**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation  Machine Learning for Data Analysis  Programming for Data Analytics  Statistics for Data Analytics |
| **Assessment Title:** | MSC\_DA\_CA1 |
| **Lecturer Name:** | **David McQuaid**  *Muhammad Iqbal*  *Sam Weiss*  *Bharathi Chakravarthi* |
| **Student Full Name:** | Syed Asad Ailia |
| **Student Number:** | 2023408 |
| **Assessment Due Date:** | 12th November,2023 |
| **Date of Submission:** | 12th November,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**

*In today's data-driven world, data analytics is becoming more and more significant, having a major impact on many facets of business, science, and society. In this report, we used data analytics techniques to clean insightful information from the mean and median hourly earnings across various economic sectors and employment statuses in Ireland. We carefully imported and checked the information to respond to specific questions. The development of a machine learning model for extracting output parameters from the validation dataset was the final step in this extensive process, which began with data preparation and continued with graphical representation using statistical techniques to identify trends. The Python framework was used to fulfil the programming requirements, and the entire project was recorded in a Jupyter Notebook.*

**Introduction**

In today's data-driven world, data analytics is an essential tool, particularly when looking at hourly salaries across various economic sectors and employment status. This analytical approach finds trends, patterns, and valuable information by methodically examining and interpreting data. Organizations, decision-makers, and scholars can gain a comprehensive grasp of the hourly wage market through the data analytics. Additionally, it offers data-driven insights into income inequality and job markets. Hourly earnings, a crucial economic metric, vary substantially between industries and job categories, reflecting the complexity of today's economy. Data analytics facilitates data-driven decision-making, highlights discrepancies, and points out causes, making it easier to navigate this complicated landscape. Our ability to analyze the data is further improved by statistical methods and machine learning models, which help us forecast future trends and the effects of policies in addition to describing the current state of affairs. When it comes to addressing economic disparities in a world where data-driven insights are essential, data analytics becomes a crucial ally that leads the way toward a more just and prosperous future. Here, we'll examine a few crucial data analytics pipelines to gain deeper understanding.

Data Preparation & Visualization

To prepare raw data for machine learning and statistical analysis, several important tasks involve data pre-processing, a fundamental step in the data analysis process. One of the first steps is data cleaning, which deals with the dataset's outliers, inconsistent values, and missing values. Imputation techniques can be used to address missing data, outliers can be identified and controlled to stop them from distorting the model, and standardization or normalization is frequently necessary to bring inconsistent data to a common scale. Another important component is data transformation, which includes feature engineering to produce new, informative attributes that can improve model performance, scaling of numerical features, and encoding of categorical variables. To reduce dimensionality and potentially improve model accuracy and computational overhead, data reduction techniques such as Principal Component Analysis (PCA) and feature selection methods are applied. To put it simply, data pre-processing is an essential step in ensuring that the data is prepared optimally for machine learning models to produce precise predictions.

A crucial step in the data analysis process is data visualization, which involves displaying data graphically to identify trends, relationships, and other patterns. Choosing the appropriate visualization techniques to match the goals and nature of the data is just as important to effective data visualization as designing visually appealing charts and graphs. There are many different types of visualizations that can be used, such as histograms for data distribution, scatter plots for relationships between variables, bar charts for comparisons, line charts for trends over time, and heatmaps for pattern recognition. Data visualization is an integral part of the data analysis process because its ultimate goal is to support data-driven decision-making and efficient dissemination of insights to stakeholders.

Machine Learning

Machine learning models are algorithms and mathematical constructs that empower computers to learn from data and make predictions or decisions without explicit programming. These models span a spectrum of techniques, with supervised learning focusing on labelled data, where the model is trained to make predictions based on input features, while unsupervised learning deals with unlabelled data to discover inherent patterns and structures. Within supervised learning, various algorithms like linear regression, decision trees, support vector machines, and deep neural networks are applied to tasks such as classification and regression. Unsupervised learning encompasses clustering techniques like K-means and dimensionality reduction methods like Principal Component Analysis (PCA) for tasks like data segmentation and feature reduction. Additionally, reinforcement learning is a branch of machine learning where agents learn to maximize cumulative rewards through a sequence of decisions in dynamic environments, serving purposes in areas like robotics and autonomous systems. The choice of the appropriate machine learning model depends on the problem at hand, and rigorous steps in model selection, hyperparameter tuning, and model evaluation are critical to ensure optimal model performance. It's essential to acknowledge that the quality of data and the effectiveness of data pre-processing and feature engineering significantly impact the overall success of machine learning models in real-world applications.

