**Employee Attrition**

by Machine Learning using Python

Created by-

Anubhab Nandy, Kalyani Government Engineering College, 10200116059 of 2016-2017

Anubhav Jain, Kalyani Government Engineering College, 10200116059 of 2016-2017

Abhinandan Singh, Kalyani Government Engineering College, 10200116059 of 2016-2017

Calvin Moktan, Kalyani Government Engineering College, 10200116059 of 2016-2017

Prity Choudhary, Kalyani Government Engineering College, 10200116059 of 2016-2017

Index

|  |  |  |
| --- | --- | --- |
| **Serial No.** | **Contents** | **Page Number** |
| 1 | Acknowledgement | 3 - 7 |
| 2 | Scope of Project | 8 |
| 3 | Introduction of  useful terms | 9 - 25 |
| 4 | Data Description | 26 - 56 |
| 5 | Code | 57 - 64 |
| 6 | Output | 65-74 |
| 7 | Conclusion |  |

**Acknowledgement**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessings, help and guidance given time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

-Anubhav Jain

**Acknowledgement**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessings, help and guidance given time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

-Anubhab Nandy

**Acknowledgement**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessings, help and guidance given time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

-Abhinandan Singh

**Acknowledgement**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessings, help and guidance given time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

-Calvin Moktan

**Acknowledgement**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessings, help and guidance given time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

-Prity Choudhary

**Project Scope**

Employers generally consider attrition a loss of valuable employees and talent. However, there is more to attrition than a shrinking workforce. As employees leave an organization, they take with them much-needed skills and qualifications that they developed during their tenure. On the other hand, junior professionals with promising qualifications can then succeed into higher level positions or business owners can introduce more diversity in experience or expertise. Accordingly, there are benefits and disadvantages to attrition. The cost of attrition can be relatively enormous.

**Employee Attrition**

Employers generally consider attrition a loss of valuable employees and talent. However, there is more to attrition than a shrinking workforce. As employees leave an organization, they take with them much-needed skills and qualifications that they developed during their tenure. On the other hand, junior professionals with promising qualifications can then succeed into higher level positions or business owners can introduce more diversity in experience or expertise. Accordingly, there are benefits and disadvantages to attrition.

### **Attrition Costs**

The cost of attrition can be relatively enormous. Attrition from retirement or resignation diminishes the workforce, demanding additional work hours and dedication from remaining employees. Whereas long-term workers have established bonds with customers and clients, attrition can reduce this rapport, running the risk of losing them to a competitor. Losing clientele affects revenue, profitability and business reputation.

Attrition from retirement or resignation diminishes the workforce, demanding additional work hours and dedication from remaining employees. Whereas long-term workers have established bonds with customers and clients, attrition can reduce this rapport, running the risk of losing them to a competitor. Losing clientele affects revenue, profitability and business reputation.

Despite the negatives of attrition, healthy attrition -- or, desirable turnover -- can positively affect organizations. Losing employees with poor performance records can boost employee morale, employee engagement and productivity among the current workforce. Moreover, attrition can be encouraging to young professionals seeking promotion and upward mobility.

**Attrition Analysis**

Employee attrition analysis refers to the employee attrition rate in a company. This analysis helps SaaS companies identify the cause of the churn and implement effective strategies for retention. Gainsight understands the negative impact that churn rate can have on company profits. Named as the "2014 cool vendor for CRM sales" (by Gartner), Gainsight’s customer intelligence and retention process automation technology:

• Gathers available employee behavior, transactions, demographics data and usage pattern

• Converts structured and unstructured data/information into meaningful insights

• Utilizes these insights to predict customers who are likely to churn

• Identifies the causes for churn and works to resolve those issues

• Engages with customers to foster relationships

• Implements effective programs for customer retention Statistics show that acquiring new customers can cost five times more than retaining existing customers. Gainsight performs customer churn analysis to reduce churn, control retention and improve performance.

**How we propose to solve:**

• Analyse the data frame

• Drop the columns which have high multi collinearity and low contribution towards ‘label’.

• Apply different classification models and note down the scores for each models.

• Select the models with the best scores

**Some machine learning methods**

Machine learning algorithms are often categorized as supervised or unsupervised.

* **Supervised machine learning algorithms**can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
* In contrast, **unsupervised machine learning algorithms**are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

**Linear Regression**

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a**regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression:

While training the model we are given:

: input training data (univariate – one input variable(parameter))

: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best and values.

: intercept

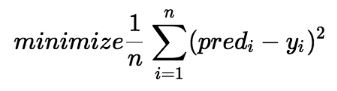
: coefficient of

Once we find the best and values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of for the input value of.

How to update and values to get the best fit line ?

**Cost Function (J):**

By achieving the best-fit regression line, the model aims to predict value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted value (pred) and true value ().



Cost function() of Linear Regression is the Root Mean Squared Error(RMSE) between predicted value (pred) and true value ().

**Gradient Descent:**

To update and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random and values and then iteratively updating the values, reaching minimum cost.

**Logistic Regression**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Sometimes logistic regressions are difficult to interpret; the Intellectus Statistics tool easily allows you to conduct the analysis, then in plain English interprets the output.

**Overfitting-** When selecting the model for the logistic regression analysis, another important consideration is the model fit.  Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds (typically expressed as R²).  However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit.

**Assumptions in Logistic Regression:**

* The dependent variable should be dichotomous in nature (e.g., presence vs. absent).
* There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29.
* There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors.

**Decision Tree**

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model.

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers the to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

**Confusion Matrix**

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

**Accuracy**:

The accuracy is not always an adequate performance measure. Let us assume we have 1000 samples. 995 of these are negative and 5 are positive cases. Let us further assume we have a classifier, which classifies whatever it will be presented as negative. The accuracy will be a surprising 99.5%, even though the classifier could not recognize any positive samples.

**Recall** **or True Positive Rate**:

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

recall =

**True Negative Rate**:

**Precision**:

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

Precision =

**Multi class case**

To measure the results of machine learning algorithms, the previous confusion matrix will not be sufficient. We will need a generalization for the multi-class case.

**Correlation**

Correlation refers to a mutual relationship or association between quantities. In almost any business, it is useful to express one quantity in terms of its relationship with others. For example, sales might increase when the marketing department spends more on TV advertisements, or a customer's average purchase amount on an e-commerce website might depend on a number of factors related to that customer. Often, correlation is the first step to understanding these relationships and subsequently building better business and statistical models.

Correlation is a useful metric because-

* Correlation can help in predicting one quantity from another
* Correlation can (but often does not, as we will see in some examples below) indicate the presence of a causal relationship
* Correlation is used as a basic quantity and foundation for many other modelling techniques

**Data Preprocessing**

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Data Preprocessing is needed for-

• For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

• Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

# **MinMax Scaling**

An alternative approach to Z-score normalization (or standardization) is the so-called Min-Max scaling (often also simply called "normalization" - a common cause for ambiguities). In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

**Random forest classifier**

Random forest is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forest creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

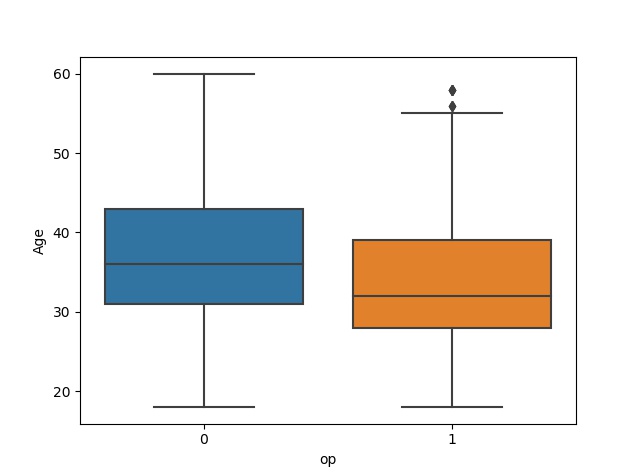
Random forest has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

