Monthly Progress Summary *(September)*

**I. Data Understanding Summary**

Question:   
Please provide a summary of your team's “Data Understanding” accomplishments during the month of September. Remember to include any relevant links to your work (e.g., a Python notebook showcasing your team’s Exploratory Data Analysis work – e.g., statistical analysis and visualizations related to key variables, patterns, relationships, data quality issues).   
Student Team Response:  
We started by reviewing the dataset documentation to understand sources, variables, etc. During our Exploratory Data Analysis (EDA), we utilized descriptive statistics to understand the distribution and summary statistics of key variables. We employed visualization techniques such as histograms, heatmaps, and pie charts, to identify patterns and relationships.. We identified and documented some data quality issues, including missing values, word changes, and noted their potential impact on our analysis). You can review our EDA findings and observations here: [*https://colab.research.google.com/drive/1pfPBTxIPOYilSb2bt0Ebs\_EAa3mdm4U2*](https://colab.research.google.com/drive/1pfPBTxIPOYilSb2bt0Ebs_EAa3mdm4U2)

**II. Data Preparation Summary**

Question:   
Please provide a summary of your team's “Data Preparation” accomplishments during the month of September. Remember to include any relevant links to your work (e.g., a Python notebook showcasing your team’s data preprocessing work – e.g., data cleaning, formatting, missing value imputation, outlier handling, feature engineering).   
Student Team Response:  
In terms of Data Preparation, we made good progress in ensuring our dataset is suitable for modeling. We performed data cleaning tasks such as handling missing values. Also, we made changes to columns so that all objects are of the same type (i.e. The column ‘Year’ includes now items of only type ‘int’). The findings from our Data Understanding phase helped us identify missing values in certain variables, and we imputed them by deleting them since their missingness does not impact the model greatly and there is no means to be calculated since these columns are non-numerical. You can review our work here: <https://colab.research.google.com/drive/1pfPBTxIPOYilSb2bt0Ebs_EAa3mdm4U2>

**III. Lessons Learned and Challenges**

Question:   
Reflecting on the Data Understanding and Data Preparation phases, what were the key insights or challenges your team encountered? How did you address them? Share any important lessons learned that can help guide future steps in the project.  
  
Student Team Response:  
  
We encountered challenges in dealing with correlations, they are quite hard to identify in this data set, and the data itself is not broad enough but with guidance from our AI Studio TA we were able to apply techniques to address them. One key lesson learned is the need for continuous iteration between the Data Understanding and Data Preparation phases. This is because insights gained during EDA can significantly influence the subsequent preprocessing steps. We will need to look back to these once we finish testing our first model and see what we can change to make it better.

**IV. Next Steps (Data Understanding and Prep)**

Question:   
Given your current progress, what additional tasks does your team need to complete in connection with the Data Understanding and Data Preparation phases of your project? What is your plan to complete these tasks?

Student Team Response:

*Based on guidance received from our Challenge Advisor and AI Studio TA, we plan to conduct more in-depth analysis to identify any additional patterns or anomalies that require further investigation or preprocessing steps. Also we need to write a proper explanation and descriptive storytelling for EDA that we conducted and results we have found to be able to present it to our supervisors. We plan to explore advanced statistical techniques such as factor analysis and cluster analysis to uncover more complex relationships within the data. Additionally, we will improve and change the feature selection process, considering the importance, correlation, and potential multicollinearity of variables when running a model, we might have to change the structure of the data inputs or group some variables to improve the model’s performance. This will help us understand whether certain variables are important to us or they are even more confusing for the model. By completing these tasks, we will be able to enhance the quality and relevance of our data for subsequent stages of the project.*

**V. Request for AI Studio TA Support**

Question(s):   
What additional support do you need from your AI Studio TA? Please structure your response as specific questions, related to the Data Understanding and Data Preparation phases of your project. Consider areas where you may require specific guidance, clarifications, suggested approaches, or suggested resources. Your AI Studio TA will review these questions and work through them with you in an upcoming meeting or chat.   
  
Student Team Response:

1. How can we address issues related to class imbalance or skewed distribution in our NLP classification task?
2. Are there any ethical considerations or bias mitigation strategies we should be aware of when working with NLP models and sensitive data?
3. What methods or metrics should we use to evaluate the performance of our NLP model, considering the nuances of natural language data?
4. Can you provide guidance on advanced statistical techniques or exploratory analysis methods that can help uncover more complex relationships/patterns in our dataset?
5. Are there any specific feature selection algorithms or best practices that you recommend we consider for our project, given our dataset characteristics and modeling objectives?
6. Are there any potential challenges or pitfalls we should anticipate during the later stages of the project, considering our current progress?

###