# **EDA Tasks**

|  |  |  |
| --- | --- | --- |
| **Tasks** | **Action** | **Results** |
| General info about the dataset | What are the dimensions? (no. of rows/columns) |  |
| Identify the target variable and predictors. Are you dropping any columns that will not contribute to further analysis/modelling process? |  |
| Your understanding of what each feature means/indicates. |  |
| Can your features be grouped in any meaningful way for analysis? (demographic, clinical, environmental, test, etc.) |  |
| Name of variable | Did you change the name or visualization or further coding purposes? |  |
| Variable types | Define by you and not by Python or R (categorical/numerical? nominal/ordinal/binary? discrete/continuous?) |  |
| Unique values | How many categories/classes are in each of your categorical variable?  \*How many unique values should a numerical column have to be considered a discrete numerical feature or a categorical feature? |  |
| Encoding | Are you encoding categorical features with numerical values? |  |
| Valid range | Do you define valid range based on existing knowledge? |  |
|  | How do you deal with the invalid range?  \*Some of you have a practice that skipping the invalid range, then using outlier detection as the way to deal with invalid range. What impacts do you think we may have? |  |
| Data quality issue | Both, characters and numerical are in the same column, what action did you investigate and what decision did you make? |  |
| Numerical variables are in different units of measurements, what implications it may have in the features? |  |
| Central tendency and variability | What is the mean, median, mode, IQR and quartiles of each feature? |  |
| Outliers | Univariate (boxplot, or histogram) |  |
| Bivariate (target variable vs [categorical, numerical] - data distribution comparison |  |
| Multivariate - Mahalanobis distance |  |
| What do you do with outliers? (remove? Impute? leave them as they are)  \*For imputation, there are multiple methods: log transformation, Winsorized, mean, median, mode, KNN, etc. |  |
| Typos and NaN/Missing values | How do you determine if entries are typos or placeholders for NaN? |  |
| How are NaN presented in the dataset? (empty cell, 0, -999, etc.) |  |
| If columns contain >50% missing values, drop from dataset. If <50%, what do you do? (if perform imputation, what method?)  \*Adjust the threshold to drop columns based on the dataset. What is your reasoning/source for choosing this threshold? |  |
| Distribution/Skewness | Histogram/Boxplot for numerical features  Bar plot for categorical features |  |
| Normality | Either visual methods (histogram, Q-Q plot, etc.) or statistical tests (Shapiro-Wilk, D’Agostino’s K², Kolmogorov-Smirnov, etc.)  \*What are the assumptions/limitations of your chosen analysis? |  |
| Correlation | Scatterplot for an overview of the relationship between numerical features. For relationship between numerical and categorical features, boxplot or bar chart |  |
| Are you dealing with linear or non-linear correlation? Other tests to detect correlation: Pearson, Spearman Rank, Kruskal-Wallis, mutual information, etc. |  |
| Multicollinearity | Use correlation matrix or VIF to detect |  |
| What do you do with predictors with high collinearity? (drop redundant variables? come up with interaction terms or polynomial terms and drop the redundant features? PCA?) |  |

For all analysis, consider when to perform stratified/class-wise analysis or overall/aggregate analysis?

Conclusion: what is the size of the so-called-cleaned dataset for construct machine learning model?