**Data Description**

**Boxplot of attrition vs age of employee**

=Attrition

=Age of employee



Data Type-Integer

Value type: continuous

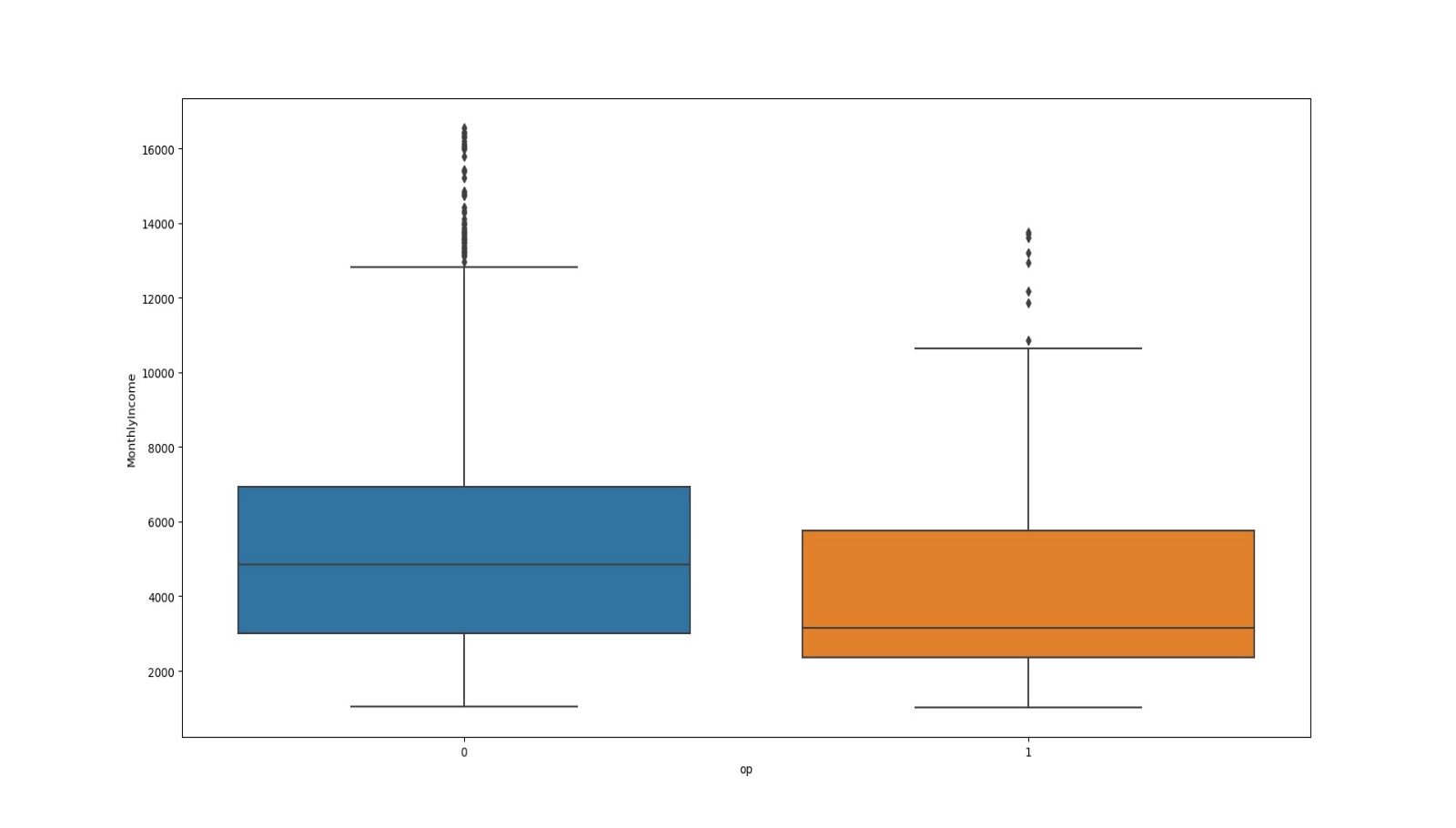
Null percentage: 0

It’s a depending graph and few outliers present

**Boxplot of attrition vs monthly income**

= Attrition

=Monthly Income

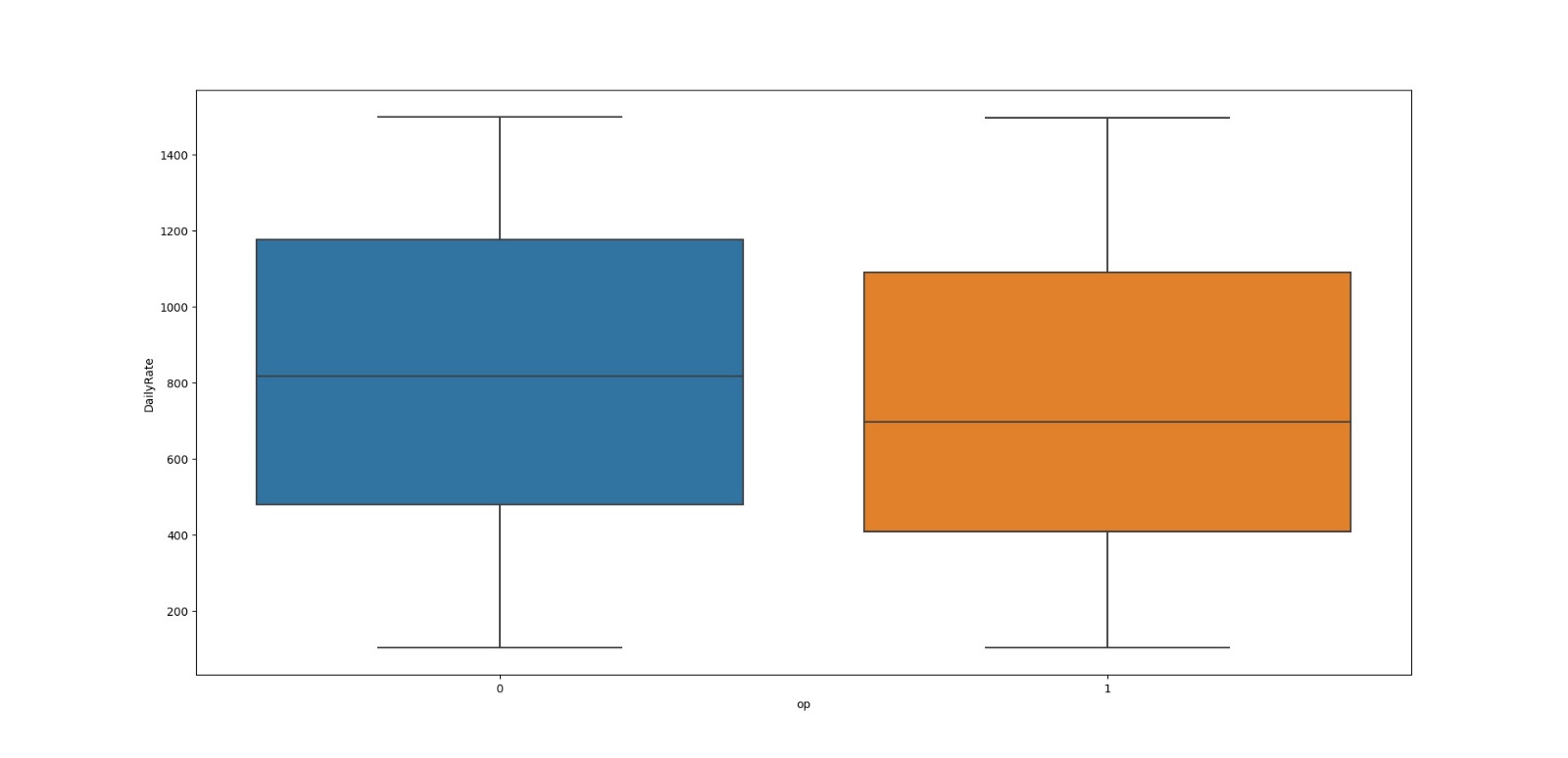


Monthly income is depending but outliers are present.

**Boxplot of attrition vs daily rate**

=Attrition

=Daily Rate

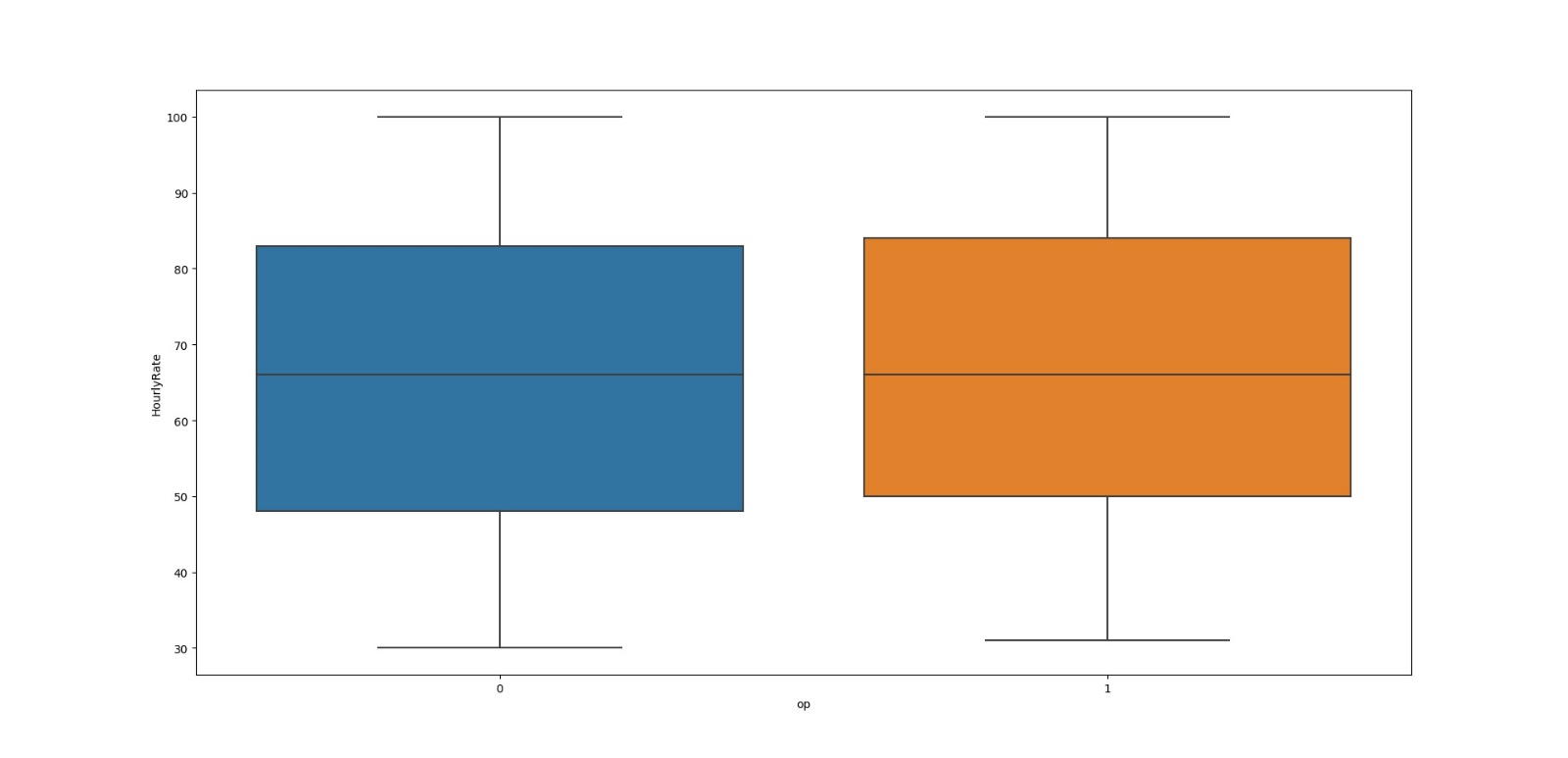


Slight dependencies can be concluded from the graph.

**Boxplot of attrition vs hourly rate.**

=Attrition

=Hourly Rate

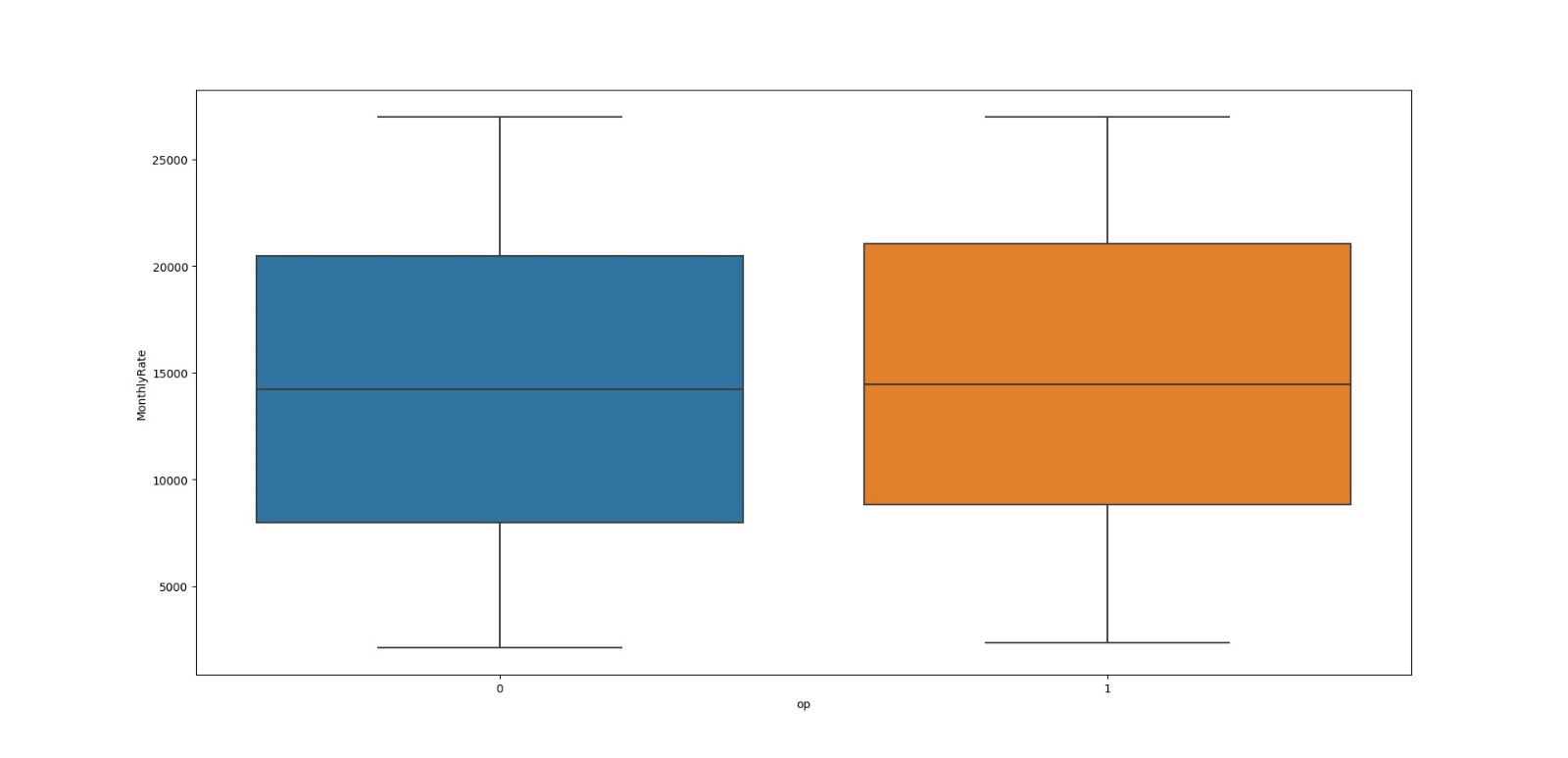


We can conclude that its almost independent.

**Boxplot of attrition vs monthly rate**

=Attrition

=Monthly Rate

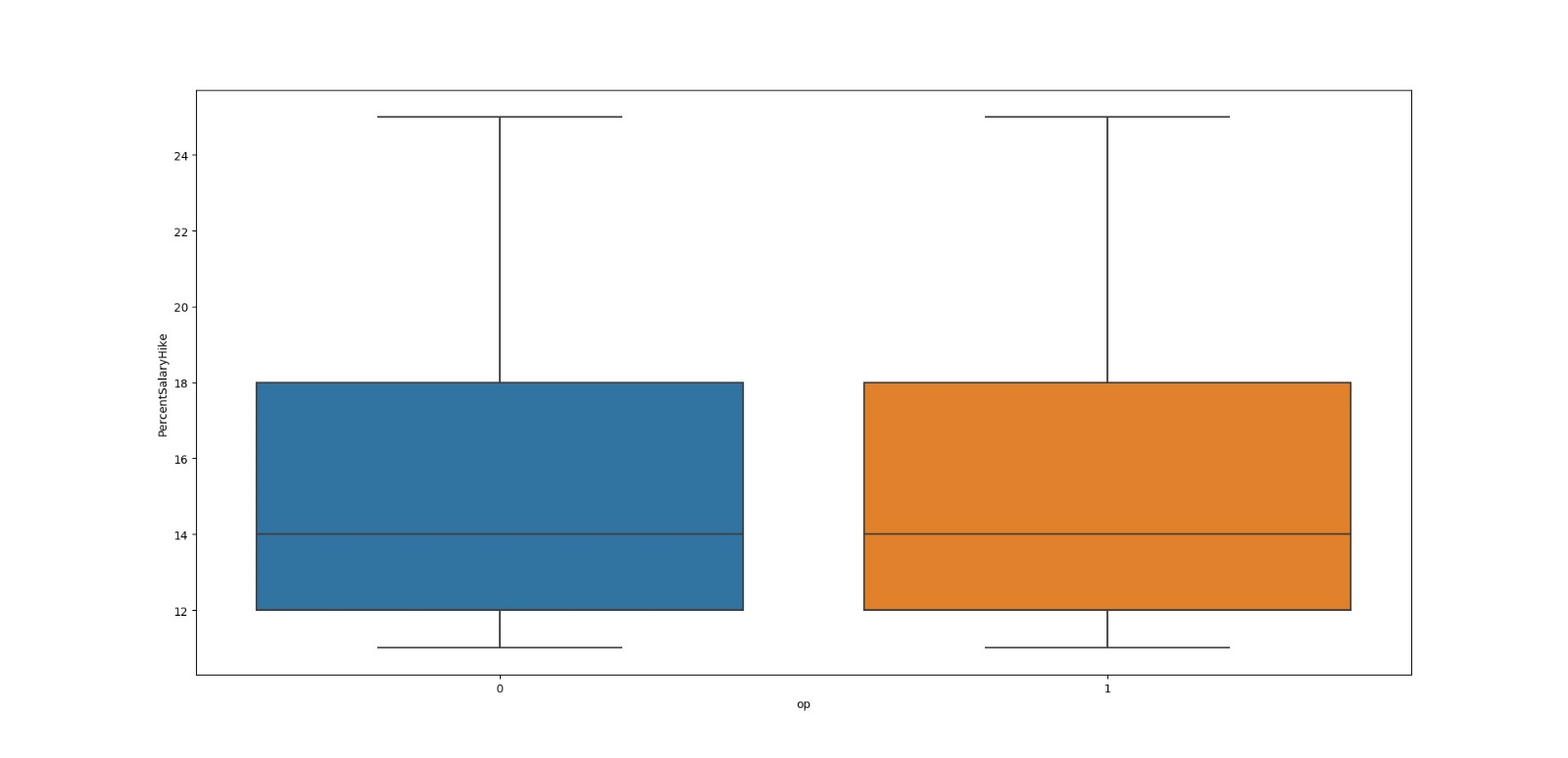


Not much dependent.

