How to face Data Science Interviews

By Varun Raste ,
Course Name : Data Science interview preparation ,Campus X

Interview Process

- Most of the companies take 3-5 rounds on an average for selecting a candidate.

- In different rounds , candidates will be tested on different skills
  - Ability to code / Technical skills
  - Ability to understand product / business
  - Story telling skills / communication skills
  - Behaviroral skills
  - Product Management skills

- Initial round is usually a coding round , it can be an online test or else could be a live technical test where coding skills related to tech stack like Python , Pyspark , SQL etc are checked.

- The next 1-3 rounds are usually the blend of technical & management skill checking , as rounds progresses candidates get interviewed by higher management so nature of questions shift to role of data science in product development based questions.

- Ultimate rounds are usually with the HR or round with the senior most authority. In these rounds basic aim is to assess intent of candidate , understanding his / her values , general personality is assessed.

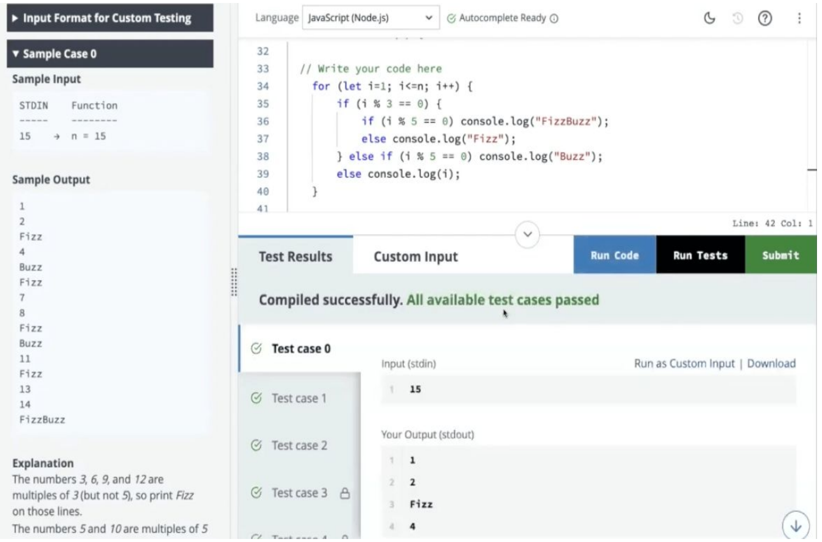
Coding Round

- This is an preliminary & eliminatory round where candidate is tested on technical skills such as Python , SQL , pyspark etc

- This round could be conducted on platforms like Hacker earth , Hacker rank or company's own platform as well. It is good to have basic familiarity with the platforms.

- This round could be MCQ based (where a code snippet will be provided to you and expected output needs to be selected from multiple choices) or a Coding based (where candidate has to code the things from scratch)

- This round could also be hosted with a panelist where he / she can assign candidate some task & candidate has to complete it while sharing his / her screen. Sometimes these interviews could be tricky as googling might not be allowed. So familiarity with basic syntaxes is expected.



Sample Input

```
STDIN    Function
-----    --------
15    →  n = 15
```

Sample Output

```
1
2
Fizz
4
Buzz
Fizz
7
8
Fizz
Buzz
11
Fizz
13
14
FizzBuzz
```

Explanation

The numbers *3, 6, 9,* and *12* are multiples of *3* (but not *5*), so print *Fizz* on those lines.

The numbers *5* and *10* are multiples of *5*

Language: JavaScript (Node.js)

```
// Write your code here
  for (let i=1; i<=n; i++) {
      if (i % 3 == 0) {
          if (i % 5 == 0) console.log("FizzBuzz");
          else console.log("Fizz");
      } else if (i % 5 == 0) console.log("Buzz");
      else console.log(i);
  }
```

Line: 42 Col: 1

Compiled successfully. All available test cases passed

Test case 0

Input (stdin)

```
15
```

Your Output (stdout)

```
1
2
Fizz
4
```

Data Science Round

- After clearing the initial coding round , next set of rounds are general rounds which will try to test your product / business understanding as well as data science skills.

- Depending on the role , candidate will be asked questions.

- For Data scientists : It could be ML algo questions along with data science case study / their own project based questions.

- For Data Analysts : Questions based on their project , BI skills (Questions related to Power BI / Tableau) could be asked

- For Data Engineers : Questions based on their project , Few Pyspark / SQL / Python questions , data pipeline questions , time –space complexity questions etc

Product Management Round

- As the candidate keeps advancing , he /she starts facing higher management folks in final rounds.

- Product management round is a round where candidate will be tested on his / her ability to understand business / product.

- In this round, candidate is usually given a product as reference and he will be asked questions on that product practices.

  Product: Google Pay (Gpay)

  Problem: Gpay has identified an increase in fraudulent transactions and misuse of the platform.

  How will you identify & solve this issue ?

- Candidate needs to make necessary assumptions , choose proper way to assess experimentation's/strategy's success and has to explain the entire approach to interviewer. Candidate will be judged on how systematically he/she structures answer..

HR Round

- This is final round in which focus is on to check candidate's intent , his overall alignment with the role and company's values.

- The famous cliches like why do you wanna join us ? What is your long term & short term goal ? Why are you leaving current company and planning to join us ? are part of this round.

- Many of the candidate underestimate this round and go in the interview with mugged up answers but this round also carries a potential to decide candiate's fate / destiny.

- Its hard to eliminate a candidate in this round but this significantly affects the salary offering if a candidate fails to show his willingness & honest intent to serve the company.

- Before appearing to this interview , it is good to do a little research on company values , its mission statement , current status (in terms of progress/ decline).

- Avoid internet copied answers instead candidate should think of their own answers which will sound realistic & diplomatic at the same time ? !

Key Data Science Concepts – Statistics , Probability , A/B Testing , Distribution analysis

- Concepts to be tested

- Measures of central tendency , Measures of Dispersion, Correlation , sampling , distribution theory , confidence interval , Inference, Hypothesis testing , A/B testing , ANOVA

- What to prepare ?

- Statistics is a relatively underrated subject of any data science interview. Usually people from non-stat background tend to pay less attention towards statistical topics but they are important for any data science interviews.

- For data scientists / data analysts , preparing for above topics is extremely important.

- Questions may not directly involve the concept application but questions might talk about a case study where we are suppose to apply the concept with respect to the situation.
  For eg : Business group A has launched their product into the market and they need to check are average sales before and after launch are same or different. This is a problem based on Hypothesis testing. Here, you need to make the necessary assumptions before applying the concept.

- Sample Questions

- What are different ways of imputation ? (Mean / median /mode)
- Can correlation imply causation? Explain your answer.
- What are the limitations of using correlation to measure the relationship between variables?
- In a password with 8 characters, how many unique passwords can be created if repetition of characters is allowed?
- A company wants to survey employee satisfaction but can't survey everyone. What sampling method would ensure a representative sample of different departments?
- A poll shows 60% of people support a new law, with a margin of error of +/- 3%. What is the confidence interval for the true population proportion?
- Tell me the difference between Simple random sampling & stratified sampling ?
- What are type 1 & Type 2 errors in hypothesis testing ?
- When would you use ANOVA instead of a simple t-test to compare means?

How to prepare for these concepts ?

- **Master the fundamentals**: Grasp core concepts like mean, median, mode, standard deviation, correlation coefficient, probability rules, and basic distributions (normal, binomial). Don’t mug up the definitions. Try to learn by relating these scenarios with day to day life.