Wages in Ireland

When examining hourly earnings based on employment status, data analytics offers the capacity for a more granular and insightful perspective on income distribution. By segmenting data into employment categories such as full-time, part-time, and aggregating all employment types, it becomes possible to unravel the intricate variations in earnings and employment dynamics. This approach allows for a more comprehensive understanding of the labour market, revealing, for instance, whether certain sectors exhibit a higher prevalence of part-time employment and how these dynamic influences hourly wages. These insights are invaluable for labour market research, providing businesses and government entities with the tools to make well-informed decisions regarding labour policies, workforce management, and strategies for addressing income disparities and economic inequality.

The same pattern of analysis is used in this report as well, with an emphasis on mean and median hourly earnings in Ireland based on different employment statuses and economic sectors. The main goal is to give the public and government a thorough understanding of the shifting patterns in earnings and the role that employment status plays in maximizing wages. The data utilized is sourced from the year 2022, offering not only hourly earnings but also information on the number of hours worked during the week, which enhances the statistical insights. In the next section, we will further break down our analysis into distinct stages to obtain a more comprehensive and nuanced understanding of hourly earnings within Ireland's economic landscape. This segmentation will empower stakeholders with the insights needed to create policies, make decisions, and take actions that foster economic growth, ensure fair wages, and improve the well-being of the workforce while addressing the vital role of employment status in shaping earnings.

**Methodology**

Data Pre-processing

We used key Python packages at different stages of the project to make sure everything ran smoothly and efficiently. We started our journey by importing and working with our dataset using Python Pandas, a potent data manipulation library. This first step's main objective was to carefully review the dataset in order to identify any potential ambiguities that might introduce errors or provide false information. In order to do this, we carefully used the sum() function in conjunction with Pandas techniques like head() and isna(), carefully going over the dataset to look for any missing values. Thankfully, our inspection produced a comforting result: we found that the dataset had no missing values, which allowed for a reliable and error-free analysis. This thorough data pre-processing step highlights the vital role Python packages play in simplifying data handling, highlighting the importance of these tools in the data analytics space.

Statistics

Because of its popularity and versatility, the normal distribution—also known as the Gaussian distribution or the bell curve—is used extensively in statistics and other scientific fields. It functions as a basic model for describing how data are distributed in both natural and artificial phenomena. The significance of statistics is highlighted by the central limit theorem, which states that even in cases where the underlying data is not normally distributed, the means of repeated random samples from any population will typically follow a normal distribution. 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 modelling 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 analysing 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.

To begin our statistical analysis, we first used the Pandas library's describe() method to get a general overview of the dataset. Nevertheless, this approach was not able to give us the thorough insights we needed because of the categorical fields. We used specially written code to overcome this restriction, making use of the SciPy package's features. We were able to derive a more accurate and insightful understanding of the dataset by using this method. Specifically, we used Scipy to plot the Normal distribution's probability density function, which was essential to our analysis.

We were able to determine the average values and variances of the dataset's parameters by extracting important information from this distribution. The bell curve that was used to represent this is clearly shown in Figure 01. It is crucial to comprehend the distribution of these parameters because it helps us identify the dataset's central tendencies and variability, which lays the groundwork for our statistical analysis. When handling complex data, the flexibility and adaptability of Python packages like SciPy prove invaluable, enabling us to derive valuable insights and support data-driven decision-making.

