Term Project: Final Submisson

Nicholas De Santos

DSC 550

Introduction

Problem Introduction and Importance:

The business problem that was explored through the process of this data analysis is the problem of trying to determne what asspects of an employee's job affect that same employee's satisfaction in their particular career. The main goal of this investigation is to possibly help in identifying what features and benefits jobs could provide for their employees in order to improve job satisfaction in a work environment. This can come as a great importance to both employees and employers as well. It can help employee's know what kind of job benefits they should look for in a new job/career in order to have a better chance in higher job satisfaction. The results can be useful to employers as well because better job satisfaction within a workforce can help with productivity as well as limiting job turnover and wasting resources constantly hiring and training new employees. Because of it's benefits for both employees and employers, this analysis could benefit just about any company wanting to improve their employees' job satisfaction.

Data Used and Background:

The data that will be used for this analysis is titled "Employee Attrition and Factors". The data provides information on 1470 different employees from various roles and career paths. The information of each employee includes things like the age of the employee, their hourly/daily/monthly rates, what department they work in, how far they travel for work and so much more. The data was aquired from the following link where more information on the data is provided:

https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors/

Main Analysis

Exploratory Data Analysis:

In [6]:
```
# loading main packages/libraries
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
```

In [7]:
```
#loading data
hrdata = pd.read_csv("HR_Analytics.csv")
hrdata
```

Out[7]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Educa
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life
4	27	No	Travel_Rarely	591	Research & Development	2	1	
...	...	...	...	...	...	...	...	
1465	36	No	Travel_Frequently	884	Research & Development	23	2	
1466	39	No	Travel_Rarely	613	Research & Development	6	1	
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life
1468	49	No	Travel_Frequently	1023	Sales	2	3	
1469	34	No	Travel_Rarely	628	Research & Development	8	3	

1470 rows × 35 columns

From the code above we learn that this particular dataset has 35 different features (columns) and 1470 different observations (rows). The different features inlcude the following.

Age:----------The age of the employee. (Numerical)
Attrition:----------Whether or not the employee has left the organization. (Categorical)
BusinessTravel:----------The frequency of business travel for the employee. (Categorical)
DailyRate:----------The daily rate of pay for the employee. (Numerical)
Department:----------The department the employee works in.

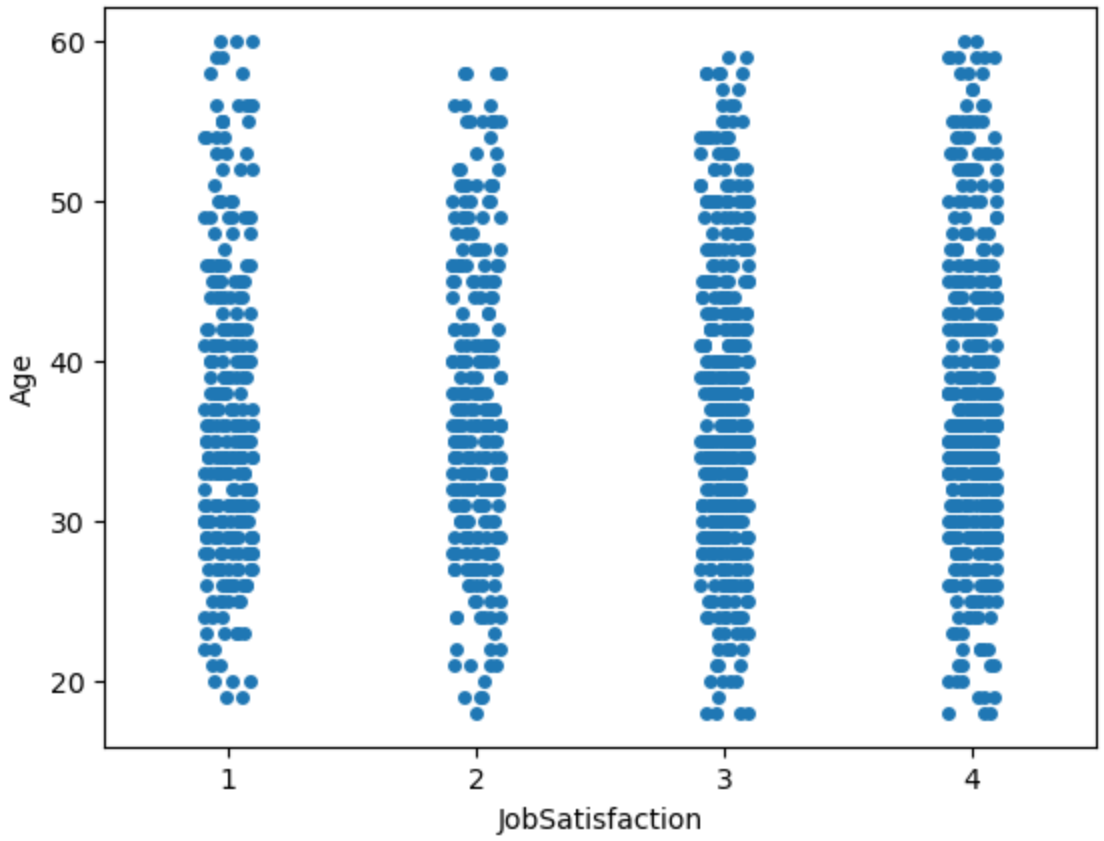
```
(Categorical)
DistanceFromHome:---------The distance from home in miles for the
employee. (Numerical)
Education:----------------The level of education achieved by the
employee. (Categorical)
EducationField:-----------The field of study for the employee's
education. (Categorical)
EmployeeCount:------------The total number of employees in the
organization. (Numerical)
EmployeeNumber:-----------A unique identifier for each employee
profile. (Numerical)
EnvironmentSatisfaction:--The employee's satisfaction with their work
environment. (Categorical)
Gender:-------------------The gender of the employee. (Categorical)
HourlyRate:---------------The hourly rate of pay for the employee.
(Numerical)
JobInvolvement:-----------The level of involvement required for the
employee's job. (Categorical)
JobLevel:-----------------The job level of the employee.
(Categorical)
JobRole:------------------The role of the employee in the
organization. (Categorical)
JobSatisfaction:----------The employee's satisfaction with their job.
(Categorical)
MaritalStatus:------------The marital status of the employee.
(Categorical)
MonthlyIncome:------------The monthly income of the employee.
(Numerical)
MonthlyRate:--------------The monthly rate of pay for the employee.
(Numerical)
NumCompaniesWorked:-------The number of companies the employee has
worked for. (Numerical)
Over18:-------------------Whether or not the employee is over 18.
(Categorical)
OverTime:-----------------Whether or not the employee works overtime.
(Categorical)
PercentSalaryHike:--------The percentage of salary hike for the
employee. (Numerical)
PerformanceRating:--------The performance rating of the employee.
(Categorical)
RelationshipSatisfaction:-The employee's satisfaction with their
relationships. (Categorical)
StandardHours:------------The standard hours of work for the
employee. (Numerical)
StockOptionLevel:---------The stock option level of the employee.
(Numerical)
TotalWorkingYears:--------The total number of years the employee has
worked. (Numerical)
TrainingTimesLastYear:----The number of times the employee was taken
for training in the last year. (Numerical)
WorkLifeBalance:----------The employee's perception of their work-
life balance. (Categorical)
YearsAtCompany:-----------The number of years the employee has been
```