- **Practice with problems**: Solve various practice problems for each concept to solidify your understanding and application. Many online resources offer question banks and tutorials. Everyday keep 30 mins for practicing this problems.

- **Visualize the data**: Utilize tools like spreadsheets or statistical software to explore data through graphs and visualizations. This helps connect concepts with real-world scenarios.

- **Brush up on statistical tests**: Understand commonly used tests like t-tests, chi-square tests, and ANOVA. Learn their assumptions and when to apply each one.

- **Focus on interpretation**: Don't just calculate statistics - interpret their meaning. What does a high standard deviation tell you? How strong is a correlation? Focus on explaining the insights derived from the data.

How to approach these questions ?

- Answering the questions based on Numbers/ calculations & assumptions could be challenging at the times and hence take a moment to think, try to write down your thought process on paper or present it on screen (if screen sharing is reqd.)

- Do not answer anything in haste , its not a rapid fire round ! But indeed a test of your patience , preciseness & situation awareness skills.

- Couple of the times you might lack access to the correct answers but even remember then that you should be able to explain the approach to the interviewer in best possible way.

- For eg : Q. Can correlation imply causation? Explain your answer.
  If you are stuck in a situation where you don’t know answer to the above question , don’t just give up and say ‘No’ .Try to recollect every information which you know about the terminologies used in the question and then simply try to frame your answer around the known information.
  In this case try talking about what is causation , what is correlation , highlight how similar / dissimilar they are.

- Assuming the worst case scenario, where you don’t know anything about question, in that case politely say that you don’t know about the answer and you will be interested in having a look at it afterwords.

SQL

- Concepts to be tested

- Data Definition Language (DDL), Data Manipulation Language (DML), advanced querying techniques, schema navigation, and potentially touch on performance optimization and bonus concepts like window functions especially LEAD , LAG , RANK , DENSE RANK , NTILE , Date Time Functions , Sub queries.

- What to prepare ?

- Grasp the core languages: Solidify your knowledge of Data Definition Language (DDL) for table creation and manipulation (CREATE, ALTER, DROP) and Data Manipulation Language (DML) for data retrieval and modification (SELECT, INSERT, UPDATE, DELETE).

- Master querying: Hone your skills in writing effective queries using WHERE clauses for filtering, JOINs for combining data from multiple tables, and GROUP BY with aggregation functions (COUNT, SUM, AVG) for data summarization.

- Explore advanced techniques: Practice using subqueries for complex data retrieval, ORDER BY for sorting results, and window functions (optional) for advanced in-query data manipulation.

- Navigate schemas: Understand database schema design principles and how to navigate relationships between tables using foreign keys.

- Optimize for efficiency: Grasp basic techniques for optimizing query performance, like using appropriate indexes and avoiding unnecessary joins.

- Bonus points: Familiarize yourself with views and stored procedures for code reusability and data security (optional).

- # Sample Questions

- Write a query to find the total number of customers from California (CA) who have placed more than 5 orders.
- Select the names of all employees who earn a salary above the department average.
- Combine data from a customers table and an orders table using a LEFT JOIN, showing all customers even if they haven't placed any orders.
- Group products by category and calculate the average price for each category.
- Find the top 3 most popular products (by quantity sold) in the last month.
- Write a query to update the email addresses for all customers in the state of New York (NY) to a new domain (e.g., "@newdomain.com").
- Delete all duplicate rows from a table (ensure you specify the duplicate identification criteria).
- Write a subquery to find all orders placed on a specific date (e.g., '2024-07-11’).
- Use the ORDER BY clause to sort employees by their last name in descending order.
- Write a stored procedure to calculate and update the total sales for each customer each month.

How to prepare for these concepts ?

- **Plan your strategy:** Practice the SQL queries on platforms like Hackerrank , Hackerearth , Leetcode.

- **Nail the fundamentals:** Master the core building blocks like DDL (CREATE, ALTER, DROP) for table structure and DML (SELECT, INSERT, UPDATE, DELETE) for data manipulation.

- **Sharpen your querying skills:** Practice writing effective queries using WHERE clauses for filtering, JOINs for combining tables, and GROUP BY with aggregation functions (COUNT, SUM, AVG) for summarizing data.

- **Explore advanced features:** Get comfortable with subqueries for complex retrievals, ORDER BY for sorting, and window functions (optional) for advanced data manipulation within a query.

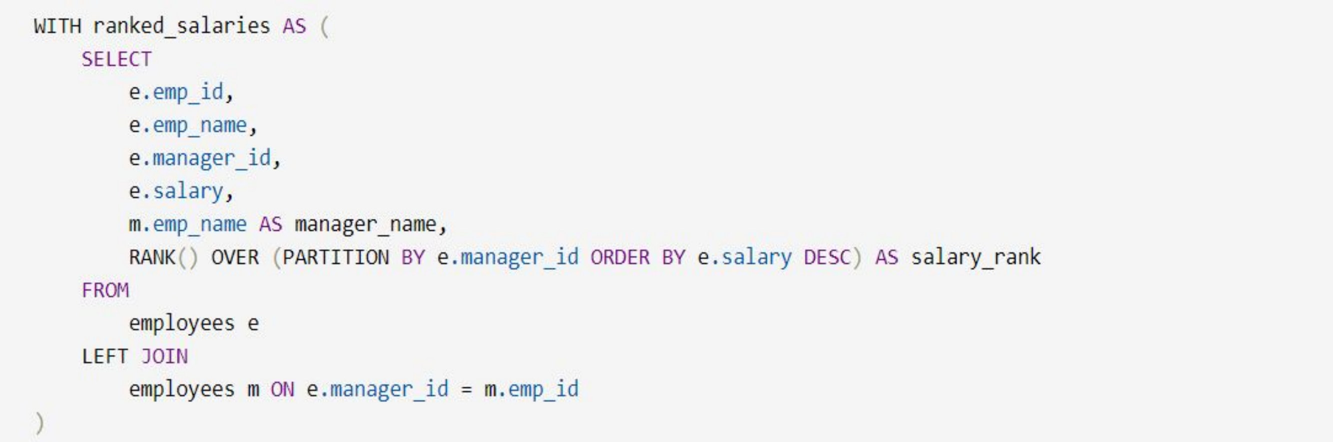
- **Master database navigation:** Understand schema design principles and how to navigate relationships between tables using foreign keys.

- **Optimize for speed:** Learn basic techniques to optimize queries, like using appropriate indexes and avoiding unnecessary joins.

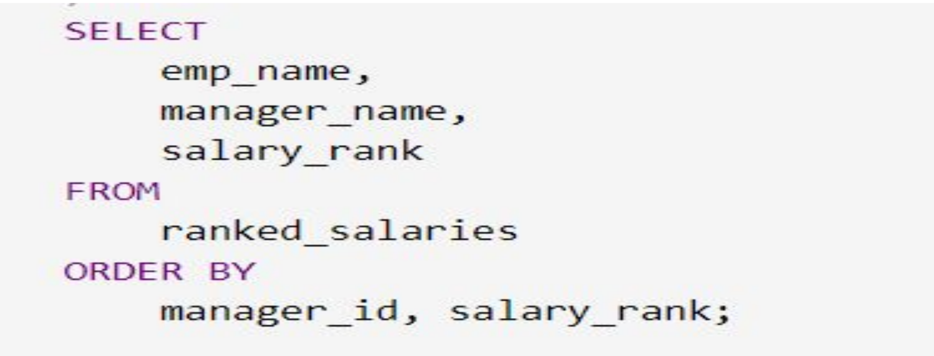
- **Bonus points:** Familiarize yourself with views and stored procedures for code reusability and data security.

