HR Analytics Data Wrangling

Data Clean Up:

The dataset used in this project is fairly clean and contains 4 csv files.

1. Employee Survey - Contains the feedback provided by the employee
2. General Data - Contains general data about the employee such as Dept, distance from home, gender, age etc
3. In Time - Contains the login time of the employee
4. Out Time - Contains the logout time of the employee
5. Manager Survey - Contains the feedback given by the manager about the employee

The file General Data contains a few columns that do not provide any relevance to the data and will not contribute towards the analysis.

```
gen_data_df = pd.read_csv('general_data.csv')
```

```
gen_data_df.isna().sum()
```

```
Age                        0
Attrition                  0
BusinessTravel             0
Department                 0
DistanceFromHome           0
Education                  0
EducationField             0
EmployeeCount              0
EmployeeID                 0
Gender                     0
JobLevel                   0
JobRole                    0
MaritalStatus              0
MonthlyIncome              0
NumCompaniesWorked        19
Over18                     0
PercentSalaryHike          0
StandardHours              0
StockOptionLevel           0
TotalWorkingYears          9
TrainingTimesLastYear      0
YearsAtCompany             0
YearsSinceLastPromotion    0
YearsWithCurrManager       0
dtype: int64
```

Columns ‘Over18’,’JobLevel’,’JobRole’ do not add any significance and are deleted from the data set.

The In Time and Out Time csv had missing column names. By comparing with the other csvs, we can conclude that the missing column corresponds to employee ID.

Missing Data:

Employee Survey:

Employees have given feedback on three parameters (Environment Satisfaction, Job Satisfaction and WorkLifeBalance). Ratings are in the range of 1-5.

```
emp_sur_df = pd.read_csv('employee_survey_data.csv')
```

```
emp_sur_df.shape
```

```
(4410, 4)
```

```
emp_sur_df.isna().sum()
```

```
EmployeeID                 0
EnvironmentSatisfaction    25
JobSatisfaction            20
WorkLifeBalance            38
dtype: int64
```

```
emp_sur_df.dropna(inplace=True)
emp_sur_df.shape
```

```
(4327, 4)
```

Since the percentage of missing values is approximately 2%, all the rows with NA will not be considered.

In Time and Out Time:

1. Non Working days (weekends, public holidays) contain NA which can be removed as it is a valid NA.
2. If the In time and Out time for a single employee is “NA” on the same date, we can conclude the employee was on a PL/SL. This is also a valid NA and can be removed.
3. Using the In time and out time, I have created a new column “Average”, which is the amount of time spent by each employee in office.

average
07:43:08.288135
07:00:47.665289
07:11:37.242553
08:00:22.228571
06:55:24.853448

Outliers:

Initial analysis does not show any outliers.