

HR Analytics Project – Understanding the Attrition In HR (Evaluation Project 2)

**Introduction**

# **Attrition in human resources refers to the gradual loss of employees' overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees. It can be voluntary as well as involuntary.**

**Introduction Of Problem Statement**

**The Human Resource Analytics Project is a classification-based problem statement to predict probability of attrition from several indicators, it also strives to study and generate as many insights as possible. Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytical processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. It also aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes. This project is concerned with answering the question, how HR analytics can help in analyzing attrition.**

**Problem Definition**

**How does HR Analytics help in analyzing attrition?**

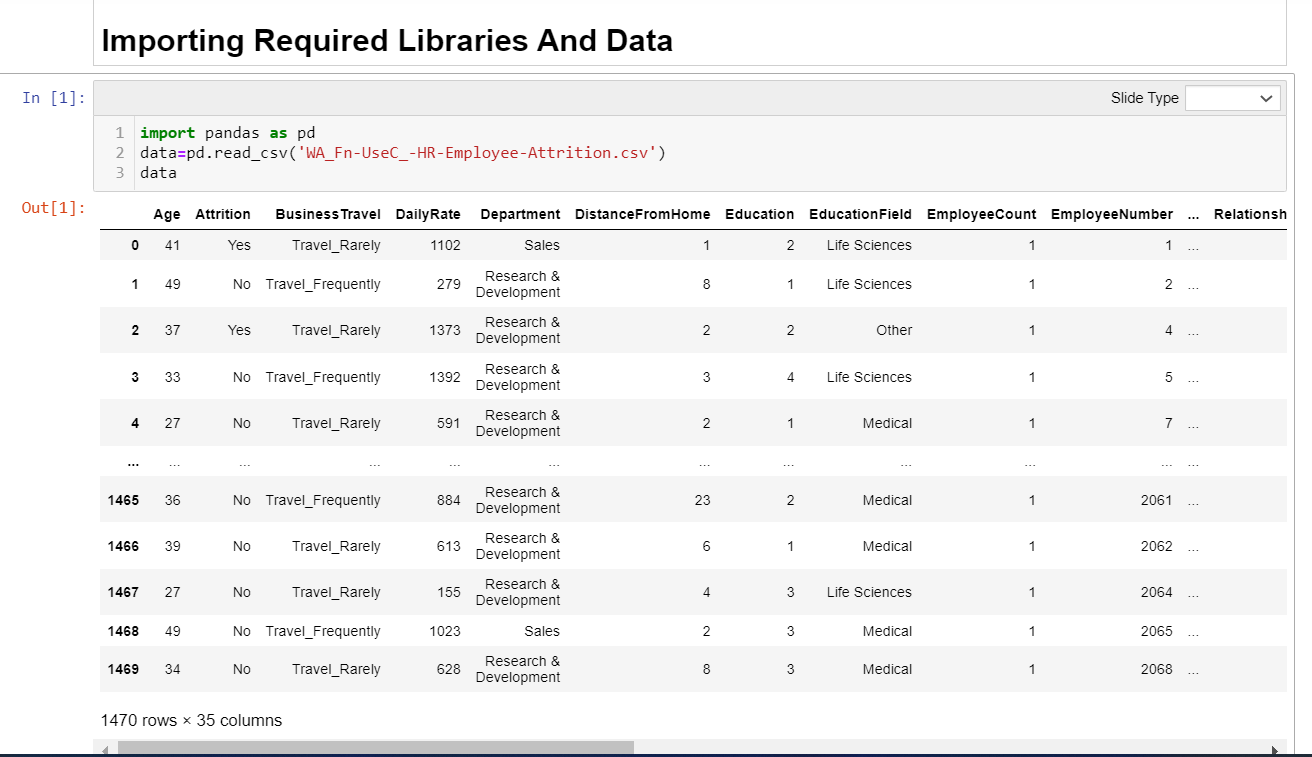
Discussion On Problem Definition

# **Analytics has taken over the human resource field. It is as famous in Human Resource as it is in other areas. HR Analytics prevent the loss of employees and hence increases the return on investment. Through HR Analytics, professionals can study how different features are involved in increasing as well as decreasing attrition rate. It can also build predictive models that can predict, based on an employee's present condition, how vulnerable he is to attrition. Based on these predictions, human resources can take required steps to prevent attrition.**

**The Insights that can be drawn with the help of data it collects, are detailed in EDA sections (it is divided in two parts for thorough study of each feature). Followed by which, I have developed ensemble, bagging and boosting based classification predictive models to provide a complete HR Analytics problem solution.**

Understanding the Data

*The data is imported from GitHub repository dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics*



*Verbal Translation:*

1. Number Of Rows: 1470
2. Number of Columns: 35
3. Column Names:

Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',  
 'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',  
 'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',  
 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',  
 'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',  
 'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',  
 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',  
 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',  
 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',  
 'YearsWithCurrManager'],  
 dtype='object')

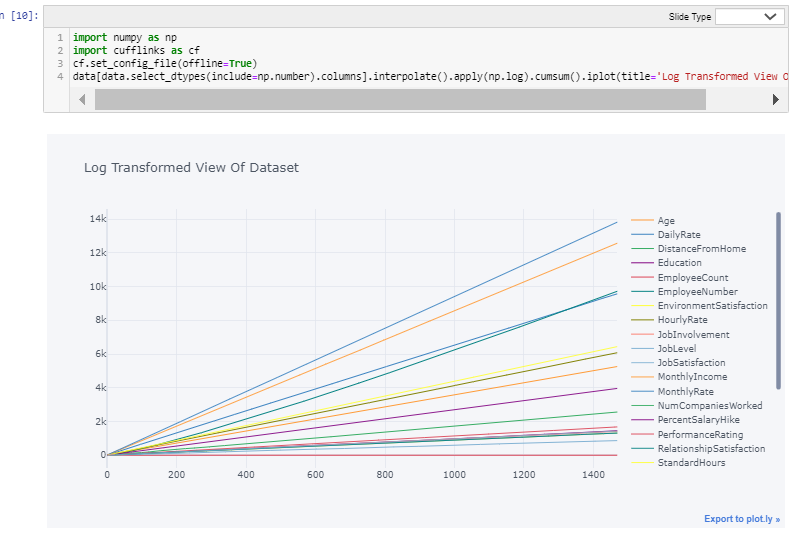
The data basically covers details of employees as features, for example, age, Department, PercentSalaryHike, etcetera and label Attrition is boolean column stating if the employee faced attrition. The idea behind studying the details of employees and training the model on historic data is to prepare a low bias low variance model pipeline that would predict with good accuracy the probability of attrition of current and prospective employees.

Thus, it would help in making good decisions about employee retention, such as, appropriate compensation plan, correct job level, promotions, etcetera.

1. The dataset has no missing values; therefore, imputation and filing data are not required.
2. The whole dataset is of 1470\*35 dimension, hence, chances of overfitting, that is, model getting trained more than essential are high so we will go through the data thoroughly and try to draw anomalies and code some derived features, that can be of help in making predictions.
3. There are 26 continuous columns and 9 categorical columns, hence, encoding is required.

Exploratory Data Analysis

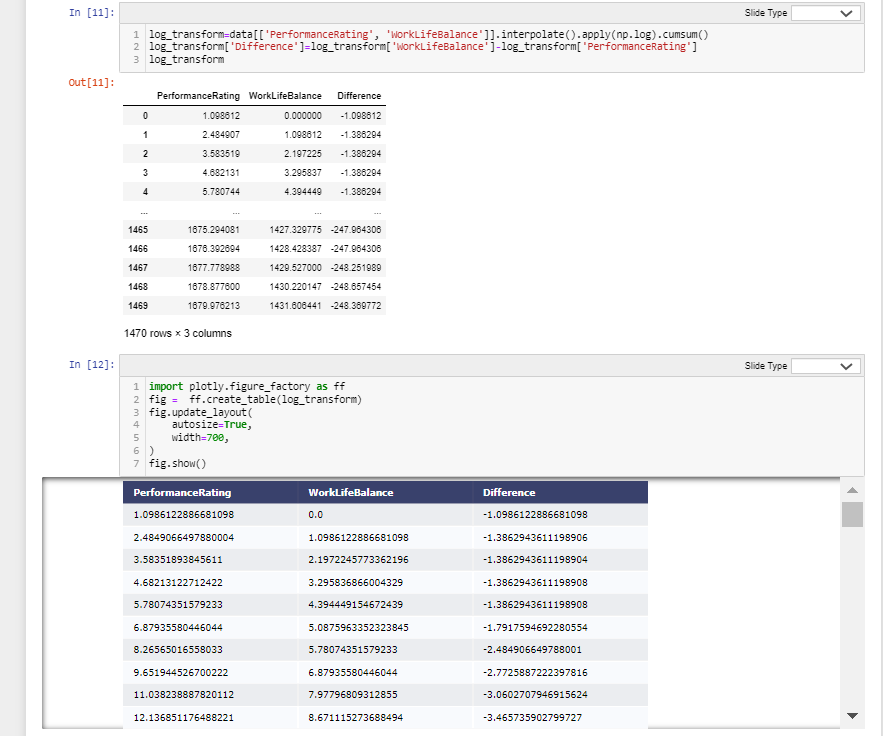
Analysis 1: Log Transformed Analysis



# **Observations:**

1. I have transformed the data into lognormal for better visual understanding of data.
2. Max of Cumulative Sum for Performance Rating and Work Life Balance are overlapping, there is just -1 difference between the two values initially which sums up to only -248.40 over 1470 data points.
3. The second most prominent value for good performance is Education with Max of Cumulative Sum 1452.87 against 1679.98 performance rating.
4. The most divergent to Performance Rating are Monthly Rate and Employee Count

Analysis 2: Log Transformed Analysis of Performance Rating and Work Life Balance



# **The above table shows each data point comparison of Work Life Balance and Performance Rating.**

**Pattern Discovered:**

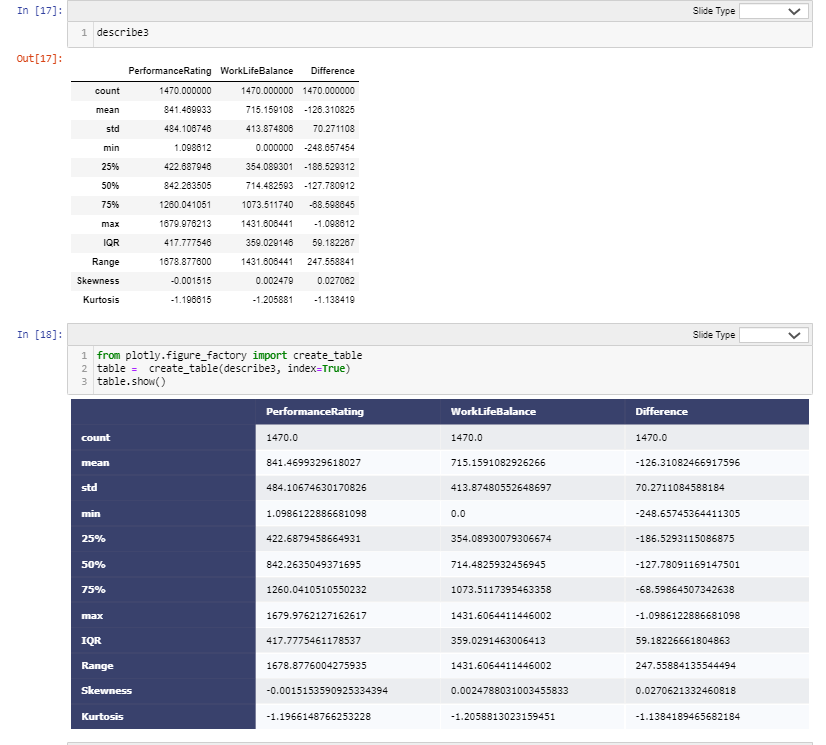
1. Performance Rating has outperformed Work Life Balance.

Q&A answered:

1. Highest Performance Rating: 4;
2. Highest cumulative Performance Rating: 1679.97;
3. Mode of Performance Rating: 3;
4. Median of Performance Rating: 3;
5. Mean of Performance Rating: 3.15 ~ 3;
6. Since mean=median=mode, it seems a normal distibution.
7. Highest Work Life Balance: 4;
8. Highest Cumulative Work Life Balance: 1431.6;
9. Mode of Work Life Balance: 3;
10. Median Work Life Balance: 3;
11. Mean of Work Life Balance: 2.8 ~ 3;
12. Since mean=median=mode, it seems a normal distibution.

**Anomalies Detected:**

1. At -1.10, 0 had the highest Difference and was 99.56% higher than 1468, which had the lowest Difference at -248.66.
2. Difference and total Performance Rating are negatively correlated with each other.
3. 1468 accounted for 0.13% of Difference.
4. Across all 1,470 Column1, Difference ranged from -248.66 to -1.10, Performance Rating ranged from 1.10 to 1,679.98, and WorkLife Balance ranged from 0 to 1,431.61.



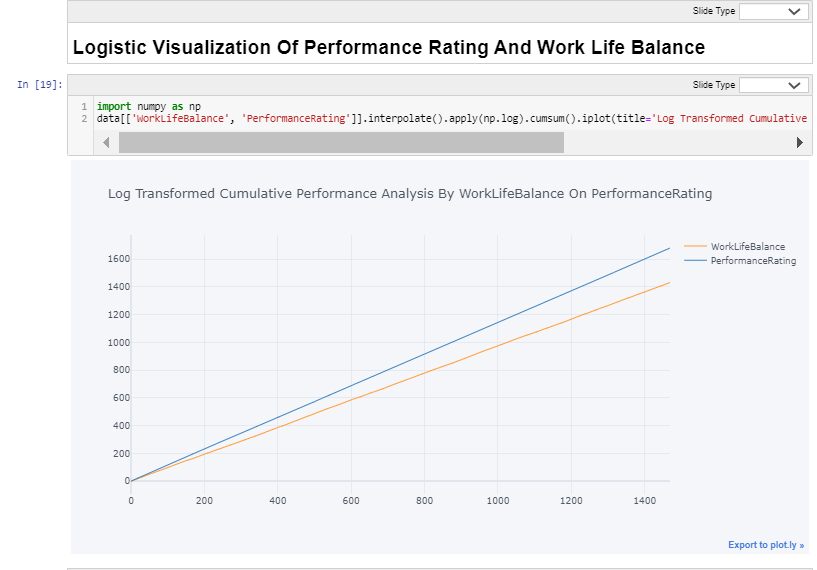
**Mathematical Notation:**

1. Mean = sum of values/count of values
2. std = sqrt (((value - mean of distribution) \*\*2 / number of values))
3. 3 quartiles are measures of variance, calculated to spot the placeholder value, it returns index of the produced value. Step 1: sort the dataset  
   Step2:  
   i) Lower Quartile (Q1: 25% distribution) = ((number of values+1)/4)th Term  
   ii) Middle Quartile (Q2: 50% distribution) = ((number of values +1)/2)th Term  
   Also, known as median (central value).  
   iii) Upper Quartile (Q3: 75% distribution) = ¾ (number of values + 1)th Term  
   iv) IQR = Upper Quartile - Lower Quartile
4. Range = Maximum Value - Minimum Value
5. Skewness = (summation (value - mean of distribution)**3)/ (number of values - 1) \* std**3)
6. Kurtosis = number of values \* (summation (value - mean of distribution)**4) / std**4)

**Verbal Translation**

# **Performance Rating = (Negative Observed Performance / Normal Performance) \* 100; As per academic theories there remains a positive relation between employee well-being and employee performance.**

1. It can be observed that there has been a constant difference between employee well-being and employee performance of -126.31.
2. Performance Rating has been at 841.47.
3. Maximum performance rating: Work Life Balance has been 1679.976213:1431. 606441..
4. The Median Performance Rating is 118% of Performance Work Life Balance (842.263505/714.482593)
5. Kurtosis for all the features is less than 3, hence, indicating platykurtic curve.

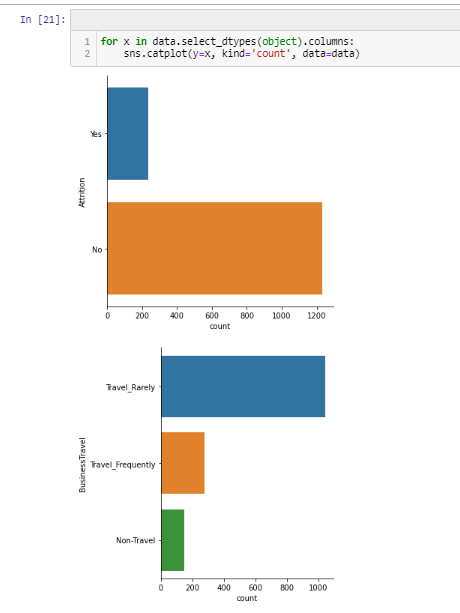


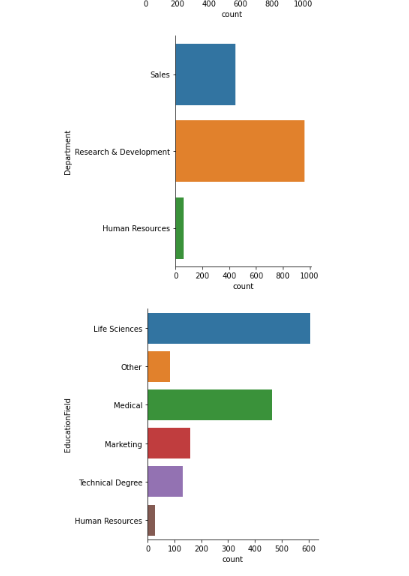
# **Observations:**

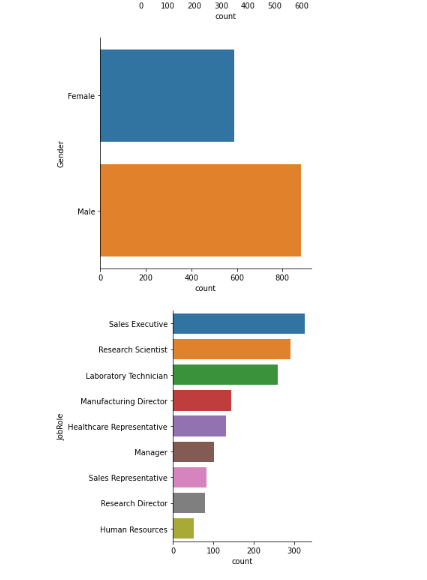
1. I have transformed the data into lognormal for better visual understanding of data.
2. The coordinates for each line are (index, value)
3. Until index 33, both the axes are overlapping. Divergence starts from index 34 and the difference keeps increasing thereon.
4. The max of cumulative sum of difference of two axes is -248.369772 (1431.606441-1679.976213)

Analysis 3: Data Exploration by Group by Method on Categorical Data

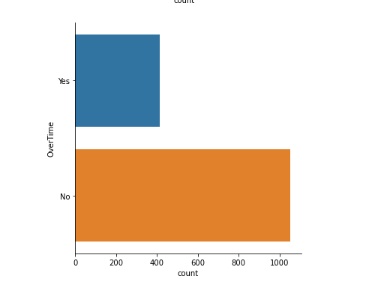
*Snippets From Notebook*



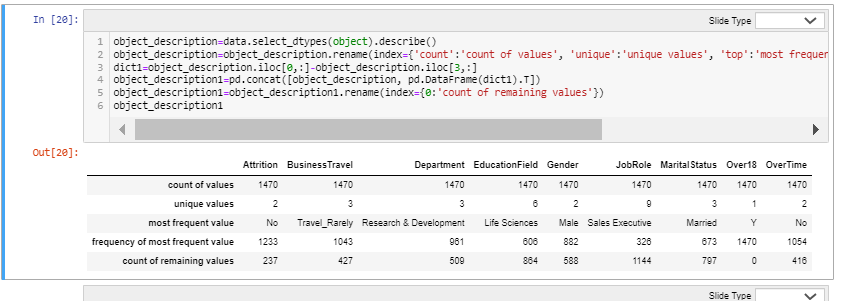








Tabulated Analysis of Count plots



Verbal Translation of Above Table

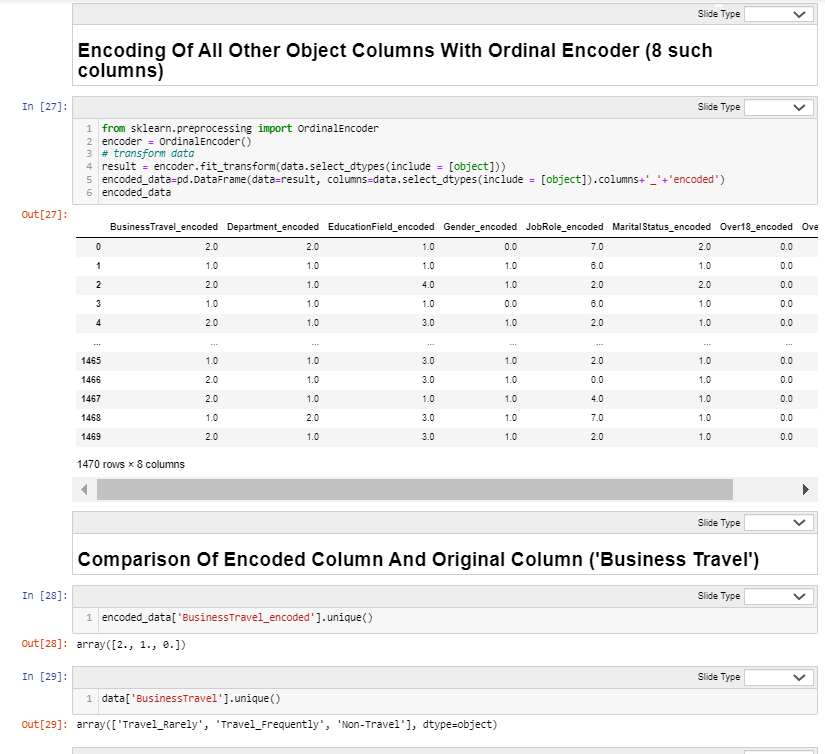
1. Unique Values in ALL The Categorical Columns

(Business Travel = array (['Travel Rarely', 'Travel Frequently', 'Non-Travel'], dtype=object),  
 Department = array(['Sales', 'Research & Development', 'Human Resources'], dtype=object),  
 Education Field = array (['Life Sciences', 'Other', 'Medical', 'Marketing',  
 'Technical Degree', 'Human Resources'], dtype=object),  
 Gender = array(['Female', 'Male'], dtype=object),  
 Marital Status = array (['Single', 'Married', 'Divorced'], dtype=object))

1. There are 9 object columns, of which 3 are spread across yes and no:
   * 1. Attrition has 1233 No Values and 237 Yes Values (237 employees are prone to attrition).
     2. Over18 has all 1470 Yes Values (entire workforce is above 18).
     3. Overtime has 1054 No values and 416 Yes values (less than half of the workforce does overtime).
2. Apart from these, the remaining 6 columns have more than 2 unique values and are arranged in the cell above:
3. Rows 3, 4 and 5 explain the most frequent value and its count and data portion occupied by other unique values.
4. For example, there are 882 Male and 588 Female, etcetera as displayed above.

Encoding Of Data

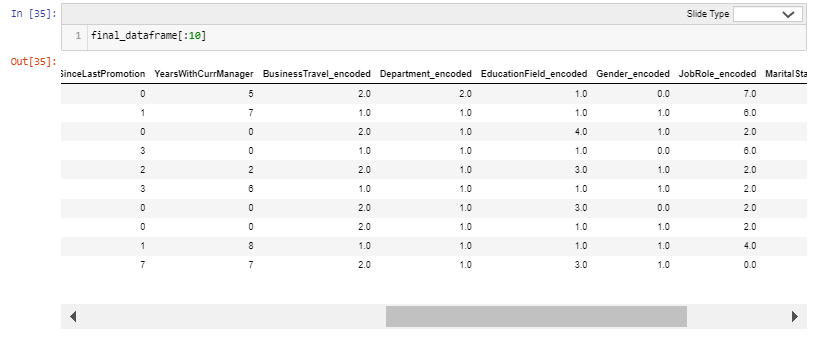
*Snippets From Notebook*



Conclusion

1. Travel Rarely is encoded to 2, Travel frequently to 1 and Non Travel to 0.
2. Likewise, all the unique values in all the columns are assigned individual codes.

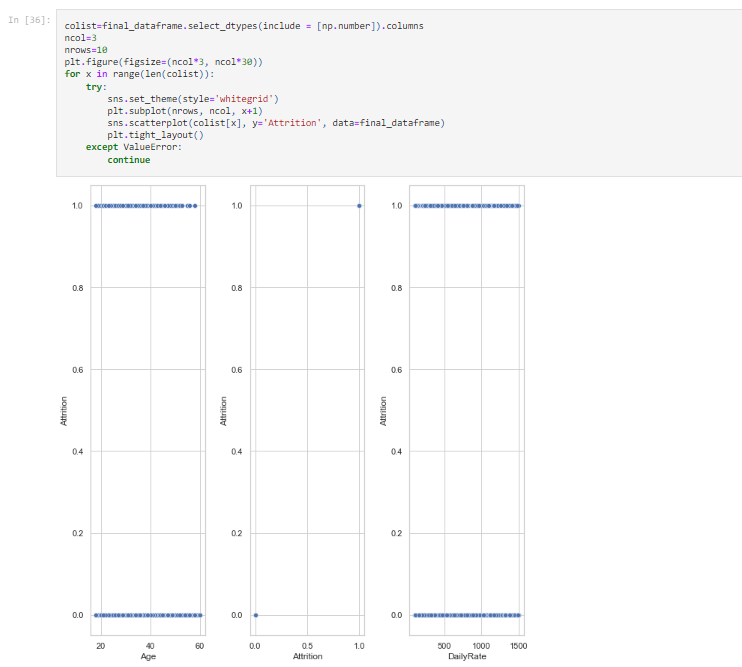
View Of Final Data Frame After Concatenation of Encoded Values.

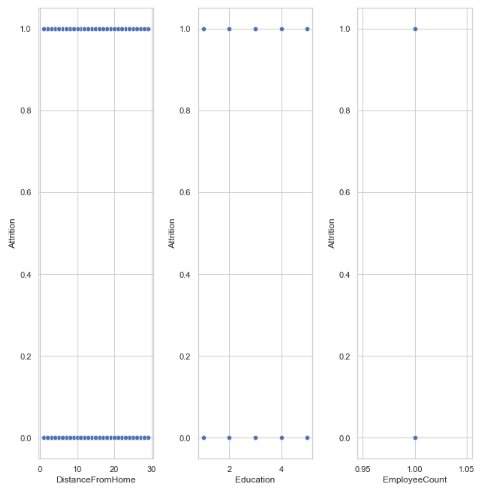


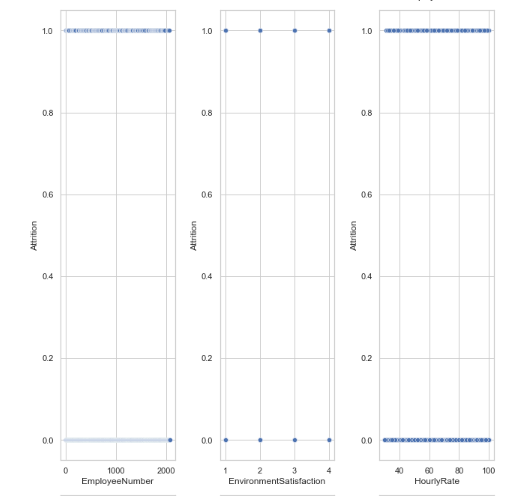
The above table represents the first 10 rows from the final data frame.

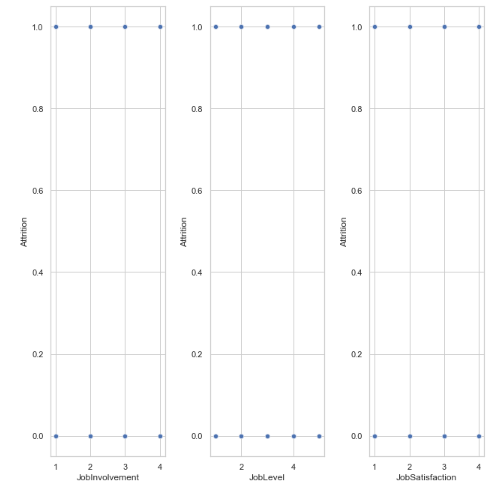
EDA Part 1

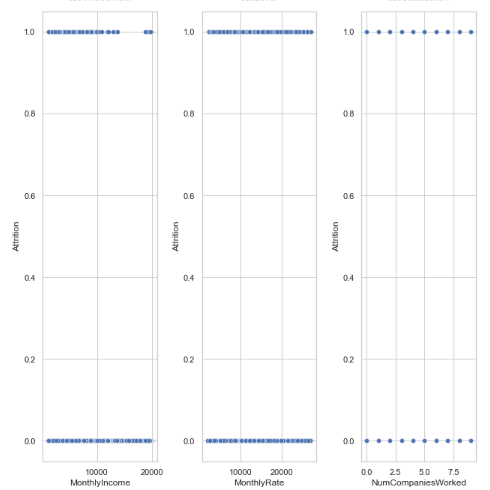
Analysis 4: Scatter Plots

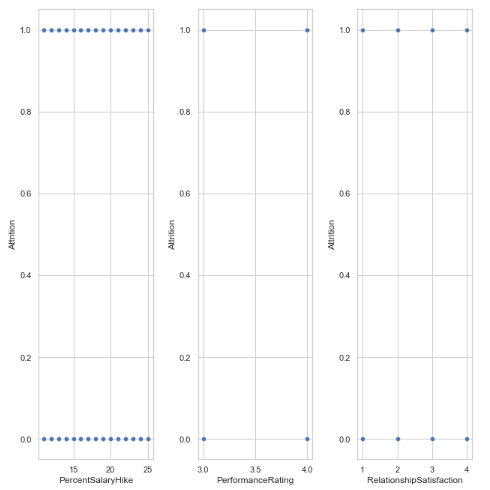


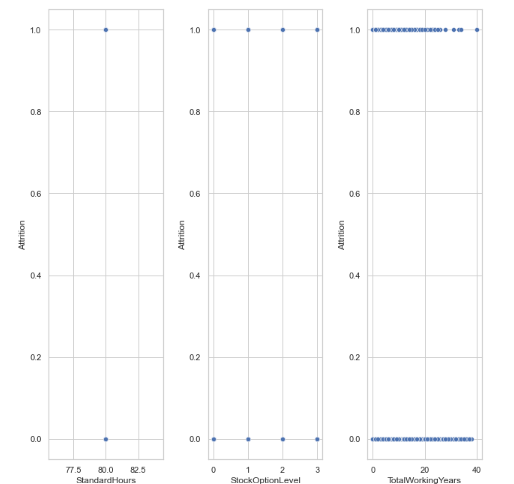


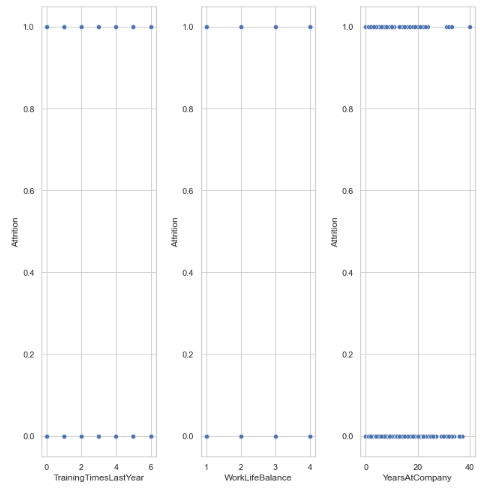


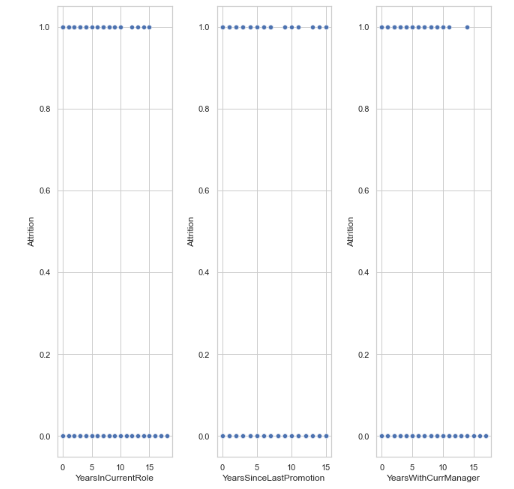


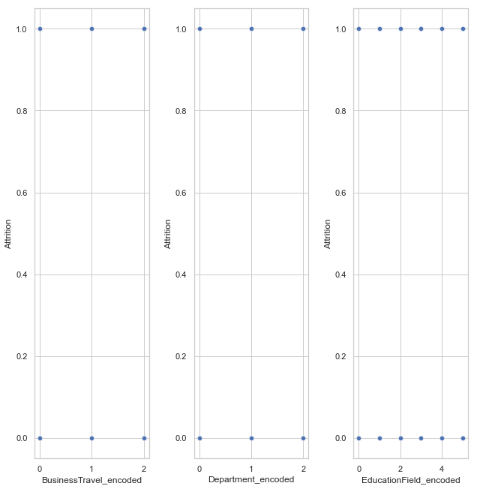












It can be visualized from above scatter plots against outcome vector that distribution is mostly evenly distributed for most of the columns. For example,

i. Age

ii. Daily Rate

iii. Distance from Home

iv. Employee Number

v. Environment Satisfaction

vi. Hourly Rate

vii. Job Involvement

viii. Job Level

ix. Job Satisfaction

x. Monthly Rate

xi. Number of Companies Worked

Likewise, all the columns.

For example,

1=Yes and 0=No

i. in years in current role, coordinate are evenly distributed between Attrition 0 and 1, except that 0 has a few extra points between 10 to 15 and beyond 15, depicting that people who have stayed in current role for more than 10 years have higher chance of attrition.

ii. people beyond 5 years since last promotion have higher chance of attrition.

iii. chances of attrition are high for total working years beyond 20.

iv. chances of attrition are high for total working years in current company beyond 20 years.

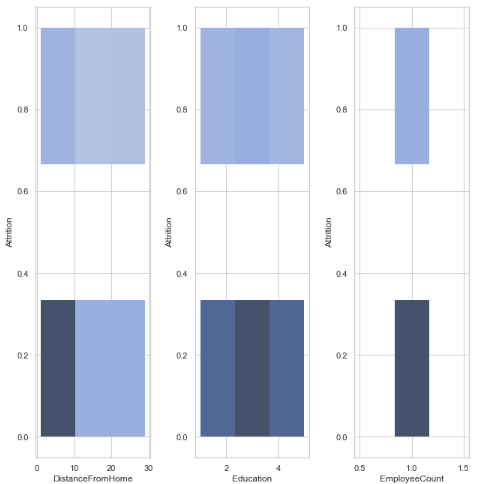
v. the chance of attrition is high for people who have been working with the current manager for more than 10 years.

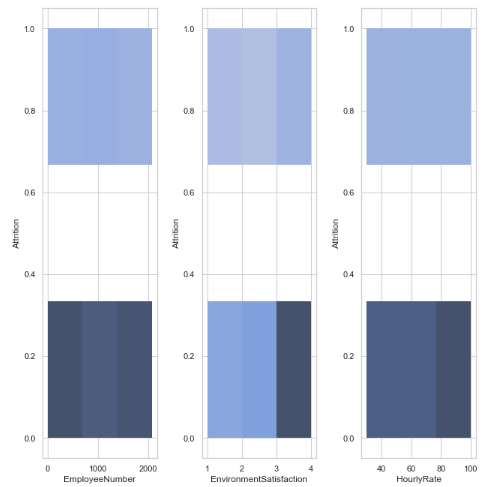
vi. Education field has equal number of yes and no.

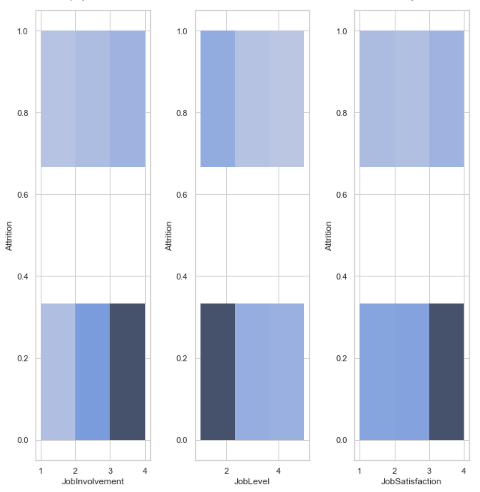
vii. Chances of attrition are high for monthly income beyond 10000.

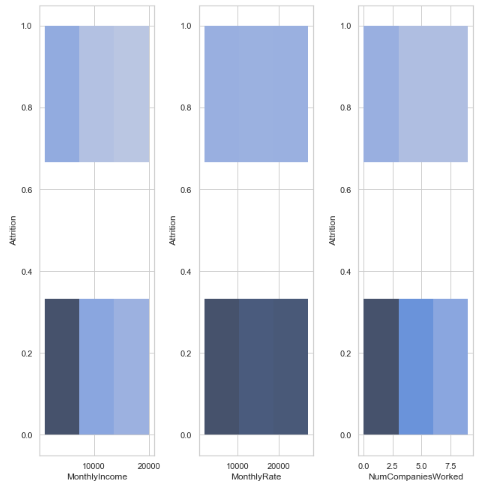
Analysis 5: Histograms

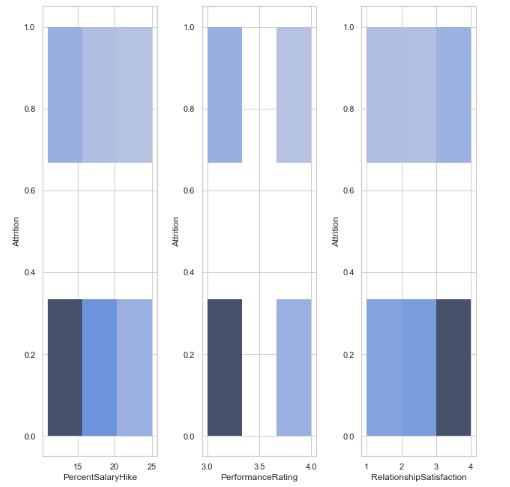




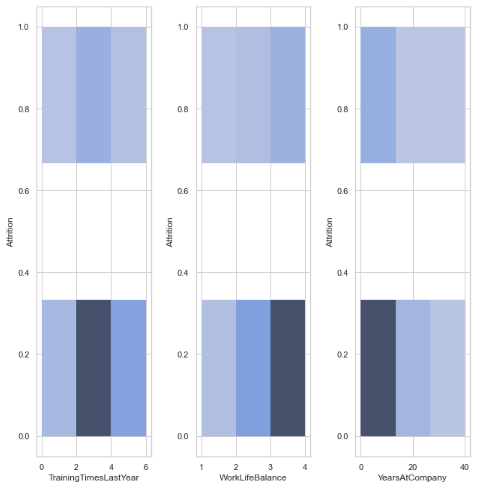


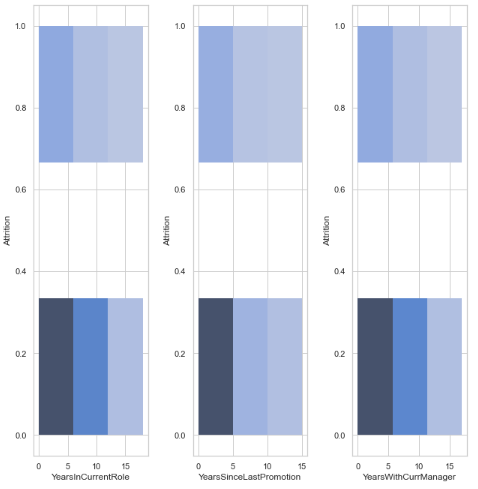


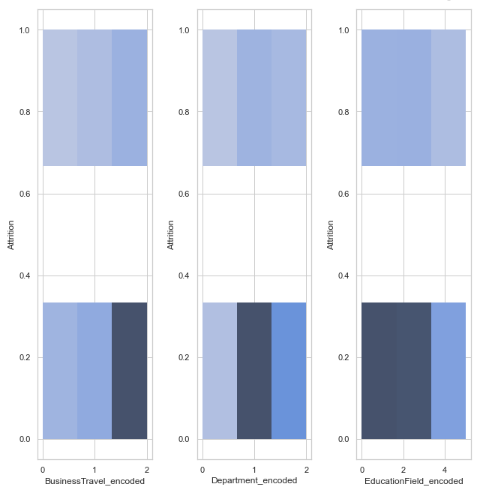












Observations:

# The above histograms show frequency distribution of features against label. It is divided into 3 bins. Only Business Travel shows skewness left. Many datapoints, WorkLife Balance, Job Involvement, Monthly Rate, etcetera seems to be close to normal with skewness threshold within +/-0.65.

3 columns have 0 skewness (as desired in normal distribution):

i. Employee Count,

ii. Standard Hours.

iii. Over 18 encoded.

Analysis 6: Correlation Analysis

# **Graphical Representation of Correlation of Features with Label**



# The above iplot represents correlation of features with label. The correlation of features with label is of high relevance. The stronger the relationship of label with axis, the more accurate the prediction. In the above line graph:

1. Highest correlation with label is observed to be 24.61%, that is shared with over time encoded. Minimum correlation is 0.3%, shared with Performance Rating.