**Boxplot of attrition vs percentage salary hike**

=Attrition

=Percentage Salary Hike

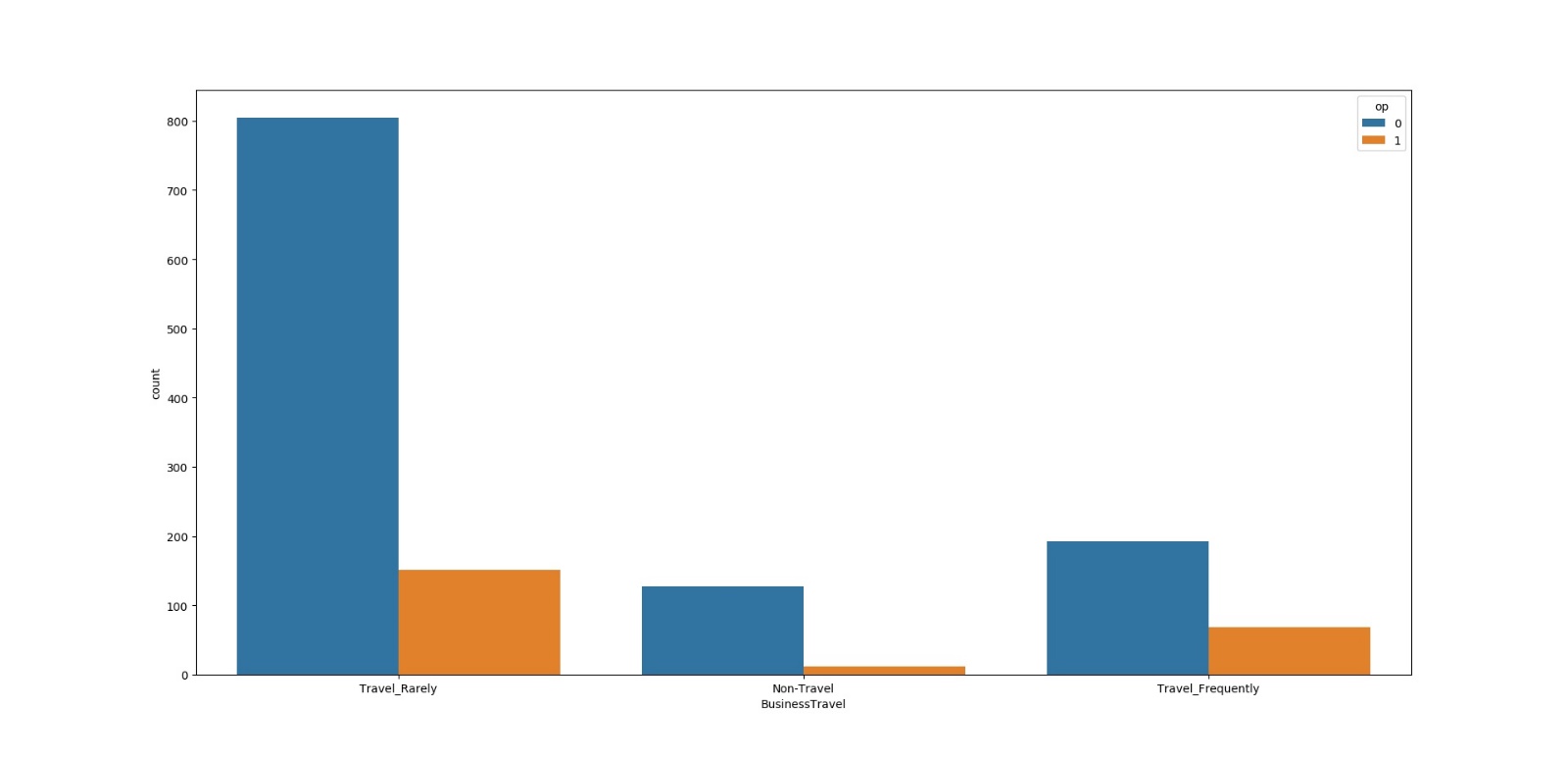


It’s almost independent.

