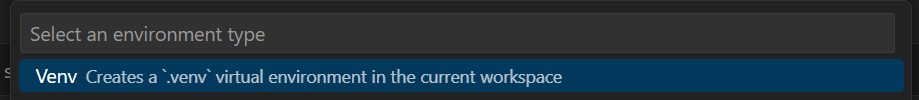
**Student Name – Sachin Shrinivas Kadam**

**Assignment Name – Data Cleaning Techniques**

* Apply data cleaning methods such as handling missing values using mean, median and mode imputation.
  + Decide whether the columns with missing values should be retained or removed. Justify your decision.

My Analysis & Response as below

* Toolset that I used
  + IDE – Visual Studio Code
  + Created a Virtual environment
    - 
  + Python
    - 
  + Download and install Pandas - pip install pandas
    - 
* I performed the below steps to handle Missing Values
  + I wrote a Python code, that I will upload when I submit the assignment.
  + Load the train.csv dataset into pandas
  + Identify the columns with Missing Values
  + Selection of appropriate imputation methods (mean, median, or mode) on the basis on the data type
  + Strategy & Justification for Imputation
    - **Age (Numerical)**: The **Age** column has a significant number of missing values. In this case, **median imputation** is used because it is less sensitive to outliers and is a good choice for age distribution, which may not always be symmetric. **Mean imputation** could be skewed by a few very high or very low values.
    - **Fare (Numerical)**: The **Fare** column generally has fewer missing values and is more likely to be symmetrically distributed. Since Fare values are mostly continuous, **mean imputation** is used here to fill the missing values.
    - **Embarked (Categorical)**: The **Embarked** column represents the port of embarkation and is categorical. **Mode imputation** is appropriate because it replaces missing values with the most frequent value, which makes sense for categorical data like embarkation port.
  + Columns to be Retained
    - **Age**: Even though it has missing values, the **Age** column is crucial for predicting survival chances, as it may play a role in survival probability. Removing this column would discard valuable information. Hence, **it should be Retained, and the missing values should be imputed.**
    - **Fare**: This column has fewer missing values and is an important feature for analysis. **Removing this column is not recommended, and should be Retained**.
    - **Embarked**: This is also an important feature for the Titanic dataset (whether the passenger embarked at C, Q, or S). Removing it would remove a key factor in determining survival probability. **It should be Retained.**
  + Columns to be Remove, Excessive Missing Values
    - **Cabin**: The **Cabin** column has a significant portion of missing values (almost 70% in the train.csv dataset). Depending on my analysis, I might choose to **drop/remove** this column because **imputing** it could introduce too much noise, and it might not provide meaningful predictive power. However, one may keep it for reason to believe that the missing values are random or can plan to engineer features from it (e.g., extracting deck information)
  + Summary
    - As per my analysis I’ll prefer to Retain **Age**, **Fare**, and **Embarked** columns**.**
    - I’ll consider dropping **Cabin** due to high percentage of missing values
    - The overall effort of Imputing Missing Values instead of removing them. We can preserve the dataset’s size and retain valuable information, which can help with Modeling (most likely predictive modeling)