2. Weak Positive to Strong Positive Relationship Is Found with Following Features:

i. BusinessTravel\_encoded 7.377694602219632e-05

ii. PerformanceRating 0.0028887517110809

iii.MonthlyRate 0.015170212530471473

iv. EducationField\_encoded 0.026845545711446116

v. Gender\_encoded 0.029453253175141608

vi. NumCompaniesWorked 0.04349373905781363

vii.Department\_encoded 0.06399059633809044

viii.JobRole\_encoded 0.0671514950495707

ix. DistanceFromHome 0.07792358295570369

x. MaritalStatus\_encoded 0.1620702346570145

xi. OverTime\_encoded 0.24611799424580436

3. Weak Negative to Strong Negative Relationship Is Found with Following Features:

i. TotalWorkingYears -0.17106324613622612

ii. JobLevel -0.16910475093102761

iii.YearsInCurrentRole -0.16054500426770077

iv. MonthlyIncome -0.15983958238498852

v. Age -0.15920500686577962

vi. YearsWithCurrManager -0.1561993159016288

vii.StockOptionLevel -0.13714491893332562

viii.YearsAtCompany -0.13439221398997708

ix. JobInvolvement -0.13001595678605374

x. JobSatisfaction -0.10348112606902123

xi. EnvironmentSatisfaction -0.10336897833793603

xii.WorkLifeBalance -0.06393904721740885

xiii. TrainingTimesLastYear -0.05947779855642057

xiv.DailyRate -0.05665199186762936

xv. RelationshipSatisfaction -0.0458722788811267

xvi.YearsSinceLastPromotion -0.033018775142584306

xvii.Education -0.031372819640049315

xviii. PercentSalaryHike -0.01347820205743911

xix.EmployeeNumber -0.010577242759242786

xx. HourlyRate -0.006845549572139952

4. Correlation of Label with itself is of no relevance

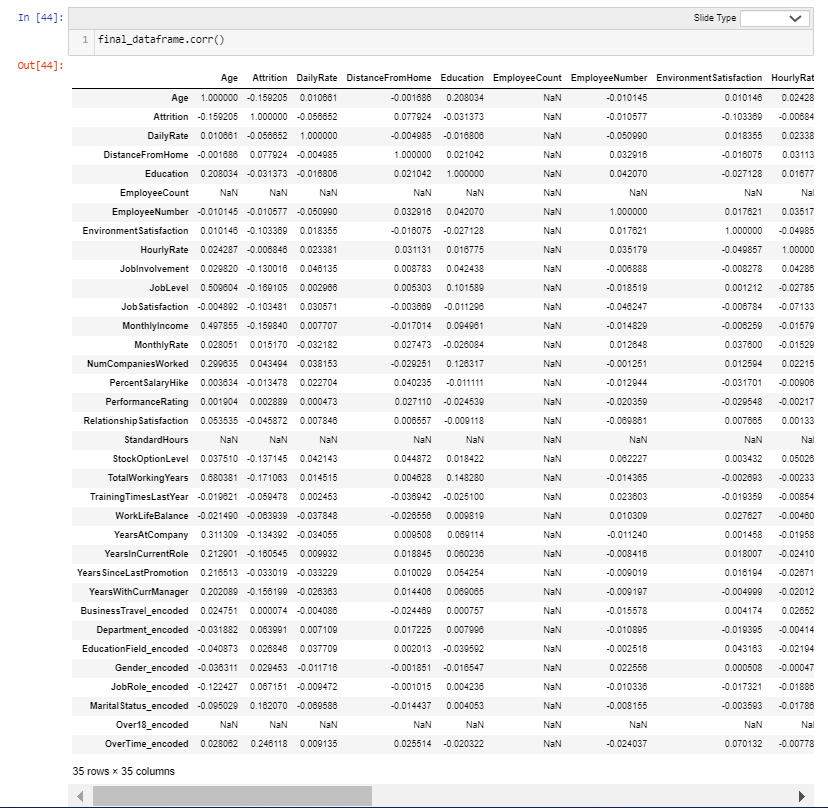
5. No Correlqtion is found with:

i. EmployeeCount

ii. Standard Hours

iii. Over 18 Encoded

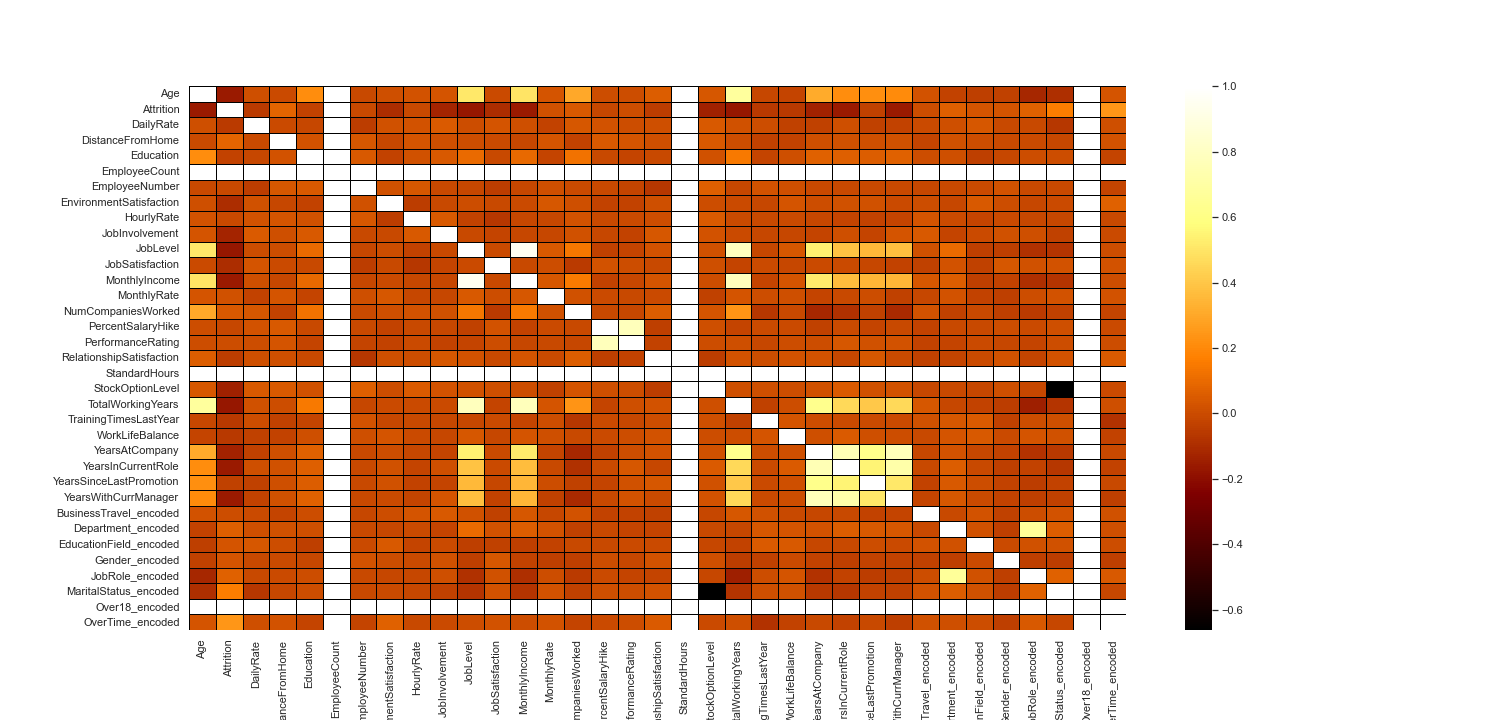
# **correlation among dataset**



**Observation:**

# **The above table represents correlation among dataset. There seems to be some multicollinearity due to presence of Job Level and Age; Marital Status Encoded and Stock Option Level; that show strong correlation between each other. Other than these columns there seems modest collinearity among features. We will do further EDA before arriving at a conclusion to delete these columns.**

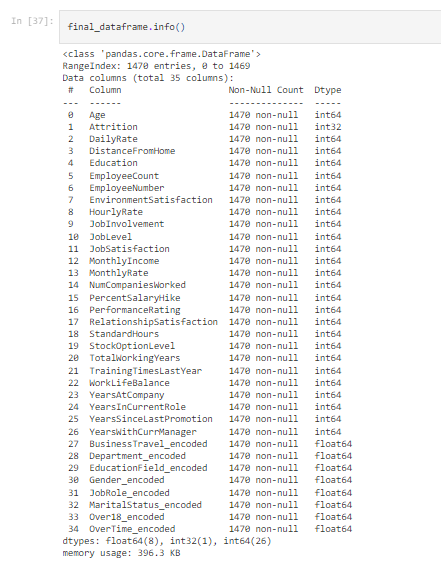
**Correlation Heatmaps:**



# **Verbal Translation of Above Graph:**

1. Label has moderate negative relationship with 3 features:
   * 1. Distance from Home
     2. Employee Number
     3. Job Satisfaction
     4. Training Times Last Year
     5. Work Life Balance
2. Label has a strong relationship with these features in this heatmap:
   * 1. Job Level
     2. Monthly Income
     3. Total Working Years
3. Strong Multicollinearity is detected among 7 features:
   * 1. Total Working Years
     2. Years at Company
     3. Years in Current Role
     4. Years Since Last Promotion
     5. Years with Current Manager
     6. Stock Option Level
     7. Marital Status Encoded
4. Multicollinearity seems to be moderate among all other data points.

