**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | **Programming for DA**  **Statistics for Data Analytics**  **Machine Learning for Data Analysis**  **Data Preparation & Visualisation** |
| **Assessment Title:** | **MSC\_DA\_CA1** |
| **Lecturer Name:** | **David McQuaid**  **Sam Weiss/ David Gonzalez**  **Bharathi Chakravarthi/ Marina Iantorno**  **Muhammad Iqbal** |
| **Student Full Name:** | **Muhammad Asif** |
| **Student Number:** | **2021402** |
| **Assessment Due Date:** | **12th November,2023** |
| **Date of Submission:** | **12th November,2023** |

**Muhammad Asif 12-11-2023**

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Abstract**

*Data analytics plays a critical and increasingly important role in today's data-driven world. Its significance can be seen in various aspects of business, science, and society. In this report, we utilized data analytics techniques to gain useful insights from the Annual Population change in Ireland. Data is imported and checked thoroughly to answer questions. During the process, it went through different stages from preparation for analysis to graphical representation, using statistics to seek trends and finally creating a Machine Learning model to seek the output parameters from validation set. Whole work is enclosed in Jupyter Notebook which used Python framework to carry out the programming requirements*

**Introduction**

When it comes to examining annual population changes, data analytics plays a crucial role in the field of population studies. In order to uncover trends, patterns, and invaluable insights, it entails carefully analyzing data. Its significance stems from the fact that it facilitates an understanding of the intricacies of population dynamics, including migration trends, birth and death rates, and other crucial demographic factors that significantly influence a region's evolutionary processes. By analyzing and interpreting massive datasets, data analytics enables academics and decision-makers to make well-informed decisions in areas like resource allocation, urban planning, and the development of successful policies. In this case, the data analysis pipeline entails several crucial procedures. These include collecting, sanitizing, and organizing data; doing exploratory analysis to look for hidden patterns; and, finally, developing predictive models. Through these procedures, data analytics contributes significantly to the transformation of raw data into insightful knowledge, improving comprehension of population changes, and enabling more informed, data-driven decision-making in the ever-changing field of demographics.

**Data Preprocessing**

Data preprocessing is a vital stage in the data analysis pipeline that involves a number of crucial tasks related to preparing raw data for machine learning and statistical analysis. Data cleansing, which addresses outliers, inconsistent values, and missing values in the dataset, is one of the initial processes. Standardization or normalization is frequently required to bring conflicting data to a common scale, outliers can be discovered and handled to stop them from influencing the model, and imputation techniques can be used to address missing data. Data transformation, which includes feature engineering to create new, informative attributes that might enhance model performance, scaling of numerical features, and categorical variable encoding, is another essential element. In order to decrease dimensionality and maybe enhance model precision and computational overhead, feature selection strategies, and data reduction techniques like Principal Component Analysis (PCA) are used. In essence, data preparation is a crucial step that ensures the data is in the best possible shape so that machine learning models can generate accurate predictions.

**Data Visualization**

Data visualization, which involves displaying data visually to detect trends, correlations, and other patterns, is an essential phase in the data analysis process. Effective data visualization requires not only the creation of aesthetically pleasing charts and graphs, but also the selection of the right visualization techniques based on the objectives and type of data. Histograms for data distribution, scatter plots for variable relationships, bar charts for comparisons, line charts for trends over time, and heat maps for pattern detection are just a few examples of the many various kinds of visualizations that can be employed. Because its ultimate purpose is to enable data-driven decision-making and effective distribution of insights to stakeholders, data visualization is an essential component of the data analysis process.

**Machine Learning**

Algorithms and mathematical structures known as machine learning models enable computers to learn from data and make predictions or judgments without the need for explicit programming. These models embrace a wide range of methods: unsupervised learning works with unlabeled data to find innate patterns and structures, whereas supervised learning concentrates on labeled data, where the model is trained to generate predictions based on input features. Several methods, including support vector machines, decision trees, deep neural networks, and linear regression, are used in supervised learning to tackle tasks like regression and classification. For tasks like data segmentation and feature reduction, unsupervised learning includes clustering algorithms like K-means and dimensionality reduction techniques like Principal Component Analysis (PCA). Furthermore, reinforcement learning is a subfield of machine learning that is used in robotics and autonomous systems to teach agents how to maximize cumulative rewards by making a series of decisions in dynamic situations. The problem at hand determines which machine learning model is best, and in order to guarantee optimal model performance, careful processes in model selection, hyper parameter tuning, and model evaluation are necessary. It's critical to recognize that the overall performance of machine learning models in real-world applications is strongly influenced by the caliber of the data as well as the efficiency of data preparation and feature engineering.

