Data Cleaning

Univariate Imputation

Till now, you have seen how missing values can create problems if present in the dataset. You have learnt about different types of missing values, namely MCAR, MAR and MNAR and the empirical rule for handling missing values. Now, there are various techniques to handle the missing values:

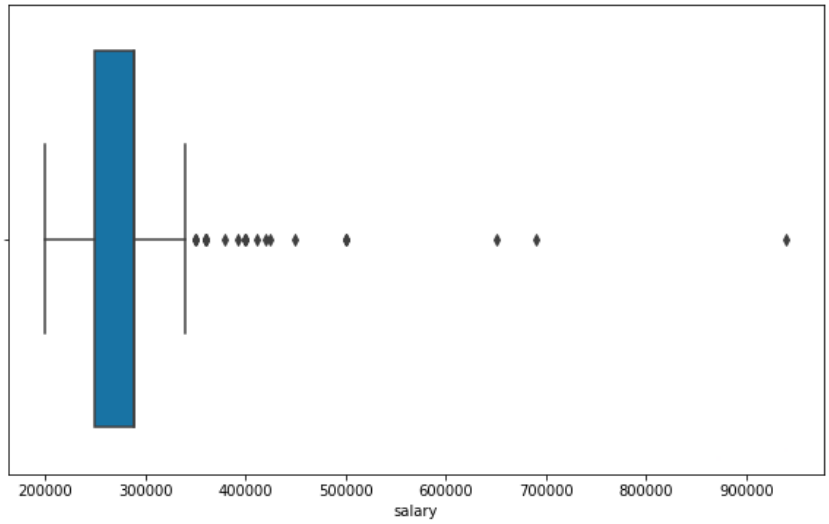
1. **Univariate Imputation**
    - Mean
    - Median
    - Mode
2. **Multivariate Imputation**
    - KNN Imputer
    - MICE

Here you will explore **univariate methods** for imputing missing values and learn how to decide which imputation technique to use.

Mean Imputation

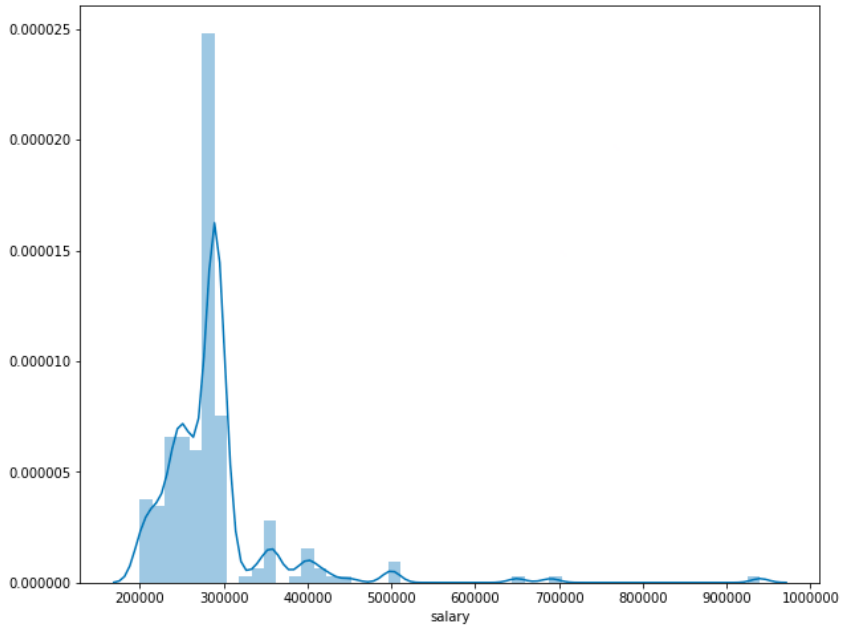
Mean Imputation replaces the missing values in a particular column using the mean of the observed values for that column. Mostly mean imputation is used when the data is distributed symmetrically around the mean. You can use box plots and distribution graphs to learn about the column and then decide if using mean imputation is a good idea in a particular case. Let's say you are working on the salary column and the boxplot for that is:

```
In [118]: fig, ax = plt.subplots(figsize=(10, 6))
          sns.boxplot(df.salary)

Out[118]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8e0ac37ed0>
```



From the box plot, you can observe that there are many data points on the right side of the maximum value of IQR which means outliers can be present in the data. In these cases, using mean imputation is not a good idea because the mean gets affected by outliers. You can also plot a distribution curve to know more about the skewness.

```
In [120]: fig, ax = plt.subplots(figsize=(10, 8))
          sns.distplot(df.salary)

Out[120]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8e0b0a7e50>
```



Here it is clear that the data is right-skewed. Mean imputation can be used when the data is normally distributed.

Median Imputation

As the mean gets affected by the outliers, this is not the case with the median. You can easily use median imputation when skewness is present in the data. It is to be noted that the median imputation is only possible with numerical data.

Mode Imputation

Another useful technique for replacing the missing values when the data is skewed is using mode. Also, the mode can be used with numerical and categorical data.