**Countplot of attrition vs business travel**

=Attrition

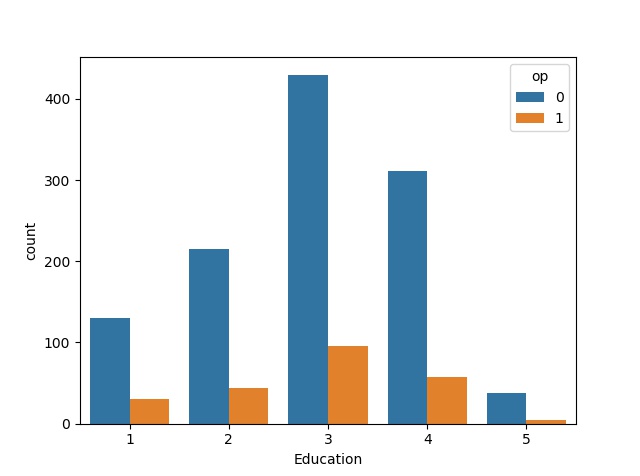
y=Business Travel



**Countplot of attrition vs count**

=Education

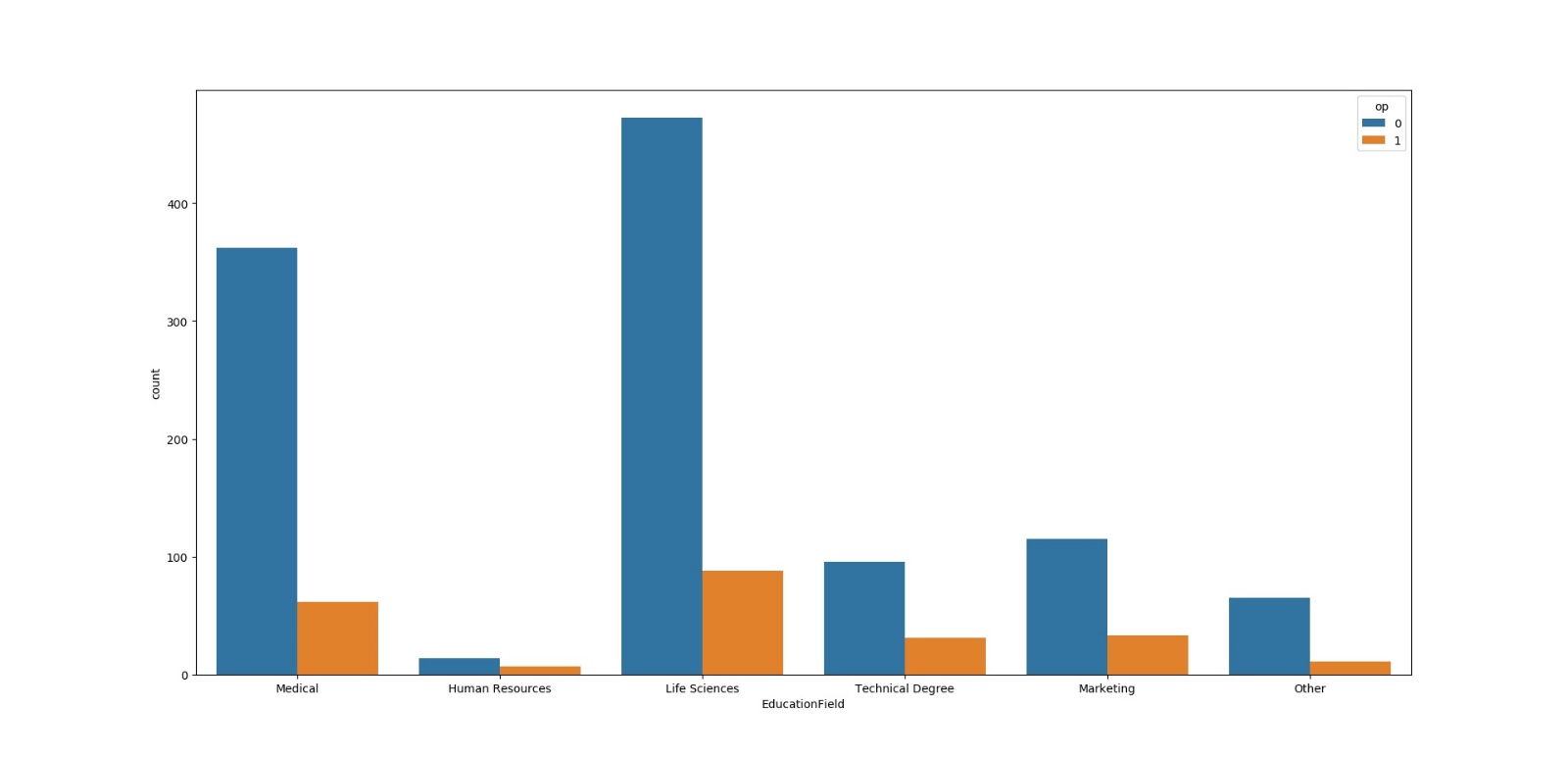
=Count



**Countplot of education field vs count**

=Education Field

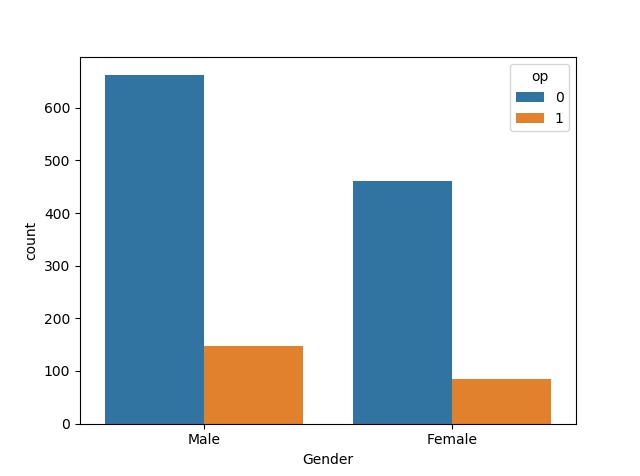
=Count



**Countplot of gender vs count**

=Gender

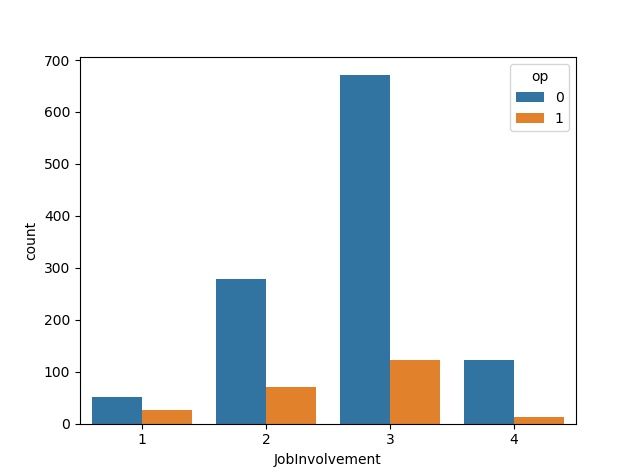
=Count



**Countplot of job involvement vs count**

=Job Involvement

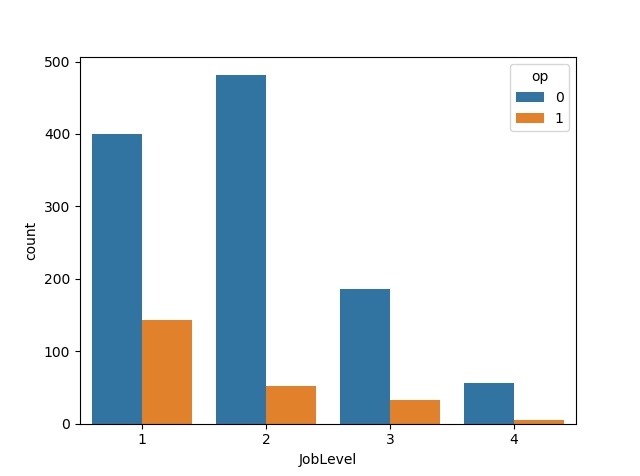
=Count



**Countplot of job level vs count**

=Job Level

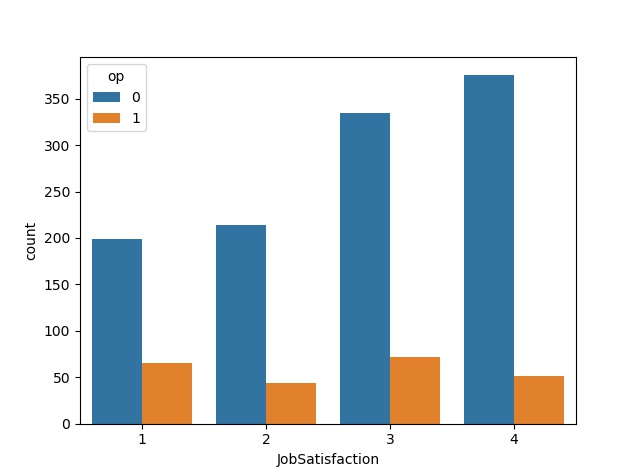
=Count



**Countplot of Job Satisfaction vs. Count**

=Job Satisfaction

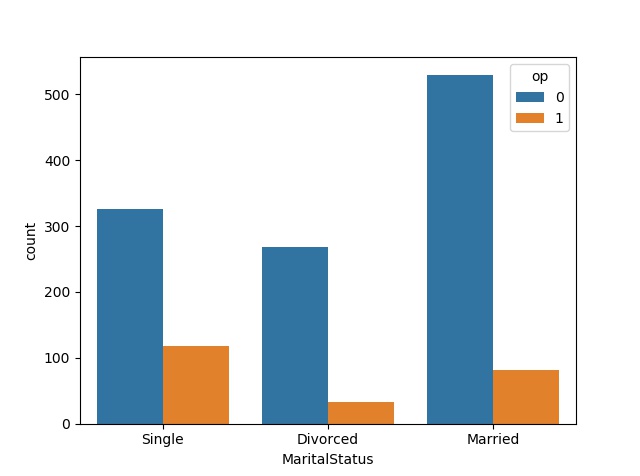
=Count



**Countplot of Divorced Martial Status vs count**

=Marital Status

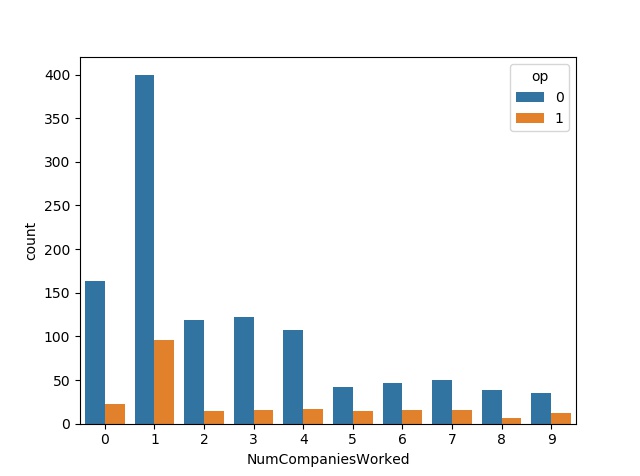
=Count



**Countplot of Number of Companies worked vs count**

=Number of Companies worked

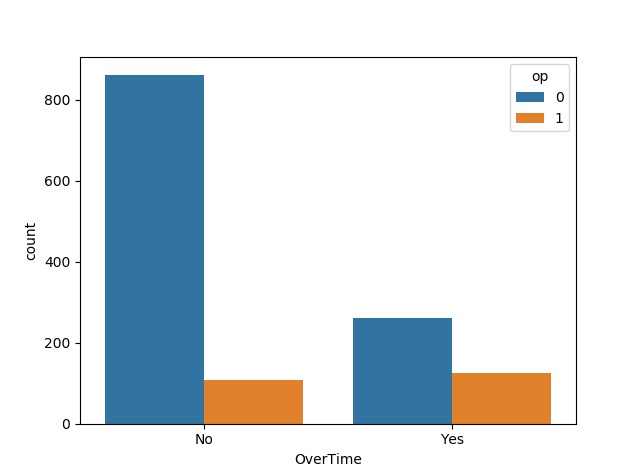
=Count



**Countplot of Overtime vs count**

=Overtime

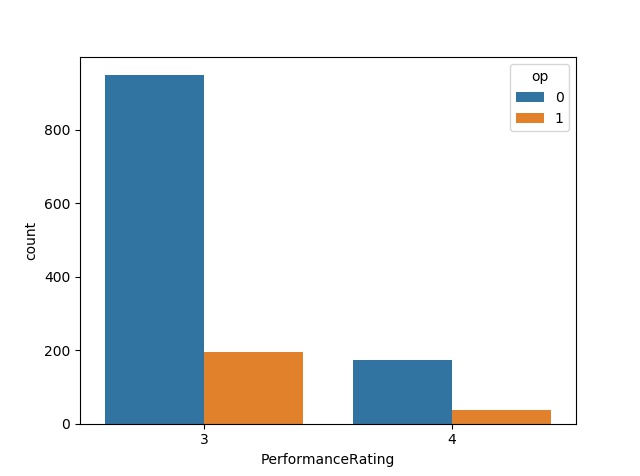
=Count



**Countplot of Performance Rating vs count**

=Performance Rating

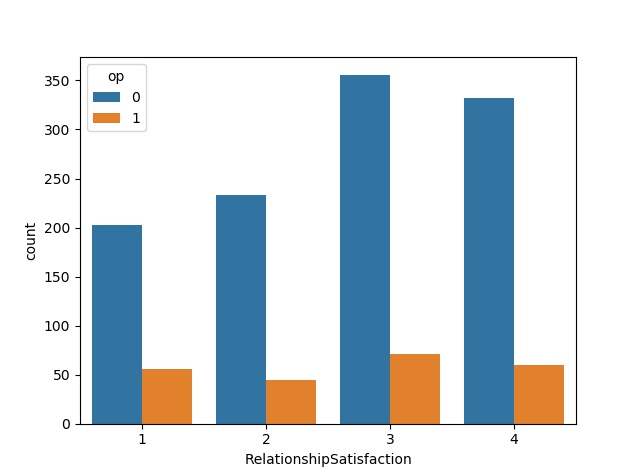
=Count



**Countplot of Relationship Satisfaction vs count**

=Relationship Satisfaction

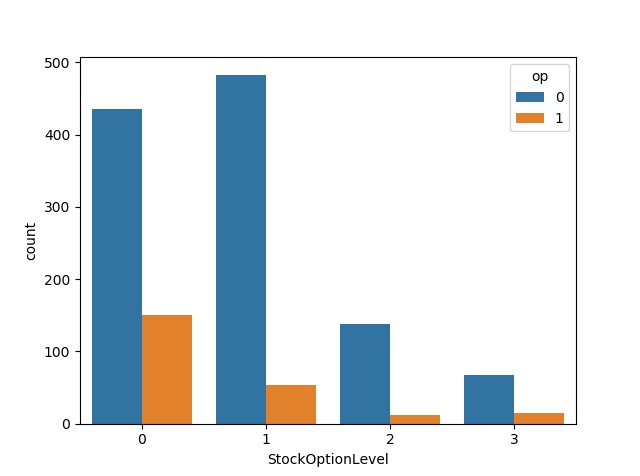
=Count



**Countplot of Stock Option Level vs. Count**

=Stock Option Level

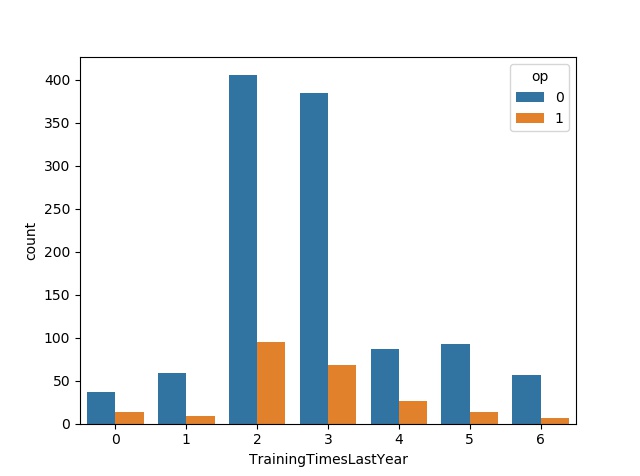
=Count



**Countplot of Training Times Last Year vs. Count**

=Training Times Last Year

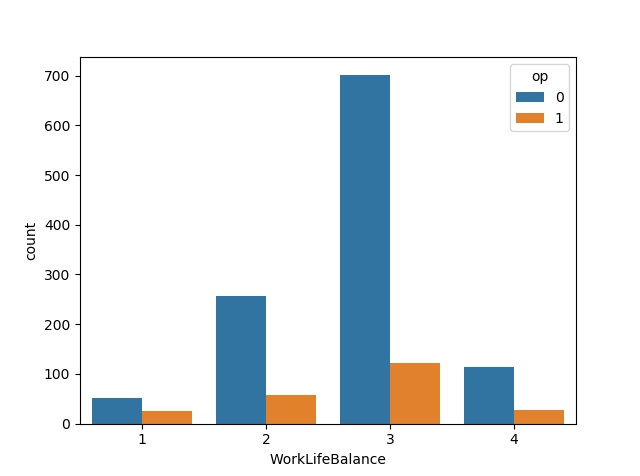
=Count



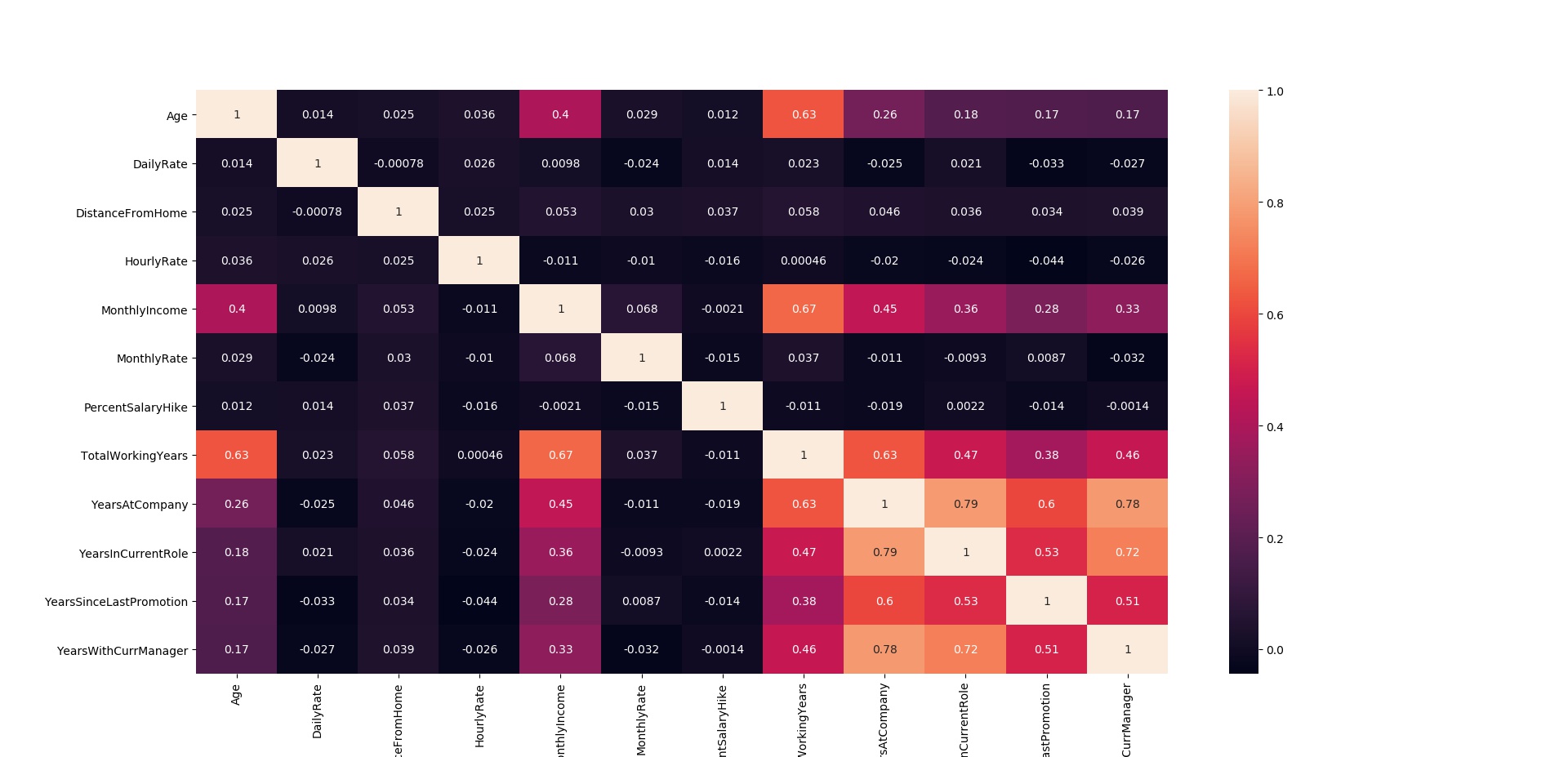
**Countplot of Work Life Balance vs. Count**

