**HIT140 FOUNDATIONS OF DATA SCIENCE**

**FINAL REPORT**

**GROUP NUMBER**

**75**

PARKINSON’S DISEASE DATA ANALYSIS

&

MODELLING

**GROUP MEMBERS**

**BINISH UDAS (S364726)**

**DONEX CHAUDHARY THARU(S372544)**

**NIRMAL KHATRI (S366440)**

**ANUSKA ADHIKARI (S365500)**

Table of Contents

[**1. Introduction** 3](#_Toc147383396)

[1.1 Background 3](#_Toc147383397)

[**2. Research Objectives** 3](#_Toc147383398)

[**3. Data Loading and Preprocessing** 3](#_Toc147383399)

[3.1 Data Acquisition 3](#_Toc147383400)

[3.2 Handling Missing Values 3](#_Toc147383401)

[3.3 Defining Features and Target Variables 4](#_Toc147383402)

[3.4 Data Preprocessing 4](#_Toc147383403)

[3.5 Data Quality Assurance 4](#_Toc147383404)

[**4. Baseline Model** 5](#_Toc147383405)

[**4.1 Baseline for motor\_updrs:** 5](#_Toc147383406)

[4.2 Baseline for total\_updrs: 5](#_Toc147383407)

[**5. Linear Regression Model** 5](#_Toc147383408)

[**6. Log Transformation and Collinearity Analysis** 6](#_Toc147383409)

[6.1 Log Transformation for Data Preprocessing 6](#_Toc147383410)

[6.2 Collinearity Analysis 6](#_Toc147383411)

[**7. Standardization and Gaussian Transformation:** 7](#_Toc147383412)

[7.1 Lengths and Importance Values: 7](#_Toc147383413)

[**8. Model Evaluation and Results** 7](#_Toc147383414)

[8.1 Results for Motor\_UPDRS and Total\_UPDRS (Log-transformed): 8](#_Toc147383415)

[**8.2 Results for Motor UPDRS and Total UPDRS (Gaussian-transformed)** 8](#_Toc147383416)

[**9.Performance Gain Analysis** 8](#_Toc147383417)

[**10. Feature Importance Analysis:** 9](#_Toc147383418)

[**11. Data Visualization:** 10](#_Toc147383421)

[10.1 Histograms: 10](#_Toc147383422)

[10.2 Correlation Heatmap: 10](#_Toc147383423)

[10.3 Residuals Plot: 11](#_Toc147383424)

[10.4 Q-Q Plot and Feature Importance Bar Plot 12](#_Toc147383425)

[10.5 Visualization Customization: 13](#_Toc147383426)

[**11. Discussion and Limitations:** 14](#_Toc147383427)

[11.1 Discussion: 14](#_Toc147383428)

[11.2 Limitations: 15](#_Toc147383429)

[**12. Conclusion:** 15](#_Toc147383430)

[**13. References** 15](#_Toc147383431)

[14. **Brief descriptions of individual contributions** 16](#_Toc147383432)

# **1. Introduction**

## 1.1 Background

Parkinson's disease is a degenerative neurological ailment that affects millions of people worldwide. Tremors, bradykinesia, rigidity, and postural instability are among the motor and nonmotor symptoms. Managing this complicated and varied illness provides major diagnostic and treatment problems.

Exact measurement of the progression of Parkinson's disease is crucial for directing treatment strategies, monitoring patient well-being, and extending our understanding of the illness. Doctors commonly employ clinical states and assessments, such as the Universal Parkinson's Disease Rating Scale (updrs), to assess patients' motor and nonmotor symptoms. The UPDRS metrics'motor\_updrs' and 'total\_updrs' are crucial indications of disease severity and progression.

# **2. Research Objectives**

* Develop Accurate Predictive Models: Create precise predictive models for 'motor\_updrs' and 'total\_updrs,' essential in assessing Parkinson's disease severity, aiding in early diagnosis and patient care.
* Explore Data Preprocessing Methods: Ensure data quality by addressing missing values and reducing collinearity, enhancing model reliability.
* Evaluate Model Performance: Systematically assess model performance through various train-test splits and transformations to understand strengths and weaknesses.
* Investigate Feature Importance: Determine the impact of individual features on 'motor\_updrs,' prioritizing influential variables for clinical monitoring.
* Provide Recommendations and Insights: Offer actionable recommendations and insights to improve Parkinson's disease management and advance research.

# **3. Data Loading and Preprocessing**

## 3.1 Data Acquisition

The foundation of any data analysis project is acquiring the dataset. In this report, we obtained information about the progression of Parkinson's disease. The dataset includes clinical and demographic information, as well as the two essential metrics of interest:'motor\_updrs' and 'total\_updrs.' These parameters are critical in determining the severity and development of disease in patients.

