****

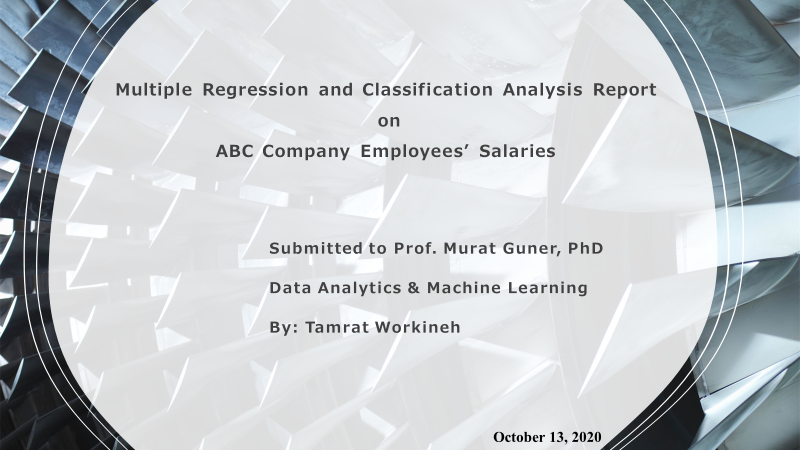
****

Table of Content

# Abstract 3

Introduction 3

# Study Objective 3

Motivation: 4

Method: 5

* Data Wrangling 5
* Train-Test Split dataset: 6
* Model: 7
* Experiments 8

Results and Interpretations 8

Conclusion and Summary 9

Limitations and later work. 9

References: 10

# 

# Abstract

This paper describes the application of multiple regression model to predict the annual salaries(target) based on sex, age, work experience and educational backgrounds of an ABC company employees. It also tries to examine if there is relationship between sex and annual income. Based on these numerical values of the features and the target variables, the performance of the model predicted with accuracy of 0.85.

# Introduction

Multiple Regression Analysis is a powerful tool for predicting and forecasting variables. Regression allows to study the relationship between dependent and independent variables thereby observe patterns to predict relating variables to each other. There are several projects out there based on this machine learning model on house sale. The researcher was motivated to use this model because it is a reliable method of identifying which variables have impact on a topic of interest.  In this project, a practical example of ABC company employees’ salaries was predicted based on respective sex, ages, work experiences and educational backgrounds. In the following section. Motivation, Related work, Proposed method, Experiments, Results and discussion, and Conclusion and summary were made based on the study conducted.

# Study Objective

The goal of the project is to determine the ability of age, work experience and education to predict the annual income of employees at ABC company by using multiple regression machine learning algorithm. It further detects the accuracy of sex and annual individual annual income relationship.

**Motivation:**

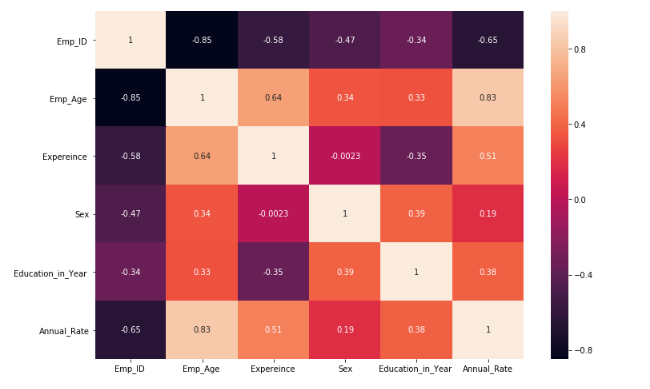
The motivation of behind using regression is may be due to its linearity, simplicity and it allows one to quantify the effect of each individual factor by also considering the interactions between the factors. Besides, in a regression one interprets the results as marginal effects.

**ABC- Company Dataset**

The ABC Company has 500 employees. The dataset of the company originally contains 500 rows and 13 columns. After data wrangling, it is reduced to 500 rows and 6 columns.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Emp\_ID | Employee \_ID |
| Emp\_Age | Age of employees |
| Experience | Work experience of the employees |
| Sex | Male or Female Employee |
| Education in Year | Educational Qualification (Diploma/Degree / Masters |
| Annual\_Rate | Annual salary of employee |

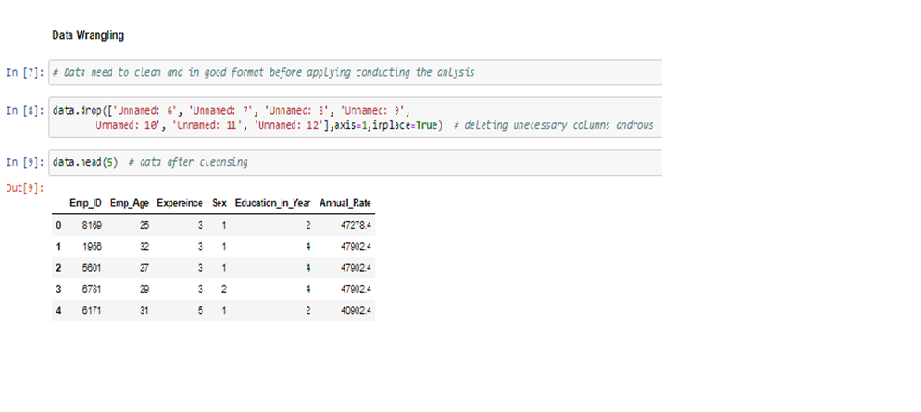
Table 1 Explanatory variables used in the Study