**Annual Population Ireland**

One of the primary functions of data analytics in the study of annual population changes is to identify trends. By scrutinizing historical data, statisticians and demographers can determine whether the population is expanding, declining, or remaining stable. This information is of utmost importance to governments, urban planners, and businesses, as it allows them to foresee shifts in resource requirements, infrastructure needs, and market opportunities. Moreover, data analytics enables the identification of the factors driving these changes, such as birth rates, death rates, and migration patterns, offering a more profound understanding of their root causes.

This report adheres to the same approach and utilizes Annual Population Change Data for Ireland to provide the government and citizens with an overview of evolving population trends and its components, including annual births and deaths, immigrants and emigrants, natural increase, net migration derived from formal variables, total population, and population distribution. Data has been collected from 1951 to 2023, with values measured in thousands.

In the next stages, we will thoroughly discuss the methods from which we drive the statistics and useful insights using various data analytics techniques. We will be discussing the important python libraries that enabled us to carry the required operation and how each analytics plays its fundamental role in identifying the patterns and trends for our dataset.

**Methodology**

**Data Preprocessing**

We utilized the important python packages at each step to carry out specific task. First, is the use of Python Pandas to import and manipulate the dataset. The first thing which is needed to be done is to look for any ambiguity in our dataset to avoid any noise or wrong information. Therefore, we used head () and sine () method with sum () to look for any missing values and we found 144 of these in the VALUE field. From the VALUE distribution plot, we found out that there is low saturation of data in the beginning rows and it gets denser afterwards which leads to the fact that data for year 1986 and below were missing. Since, we got enough years to carry out our analysis so we discarded that row and cleaned our dataset.

**Statistics**

The normal distribution, often referred to as the Gaussian distribution or the bell curve, holds paramount importance in statistics and various scientific disciplines due to its versatility and prevalence. It serves as a fundamental model for characterizing the distribution of data in natural and man-made phenomena. The central limit theorem, a cornerstone of statistics, underscores its significance, as it dictates that the means of repeated random samples from any population tend to follow a normal distribution, even when the underlying data is not normally distributed. This crucial characteristic allows researchers and statisticians to conduct hypothesis testing, estimate parameters accurately, and draw strong conclusions about populations. The normal distribution offers a common framework for comprehending and modeling variability, which helps to simplify complicated real-world issues and aids in forecasting and decision-making. The normal distribution is a universal and essential concept in the field of statistics and empirical research, with applications spanning from physics and engineering to economics and biology. It is a potent tool for analyzing and forecasting data. It is the cornerstone of probability and statistics due to its symmetrical, well-defined properties and broad applicability, which enhances our comprehension of the world and informs innumerable practical applications.

We plotted the normal distribution's probability density function for our dataset using the spicy package. This information was crucial because it allowed us to understand the variance and average estimates of each parameter along the bell curve, as depicted in figure 01.

