Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

Employee Attrition Analysis

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation

For the Internship Project work done during
DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by
Rajeshwari Nalbalwar (TCRIB2R131)
Omkar Kalekar (TCRIB2R134)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Preparation of Papers in Two-Column Format for Conference Proceedings Published by IEEE[1]

Bart Simpson and Homer Simpson
Department of Nuclear Power Engineering
University of Springfield
Springfield, Nostate 12345, USA
{bart.simpson & homer.simpson}@uspringfield.edu

Monkey King, Bajie Zhu and Seng Tang
Department of Intelligent Robotics
University of Huaguoshan
Huaguoshan, Jileshijie Province, China
monkey.king@uhuaguoshan.edu.cn

[1] This work is partially supported by NSF Grant #2003168 to H. Simpson and CNSF Grant #9972988 to M. King.

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

Employee Attrition Analysis

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation

For the Internship Project work done during
DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by
Rajeshwari Nalbalwar (TCRIB2R131)
Omkar Kalekar (TCRIB2R134)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Preparation of Papers in Two-Column Format for Conference Proceedings Published by IEEE[1]

Bart Simpson and Homer Simpson
Department of Nuclear Power Engineering
University of Springfield
Springfield, Nostate 12345, USA
{bart.simpson & homer.simpson}@uspringfield.edu

Monkey King, Bajie Zhu and Seng Tang
Department of Intelligent Robotics
University of Huaguoshan
Huaguoshan, Jileshijie Province, China
monkey.king@uhuaguoshan.edu.cn

[1] This work is partially supported by NSF Grant #2003168 to H. Simpson and CNSF Grant #9972988 to M. King.

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022
Aim-. To analyse employee attrition.

I. Abstract

Comprehending the most likely reasons for employees leaving, will aid the organization to take appropriate actions to reduce the level of Attrition.
We have constructed a very simple pipeline of predicting employee attrition, from some basic Exploratory Data Analysis to feature engineering as well as implementing one learning model in the form of a Logistic Regression model. This entire notebook takes less than a minute to run and it actually returns 90% accuracy in its predictions.

II. INTRODUCTION

Our aim is to study employee attrition in an organisation, analyse the reasonings and predict future attrition.
We used an existing database containing employee information and their attrition rates.
After reading the given CSV file we performed various Numpy and Pandas operations to find null values, describe the table, etc.
Then after, we performed data visualization in which we created various tables like 'count of age', 'count of the gender of employees', 'Business Travel and attrition', 'dependency of overtime attrition', 'dependency of RelationshipSatisfaction',etc.
We also created pie-charts for analysis of various fields like 'education', 'business travel', 'job role', 'employee marital info', etc.
Our project is also capable of analysing continuous values as we have created a Histogram for distribution.
Once we performed data analysis we created a machine learning model which is capable of predicting the reasoning and attrition rate with high accuracy.

III. Theory:

Python:
Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small- and large-scale projects.[1]

Numpy:
NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. NumPy is a Python package. It stands for 'Numerical Python'. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.[2]

Pandas:
Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.
pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.[3]

Matplotlib:
Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Seaborn:
Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

Name- Rajeshwari Nalbalwar
Internship Program- DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM
Batch- JAN
Certificate Code- TCRIB2R131
Date of submission- 4.4.2022

Machine learning algorithm:

Linear Regression:
Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.[4]

Techniques used:

Data Preprocessing:
Data preprocessing is a step in the data mining and data analysis process that takes raw data. and transforms it into a format that can be understood and analysed by computers and machine learning. Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design. Machines like to process nice and tidy information - they read data as 1s and Os. So, calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.[5]

Data Visualization:
Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends, and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics. Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.[6]

IV. METHODOLOGY:

We need to perform the following steps:
1. Import the relevant Packages and Libraries.
2. Download and explore the datasheet.
3. Perform EDA, apply a dataset for processing.
4. Predict the target columns.

V. RESULTS:

Top Reasons why Employees leave the Organization:

1. No Overtime: This is an unexpected cause. According to the performed studies, the employees not working overtime are more likely to leave the organization. An explanation for this could be, employees supposing their work less valued or not up to the mark.
2. Monthly Income: As predicted, income is the 2nd most reason for employees leaving the organization. Getting paid with deserved wedges is expected by everyone. Not getting paid a merited salary, bonuses, etc. on time makes an employee vacate.
3. Age: This again is very expected reasoning. This is because of the employees seeking to retire. These employees will leave the organization.

VI. Output:

(Result Photos on last pages)

Precautions to be followed
1. Make sure the data is complete when uploaded.
2. All the sequential steps must be run one after the other. Otherwise, errors might occur.