Analysis 7: Dataset Information Table



# The above table reprents information dataset:

i. Columns

ii. Count of Non-Null Data Points

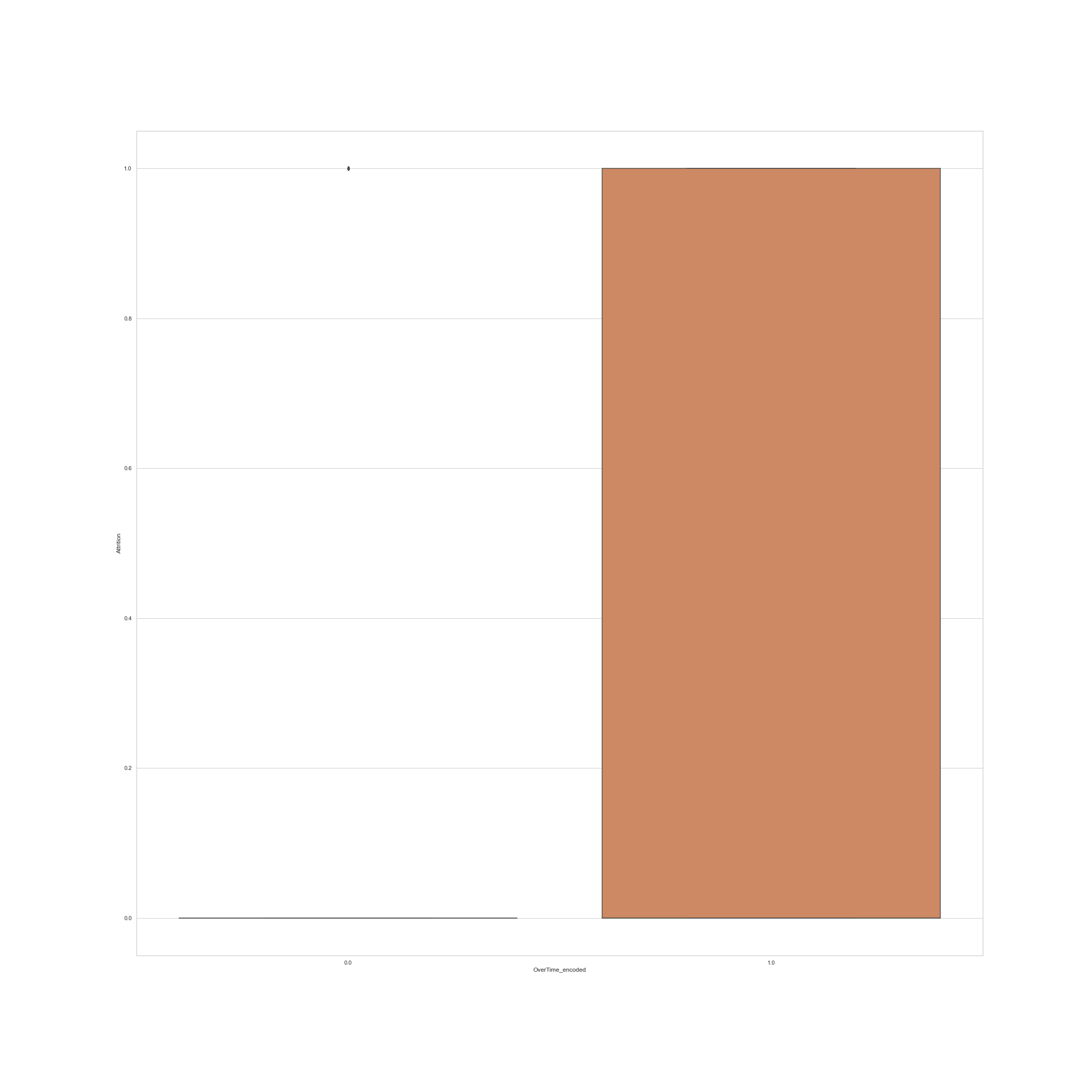
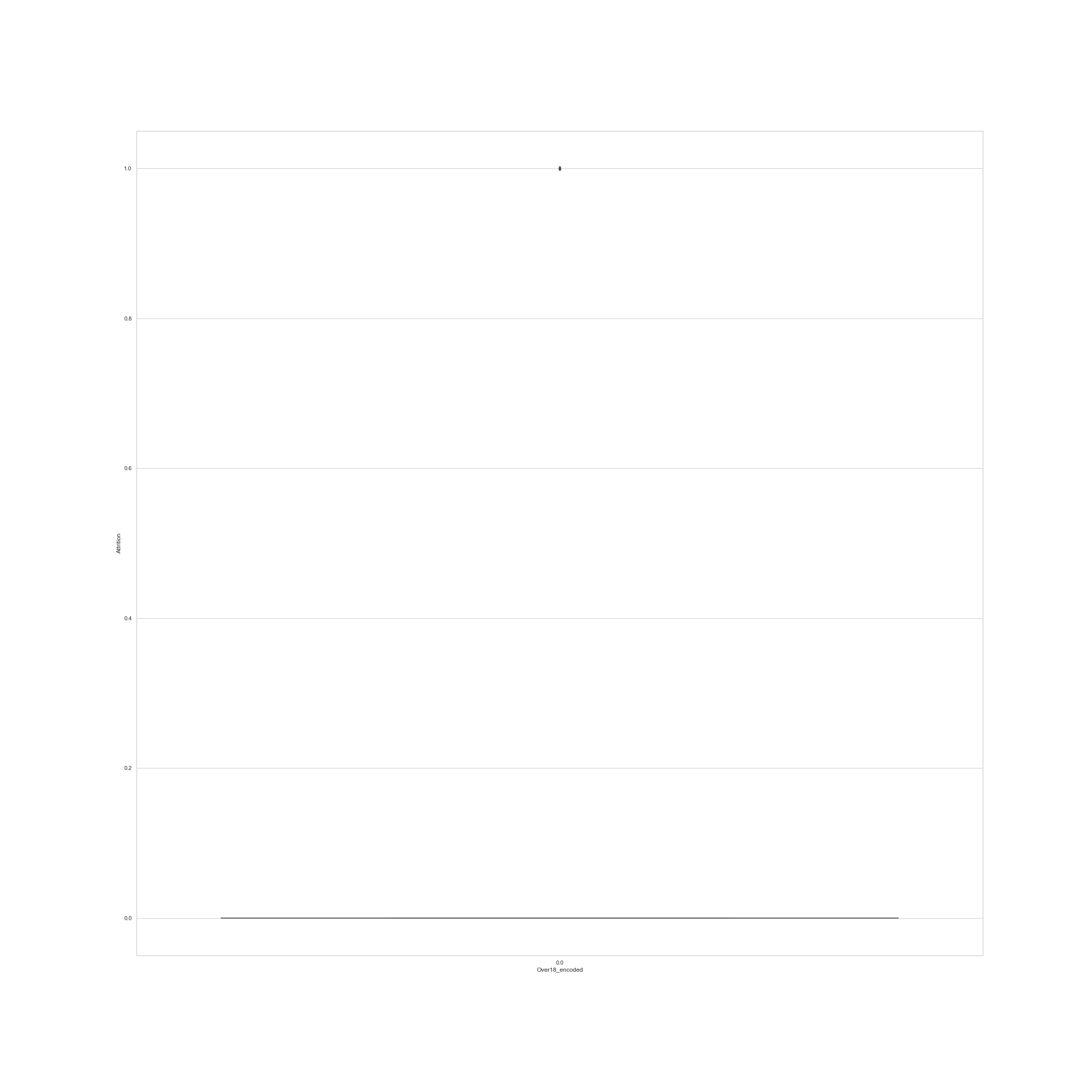
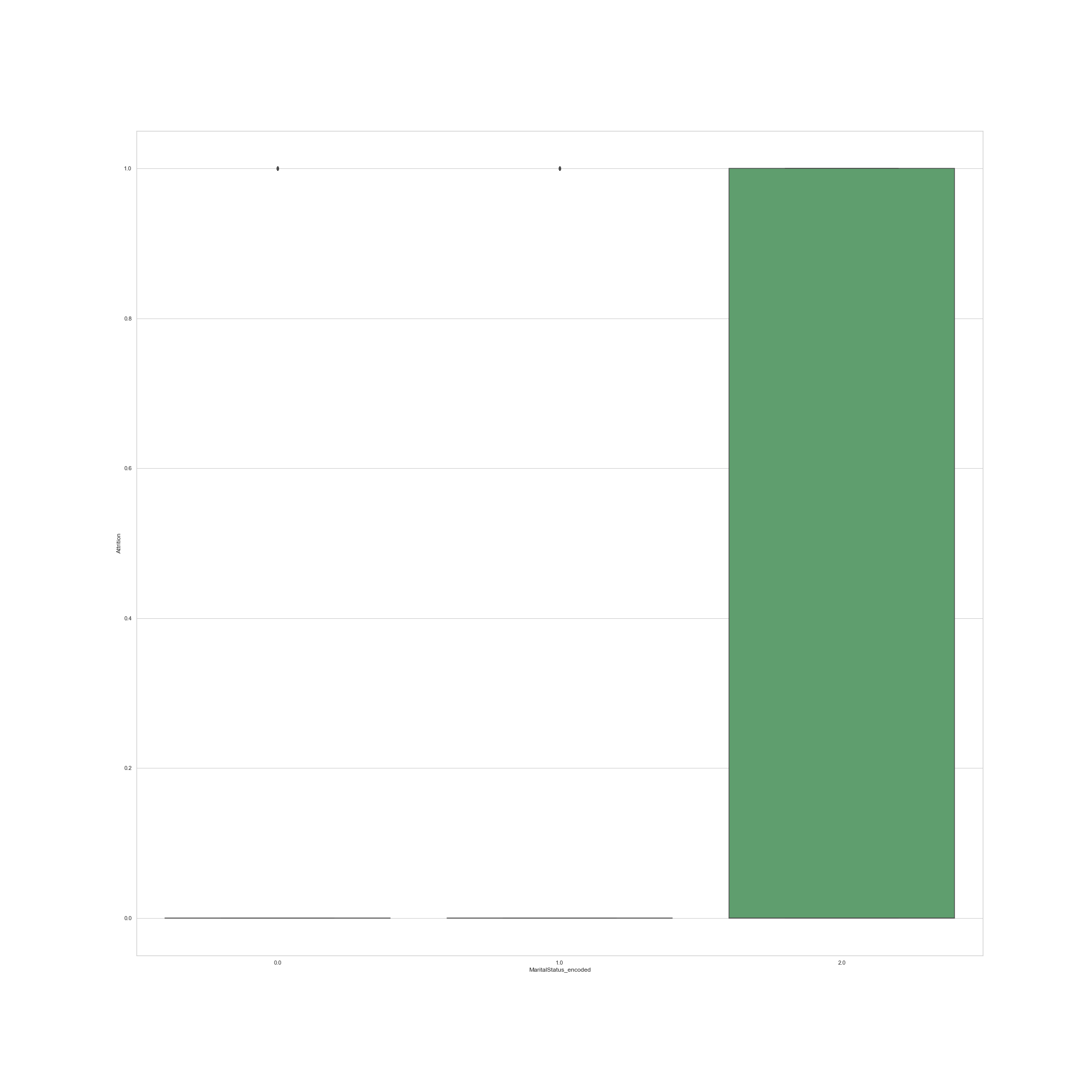
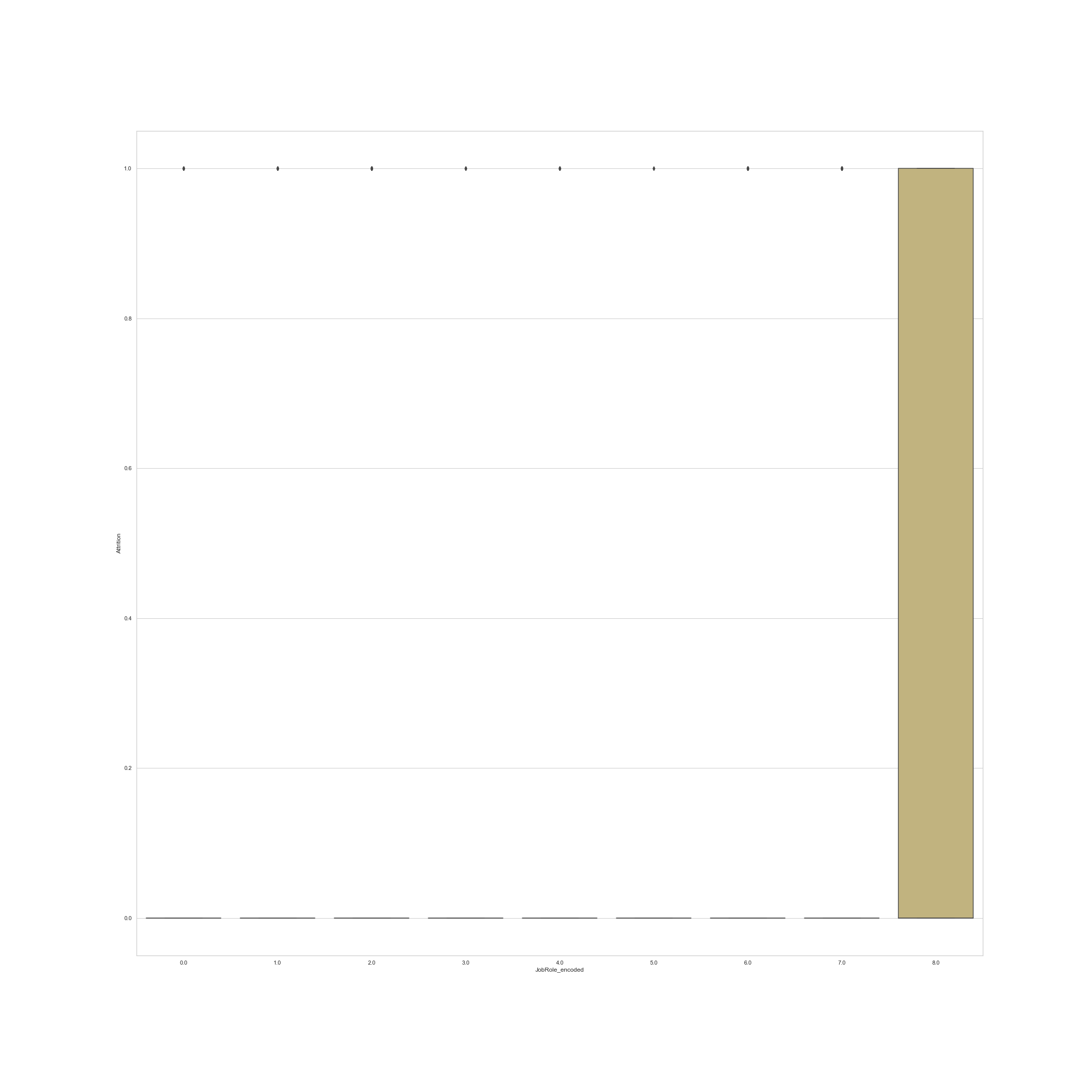
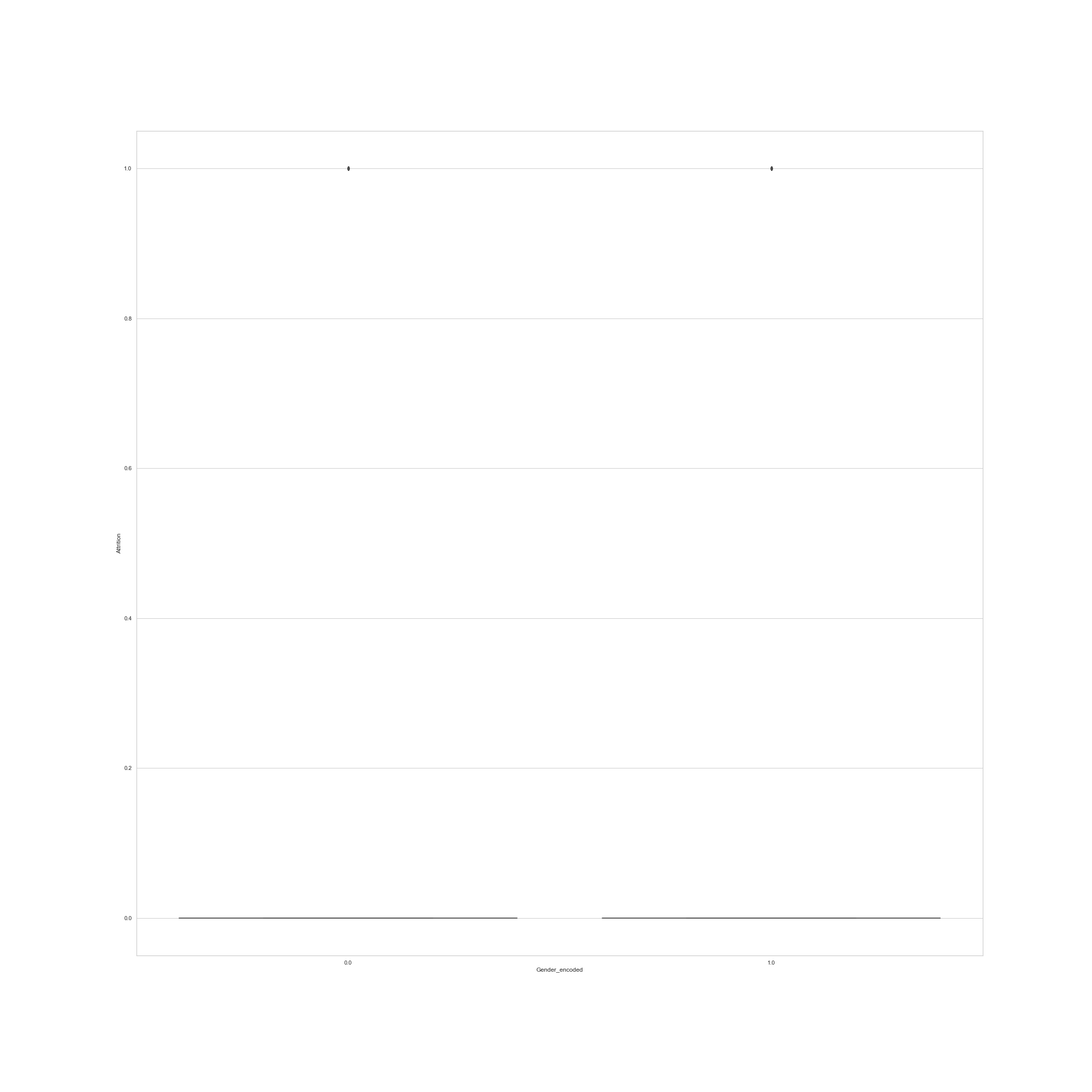
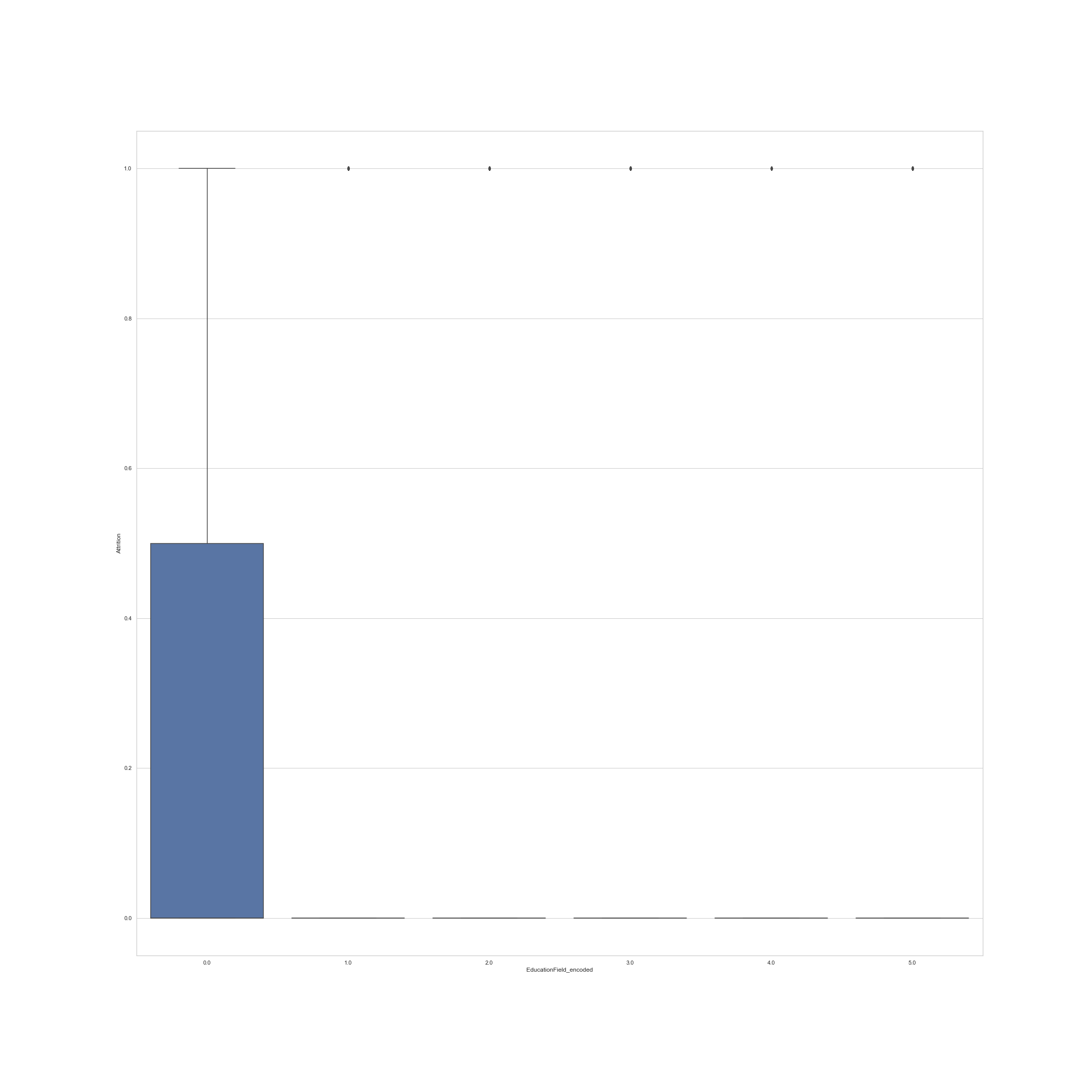
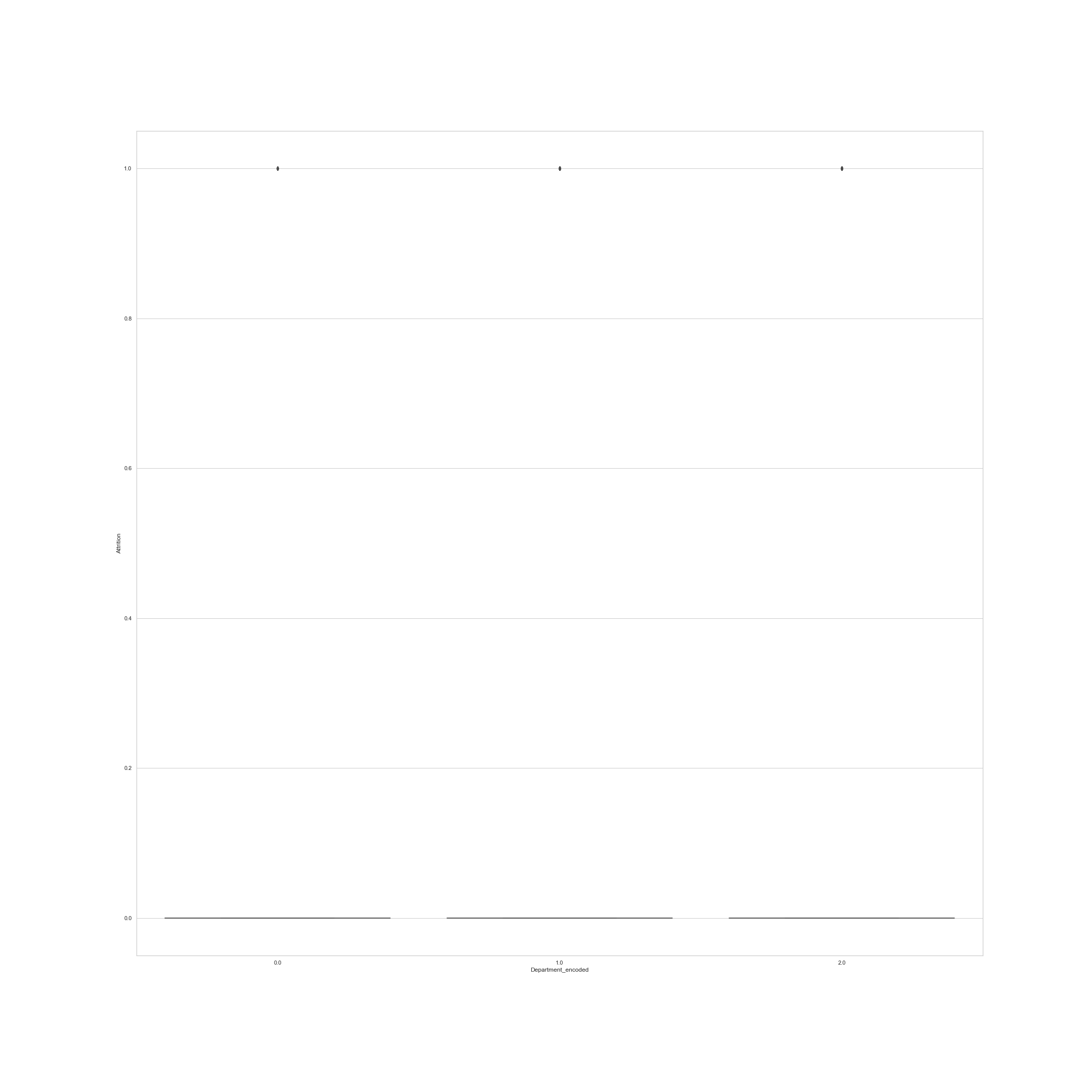
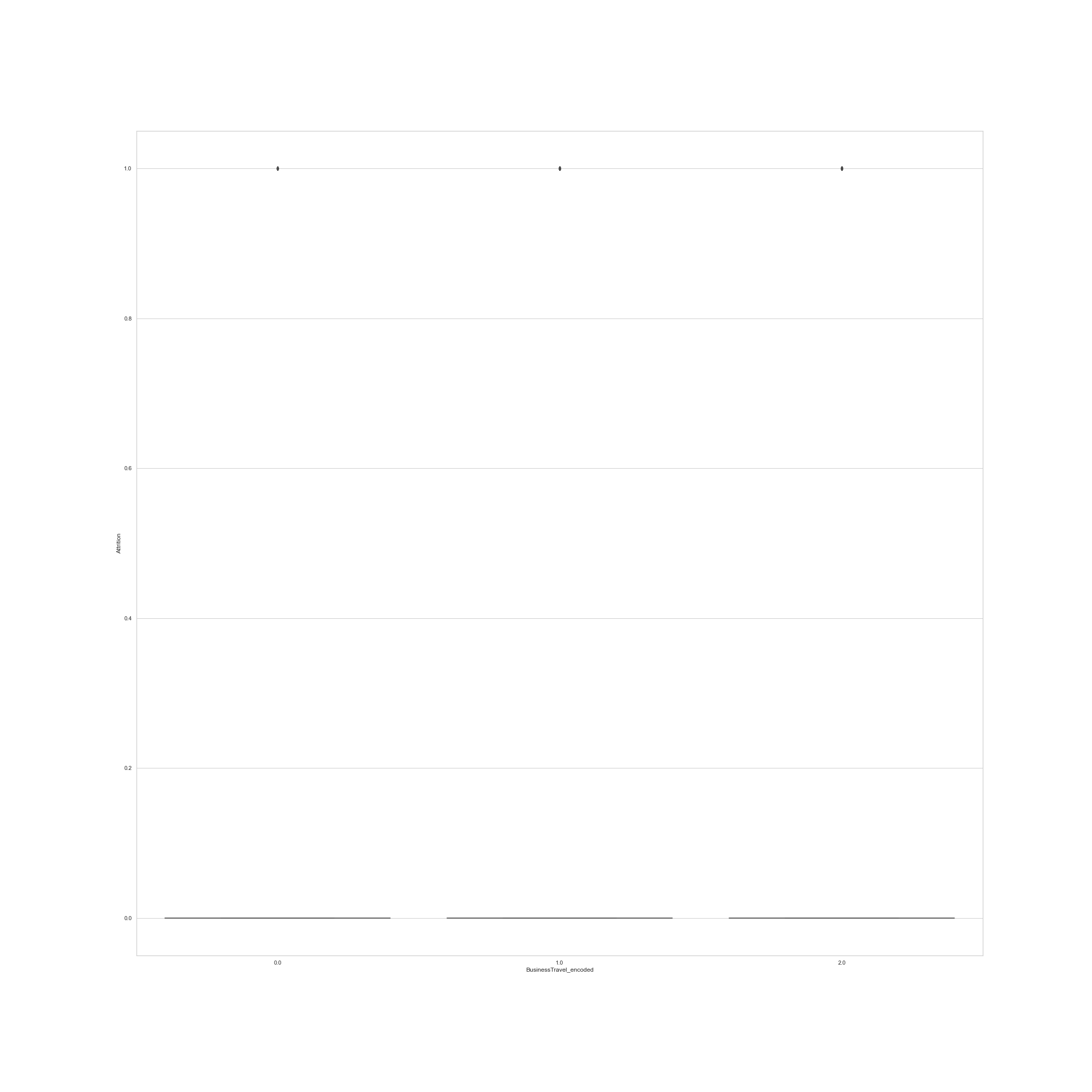
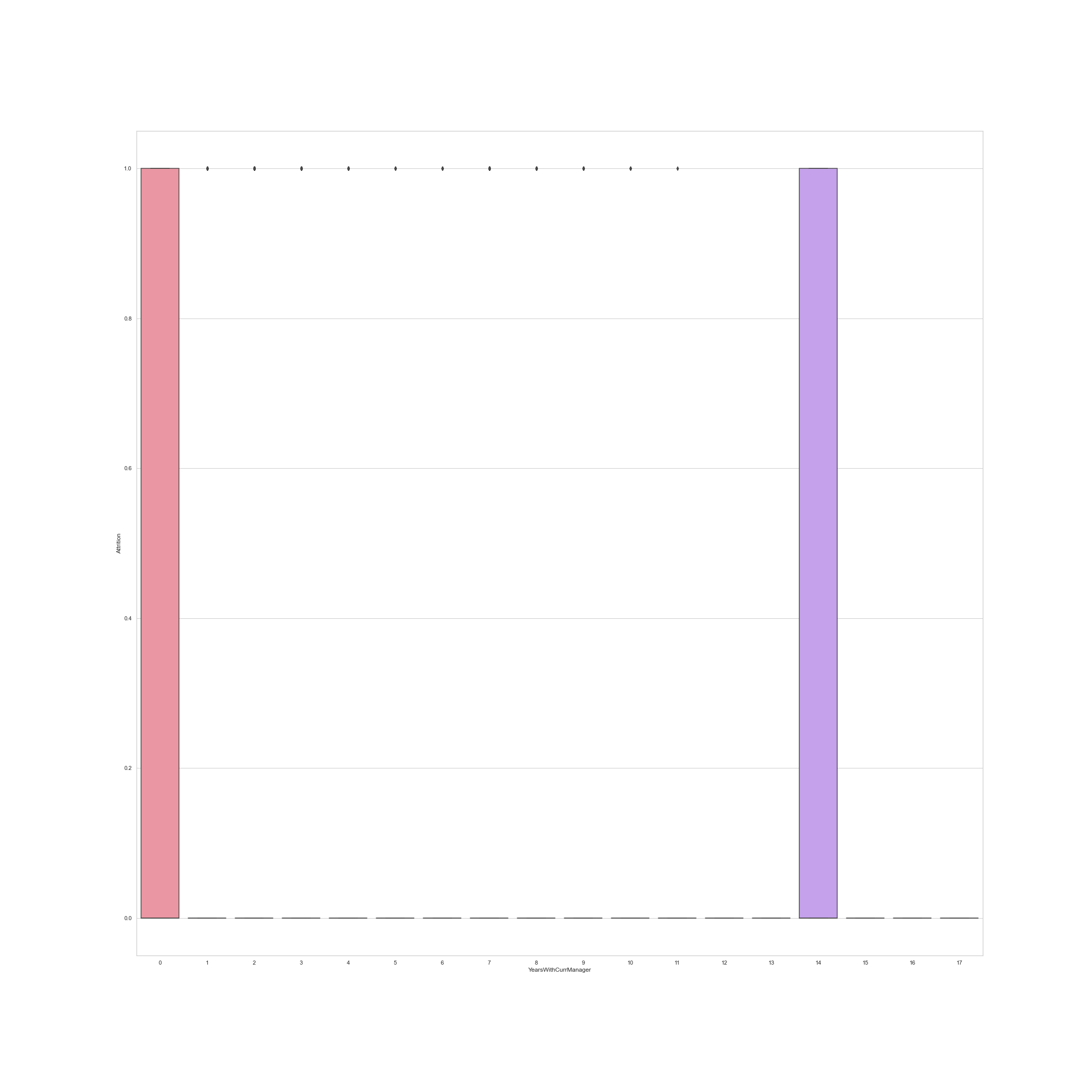
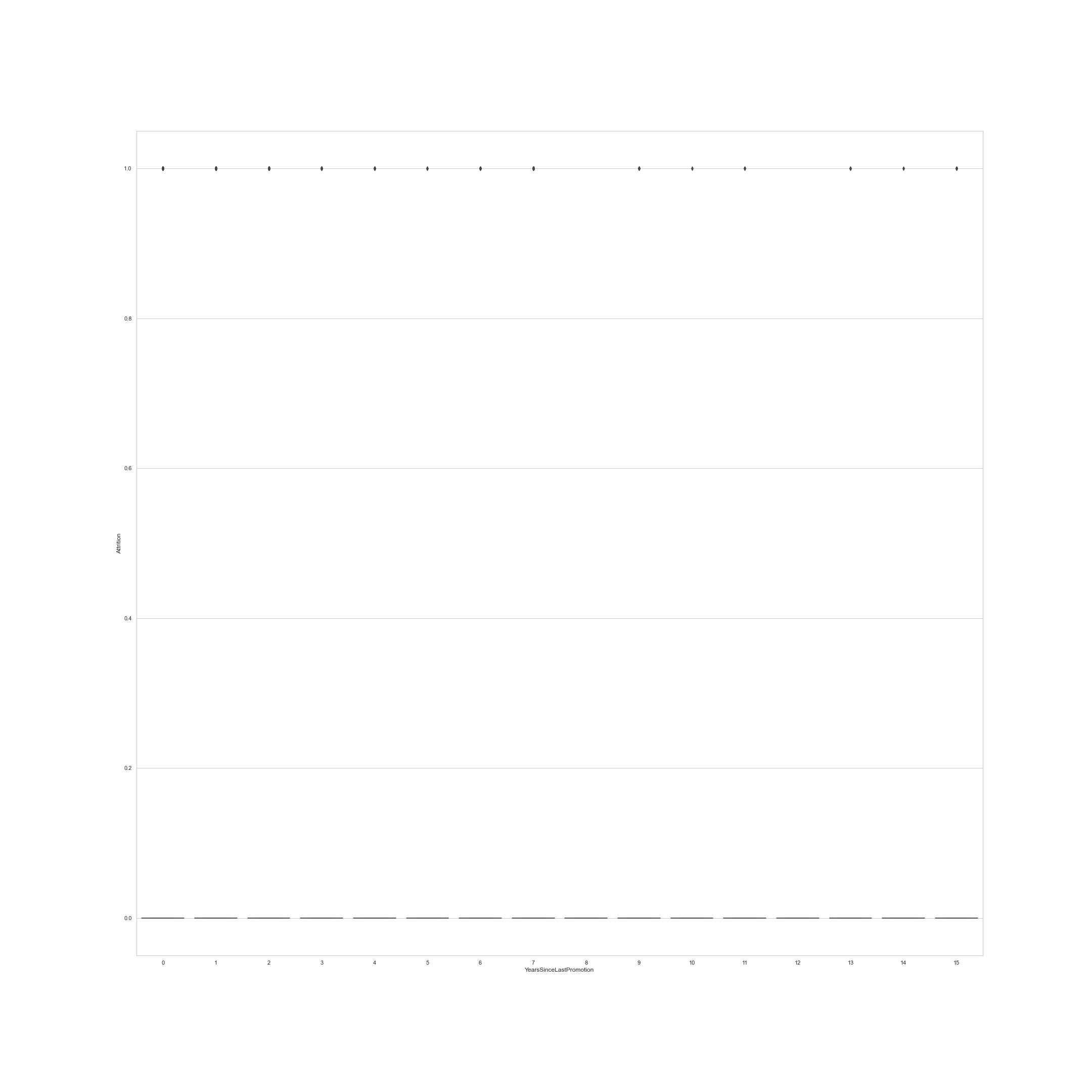
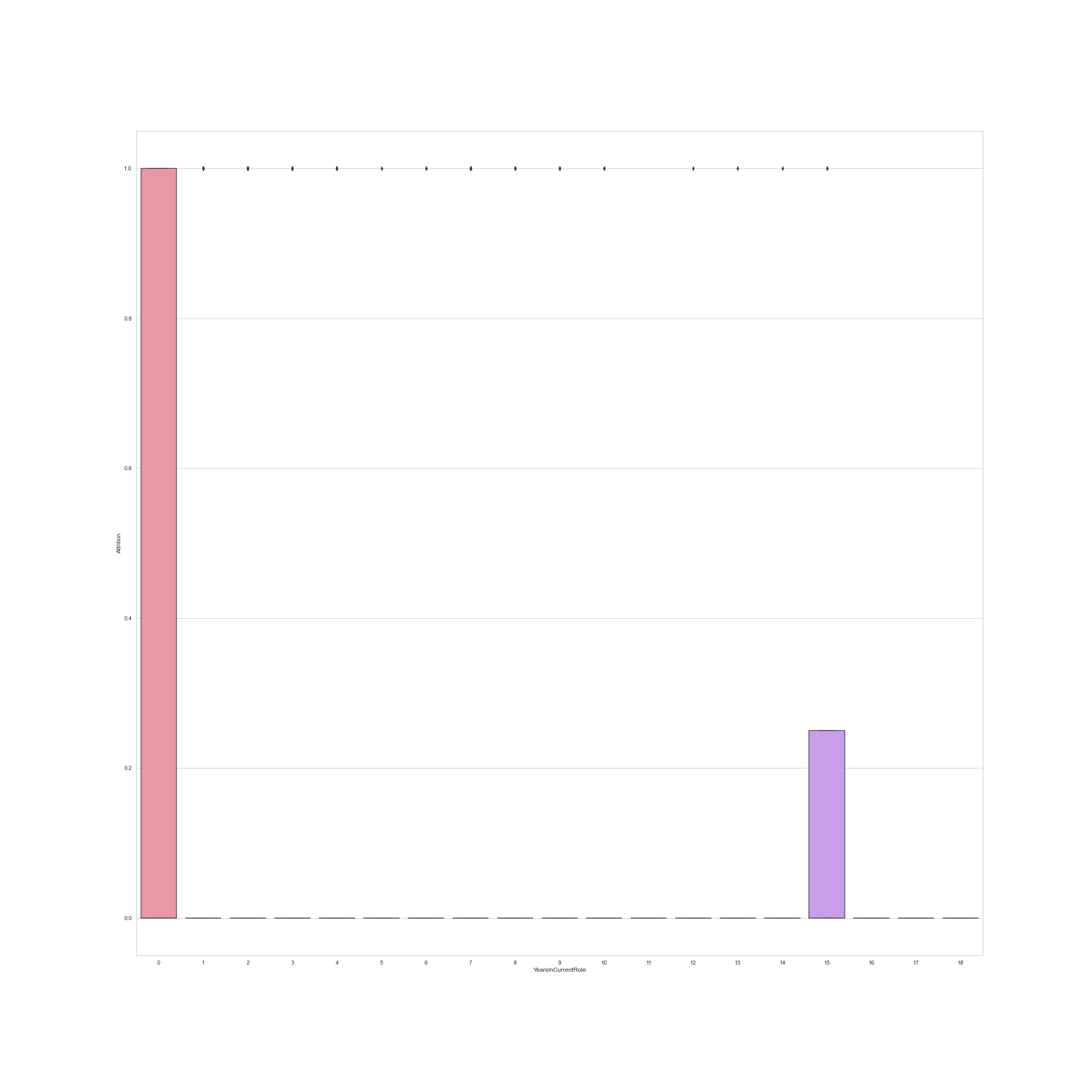
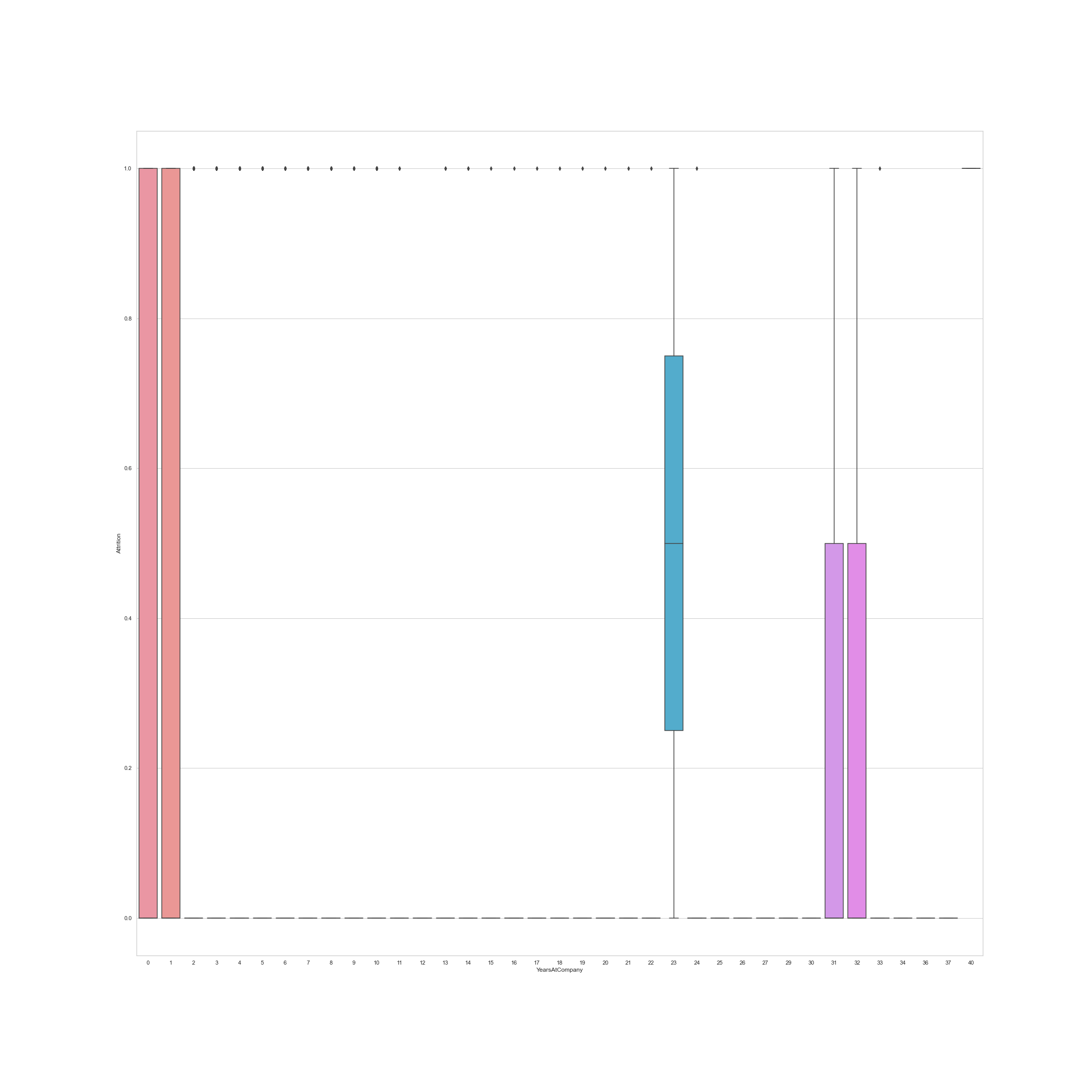
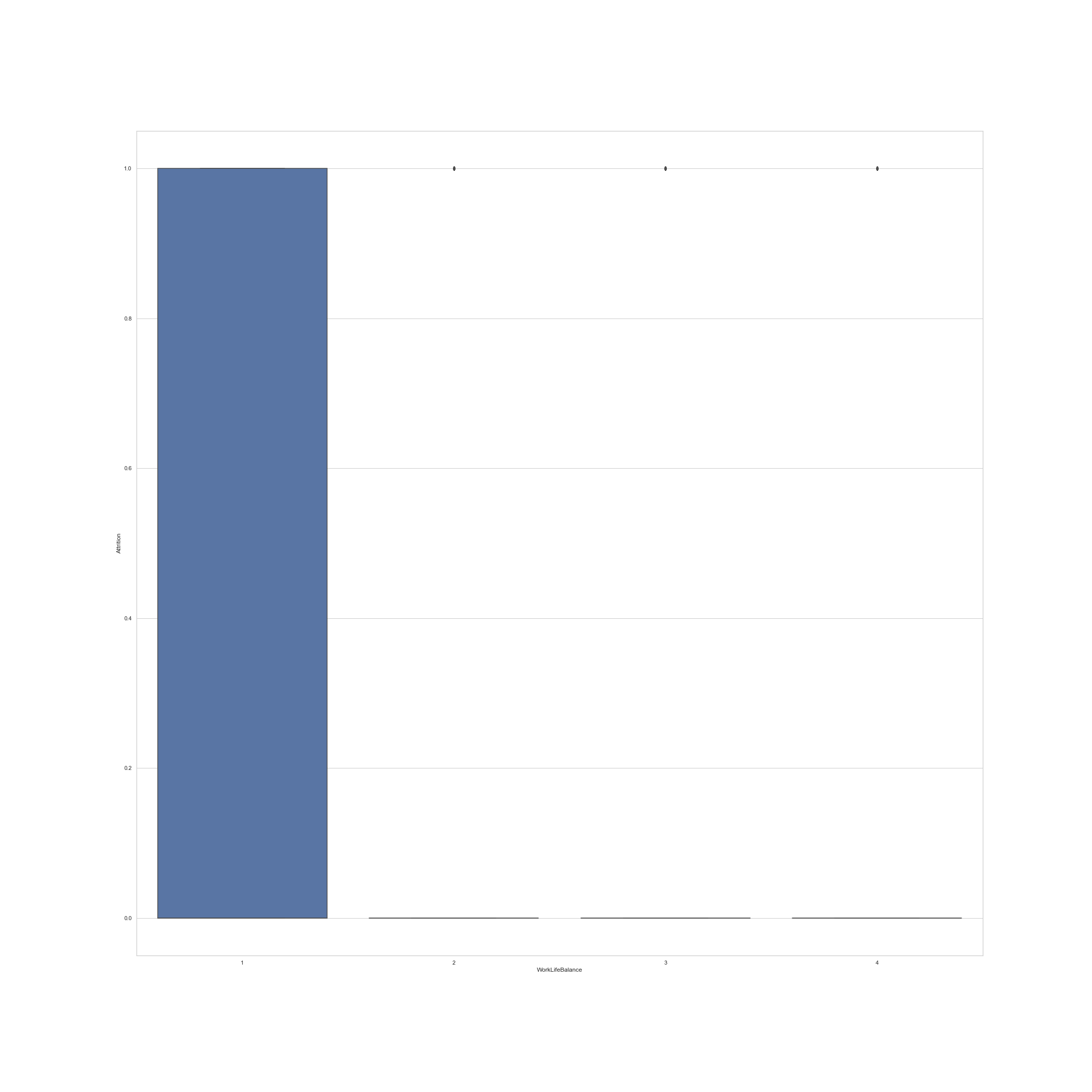
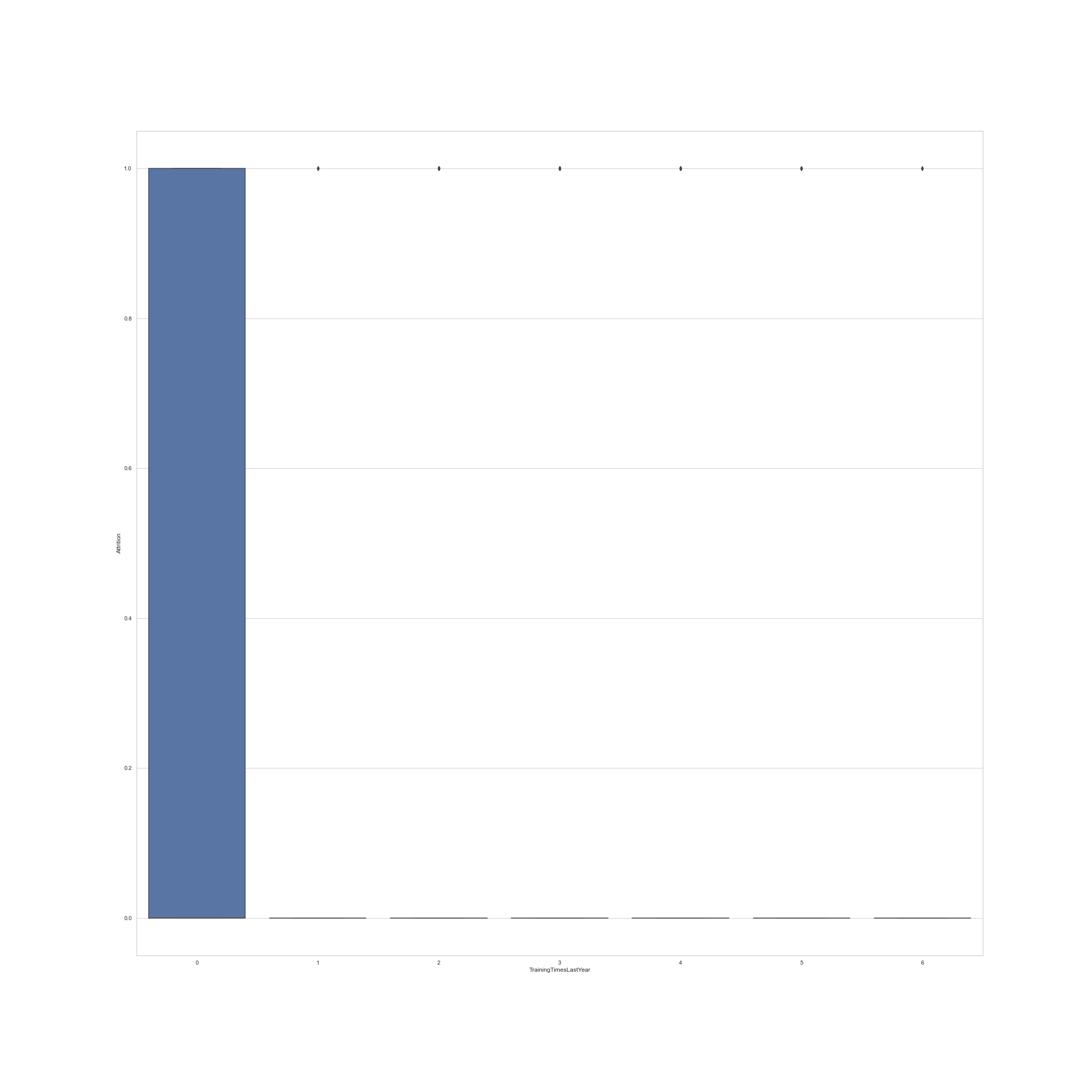
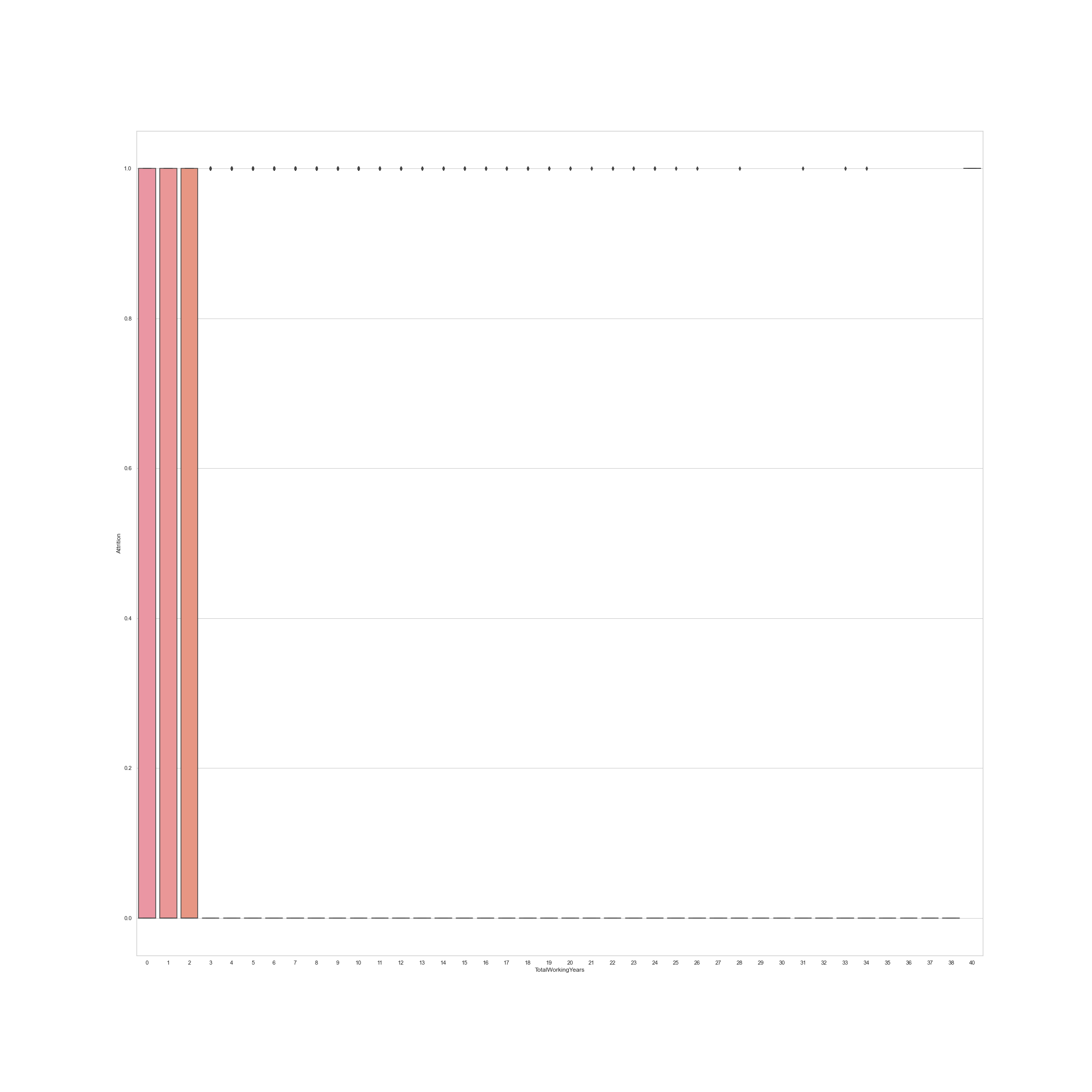
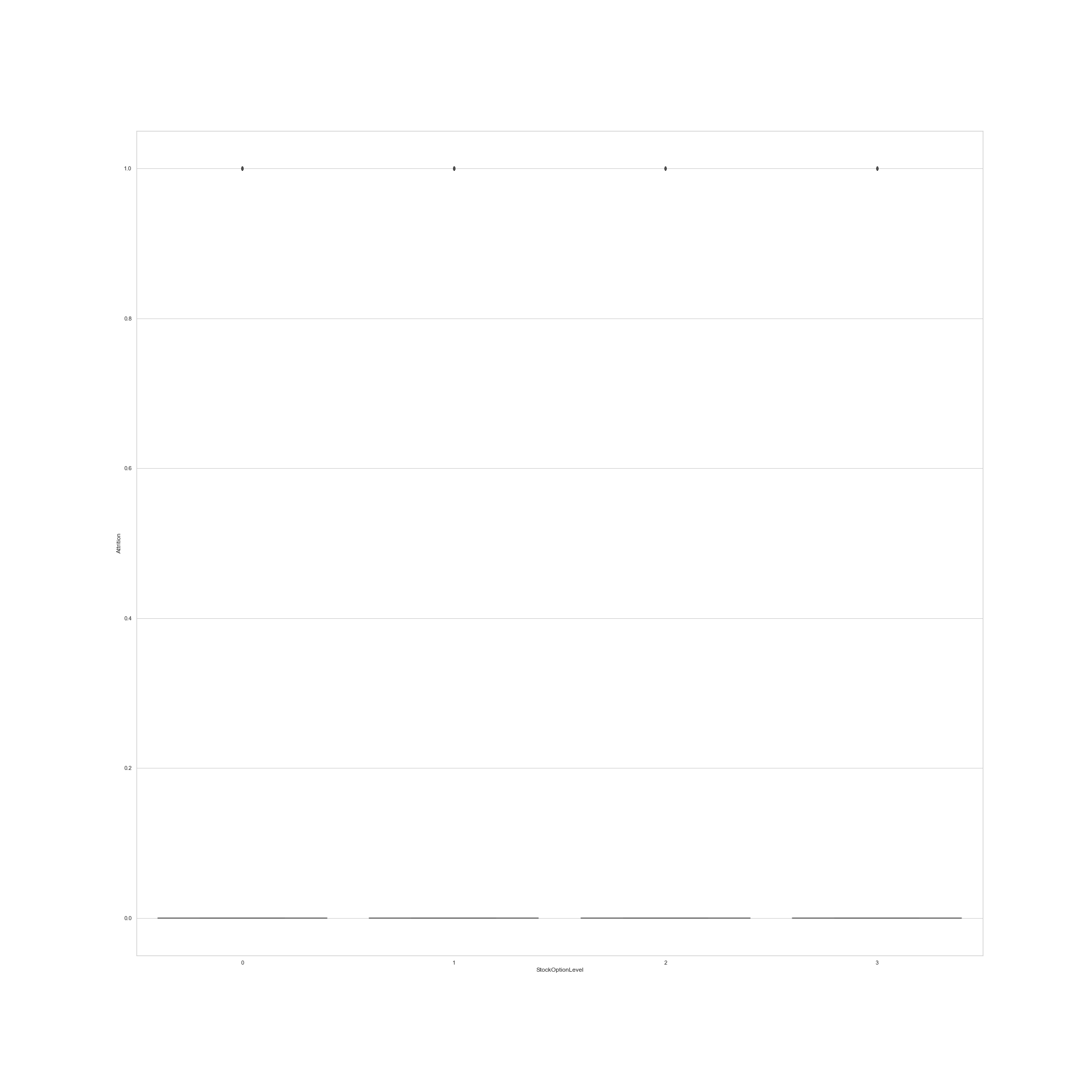