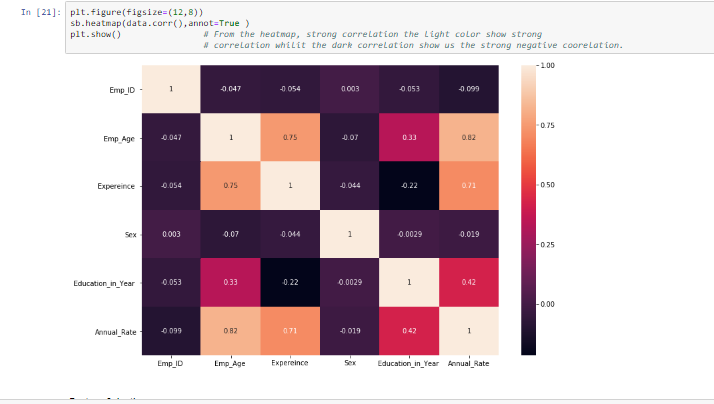
  
 Fig 1 : Heatmap : The heatmap shows the plosive correlation as we go up to lighter color.

**Method**:

**Data Wrangling:**

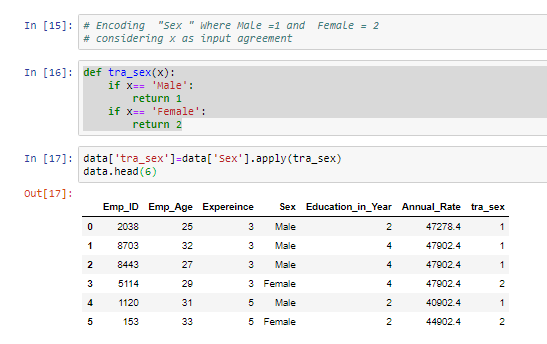
The ABC Company has 500 employees. The dataset of the company originally contains 500 rows and 13 columns. After data wrangling, it is reduced to 500 rows and 6 columns.





Label Encoding:

Since machine learning algorithm can only read numerical values, there was a need to convert categorical values into numbers. Accordingly, the Sex feature was encoded as Male = 1 and Female= 2 as shown hereunder.

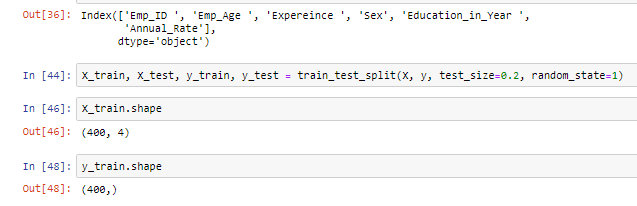


Data encoding(Male= 1 and Female = 2)

Fig : 2 *The heatmap shows the correlation between the predictor and predictand variables. The lighter color represents more positive correlation.*

**Train-Test Split dataset:**

This technique was used to take a sample data to evaluating the performance of the linear regression model. To this effect, the feature and target variables of a dataset were split into two subsets 80% to 20 % training and test respectively. Accordingly, the dataset was divided to contains X and Y trains 400 data rows each whereas X and Y test to have 100 rows each.



**Model**:

**Theory for multiple linear regression**

In multiple linear regression, there are p explanatory variables, and the relationship between the dependent variable (Y) and the explanatory variables (X) is represented by the following equation:

 ,

**Experiments** :

In this project, a standard multiple regression analysis was performed to assess the ability of age, sex, work experience and education to predict Annual rate of employees.

**Results and Interpretations**

A multiple regression model was calculated substituting the formula with the coefficients and Y-intercept to preditct the target employees’ annual rate. Based on numerical values of age, sex, work experience and education predictors. The accuracy of the model is found to be 0.85 as shown below:



**Classification and Regression Analysis with “Sex” vs “Annual Salary”:**

The second scenario was to detects the accuracy of sex and annual salary relationship. After the encoding the sex variable, the dataset containing the sex (feature) and annual salary (target)were split into two subsets 80% to 20 % training and test respectively to find out the accuracy of the regression model as described on the Jupyter notebook hereunder.



**Conclusion and Summary:**

The goal of the project is to determine the ability of age, work experience and education to predict the annual income of employees at ABC company by using multiple regression machine learning algorithm. The model, based on numerical values of age, work experience and education predictors. The accuracy of the model is found to be 0.85. The study further implemented the regression analysis model to see the accuracy of the model in this regard showed 0.0045.

**Limitations and later work.**

Due to ongoing class and limited data model knowledge, the study was conducted based on regression model and classification to determine if there is a relationship between sex and the target variable. It has a constraint on generalizability of findings. However, the study will be validated as after additional models are taught.

**References:**

1. [https://www.valuebasedmanagement.net/methods\_regression\_analysis.html](about:blank)
2. https://www.statisticssolutions.com/sample-size-formula/