VII. Conclusion :

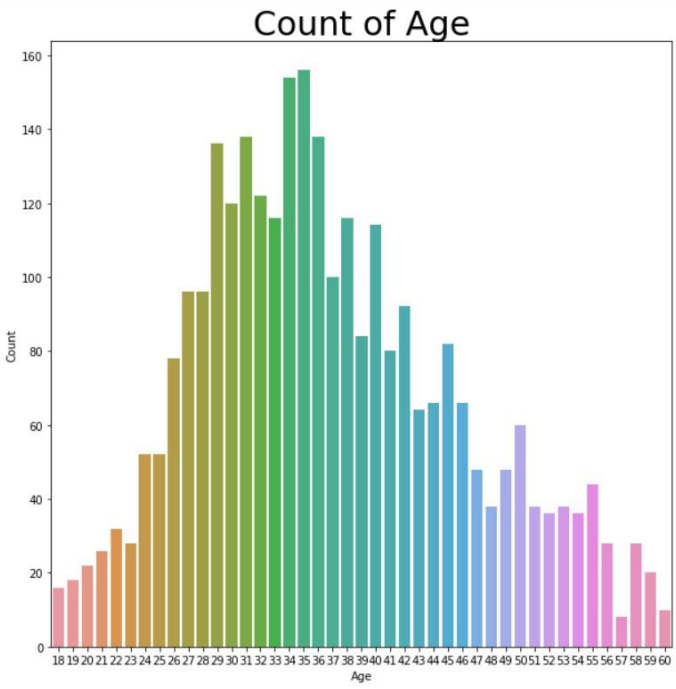
Our project "EMPLOYEE ATTRITION ANALYSIS" analyzes and studies various aspects of the reasonings of the employees leaving the organization with the help of linear regression model.

VIII. References:

1. *Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python*

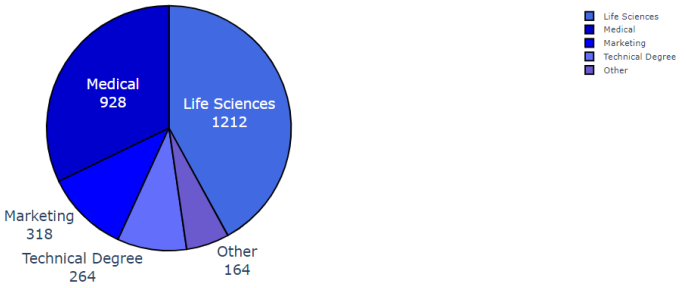
Exercises". *Section 1.1. Archived from the original (PDF) on 23 June 2012.*

2. Charles R Harris; K. Jarrod Millman; Stéfan J. van der Walt; et al. (16 September 2020). "Array programming with NumPy" (PDF). *Nature*. **585** (7825): 357–362. doi:10.1038/S41586-020-2649-2. ISSN 1476-4687. PMC 7759461. PMID 32939066. Wikidata Q99413970.
3. "License – Package overview – pandas 1.0.0 documentation". *pandas*. 28 January 2020. Retrieved 30 January 2020.
4. David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 26. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right hand side, each with its own slope coefficient.
5. "Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data". *Tableau*. Retrieved 2021-10-17.
6. Nussbaumer Knaflic, Cole (2 November 2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. ISBN 978-1-119-00225-3.



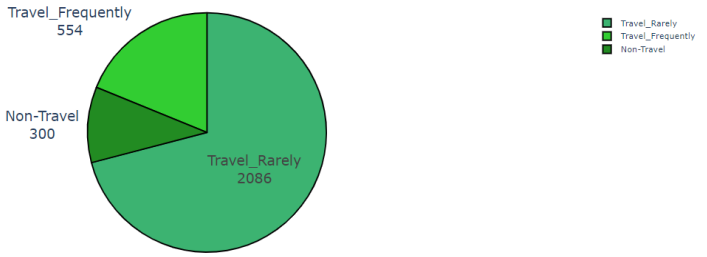
EducationField

```
[ ]  colors = ['royalblue ', 'mediumblue ', 'blue ','arine','slateblue']

     fig = go.Figure(data=[go.Pie(labels=['Life Sciences','Medical','Marketing', 'Technical Degree', 'Other', 'Human Resources'],
                                  values=[1212,928,318,264,164])])
     fig.update_traces(hoverinfo='label+percent', textinfo='label + value', textfont_size=20,
                       marker=dict(colors=colors, line=dict(color='#000000', width=2)))
     fig.show()
```



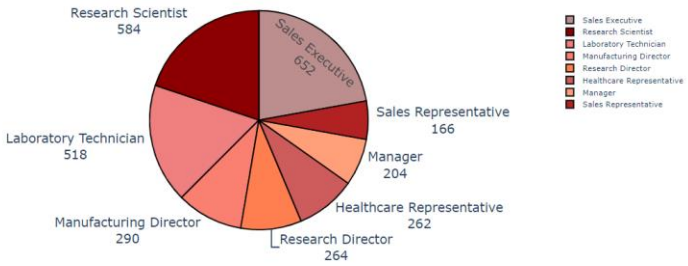
BusinessTravel

```
colors = ['mediumseagreen','limegreen', 'forestgreen']

fig = go.Figure(data=[go.Pie(labels=['Travel_Rarely','Travel_Frequently','Non-Travel'],
                             values=[2086,554,300])])
fig.update_traces(hoverinfo='label+percent', textinfo='label + value', textfont_size=20,
                  marker=dict(colors=colors, line=dict(color='#000000', width=2)))
fig.show()
```