=Work Life Balance

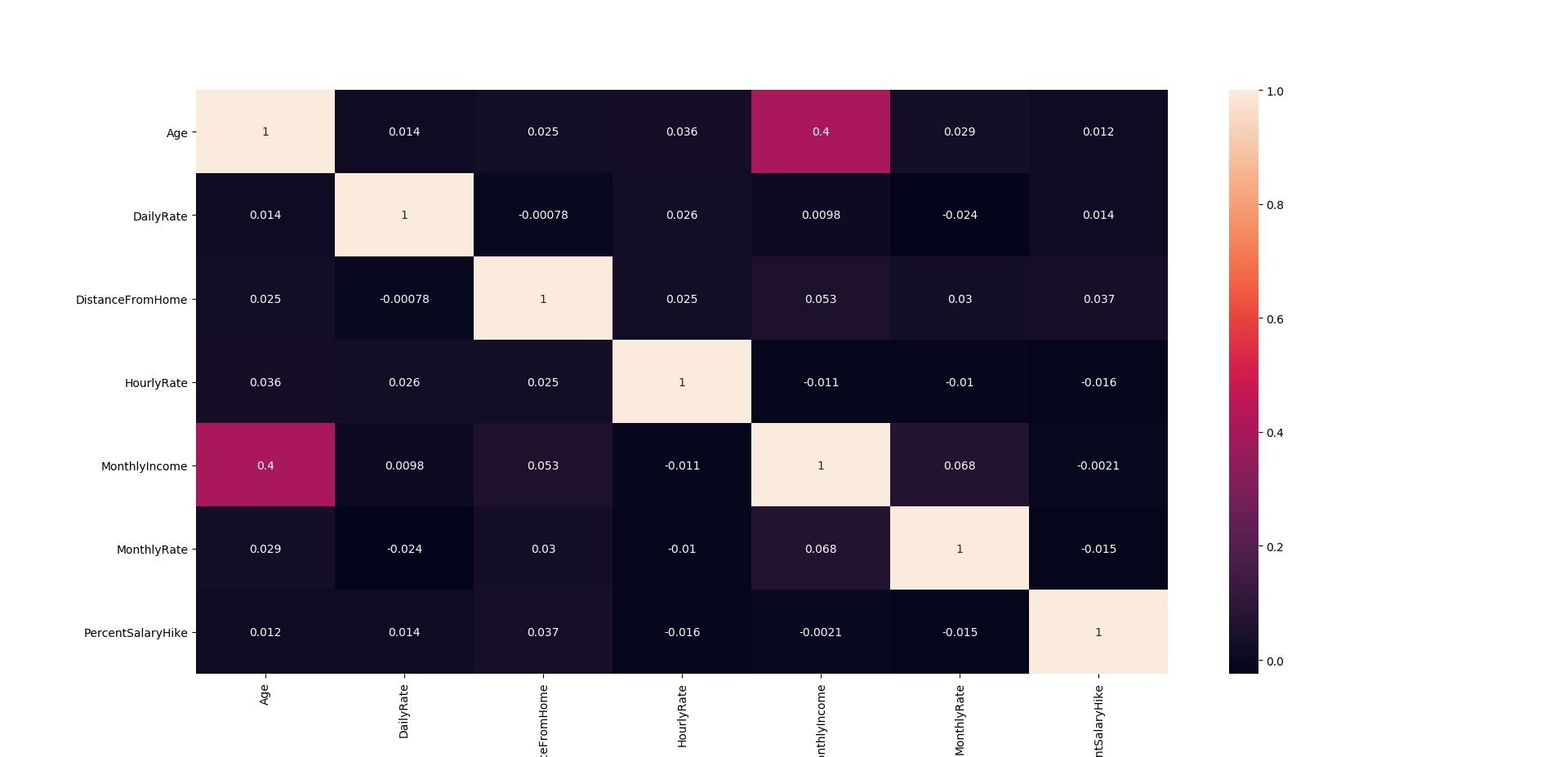
=Count



**HeatMap before deletion of some columns:**

****

**HeatMap after deletion of some columns:**



**Before some columns are deleted, we gather that:**

1.age and total working years are correlated:0.63

2.Monthly income and total working years are

correlated:0.67

3.years at company and total working years are correlated: 0.63

4.years at company years, in current role, years since last promotion, years with current manager all 4 are correlated >0.6 coefficient

that's why we are deleting 4 continuous columns

as

1. they are highly correlated

2.they have too many unique values so we cannot consider them as categorical columns

they are:

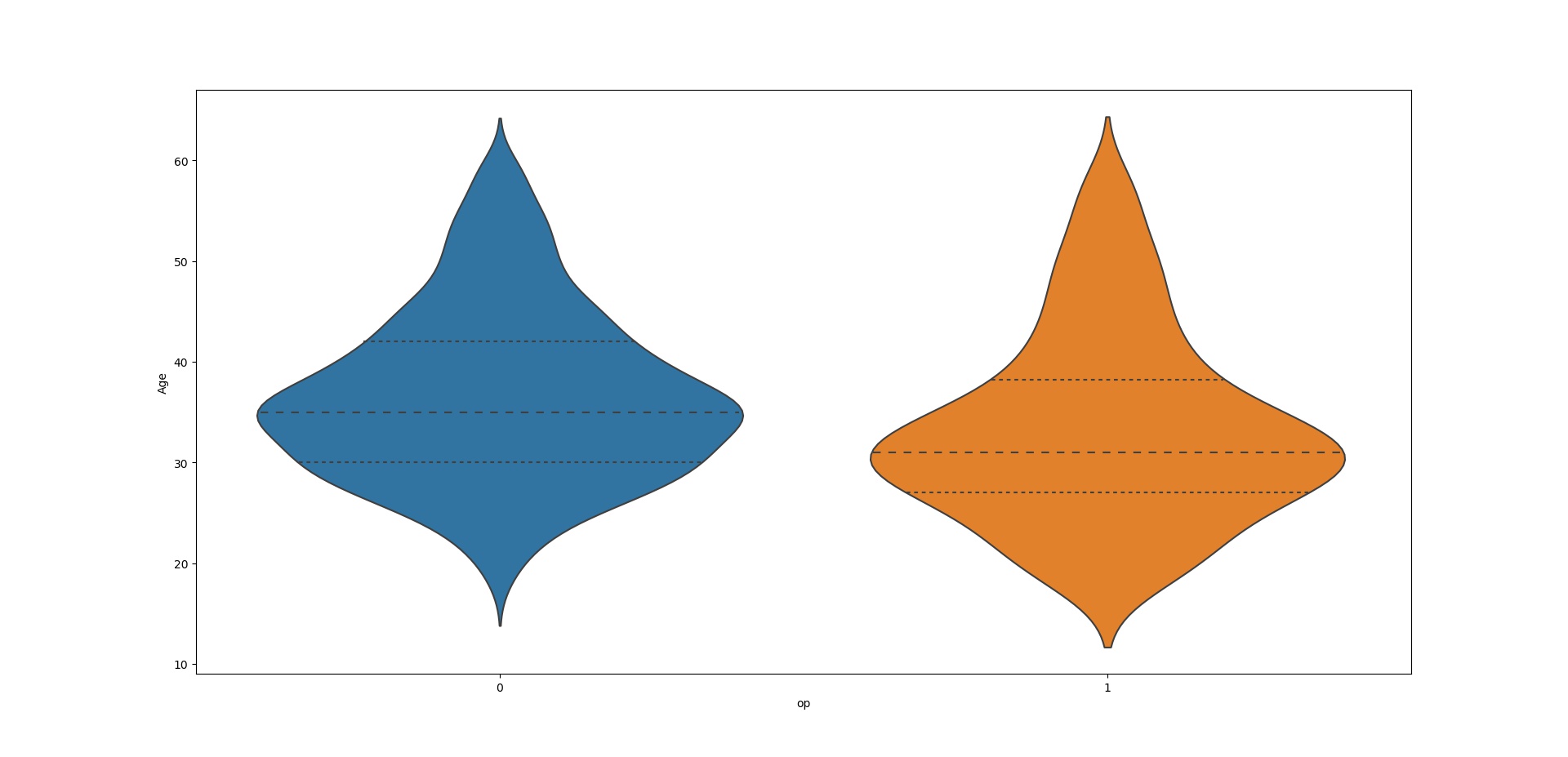
1. Years at Company
2. Years in Current Role
3. Years Since Last Promotion
4. Years with Current Manager