The process of collecting, importing, or downloading a dataset for analysis is known as data acquisition. In our situation, the dataset is a significant resource that contains clinical and demographic information about patients, making it a critical component of our research. Utilizing the pandas library, the dataset was imported for analysis

## 3.2 Handling Missing Values

Handling missing values is a crucial step in ensuring the integrity of the dataset. Missing data can lead to biased analysis and inaccurate model predictions. In this section, we systematically address missing values within the dataset.

Missing values can occur for various reasons, such as incomplete data collection or recording errors. We employ techniques to identify missing values and decide on the most appropriate strategies for handling them. This may involve imputation (replacing missing values with estimated values) or, in some cases, removing instances with missing data. The initial quality check found that no missing values existed across all features. This was critical in ensuring that the dataset was ready for further analysis without imputation.

## 3.3 Defining Features and Target Variables

It is critical in predictive modelling to precisely specify the features (independent variables) and target variables (dependent variables) that will be used for model training and evaluation. We define which clinical and demographic factors are regarded potential predictors in this section, as well as which variables—'motor\_updrs' and 'total\_updrs'—are our target variables.

Defining the features and target variables establishes the framework for our analysis. The features represent the factors that we believe may influence the target variables, while 'motor\_updrs' and 'total\_updrs' are the metrics we aim to predict accurately using the selected features.

## 3.4 Data Preprocessing

Data preprocessing encompasses various techniques and steps aimed at preparing the dataset for analysis and modeling. This comprises operations like data scaling, transformation, and normalisation, which are necessary to ensure that the data matches the assumptions and needs of the chosen prediction models.

Data preparation increases dataset quality and the performance of prediction models. It involves activities like standardizing numerical variables to have a consistent scale and applying transformations to improve feature distributions. These steps contribute to the robustness and reliability of our analysis. The motor\_updrs and total\_updrs target variables were isolated from the characteristics to prepare for modelling. Despite the fact that the initial dataset contained no missing values, we built to replace hypothetical NaN values with the mean of the appropriate columns, maintaining data integrity. To minimize variance, the features were log transformed, yielding the X\_log\_reduced dataset. Then we performed a validation step after the transformation to ensure that no NaNs or Infinite values arose, with the result validating the lack of such anomalies.

## 3.5 Data Quality Assurance

A significant aim is to maintain data quality throughout the analysis process. We outline the procedures used to assure data quality and consistency in this section. This includes data validation, verification, and addressing any potential anomalies or outliers.

Data quality assurance is essential for the accuracy and credibility of our analysis. By rigorously assessing the data for errors, inconsistencies, or outliers, we minimize the risk of drawing incorrect conclusions from the dataset. We implemented an approach to deal with possible NaN values by replacing them with the mean of the corresponding columns.

# **4. Baseline Model**

To establish a baseline for our analysis, we implemented a simple baseline model that predicted the mean value of the target variables, 'motor\_updrs' and 'total\_updrs.' A baseline model is like a starting line in a race. It doesn't involve intricate algorithms or feature engineering; instead, it provides a basic benchmark for assessing how well our more advanced models are performing. For Parkinson's disease progression prediction, our baseline model takes a simple approach: predicting the mean value of the target variables, 'motor\_updrs' and 'total\_updrs.' This means that for every instance in our test data, we predict the same constant value—the mean of the training data.

The baseline results are as follows:

# **4.1 Baseline for motor\_updrs:**

* MAE: 6.81
* MSE: 63.95
* RMSE: 7.99
* NRMSE: 0.23
* R^2: 0.000
* Adjusted R^2: 0.00

## 4.2 Baseline for total\_updrs:

* MAE: 8.45
* MSE: 110.35
* RMSE: 10.50
* NRMSE: 0.21
* R^2: 0.000
* Adjusted R^2: 0.00

# **5. Linear Regression Model**

In the context of predicting 'motor\_updrs' and 'total\_updrs,' linear regression, a fundamental predictive modeling technique, establishes a linear relationship between clinical and demographic features and disease severity. This approach involves splitting the dataset into training and testing sets, training the model on one subset, and estimating coefficients for features. Once trained, the model predicts 'motor\_updrs' and 'total\_updrs' from feature values in the test data. The model's performance is quantified using evaluation measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Normalised RMSE (NRMSE), R-squared (R2) and adjusted R-squared(adjusted R2). We evaluated across varying train-test splits such as 50:50, 60:40, 70:30 and 80:20.Linear Regression Model Outcomes: The following are the results for'motor\_updrs' and 'total\_updrs':