JOB roles

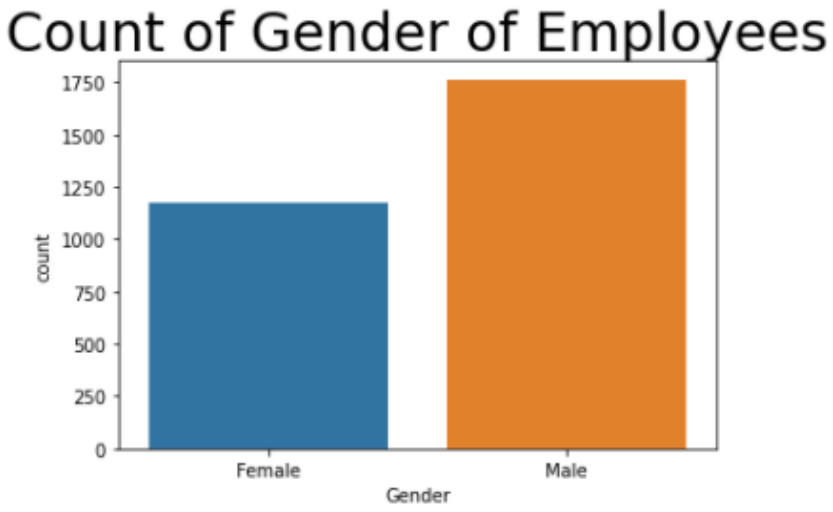
```
colors = ['rosybrown', 'darkred', 'lightcoral', 'salmon', 'indianred', 'lightsalmon', 'firebrick', 'coral']

fig = go.Figure(data=[go.Pie(labels=['Sales Executive','Research Scientist','Laboratory Technician', 'Manufacturing Director', 'Healthcare Representative', 'Manager', 'Sales Representative', 'Research Director', 'Research Director', 'Human Resources
                             values=[652,584,518,290,262,204,166,160,104])])
fig.update_traces(hoverinfo='label+percent', textinfo='label + value', textfont_size=20,
                  marker=dict(colors=colors, line=dict(color='#000000', width=2)))
fig.show()
```



Gender

```
[ ]  sns.countplot('Gender', data=df)
     plt.title('Count of Gender of Employees', fontsize=30)
     plt.xlabel('Gender')
     plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:

Pass the following variable as a keyword arg: x. From version 0.12, the only valid pos
```