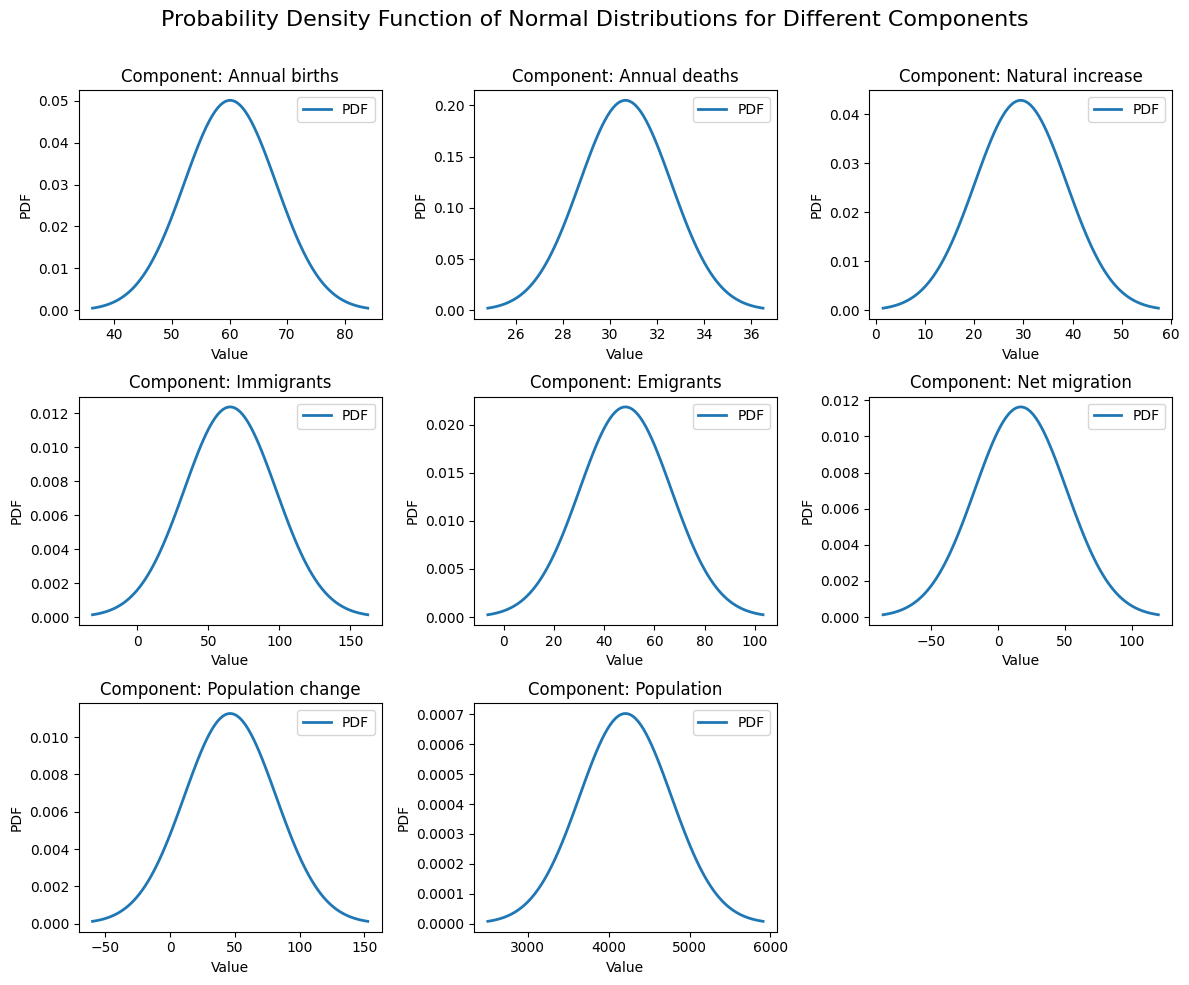
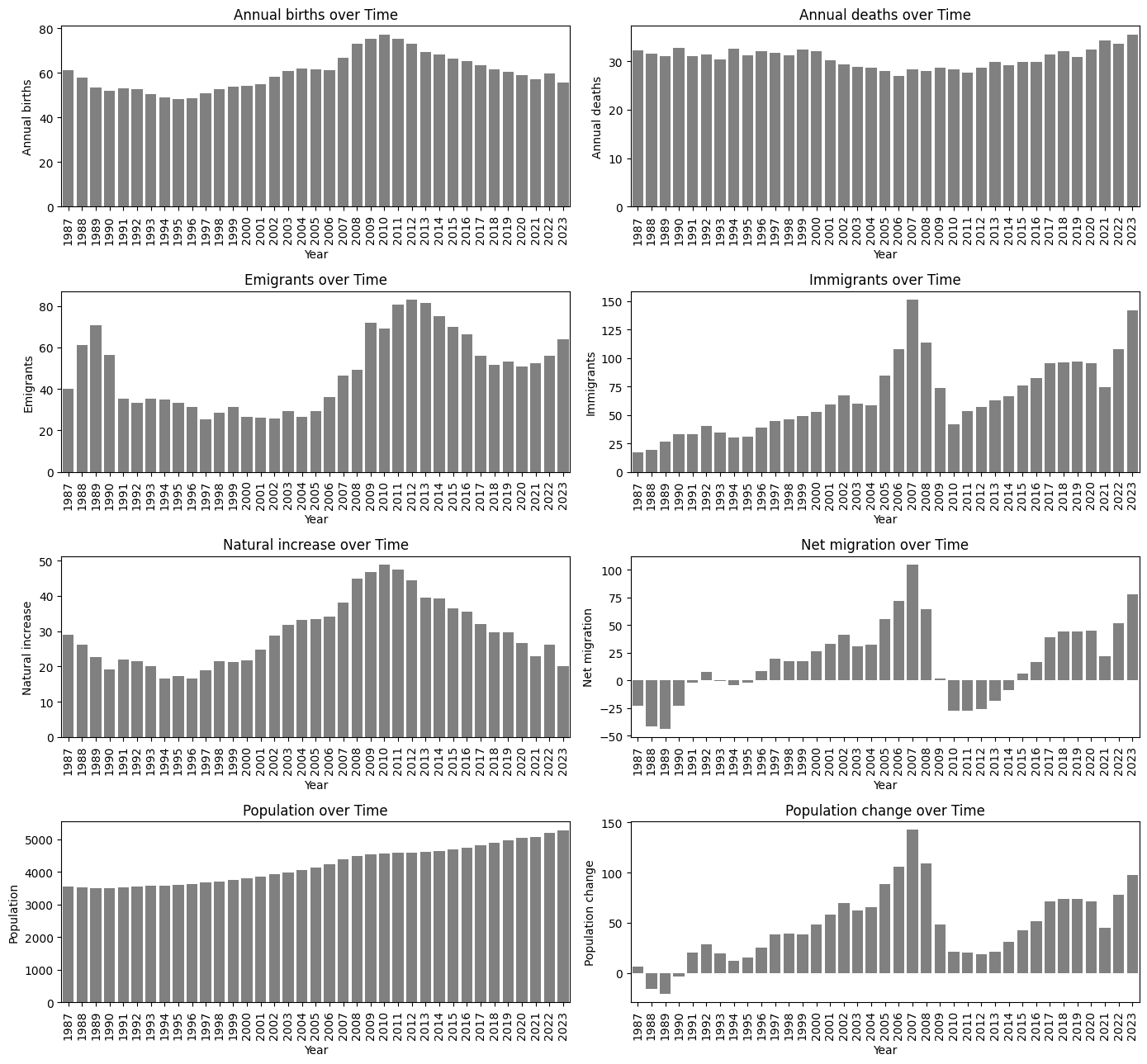