Sample SQL problem

Question: Write a query to find:The name of each employee.The name of their manager.The rank of each employee's salary within their department (employees with the same manager are considered in the same department).

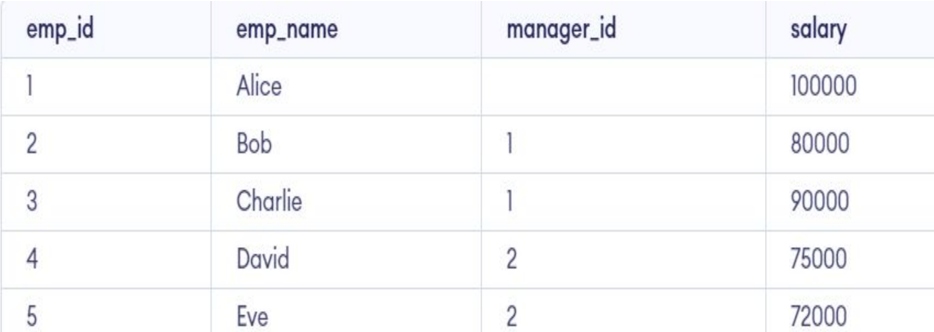


```
WITH ranked_salaries AS (
    SELECT
        e.emp_id,
        e.emp_name,
        e.manager_id,
        e.salary,
        m.emp_name AS manager_name,
        RANK() OVER (PARTITION BY e.manager_id ORDER BY e.salary DESC) AS salary_rank
    FROM
        employees e
    LEFT JOIN
        employees m ON e.manager_id = m.emp_id
)
```



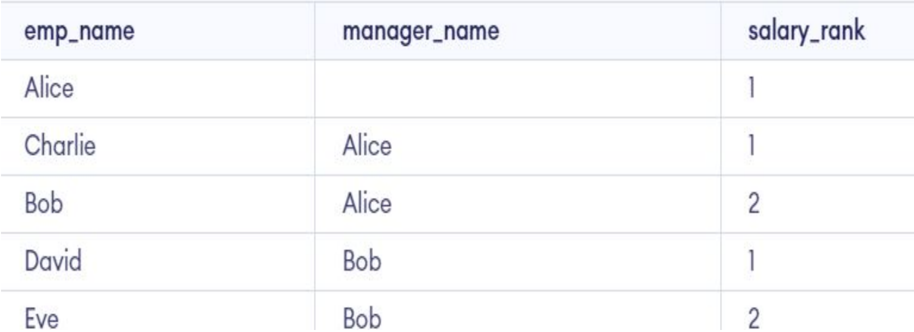
```
SELECT
    emp_name,
    manager_name,
    salary_rank
FROM
    ranked_salaries
ORDER BY
    manager_id, salary_rank;
```

Input Table



emp_id	emp_name	manager_id	salary
1	Alice		100000
2	Bob	1	80000
3	Charlie	1	90000
4	David	2	75000
5	Eve	2	72000

Output Table



emp_name	manager_name	salary_rank
Alice		1
Charlie	Alice	1
Bob	Alice	2
David	Bob	1
Eve	Bob	2

Python

- Concepts to be tested

- DSA*, OOPS*, Data manipulation, Numpy, pandas, scikit-learn , Tensorflow , Keras , Matplotlib-pyplot , seaborn , lambda , apply , map , filter

- What to prepare ?

- **Master the fundamentals**: Solidify your grasp of variables, data types, operators, control flow (if/else, loops), and functions. This strong foundation is crucial for building upon later concepts.
- **Practice problem-solving**: Hone your ability to break down complex problems into smaller, manageable steps. This skill is essential for tackling coding challenges and real-world scenarios.
- **Learn essential data structures**: Understand lists, tuples, dictionaries, and sets. Explore their creation, manipulation (adding/removing elements), and searching techniques.
- **Embrace object-oriented programming (OOP):** Grasp the concepts of classes, objects, attributes (data), and methods (functions). OOP helps organize code and promotes reusability.
- **Leverage libraries**: Explore popular libraries like NumPy (numerical computing), Pandas (data analysis), and Matplotlib (visualization). These tools enhance your Python capabilities.
- **Practice consistently**: The key to mastering Python is consistent practice. Utilize online coding platforms, work on personal projects, and participate in coding challenges to solidify your knowledge.

How to prepare for Pandas, Numpy , Matplotlib ,Scikit-learn for Data Analysis

Empower Yourself with Pandas:

- Explore DataFrames and Series for data analysis and manipulation.
- Master data loading, cleaning, filtering, and transformation techniques.

Master NumPy:

- Learn array creation, manipulation, and mathematical operations.
- Understand broadcasting for efficient vectorized computations.

Visualize with Matplotlib and Seaborn:

- Utilize Matplotlib for basic plots (scatter, line, bar).
- Leverage Seaborn for high-level statistical visualizations.

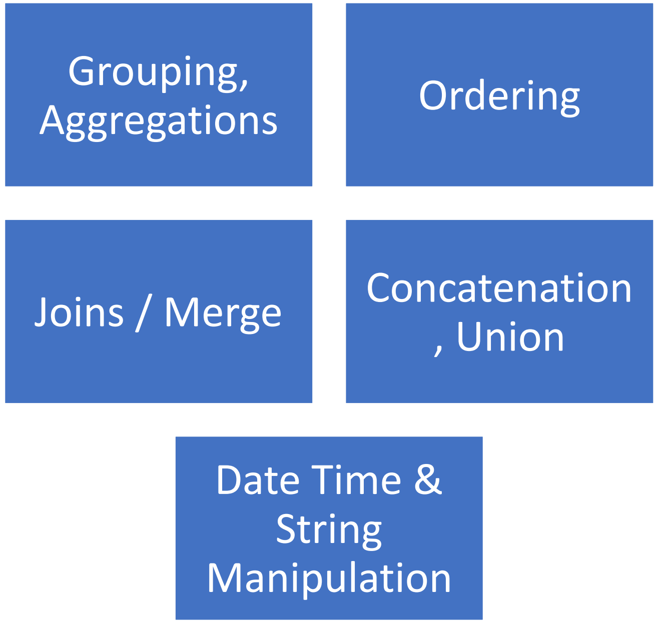
Dive into TensorFlow:

- Grasp foundational concepts of machine learning and deep learning.
- Understand tensors, neural networks, and building basic models.

Learning Resources:

- Online tutorials and courses (Coursera, edX, Kaggle Learn).
- Books ("Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow" by Aurélien Géron).

Must to prepare for any interview



Grouping, Aggregations

Ordering

Joins / Merge

Concatenation , Union

Date Time & String Manipulation

Sample Questions

1. **Pandas:** Can you explain how vectorized operations in pandas DataFrames leverage broadcasting for efficient data manipulation?
2. **NumPy:** Describe the concept of memory-mapped files and how they can be utilized with NumPy arrays for large datasets.
3. **Scikit-learn:** Briefly explain the Bias-Variance tradeoff in machine learning models trained with scikit-learn and how it impacts generalization.
4. **Matplotlib:** How can custom colormaps be defined in Matplotlib to represent specific data patterns or categories?
5. **Seaborn:** Explain the statistical reasoning behind using violin plots in Seaborn for visualizing the distribution of data with potential outliers.
6. **TensorFlow:** Briefly describe the backpropagation algorithm used in TensorFlow for training neural networks and its role in optimizing model parameters.
7. **Pandas:** Can you write a one-line pandas expression to calculate the rolling standard deviation of a specific column within a DataFrame for a window size of 5?
8. **NumPy:** Given two NumPy arrays of different shapes, how can you achieve element-wise multiplication using broadcasting for efficient computation?
9. **Scikit-learn:** How can the GridSearchCV function from scikit-learn be used to hyperparameter tune a machine learning model to find the best configuration?
10. **TensorFlow:** Briefly describe the concept of regularization techniques like L1 and L2, and how they can be implemented in TensorFlow to prevent overfitting in neural networks.