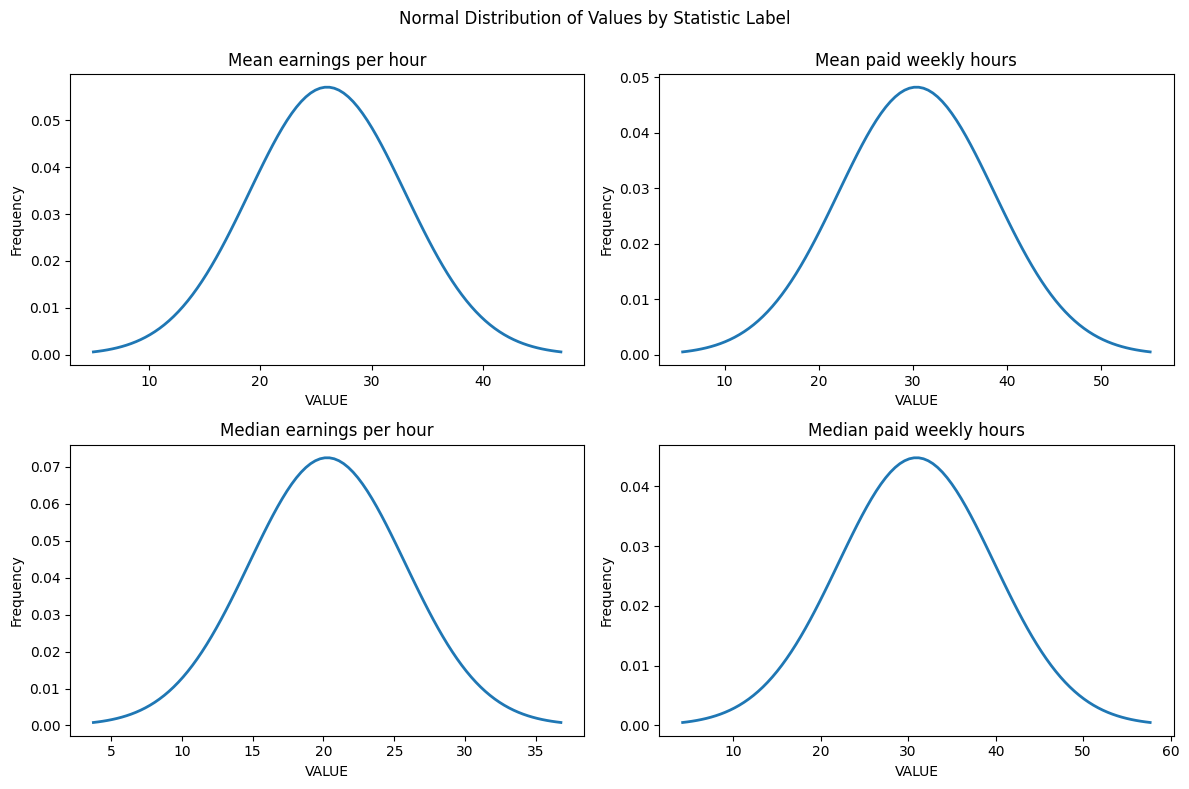


Figure 01. PDF of Normal Distribution for each Statistics label

The distribution is balanced across the average value of each statistic label and spread across the standard deviation. This is useful to seek the average wage in per hour earning according to average mean hours paid weekly across all economic sector combined.

Data Preparation & Visualization

Data visualization, a powerful method for communicating complex information, uses bar graphs as a key component of this visual communication toolkit. Bar graphs provide a clear and easy way to represent categorical data and compare values between different categories. They are typically composed of rectangular bars, each of which can be vertical or horizontal, and whose length is determined by the quantity it represents. Because of their effectiveness and simplicity, bar graphs are a great choice for a wide range of applications, from science and education to business and finance. They are the ideal choice for showing discrete data, like sales numbers for various goods, academic achievement across subjects, or regional population distribution, because they make it possible for viewers to discern patterns, variances, and connections among data quickly. Bar graphs can be made even better by highlighting specific data insights with colour coding, grouping, and stacking. Because of their versatility, ease of use, and straightforward design, they are a mainstay in the field of data visualization, helping analysts, researchers, and decision-makers to present their findings in an understandable and visually appealing manner. This encourages more in-depth understanding of the underlying data and improved decision-making.