Figure 01. PDF of Normal Distribution for each Component

**Data Preparation and Visualization**

Bar graphs are an essential part of this visual communication toolkit, which is used in data visualization, a potent technique for communicating complex information. Bar graphs offer a simple and straightforward method for comparing values between various categories and representing categorical data. Usually, they are made up of rectangular bars that can be vertical or horizontal, and each bar's length corresponds to the quantity it represents. Bar graphs are a vital option for many applications, ranging from science and education to business and finance, due to their simplicity and efficacy. They are the perfect option for displaying discrete data, such as sales figures for different products, student performance across subjects, or population distribution across regions, because they allow viewers to quickly understand trends, variations, and relationships within the data. Bar graphs can be further enhanced by incorporating color coding, grouping, and stacking to highlight specific insights within the data. Their adaptability, accessibility, and straightforward design make them a staple in the world of data visualization, helping analysts, researchers, and decision-makers present their findings in a manner that is both visually engaging and easy to interpret, thereby facilitating more informed decisions and a deeper understanding of the underlying informationTop of Form. After statistical analysis, we prepared our data for the exploratory analysis by eliminating the STATISTIC Label and UNIT column as they didn’t serve any purpose. Our data required reshaping since the categories in component field needed to be align in columns not rows so we used pivot () method from pandas to reshaped our data. We started with the describe () method that gave the additional statistical information with 25%, 50% and 75% quantiles which are important to draw projects at different intervals of data. The bar plot is plotted using the sea born package to get historical trend of each component as shown in figure 02