How to prepare for DSA

Solidify Your Python Foundations:

- Grasp core concepts like variables, data types, operators, control flow, and functions.

Master Fundamental Data Structures:

- Focus on Lists, Tuples, Dictionaries, Sets, Stacks, Queues.
- Understand their creation, operations (insertion, deletion, searching), and time/space complexitie

Learn Core Algorithms:

- Grasp searching algorithms (Linear Search, Binary Search).
- Master sorting algorithms (Bubble Sort, Selection Sort, Insertion Sort, Merge Sort, Quick Sort).

Sharpen Your Problem-Solving Skills:

- Break down problems into smaller, solvable steps.
- Analyze time and space complexities of algorithms (Big O Notation).

Practice Makes Perfect:

- Utilize online coding platforms (LeetCode, HackerRank).
- Start with easy problems and gradually progress to harder ones.

Valuable Resources:

- Online Courses (Coursera, edX).
- Books ("Grokking Algorithms" by Aditya Bhargava, "Introduction to Algorithms" by Cormen et al.).

Ultimate Saviour

Sample Questions

1. **Find the longest substring without repeating characters:** Given a string, find the length of the longest substring that does not contain repeating characters. (This can be solved with a sliding window technique, but requires handling edge cases)
2. **Clone a linked list with a random pointer:** Given a linked list where each node has a pointer to another random node in the list, create a deep copy of the list. (This requires keeping track of a map between original and copied nodes)
3. **Detect cycle in an undirected graph:** Given an undirected graph, determine if there exists a cycle formed by edges. (This can be solved using techniques like Union-Find or Depth-First Search with cycle detection)
4. **Merge K sorted lists:** Given an array of K sorted linked lists, merge them into one single sorted linked list. (This requires utilizing a priority queue or a divide-and-conquer approach)
5. **LRU Cache implementation:** Design and implement a Least Recently Used (LRU) cache in Python with a limited capacity. (This involves manipulating a dictionary and doubly linked list)
6. **Flatten a nested dictionary:** Given a nested dictionary with arbitrary levels of nesting, write a function to flatten it into a single level dictionary with appropriate keys. (This can be solved recursively)
7. **Find the maximum sum subarray:** Given an array of integers, find the subarray with the maximum sum. (This can be solved using Kadane's algorithm or dynamic programming)
8. **Egg Drop Problem:** You are given two eggs and a building with n floors. From which floor should you drop the first egg to minimize the number of drops needed to find the floor that will break the egg? (This is a classic dynamic programming problem)
9. **Validate a binary search tree:** Given a binary tree, determine if it is a valid binary search tree where each node's value is greater than its left child and less than its right child. (This requires in-order traversal and value comparison)
10. **Find the kth largest element in an unsorted array:** Given an unsorted array of integers and a value k, find the kth largest element in the array. (This can be solved using a min-heap or by partitioning the array)

How to prepare for OOPS

Solidify your Python basics:

- Grasp data types, variables, operators, control flow, and functions.

Understand Object-Oriented Concepts:

- Focus on Classes, Objects, Attributes, and Methods.

Brush up on:

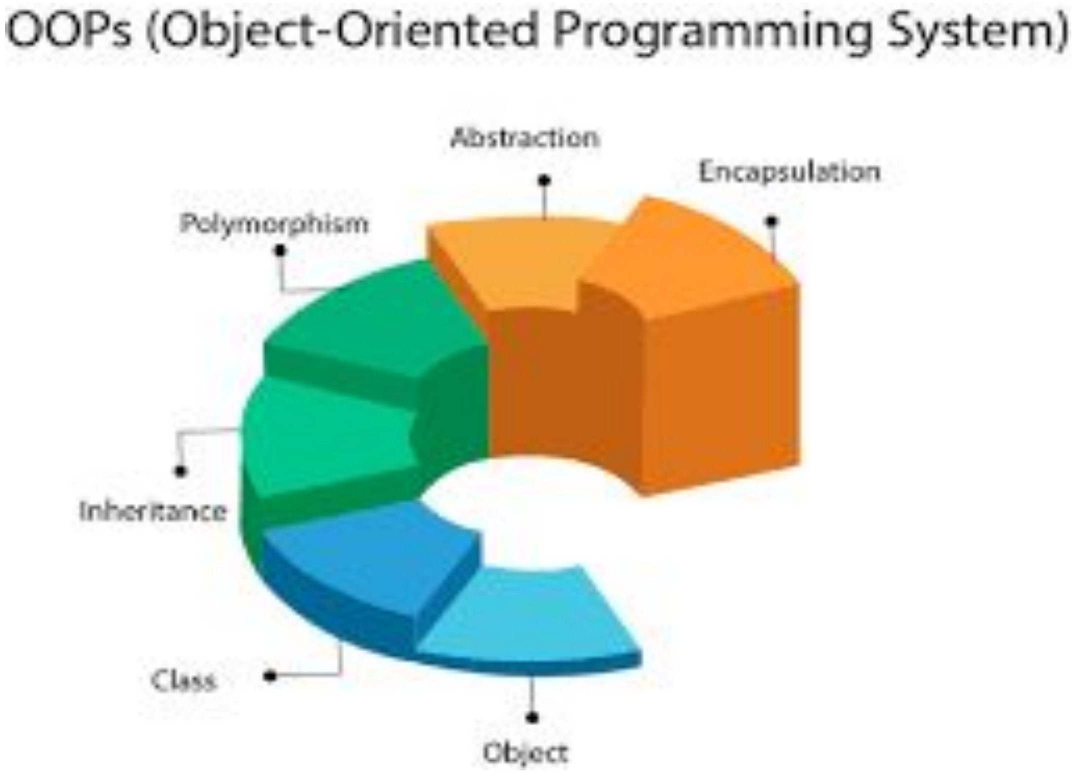
- Inheritance: Reusing code through parent-child relationships.
- Encapsulation: Protecting data integrity within a class.
- Polymorphism: Making objects behave differently in similar situations.

Resources:

- Online tutorials and courses (mention popular platforms like Coursera or edX).
- Books on Python OOP

Practice Makes Perfect:

- Start with small projects implementing OOP concepts.
- Utilize online coding challenges and exercises.



Sample Questions

- What is the difference between a class and an object?
- Explain the concept of encapsulation with an example.
- Describe the four pillars of OOP (Inheritance, Polymorphism, Encapsulation, Abstraction). Briefly explain each.

- Explain the concept of polymorphism and its benefits.
- Differentiate between static and dynamic polymorphism.
- How can method overriding achieve polymorphism?

- What are the different types of inheritance (Single, Multilevel, Hierarchical, Multiple)?
- Differentiate between method overriding and overloading.
- In inheritance, when would you use a superclass method?

- What is the purpose of abstract classes and interfaces?
- Describe the advantages and disadvantages of inheritance.
- Explain how exception handling is used in OOP.

Excel

Concepts to be tested

- Formatting , data imports , datatype column conversions , visualizations , pivot , vlookup ,data analysis , MATCH , Index , Excel shortcuts

What to prepare ?