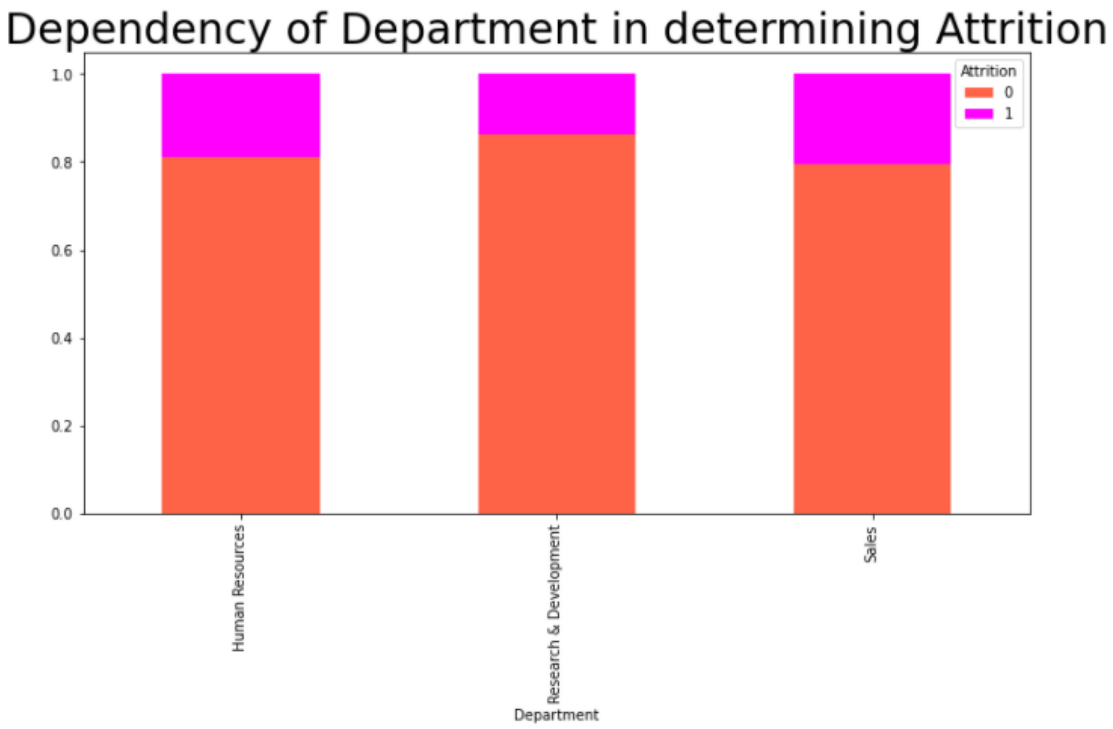


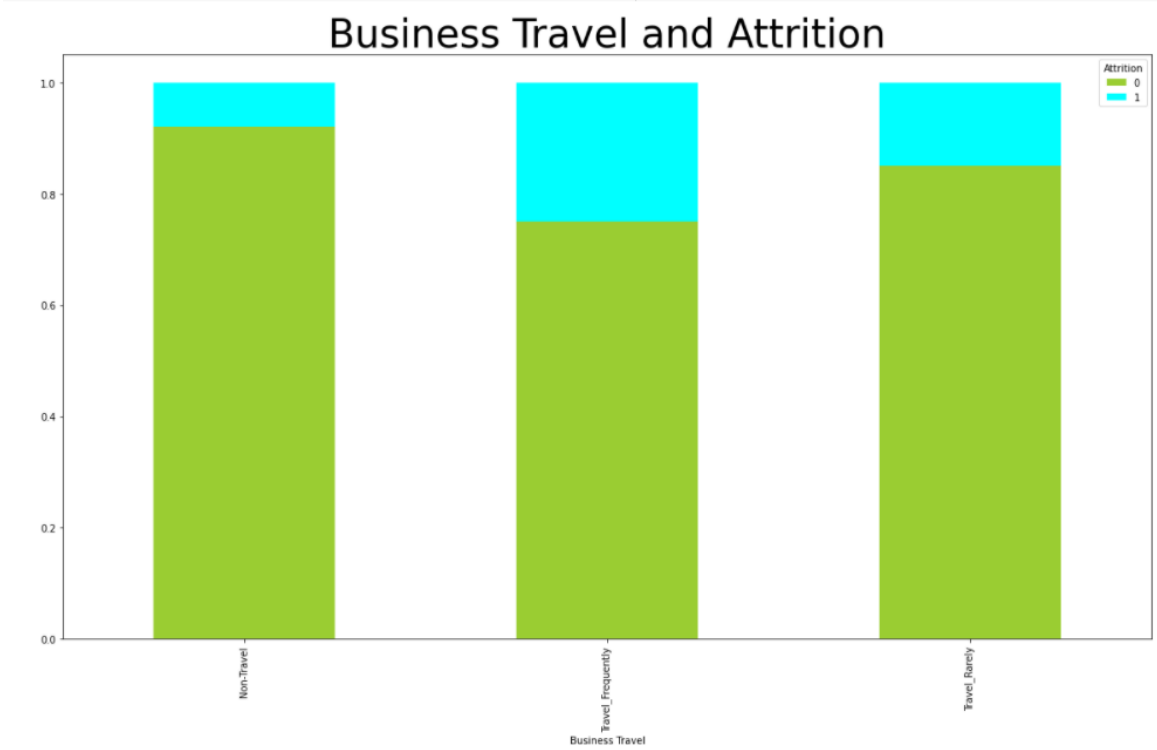
Employee Marital info

```
[ ]  colors = ['palevioletred', 'pink', 'mistyrose']

     fig = go.Figure(data=[go.Pie(labels=['Married','Single','Divorced'],
                                  values=[1346,940,654])])
     fig.update_traces(hoverinfo='label+percent', textinfo='label + value', textfont_size=20,
                       marker=dict(colors=colors, line=dict(color='#000000', width=2)))
     fig.show()
```

Department table

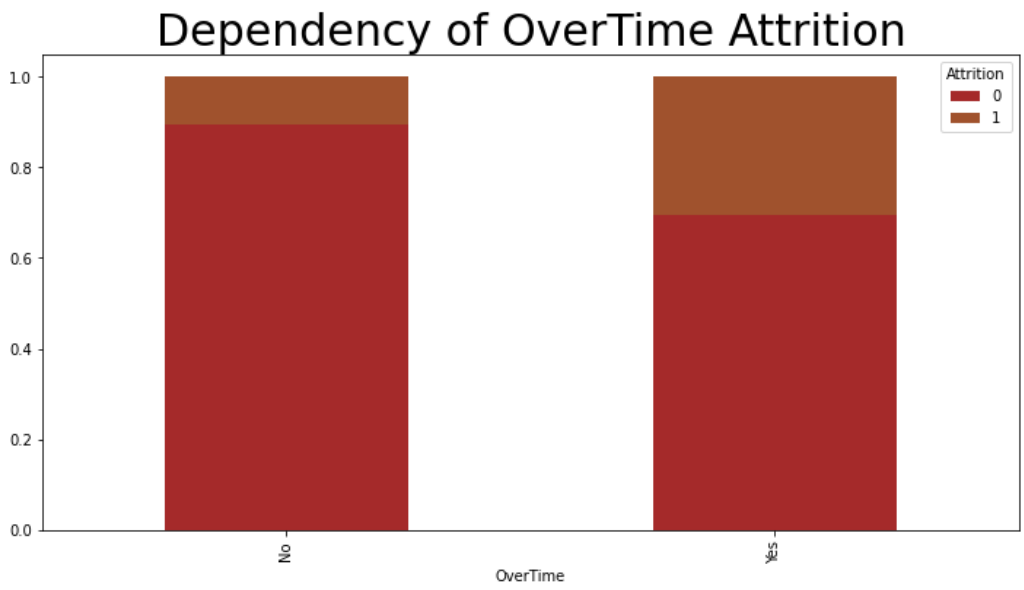
```
[ ]  data=pd.crosstab(df['Department'], df['Attrition'])
     data.div(data.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True, color=['tomato', 'magenta'],
                                                      figsize=(12,6))
     plt.title('Dependency of Department in determining Attrition', fontsize=30)
     plt.xlabel('Department')
     plt.show()
```





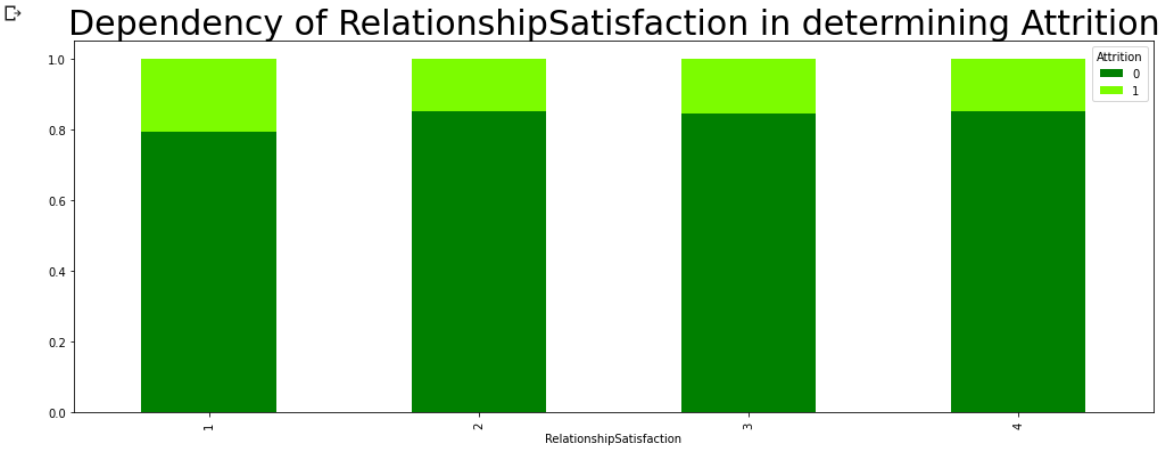
OverTime

```
[ ]  data=pd.crosstab(df['OverTime'], df['Attrition'])
     data.div(data.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True, color=['brown', 'sienna'],
                                                      figsize=(12,6))
     plt.title('Dependency of OverTime Attrition', fontsize=30)
     plt.xlabel('OverTime')
     plt.show()
```