We ready our data for a thorough exploratory analysis after finishing our first statistical analysis. We opted to remove the year column in order to simplify the process because we didn't think it was relevant to the patterns we were trying to find. Earnings trends based on Employment Status and Economic Sector were the two primary categorical attributes that dominated our dataset and required our attention. Analysing these characteristics revealed distinct wage patterns, which are shown graphically in Figures 02 and 03.

Figure 02 provided us with a thorough understanding of how per-hour earnings distributed across different employment statuses for each statistical label. We used the Seaborn package, a flexible data visualization tool, to accomplish this. The development of interactive visualizations specifically for categorical attributes was made easier by this package. Every subplot was closely linked to a particular statistical label, and the data distribution was represented graphically using histograms. These plots provided insights into the relationships between earnings and employment status by using the lines as indicators of ongoing changes in value trends.

Another output from the Seaborn package, Figure 03, provided a clear visual representation of the complex relationship between hourly wages and the quantity of hours worked per week. For every unique statistical label, this relationship was broken down and classified according to the economic sector. The graphical display exhibited the distinct salary trends present in the dataset, augmenting our comprehension of the ways in which employment across diverse economic domains impacted earnings.

These illustrations, which were made possible by the Seaborn package's power, were essential in revealing and showcasing the complex insights that were present in our dataset. They offered a visual story that helped us explore the connections and trends in the data in greater detail, which improved our capacity to make data-driven decisions and obtain a more thorough grasp of hourly wages in Ireland across a range of employment situations and economic sectors.

