**Data Science Intern Case Study Document**

**Project Overview**

This Data Science project aims to analyze and draw meaningful insights from the given dataset. The project follows a structured data analysis process, including data cleaning, exploratory data analysis (EDA), and modeling. This project aims to analyze a dataset containing medical records of individuals who experienced side effects from various medications. The goal is to explore the relationships between demographic factors, medication use, and the occurrence of side effects. We also aim to uncover patterns related to chronic illnesses and allergies, and how these factors contribute to the likelihood of experiencing specific side effects.

**Dataset Overview**  
The dataset consists of patient demographics, medication history, and recorded side effects. It also includes information about allergies and chronic illnesses, both personal and familial.

**Libraries and Tools**

The following libraries were utilized to complete this project:

* *Pandas*: For data manipulation, cleaning, and aggregation.
* *NumPy*: For numerical operations and statistical calculations.
* *Matplotlib & Seaborn*: For visualizations to uncover patterns and relationships.
* *Scikit-learn*: For building machine learning models, where necessary.
* *Statsmodels*: For statistical testing and hypothesis verification.

**Environment**

Google Colab was used as the primary environment to write and execute the code.

**Steps Followed**

1. **Data Loading**

* The dataset was imported into the environment using Pandas. This allowed us to handle the dataset efficiently for exploration and cleaning.
* After the data set was loaded, the first and last 5 rows were fetched using the *head* and *tail* function. With these functions, the data set was examined.
* Using the *info* function, we obtained the data type of the columns in the data set and the number of non-empty rows.
* Using the *format* function, the number of rows and columns in our data set was reached. The column naming in the dataset has been adjusted.
* After adequately examining the data set, the total number of null values was obtained by using the *isnull* function to determine the number of empty fields.
* Since the naming of some columns was not correct, renaming was done with the *rename* function.

**2. Exploratory Data Analysis (EDA)**

EDA is performed to understand the structure and nature of the dataset. This includes:

* Checking for missing values
* Examining data distributions
* Identifying correlations between features
* Visualizing key features

**3. Data Preprocessing**

Several steps are undertaken to prepare the data for modeling:

* **Handling missing data**: Rows with missing values are either filled using statistical methods or removed.
* **Feature encoding**: Categorical variables are transformed into numerical representations using encoding techniques like one-hot encoding or label encoding.
* **Scaling**: Continuous features are scaled using standardization or normalization techniques to improve model performance.

**Conclusion**

This notebook offers a structured approach to conducting a complete data science case study, covering key aspects such as loading, cleaning, exploring, analyzing, visualizing, and interpreting health-related data. The goal is to extract valuable insights and make informed decisions based on the results obtained.