Relationship Satisfaction attrition

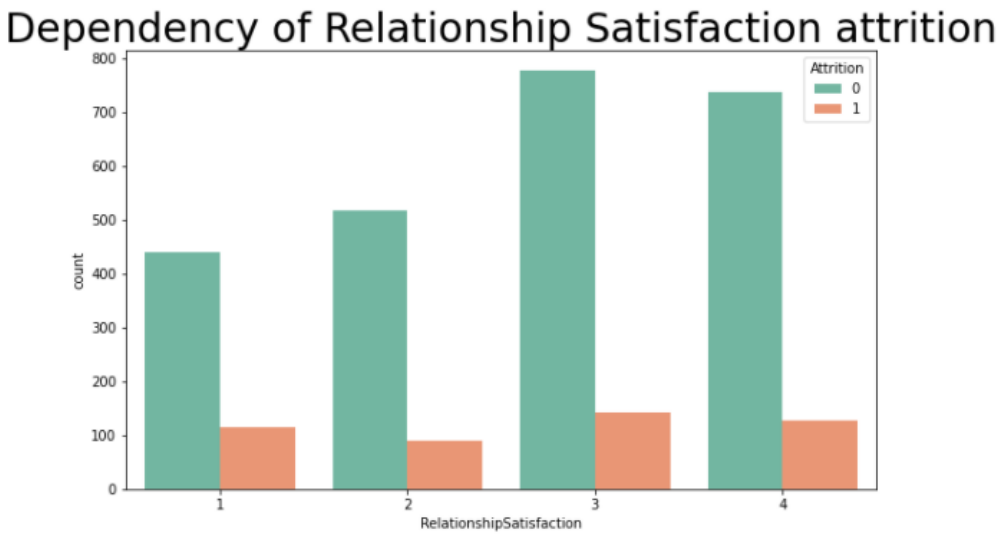
```
data=pd.crosstab(df['RelationshipSatisfaction'], df['Attrition'])
data.div(data.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True, color=['green', 'lawngreen'],
                                                 figsize=(17,6))
plt.title('Dependency of RelationshipSatisfaction in determining Attrition', fontsize=30)
plt.xlabel('RelationshipSatisfaction')
plt.show()
```



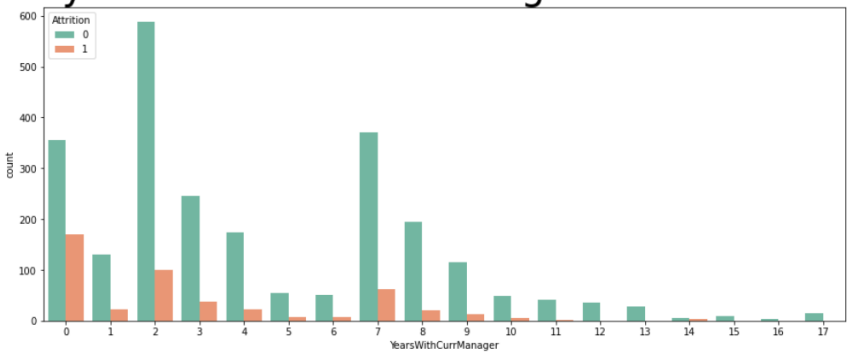
RelationshipSatisfaction

```
[ ] plt.figure(figsize=(10,6))
    sns.countplot('RelationshipSatisfaction', hue='Attrition', data=df, palette='Set2')
    plt.title('Dependency of Relationship Satisfaction attrition', fontsize=30)
    plt.xlabel('RelationshipSatisfaction')
    plt.show()
```

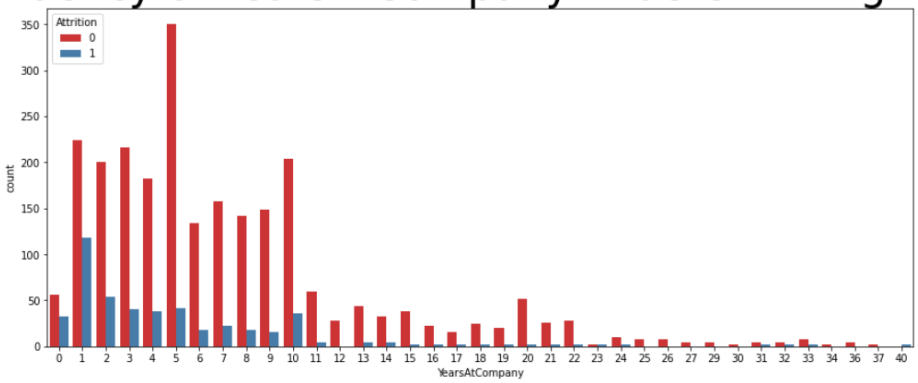
```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:

Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `da
```