and converting one continuous column as categorical Years With Current Manager

**ViolinPlot**

=Attrition

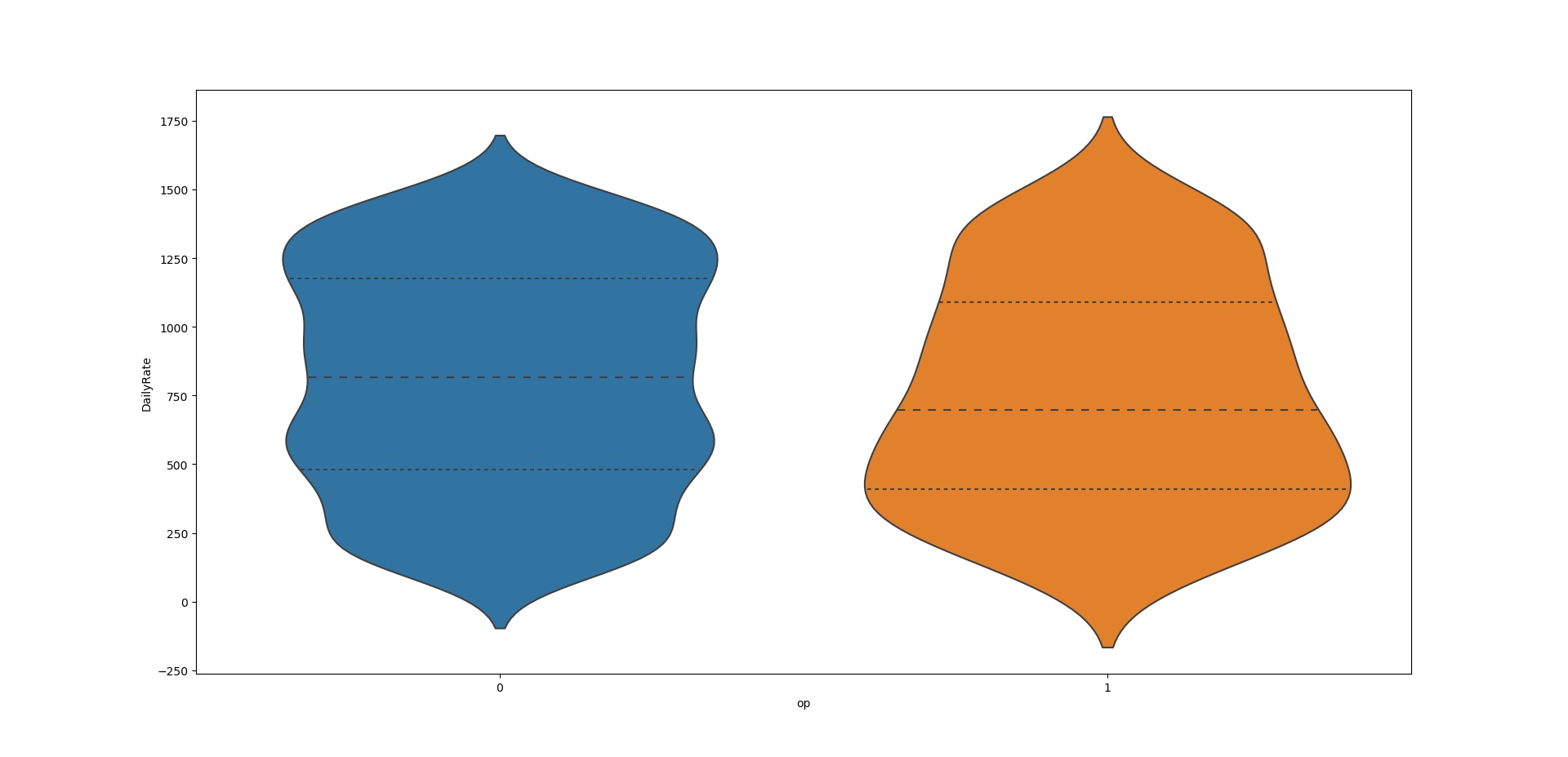
=Age



**ViolinPlot**

=Attrition

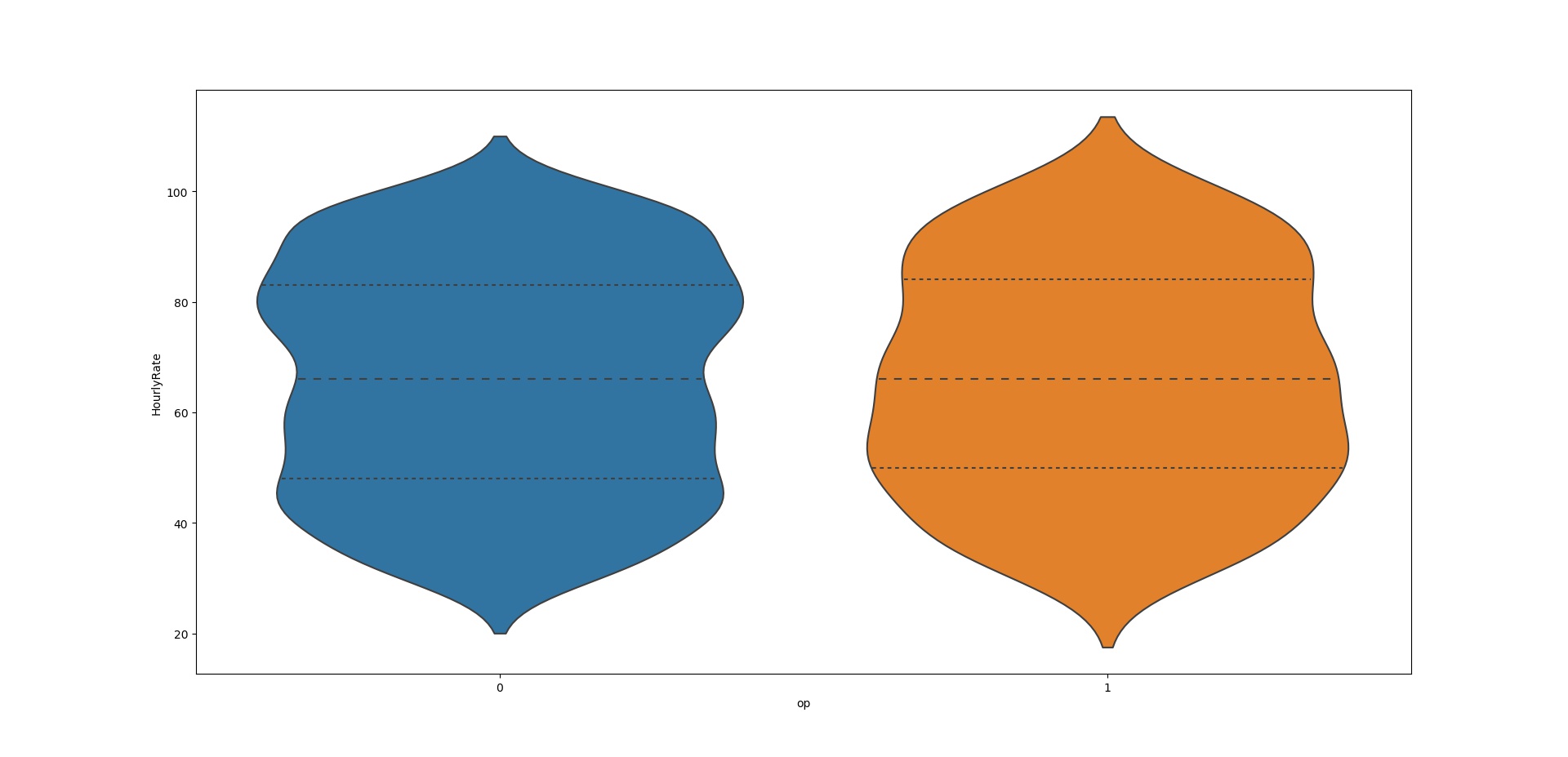
Y=Daily Rate



**ViolinPlot**

=Attrition

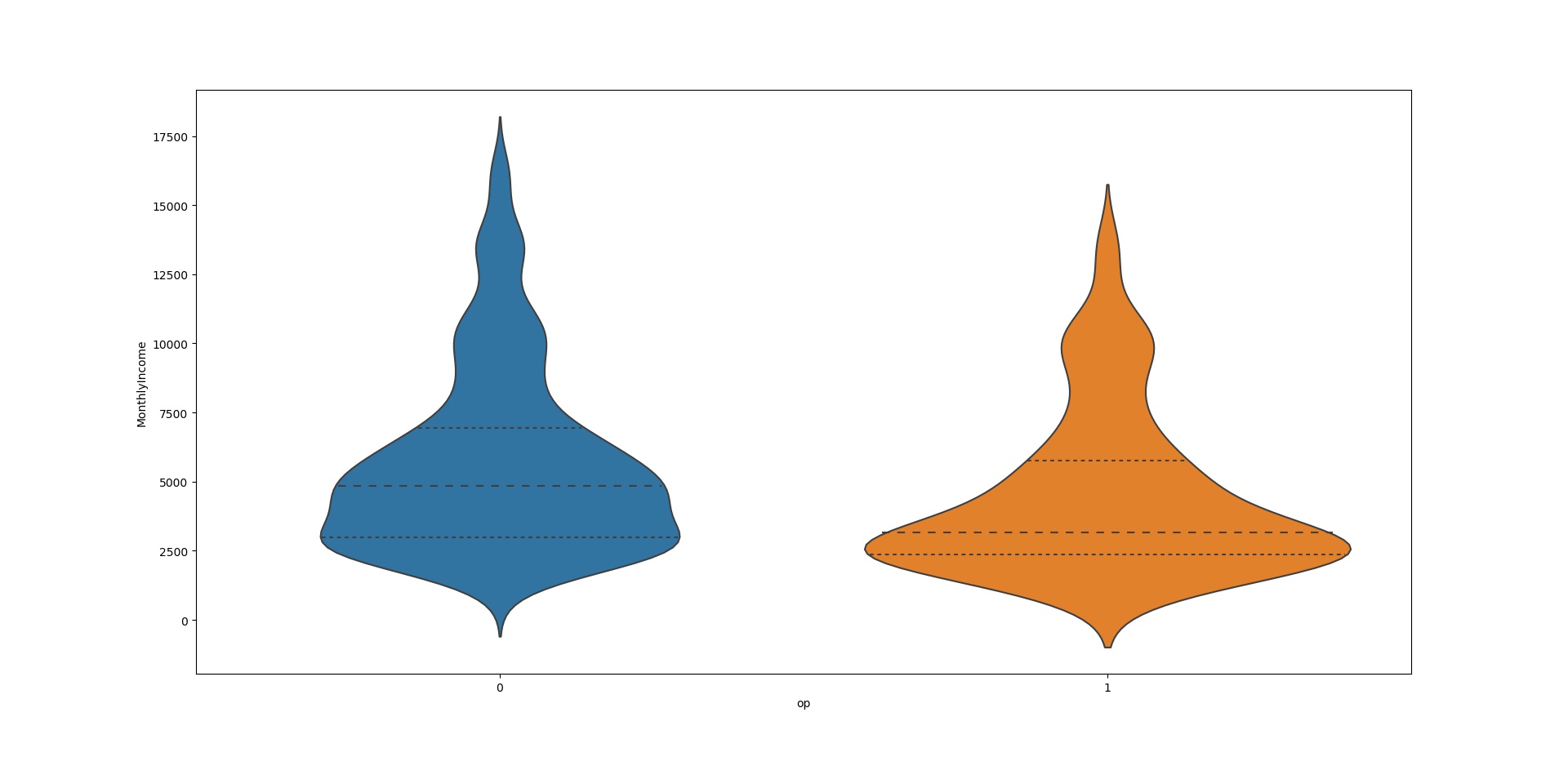
=Hourly Rate



**ViolinPlot**

=Attrition

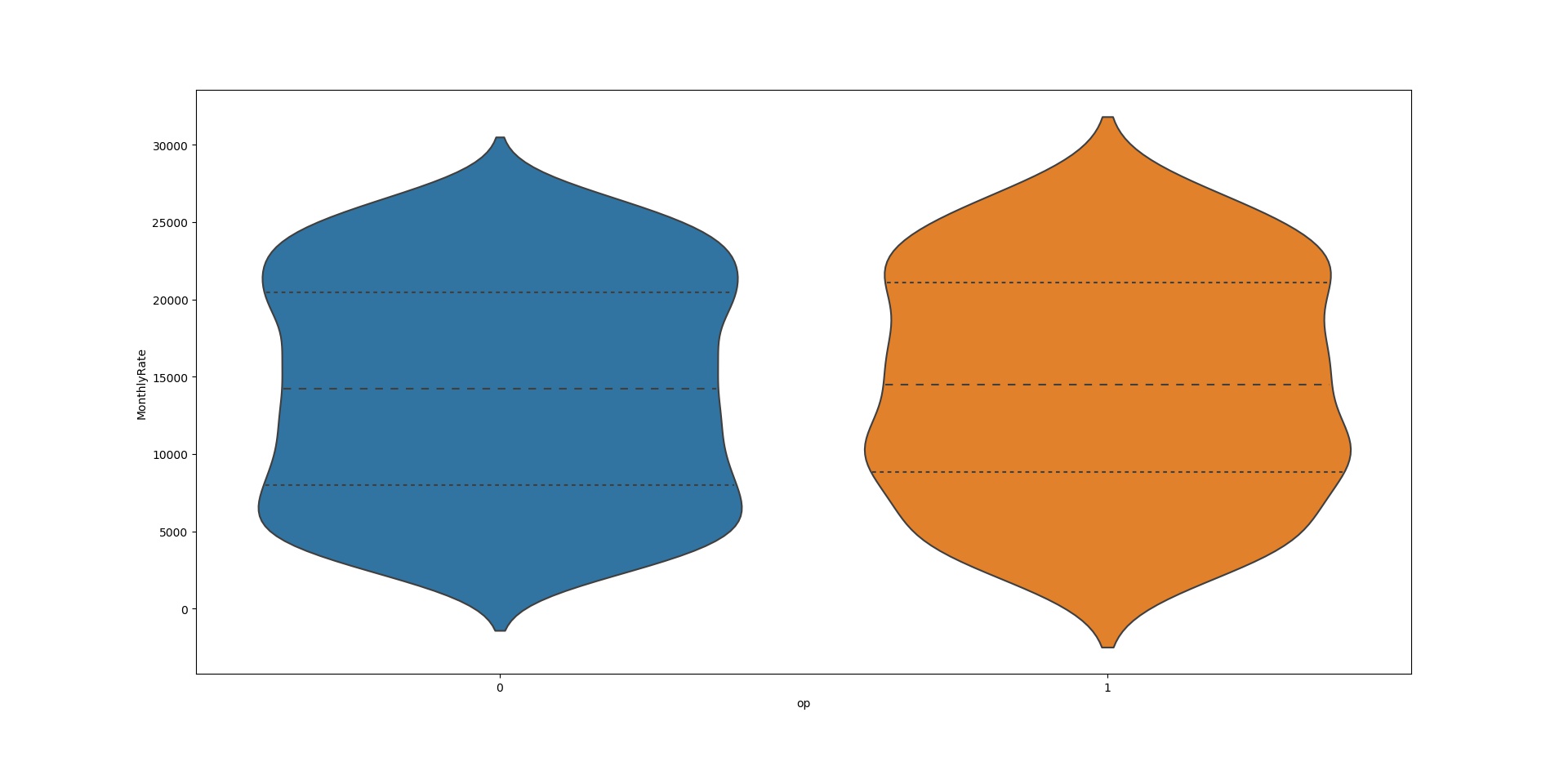
=Monthly Income



**ViolinPlot**

=Attrition

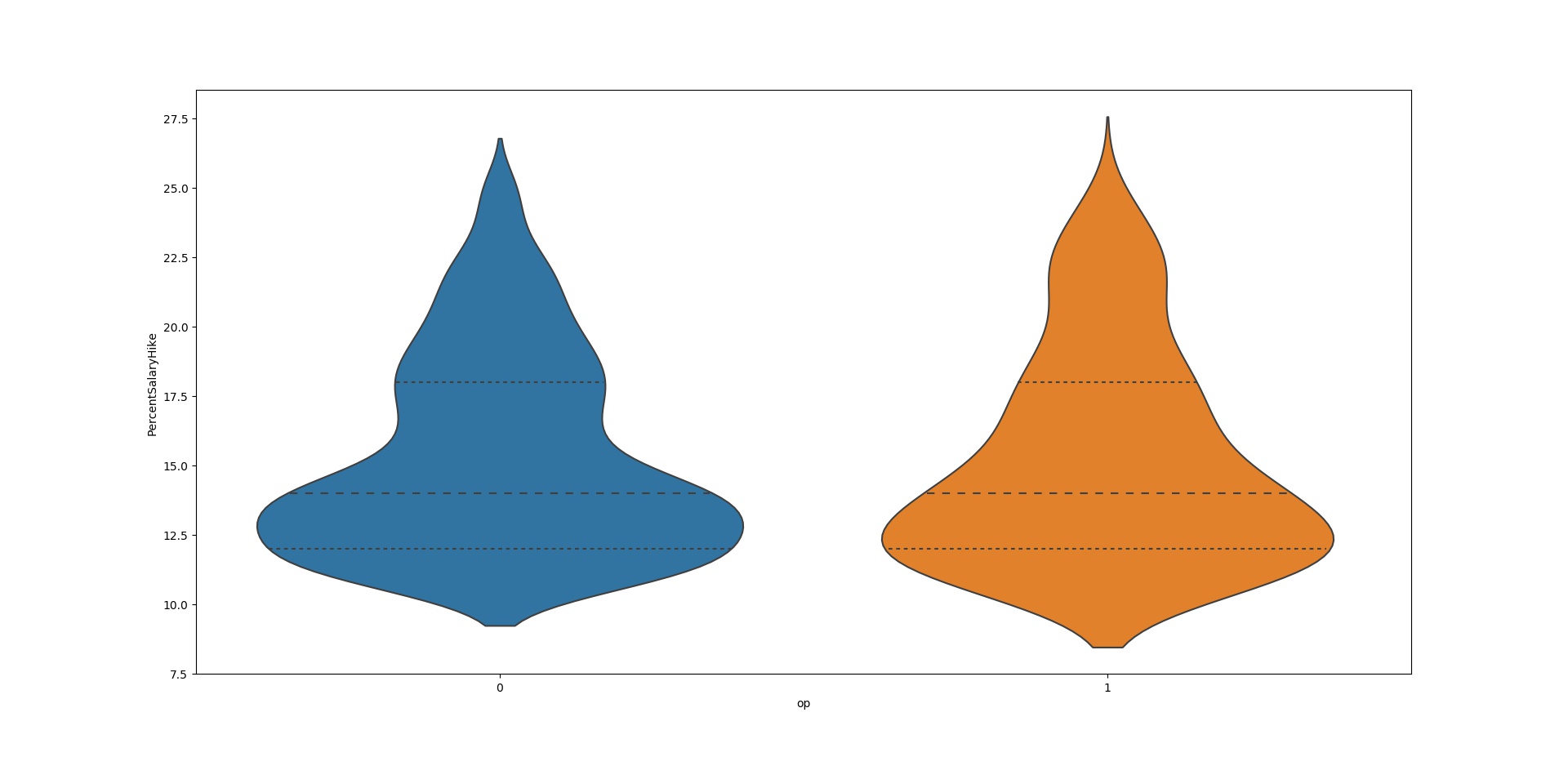
=Monthly Rate



**ViolinPlot**

=Attrition

=Percentage Salary Hike



**Code**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import model\_selection

from sklearn import metrics

from sklearn import preprocessing

from datetime import datetime

from sklearn import feature\_selection

from sklearn import naive\_bayes

from sklearn import tree

from sklearn import utils

from sklearn import ensemble

from sklearn import linear\_model

from sklearn import neighbors

import random

import warnings

from scipy import stats

warnings.simplefilter(action='ignore', category=FutureWarning)

df\_p=pd.read\_csv("EmployeeAttrition.csv")

df\_p.info()

#no null values

df\_p=utils.shuffle(df\_p,random\_state=42)

df\_p["op"]=(df\_p["Attrition"]=="Yes").astype(np.int)

df\_p.drop("Attrition",axis=1,inplace=True)

l\_drop=[]

for col in df\_p.columns.values:

if df\_p[str(col)].unique().shape[0]==1:

l\_drop.append(str(col))

