model.

- Despite this assumption, it's still a powerful & widely used algorithm for classification task.

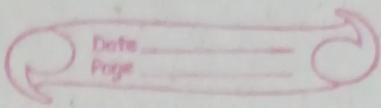
Q. What's simple difference between Standardized & unstandardized coefficient?

- => • Standardized coefficient: interpreted based on their standard deviation values.
• unstandardized coefficient: measured based on actual value present in dataset.

Q. How to detect outliers?

=> Outliers can be detected with various statistical methods, such as:

- ① visualization technique: (Box plot & scatter plot) To identify observations that are far away from majority of data points.
- ② Using statistical measures as standard deviation & interquartile range that observation that fall outside a certain range.
- ③ Using ml like clustering & density based methods to identify observation that are distinct from main clusters.



Q. Why KNN preferred when determining missing numbers in data?

⇒ KNN is preferred when determining missing no. of data because it's finds similarity between datapoints & based on similarity it can estimate missing value which closest to it.

Q. How one can handle suspicious or missing data in dataset while performing analysis?

⇒ ① Removing variables or observations that contain missing data, if represents small proportion of total data.

② Imputing missing value using method like mean, median/mode imputation, KNN, regression imputation.

③ Using advance techniques like data mining, machine learning & statistical method to handle missing data.

④ Creating a new category for missing values, in case if missing values are not missing at random.

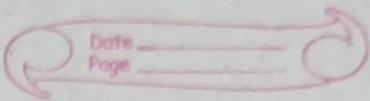
⑤ using technique like data scrubbing, data cleaning, data validation

Q. what is simple difference between PCA & Factor analysis (FA)?

- ⇒ - PCA technique to reducing dimensionality of a data set by identifying a set of uncorrelated variable (principle).
- Factor Analysis aim to identify factors affecting relationship between set of observation sets.

Q. How it's beneficial to make use of version control? (github)

- ⇒ 1) Allowing multiple people work on same project simultaneously.
2) creates history of all changes made to project, so previous version can easily retrieved by needed.
3) It makes collaboration & code review easily by allowing multiple people to make suggestions & approve changes before added to project.
4) Helps In a disaster recovery, if you crashes or you get data get lost, version control system to recover previous



version of code.

Q. what're future trends in data analysis?

→ Future Trends in data analysis increase the use of machine learning & artificial intelligence, as well as growth in Internet of Things which generate more data than ever before, & integrate with data analysis, other technologies blockchain & virtual reality.

Q. why're you applying for data analysis role in our company?

⇒

- I have hands on data related field during college time & I enjoy it more than development.

- Also, our company has good reputation, it will help to use my skills, as well as technology changes quickly to drive business decision & planning in right direction.

Q. Can you rate yourself on scale of 1-10 depending on your proficiency in data analysis?

- ⇒ - I would rate myself 7 out of 10 in terms of proficiency in data analysis
- I have good understanding of concepts & techniques used in data analysis
- But always looking to learn & improve my skills.

Q. What is your plan after joining for data analysis?

- ⇒
 - ① Privacy concern may arise if data contains sensitive information.
 - ② Cost of collecting, cleaning & storing data can be high
 - ③ Over-reliance on data & analysis can lead to neglecting human intuition & expertise.

Q. What skills should a successful data analyst possess?

- ⇒ A successful data analyst should possess a combination of technical skills as Statistics, programming language

Data visualization tools.

- Soft skills or problem solving communication & ability to work well in a team.

Q. Why do you think you're right fit for data analyst role?

⇒ I am a good fit for data analyst because I have necessary technical skills like problem-solving, communication & ability to work well in a team. I am tends my skills & take company to new heights.