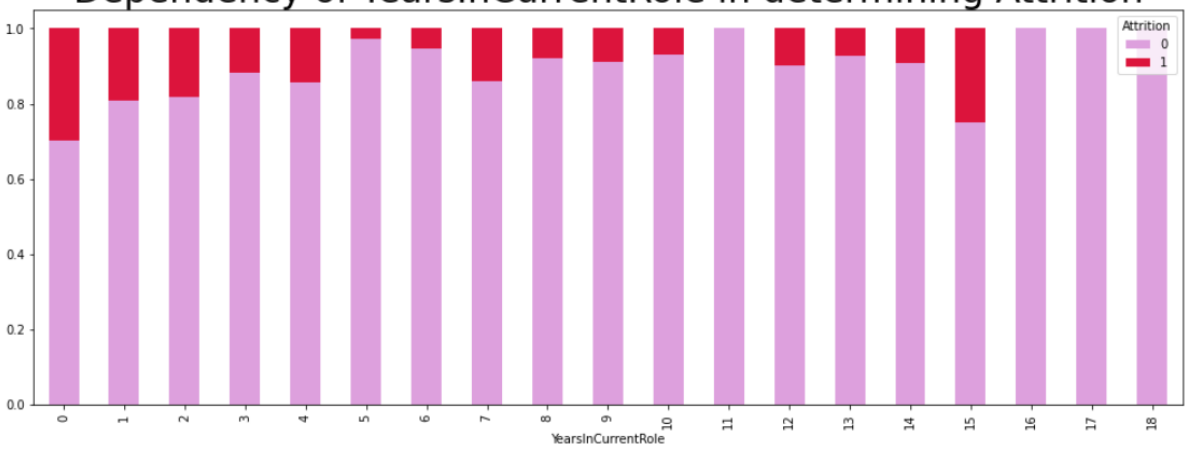
Dependency of YearsWithCurrManager in determining Attrition

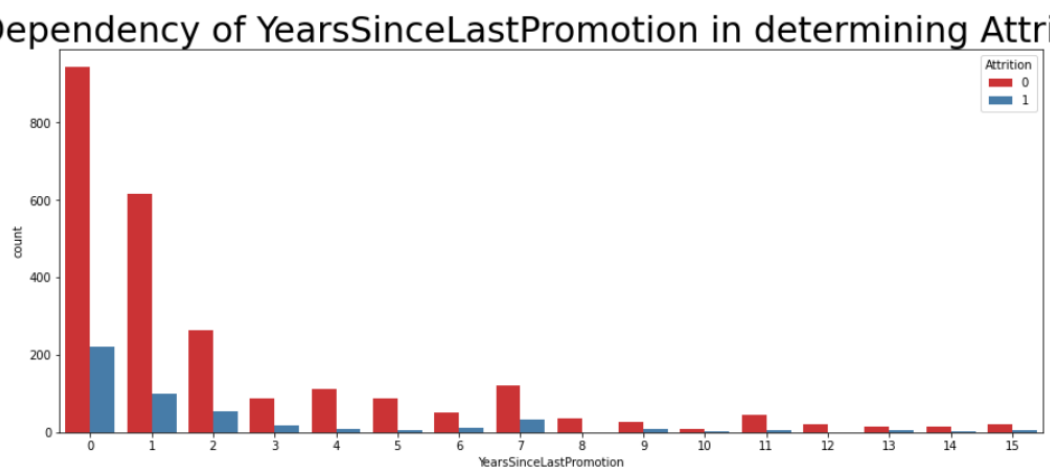


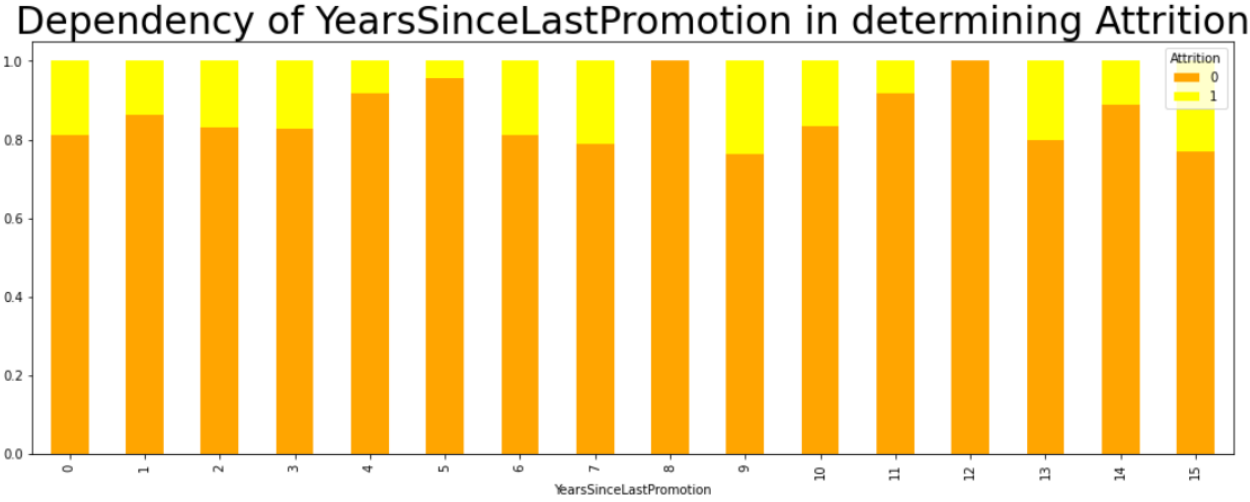
Dependency of YearsAtCompany in determining Attrition

