ShadowFox

LEARN • CREATE • LEAD

Data Science Internship Task List

Internship Pre-requisites before starting your tasks:

1.LinkedIn Profile Update: Ensure that your LinkedIn profile is updated to reflect your technical skills, and update your experience section to include **"ShadowFox Data Science Intern."**

2.LinkedIn Post: It's not mandatory to post the offer letter on LinkedIn, but if you wish to receive **extra swags** at the end of the internship, you can post your offer letter and **tag us** to receive assured swags.

3.GitHub Repository: You are required to create a separate GitHub repository named **"ShadowFox"** for all tasks. You can use any IDE to write your code and upload it to the respective repository.

4. Completion of Tasks: Complete the required tasks as specified in this Task List.

5. Proof of Work: At ShadowFox, we value **Proof of Work** (POW). You are required to post a video explanation of your respective tasks in LinkedIn. Screenshots must be submitted during your task submission.

After completing all the above steps, proceed with your task completion.Kindly note that all the details and screenshots you submit will be thoroughly verified.

Task Level (Beginner):

Visualization Library Documentation

Objective: Create a comprehensive documentation guide for **2** of the following Python visualization libraries: **Matplotlib, Seaborn, Plotly, Bokeh, and Pandas**. Your guide should focus on the variety of graphs each library can generate and include practical examples with code snippets.

Requirements:

1. **Library Overview:** Provide a brief introduction to the selected libraries, highlighting their unique features and typical use cases.
2. **Graph Types:**
- Document the different types of graphs available in each library, such as line plots, scatter plots, bar charts, histograms, pie charts, etc.
- For each graph type, include a brief description, potential use case, and a simple code example demonstrating how to generate the graph.
3. **Comparison:** Offer a comparison section discussing the strengths and weaknesses of each library regarding ease of use, customization options, interactivity, and performance with large datasets.

4.Resources

Matplotlib:https://matplotlib.org/stable/users/explain/quick_start.html#quick-start
Seaborn-https://seaborn.pydata.org/tutorial/introduction.html
Plotly-https://plotly.com/python/distplot/
Bokeh:https://docs.bokeh.org/en/latest/docs/user_guide/basic.html
Pandas-https://pandas.pydata.org/docs/user_guide/index.html

Deliverable:
A PDF or Markdown file containing the compiled guide, ensuring the content is clear, concise, and informative for new users.
Matplotlib:https://matplotlib.org/stable/users/explain/quick_start.html#quick-start
Seaborn-https://seaborn.pydata.org/tutorial/introduction.html
Plotly-https://plotly.com/python/distplot/
Bokeh:https://docs.bokeh.org/en/latest/docs/user_guide/basic.html
Pandas-https://pandas.pydata.org/docs/user_guide/index.html

Task Level (Intermediate): Do any 1 of 2

1. Conduct an in-depth analysis of the Air Quality Index (AQI) in Delhi, addressing the specific environmental challenges faced by the city. Define research questions centered around key pollutants, seasonal variations, and the impact of geographical factors on air quality. Utilize statistical analyses and visualizations to gain insights into the dynamics of AQI in Delhi, offering a comprehensive understanding that can inform targeted strategies for air quality improvement and public health initiatives in the region.
Dataset-Link:
https://drive.google.com/drive/folders/1Bjn2YEmafyYckkJAwsOSqngpEMGWyIXc? usp=drive_link

2. Undertake a sentiment analysis of X data to gain insights into the prevailing sentiments expressed on the platform. Define research questions focusing on specific topics, trends, or events, and leverage natural language processing techniques to categorize tweets as positive, negative, or neutral. Employ statistical analyses and visualizations to discern patterns in sentiment over time, contributing to a comprehensive understanding of public opinion on X regarding diverse subjects.
Dataset-Link:
https://drive.google.com/file/d/1cJlTwbx7mUxKCZS32uduCX0_850UiNKv/view?usp=sharing

Task Level(Advanced): Do any 1 out of 2

1. Cricket Fielding Analysis Data Collection Objective:
As a budding sports analyst with an interest in cricket, your task is to conduct a detailed fielding performance analysis for three players of your choice from any innings of a T20 match. This analysis will help to gauge individual fielding contributions and their impact on the team's defensive play. A detailed sample data and sample performance matrix is attached along the mail of task list.

Dataset Features:
- Match No.: Identifier for the match.
- Innings: Which innings the data is being recorded for.
- Team: The team in the field.
- Player Name: The fielder involved in the action.
- Ballcount: Sequence number of the ball in the over.
- Position: Fielding position of the player at the time of the ball.
- Short Description: Brief description of the fielding event.
- Pick: Categorize the pick-up as clean pick, good throw, fumble, bad throw, catch, or drop catch.
- Throw: Classify the throw as run out, missed stumping, missed run out, or stumping.
- Runs: Enter the number of runs saved (+) or conceded (-) through the fielding effort.
- Overcount: The over number in which the event occurred.
- Venue: Location of the match.