- **Prepare Formulas:**

Explore formulas (like SUM, AVERAGE, VLOOKUP) to automate calculations and save time. Mastering basic formulas can unlock the power of Excel.

- **Leverage Formatting:**

Apply formatting (bold, colors, conditional formatting) to make your spreadsheet visually appealing and easier to understand. Highlight important data points for better readability.

- **Consider PivotTables and Charts:**

For complex data analysis, leverage PivotTables to summarize data and create interactive reports. Use charts (bar graphs, pie charts) to visually represent trends and patterns.

- **Explore Advanced Features:**

As you become comfortable, delve into advanced features like data validation, macros, and custom functions to further enhance your spreadsheets and automate repetitive tasks.

Sample Questions

1. **Circular References:** How can circular references in Excel lead to incorrect calculations, and what strategies can be employed to identify and resolve them?
2. **Volatile Functions:** Explain the concept of volatile functions in Excel and how their automatic recalculation can impact spreadsheet performance. How can you optimize spreadsheets with volatile functions?
3. **Conditional Formatting:** Describe a scenario where advanced conditional formatting techniques, like using formulas within formatting rules, can be used to create a dynamic and informative data presentation.
4. **Data Validation:** How can data validation rules be used to enforce specific data types (e.g., only allow numbers, dates) or restrict data entry within a certain range in an Excel sheet? Explain the benefits of data validation.
5. **Macros vs. VBA:** Briefly differentiate between macros and VBA in Excel. When would you use one over the other for automating tasks within a spreadsheet?
6. **Excel PivotTable Calculations:** How can custom calculations be defined within Excel PivotTables to derive new insights from the summarized data? Explain an example scenario.
7. **Goal Seek vs. Solver:** Describe the functionalities of Goal Seek and Solver in Excel. When would you use Goal Seek versus Solver for finding solutions based on specific target values?
8. **Error Handling:** Explain the purpose of different error codes displayed in Excel (e.g., #DIV/0!, #REF!) and how custom error handling functions can be used to provide informative messages for users encountering errors.
9. **Dynamic Array Formulas:** Since Excel 365, dynamic array formulas have been introduced. How do these differ from traditional array formulas, and what are some advantages of using them?
10. **Data Consolidation:** Imagine you have multiple Excel workbooks with similar data structures. Describe two or more techniques for consolidating this data into a single master workbook for comprehensive analysis.

How to prepare a ML algorithm

1. Saggregate study into Supervised & Unsupervised setup , classification & regression setup.

2. Learn about assumptions about algorithms.

3. Read theory about that algorithm

4. Go to scikit-learn documentation page and try ro read about all the hyperparameters , parameters

5. Implement that algorithm on at least a dataset.

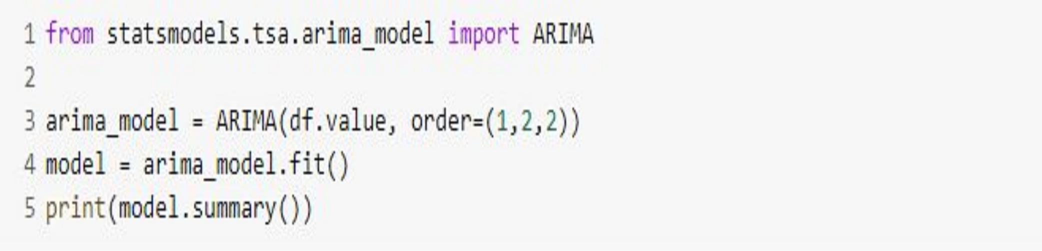
6. Understand How it could be evaluated.

7. Learn about limitations & advantages of the algorithm

Important aspects of ML study

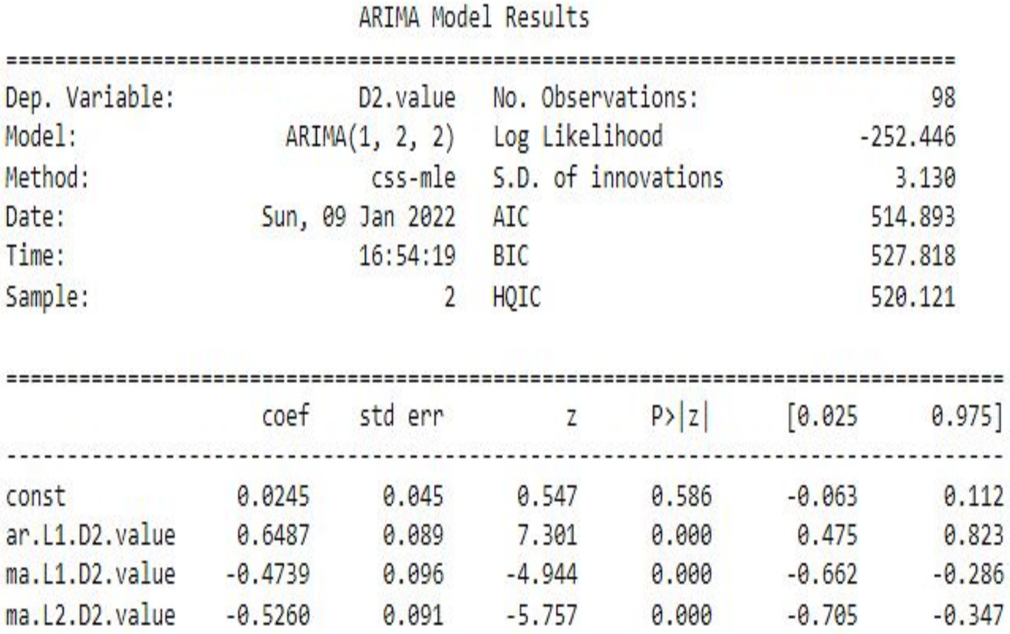
1. If everything is already coded in scikit-learn why should I learn the ML algorithm from scratch ?
2. Missing values are so easy to be treated , well is that the case really ?
3. Understanding Class imbalance / What can happen if not taken care of? Role of Sampling
4. Interpretation of Clustering
5. How do you use PCA for supervised problems
6. Role of Multicollinearity in ML modelling
7. Sparse features / categorical features
8. Scaling: where to do /where not to
9. Interpreting evaluation metrics
10. Model Explanability / Accuracy (Classic Trade-off)
11. Lets have chit-chat about NLP