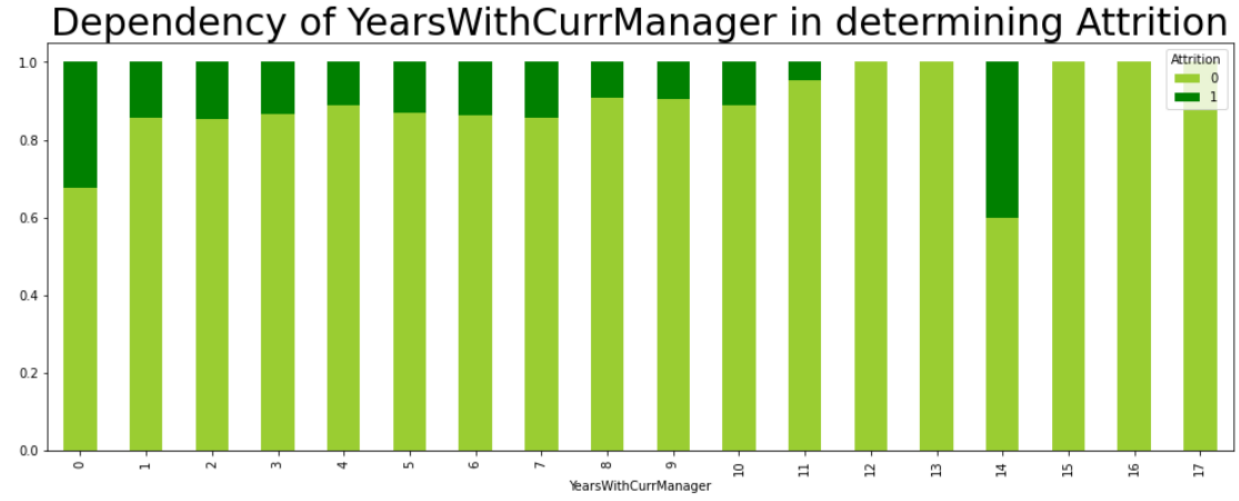


Dependency of YearsInCurrentRole in determining Attrition




Dependency of YearsSinceLastPromotion in determining Attrition
Attrition
0
1
count
YearsSinceLastPromotion

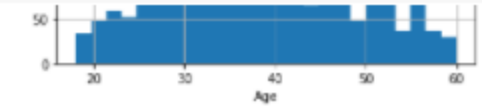

Dependency of YearsSinceLastPromotion in determining Attrition
Attrition
0
1
YearsSinceLastPromotion

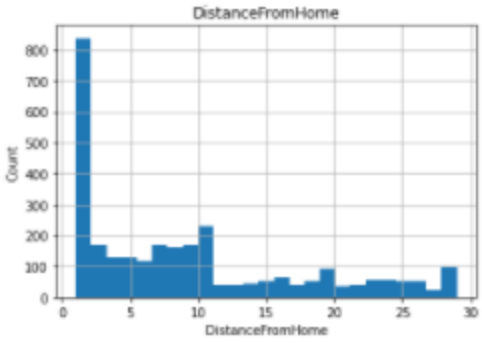

Dependency of YearsWithCurrManager in determining Attrition
Attrition
0
1
YearsWithCurrManager

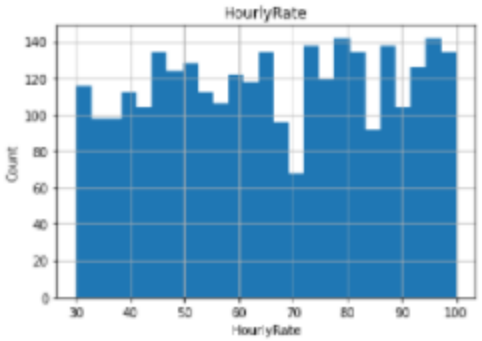
Feature Engineering

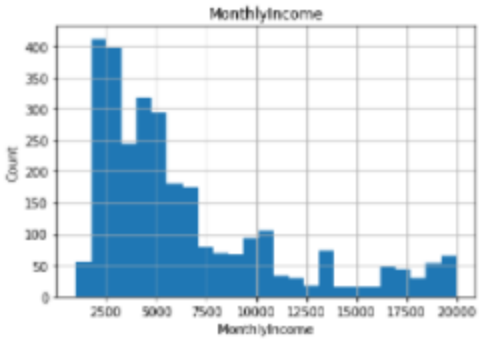
analysing continuous values by creating histograms to understand distribution

```
for feature in continuous_features:
  data= df.copy()
  data[feature].hist(bins = 25)
  plt.xlabel(feature)
  plt.ylabel("Count")
  plt.title(feature)
  plt.show()
```