Figure 02. Bar plot of each component over years (Time)

A correlation plot, a fundamental tool in data analysis and visualization, is a graphical representation of the relationships and dependencies between variables within a dataset. It is especially helpful in examining and evaluating the strength and type of relationships between different sets of variables. Each variable is usually represented as a row and a column in a matrix for a correlation plot, with correlation coefficients shown at the intersections. A color-coded matrix is produced when positive correlations are frequently represented in one color and negative correlations in another. It is visually evident how strong and in which direction these correlations are, with darker or more intense colors denoting stronger associations. For a variety of uses, including determining important characteristics during data preprocessing and comprehending multi collinearity in regression analysis, correlation plots are invaluable. They give researchers, scientists, and data analysts a thorough understanding of how variables interact, empowering them to make deft choices regarding feature selection, model construction, and data-driven insights. Correlation plots are essentially an effective diagnostic and exploratory tool that help to extract meaningful patterns and relationships from large, complicated datasets. They also play a crucial role in the process of finding hidden insights that can be used for further analysis and well-informed decision-making.

To check linear trend between our numerical field, we plotted a correlation plot to seek relationships between different component fields. This is done using the correlation matrix provided by core () method and is sent to sea born for interactive visualization as shown in figure 03

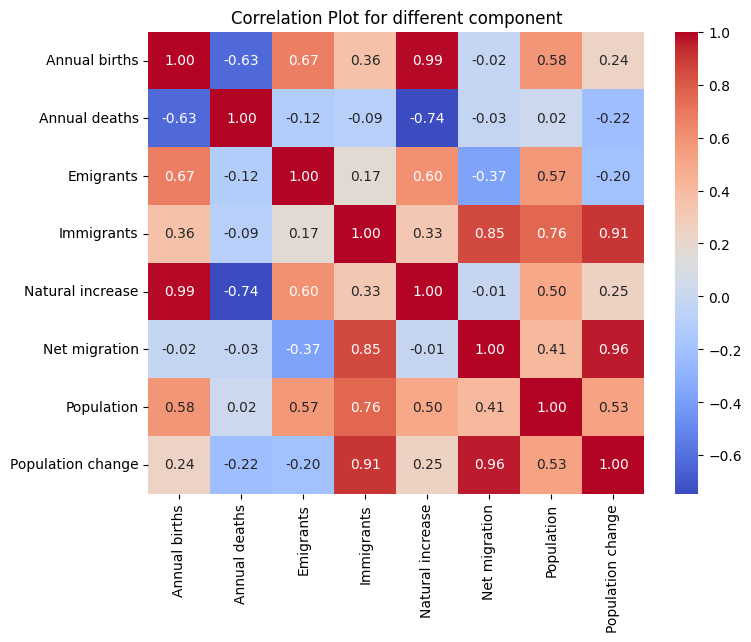


Figure 03. Correlation plot between different components

**Machine Learning for Data Analytics**

With wide-ranging effects, data scaling is a crucial machine learning preprocessing step. Data scaling involves standardizing and adjusting a dataset's numerical features to ensure that they fall within a predefined range, typically between -1 and 1. This process is crucial when working with algorithms that are sensitive to feature magnitudes, such as neural networks, support vector machines, and k-nearest neighbors. By scaling, one can make sure that every feature makes a proportionate contribution to the model's performance, avoid the unwarranted influence of variables with larger scales, and facilitate a faster convergence of the model during training. It also makes the model coefficients easier to read, which makes it easier to understand the relative importance of each feature. Data scaling is also crucial for distance-based algorithms, since unscaled features can drastically alter the metric used to determine similarity. Ultimately, data scaling is a crucial technique that improves the stability and performance of machine learning models, facilitates better comprehension of the outcomes, and fully utilizes the data to produce more insightful and precise predictions.