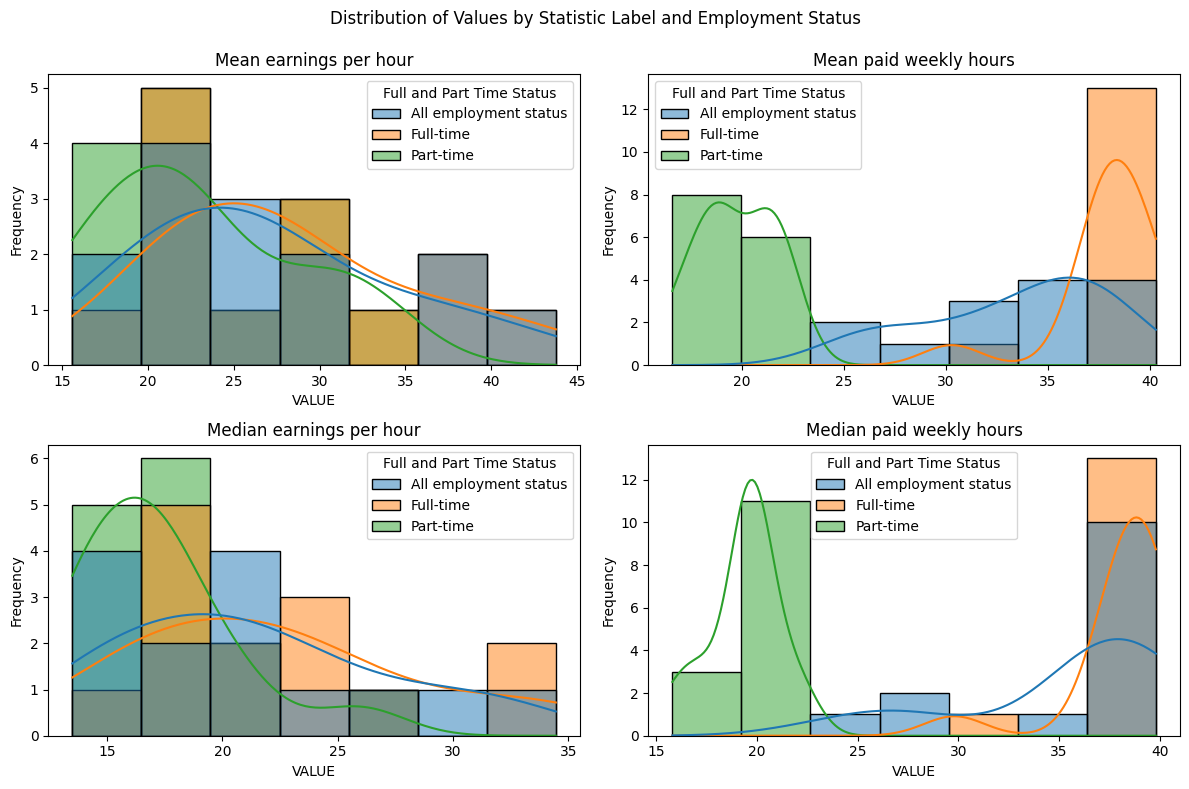


Figure 02. Bar plot of each component over years (Time)