with the company. (Numerical)
YearsInCurrentRole:-------The number of years the employee has been in their current role. (Numerical)
YearsSinceLastPromotion:--The number of years since the employee's last promotion. (Numerical)
YearsWithCurrManager:-----The number of years the employee has been with their current manager. (Numerical)

For the sake of our analysis, out target varaible will be JobSatisfaction since we are trying to see what other features affect the amount of job satisfaction an employee feels in their career. Not all of these features will be used in the model building process so we first must see if we can decide what features affect our target variable the most if at all.

In [8]:
```
#Graph 1: JobSatisfaction vs Age
import seaborn as sns
sns.stripplot(x= "JobSatisfaction", y= "Age", data=hrdata)
```

Out[8]:
```
<Axes: xlabel='JobSatisfaction', ylabel='Age'>
```

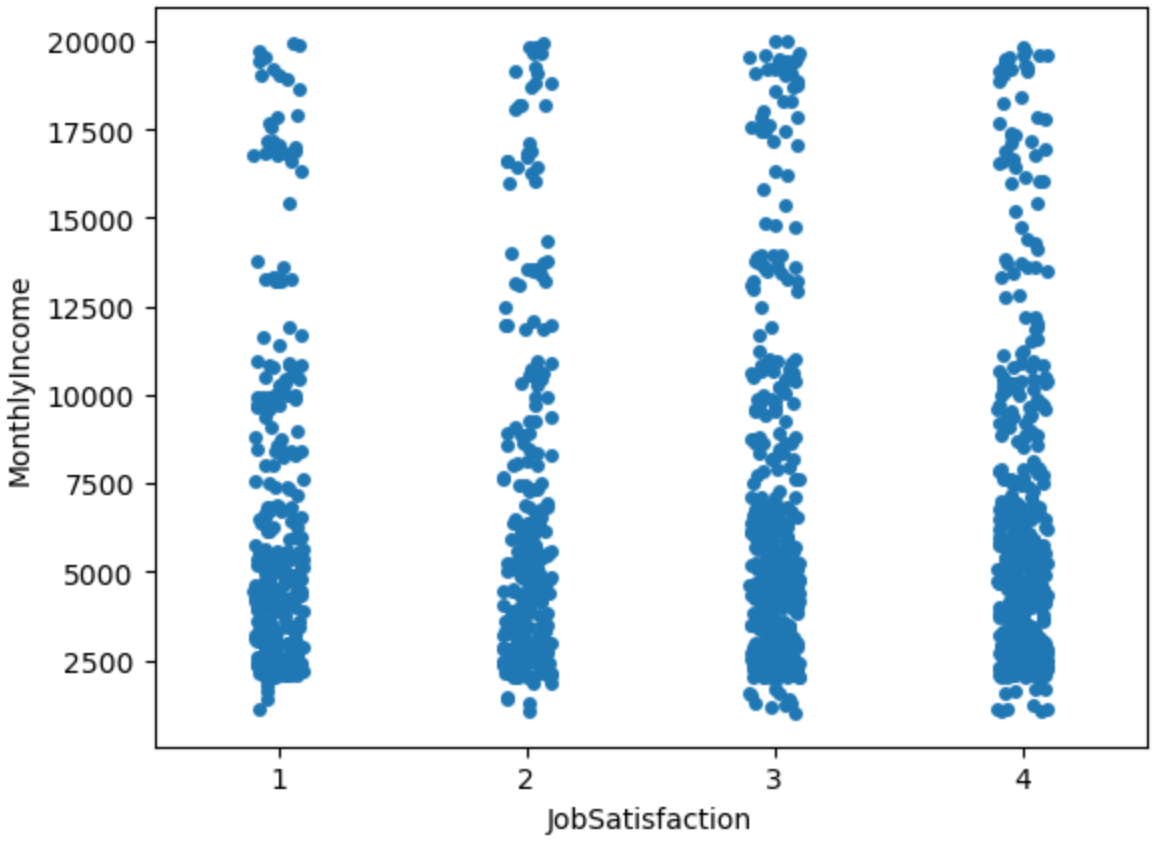


