**Insights and Q&A**

This file would act as a tracking book of the Insights and challenges I acquired/ faced while building the **career-counsellor-ml** project. Along with the Insights, I would also keep the record of a few Questions and Answers which I think are important, and which really forced me to think about the concept deeply.

**I – Insights**

**Q – Questions**

**C – Challenges**

**I1 -** In the describe function, include = "all" parameter makes sure to include all the columns while drawing the statistics table.

**Q1 – Which type of graphs are used for which type of data?**Ans) Categorical variables 🡪 Bar plot (Matplotlib), Count plot (Seaborn)  
 Numerical variables 🡪 Histogram, Boxplot  
 Time-series 🡪 Line plot  
 Multiple Numeric variables 🡪 Heatmap (Just like in the project)

**Q2 – How to detect and handle missing values?**Ans) Missing values can be detected by “df.isnull( ).sum( )”. There are multiple ways to handle missing values – Dropping rows/ columns, Imputation methods (Replacing with Mean, Median and Mode, Interpolation methods, model-based imputation), Missingness as a feature.

**Q3 – Explain different Imputation methods for handling missing values.**Ans) Imputation methods are basically for filling missing values with estimations which helps in preserving data (dropping rows/ columns can lead to removing critical information).  
1) **Simple Imputation methods** 🡪  
**Mean Imputation** – Replace missing values with the mean of the non-missing values. Quite easy to implement but can lead to distortion in distributions if outliers are present. *Also, not suitable for categorical variables.*  
**Median Imputation** – Replacing the missing values with the median of the non-missing values. This is more robust to outliers as compared to mean. *Doesn’t consider relationship between variables and not suitable for categorical variables.*  
**Mode Imputation** – Replacing missing values with the mode of the non-missing values. Suitable for categorical variables but can reduce variability in data.

2) **Forward fill/ Backward fill** – Filling the values with the last (forward fill) /next (backward fill) observed non-missing value. *Useful for sequential data, when values are expected to be similar over time. But this assumes continuity, what if the data changes every second? This method would fail then*.

3) **Interpolation Techniques (Linear Interpolation)** – Estimates missing values based on values of neighbouring data points. *This method can capture the relationships in data more effectively. The only issue is it assumes linear relationship between data points*.

4) **Model based Imputation** -  
**Regression Imputation** – Missing values are predicted using a regression model based on other values in the dataset.   
**KNN Imputation –** Fills missing values by finding the K-nearest neighbours and observes their values to impute the missing data. (By maybe taking their mean, median or mode). *This method considers similarity and relationship between variables but can be computationally very expensive for large datasets.*

**Q4 – Explain missingness as feature and its types.**Ans) In some cases, the fact that the value is missing can be an information in itself, so a new binary variable can be created (0 for present, 1 for missing) for variables with missing values. Types of missing values -  
1) **MCAR (Missing Completely at Random) –** Standard deletion methods can be used for large dataset.  
2) **MAR (Missing at Random) –** Imputation methods are generally preferred.  
3) **MNAR (Missing Not at Random) –** Most challenging, requires complex modelling/ data collection efforts.

*Important: Write about the value error faced in step 6 soon after writing train\_test\_split function -* ***ValueError****: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less than 2.*

*This is due to stratify = y not having enough members in the class to distribute evenly*