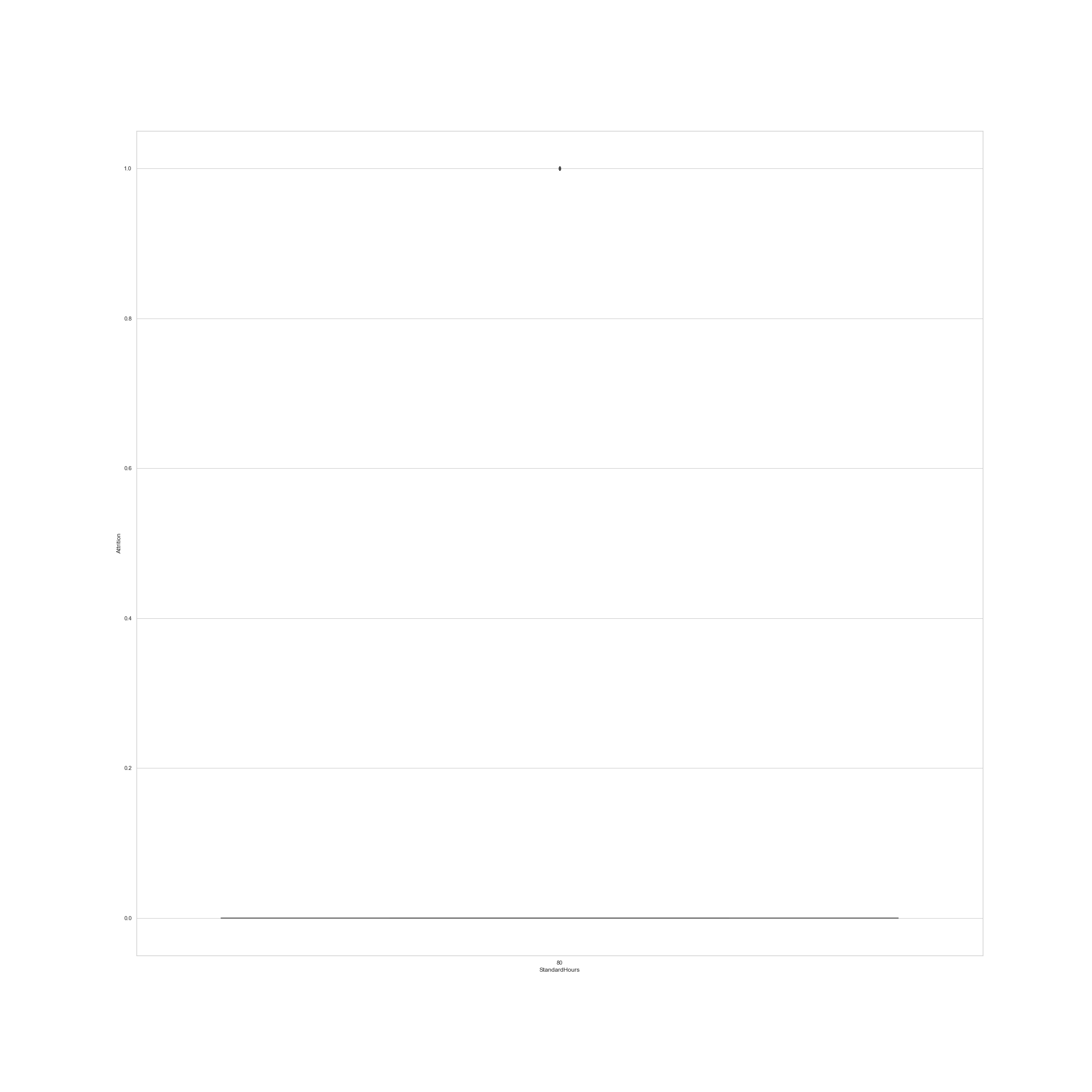
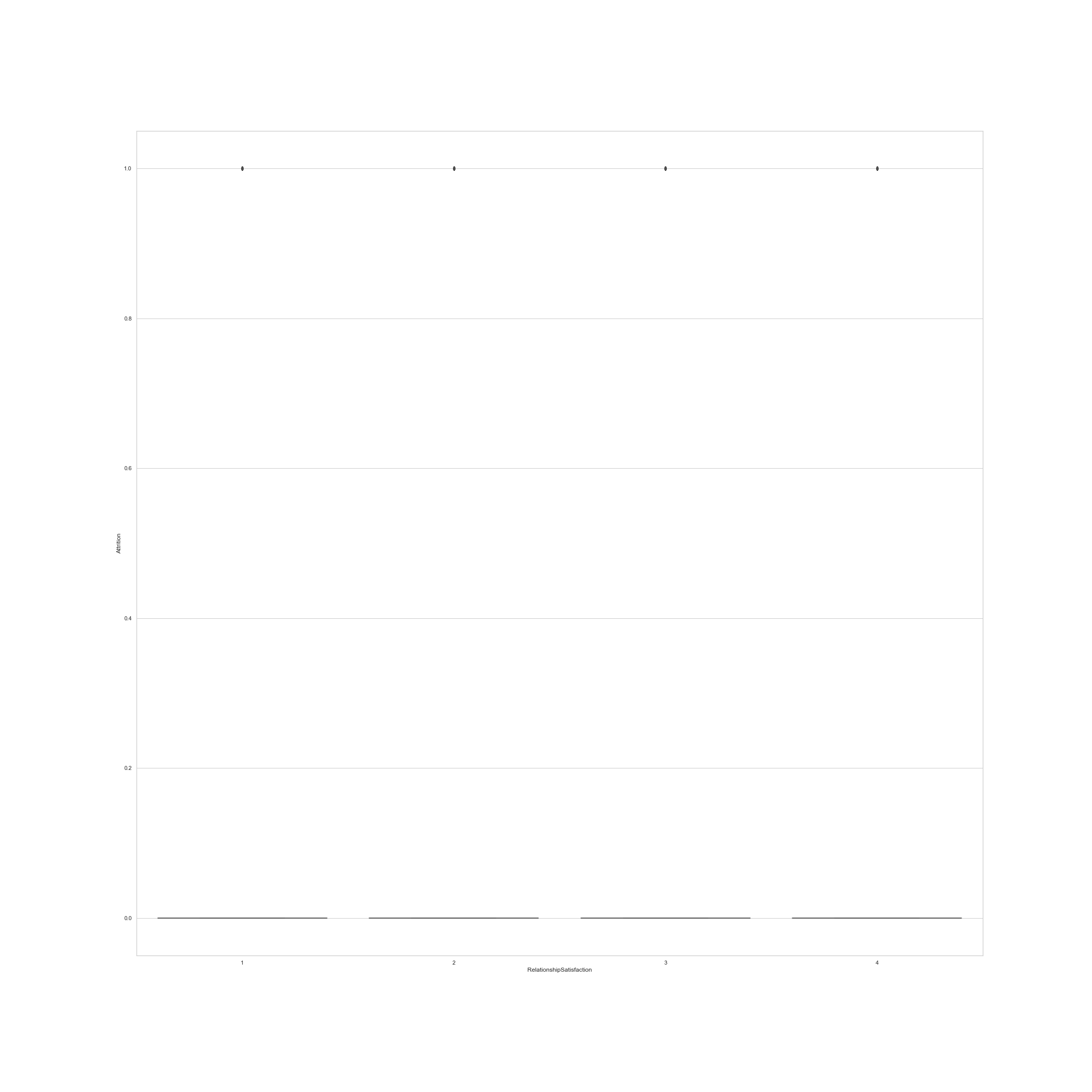
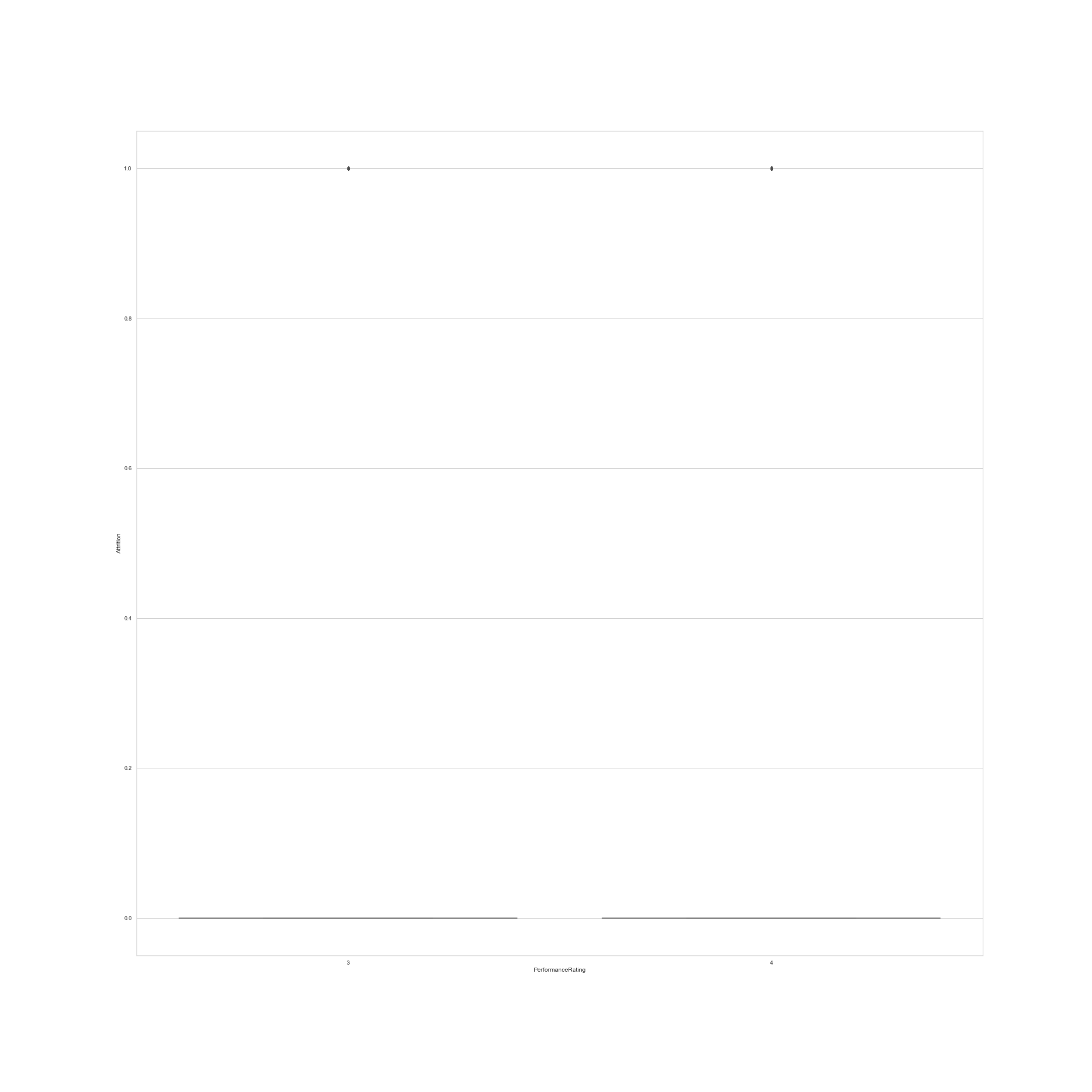
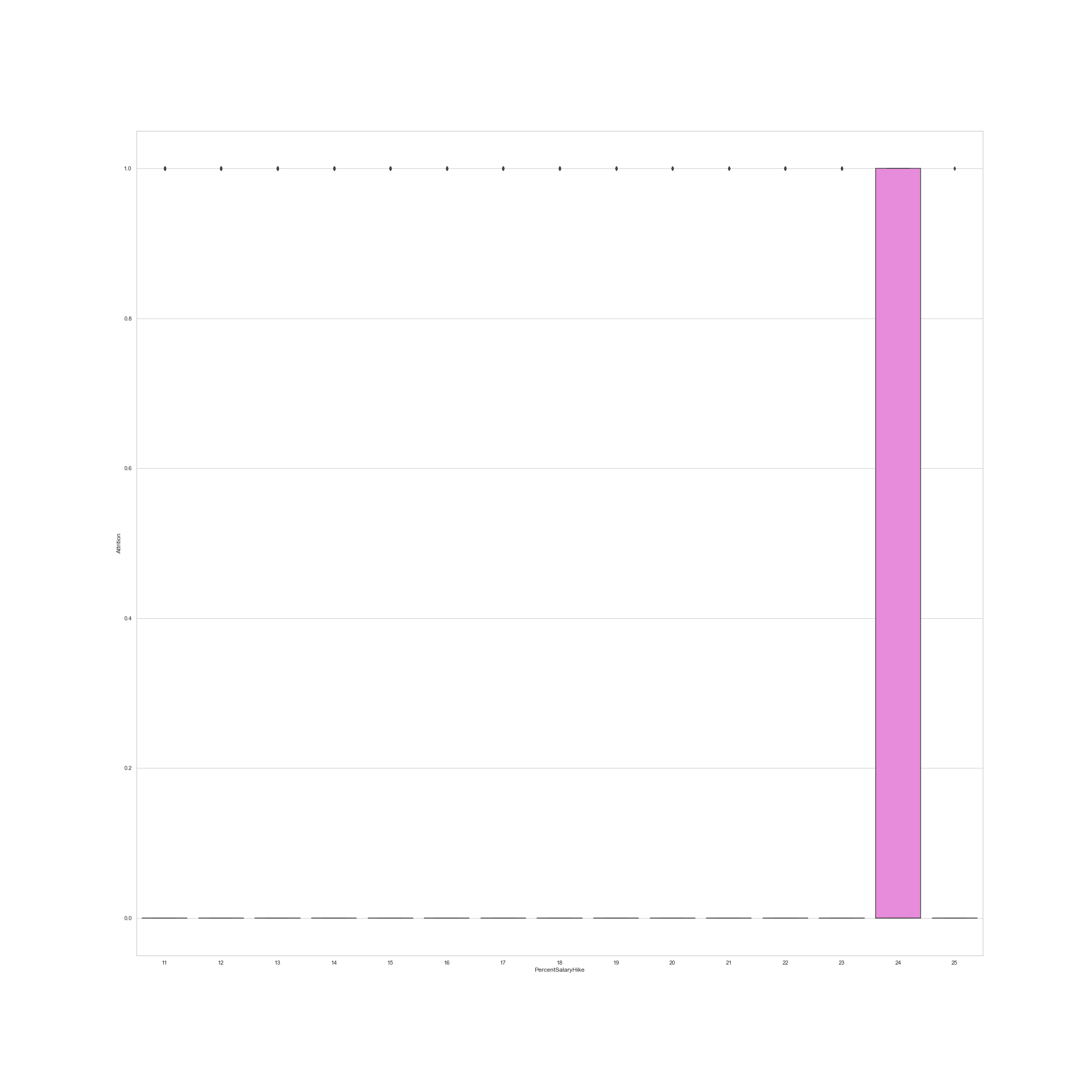
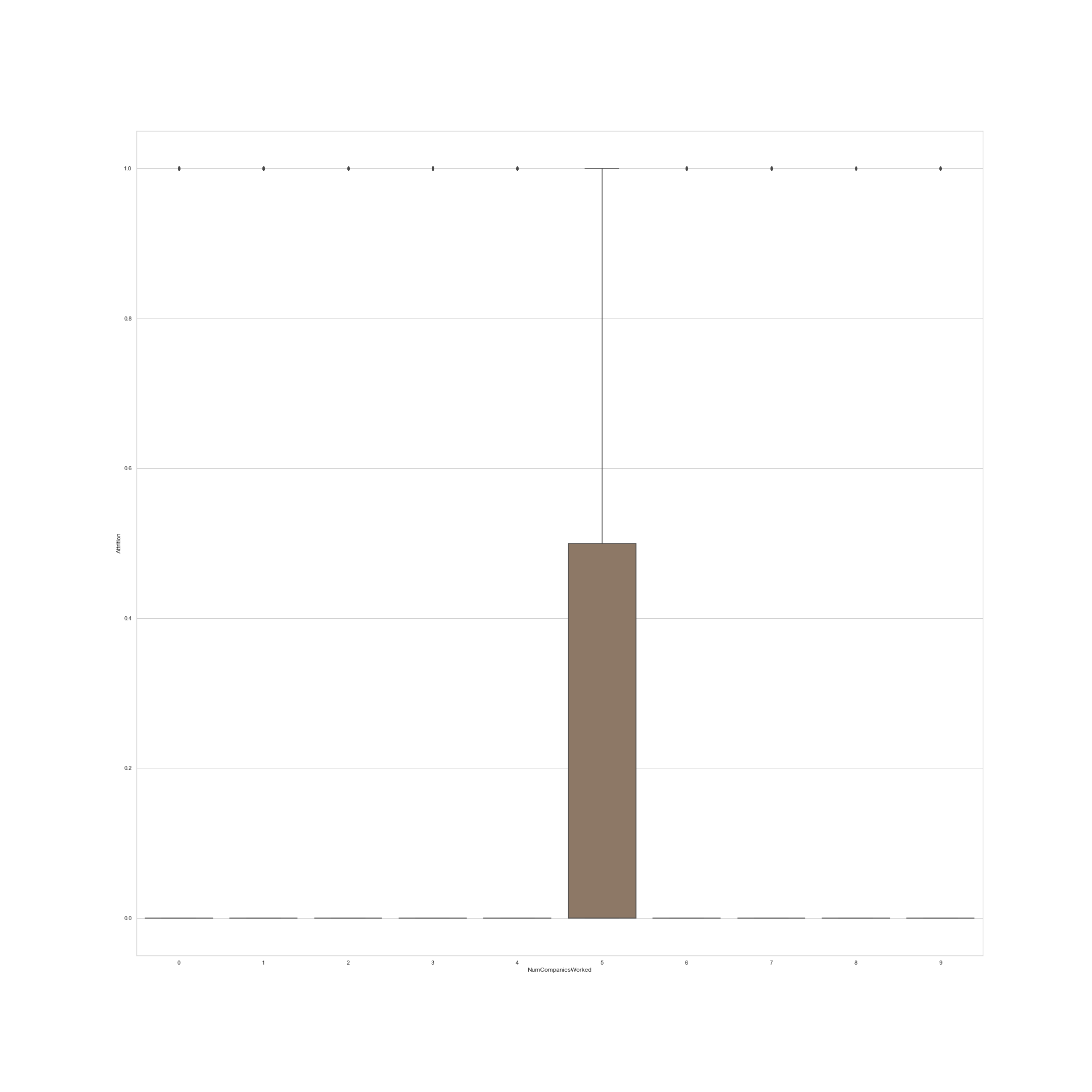
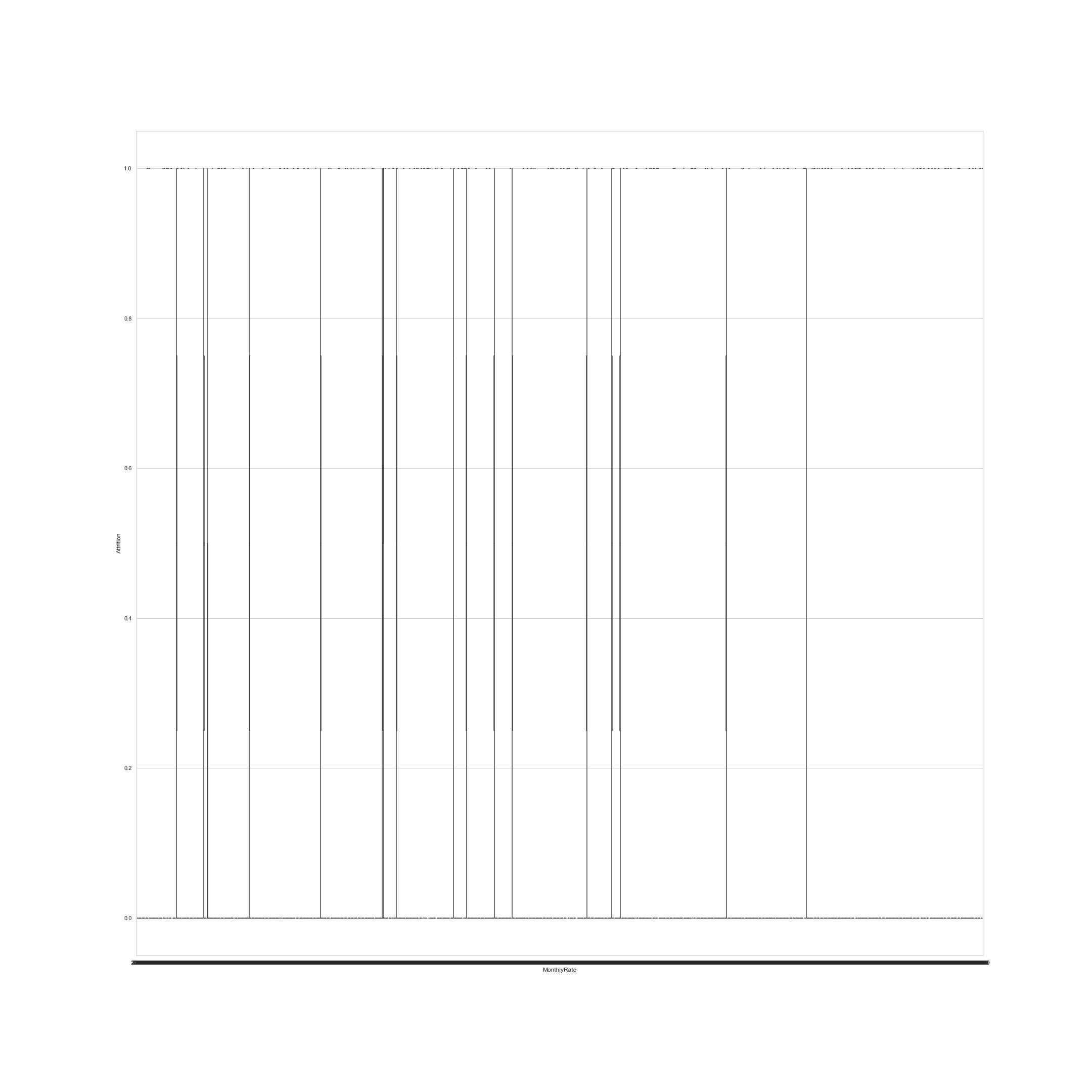
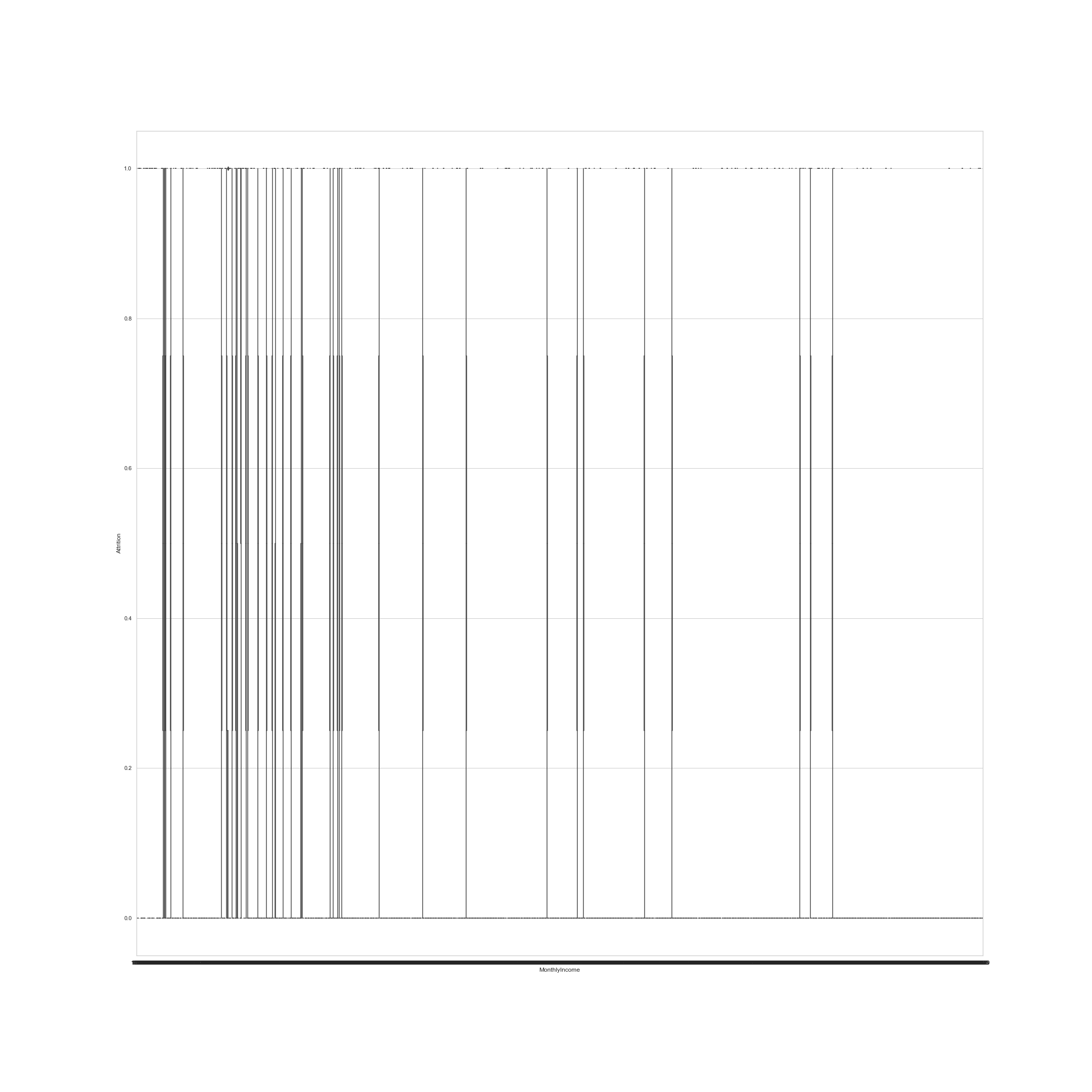
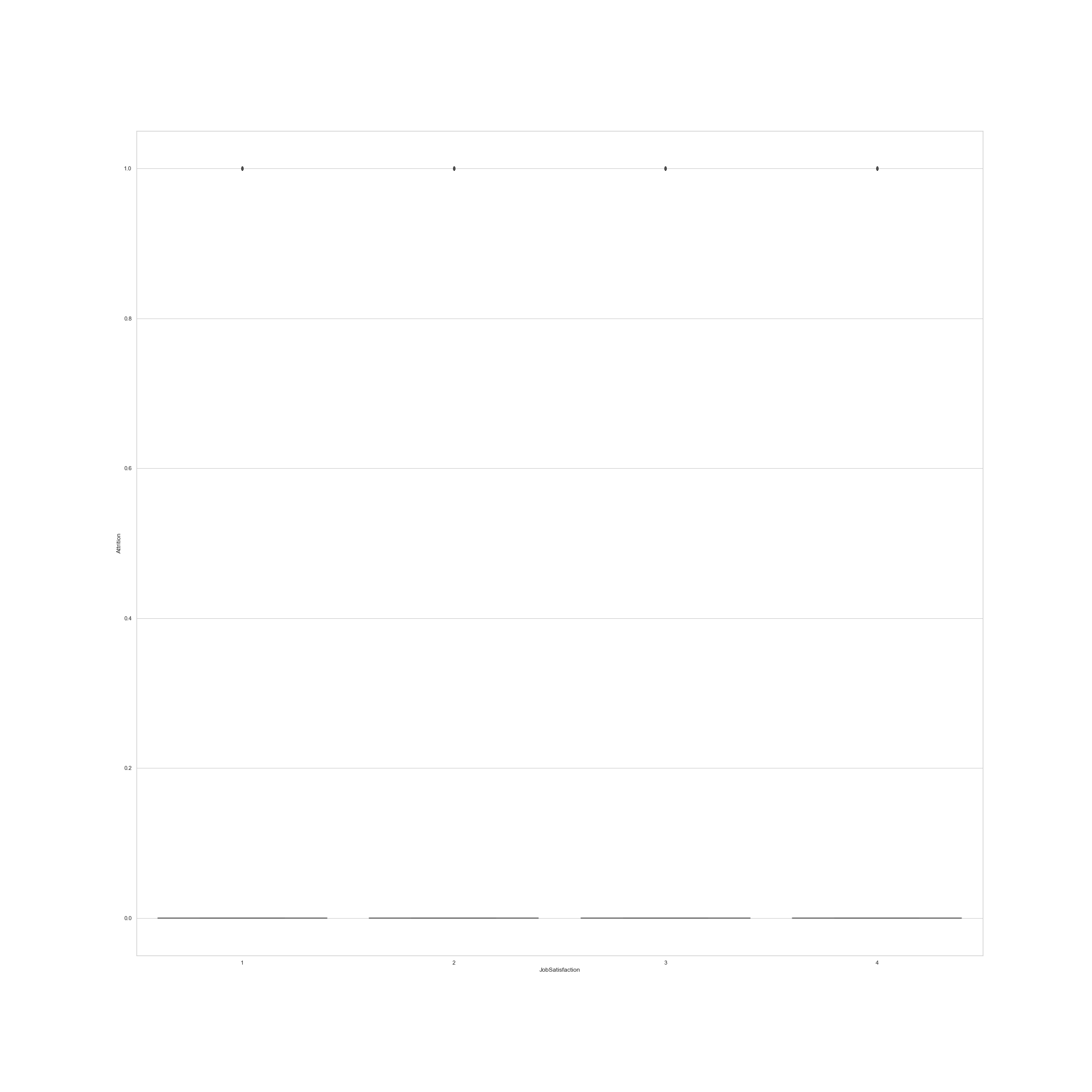
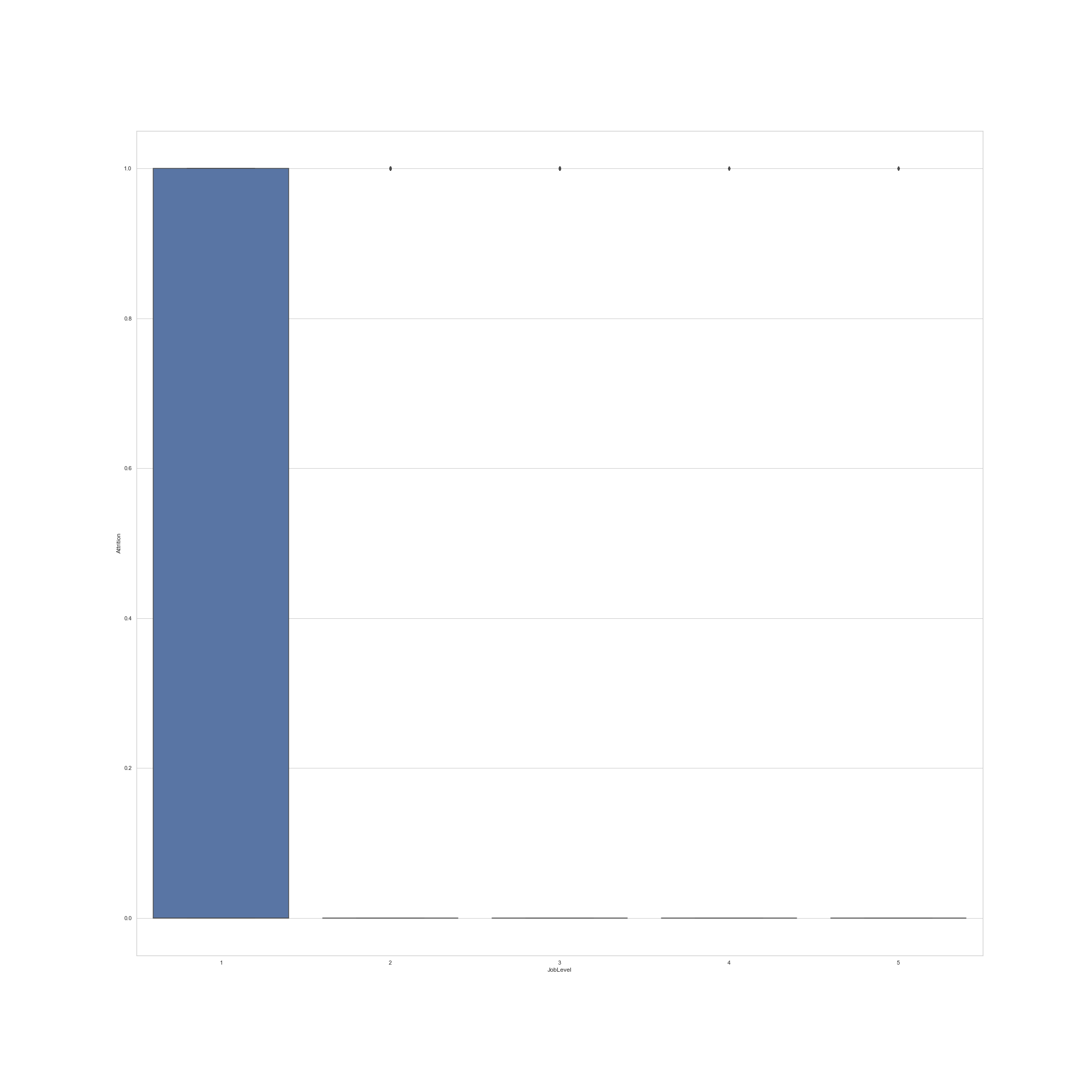
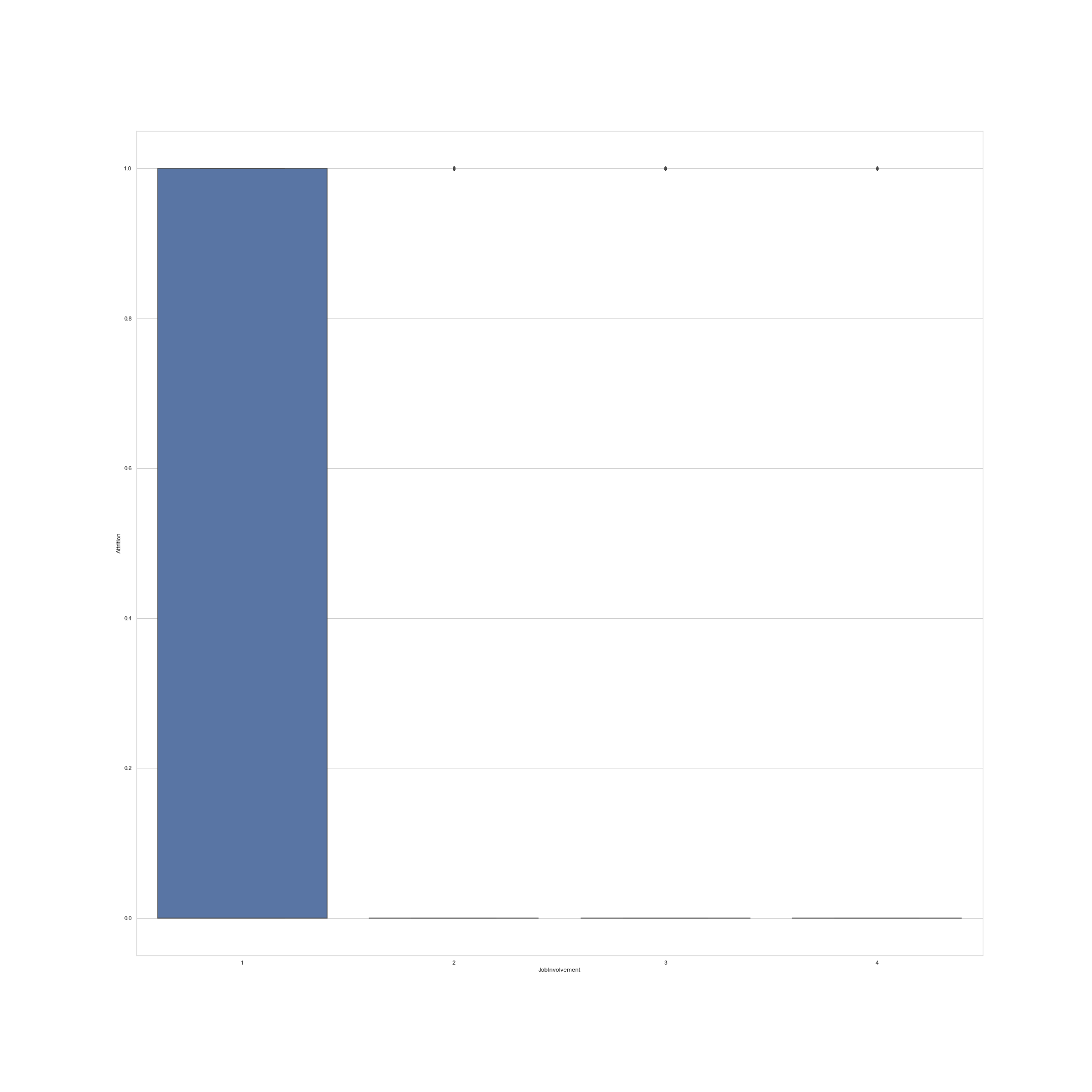
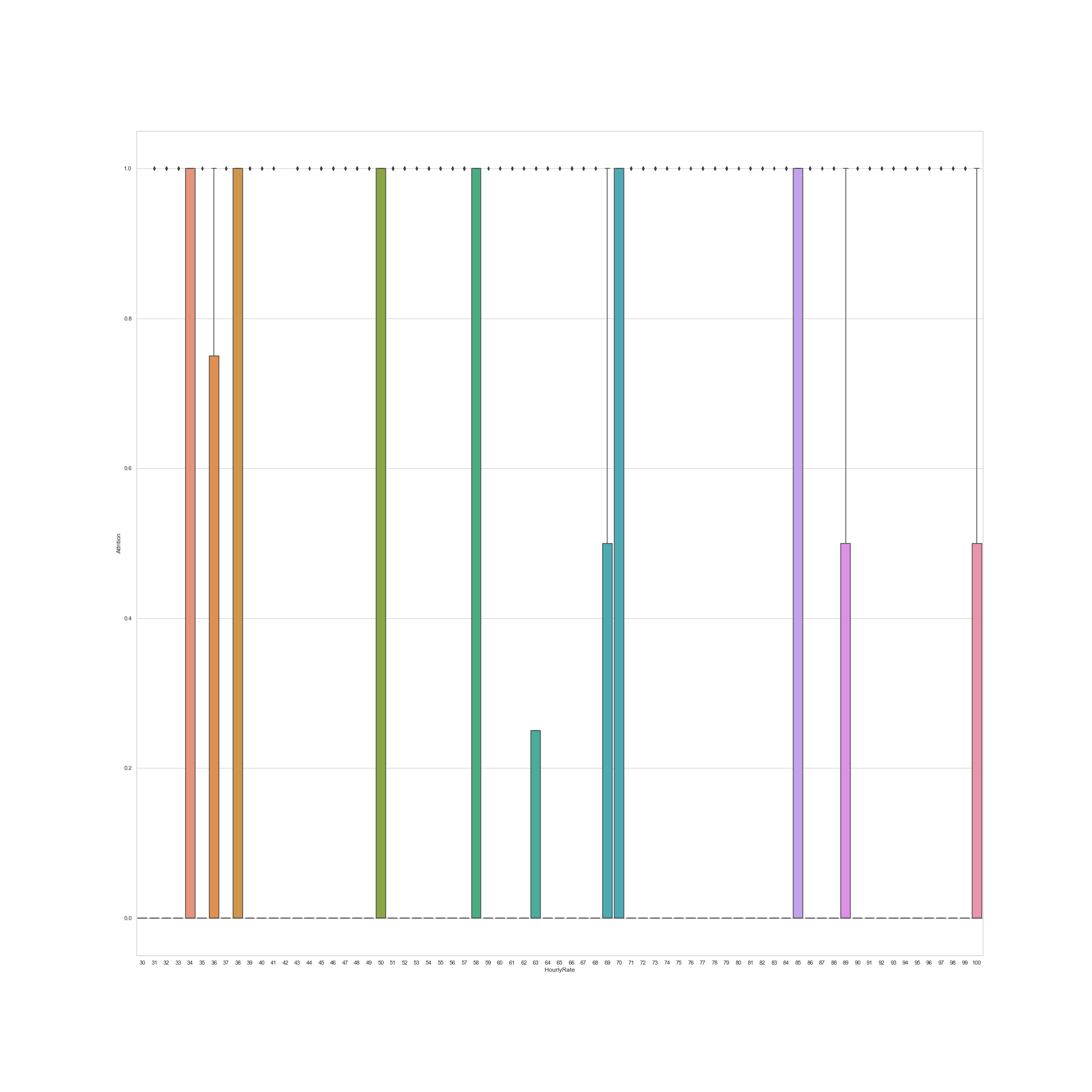
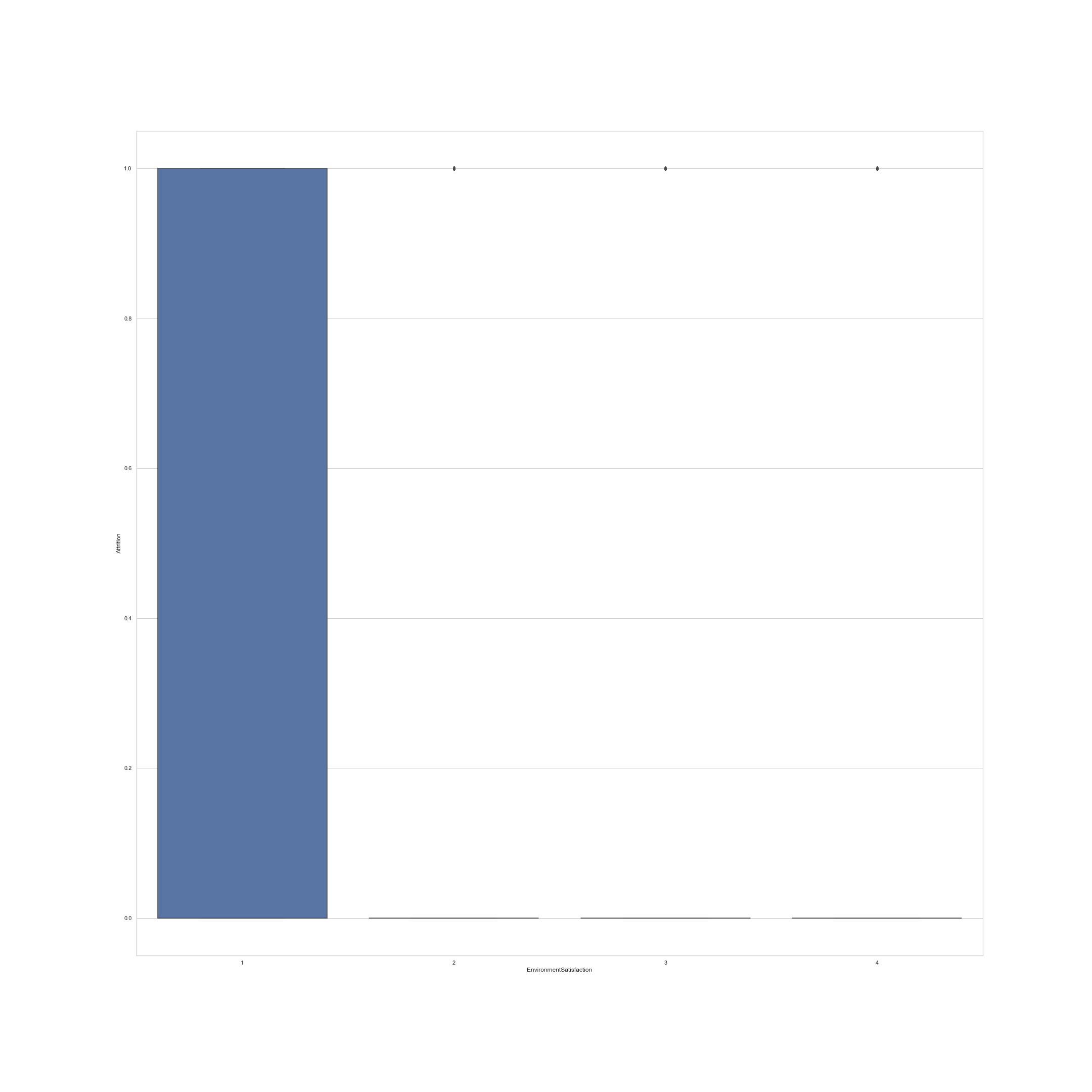
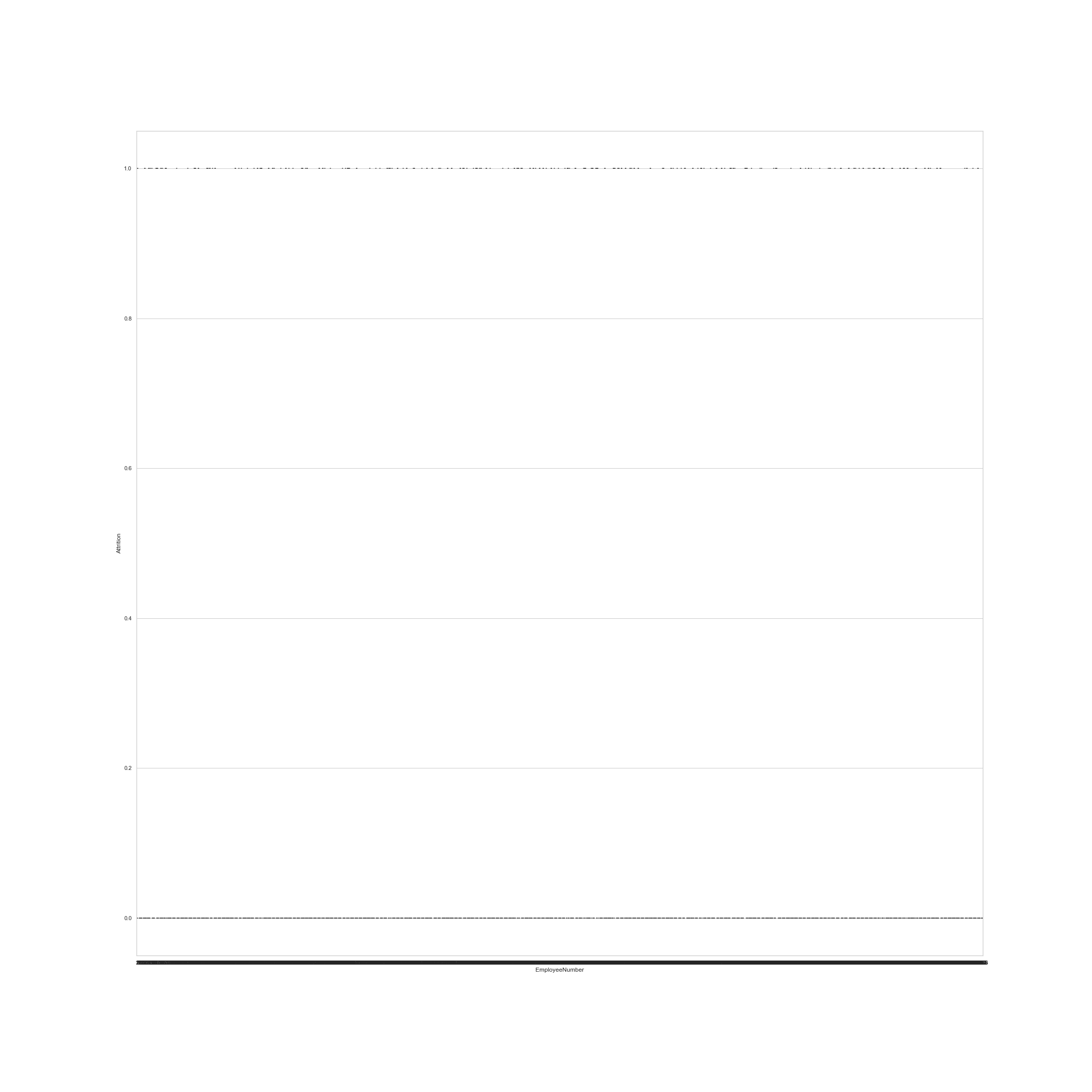
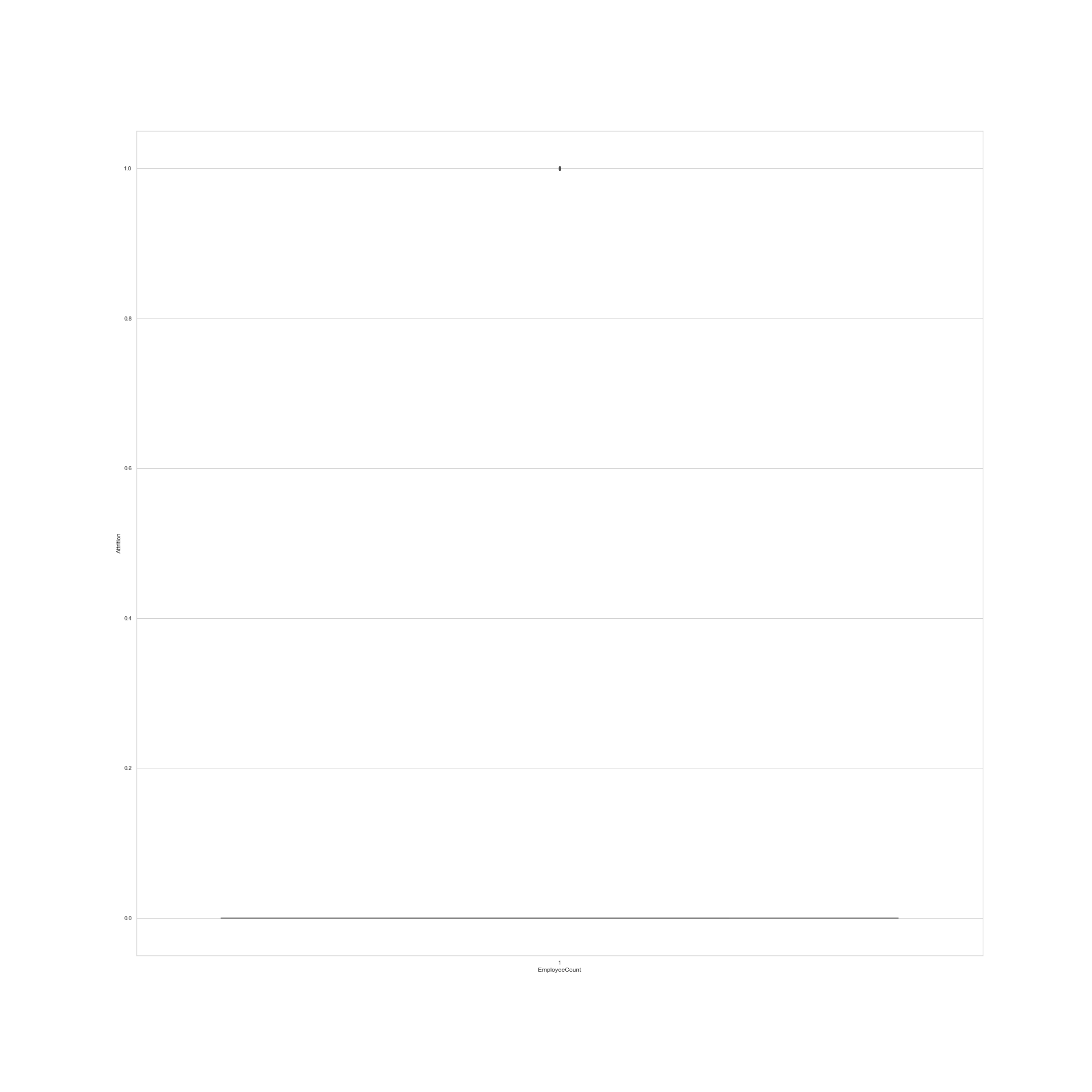
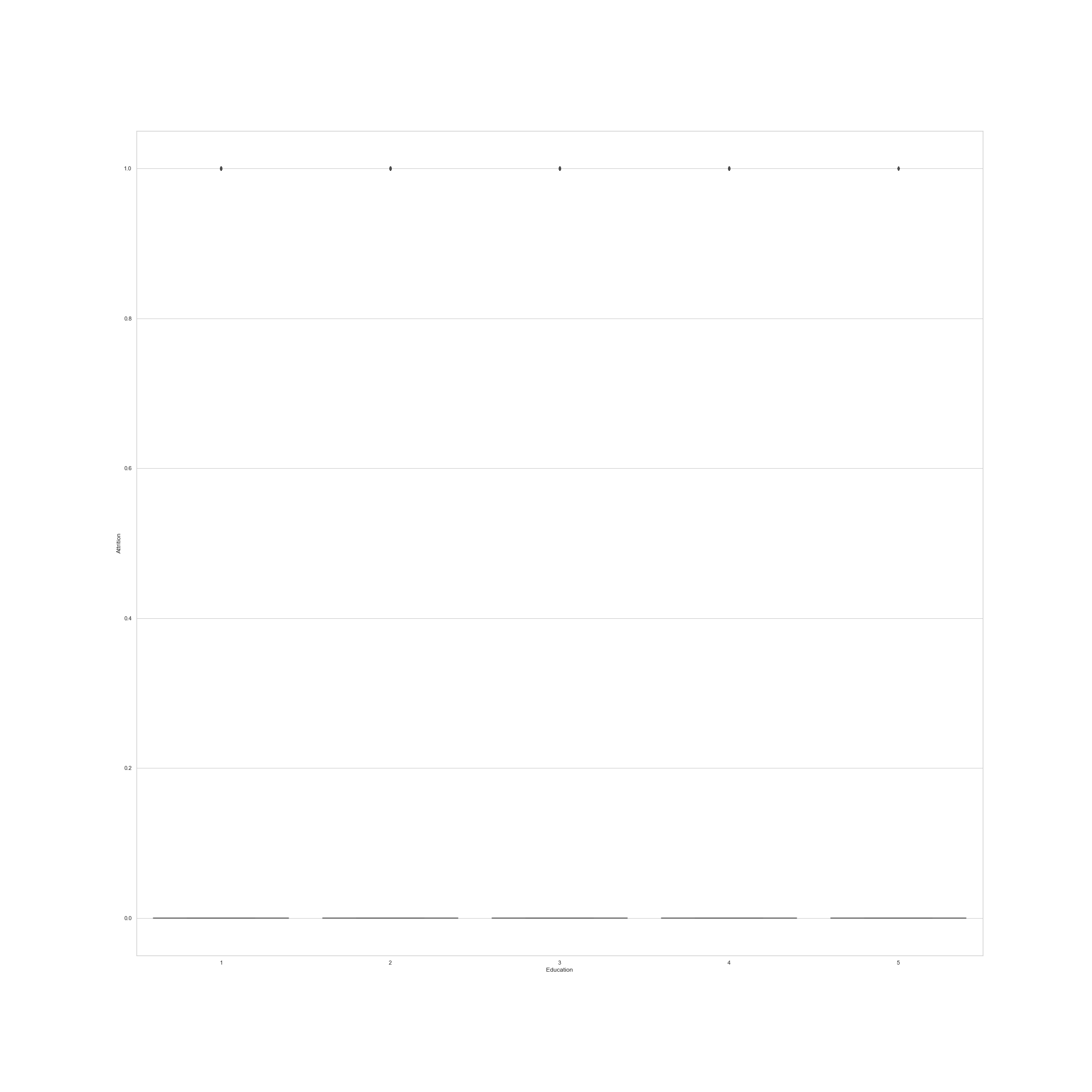
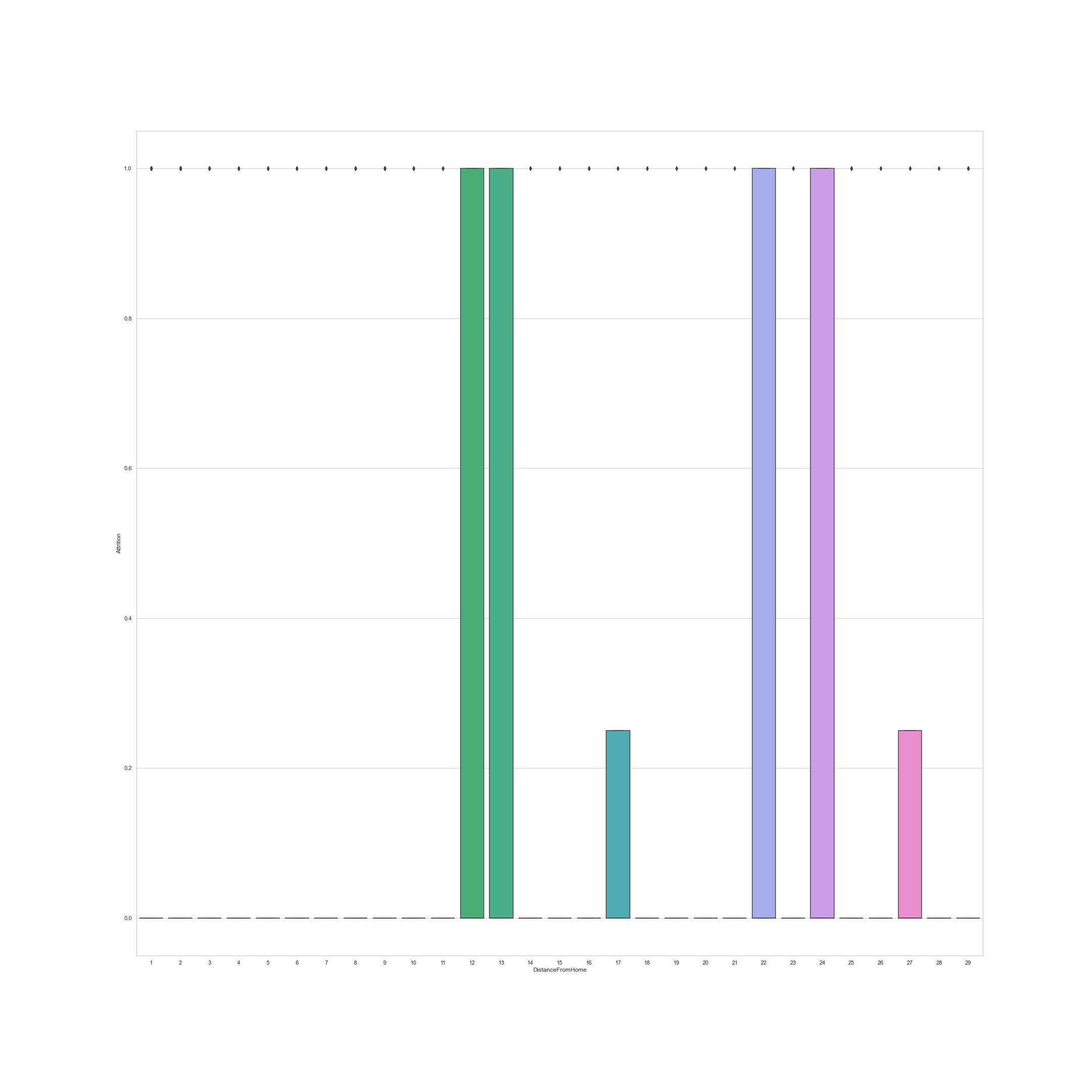
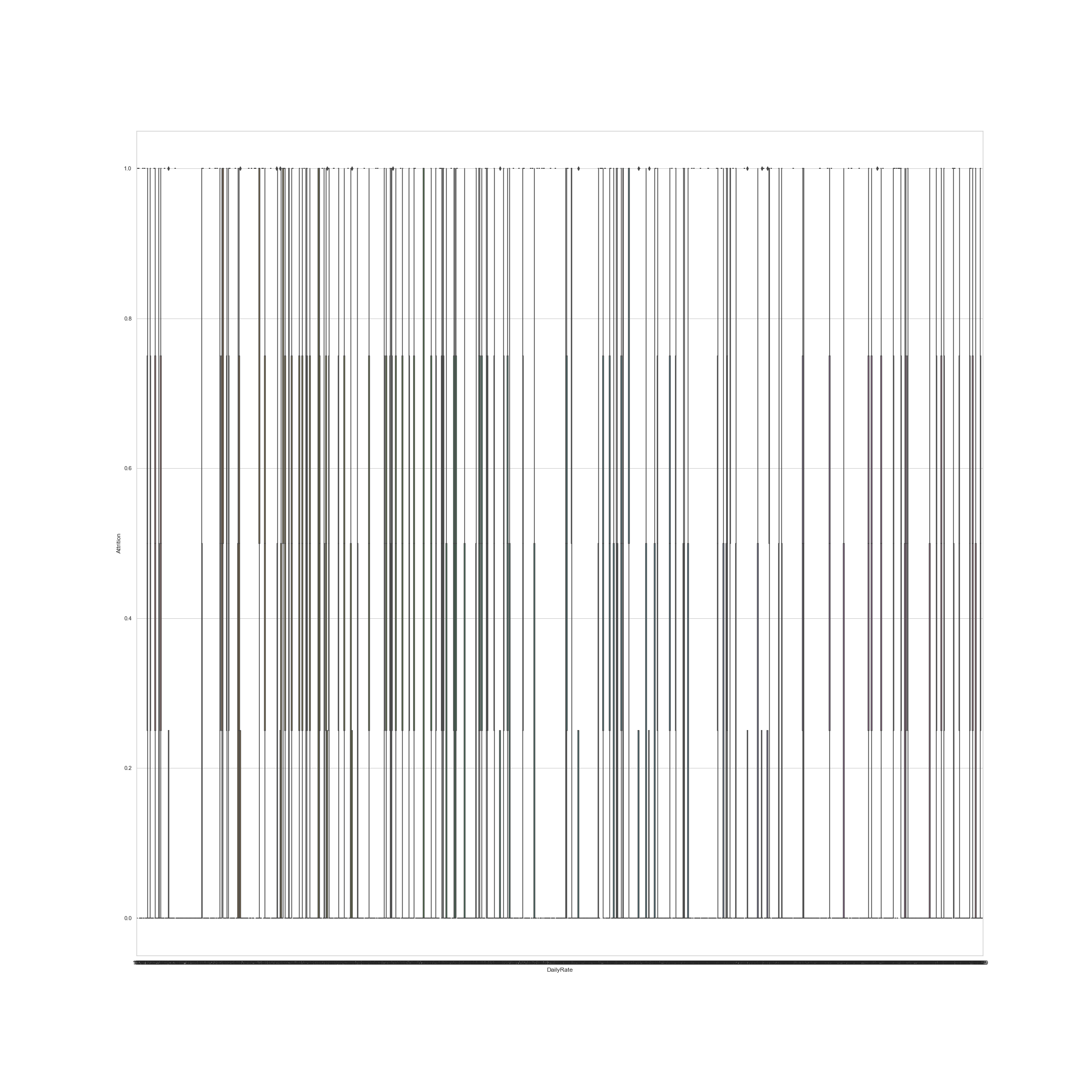
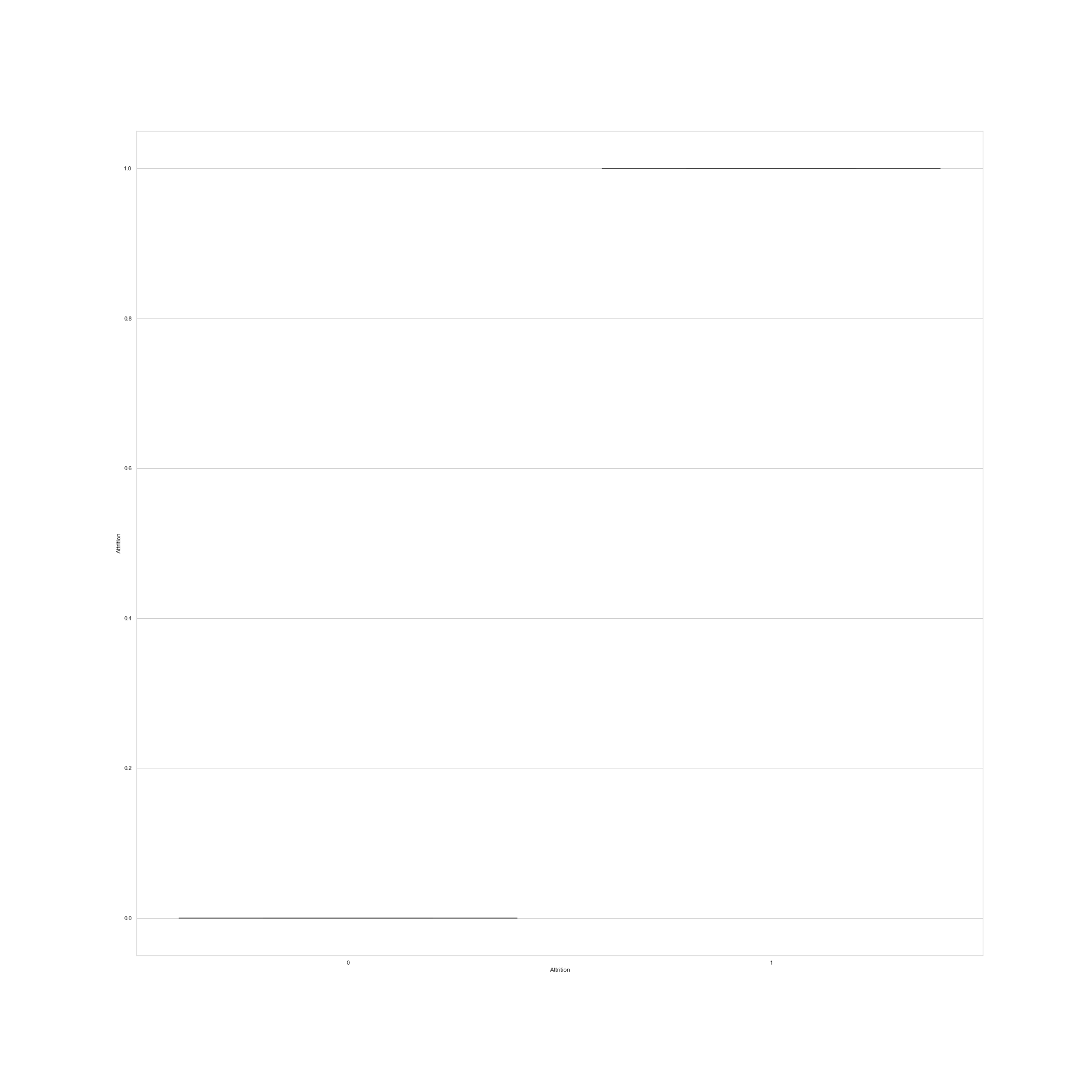
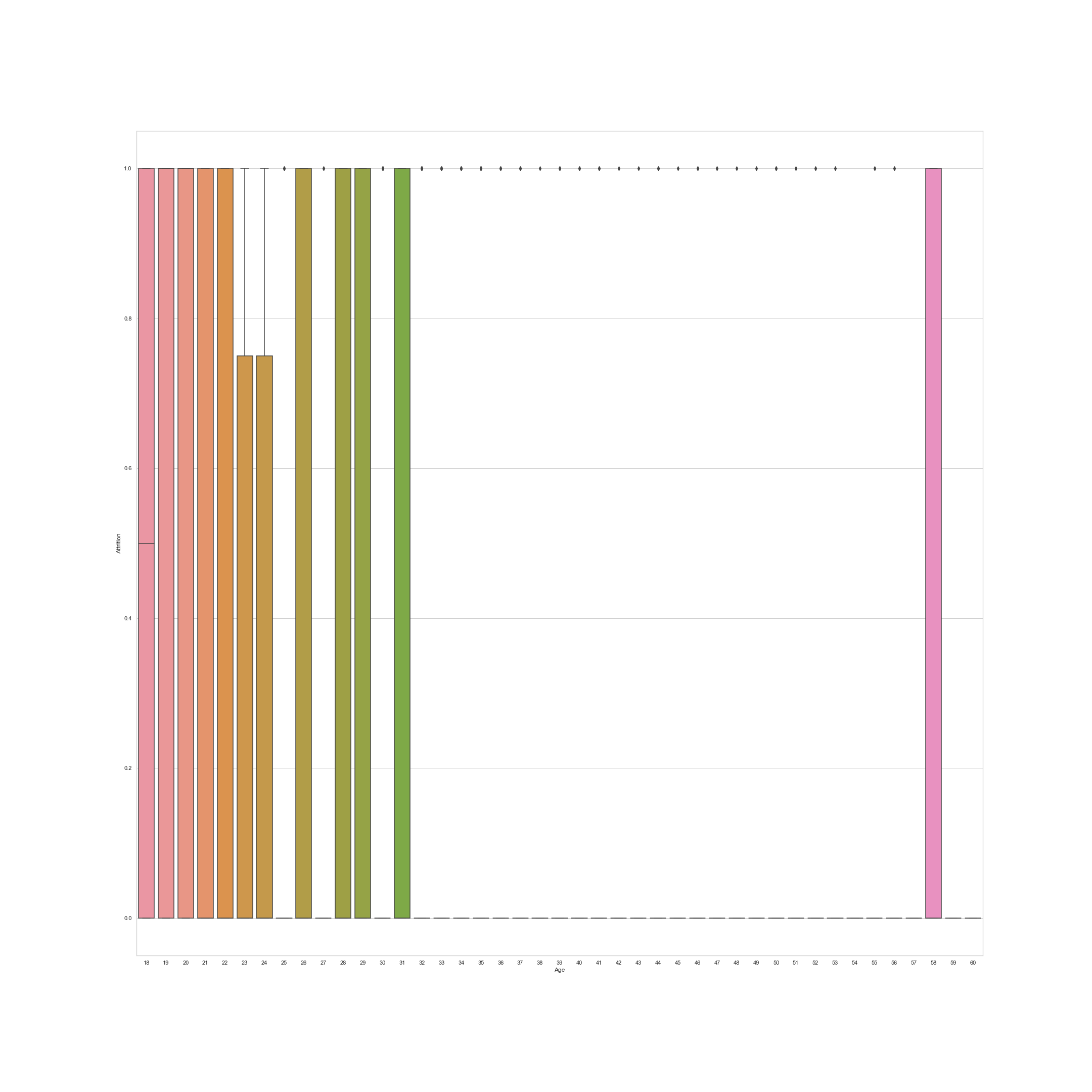
iii. Values data type for each column