Graph 1: Explaination:

Looking at the relationship between job satisfaction and employee age, the two higher job satisfaction ratings have a larger density of employees around the age of 25 to 45. It's not a super strong relationship and it's not obvious so perhaps we'll have to dive into these relationships further.

In [9]:
```
#Graph 2: MonthlyIncome vs JobSatisfaction (monetary motivation? Could influence perfo
sns.stripplot(x= "JobSatisfaction", y= "MonthlyIncome", data=hrdata)
```

Out[9]:
```
<Axes: xlabel='JobSatisfaction', ylabel='MonthlyIncome'>
```

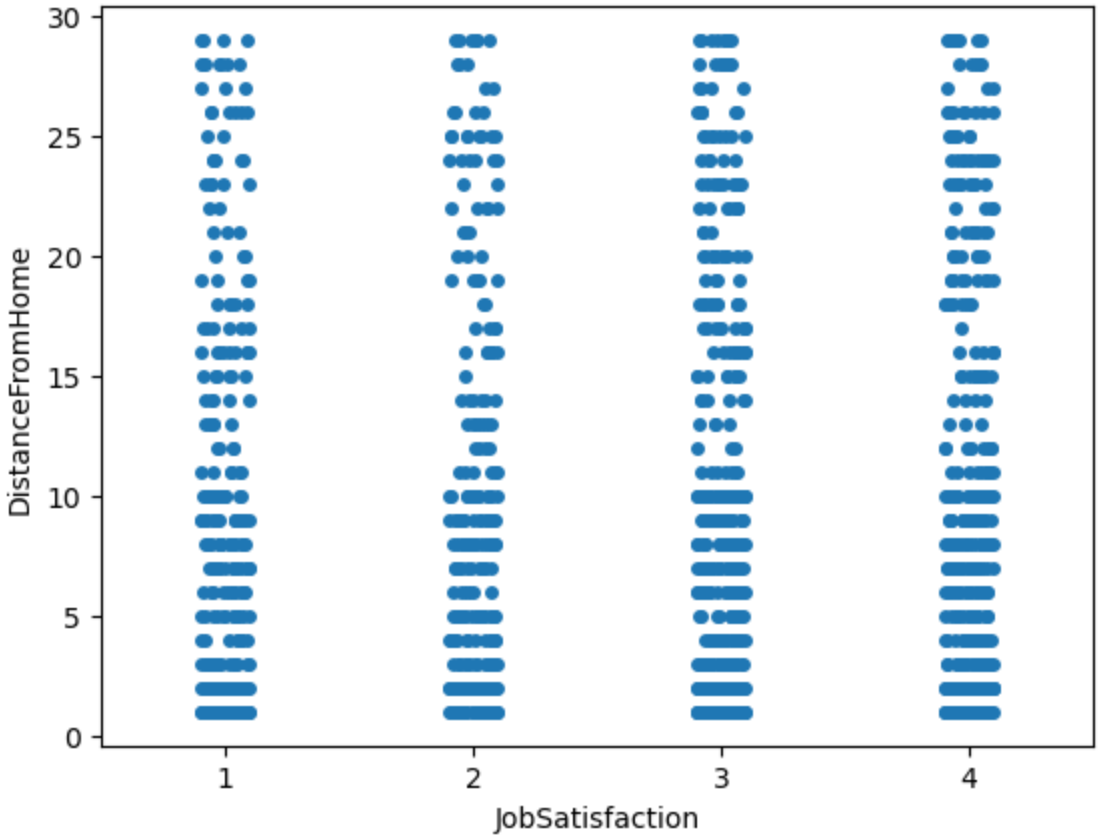


Graph 2: Explaination:

As the monthly income starts to increase we can see a slight increase in density for the higher job satisfaction ratings. This graph might have to be checked for outliers during further analysis due to the fact that salaries are often skewed to the right. It might also be useful to only use data of "average" employees.