Reading Outputs



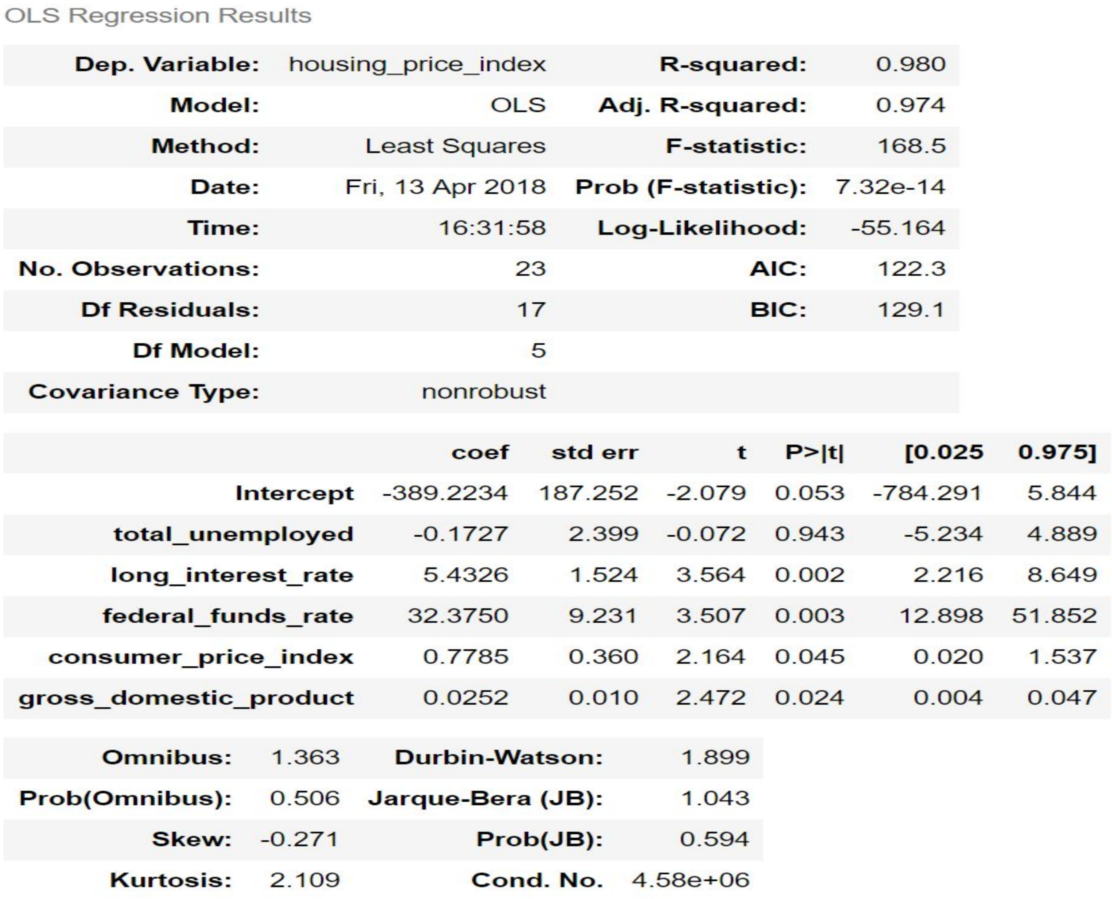
```python
from statsmodels.tsa.arima_model import ARIMA

arima_model = ARIMA(df.value, order=(1,2,2))
model = arima_model.fit()
print(model.summary())
```



```
                             ARIMA Model Results
==============================================================================
Dep. Variable:               D2.value   No. Observations:                   98
Model:                 ARIMA(1, 2, 2)   Log Likelihood                -252.446
Method:                       css-mle   S.D. of innovations              3.130
Date:                Sun, 09 Jan 2022   AIC                            514.893
Time:                        16:54:19   BIC                            527.818
Sample:                             2   HQIC                           520.121

==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.0245      0.045      0.547      0.586      -0.063       0.112
ar.L1.D2.value     0.6487      0.089      7.301      0.000       0.475       0.823
ma.L1.D2.value    -0.4739      0.096     -4.944      0.000      -0.662      -0.286
ma.L2.D2.value    -0.5260      0.091     -5.757      0.000      -0.705      -0.347
```



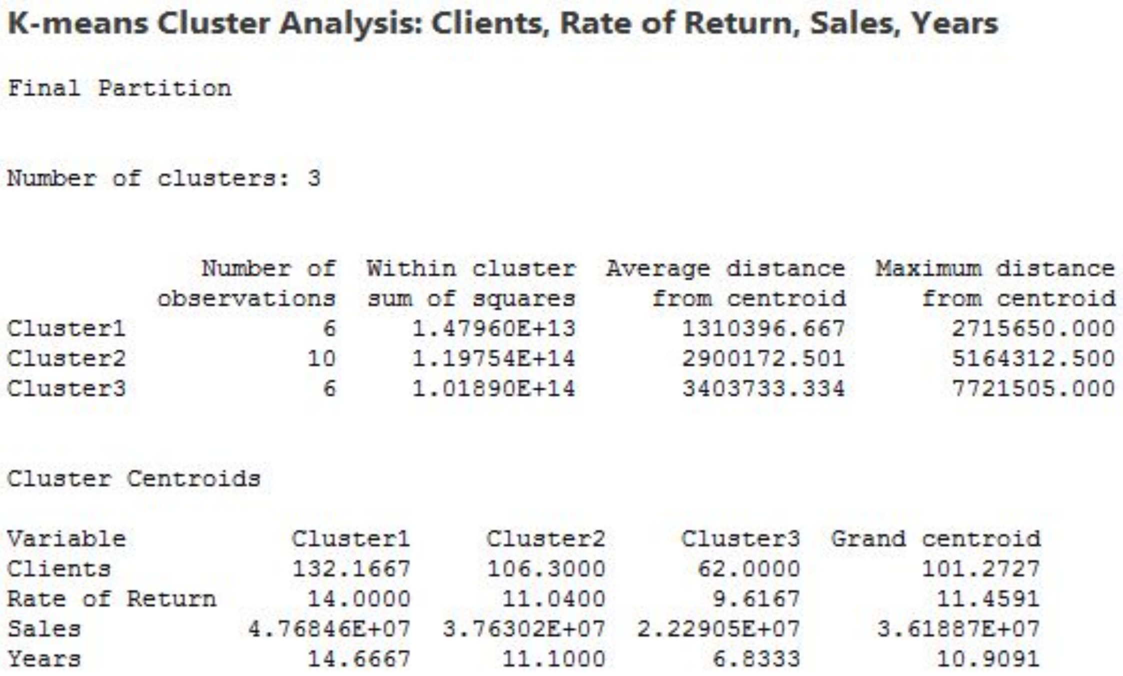
OLS Regression Results

Dep. Variable:	housing_price_index	**R-squared:**	0.980
Model:	OLS	**Adj. R-squared:**	0.974
Method:	Least Squares	**F-statistic:**	168.5
Date:	Fri, 13 Apr 2018	**Prob (F-statistic):**	7.32e-14
Time:	16:31:58	**Log-Likelihood:**	-55.164
No. Observations:	23	**AIC:**	122.3
Df Residuals:	17	**BIC:**	129.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-389.2234	187.252	-2.079	0.053	-784.291	5.844
total_unemployed	-0.1727	2.399	-0.072	0.943	-5.234	4.889
long_interest_rate	5.4326	1.524	3.564	0.002	2.216	8.649
federal_funds_rate	32.3750	9.231	3.507	0.003	12.898	51.852
consumer_price_index	0.7785	0.360	2.164	0.045	0.020	1.537
gross_domestic_product	0.0252	0.010	2.472	0.024	0.004	0.047

Omnibus:	1.363	**Durbin-Watson:**	1.899
Prob(Omnibus):	0.506	**Jarque-Bera (JB):**	1.043
Skew:	-0.271	**Prob(JB):**	0.594
Kurtosis:	2.109	**Cond. No.**	4.58e+06