iv. Count of data types in iii.

v. Memory Usage

It can be deduced that the dataset has 1470 rows and each column has no missing values. The datatypes used are float64, int32 and int64. There are 8 float64 columns, 1 int 32 column and 26 int64 columns. Memory usage is 396.3 KB.

Analysis 8: Boxplot



Observations:

1. The data has many outliers in most of the columns, like, Years since last promotion, etcetera.

2. A few outliers are detected in:

i. Hourly Rate

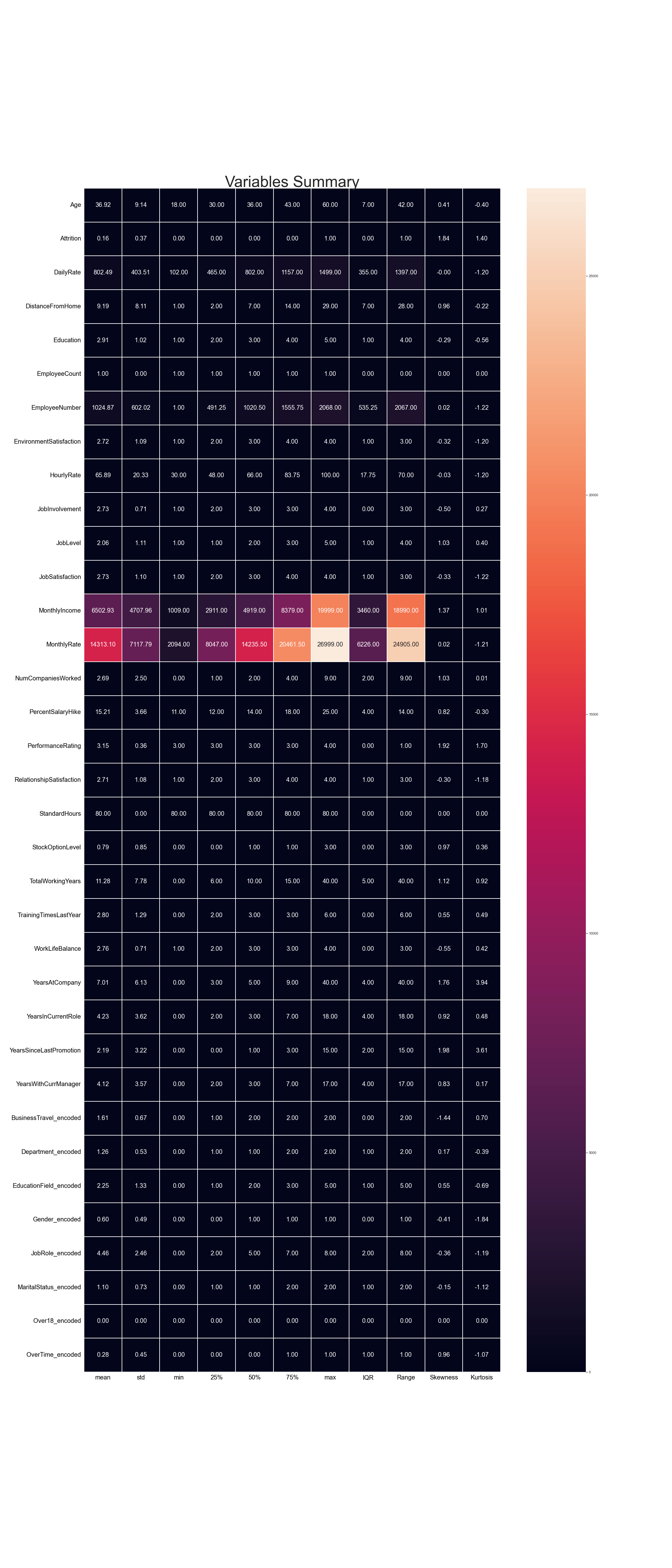
ii. Years at Company

iii. Education Field Encoded

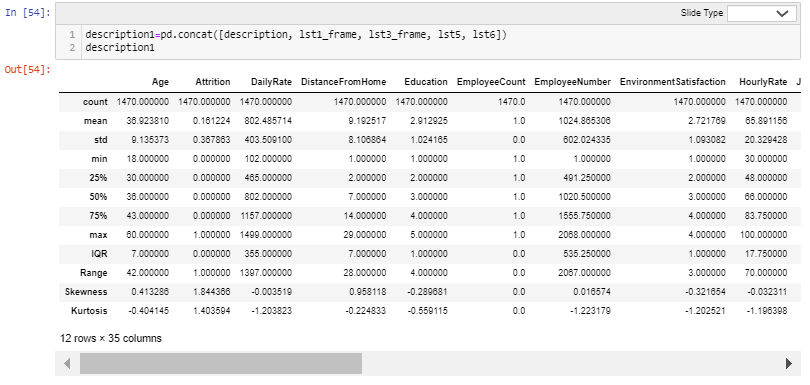
Solution:

**Data Pre-Processing is required to reduce the number of outliers in the data.**

Analysis 9: Descriptive Statistics



*Snippet From Notebook*



# Verbal Translation of Descriptive Table and Graphic

1. For values that are scaled up to 1, the mean is mostly around 0 and standard deviation is comparatively low. Hence, making the data much more acceptable by algorithms to process it more accurately.

2. The entire dataset ranges from 0 till 19999.

3. IQR, Range and Skewness and kurtosis are much more condensed in data that is scaled upto 1.

4. Skewness is 0 and within +/- 0.65 for:

WorkLifeBalance

JobInvolvement

Gender\_encoded

JobRole\_encoded

JobSatisfaction

EnvironmentSatisfaction

RelationshipSatisfaction

Education

MaritalStatus\_encoded

HourlyRate

DailyRate

EmployeeCount

StandardHours

Over18\_encoded

EmployeeNumber

MonthlyRate

Department\_encoded

Age

EducationField\_encoded

TrainingTimesLastYear

5. Acceptable skewness is +/- 0.65 and skewness for bell shaped curve should be 0.

6. Kutosis is upto 3 for most dataset, indicating platykurtic curves.

7. Kurtosis is greater than 3 obly for Years Since Last Promotion, indicating, leptokurtic curve.

8. Kurtosis for bell shaped curve should be 3.

Mathematical Notation:

1. Mean = sum of values/count of values

2. std = sqrt (((value - mean of distribution) \*\*2 / number of values))

3. 3 quartiles are measures of variance, calculated to spot the placeholder value, it returns index of the produced value.

Step 1: sort the dataset

Step2:

i) Lower Quartile (Q1: 25% distribution) = ((number of values+1)/4)th Term

ii) Middle Quartile (Q2: 50% distribution) = ((number of values +1)/2)th Term

Also, known as median (central value).