In [10]:
```
#Graph 3: JobSatisfaction vs DistanceFromHome (does less hours improve performance?)
sns.stripplot(x= "JobSatisfaction", y= "DistanceFromHome", data=hrdata)
```

Out[10]:
```
<Axes: xlabel='JobSatisfaction', ylabel='DistanceFromHome'>
```

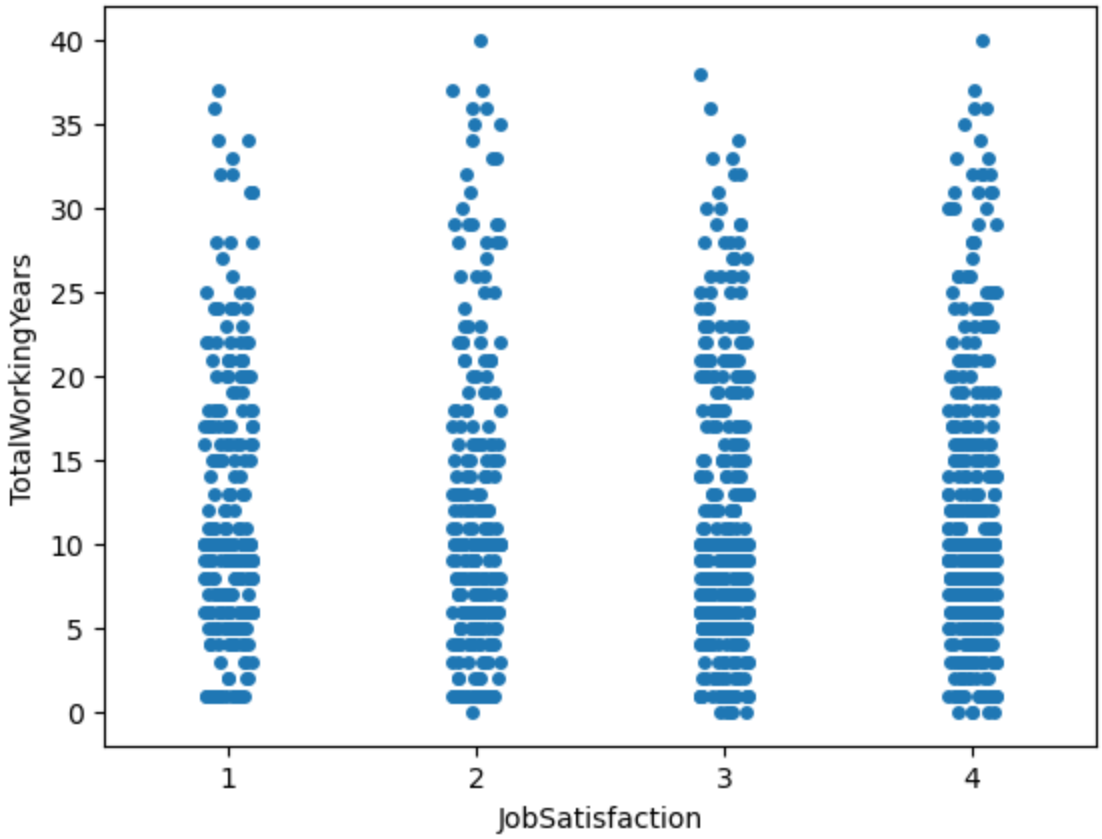


Graph 3: Explaination:

This graph is interesting considering I would've guessed that the closer you are to your job the more job satisfaction you'd have considering the fact that you wouldn't have to sit through more traffic before and after your shift but depending on location, traffic might not even be a deciding factor in how mch you enjoy your job. But looking at the graph, job satisfaction seems pretty constant despite difference in distance from home.