Task Level(Advanced): Continued...

Performance Metrics Formula:
To assess the fielding performance, use the following formula:
PS=(CP×WCP)+(GT×WGT)+(C×WC)+(DC×WDC)+(ST×WST)+(RO×WRO)+(MRO×WMRO)+(DH×WDH)+RS
Where:
PS: Performance Score
CP: Clean Picks
GT: Good Throws
C: Catches
DC: Dropped Catches
ST: Stumpings
RO: Run Outs
MRO: Missed Run Outs
DH: Direct Hits
RS: Runs Saved (positive for runs saved, negative for runs conceded)

Task Instructions:
1. Data Collection: For each ball bowled in the match, record the fielding effort according to the dataset features outlined above. Pay close attention to the effectiveness of fielding actions and their outcomes.
2. Analysis Preparation: Your collected data will be used for advanced fielding analysis, identifying key areas of improvement and fielding strengths within the team.

Deliverable: A well-organized spreadsheet or database containing the complete fielding data for the match.

This task requires meticulous attention to detail and an understanding of cricket fielding dynamics. Your analysis will contribute to strategic fielding placements and improvements in team performance.

View Sample Dataset with Calculation here

Resources: https://roadmap.sh/python

Task Level(Advanced):

2. Present your findings on the final project, where you are tasked with creating a Jupyter notebook from scratch and conducting a data analysis on a dataset of your choice. This comprehensive process involves selecting a dataset that piques your interest, exploring its contents within a Jupyter notebook, and identifying research questions that the data might help answer.

Guidelines:

1. Begin by finding a dataset that piques your interest. You can choose from a list of places with valuable datasets provided in our reading, or feel free to select data related to your hobbies or work if it is publicly available.
2. Explore the dataset in a Jupyter notebook, gaining a deep understanding of its contents. This exploration phase will help you identify the types of questions that can be addressed using the available data. Remember that data cleaning may be necessary during this step.
3. Based on your exploration, narrow down and define a research question. Consider what specific information or insights you want to derive from the dataset. Your research question should be tailored to the characteristics of the data you have chosen.
4. Utilize various visualization techniques to analyze the dataset and find a solution for your research problem. Visualization is a powerful tool for uncovering patterns, trends, and relationships within the data.
5. Throughout this process, pay close attention to the project description and grading rubric to ensure that your work aligns with the project's requirements and expectations.
6. By the end of this week, you should have not only selected a dataset and formulated a clear research question but also utilized visualizations to derive meaningful insights and solutions to your research problem.

Task Level(Advanced): Continued...

Example: In the intricate realm of sales data analytics, the dataset under consideration is a compilation of transactional intricacies encompassing transaction ID, date, gross sales, net sales, profit/loss, and additional factors such as cost of goods sold (COGS), manufacturing costs, and freight costs. The analytical odyssey commences with a meticulous exploration, encompassing exploratory data analysis (EDA) techniques to unravel underlying patterns, detect outliers, and address data anomalies, with a particular focus on intricate financial metrics like COGS.

Fiscal years, representing the financial reporting period distinct from the conventional calendar year, add an additional layer of complexity to the analysis.

Understanding the cyclical nature of fiscal periods is crucial for discerning seasonality effects, identifying peak sales periods, and aligning data insights with broader financial reporting structures.

The formulation of a research question in this sophisticated landscape requires consideration of multifaceted variables. For instance, one might delve into understanding the correlation between manufacturing costs and net sales or investigate the impact of freight costs on overall profitability. This step involves not only identifying questions relevant to the sales domain but also establishing nuanced connections between financial variables to unearth strategic insights.

Task Level(Advanced): Continued...

Visualization techniques, implemented through Python libraries like Matplotlib and Seaborn, extend beyond conventional line plots. They may include sophisticated financial visualizations such as cost breakdowns, profit margin trends, and comparative analyses between fiscal years. Time series analyses, integrating fiscal timelines, offer a comprehensive perspective on the evolution of financial metrics over distinct reporting periods.

The ultimate presentation in the Jupyter notebook encapsulates not only the temporal trends in gross sales, net sales, and profit/loss but also incorporates detailed breakdowns of COGS, manufacturing costs, and freight costs. This holistic approach to sales data analysis, integrating both financial and temporal dimensions, demonstrates proficiency not only in conventional data analytics but also in navigating the intricacies of fiscal reporting and complex financial metrics within the sales domain.