Regression analysis is a statistical technique for modeling the relationship between a dependent variable and one or more independent variables. Often used to understand how changes in one or more predictors affect the dependent variable, it is a crucial tool in data analysis. Linear regression is among the most widely used and fundamental forms of regression. In linear regression, it is assumed that the independent and dependent variables have a linear relationship that can be represented by a straight-line equation. It is necessary to identify the best-fitting line, or hyperplane in the case of multiple predictors, in order to minimize the sum of squared differences between the predicted and actual values. Numerous disciplines, including economics, finance, and the social sciences, use linear regression.

Conversely, ridge regression is a regularized version of linear regression. It is intended to deal with the situation of highly correlated independent variables, or multi collinearity. Large predictor coefficients are discouraged by ridge regression, which adds a penalty term to the linear regression objective function. By reducing overfitting, this regularization term makes the model more stable. Ridge regression comes in especially handy when working with datasets that have a lot of features and multi collinearity problems. It achieves a balance between preserving model interpretability and lowering the model's variance (overfitting). Ridge regression enhances the generalization and robustness of linear regression models by being a member of the family of regularization methods known as "shrinkage methods". In conclusion, ridge regression is a useful strategy in the field of statistical analysis and predictive modeling. While linear regression is a basic tool for modeling linear relationships, ridge regression increases its usefulness by reducing multi collinearity and improving model stability.

Regression analysis's output parameters offer important information about how the independent and dependent variables are related to one another. The regression coefficients, intercept, R-squared (coefficient of determination), p-values, and standard errors are important output parameters. Regression coefficients provide quantitative measures of the influence of the predictors by expressing the change in the dependent variable associated with a one-unit change in the corresponding independent variable. The intercept is the value of the dependent variable when all independent variables are zero. R-squared quantifies the goodness-of-fit, indicating the proportion of variance explained by the model. P-values assess the significance of each coefficient, helping to determine whether predictors are statistically significant. Standard errors provide a measure of the precision of the coefficients' estimates. Collectively, these output parameters guide the interpretation of regression models, facilitating the identification of significant predictors, understanding the model's explanatory power, and assessing its overall quality, making them indispensable in the process of drawing meaningful insights and informed decisions from regression analyses.

The data prepared in EDA is used to create regression models with population change as target variables and annual birth, annual death, immigrants and emigrants as features. This selection of features comes from the correlation plot from figure 03 that provided us with best possible co-efficient. Before modeling and splitting, it is best practice to scale the data for better computation therefore, we used MinMaxScalar () from sk-learn package and split data into training and testing using train\_test\_split (). First, we created a simple linear regression and calculated mean square error and R square to evaluate the model parameters. Next, we used Ridge regression and obtained results for different hyper parameter alpha i.e., the regularization co-efficient to plotted the output parameters as shown in figure 04.

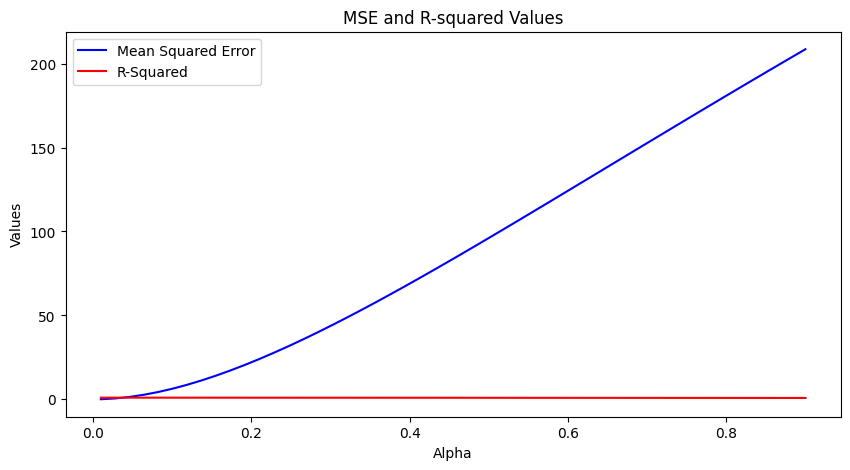


Figure 04. Mean Square Error and R Square for different regularization constant in Ridge model