Reading Outputs



K-means Cluster Analysis: Clients, Rate of Return, Sales, Years

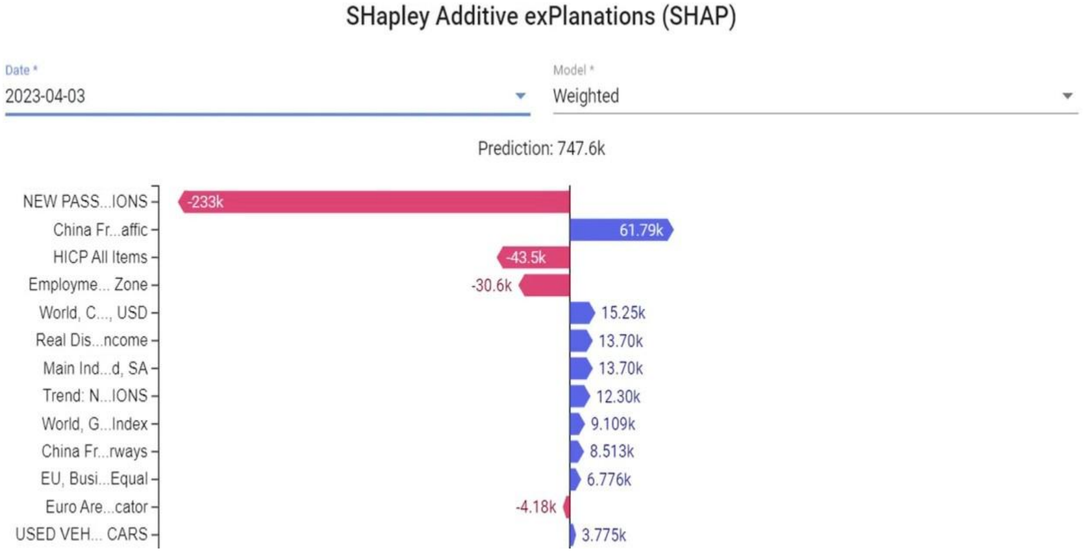
```
Final Partition


Number of clusters: 3


           Number of  Within cluster  Average distance  Maximum distance
        observations  sum of squares     from centroid     from centroid
Cluster1           6     1.47960E+13       1310396.667       2715650.000
Cluster2          10     1.19754E+14       2900172.501       5164312.500
Cluster3           6     1.01890E+14       3403733.334       7721505.000


Cluster Centroids

Variable           Cluster1     Cluster2     Cluster3  Grand centroid
Clients            132.1667     106.3000      62.0000        101.2727
Rate of Return      14.0000      11.0400       9.6167         11.4591
Sales           4.76846E+07  3.76302E+07  2.22905E+07     3.61887E+07
Years               14.6667      11.1000       6.8333         10.9091
```



SHapley Additive exPlanations (SHAP)

Date *
2023-04-03

Model *
Weighted

Prediction: 747.6k

Data Science Case study

- Case Study 1:

An airline company has experienced a significant churn of customers and thus drop in revenues over the past year. The management has noticed this trend but cannot pinpoint the exact causes. As a data scientist, you are tasked with analyzing the situation. Additionally, you are expected to develop a machine learning model that can predict the potential churn customers and provide actionable insights to help the company to retain them.

- Pre- Execution phase

1. Objective understanding / Product understanding / Selecting success criterias
2. Response definition
3. Data Collection & Data check

---------------------------------- Approach --------------------------------

Response variable : Churn

All the customers :

- At least 2 years in last 4 years & haven’t taken a flight with us since last 1 year, AFV > 100$, Min No of flights : 5

Churn

-Data variables : 1. Demographic features (Customer oriented) : Age , gender , Destination , Departure , Class , Average amount , Opts for Meal / Not , Rating , Average time between flights, Average Journey time

2. Transaction : Mode of payment , Reward , Transaction time

3. Flight : Timing of flight , Weekend/ weekday , Average delay ,Flight duration

- Data Science Execution

1. Data Cleaning/ standardization
2. Outliers Treatment / Missing value analysis
3. Aggregations
4. EDA
5. Splitting
6. ML Algo
7. Evaluation Sign-off / Strategy

- Post DS execution

1. Model Deployment
2. Model Monitoring (Checking Model drift , Data drift)

Data Science Case study

- Case Study 2:

A retail company wants to enhance its marketing strategies by better understanding its diverse customer base. They have collected a rich dataset containing various features related to customer demographics, purchase behavior, and product preferences. As a data scientist, your task is to perform a thorough analysis of this dataset and use clustering techniques to segment the customers into distinct groups. The goal is to identify unique customer segments that can be targeted with tailored marketing campaigns to improve customer engagement and increase sales.

.

Solution

Data Science Case study

- Case Study 3:

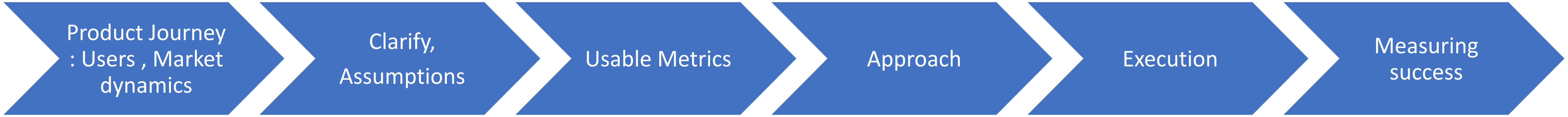
An e-commerce platform has observed fluctuating sales figures over the past few years, which makes inventory management and financial planning challenging. The company has comprehensive historical sales data, including seasonal trends, promotional events, and other external factors. As a data scientist, you are required to analyze this time series data to identify patterns and trends. Your objective is to develop a robust forecasting model that can accurately predict future sales. This forecast will help the company in optimizing inventory levels, planning marketing strategies, and making informed business decisions.

.

Solution

Product Round Preparation

Product Round Preparation Goes around these fundamental concepts:



Product Journey : Users , Market dynamics

Clarify, Assumptions

Usable Metrics

Approach

Execution

Measuring success

Step -1 : Product Breakdown, User Journey, Market Dynamics

- **Product Breakdown:** Define the product, its features, and target audience.
- **User Journey:** Map out the user's interaction with the product, from acquisition to retention.
- **Market Dynamics:** Analyze competitors, industry trends, and customer behavior.
- **Identify Problem Areas:** Pinpoint challenges or opportunities within the product or market.
- **Define Business Objectives:** Clearly articulate the desired outcome of the data analysis.

Step- 2: Clarifications and Assumptions

- **Data Availability:** Specify the available data sources and formats.
- **Data Quality:** Assess data accuracy, completeness, and consistency.
- **Assumptions:** Clearly state underlying assumptions about user behavior, market trends, etc.
- **Data Limitations:** Acknowledge potential data gaps or biases.
- **Scope of Analysis**: Define the boundaries of the analysis.

Step-3 : Usable Metrics

- **Key Performance Indicators (KPIs):** Identify relevant metrics to measure product success.
- **User Behavior Metrics:** Analyze user actions and engagement within the product.
- **Business Metrics:** Track revenue, growth, and profitability.
- **Data-Driven Insights:** Explain how metrics will inform decision-making.
- **Data Visualization:** Consider effective ways to present metrics (charts, graphs).

Step- 4: Approach

- **Data Collection:** Outline methods for gathering necessary data.
- **Data Cleaning and Preparation:** Describe data preprocessing steps.
- **Exploratory Data Analysis (EDA):** Summarize initial findings and patterns.
- **Modeling Techniques:** Select appropriate statistical or machine learning models.
- **Evaluation Methodology:** Determine how to assess model performance

Step-5 : Execution

- **Data Analysis:** Apply statistical methods and machine learning algorithms.
- **Model Building:** Develop predictive or explanatory models.
- **Data Visualization:** Create compelling visuals to communicate insights.
- **Storytelling:** Craft a narrative that effectively conveys findings.
- **Iterative Process:** Emphasize the importance of refining analysis based on results.