iii) Upper Quartile (Q3: 75% distribution) = ¾ (number of values + 1)th Term

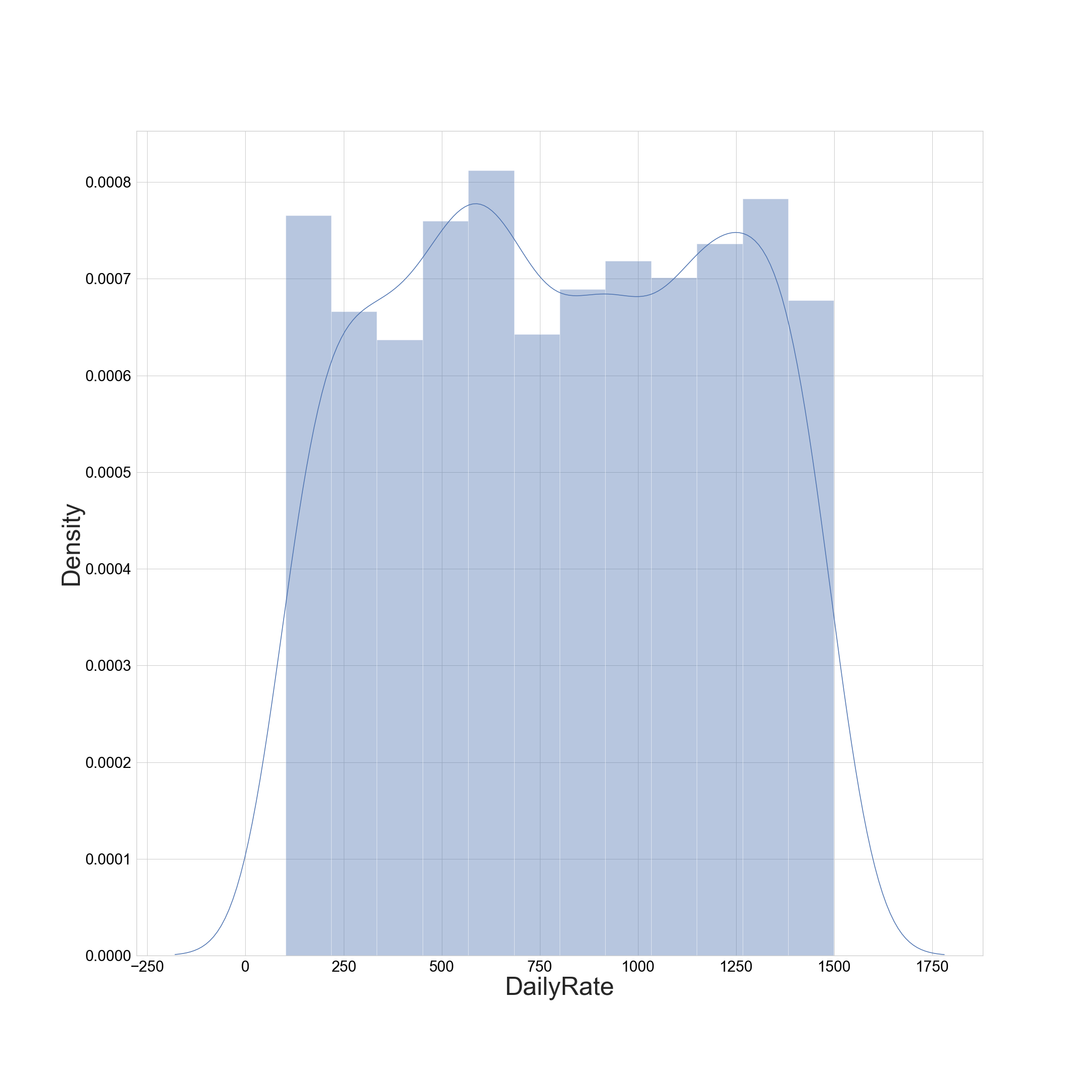
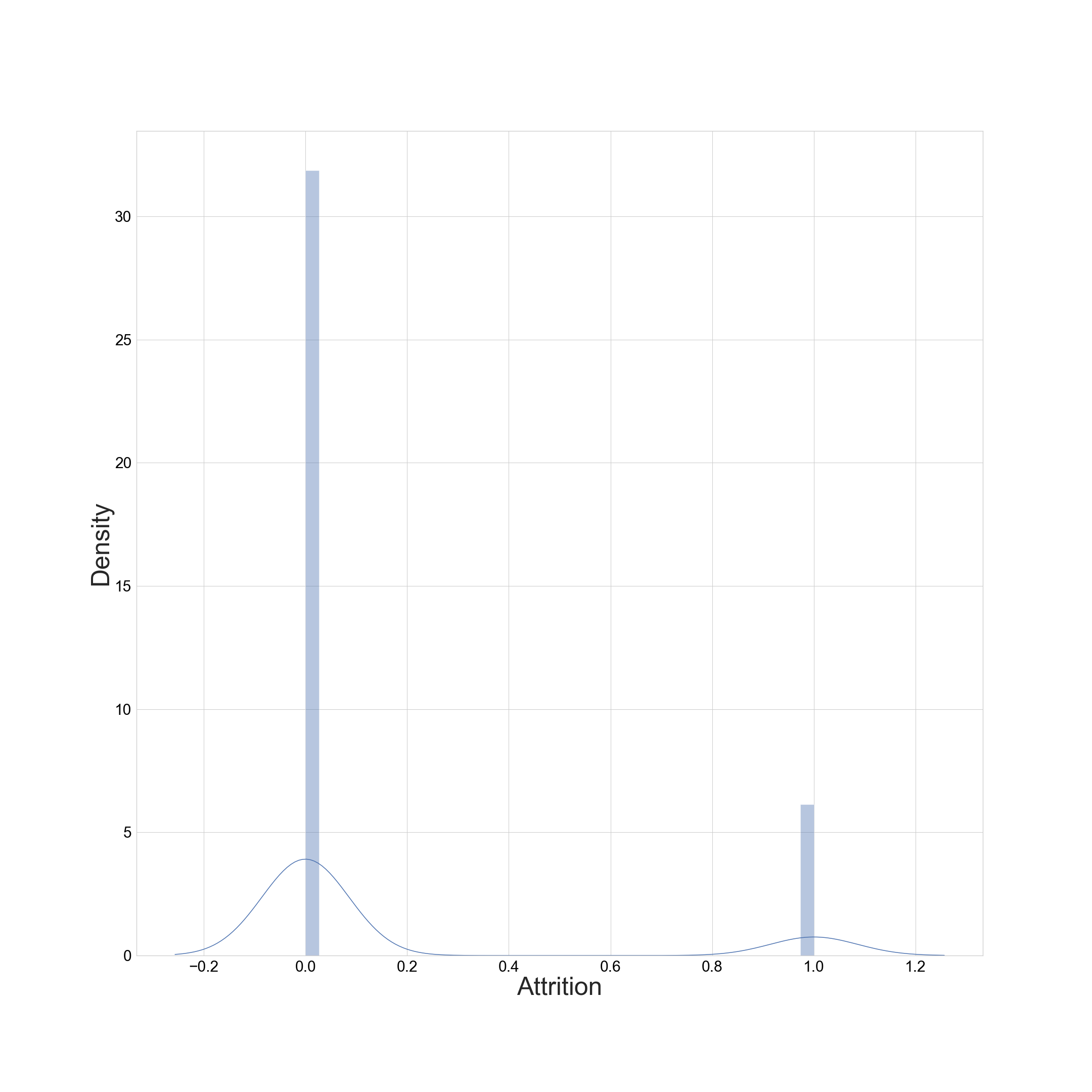
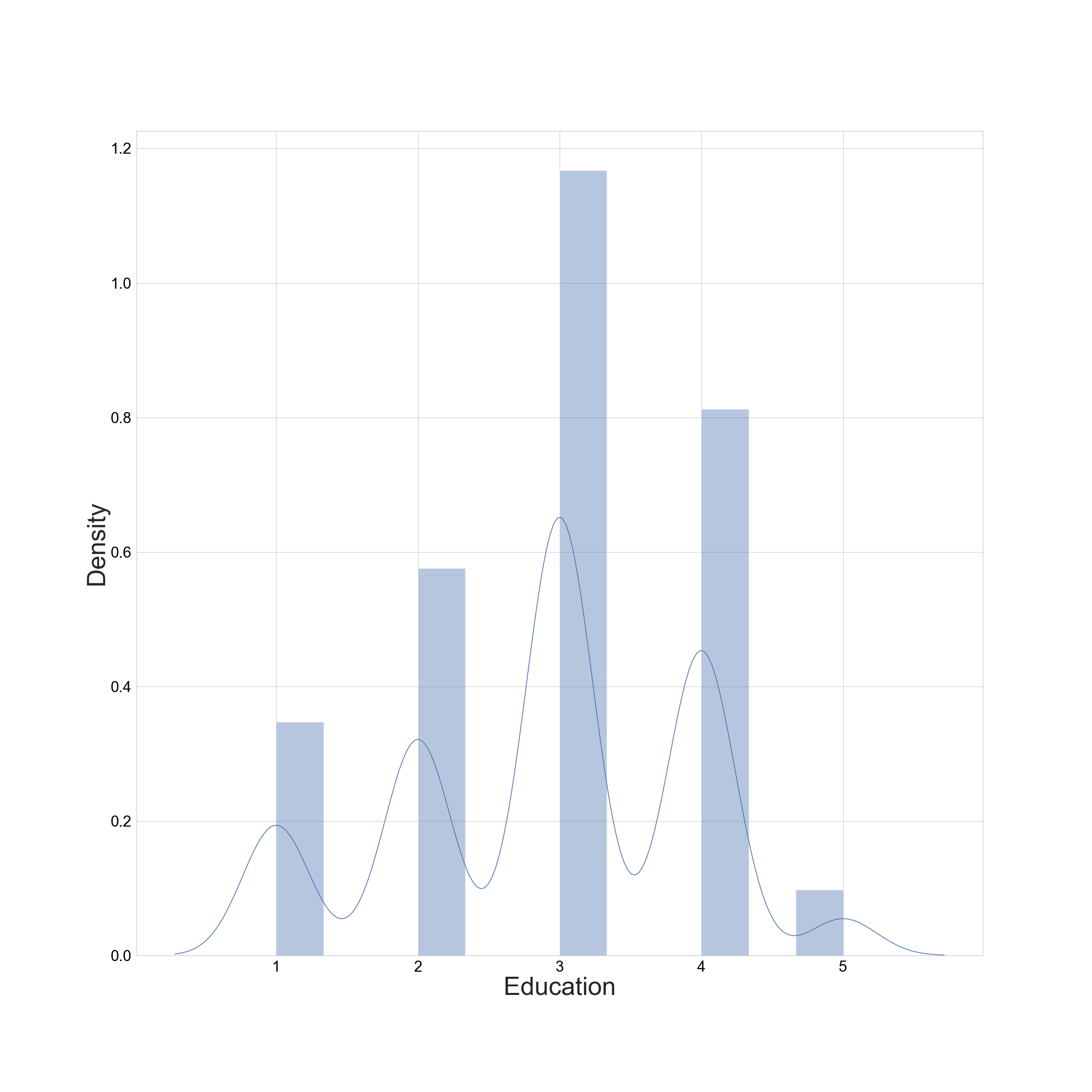
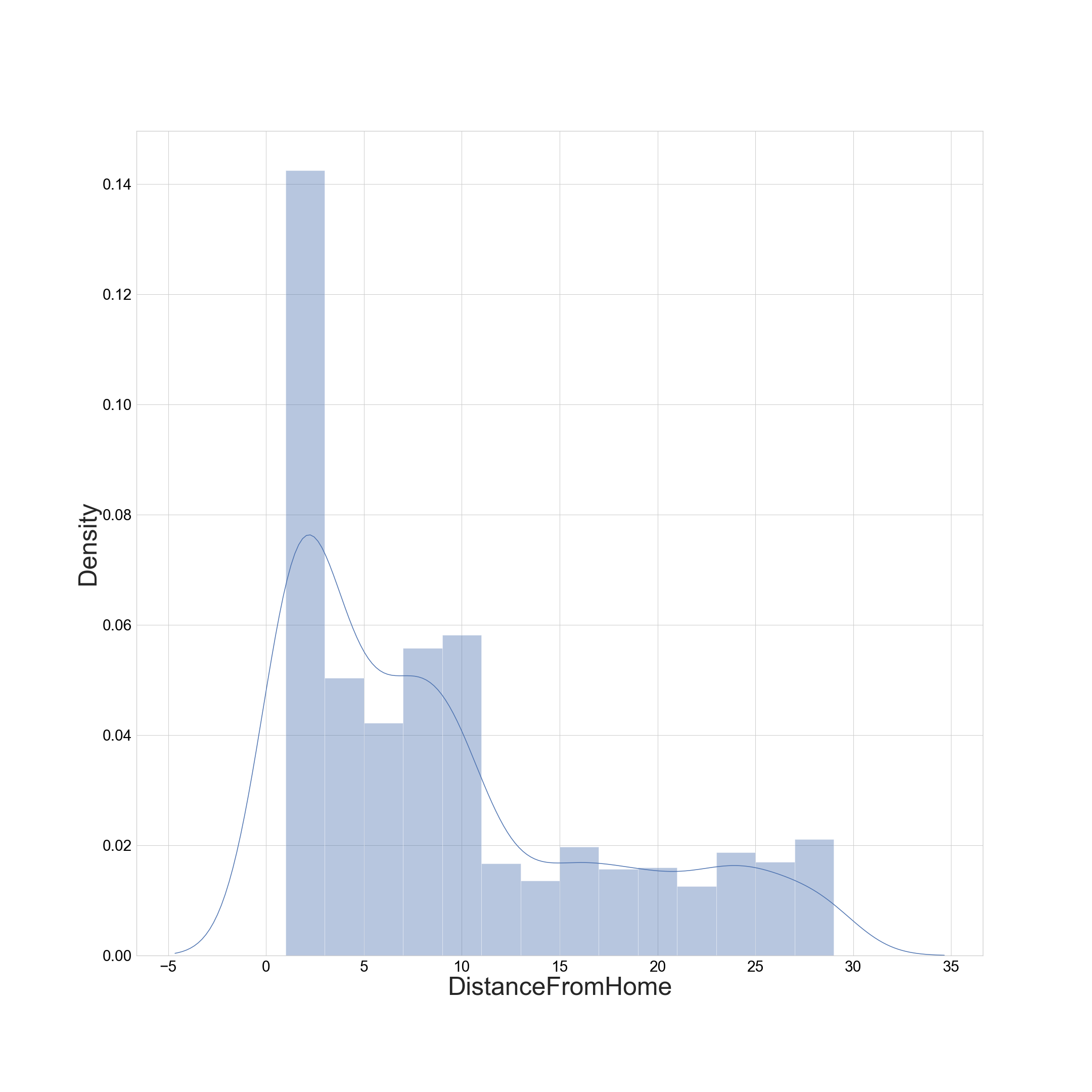
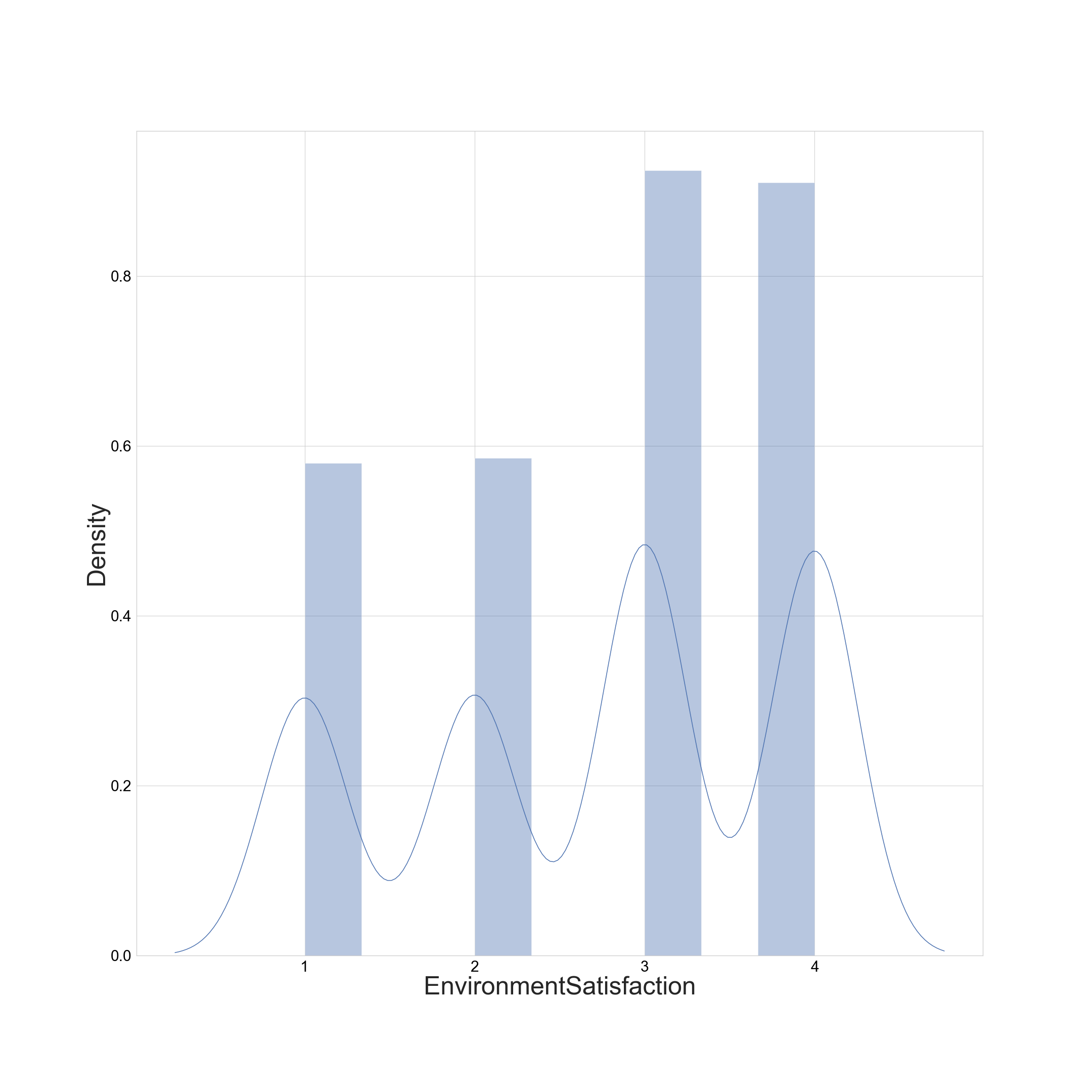
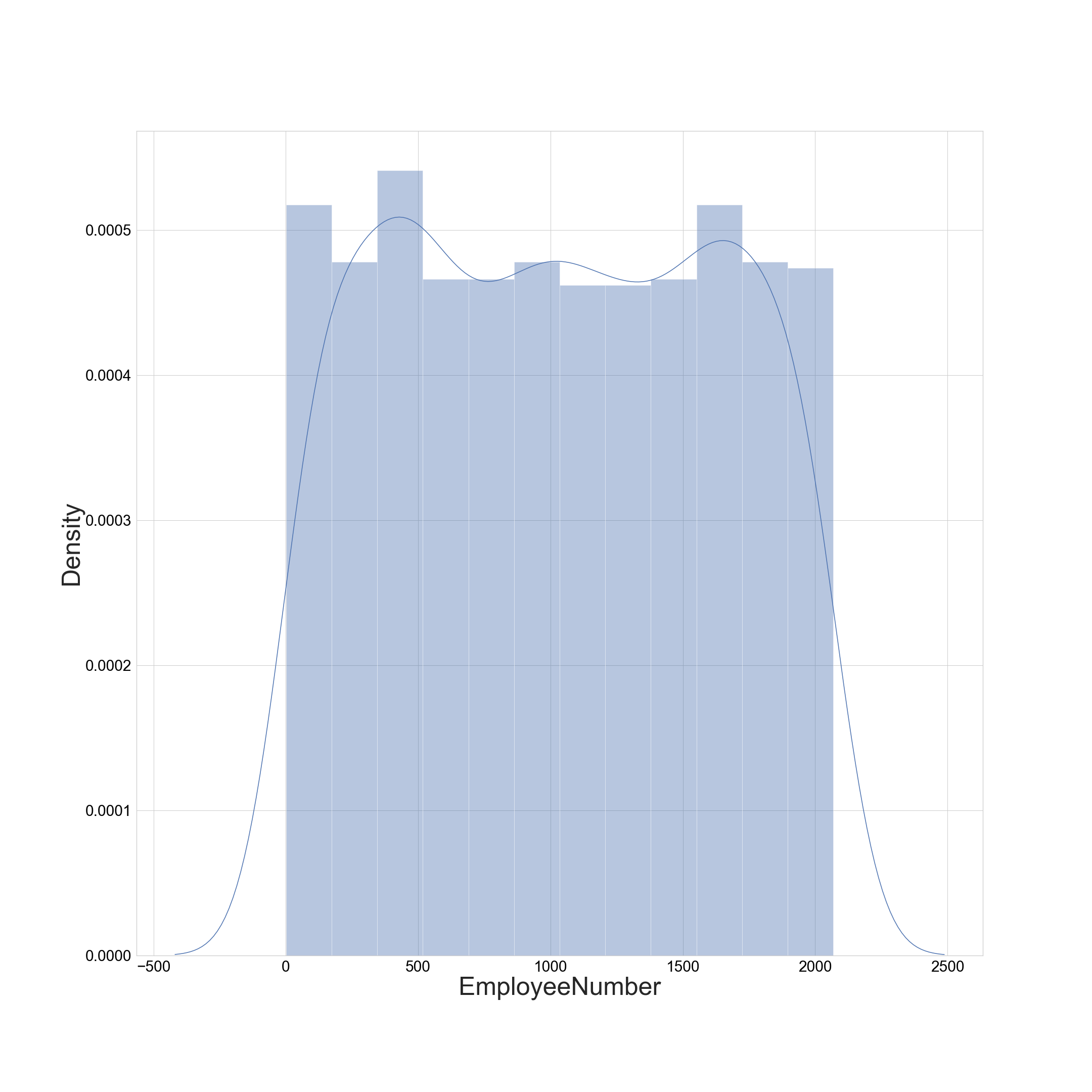
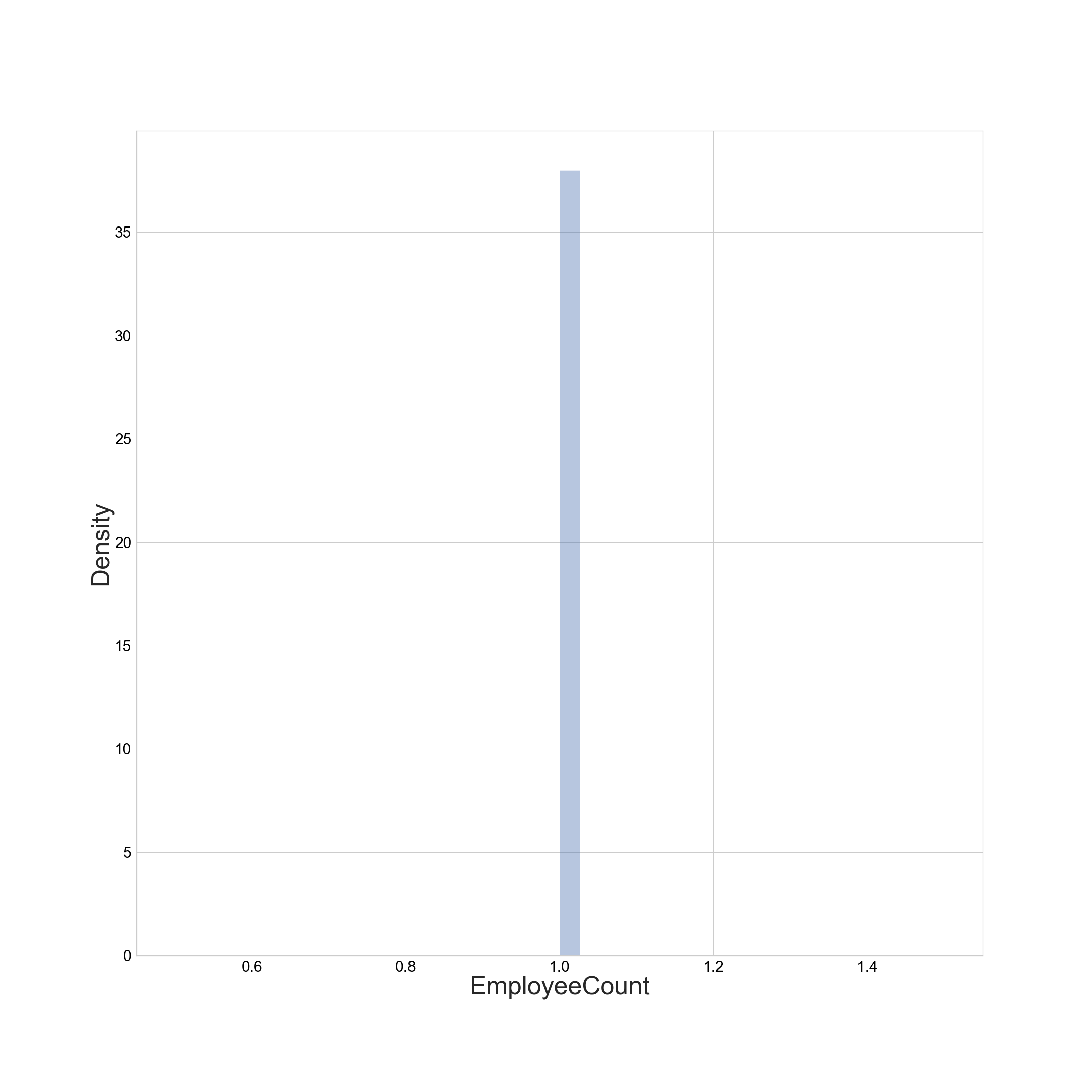
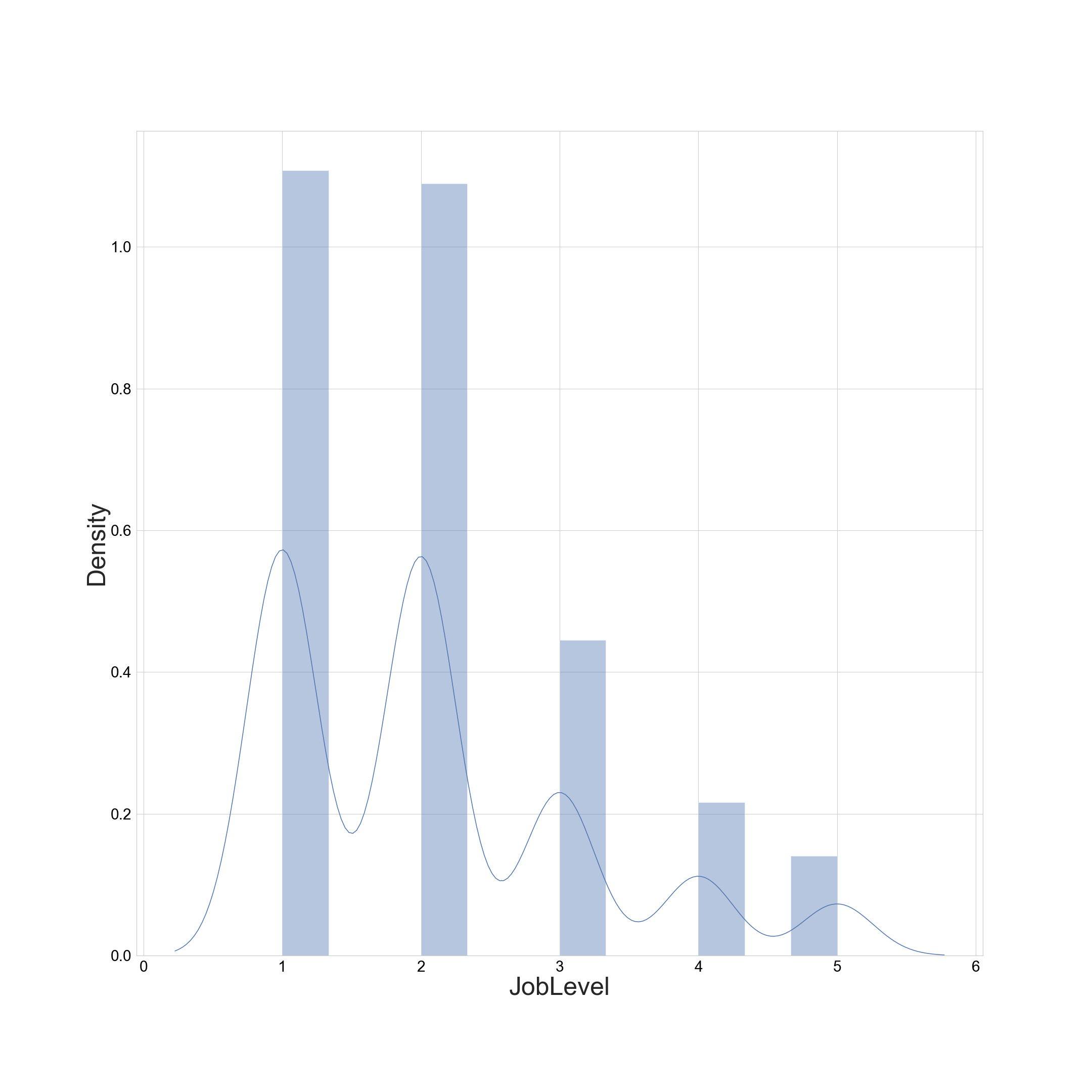
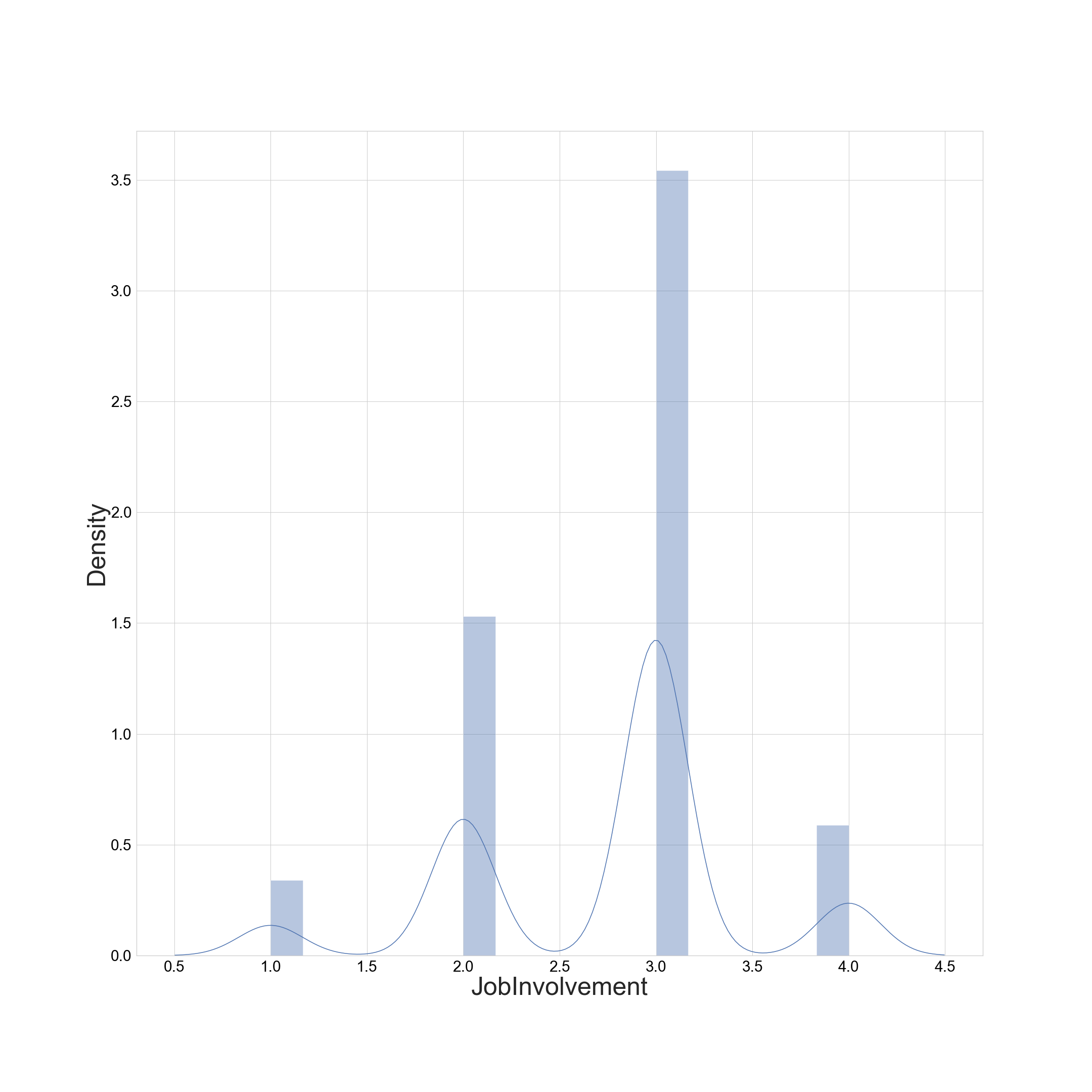
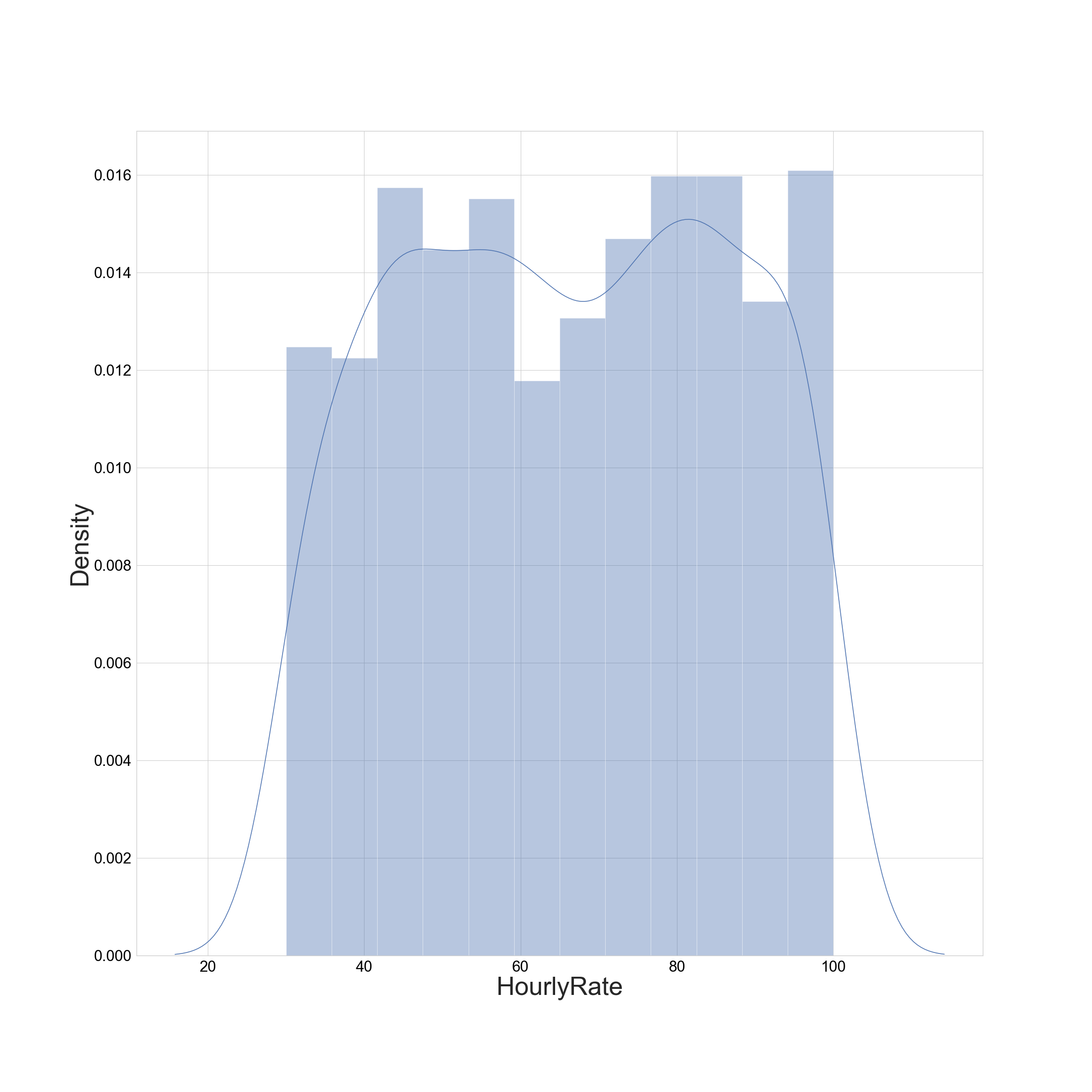
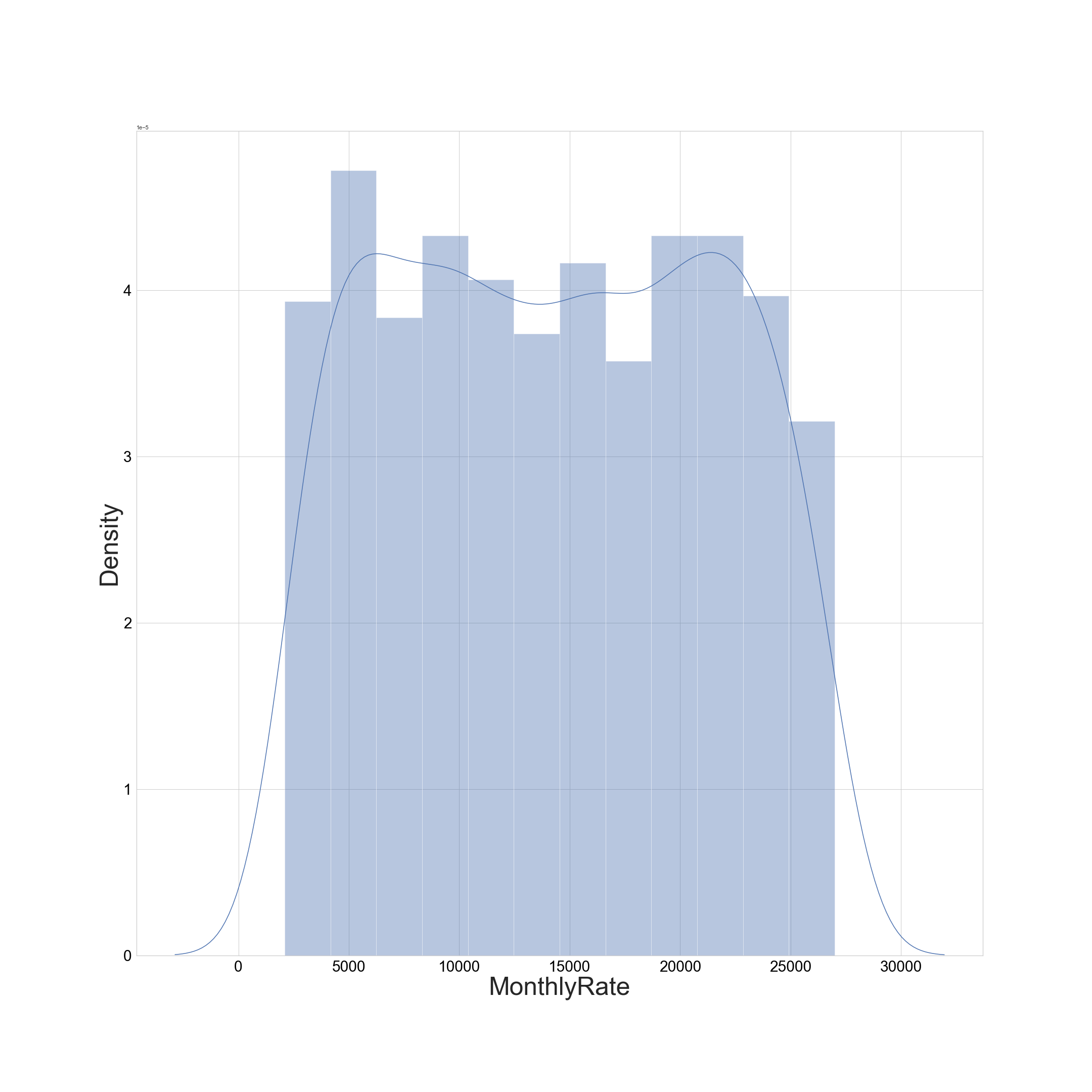
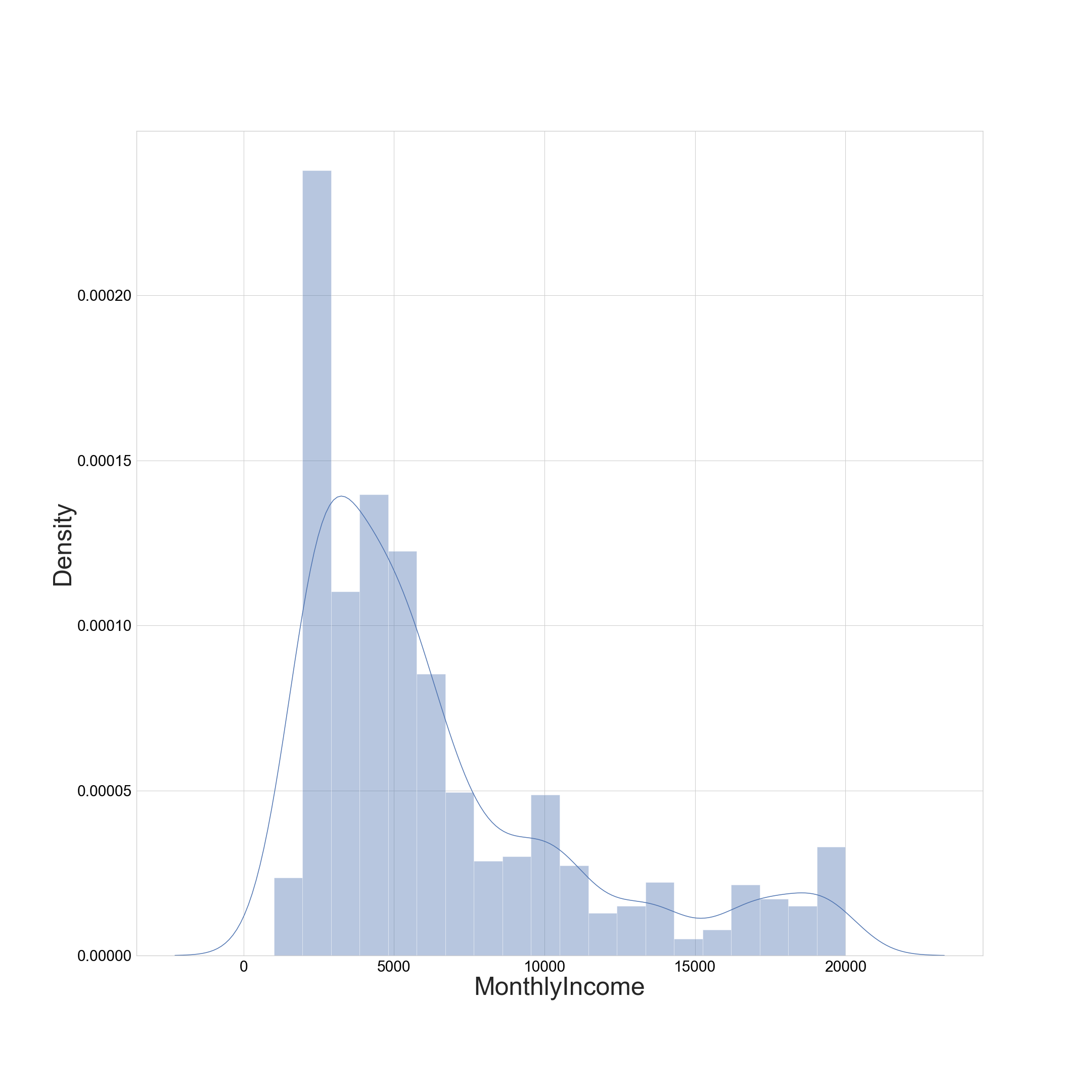
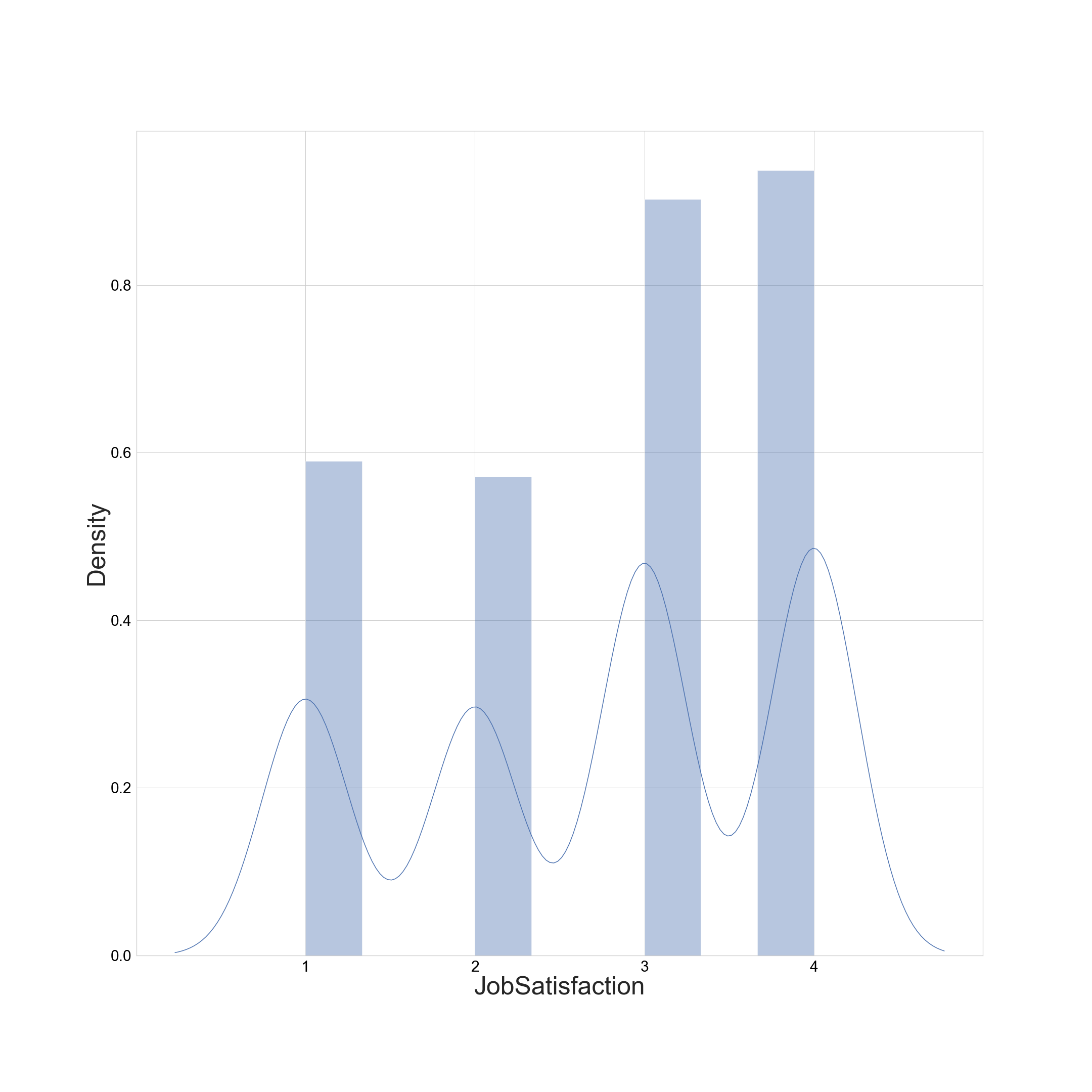
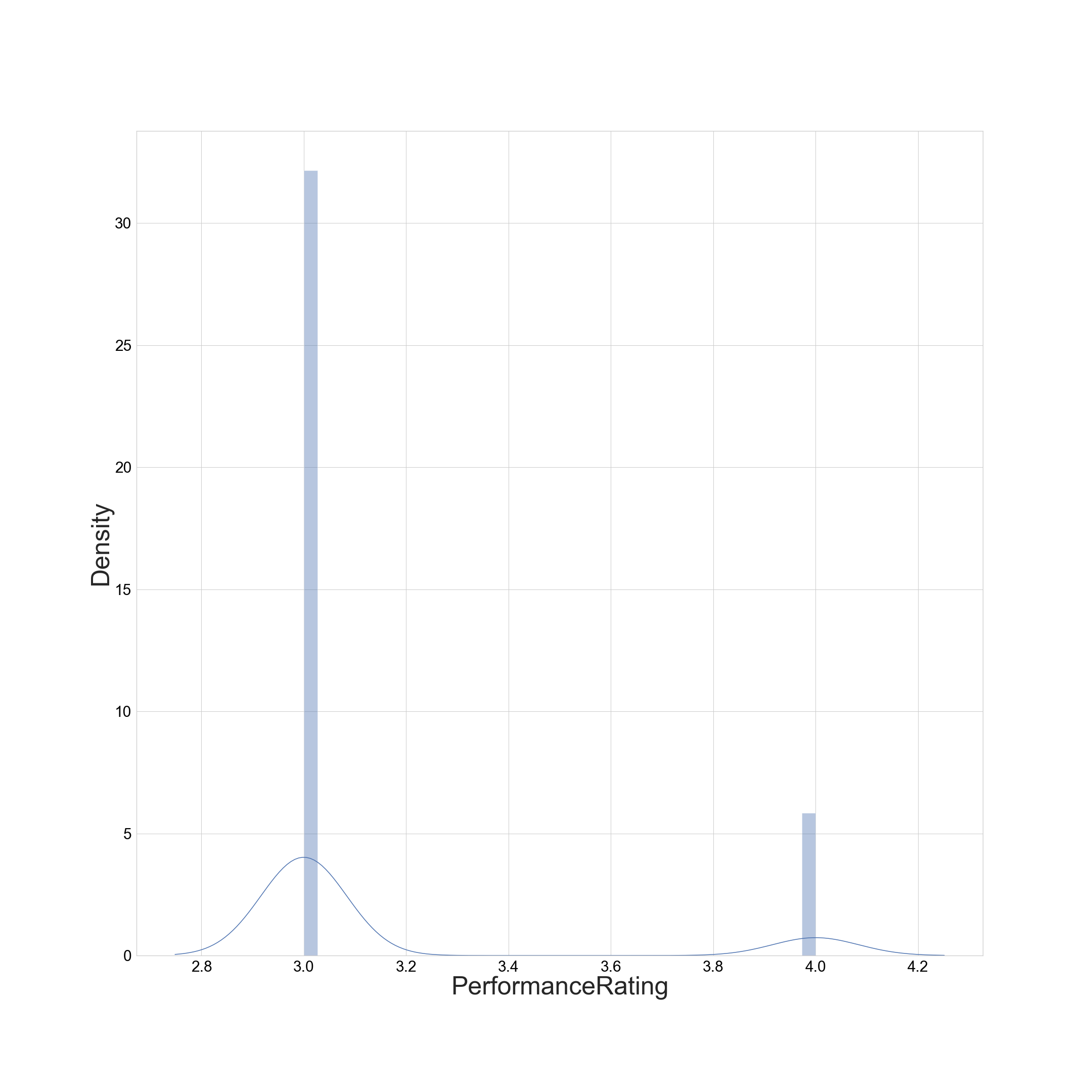
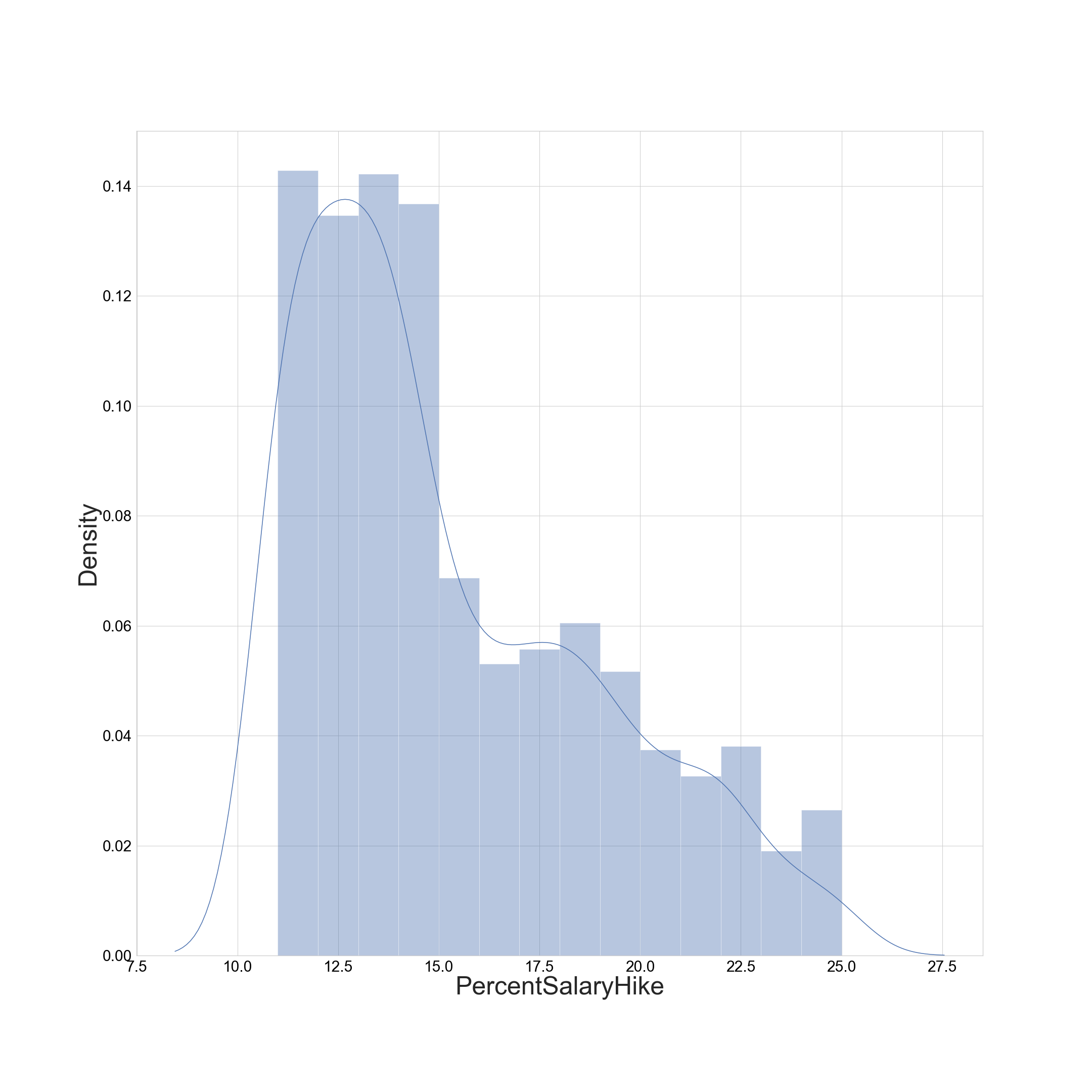
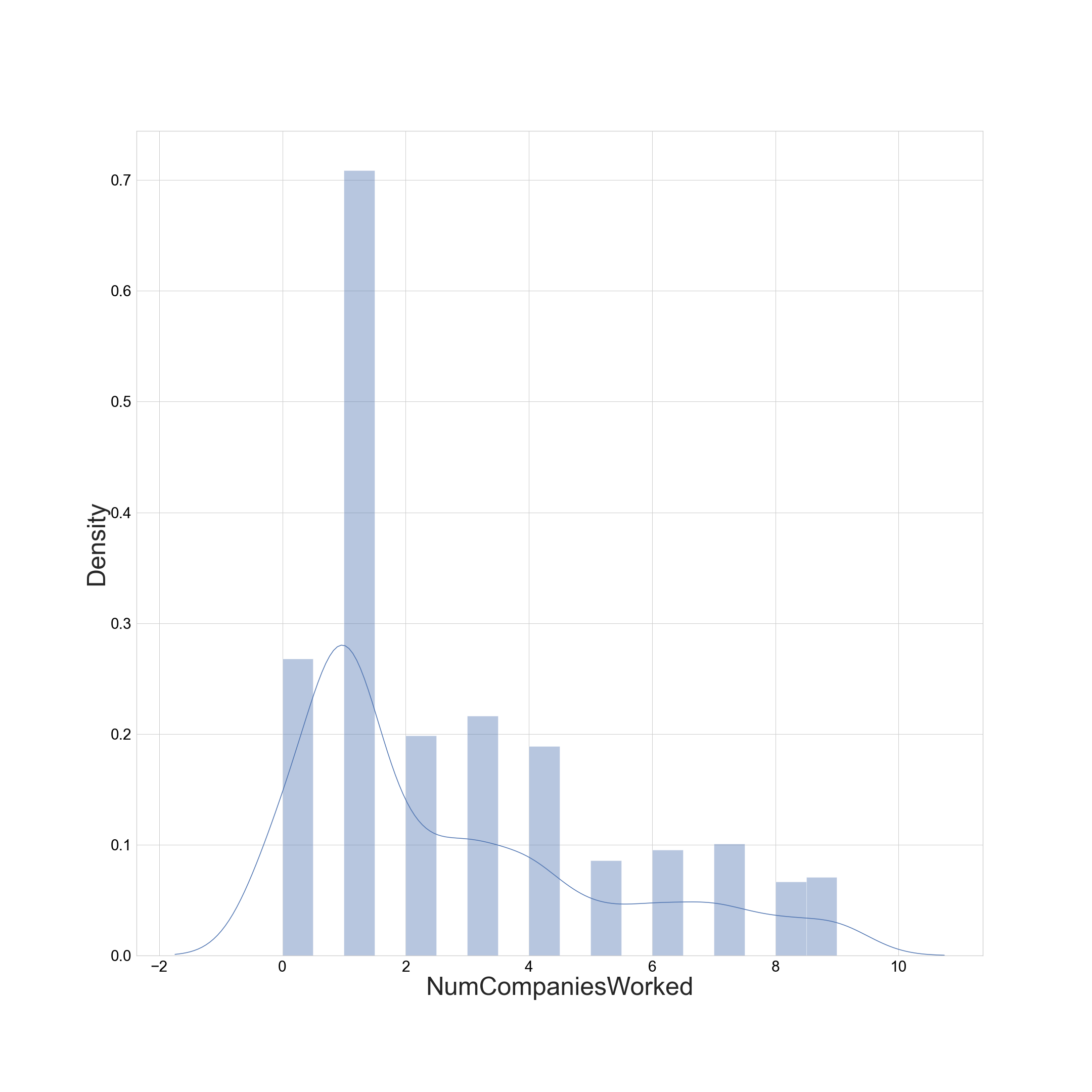
