**Data Analytics Week 3**

**Week 3:  
  
Tasks & Resources: -**

“Welcome to the Week 3 of PrepInsta’s Data Analytics Internship program.  In this module you will be working on data cleaning and manipulation using Pandas and NumPy. You will find the task and resources both on this page.”

**Project 3: -**

Dive into data cleaning and manipulation using Pandas and NumPy. Apply these tools to pre-process and transform a provided dataset, preparing it for analysis.

**Takeaway: -**

This task is designed to assess your ability to clean and manipulate data using Pandas and NumPy, essential skills in any data analytics role.

Effective data cleaning lays the foundation for accurate and meaningful analysis, so approach this task with attention to detail and thoroughness.

Best of luck!

**Pre-requisites**

* Basic understanding of Python programming language.
* Familiarity with Pandas and NumPy libraries for data manipulation.
* Knowledge of common data cleaning techniques, such as handling missing values and outliers.

Note: - In case you want to revise the pre-reqs, just head over to the resources section for a quick brush up.

**What you need to do?**

* **Dataset Exploration:**

Begin by exploring the provided dataset. Understand its structure, column names, and the types of data it contains.

Identify any missing values, outliers, or inconsistencies within the dataset.

* **Library Setup:**

Set up your Python environment with Pandas and NumPy. Ensure you have a Jupyter Notebook or another suitable platform for interactive coding.

* **Data Cleaning:**

Address missing values by deciding on appropriate strategies such as imputation or removal.

Handle any duplicate records in the dataset.

Correct any inconsistencies or errors in data entry.

* **Data Transformation:**

Utilise Pandas and NumPy functions to transform the dataset according to the analysis requirements.

Consider creating new features, normalizing data, or handling categorical variables.

* **Handling Outliers:**

Identify and address outliers in the dataset using the NumPy and Pandas methods.

Decide on an appropriate approach, whether it’s removing outliers or transforming them.

* **Data Validation:**

Validate the cleaned and transformed dataset to ensure that it aligns with the intended analysis goals.

Perform sanity checks on key variables and relationships within the data.

* **Documentation:**

Document the steps you took to clean and manipulate the dataset, including the reasoning behind your decisions.

Provide explanations for any transformations or imputations made during the process.

**Points to Remember**

1. Use meaningful variable names and comments to enhance code readability.
2. Experiment with different approaches to data cleaning and transformation to understand their impact on analysis.
3. Seek input from mentors or colleagues for best practices in data cleaning.