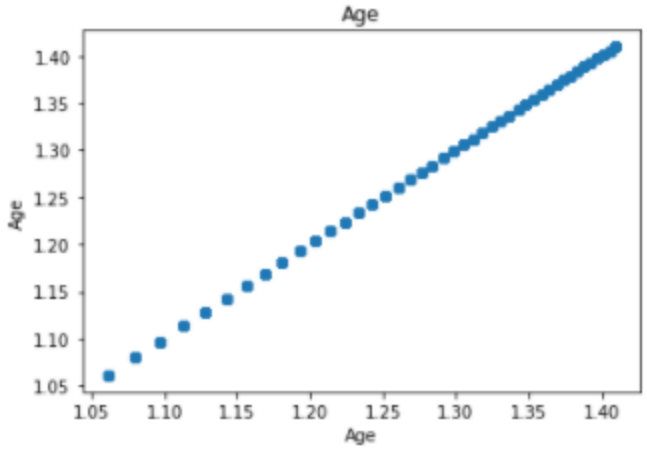


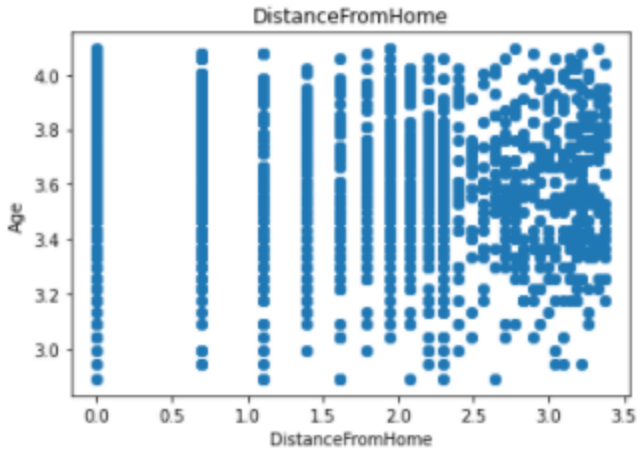


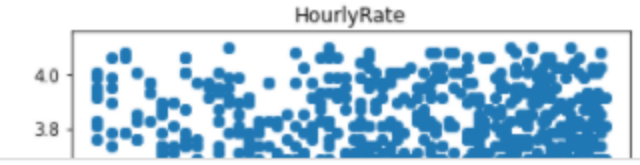


logarithmic transformation

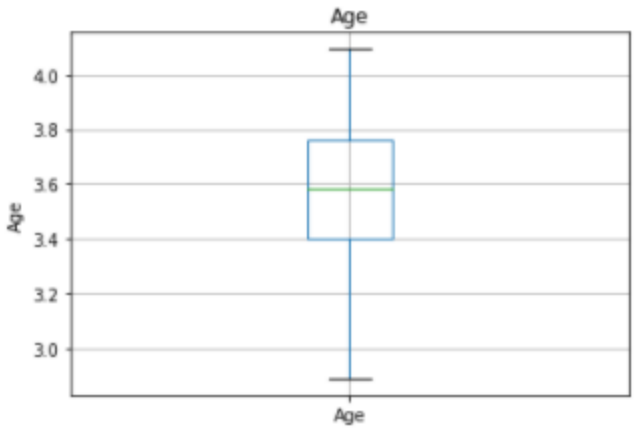
```
[ ]  for feature in continuous_features:
       data = df.copy()
       if 0 in data[feature].unique():
         pass
       else:
         data[feature] = np.log(data[feature])
         data["Age"] = np.log(data["Age"])
         plt.scatter(data[feature] , data["Age"])
         plt.xlabel(feature)
         plt.ylabel('Age')
         plt.title(feature)
         plt.show()
```

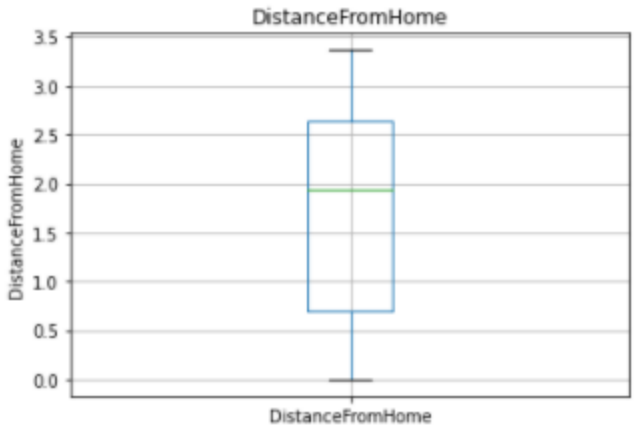


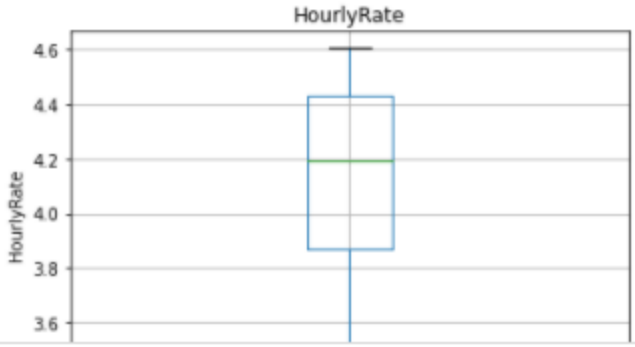




```
for feature in continuous_features:
  data = df.copy()
  if 0 in data[feature].unique():
    pass
  else:
    data[feature] = np.log(data[feature])
    data.boxplot(column = feature)
    plt.ylabel(feature)
    plt.title(feature)
    plt.show()
```







```
[ ]  for feature in categorical_features:
       data = df.copy()
       data.groupby(feature)["Age"].median().plot.bar()
       plt.xlabel(feature)
       plt.ylabel("Age")
       plt.title(feature)
       plt.show()
```