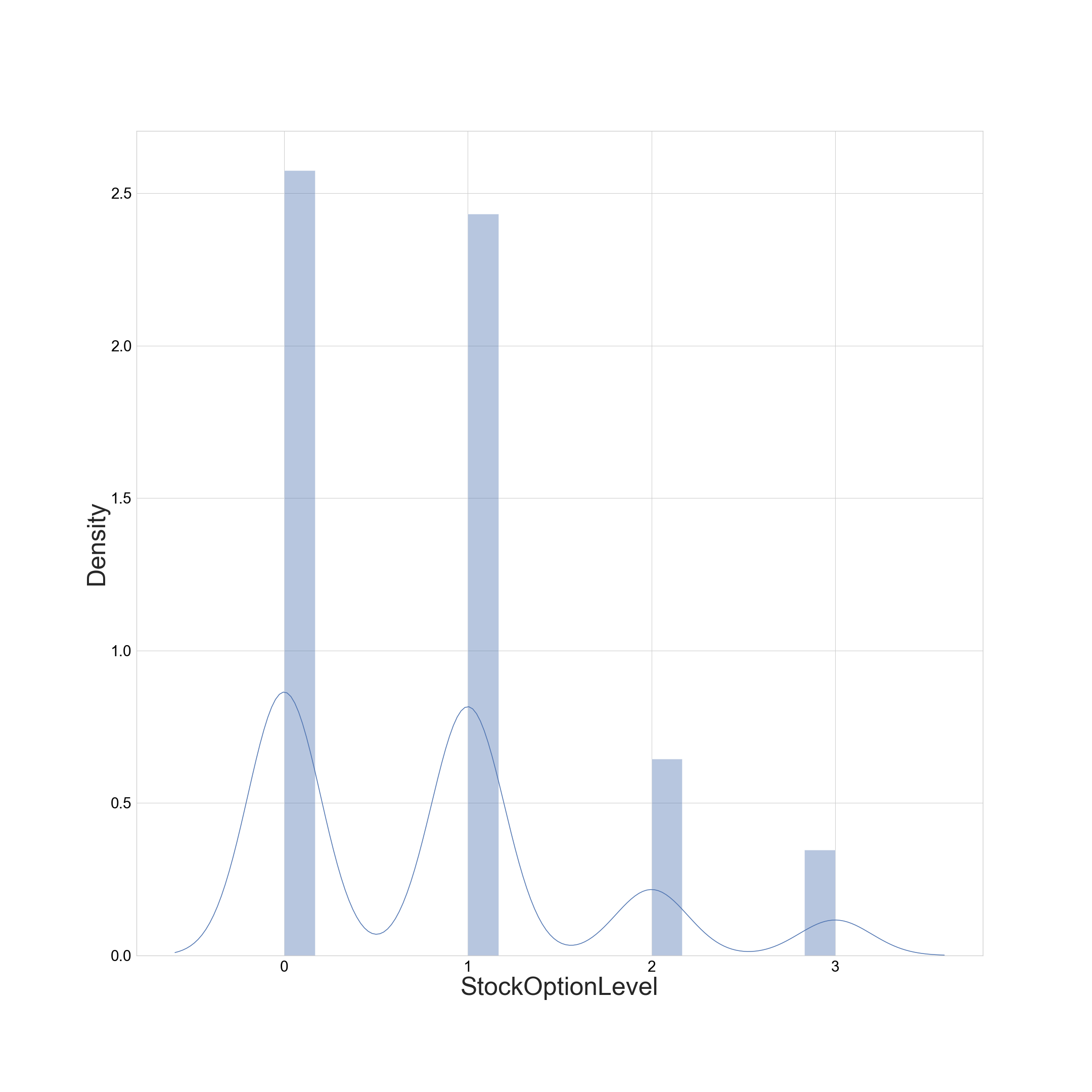
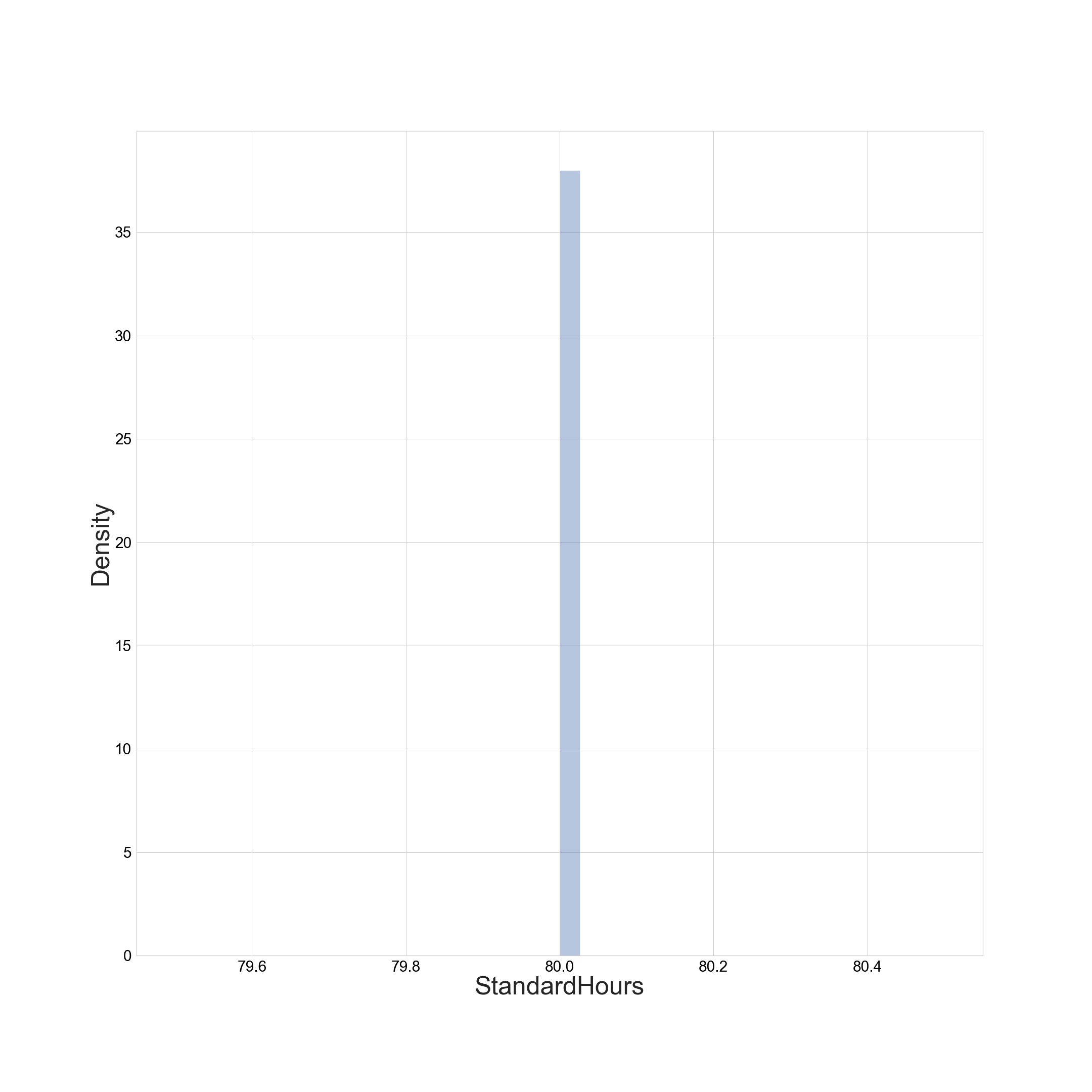
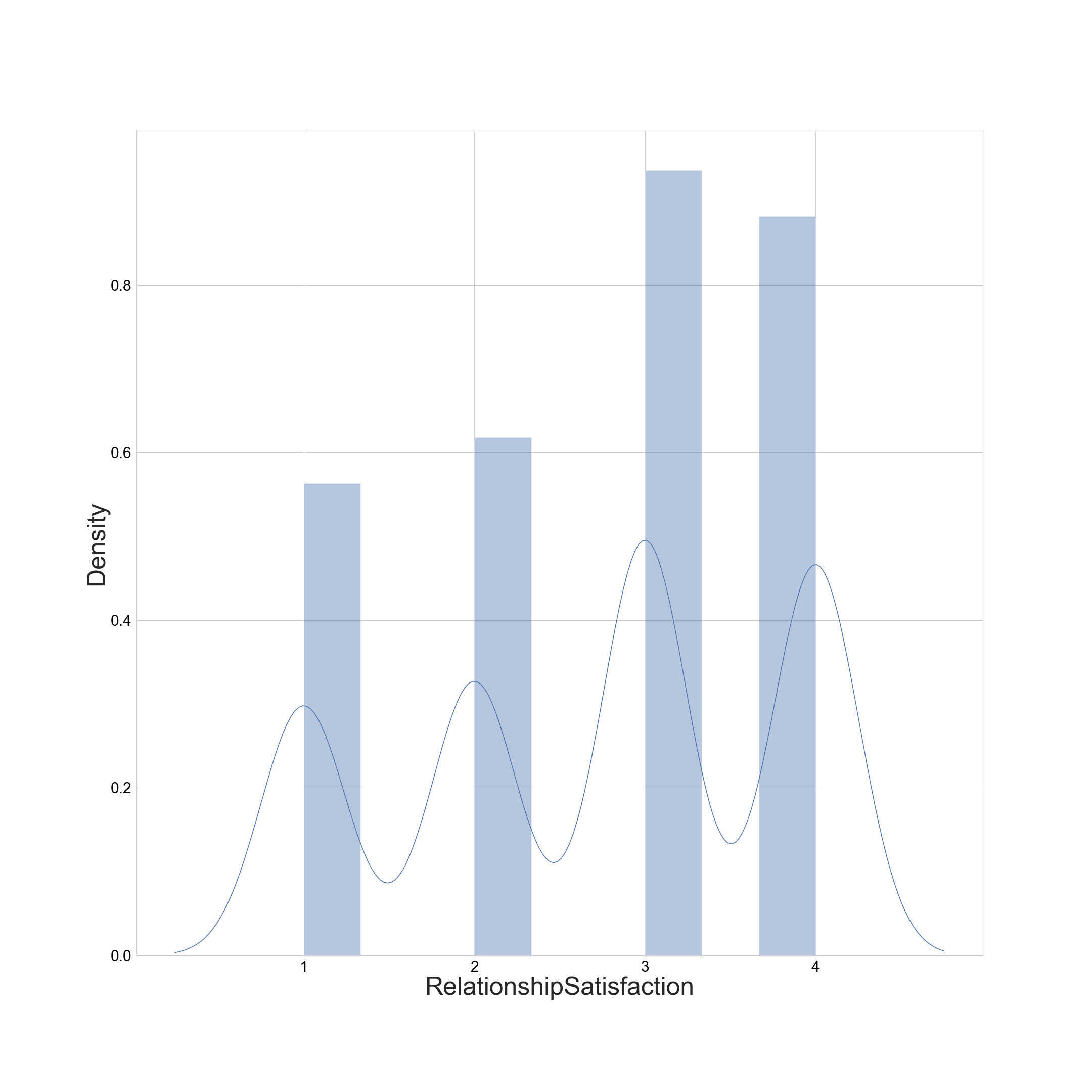
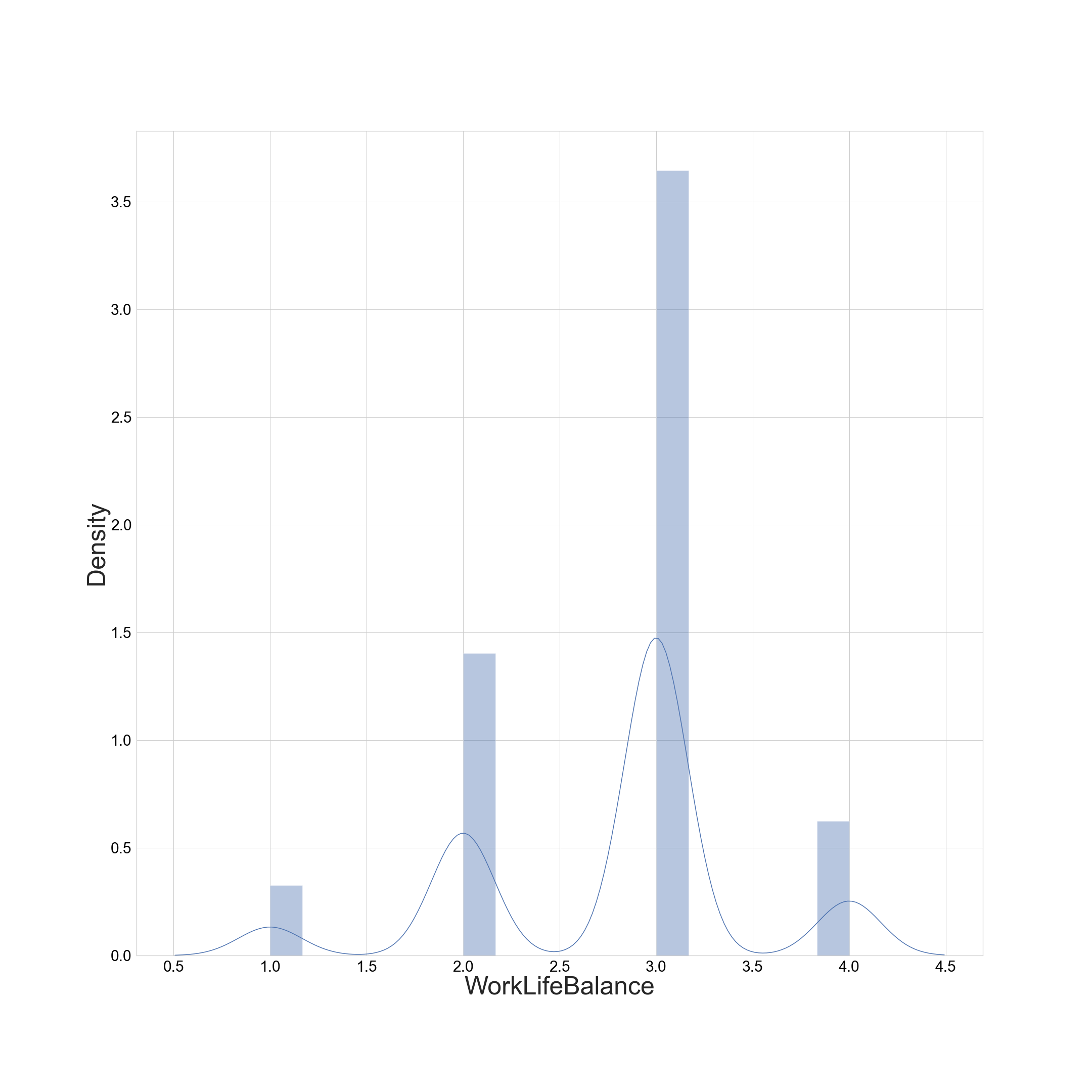
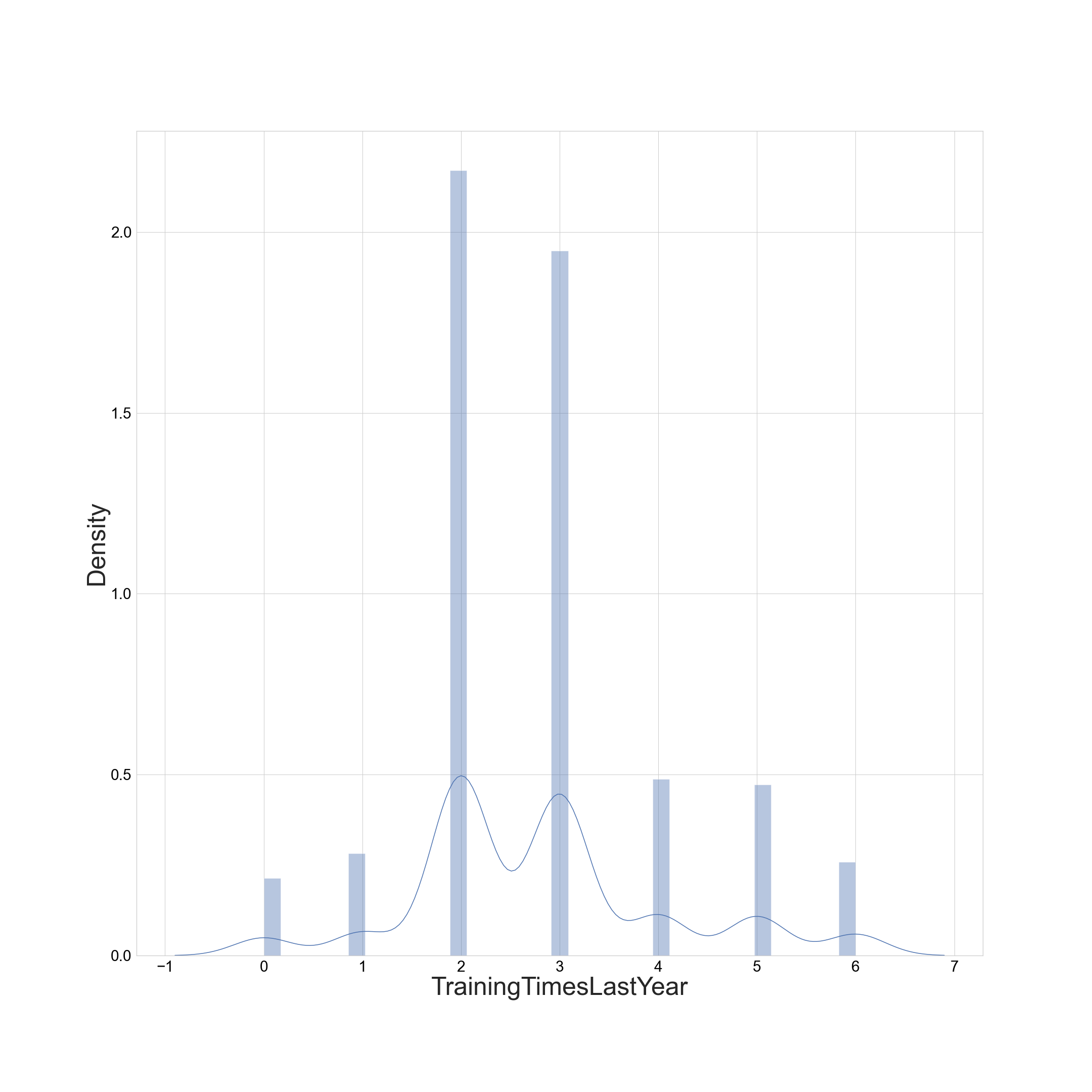
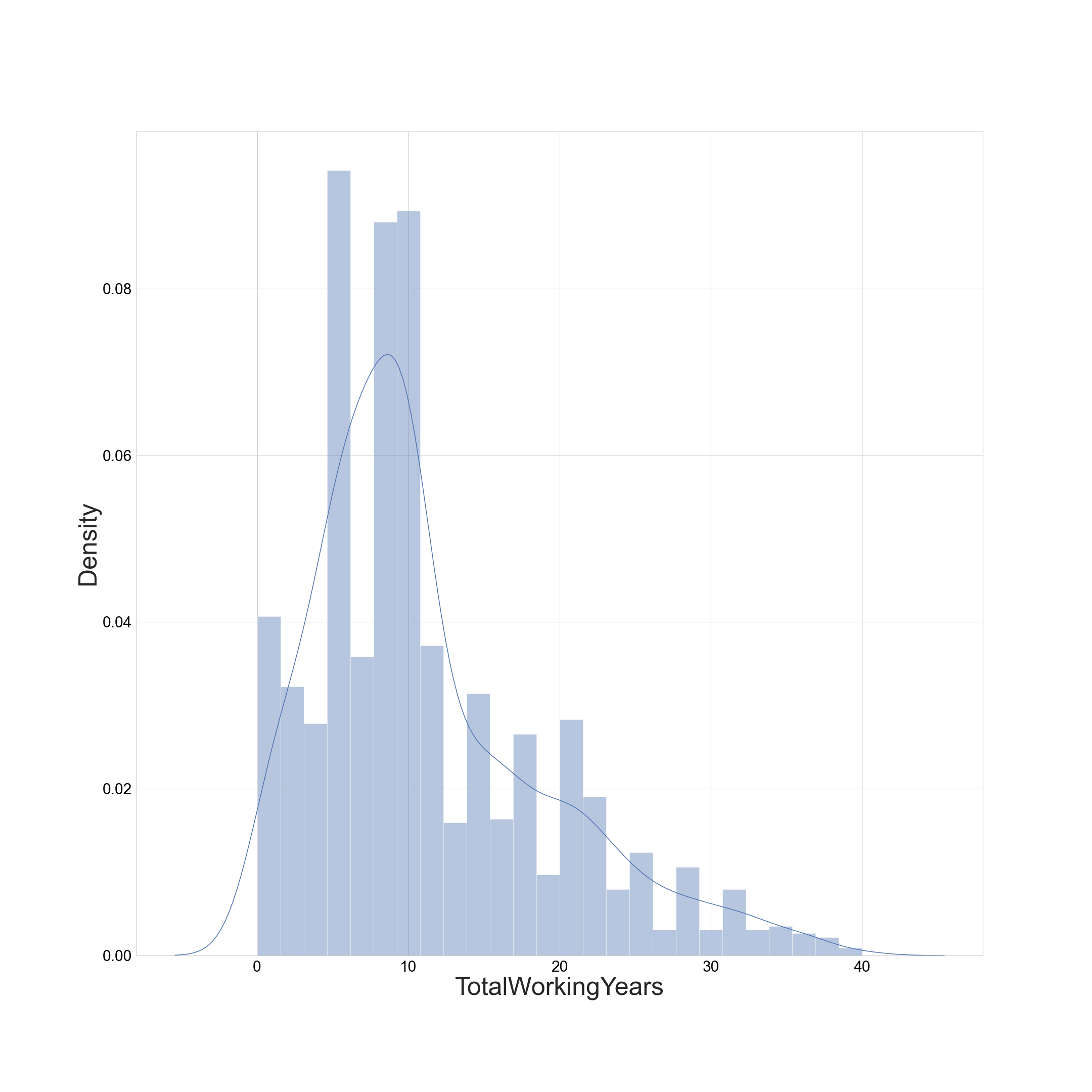
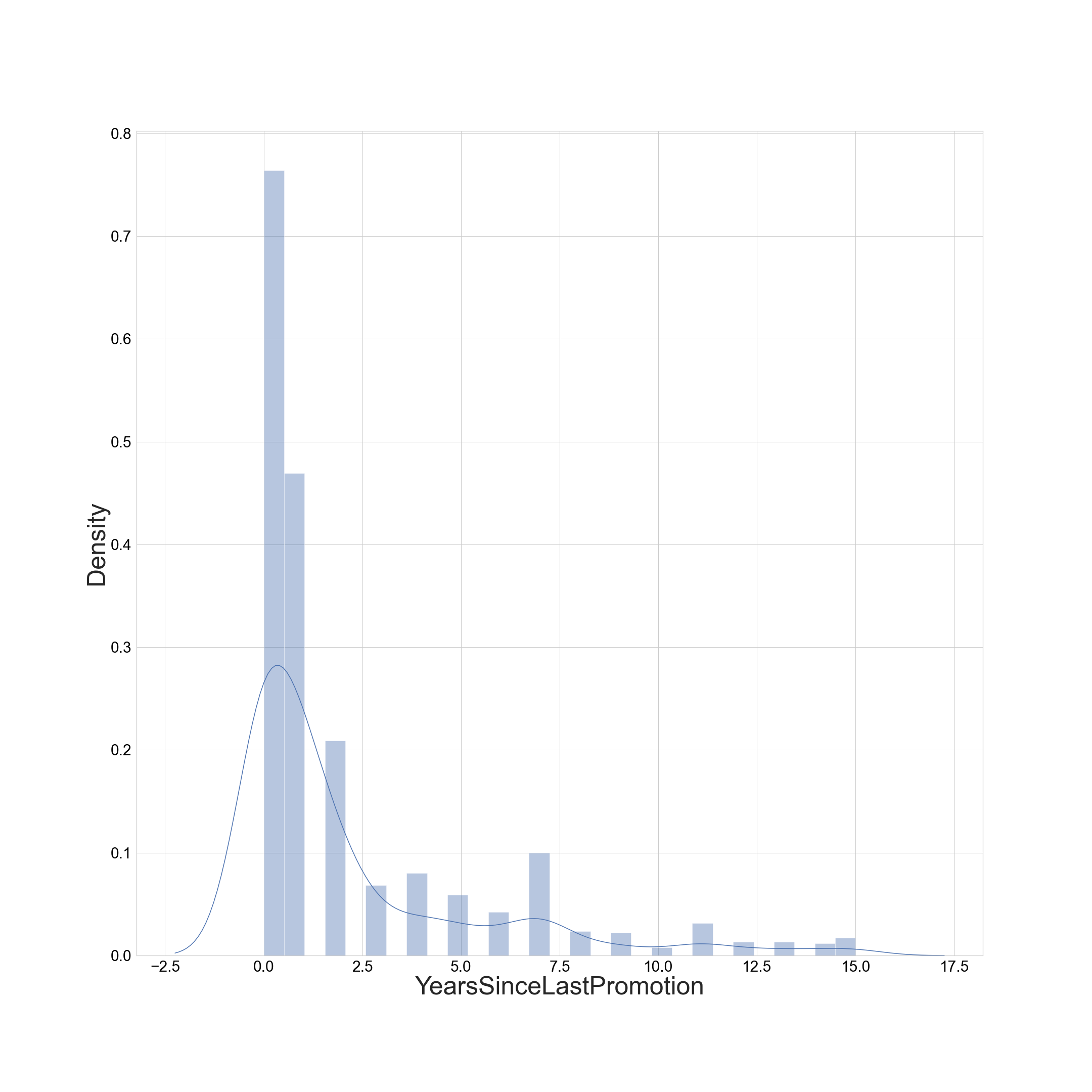
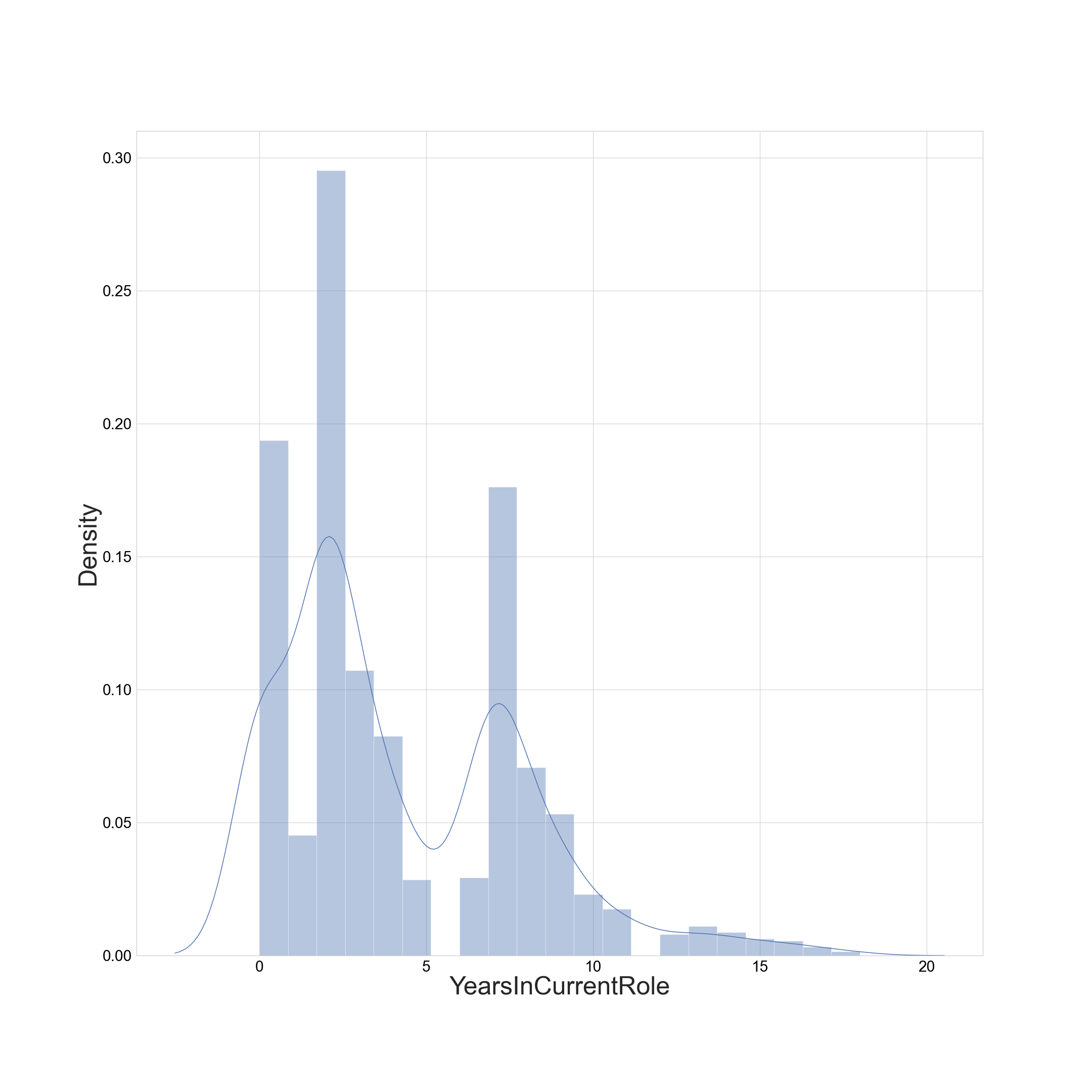
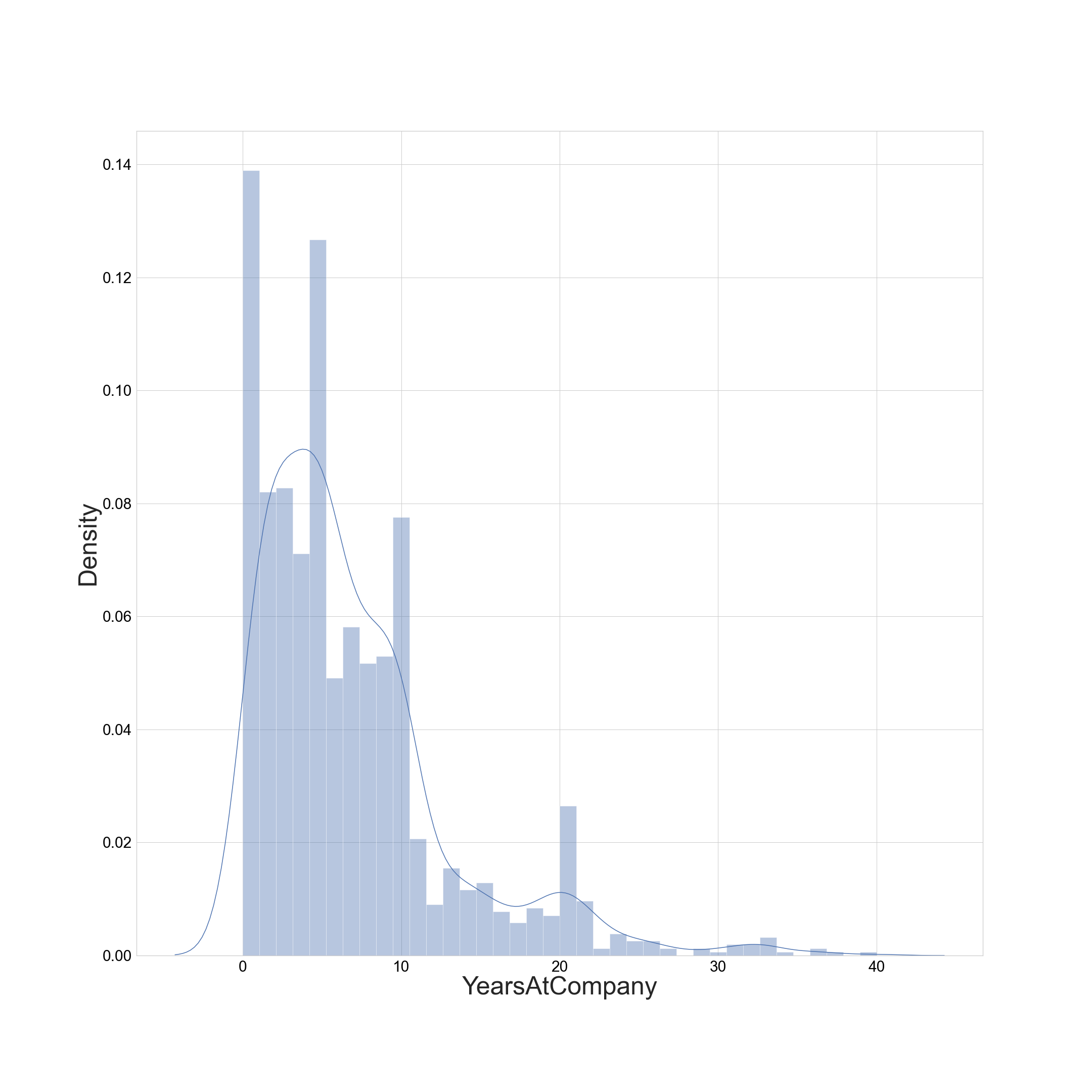
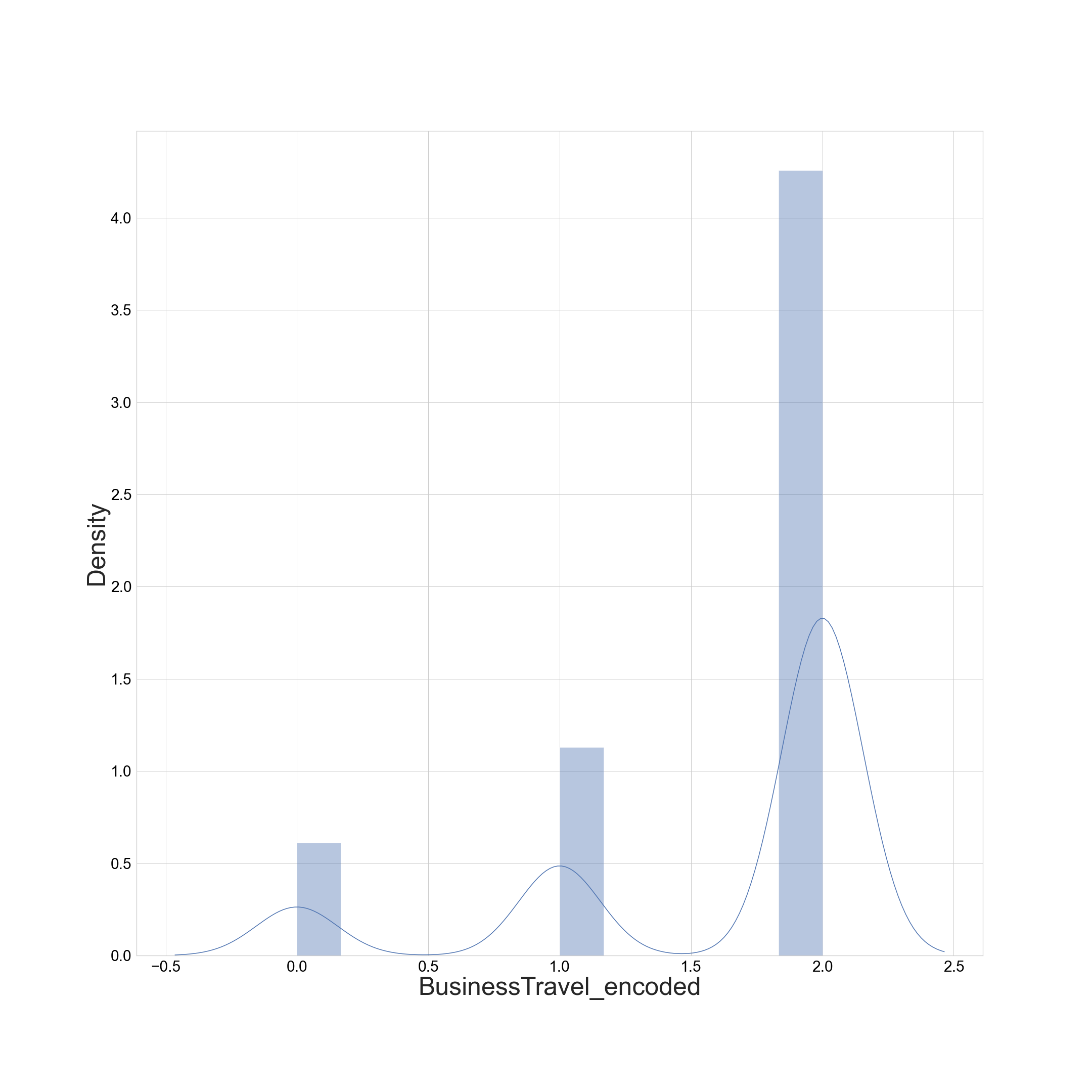
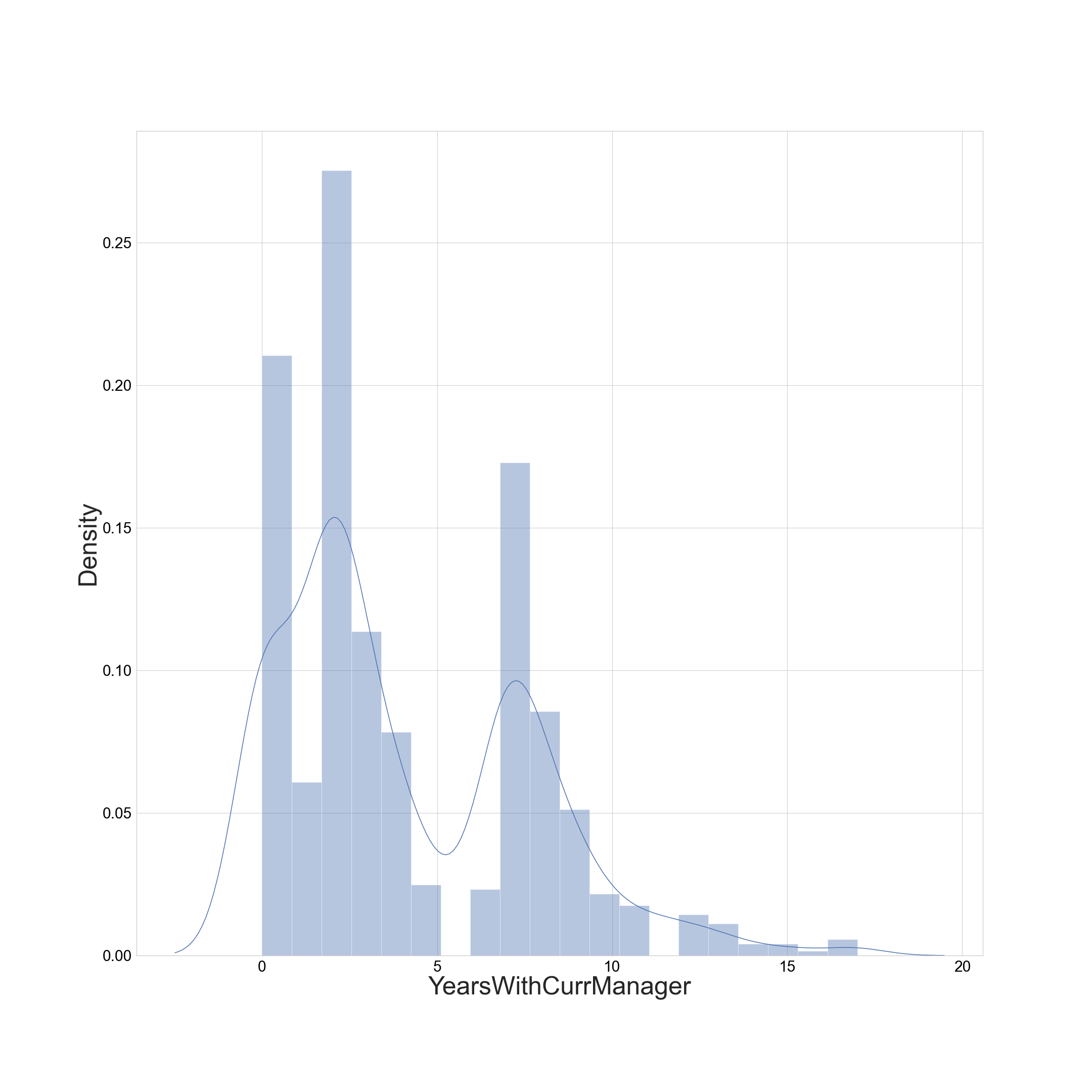
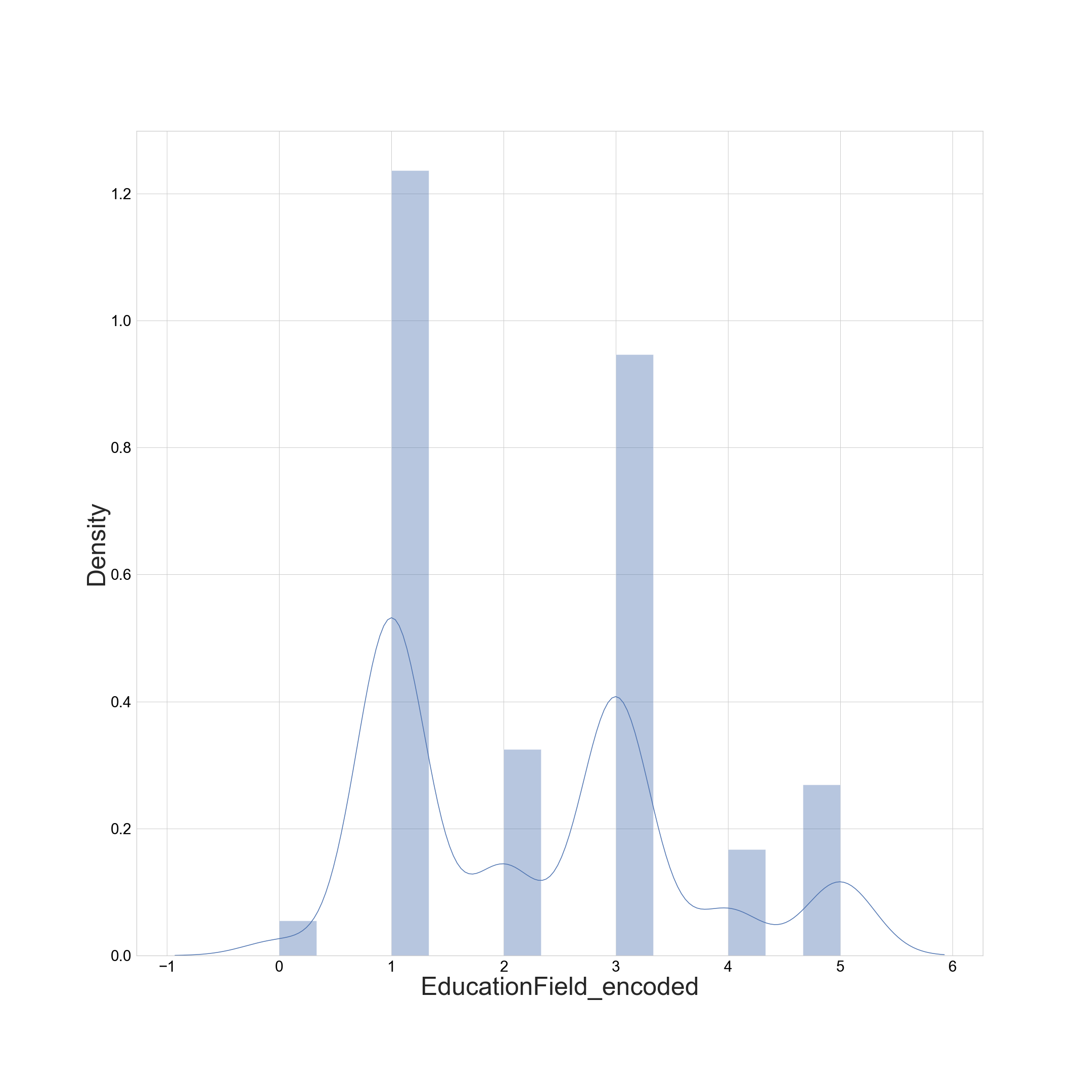
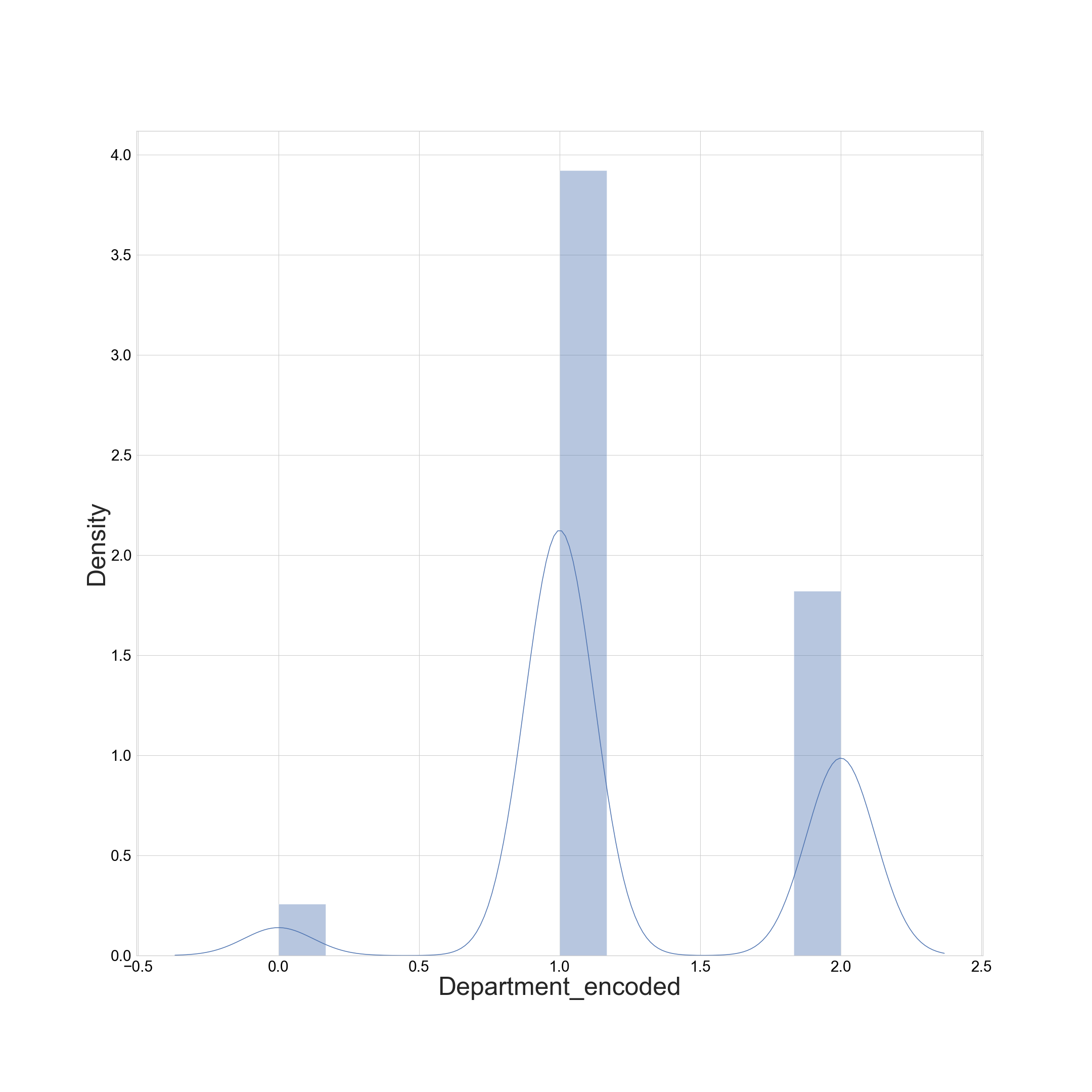
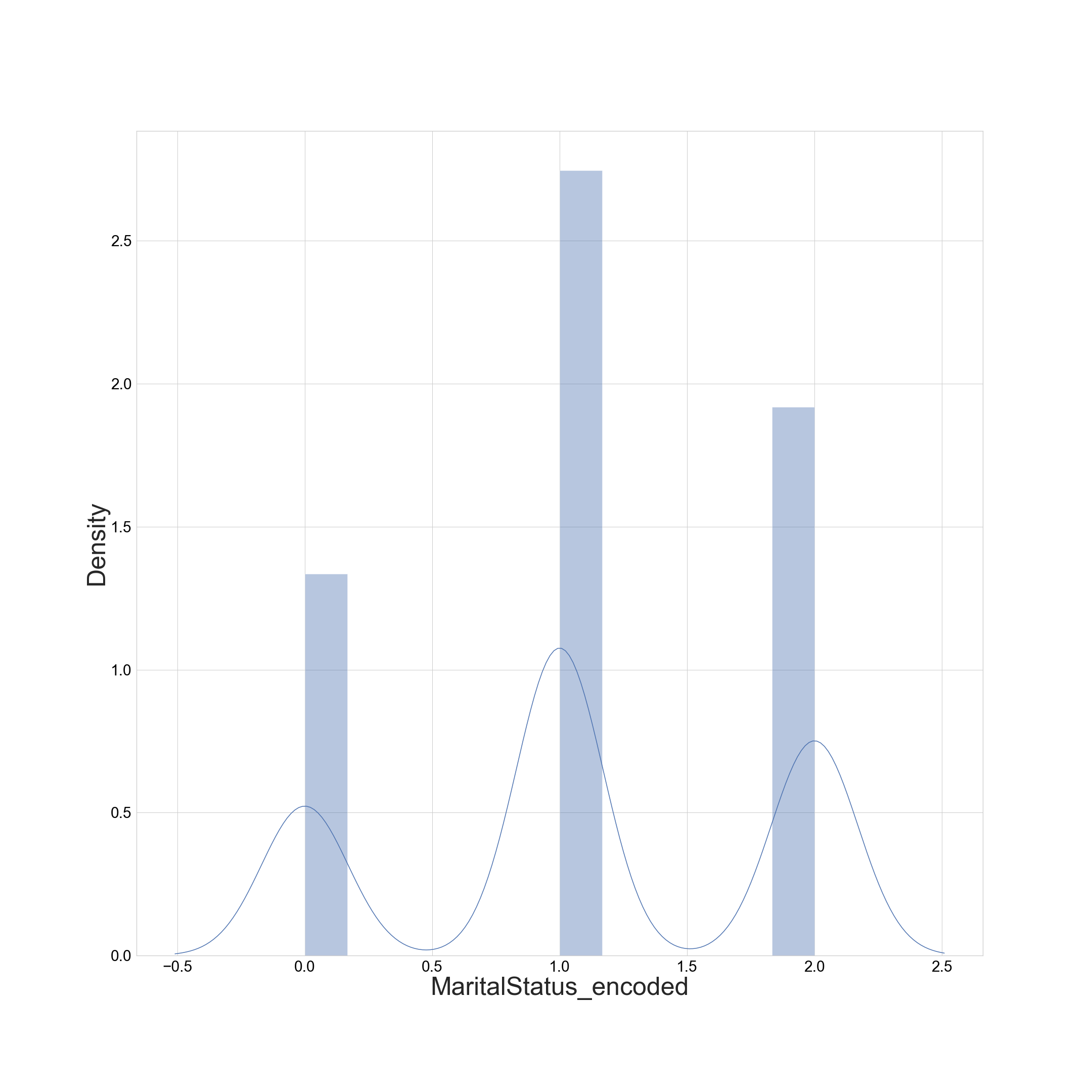
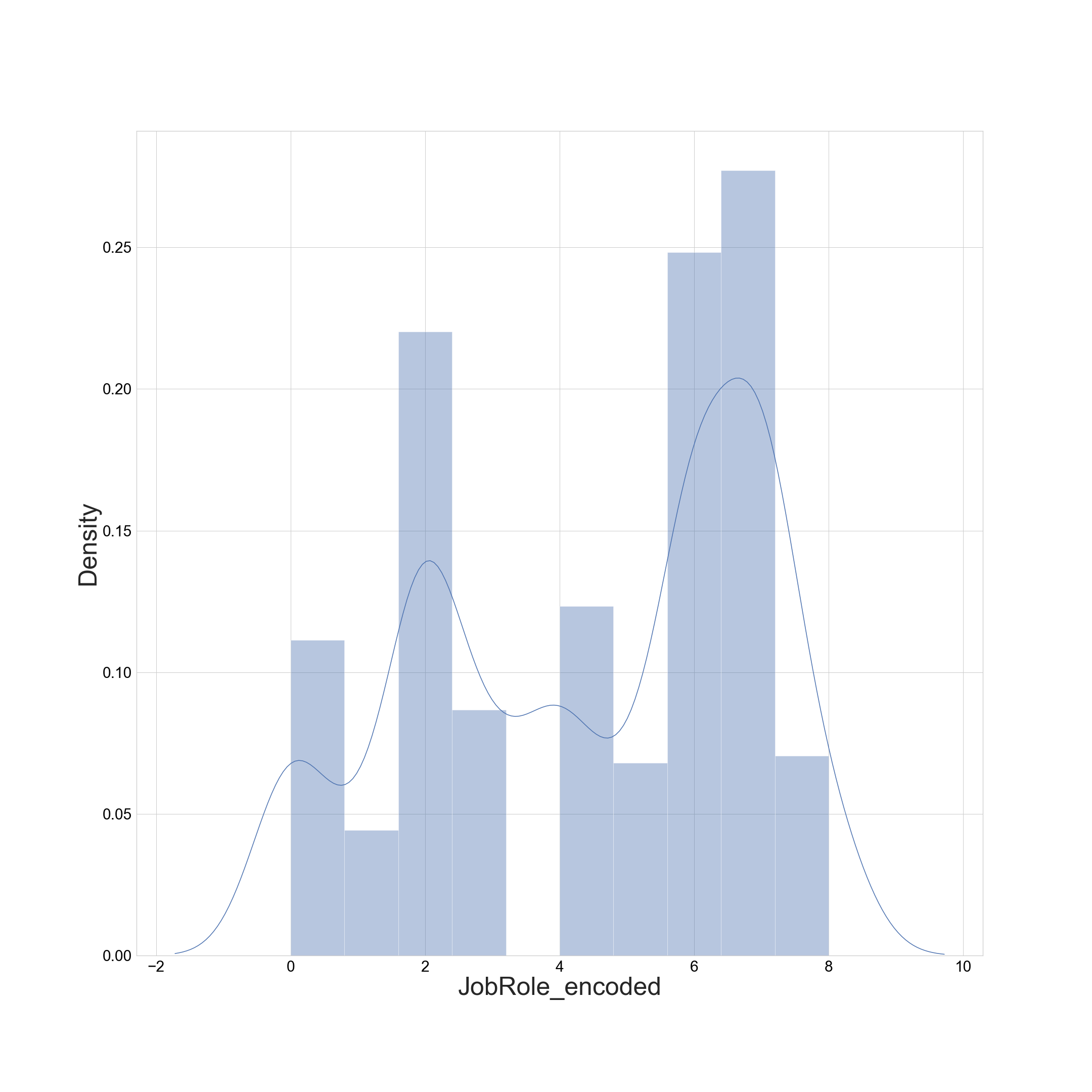
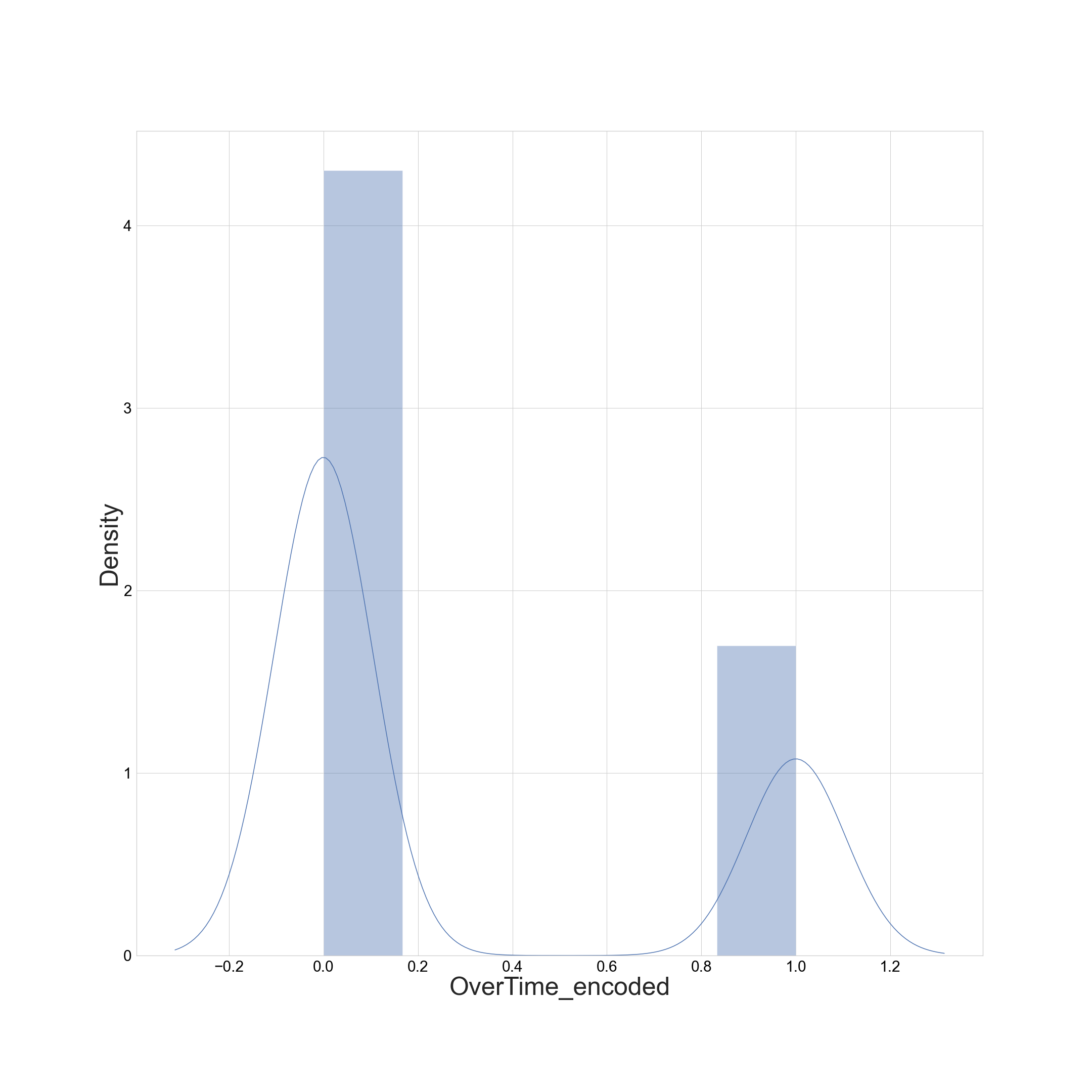
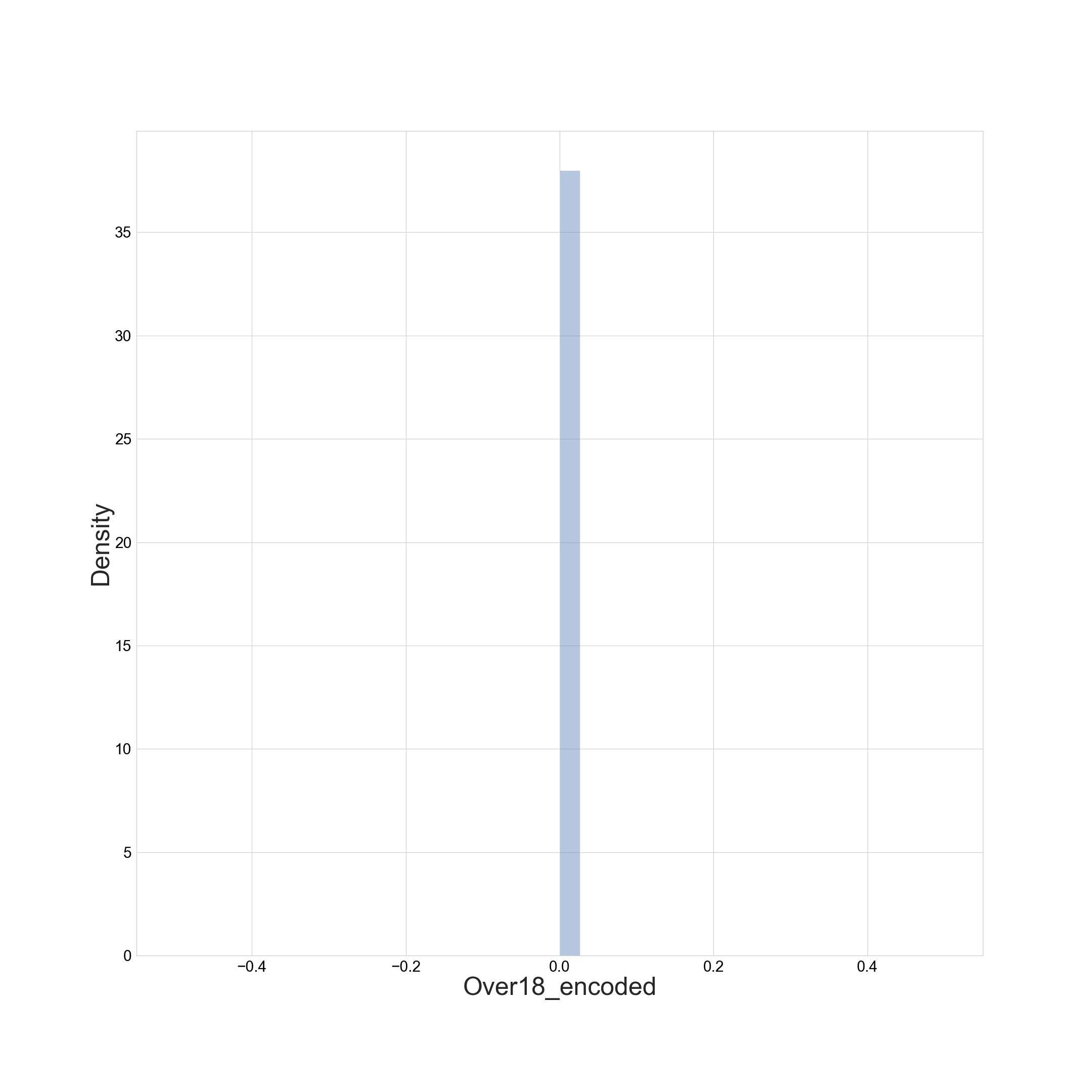
iv) IQR = Upper Quartile - Lower Quartile