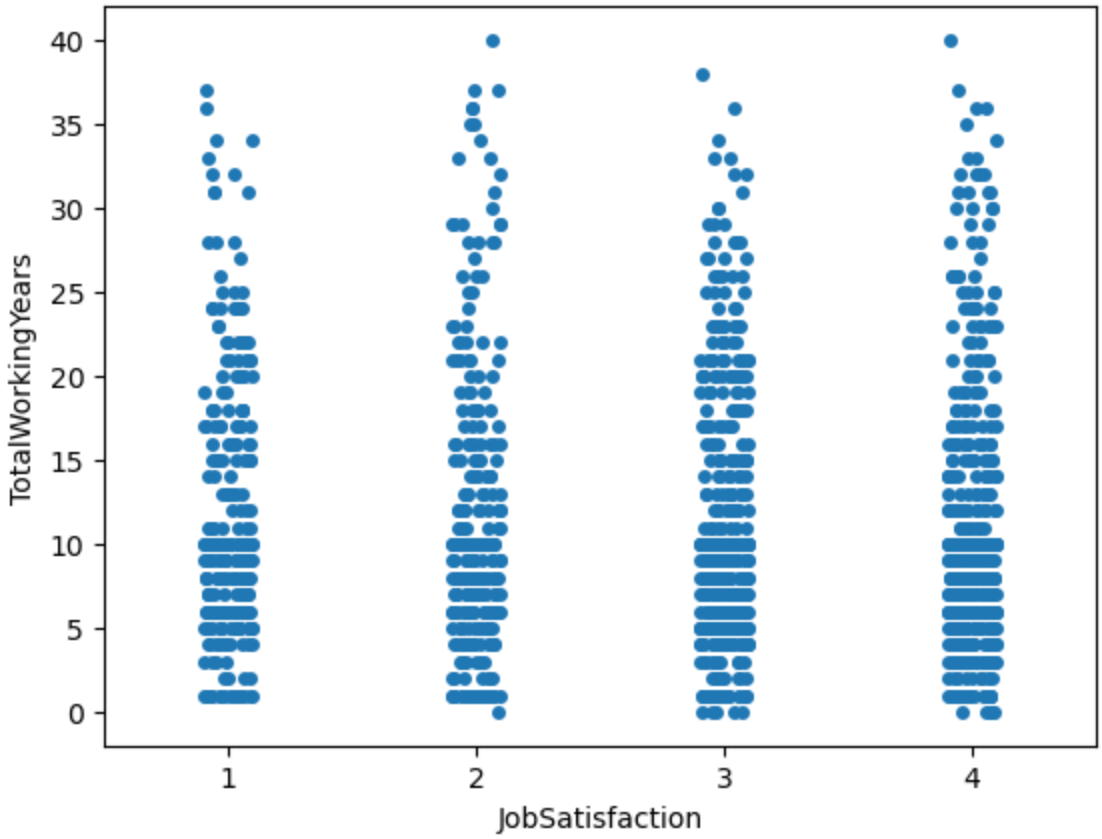
In [11]:
```
#Graph 4: JobSatisfaction vs TotalWorkingYears (does traffic and commnity affect JobSa
sns.stripplot(x= "JobSatisfaction", y= "TotalWorkingYears", data=hrdata)
```

Out[11]:
```
<Axes: xlabel='JobSatisfaction', ylabel='TotalWorkingYears'>
```



In [12]:
```
#Graph 4: JobSatisfaction vs TotalWorkingYears (does traffic and commnity affect JobSa
sns.stripplot(x= "JobSatisfaction", y= "TotalWorkingYears", data=hrdata)
```

Out[12]:
```
<Axes: xlabel='JobSatisfaction', ylabel='TotalWorkingYears'>
```



Graph 4: Explaination:

If you look at the graph, you can see a slight increase in density as the amount of years worked increases. You have more observations of higher job satisfaction ratings the more years an employee has worked. This could be due to the fact that these people have more experience either in their current job or exploring different job options and now know what works for them. It'll be interesting looking further into this.

Conclusion of Graphical Analysis:

As of right now, our analysis has showed us that job satisfaction has some kind of positive realtionship with total working years, monthly income, and employee age. Of course this is not an extensive list and further analysis is still needed within those observations. I would also like focusing my analysis on specific subgroups of the data to see if any new relationships become visible in that way.

Data Preparation:

In the data preparation portion of the analysis, we more so look for problems in the data. First we look at the different data types that we're working with and make sure that each feature's data type is consistent with what we expect.

In [13]:
```
#Variable names and data types
result = hrdata.dtypes

print("Output:")
print(result)
```

```
Output:
Age                        int64
Attrition                 object
BusinessTravel            object
DailyRate                  int64
Department                object
DistanceFromHome           int64
Education                  int64
EducationField            object
EmployeeCount              int64
EmployeeNumber             int64
EnvironmentSatisfaction    int64
Gender                    object
HourlyRate                 int64
JobInvolvement             int64
JobLevel                   int64
JobRole                   object
JobSatisfaction            int64
MaritalStatus             object
MonthlyIncome              int64
MonthlyRate                int64
NumCompaniesWorked         int64
Over18                    object
OverTime                  object
PercentSalaryHike          int64
PerformanceRating          int64
RelationshipSatisfaction   int64
StandardHours              int64
StockOptionLevel           int64
TotalWorkingYears          int64
TrainingTimesLastYear      int64
WorkLifeBalance            int64
YearsAtCompany             int64
YearsInCurrentRole         int64
YearsSinceLastPromotion    int64
YearsWithCurrManager       int64
dtype: object
```

After looking at each feature's data types and getting a good idea of some relationships between those features and our target varaible JobSatisfaction, now it's time to decide what features we want to keep in our dataset that might be helpful in our model building stage later.

First we want to subset the data for employees that are still working at their current job since this can add some bias towards satisfaction but also the goal is to improve job satisfaction within an company and then we want to get rid of that variable