###4 columns have only 1 values

#so dropping

l\_drop.append("EmployeeNumber")

df\_p.drop(l\_drop,axis=1,inplace=True)

l\_str=[]#list of column names having sering data

for col in df\_p.columns.values:

if type(df\_p[str(col)][0])==str or df\_p[str(col)].unique().shape[0]==1:

l\_str.append(str(col))

l\_cat=[]

for col in df\_p.drop(l\_str,axis=1).columns.values:

if df\_p[str(col)].unique().shape[0]<=5:

l\_cat.append(str(col))

l\_cat.remove("op")

l\_con=[]

for col in df\_p.drop(l\_str,axis=1).columns.values:

if df\_p[str(col)].unique().shape[0]>5:

l\_con.append(str(col))

def IQR(data):

upper\_quantile=data.quantile(0.75)

lower\_quantile=data.quantile(0.25)

IQR=upper\_quantile-lower\_quantile

outlier1=upper\_quantile+1.5\*IQR

outlier2=lower\_quantile-1.5\*IQR

return (IQR,outlier1,outlier2)

sns.distplot(df\_p.MonthlyIncome)

i,outlier1,outlier2=IQR(df\_p["MonthlyIncome"])

df\_p.drop(df\_p[df\_p["MonthlyIncome"]>outlier1].index,inplace=True)

sns.heatmap(df\_p[l\_con].corr(),annot=True)

df\_p.drop([ 'YearsAtCompany',

'YearsInCurrentRole',

'YearsSinceLastPromotion',

'YearsWithCurrManager'],axis=1,inplace=True)

for i in [ 'YearsAtCompany',

'YearsInCurrentRole',

'YearsSinceLastPromotion',

'YearsWithCurrManager']:

l\_con.remove(i)

sns.heatmap(df\_p[l\_con].corr(),annot=True)

df\_p.drop("TotalWorkingYears",axis=1,inplace=True)

l\_con.remove("TotalWorkingYears")

sns.distplot(df\_p['Age'])

sns.distplot(stats.boxcox(df\_p['DailyRate']))

sns.distplot(np.log(df\_p['DistanceFromHome']))

sns.distplot(np.log(df\_p["MonthlyIncome"]))

sns.distplot(np.log(df\_p['MonthlyRate']))

sns.distplot(df\_p['NumCompaniesWorked'])

sns.distplot(df\_p['PercentSalaryHike'])

sns.distplot(df\_p['TrainingTimesLastYear'])

cat=l\_cat+l\_str

df\_all\_cat=df\_p[cat]

#one hot encoding categorical dta

df\_p=pd.get\_dummies(df\_p,columns=cat)

df\_all\_cat.columns.values

df\_p.drop(l\_con,axis=1).columns.values

**Data Analysis**

|  |  |  |
| --- | --- | --- |
| **Index** | **Column Name** | **Number of unique values** |
| 1 | Age | 43 |
| 2 | Attrition | 2 |
| 3 | Business Travel | 3 |
| 4 | Daily Rate | 886 |
| 5 | Department | 3 |
| 6 | Distance From Home | 29 |
| 7 | Education | 5 |
| 8 | Education Field | 6 |
| 9 | Employee Count | 1 |
| 10 | Employee Number | 1470 |
| 11 | Environment Satisfaction | 4 |
| 12 | Gender | 2 |
| 13 | Hourly Rate | 71 |
| 14 | Job Involvement | 4 |
| 15 | Job Level | 5 |
| 16 | Job Role | 9 |
| 17 | Job Satisfaction | 4 |
| 18 | Marital Status | 3 |
| 19 | Monthly Income | 1349 |
| 20 | Monthly Rate | 1427 |
| 21 | NumCompaniesWorked | 10 |
| 22 | Over 18 | 1 |
| 23 | Overtime | 2 |
| 24 | Percent Salary Hike | 15 |
| 25 | Performance Rating | 2 |
| 26 | Relationship Satisfaction | 4 |
| 27 | Standard Hours | 1 |
| 28 | Stock Option Level | 4 |
| 29 | Total Working Years | 40 |
| 30 | Training Times Last Year | 7 |
| 31 | Work Life Balance | 4 |
| 32 | Years At Company | 37 |
| 33 | Years In Current Role | 19 |
| 34 | Years Since Last Promotion | 16 |
| 35 | Years With Current Manager | 18 |

There are 4 columns which have either only one unique values(Employee Count, Over 18, Standard Hours) and one column having 1470 different values means different values for all columns-Employee Number. These four columns are therefore dropped.

**Distinguishing categorical and continuous columns**

We considered the columns having more unique values than 10 as continuous or else categorical.

Initially, the continuous columns were

* Age
* Daily Rate
* Distance from Home
* Hourly Rate
* Monthly Income
* Monthly Rate
* Percent Salary Hike
* Total Working Years
* Years At Company
* Years In Current Role
* Years Since Last Promotion
* Years with Current Manager

and categorical columns were

1. String categorical

* Business Travel
* Department
* Education Field
* Gender
* Job Role
* Marital Status
* Overtime

1. Numeric categorical

* Education
* Environment Satisfaction
* Job Involvement
* Job Level
* Job Satisfaction
* NumCompaniesWorked
* Performance Rating
* Relationship Satisfaction
* Stock Option Level
* Training Times Last Year
* Work Life Balance

**Zero or negative value checking ::**

But further changes and tests are made later like checking for zero or negative values and we find that-

|  |  |  |
| --- | --- | --- |
| **Index** | **Column Name** | **Number of Zeroes** |
| 1 | Total Working Years | 11 |
| 2 | Years At Company | 44 |
| 3 | Years In Current Role | 244 |
| 4 | Years Since Last Promotion | 581 |
| 5 | Years With Current Manager | 263 |

After further investigating zeroes and checking the minimum, maximum and number of zeros in the particular columns which contains zeroes, we get

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Column Name** | **No. of zeroes** | **Maximum** | **Minimum** |
| 1 | Total Working Years | 40 | 40 | 0 |
| 2 | Years At Company | 37 | 40 | 0 |
| 3 | Years In Current Role | 19 | 18 | 0 |
| 4 | Years Since Last Promotion | 16 | 15 | 0 |
| 5 | Years With Current Manager | 18 | 17 | 0 |

As none of these are odd values, so we do not drop these values

**Outliers checking::**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Index** | **Column Name** | **Upper Outliers** | **Lower Outliers** | **Theoretical Maximum** | **Maximum** | **Theoretical Minimum** | **Minimum** |
| 1 | Age | 0 | 0 | 62.5 | 60 | 10.5 | 18 |
| 2 | Daily Rate | 0 | 0 | 2195 | 1499 | -573 | 102 |
| 3 | Distance from Home | 0 | 0 | 32 | 29 | -16 | 1 |
| 4 | Hourly Rate | 0 | 0 | 137.375 | 100 | -5.625 | 30 |
| 5 | Monthly Income | 114 | 0 | 16581 | 19999 | -5291 | 1009 |
| 6 | Monthly Rate | 0 | 0 | 39083.25 | 26999 | -10574.75 | 2094 |
| 7 | Percent Salary Hike | 0 | 0 | 27 | 25 | 3 | 11 |
| 8 | Total Working Years | 63 | 0 | 28.5 | 40 | -7.5 | 0 |
| 9 | Years At Company | 104 | 0 | 18 | 40 | -6 | 0 |
| 10 | Years In Current Role | 21 | 0 | 14.5 | 18 | -5.5 | 0 |
| 11 | Years Since Last Promotion | 107 | 0 | 7.5 | 15 | -4.5 | 0 |
| 12 | Years With Current Manager | 14 | 0 | 14.5 | 17 | -5.5 | 0 |

Except Monthly Income columns in other columns which contains upper outliers, those values may not be considered as outliers because they are not anomalies and are totally acceptable values.

So only outliers of Monthly Income values are deleted.

**checking corelation between continous data**::

1.age and total working years are corelated::0.63

2.monthly inccome and totalworkingyears are corelated::0.67

3.years at company and totalworking years are corelated :: 0.63

4.years at company years ,in current role ,years since last promotion ,years with current manager all 4 are corelated >0.6 coefficient

that's why we r deleting 4 continous columns as

1.they are highly corelated

2.they have too many unique values so we cannot consider them as categorical columns

they are

1.'YearsAtCompany',

2.'YearsInCurrentRole',

3.'YearsSinceLastPromotion',

4.'YearsWithCurrManager'

and converting one continous column as categorical

1.YearsWithCurrManager

investigating over different categories of cateorical columns::

**the percentage of yes and no on total dataset is**

**yes ::0.1710914454277286 ,no:: 0.8289085545722714**

**checking yes and no for differnent categories of all categorical columns to find anomaly::**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Category name** | **% of yes** | **% of no** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Education | 1 | 0.1875 | 0.8125 |
| Education | 2 | 0.1699 | 0.8301 |
| Education | 3 | 0.1825 | 0.8174 |
| Education | 4 | 0.1549 | 0.8451 |
| Education | 5 | 0.1162 | 0.8837 |
| EnvironmentSatisfaction | 1 | 0.2671 | 0.7328 |
| EnvironmentSatisfaction | 2 | 0.1567 | 0.8432 |
| EnvironmentSatisfaction | 3 | 0.1476 | 0.8523 |
| EnvironmentSatisfaction | 4 | 0.1428 | 0.8571 |
| JobImvolvement | 1 | 0.3376 | 0.6623 |
| JobImvolvement | 2 | 0.2034 | 0.7965 |
| JobImvolvement | 3 | 0.1536 | 0.8463 |
| JobImvolvement | 4 | 0.0958 | 0.9044 |
| JobLevel | 1 | 0.2633 | 0.7366 |
| JobLevel | 2 | 0.0973 | 0.9026 |
| JobLevel | 3 | 0.1467 | 0.8532 |
| JobLevel | 4 | 0.0819 | 0.9180 |
| JobSatisfaction | 1 | 0.2462 | 0.7537 |
| JobSatisfaction | 2 | 0.1705 | 0.8294 |
| JobSatisfaction | 3 | 0.1769 | 0.8230 |
| JobSatisfaction | 4 | 0.1194 | 0.8805 |
| NumCompaniesWorked | 0 | 0.1229 | 0.8770 |
| NumCompaniesWorked | 1 | 0.1935 | 0.8064 |
| NumCompaniesWorked | 2 | 0.1119 | 0.8880 |
| NumCompaniesWorked | 3 | 0.1159 | 0.8840 |
| NumCompaniesWorked | 4 | 0.1370 | 0.8629 |
| NumCompaniesWorked | 5 | 0.2631 | 0.7368 |
| NumCompaniesWorked | 6 | 0.2539 | 0.7460 |
| NumCompaniesWorked | 7 | 0.2424 | 0.7575 |
| NumCompaniesWorked | 8 | 0.1363 | 0.8636 |
| NumCompaniesWorked | 9 | 0.2553 | 0.7446 |
| PerformanceRating | 3 | 0.1701 | 0.8298 |
| PerformanceRating | 4 | 0.1761 | 0.8238 |
| RelationshipSatisfaction | 1 | 0.2162 | 0.7837 |
| RelationshipSatisfaction | 2 | 0.1618 | 0.8381 |
| RelationshipSatisfaction | 3 | 0.1662 | 0.8337 |
| RelationshipSatisfaction | 4 | 0.1530 | 0.8469 |
| StockOptionLevel | 0 | 0.2572 | 0.7427 |
| StockOptionLevel | 1 | 0.1005 | 0.8994 |
| StockOptionLevel | 2 | 0.08 | 0.92 |
| StockOptionLevel | 3 | 0.1829 | 0.8170 |
| TrainingTimesLastYear | 0 | 0.2745 | 0.7254 |
| TrainingTimesLastYear | 1 | 0.1323 | 0.8676 |
| TrainingTimesLastYear | 2 | 0.1896 | 0.8103 |
| TrainingTimesLastYear | 3 | 0.1501 | 0.8498 |
| TrainingTimesLastYear | 4 | 0.2300 | 0.7699 |
| TrainingTimesLastYear | 5 | 0.1308 | 0.8691 |
| TrainingTimesLastYear | 6 | 0.0952 | 0.9047 |
| WorkLifeBalance | 1 | 0.3246 | 0.6753 |
| WorkLifeBalance | 2 | 0.1841 | 0.8158 |
| WorkLifeBalance | 3 | 0.1480 | 0.8519 |
| WorkLifeBalance | 4 | 0.1928 | 0.8071 |
| YearsWithCurrManager | 0 | 0.3386 | 0.6613 |
| YearsWithCurrManager | 1 | 01617 | 0.2382 |
| YearsWithCurrManager | 2 | 0.1501 | 0.8498 |
| YearsWithCurrManager | 3 | 0.1407 | 0.8592 |
| YearsWithCurrManager | 4 | 0.1063 | 0.8936 |
| YearsWithCurrManager | 5 | 0.1333 | 0.8666 |
| YearsWithCurrManager | 6 | 0.1739 | 0.8260 |
| YearsWithCurrManager | 7 | 0.155 | 0.845 |
| YearsWithCurrManager | 8 | 0.08247 | 0.9175 |
| YearsWithCurrManager | 9 | 0.098 | 0.9019 |
| YearsWithCurrManager | 10 | 0.1578 | 0.8421 |
| YearsWithCurrManager | 11 | 0.0 | 1.0 |
| YearsWithCurrManager | 12 | 0.0 | 1.0 |
| YearsWithCurrManager | 13 | 0.0 | 1.0 |
| YearsWithCurrManager | 14 | 0.6666 | 0.333 |
| YearsWithCurrManager | 15 | 0.0 | 1.0 |
| YearsWithCurrManager | 17 | 0.0 | 1.0 |
| BusinessTravel | Non-Travel | 0.0863 | 0.9136 |
| BusinessTravel | Travel\_Frequently | 0.2633 | 0.7366 |
| BusinessTravel | Travel\_Rarely | 0.15811 | 0.8418 |
| Department | Human\_Resource | 0.2222 | 0.7777 |
| Department | R and D | 0.1475 | 0.8524 |
| Department | Sales | 0.2137 | 0.7862 |
| EducationField | Human\_Resources | 0.3333 | 0.6666 |
| EducationField | Life Sciences | 0.1571 | 0.8428 |
| EducationField | Marketing | 0.2229 | 0.777 |
| EducationField | Medical | 0.1462 | 0.8537 |
| EducationField | Other | 0.1447 | 0.8552 |
| EducationField | Technical | 0.2440 | 0.7559 |
| Gender | Male | 0.1541 | 0.8458 |
| Gender | Female | 0.1824 | 0.8175 |
| JobRole | Healthcare Representative | 0.0687 | 0.9313 |
| JobRole | Human Resource | 0.2307 | 0.7692 |
| JobRole | Laboratory Technician | 0.2393 | 0.7606 |
| JobRole | Manager | 0.0714 | 0.9285 |
| JobRole | Manufacturing  Director | 0.068 | 0.9285 |
| JobRole | Research Director | 0.0 | 1.0 |
| JobRole | Sales  Representative | 0.3975 | 0.6024 |
| JobRole | Sales  Executive | 0.1748 | 0.8251 |
| JobRole | Research Scientiest | 0.1609 | 0.8390 |
| MaritalStatus | Divorced | 01609 | 0.839 |
| MaritalStatus | Married | 0.1748 | 0.8251 |
| MaritalStatus | Single | 0.3975 | 0.6024 |
| OverTime | No | 0.1113 | 0.8886 |
| OverTime | Yes | 0.3212 | 0.6787 |