4. Range = Maximum Value - Minimum Value

5. Skewness = (summation (value - mean of distribution) \*\*3)/ (number of values - 1) \* std\*\*3)

6. Kurtosis = number of values \* (summation (value - mean of distribution) \*\*4) / std\*\*4)

Analysis 10: Distribution Plots



Observations

1. Acceptable skewness is +/-0.65 and Right skewness for bell shaped curve is 0

2. Acceptable and Outliers Prone Left skewness is observed in:

BusinessTravel\_encoded -1.4390059727642035

WorkLifeBalance -0.5524802990965146

JobInvolvement -0.4984193640419493

Gender\_encoded -0.4086654142437622

JobRole\_encoded -0.3572699195636241

JobSatisfaction -0.3296719586636647

EnvironmentSatisfaction -0.32165444773937907

RelationshipSatisfaction -0.30282756517072296

Education -0.2896810819684331

MaritalStatus\_encoded -0.1521746207726309

HourlyRate -0.03231095290044942

DailyRate -0.003518568352325854

3. Acceptable And Outliers Prone Right Skewness is observed in:

EmployeeNumber 0.016574019580105036

MonthlyRate 0.018577807891132458

Department\_encoded 0.1722308111183741

Age 0.4132863018563338

EducationField\_encoded 0.5503712491120529

TrainingTimesLastYear 0.5531241710537028

PercentSalaryHike 0.8211279755780908

YearsWithCurrManager 0.8334509919918475

YearsInCurrentRole 0.9173631562908262

DistanceFromHome 0.9581179956568269

OverTime\_encoded 0.9644888640425097

StockOptionLevel 0.9689803167738937

JobLevel 1.0254012829518246

NumCompaniesWorked 1.026471111968205

TotalWorkingYears 1.1171718528128527

MonthlyIncome 1.3698166808390662

YearsAtCompany 1.7645294543422085

Attrition 1.8443661240010911

PerformanceRating 1.921882702142603

YearsSinceLastPromotion 1.9842899833524859

4. Bell Shaped Curve (O skewness) is observed in:

EmployeeCount 0.0

StandardHours 0.0

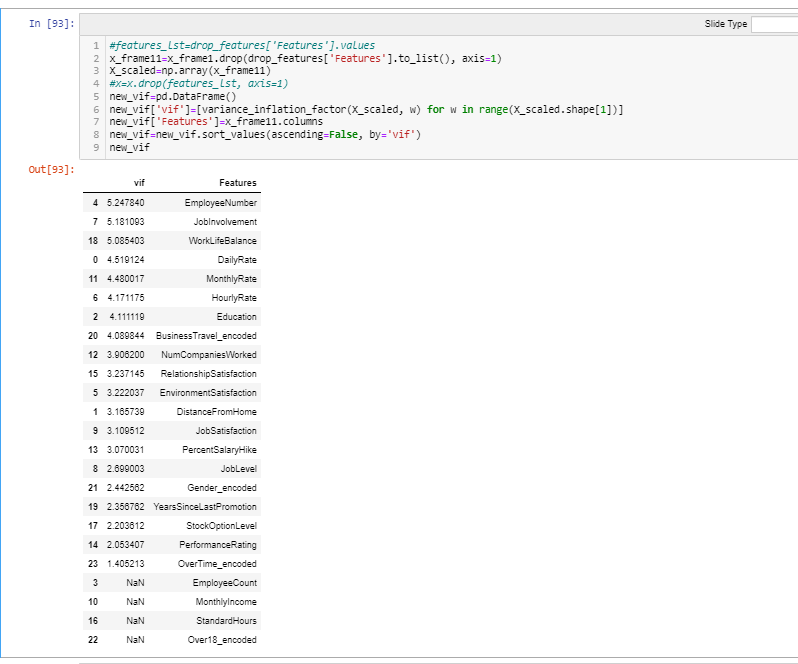
Over18\_encoded 0.0

Analysis 11: Variance Inflation Factor (VIF)

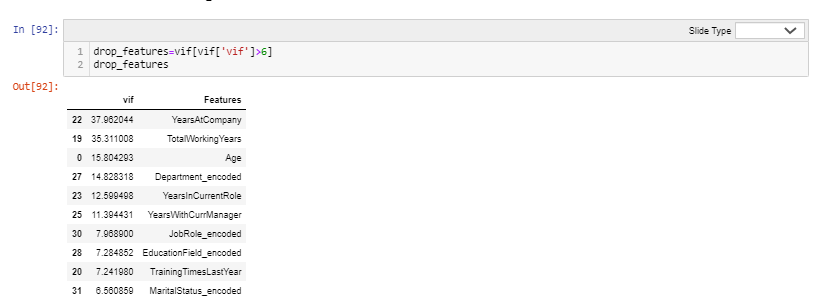
Variance Inflation Factor (VIF) is a measure to spot multicollinearity in the data.

Based on VIF Analysis, we can remove biased columns in the dataset.

VIF Score of each factor, is showcased in the snip below:

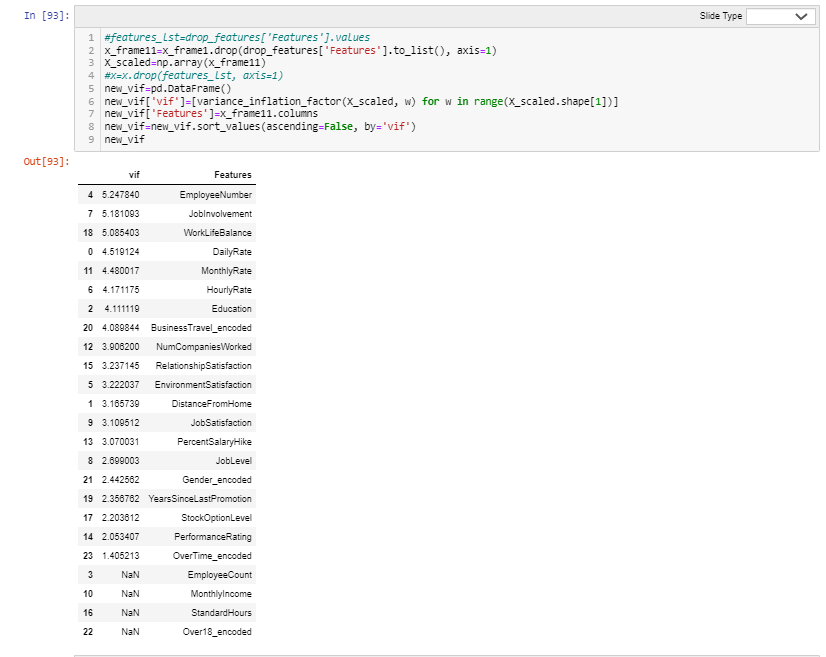


Based on these scores, I have removed 10 features, as highlighted in the snip below:



The dataset now seems free of multicollinearity because VIF scores of remaing columns are less than 6:

*Snippet of Final Scores:*



Exploratory Data Analysis Concluding Remarks

# **Based on the above analysis, these columns have significant outliers, hence, are platykurtic, hence, I am removing these columns.**

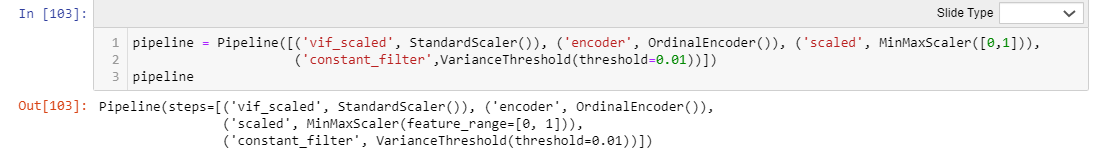
* + 1. Hourly Rate
    2. Education Field Encoded

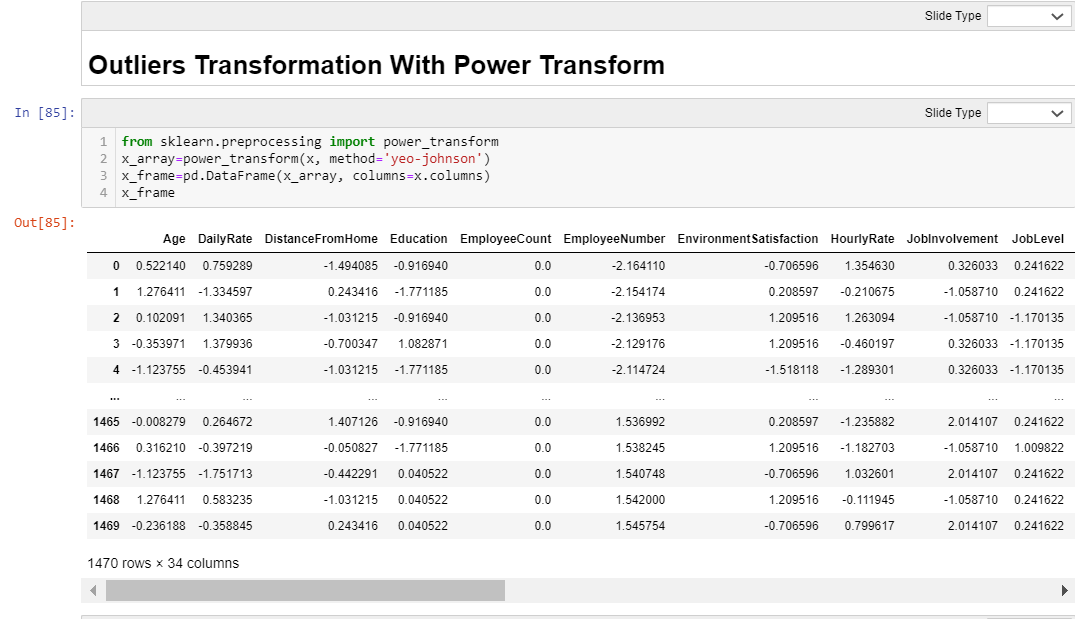
1. The above-mentioned columns can also be called highly biased columns.
2. The column that has the most explanatory power is Business Travel.
3. Multicollinearity irregularities are caused due to multicollinear pairs:
   * 1. Job Level and Age
     2. Marital Status Encoded and Stock Option Level
4. But their descriptive stats do not reflect very high irregularities so we will use variance inflation factor (VIF) scores to detect which of the 4 columns is the root cause of the biasness caused by these pairs.
5. Based on VIF Analysis, I have found 10more biased features and removed those, namely:
   * 1. YearsAtCompany
     2. TotalWorkingYears
     3. Age
     4. Department\_encoded
     5. YearsInCurrentRole
     6. YearsWithCurrManager
     7. JobRole\_encoded
     8. EducationField\_encoded
     9. TrainingTimesLastYear
     10. MaritalStatus\_encoded
6. In total, I have removed 12 of 34 features because those were causing bias in the model.
7. As part of data handling, I have removed features with high outliers (by analyzing box plots, dist plots, variable plot and scatter plots).
8. I have removed multicollinearity by analyzing correlation, correlation heatmaps and variance inflation factor.
9. Hence, the model can be expected to be low bias and low variance model.

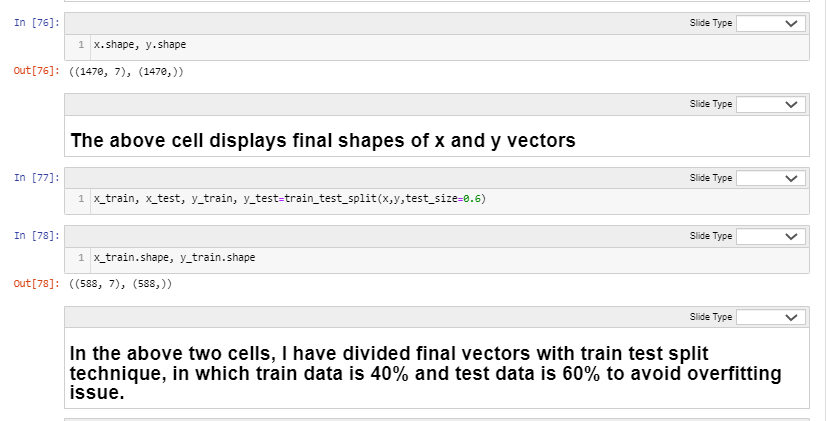
Data Pre-Processing and Pre-Processing Pipelines

4 pre-processing steps executed are:

1. Ordinal Encoding of Categorical Data.
2. Power Transform for standard scaling of features and outliers' transformation.=does not apply fit and transform method and hence cannot be stored in pipeline.
3. Min Max Scaling for scaling of features from 0 to 1.
4. Standard Scaling For estimating Variance Threshold.
5. Variance Threshold using Feature Selection.
6. Train Test Split= addresses overfitting issues in the model. Since it does not apply fit and transform method and hence cannot be stored in pipeline.







Building Machine Learning Models

# **Model Development, Evaluation and Saving (Total Models = 10)**

# **Selection Reasoning of Models 1 to 5**

# **The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.**

Two families of ensemble methods are usually distinguished:

In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimators because its variance is reduced.

Examples: Bagging methods, Forests of randomized trees, etcetera

By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.

Examples: AdaBoost, Gradient Tree Boosting, etcetera

The assigned use case revolves around uneven label points hence there is high probability of not achieving a good fit. Hence, I have tried different ensemble techniques that can lower variance and bias and help achieve a goodness of fit. (Just as a reminder, I have already applied variance threshold of 0.01 to ensure that risk of models is low).

The theories in the two cells above explain why I have chosen Model 1, Model 2, Model 3, Model 4 and Model 5 for this use case. Model 5 is a support vector classifier and hence, is again a very powerful implementation to prevent illness of fitness and make appropriate decision boundaries.

Model 1: Random Forest Classifier with Intuitional Hyper Parameter Tuning

Model 2: Random Forest Classifier with Default Hyper Parameter Tuning

Model 3: RFC With Grid Search CV

Model 4: Bagging Classifier with Grid Search CV Hyper Parameter Tuning

Model 5: Support Vector Classifier with Sigmoid

# **Resampling, Rescaling and Re Feature Selection and Re Splitting for Model Optimization**

# **Reasoning Of Applying Resampling and developing models 6 to 10.**

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.

Two families of ensemble methods are usually distinguished:

In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

Examples: Bagging methods, Forests of randomized trees, etcetera

By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.

Examples: AdaBoost, Gradient Tree Boosting,etcetera

Since, original y vector is highly unsampled, I will do resampling to balance both 1 and 0 data points to ensure that along with being low on variance, the model is also low on bias. Hence, I can achieve bias variance trade off.

Hence, I have developed models 6 to 10 based on rfc, ensemble and svm methods on resampled data

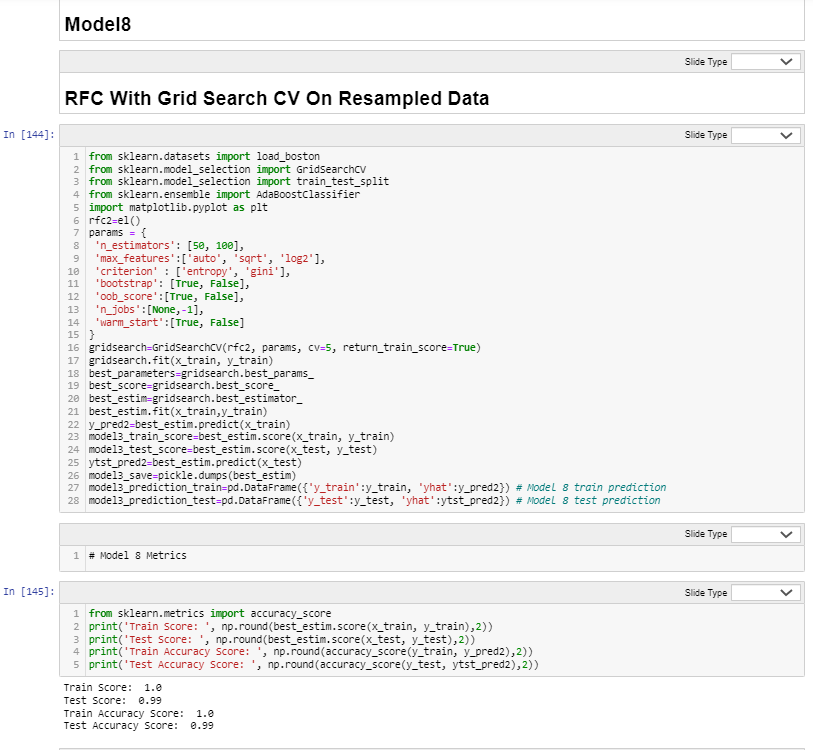
Model 6: RFC On Resampled Data with Intuitional Hyper Parameter Tuning

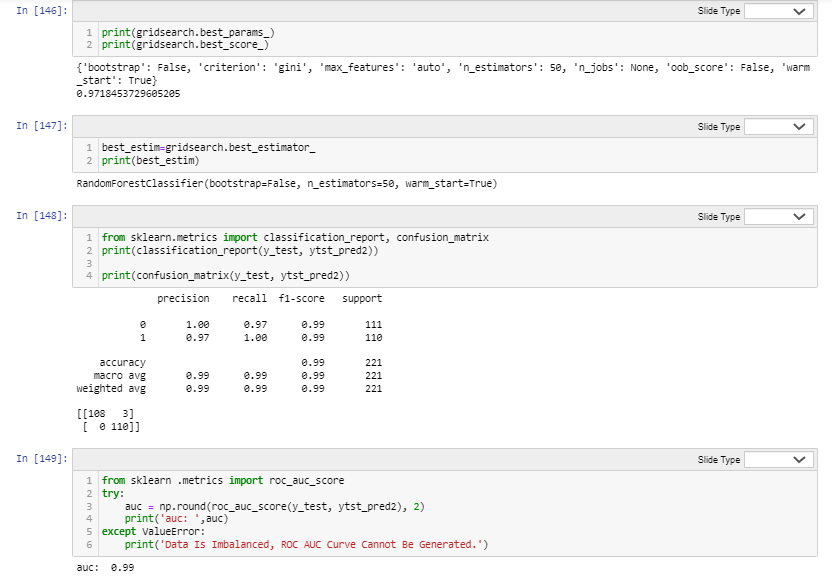
Model 7: RFC On Resampled Data with Default Hyper Parameter Tuning

Model 8: RFC With Grid Search CV On Resampled Data

Model 9: Bagging Classifier with Grid Search CV Hyper Parameter Tuning on Resampled Data

Model 10: Support Vector Classifier with sigmoid to prevent overfitting on resampled data







In the above snippets, I have elaborated Model Development, Evaluation and Saving of The Successful Model that will be sent to production.

Successful Model is saved in pickle for production.

# **Conclusion**

There is 1 successful mode with right fit The model daisplays Low Variance And Low Biance with good accuracy scores.

Model 8: RFC With Grid Search CV On Resampled Data

Train Score: 1.0

Test Score: 0.99

Train Accuracy Score: 1.0

Test Accuracy Score: 0.99

precision recall f1-score support  
  
 0 1.00 0.97 0.99 111  
 1 0.97 1.00 0.99 110  
  
accuracy 0.99 221

macro avg 0.99 0.99 0.99 221 weighted avg 0.99 0.99 0.99 221

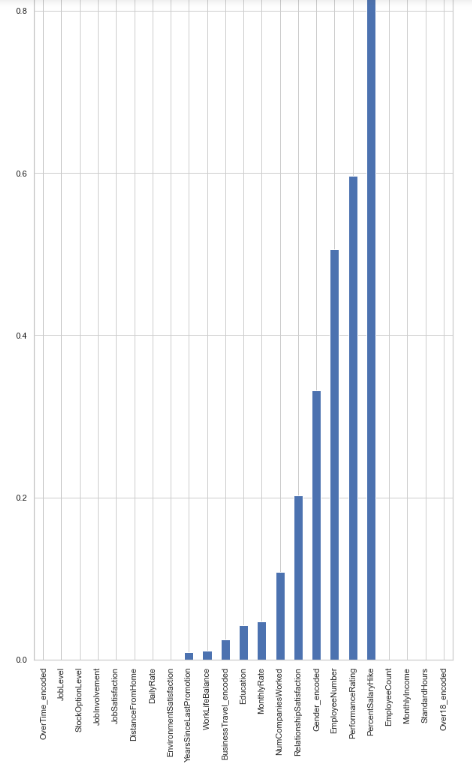
[[108 3] [ 0 110]]

Please note, metrics calculation is explained above while development and evaluation.

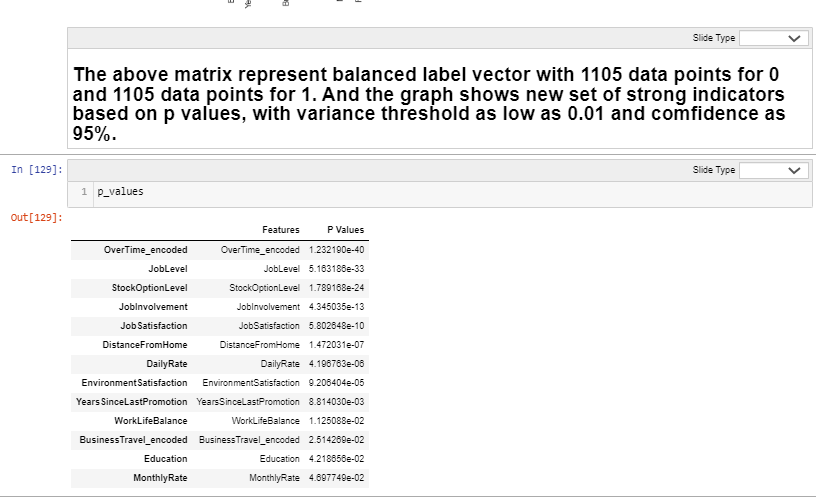
Concluding Remarks

An onlooker can say that Business Travel is the most important feature. But to create a low variance and low bias model is not just about onlooking, it requires thorough analysis and hypothesis testing. The model oozes from one step to another and on its way onwards, does hypothesis testing and grid search cv best estimators' selection.

To reduce bias in the model, I have performed resampling of data and done feature selection hypothesis test.



Now, a keen scientist can say that Over Time Encoded is the most important feature.



The above snippet highlights all the important features and their p values.

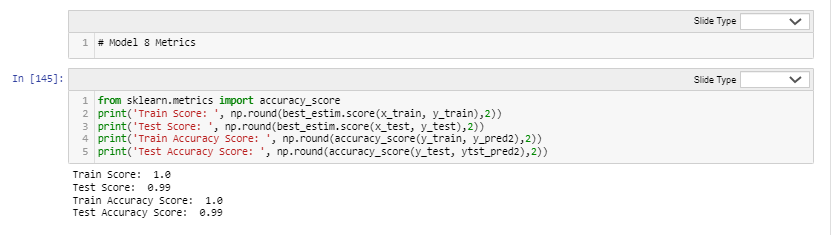
Best estimators and best score as per GridSearchCV are:



I have achieved a model with:

Train Score: 100%, and

Test Score: 99%



The above snippet highlights the scores attained by the model in training and testing phases.

The model could achieve this accuracy score by applying these fixtures:

1. Removing highly skewed and inappropriately peaked curves With Exhaustive EDA.
2. Removing multicollinearity with Correlation Analysis and VIF scores.
3. Resampling Of Dataset.
4. Applying Grid Search CV for Hyper Parameter Tuning.

Credits

Python Documentation and Data Trained.