In [14]:
```
#removing employees that aren't working for the company anymore

hrdataa = hrdata[hrdata['Attrition'] == "No"]
hrdataa = hrdataa.drop('Attrition', axis=1)

print(hrdataa)
```

```
      Age     BusinessTravel  DailyRate              Department  \
1      49  Travel_Frequently        279  Research & Development
3      33  Travel_Frequently       1392  Research & Development
4      27      Travel_Rarely        591  Research & Development
5      32  Travel_Frequently       1005  Research & Development
6      59      Travel_Rarely       1324  Research & Development
...   ...                ...        ...                     ...
1465   36  Travel_Frequently        884  Research & Development
1466   39      Travel_Rarely        613  Research & Development
1467   27      Travel_Rarely        155  Research & Development
1468   49  Travel_Frequently       1023                   Sales
1469   34      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
1                    8          1  Life Sciences              1
3                    3          4  Life Sciences              1
4                    2          1        Medical              1
5                    2          2  Life Sciences              1
6                    3          3        Medical              1
...                ...        ...            ...            ...
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  EnvironmentSatisfaction  ... RelationshipSatisfaction  \
1                  2                        3  ...                        4
3                  5                        4  ...                        3
4                  7                        1  ...                        4
5                  8                        4  ...                        3
6                 10                        3  ...                        1
...              ...                      ...  ...                      ...
1465            2061                        3  ...                        3
1466            2062                        4  ...                        1
1467            2064                        2  ...                        2
1468            2065                        4  ...                        4
1469            2068                        2  ...                        1

      StandardHours  StockOptionLevel  TotalWorkingYears  \
1                80                 1                 10
3                80                 0                  8
4                80                 1                  6
5                80                 0                  8
6                80                 3                 12
...             ...               ...                ...
1465             80                 1                 17
1466             80                 1                  9
1467             80                 1                  6
1468             80                 0                 17
1469             80                 0                  6

     TrainingTimesLastYear  WorkLifeBalance YearsAtCompany  \
1                        3                3             10
3                        3                3              8
4                        3                3              2
5                        2                2              7
6                        3                2              1
...                    ...              ...            ...
1465                     3                3              5
```

```
1466                       5                        3                        7
1467                       0                        3                        6
1468                       3                        2                        9
1469                       3                        4                        4

      YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
1                      7                        1                     7
3                      7                        3                     0
4                      2                        2                     2
5                      7                        3                     6
6                      0                        0                     0
...                  ...                      ...                   ...
1465                   2                        0                     3
1466                   7                        1                     7
1467                   2                        0                     3
1468                   6                        0                     8
1469                   3                        1                     2

[1233 rows x 34 columns]
```

Next we want to drop some other features just based on their background and description we have. For example, there are three different ways in which an employee's compensation is accounted for. DailyRate, HourlyRate, and MonthlyRate. DailyRate and HourlyRate will be dropped because the data also includes monthly rate. Because MonthlyRate and MonthlyIncome (total income including other incomes aside from this job) are techically different, we will keep both of them. Because part of the purpose of this analysis is to help employers see what they can possibly provide for employees in order to increase job satisfaction, we will also drop MaritalStatus because that is out of the control of what companies can do for their employees. We will also drop PerformanceRating because the answers all seemed unimportant (everyone was either 3 or 4 so there wasn't enough significance in the differences). Lastly, we drop RelationshipSatisfaction for the same reason as MaritalStatus.

In [15]:
```
#Dropping features listed above

hrdataa = hrdataa.drop(['DailyRate', 'HourlyRate', 'MaritalStatus', 'PerformanceRating
                        'RelationshipSatisfaction'], axis=1)

result = hrdataa.dtypes

print("Output:")
print(result)
```

```
Output:
Age                         int64
BusinessTravel             object
Department                 object
DistanceFromHome            int64
Education                   int64
EducationField             object
EmployeeCount               int64
EmployeeNumber              int64
EnvironmentSatisfaction     int64
Gender                     object
JobInvolvement              int64
JobLevel                    int64
JobRole                    object
JobSatisfaction             int64
MonthlyIncome               int64
MonthlyRate                 int64
NumCompaniesWorked          int64
Over18                     object
OverTime                   object
PercentSalaryHike           int64
StandardHours               int64
StockOptionLevel            int64
TotalWorkingYears           int64
TrainingTimesLastYear       int64
WorkLifeBalance             int64
YearsAtCompany              int64
YearsInCurrentRole          int64
YearsSinceLastPromotion     int64
YearsWithCurrManager        int64
dtype: object
```

After deciding which varaibles we're going to keep in our final model, next we want to convert the categorical variables to dummy varaibles. We can check which variables are supposed to be categorical variables in the results above as well as which variables are supposed to be numerical. The categorical variable will have to be converted into dummy varaible for them to be used in any type of model building process.

Now usually we wouldn't use 0,1 and 2 as dummy values for a varaible when creating dummy varaibles but since the BusinessTravel varaible has rankable categories. In this case the bigger the number the more business travel was required in that particular job. Also notice that we had to account for an extra value in Over18. This was because there was an error in the data input. Most observations were either "Yes" or "No" but one was simply "Y" meaning yes so we had to account for that when changing the varaible to a dummy variable.

For the categorical variables that we wouldn't convert to dummy variables manually, we were able to use the get_dummies attribute for pandas dataframes and convert those variables that way.

