**Project Group Name: VeriSpor**

**Abstract**

In this project, we worked on the data in the train.csv file. We first examined the data and used various methods to make the data analyzable. We have done the necessary work to analyze, organize and clean up the missing data. After cleaning the data and making it analyzable, we observed the relationship of the data with each other with some correlation methods. Finally, we tried to visually analyze and understand the correlation results we observed.

**Steps**

1. **Load the train.csv:** The train.csv file was read and imported into a pandas dataframe (df). Various libraries are imported and the file is made ready to be read.

1. **Explore Data**: Firstly we used various codes to explore the dataset. We got a general information about the dataset, for example, what are the data types and how many are there. We learned about the data types in the dataset and observed the first 5 rows and columns in the dataset. Finally we displayed how many rows and columns the dataset consists of.

1. **Rename Columns:** After reviewing the data, we observed that the columns appeared unnamed and very confused, and we renamed the column names and columns are named and more regular.

1. **Analyze, Edit and Delete Missing Data:**  this stage, we first observed the missing data and total number of columns with missing data. After that We observed the missing values in all columns as a ratio because deleting data with more than half missing values will make our dataset cleaner and more meaningful. After deleting the columns with more than half missing values, we displayed the columns with the actual missing value.

1. **Fix Missing data with Pipeline:** Since the dataset has too many columns and too many missing data, we used the pipeline method. For numeric values, we could look at their averages for each column separately and manually fill in the missing values with the average, but there are too many columns and too many missing values. So, we used the pipeline method. in addition, the pipeline also takes the most frequent values instead of the mean for categorical values.
2. **Correlation Matrix for Data:**  We used a correlation heatmap to understand the relationship between the columns. We have enabled heatmap and required packages. however, the correlation heatmap with too many columns was not effective and it was not fully understood which columns could be analyzed with each other because there were too many columns.
3. **Visualization Methods:** Based on the correlation heatmap, we applied some visualizations methods.

**Challenges Encountered**

The data columns were not named and we had a hard time analyzing them because we didn't fully understand the purpose of the dataset. For example, we could not understand which columns to analyze with each other for the correlation heatmap. When we look at the correlation heatmap of the whole data set, there is no meaningful result because there are too many columns, the resulting graph cannot be read.

It was also difficult to analyze and organize the missing data correctly, as the number of missing data was quite high. To address this issue, more research has been done to identify the causes of missing data and how this data can be changed.

**REFERENCES**

Marloz. "Handling Missing Values with Scikit-Learn Pipelines". (2020). URL: <https://marloz.github.io/projects/sklearn/pipeline/missing/preprocessing/2020/03/20/sklearn-pipelines-missing-values.html>.

DataCamp. "Introduction to Data Visualization with Matplotlib". URL: <https://app.datacamp.com/learn/courses/introduction-to-data-visualization-with-matplotlib>.