Data Wrangling Report

Introduction:

This document particularly describes the data wrangling steps that I undertook to prepare the IBM HR Analytics Employee Attrition & Performance dataset for the further process in the project. It explains what kind of steps were performed on this particular data set, how the missing values or the outliers handled.

Data Retrieval:

Dataset is in the open source Kaggle website and can be reached from this link. I loaded the dataset from here in csv format and read it in the jupyter notebook after importing necessary libraries.

Data Specifications:

The dataset has 1470 rows and 35 columns. Rows are observations from each employee and columns are from different features which are obtained in order to explain the employee attrition. The features data types consist of 27 integers and 8 objects. For some features, It is important to figure out their identity.

Field	1	2	3	4
Education*	Below College	College	Bachelor	Master
Environment Satisfaction	Low	Medium	High	Very High
Job Involvement	Low	Medium	High	Very High
Job Satisfaction	Low	Medium	High	Very High
Performance Rating	Low	Good	Excellent	Outstanding
Relationship Satisfaction	Low	Medium	High	Very High
Work Life Balance	Bad	Good	Better	Best

* For ‘Education’ field, 5 stands for ‘Doctor’.

List of attributes are presented below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
Age                         1470 non-null int64
Attrition                   1470 non-null int64
BusinessTravel              1470 non-null object
DailyRate                   1470 non-null int64
Department                  1470 non-null object
DistanceFromHome            1470 non-null int64
Education                   1470 non-null int64
EducationField              1470 non-null object
EmployeeCount               1470 non-null int64
EmployeeNumber              1470 non-null int64
EnvironmentSatisfaction     1470 non-null int64
Gender                      1470 non-null object
HourlyRate                  1470 non-null int64
JobInvolvement              1470 non-null int64
JobLevel                    1470 non-null int64
JobRole                     1470 non-null object
JobSatisfaction             1470 non-null int64
MaritalStatus               1470 non-null object
MonthlyIncome               1470 non-null int64
MonthlyRate                 1470 non-null int64
NumCompaniesWorked          1470 non-null int64
Over18                      1470 non-null object
OverTime                    1470 non-null object
PercentSalaryHike           1470 non-null int64
PerformanceRating           1470 non-null int64
RelationshipSatisfaction    1470 non-null int64
StandardHours               1470 non-null int64
StockOptionLevel            1470 non-null int64
TotalWorkingYears           1470 non-null int64
TrainingTimesLastYear       1470 non-null int64
WorkLifeBalance             1470 non-null int64
YearsAtCompany              1470 non-null int64
YearsInCurrentRole          1470 non-null int64
YearsSinceLastPromotion     1470 non-null int64
YearsWithCurrManager        1470 non-null int64
dtypes: int64(27), object(8)
memory usage: 402.0+ KB
```

Data Preprocessing:

I searched for missing values in every features of dataset, all features look like having 1470 non-null entries. However, missing values can be encoded in a number of different ways, such as by zeroes, or questions marks. For that reason, I checked both missing values and duplicate values in the dataset. Luckily, it was okay to continue to next step.

I observed 5 random sample records in the dataset to grasp the general intuition about whole picture. Besides that, I explored the statistical attributes of each features such as their mean, standard deviation, interquartile values in order to detect outliers. This research also gave me a general impression about unique and top values for each attributes in addition to their frequencies in the dataset. I made double checks on some of features in order to make sure that everything is good to go. Those results were also okay.

I inspected the useless features in order to drop in the dataset. “Over 18”, ”StandardHours”, and “EmployeeCount” had only one unique value for each observations and that did not impact or change anything in the data. For that reason, I dropped those three useless columns.

To be able to use effectively in the further steps, I reassigned the response variable (Attrition) which had “Yes” and “No” values previously. They were assigned to 1 and 0 respectively. After that, I moved the response variable to the last column place.

The dataset has 8 object types which are 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'. To be able have more memory usage and become fast, I changed object type to category type in the dataset. At first memory usage was 402.0+ KB, and after changing the data types, it became 298.3 KB.