In [16]:
```
#Create dummy variables if necessary.
#changing Business Travel - since this is quantifiable/rankable we will simply change
hrdataa['BusinessTravel'] = hrdataa['BusinessTravel'].replace(['Non-Travel'], 0)
```

```
hrdataa['BusinessTravel'] = hrdataa['BusinessTravel'].replace(['Travel_Rarely'], 1)
hrdataa['BusinessTravel'] = hrdataa['BusinessTravel'].replace(['Travel_Frequently'], 2

#Changing Gender to Binary
hrdataa['Gender'] = hrdataa['Gender'].replace(['Male'], 0)
hrdataa['Gender'] = hrdataa['Gender'].replace(['Female'], 1)

#Changing Over18 to Binary
hrdataa['Over18'] = hrdataa['Over18'].replace(['No'], 0)
hrdataa['Over18'] = hrdataa['Over18'].replace(['Yes'], 1)
hrdataa['Over18'] = hrdataa['Over18'].replace(['Y'], 1)

#Changing OverTime to Binary
hrdataa['OverTime'] = hrdataa['OverTime'].replace(['No'], 0)
hrdataa['OverTime'] = hrdataa['OverTime'].replace(['Yes'], 1)


#We have to do Department and EducationField using a different method

hrdataadum = pd.get_dummies(hrdataa)

result = hrdataadum.dtypes
print("Output:")
print(result)
print(hrdataadum)
```

```
Output:
Age                                int64
BusinessTravel                     int64
DistanceFromHome                   int64
Education                          int64
EmployeeCount                      int64
EmployeeNumber                     int64
EnvironmentSatisfaction            int64
Gender                             int64
JobInvolvement                     int64
JobLevel                           int64
JobSatisfaction                    int64
MonthlyIncome                      int64
MonthlyRate                        int64
NumCompaniesWorked                 int64
Over18                             int64
OverTime                           int64
PercentSalaryHike                  int64
StandardHours                      int64
StockOptionLevel                   int64
TotalWorkingYears                  int64
TrainingTimesLastYear              int64
WorkLifeBalance                    int64
YearsAtCompany                     int64
YearsInCurrentRole                 int64
YearsSinceLastPromotion            int64
YearsWithCurrManager               int64
Department_Human Resources         uint8
Department_Research & Development  uint8
Department_Sales                   uint8
EducationField_Human Resources     uint8
EducationField_Life Sciences       uint8
EducationField_Marketing           uint8
EducationField_Medical             uint8
EducationField_Other               uint8
EducationField_Technical Degree    uint8
JobRole_Healthcare Representative  uint8
JobRole_Human Resources            uint8
JobRole_Laboratory Technician      uint8
JobRole_Manager                    uint8
JobRole_Manufacturing Director     uint8
JobRole_Research Director          uint8
JobRole_Research Scientist         uint8
JobRole_Sales Executive            uint8
JobRole_Sales Representative       uint8
dtype: object
      Age  BusinessTravel  DistanceFromHome  Education  EmployeeCount  \
1      49               2                 8          1              1
3      33               2                 3          4              1
4      27               1                 2          1              1
5      32               2                 2          2              1
6      59               1                 3          3              1
...   ...             ...               ...        ...            ...
1465   36               2                23          2              1
1466   39               1                 6          1              1
1467   27               1                 4          3              1
1468   49               2                 2          3              1
1469   34               1                 8          3              1

      EmployeeNumber  EnvironmentSatisfaction  Gender  JobInvolvement  \
```

```
1                2                       3       0              2
3                5                       4       1              3
4                7                       1       0              3
5                8                       4       0              3
6               10                       3       1              4
...            ...                     ...     ...            ...
1465           2061                      3       0              4
1466           2062                      4       0              2
1467           2064                      2       0              4
1468           2065                      4       0              2
1469           2068                      2       0              4

      JobLevel  ...  EducationField_Technical Degree  \
1            2  ...                                0
3            1  ...                                0
4            1  ...                                0
5            1  ...                                0
6            1  ...                                0
...        ...  ...                              ...
1465         2  ...                                0
1466         3  ...                                0
1467         2  ...                                0
1468         2  ...                                0
1469         2  ...                                0

      JobRole_Healthcare Representative  JobRole_Human Resources  \
1                                     0                        0
3                                     0                        0
4                                     0                        0
5                                     0                        0
6                                     0                        0
...                                 ...                      ...
1465                                  0                        0
1466                                  1                        0
1467                                  0                        0
1468                                  0                        0
1469                                  0                        0

      JobRole_Laboratory Technician  JobRole_Manager  \
1                                 0                0
3                                 0                0
4                                 1                0
5                                 1                0
6                                 1                0
...                             ...              ...
1465                              1                0
1466                              0                0
1467                              0                0
1468                              0                0
1469                              1                0

      JobRole_Manufacturing Director  JobRole_Research Director  \
1                                  0                         0
3                                  0                         0
4                                  0                         0
5                                  0                         0
6                                  0                         0
...                              ...                       ...
1465                               0                         0
1466                               0                         0
```

```
1467                              1                          0
1468                              0                          0
1469                              0                          0

      JobRole_Research Scientist  JobRole_Sales Executive  \
1                              1                         0
3                              1                         0
4                              0                         0
5                              0                         0
6                              0                         0
...                          ...                       ...
1465                           0                         0
1466                           0                         0
1467                           0                         0
1468                           0                         1
1469                           0                         0

      JobRole_Sales Representative
1                                0
3                                0
4                                0
5                                0
6                                0
...                            ...
1465                             0
1466                             0
1467                             0
1468                             0
1469                             0

[1233 rows x 44 columns]
```

Now once we have chosen our varaibles and made the necessary changes to our data, the last thing that is left is to make sure that there aren't any missing values in our data and then we are able to move onto the model building process.