**Anomaly from above observations::**

1.pepole who has overtime are more prawn to leave the company,the ratio becomes 1:2 in this case

2.people having job role as research director doesnot leave company

3.people who are working for current manager for 11 12 13 15 16 17 years that means for a long time generally don’t leave the company except current manager year 14 where this ratio is drastically opposite 2:1

4.people who has worklife balance status as 1 generally more prawn to leave the company

The ratio is 1:2 for no attrition vs attrition

5.people havinf human resources as education field are more prawn to leave the company

The ratio is again 1:2

6.people in manager or manufacturing director post has very less tendency to leave,the ratio is almost 1:10

**Categorical data are one hot encoded,after one hot encoding number of total columns become 105**

**Feature scalling is applied on the**

**applying different Tree Classifier and getting feature importance:::**

among 3 applied tree classifiers

1.decisiontree

2.randomforest

3.extratreeclassifier

**Applying differnet model classification algorithm to get result**

1.logistic regression

2.decision tree classifier

3.randomforrest classifier

3.k nearest neighbours

4.naive byase

**test 1:**

we applied decision tree classifier for feature importance and passed the highest important 10 columns for these 4 algorithms

output:

features passed::: 'DistanceFromHome', 'Age', 'OverTime\_No', 'JobLevel\_1',

'MonthlyRate', 'MonthlyIncome', 'YearsWithCurrManager\_0',

'StockOptionLevel\_0', 'DailyRate', 'PercentSalaryHike'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.857377 0.709677 0.22000 0.601176

1 LogisticReg-test 0.794118 0.750000 0.18750 0.584135

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.750000 0.450000 0.28125 0.587740

4 KNN(5)-train 0.877869 0.774194 0.36000 0.669706

5 KNN(5)-test 0.801471 0.727273 0.25000 0.610577

6 NB-train 0.841803 0.522293 0.41000 0.668235

7 NB-test 0.816176 0.705882 0.37500 0.663462

**Test 2::**

We applied random forrest to measure feature importance and again passed 10 highest important columns

Output::

Columns ::: 'MonthlyIncome', 'Age', 'MonthlyRate', 'OverTime\_Yes',

'OverTime\_No', 'DistanceFromHome', 'DailyRate', 'HourlyRate',

'JobLevel\_1', 'BusinessTravel\_Travel\_Frequently'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.850000 0.673469 0.16500 0.574657

1 LogisticReg-test 0.779412 0.625000 0.15625 0.563702

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.801471 0.666667 0.31250 0.632212

4 KNN(9)-train 0.850820 0.625000 0.22500 0.599265

5 KNN(9)-test 0.757353 0.454545 0.15625 0.549279

6 NB-train 0.832787 0.489011 0.44500 0.676912

7 NB-test 0.764706 0.500000 0.34375 0.618990

**Test 3:::**

Output::

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.854918 0.716981 0.19000 0.587647

1 LogisticReg-test 0.779412 0.666667 0.12500 0.552885

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.757353 0.473684 0.28125 0.592548

4 KNN(5)-train 0.872131 0.761905 0.32000 0.650196

5 KNN(5)-test 0.750000 0.416667 0.15625 0.544471

6 NB-train 0.830328 0.482759 0.49000 0.693529

7 NB-test 0.830882 0.680000 0.53125 0.727163

Accuracy and AUC value is high but recall is not much

**Inference from these 3 tests::**

Number of columns are all total 105 ,so only passing 10 columns will not help incrementing recall,AUC and Accuracy is already pretty high,we need to focus on recall

**Test 4::**

Applying all 4 algorithms on highest 30 columns using feature importance as decision tree classifier

Output::

Name of 30 columns those are used:

'DistanceFromHome', 'Age', 'OverTime\_No', 'JobLevel\_1',

'MonthlyRate', 'MonthlyIncome', 'YearsWithCurrManager\_0',

'StockOptionLevel\_0', 'DailyRate', 'PercentSalaryHike',

'HourlyRate', 'WorkLifeBalance\_2', 'WorkLifeBalance\_1',

'JobSatisfaction\_4', 'JobRole\_Sales Executive',

'MaritalStatus\_Single', 'EducationField\_Technical Degree',

'NumCompaniesWorked\_7', 'RelationshipSatisfaction\_1',

'TrainingTimesLastYear\_4', 'WorkLifeBalance\_4', 'JobInvolvement\_1',

'RelationshipSatisfaction\_3', 'RelationshipSatisfaction\_4',

'NumCompaniesWorked\_3', 'NumCompaniesWorked\_0',

'TrainingTimesLastYear\_1', 'JobRole\_Sales Representative',

'JobRole\_Laboratory Technician', 'JobRole\_Human Resources'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.854918 0.716981 0.19000 0.587647

1 LogisticReg-test 0.779412 0.666667 0.12500 0.552885

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.757353 0.473684 0.28125 0.592548

4 KNN(5)-train 0.872131 0.761905 0.32000 0.650196

5 KNN(5)-test 0.750000 0.416667 0.15625 0.544471

6 NB-train 0.830328 0.482759 0.49000 0.693529

7 NB-test 0.830882 0.680000 0.53125 0.727163

**Test 5::**

Applying all 4 algorithms on highest 30 columns using feature importance as random forrest

Classifier

Output::

Columns names::

'MonthlyIncome', 'Age', 'MonthlyRate', 'OverTime\_Yes',

'OverTime\_No', 'DistanceFromHome', 'DailyRate', 'HourlyRate',

'JobLevel\_1', 'BusinessTravel\_Travel\_Frequently',

'PercentSalaryHike', 'JobRole\_Sales Representative',

'MaritalStatus\_Single', 'StockOptionLevel\_1',

'YearsWithCurrManager\_0', 'EducationField\_Medical',

'EnvironmentSatisfaction\_3', 'JobSatisfaction\_1',

'NumCompaniesWorked\_1', 'WorkLifeBalance\_1', 'WorkLifeBalance\_2',

'Education\_4', 'TrainingTimesLastYear\_3',

'EnvironmentSatisfaction\_1', 'WorkLifeBalance\_3',

'TrainingTimesLastYear\_2', 'JobLevel\_2',

'RelationshipSatisfaction\_1', 'Gender\_Male',

'RelationshipSatisfaction\_2'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.873770 0.730000 0.36500 0.669265

1 LogisticReg-test 0.823529 0.900000 0.28125 0.635817

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.750000 0.454545 0.31250 0.598558

4 KNN(7)-train 0.866393 0.813559 0.24000 0.614608

5 KNN(7)-test 0.779412 1.000000 0.06250 0.531250

6 NB-train 0.822951 0.466387 0.55500 0.715245

7 NB-test 0.830882 0.666667 0.56250 0.737981

**Test 6::**

Using extra tree classifier for feature importance and passing highest 30 columns

Output::

Columns::

‘OverTime\_No', 'OverTime\_Yes', 'JobLevel\_1', 'Age',

'YearsWithCurrManager\_0', 'MaritalStatus\_Single',

'DistanceFromHome', 'MonthlyIncome', 'HourlyRate',

'BusinessTravel\_Travel\_Frequently', 'DailyRate',

'EnvironmentSatisfaction\_3', 'MonthlyRate',

'EducationField\_Medical', 'JobSatisfaction\_1',

'EnvironmentSatisfaction\_1', 'WorkLifeBalance\_2',

'RelationshipSatisfaction\_1', 'Gender\_Female',

'RelationshipSatisfaction\_4', 'JobSatisfaction\_2',

'Department\_Research & Development', 'PercentSalaryHike',

'EnvironmentSatisfaction\_4', 'EducationField\_Marketing',

'JobSatisfaction\_3', 'JobInvolvement\_3', 'NumCompaniesWorked\_1',

'Education\_3', 'JobRole\_Research Scientist'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.889344 0.835052 0.40500 0.694657

1 LogisticReg-test 0.816176 0.888889 0.25000 0.620192

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.735294 0.423077 0.34375 0.599760

4 KNN(7)-train 0.867213 0.895833 0.21500 0.605049

5 KNN(7)-test 0.779412 0.750000 0.09375 0.542067

6 NB-train 0.850000 0.542289 0.55500 0.727402

7 NB-test 0.823529 0.722222 0.56625 0.679087

**Inference from test 4,5,6::**

1.extra tree classifier gives better feature importance than random forrest and decision tree classifier

2.recall is slightly underfitted,so we need to increase number of features aa bit,but accuracy and AUC score is not bad

**Test 7::**

Using extratree classifier to get feature importances and passing total 35 columns to get the result

Output::

columns are

'OverTime\_Yes', 'OverTime\_No', 'MonthlyIncome',

'YearsWithCurrManager\_0', 'DistanceFromHome',

'MaritalStatus\_Single', 'HourlyRate', 'JobLevel\_1',

'JobSatisfaction\_1', 'MonthlyRate', 'PercentSalaryHike',

'JobLevel\_2', 'BusinessTravel\_Travel\_Frequently', 'Age',

'StockOptionLevel\_1', 'RelationshipSatisfaction\_1',

'StockOptionLevel\_0', 'TrainingTimesLastYear\_3',

'RelationshipSatisfaction\_3', 'JobRole\_Laboratory Technician',

'EnvironmentSatisfaction\_3', 'EnvironmentSatisfaction\_1',

'JobSatisfaction\_4', 'DailyRate', 'WorkLifeBalance\_3',

'Gender\_Male', 'WorkLifeBalance\_1', 'EducationField\_Medical',

'BusinessTravel\_Travel\_Rarely', 'WorkLifeBalance\_2',

'EnvironmentSatisfaction\_4', 'EducationField\_Life Sciences',

'JobSatisfaction\_3', 'NumCompaniesWorked\_1',

'JobRole\_Sales Representative'

MODELNAME ACCURACY PRECISION RECALL AUC

0 LogisticReg-train 0.883607 0.758929 0.42500 0.699265

1 LogisticReg-test 0.852941 1.000000 0.37500 0.687500

2 DecisionTree-train 1.000000 1.000000 1.00000 1.000000

3 DecisionTree-test 0.750000 0.450000 0.28125 0.587740

4 KNN(5)-train 0.886885 0.887500 0.35500 0.673088

5 KNN(5)-test 0.772059 0.571429 0.12500 0.548077

6 NB-train 0.830328 0.486056 0.61000 0.741765

7 NB-test 0.860294 0.760000 0.59375 0.768029

**Inference::**

Naïve byase showed the best results

**CONCLUSION**

**1.used naïve baise classification algorithm**

**2.using extra tree classifier decided highest importatnt features**

**3.total 35 columns out of 105 features(that is present after one hot encode of the categorical columns) are used**

**4.name of the columns::**

**1.'OverTime\_Yes'**

**2. 'OverTime\_No'**

**3. 'MonthlyIncome'**

**4. 'YearsWithCurrManager\_0'**

**5. 'DistanceFromHome'**

**6. 'MaritalStatus\_Single'**

**7. 'HourlyRate'**

**8.'JobLevel\_1',**

**9.'JobSatisfaction\_1'**

**10. 'MonthlyRate'**

**11. 'PercentSalaryHike',**

**12.'JobLevel\_2'**

**13. 'BusinessTravel\_Travel\_Frequently'**

**14.'Age'**

**15.'StockOptionLevel\_1'**

**16.'RelationshipSatisfaction\_1',**

**17.'StockOptionLevel\_0'**

**18.'TrainingTimesLastYear\_3'**

**19.'RelationshipSatisfaction\_3'**

**20.'JobRole\_Laboratory Technician'**

**21. 'EnvironmentSatisfaction\_3'**

**22.'EnvironmentSatisfaction\_1',**

**23.'JobSatisfaction\_4'**

**24.'DailyRate'**

**25.'WorkLifeBalance\_3'**

**26 .'Gender\_Male'**

**27.'WorkLifeBalance\_1'**

**28.'EducationField\_Medical'**

**29.'BusinessTravel\_Travel\_Rarely'**

**30.'WorkLifeBalance\_2',**

**31.'EnvironmentSatisfaction\_4'**

**32.'EducationField\_Life Sciences',**

**33.'JobSatisfaction\_3'**

**34.'NumCompaniesWorked\_1',**

**35.'JobRole\_Sales Representative'**

**Results::**

**1.accuracy::86%(over test data)**

**2.recall score::59%(over test data)**

**3.AUC value::0.76(over test data)**

**ROC\_AUC curve::**

