Mini project: Performing EDA on a dataset



You are provided with the above dataset. You have to produce an exploratory data analysis on the dataset. Following are some of the information which you need to document into a (word or PPT) file. NOTE: For many of the questions there is no definite answer, in other words answers are subjective. So, feel free to make guesses too at places of required, but you need to provide reliable explanation for your guesses too.

1. What the dataset is about? What type of information can you gather by simply reading the column names and the values contained in the columns? SUGGESTION: Try to see through the data manually in Excel, before treading the data in Python.
2. What is the percentage of missing values in the whole dataset? Which column has the highest number of missing values?
3. How many columns have not more than just 1 value?
4. Calculate the statistical properties of columns having numerical values as well as at least 5 unique column values.
5. Create a bar plot for some selected categorical columns.
6. Create a histogram plot for some selected numerical columns.
7. Create a pair plot for the numerical columns in the dataset.
8. List down those pairs of numerical columns having the top 3 positive correlation values. Create a scatter plot for these 3 pairs.
9. List down those pairs of numerical columns having the top 3 negative correlation values. Create a scatter plot for these 3 pairs.
10. Bring up any other additional data insights from the EDA you perform.

Feel free to refer to documentations on the internet. But please do not copy any explanations or codes from other students, because even if some codes will be similar, explanations are subjective.