In [17]:
```
#Deal with missing data (do not just drop rows or columns without justifying this).
# Applying the method
check_nan = hrdataadum.isnull().values.any()

# printing the result
print(check_nan)
```

```
False
```

Model building and Evaluation:

Before we start building our model, we first want to separate our dataset into training and testing sets so we are able to evaluate our model later one. For this analysis the training set will be 80% of the original dataset and then 20% of the data goes to the testing set which will be used in order to compare to the model predictions we get. In previous models, we used 75% of the data for the training set but decided to increase the size of the training set with the hopes of training a better and more accurate model.

In [18]:
```
#Importing necessary packages to create our model and splitting data into training and
from sklearn.model_selection import train_test_split

X = hrdataadum.loc[:, hrdataadum.columns != "JobSatisfaction"]
y = hrdataadum.loc[:, ['JobSatisfaction']]

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, train_size =
```

The first round of model building, we used a linear regression model and treated our target variable as a numerical one. But this time around we tried using a logistic regression model and treated our target varaible as a categorical variable once again since the values are either 1, 2, 3 or 4. We then fit our logistic regression model with our training data.

In [20]:
```
#Package for linear regression model and training model with trianing set
from sklearn.linear_model import LogisticRegression
logmod = LogisticRegression()
logmod.fit(X_train, y_train)
```

```
C:\Users\nickx\anaconda3\lib\site-packages\sklearn\utils\validation.py:1143: DataConv
ersionWarning: A column-vector y was passed when a 1d array was expected. Please chan
ge the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
C:\Users\nickx\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:458: Con
vergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[20]:
```
▾ LogisticRegression

LogisticRegression()
```

Once the model has been fit, we are able to get predictions from this model by plugging in our testing x values to the model.

In [21]:
```
#Using our model to get predictions
import sklearn.metrics as metrics
predvals = logmod.predict(X_test)
y = y_test
y = y.to_numpy()
yhat = predvals
```

Finally, once we have our predicted values from our model, we are able to evalate our model by comparing our testing y values to the predicted values and getting our accuracy score. We end up getting an accuracy score of about 0.3117. This means that our model predicted the correct JobSatisfaction ranking 31.17% of the time which it

better than random guessing (25% chance of guessing correctly since there are four possible values to choose from). This is a major improvement from our previous model. While there was no measure of accuracy

In [22]:
```
#Model Evaluation
from sklearn.metrics import accuracy_score
print(accuracy_score(y, predvals))
```

```
0.3117408906882591
```

In [24]:
```
confusion_matrix = metrics.confusion_matrix(y, predvals)
confusion_matrix
```

Out[24]:
```
array([[ 0,  0, 17, 31],
       [ 0,  0, 19, 22],
       [ 0,  0, 28, 46],
       [ 0,  0, 35, 49]], dtype=int64)
```

Looking at the confusion matrix we can see that the model struggles in predicting the 1 and 2 values for the JobSatisfaction rankings. This information can possibly help us in creating a better model in the future.

Conclusion:

What does the analysis/model building tell you?

Overall, the model tells us that a company can possibly predict their emploee's job satisfaction based on those varaibles that we decided to keep in our model with an accuracy of around 31%. While this might not sound like too good of a model based on that accuracy, it's still better than randomly guessing (25% chance of guessing correctly).

Is this model ready to be deployed?

While the model does theoretically perform better than just randomly guessing the results, we don't believe that this model is ready to be deployed and used in business practices. The model struggles when predicting the lower ranking values so that is an area that should be inproved before being deployed. While this model isn't necessarily tied with business profits, it might also be better to strive for a model that has a higher accuracy score. Somewhere around 70% accuracy would be something to strive for and something a company might consider with more confidence.

What are your recommendations?

As stated before, I would recommend including more data that includes more JobSatisfaction values of 1 and 2. This could possibly help in raising that accuracy score for those values as we saw that was a

struggle in the confusion matrix. Not only the accuracy for those values but the accuracy score in general as well.

What are some of the potential challenges

Some potential challenges could be replication of this data. This data included observations of different employees from a single company. Ideally this information could be gathered from different companies. A larger amount of observations would also be ideal but making sure that there are enough of each ranking values in the data in order to improve the accuracy of our model across all values might be a challenge as well. Perhaps there should be a different ranking system to be put in place. Depending on who is collecting the data and with what methods they are using can also affect the integrity of the data which is also a challenge. Is it the employers and bosses collecting this data? Do the employee's remain anonymous while providing their information? These are some questions to consider which might affect how truthful employees are when providing the information provided in the dataset.

Are there additional opportunities to be explored?

Naturally there are always additonal opportunities to be explored. As I mentioned, something more on a larger scale can be greatly beneficial in improving the working enjoronment and job satisfaction employees feel in their careers. Being able to pinpoint what type of variables are associated with a person's job satisfaction can be beneficial to both the employees as well as the employers. More variables can be added, possibly employee benefits. Companies can then see what kind of benefits they can provide for their employees that would drive up job satisfaction and in turn lower job turnover and increase productivity.