Step- 6: Measuring Success

- **Impact Assessment:** Evaluate the impact of data-driven recommendations.
- **ROI Calculation:** Quantify the return on investment of the project.
- **Continuous Monitoring:** Establish a system for tracking ongoing performance.
- **Feedback Loop:** Incorporate user feedback to refine the product.
- **Data-Driven Culture:** Promote data-informed decision-making within the organization

As a product manager at Meta AI , you are planning to propose a new video feature in the market. How will you assess the situations and dynamics behind these decision ?

1. Product Journey / User journey , Market Dynamics :   Meta AI : Subscription , Free to use , Anyone can create account, user might list his/her business , app can get revenue advertisement , subscriptions , Entertainment , Tiktok , Youtube ,Linkedin , Snapchat , Objective : To assess effectiveness of the video feature rollout

2. Clarifications / Assumptions : Userbase it is spending on an avg 4 hrs on platform , This userbase includes xyz regions it is banned in some of the regions , What are more details about feature ? (Duration :30 sec , Theme , genre(concept) , How is it different existing feature , What userbase it will try to attract ? AI involvement ? Promotions , sensitive content

3. Usable Metrics : Engagagement (Improvement of ratings) ,(Average user time) , Average session time for a single session, Likes , Comments , Shares , Acquisition /Churn (New users/exiting users leaving), Average Monthly users, Average Yearly users, Video completion rate, ad click through rate , Revenue (Business model : Subscription , ad-revenue)

4. Approach : Forecasting (Revenues/ acquisition/churn, user time) , A/B Testing (Effectiveness of the feature) , Dashboards (Improvement / Decline over time) , Classification, Optimization

5. Execution : A/B Testing : Data , Control / treatment , cleaning , Population assumptions , sampling , level of confidence , Z/T/F/Chi test, p-value

6. Measuring success : Delta (All sort of differences !) , change in reveues , customer retention /acquisition

As a product manager at you tube , you are given a task to detect the abusive comments. How will you deal with this situation ? What all factors will you be considering ?

As a Data Scientist at big payment app company , you wish to analyze the price sensitivity for different merchants. How will you go about it ?

Guess-estimates preparation

- **Guesstimates** are estimation problems often presented in data science and business analyst interviews.
- They assess your ability to structure a problem, make logical assumptions, and perform quick calculations.
- Examples include estimating the market size of a product, the number of cars in a city, or the revenue of a company.
- They test your analytical thinking, problem-solving skills, and ability to communicate your thought process.
- Guesstimates are not about finding the exact answer but demonstrating a structured approach to arrive at a reasonable estimate.

Guess-estimates preparation

- Structure the problem: Break down the problem into smaller, manageable parts.
- Make assumptions: Clearly state your assumptions and their rationale.
- Use a framework: Consider frameworks like the 'Top-Down' or 'Bottom-Up' approach.
- Perform calculations: Use basic arithmetic and estimation techniques to arrive at a final estimate.
- Communicate clearly: Explain your thought process and assumptions confidently.
- Practice: Regularly practice solving guesstimate problems to improve your skills.

Top-Down Approach

- **Start with a big picture:** Begin with the overall market size or total number.

- **Break it down:** Divide the total into smaller, more manageable segments based on relevant factors (e.g., demographics, geography, product categories).

- **Apply percentages:** Estimate the percentage of the total that each segment represents.

- **Calculate the final estimate:** Multiply the percentages by the total to arrive at the final estimate.

- **Example:** Estimating the number of smartphones in a city.
  Start with the total population of the city.
  Break down the population by age groups (e.g., 18-24, 25-34, etc.).
  Estimate the percentage of smartphone ownership in each age group.
  Calculate the total number of smartphones by multiplying the population of each age group by their
  respective smartphone ownership percentage.

Top-Down
Highly Structured
Clear Management
VS
Focus on Collaboration
Creativity Encouraged
Bottom-Up

Bottom-Up Approach

- **Start with the smallest units:** Begin by estimating the number or value of individual components.

- **Aggregate upwards:** Combine the estimates for individual components to arrive at a larger total.

- **Consider all relevant factors:** Ensure that all necessary components are included in the calculation.

- **Example:**
  Estimating the revenue of a pizza delivery company.
  Start by estimating the average number of pizzas sold per day per store.
  Calculate the average price per pizza.
  Estimate the number of stores.
  Multiply the number of pizzas per day, average price, and number of stores to get the daily revenue.
  Multiply the daily revenue by the number of days in a year to get the annual revenue.

Top-Down
Highly Structured
Clear Management
VS
Focus on Collaboration
Creativity Encouraged
Bottom-Up

Guess-estimate Example

Estimate the total revenue generated by online food delivery services in Mumbai.

Top to Bottom Approach:

Population of Mumbai: 20 million
Working Professionals (40%): 20 million *40% : 80 lac , Others ():120 lac

Age category : Working Professionals : 20-30 : 35 lac , 30-45 : 25 lac , 45+:

Working Professionals : Avg 2 orders per week
Non-working : Avg 0.5 orders per week

Avg Dish order value : Rs 300 (Working)
Non-working : Rs 150 (Non-working)

80 * 2 * 300= xyz (weekly) revenue for Working Professionals
+
120 lac *0.5* 150 = abc(weekly) revenue for Others
- Market_share (40%) : 40% * (number)

Guess-estimate

Sample Problems:

1. Estimate the number of smartphones sold in the India in a year.

2. Estimate the number of rides a typical ridesharing service (like Uber) completes in Mumbai City on an average day.

3. Estimate the population of United States given that population of world is 7 billion. Estimate the annual consumption of coffee in the India.

4. Estimate the number of sanitizers used in the India each year.

5. Estimate the number of streetlights in a Bangalore city with a population of 1 million people.

Psychological Preparation

- **Build Confidence**: Believe in your abilities and preparation. Practice mock interviews to boost your confidence.
- **Manage Anxiety**: Employ relaxation techniques like deep breathing or meditation to calm nerves. Focus on the learning experience rather than just the outcome.
- **Positive Mindset**: Approach the interview with a positive attitude. Emphasize your passion for data and problem-solving.
- **Effective Communication**: Practice clear and concise articulation of your thoughts. Active listening is key.
- **Storytelling**: Develop compelling narratives around your projects to showcase your impact.
- **Embrace Challenges**: View challenges as opportunities to demonstrate problem-solving skills and adaptability.

HR Interview Preparation

General Questions

- Tell me about yourself.
- Why are you interested in a data science role?
- Why our company?
- What do you know about our company and its data science team?
- What excites you most about data science?
- What are your strengths and weaknesses?
- How do you handle pressure and deadlines?

HR Interview Preparation

Behavioral Questions

- Describe a challenging data project you worked on.
- How do you handle conflicts within a team?
- Give an example of a time you failed. What did you learn from it?
- How do you stay updated with the latest trends in data science?
- Describe a time you had to explain complex data insights to a non-technical audience.

Role-Specific Questions

- What is your experience with data cleaning and preprocessing?
- How do you choose the right algorithm for a given problem?
- Can you explain the difference between supervised and unsupervised learning?
- What is your experience with data visualization tools?
- How do you handle missing data?

HR Interview Preparation

Miscellaneous questions

- What parameters you look for while selecting any company's offer
- What are your future aspirations?
- Your resume has a gap why is it so?
- What is your total experience ? What is Relevant experience ?
- What is your expected salary ?
- Your demanded hike is way high , perhaps not as per market expectations why is it so?
- Your notice period is too long , we want someone who can join us immediately
- You already have an offer in hand , then why should we trust you ? What are the chances that you will join us only ?
- Some tips on negotiations , Fixed , Variable , ESOPS
- You have been switching organizations quite frequently why should we trust you?
- What is reason that you are looking out for job change ?