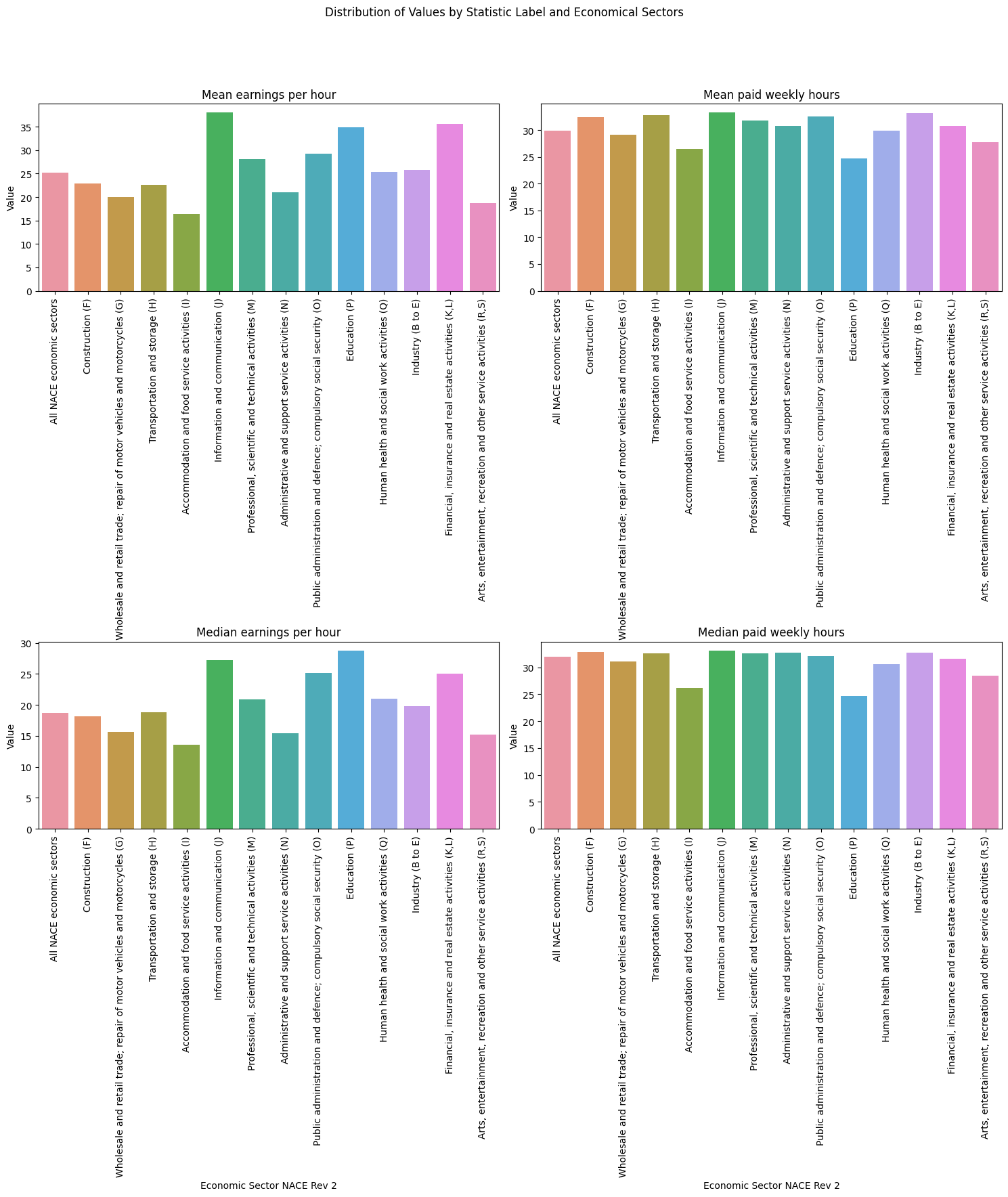


Figure 03. Bar plot of total earnings for distinct economic sectors

Machine Learning for Data Analytics

Data scaling is an important pre-processing step in machine learning with broad implications. Data scaling is the process of changing and standardizing a dataset's numerical features to make sure they fall within a predetermined range, usually between 0 and 1 or -1 and 1. When working with algorithms that are sensitive to feature magnitudes, like support vector machines, k-nearest neighbors, and neural networks, this procedure is essential. By scaling, one can ensure that each feature contributes proportionately to the model's performance, prevent the undue influence of variables with larger scales, and help the model converge more quickly during training. Additionally, it improves the readability of model coefficients, which facilitates comprehension of the relative significance of every feature. In distance-based algorithms, where unscaled features can significantly affect the metric used to calculate similarity, data scaling is also essential. In the end, data scaling is an essential method that helps machine learning models perform better and remain stable while also making it easier to understand the results and utilizing all of the data's potential to make more insightful and accurate predictions.

A statistical method for simulating the relationship between a dependent variable and one or more independent variables is regression analysis. It is a key instrument in data analysis, frequently used to comprehend how variations in one or more predictors impact the dependent variable. One of the most basic and popular types of regression is linear regression. It is assumed in linear regression that there is a linear relationship between the independent and dependent variables, which can be expressed as a straight-line equation. In order to minimize the sum of squared differences between the predicted and actual values, the best-fitting line—or hyperplane, in the case of multiple predictors—must be found. Linear regression has applications in various fields, such as economics, finance, and social sciences.

The output parameters of regression analysis provide valuable insights into the relationship between the dependent variable and independent variables. Key output parameters include the regression coefficients, intercept, R-squared (coefficient of determination), p-values, and standard errors. The regression coefficients represent the change in the dependent variable associated with a one-unit change in the corresponding independent variable, offering quantitative measures of the predictors' impact. 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 meticulously prepared during the exploratory data analysis (EDA) phase serves as the foundation for constructing regression models that aim to predict earnings based on employment status. This analysis employs both Linear Regression and K-nearest neighbors (KNN) Regressor models, allowing us to assess their performance and capabilities. To ensure a robust evaluation, our data was initially partitioned into training and validation sets using the train\_test\_split() function from the sklearn package, with a 20% allocation for validation. The outcomes of these regression models are comprehensively presented in Table 01, featuring essential metrics such as mean square error and R-squared values for each statistical label.

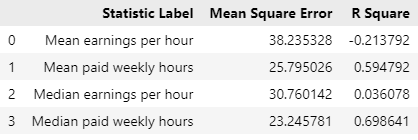


Table 01. Output parameters of linear regression for each statistic label

In the realm of K-nearest neighbors (KNN) Regression, the pivotal hyperparameter "k" holds the key to determining the plane of nearest neighbors that is assigned to an incoming data sample. To ensure that our model is finely tuned and provides optimal regression results, we embarked on a rigorous analysis of the K value, ranging from 1 to 20. The results of this exploration are meticulously captured in Figure 04 and 05, showcasing the dynamic trends of Mean Square Error (MSE) and R-squared values across the spectrum of K values.

The importance of parameter selection in the KNN model is highlighted by this thorough analysis of the ideal hyperparameter, which also offers a graphic depiction of how various K values affect the precision and accuracy of the predictions. We can determine the hyperparameter configuration that yields the most robust and dependable regression results by finding the K value that equates to the lowest MSE and the most advantageous R-squared value. The significance of data analytics in the creation and improvement of machine learning models is highlighted by these insights, which are crucial for improving the model's predictive power.