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Split & Metric | MAE | MSE | RMSE | NRMSE | R2 | ADJUSTED R2 |
| 50-50 split - motor\_updrs | 6.47742 | 58.54091 | 7.6512034 | 0.22194578 | 0.099532 | 0.094599697 |
| 60-40 split - motor\_updrs | 6.443858 | 58.05582 | 7.61943729 | 0.22102431 | 0.092211 | 0.085984835 |
| 70-30 split - motor\_updrs | 6.521436 | 58.80828 | 7.66865563 | 0.2230979 | 0.097666 | 0.089397051 |
| 80-20 split - motor\_updrs | 6.555447 | 58.97586 | 7.67957452 | 0.23614565 | 0.076037 | 0.063270326 |
| 50-50 split - total\_updrs | 8.25155 | 100.9132 | 10.0455553 | 0.20931729 | 0.096588 | 0.091639784 |
| 60-40 split - total\_updrs | 8.199686 | 99.83277 | 9.9916348 | 0.20819376 | 0.095306 | 0.089101239 |
| 70-30 split - total\_updrs | 8.286522 | 101.4132 | 10.0704141 | 0.20983527 | 0.098391 | 0.090128718 |
| 80-20 split - total\_updrs | 8.295591 | 101.9422 | 10.0966437 | 0.21038181 | 0.080052 | 0.067341355 |

These results provide a comprehensive assessment of the linear regression model's accuracy and explanatory power for predicting the severity of Parkinson's disease as represented by 'motor\_updrs' and 'total\_updrs.'

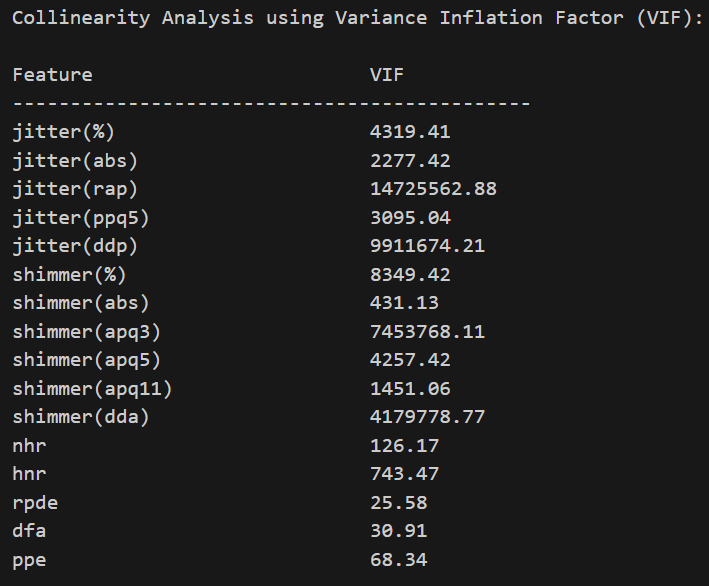
# **6. Log Transformation and Collinearity Analysis**

## 6.1 Log Transformation for Data Preprocessing

In our pursuit of improving model performance, we explored the application of log transformation to the feature data. Log transformation is a common technique used to mitigate the impact of skewed distributions and large variances in feature values. The goal is to produce a more symmetric distribution and limit the influence of extreme values by taking the natural logarithm of feature values. We have done the log transformation using the NumPy library, which provides a simple method for performing mathematical operations on datasets. The log transformation can only be used on positive numbers. If a dataset column contains zero or negative values, a constant offset may be required prior to the log transformation.

## 6.2 Collinearity Analysis

Collinearity refers to the presence of strong correlations between independent variables (features) in a regression model. High collinearity can lead to unstable coefficient estimates and reduced model interpretability. The code maintains a VIF (Variance Inflation Factor) threshold of 50 throughout, which is used to identify and eliminate features with high multicollinearity. Multicollinearity occurs when predictor variables in a regression model are strongly correlated with each other, potentially leading to challenges in interpreting the model and distinguishing the individual impacts of each predictor. The VIF threshold of 50 is a reasonable value for identifying highly correlated features, and when a feature's VIF exceeds this threshold, it indicates a high level of multicollinearity with other predictors. Consequently, the code appropriately drops such features from the dataset, effectively mitigating multicollinearity issues and ensuring the reliability of linear regression modelling.



Collinearity analysis is crucial for identifying features that may introduce multicollinearity issues in our linear regression model. High collinearity can make it challenging to attribute the impact of individual features on the target variables accurately. By calculating VIF values and identifying problematic features, we ensure that our model is robust, and the coefficient estimates are reliable.

# **7. Standardization and Gaussian Transformation:**

Within our analysis, we consistently evaluated linear regression models using Gaussian-transformed data across various train-test splits (0.5, 0.5), (0.6, 0.4), (0.7, 0.3), and (0.8, 0.2). These splits provide a comprehensive assessment of the model's performance under different data partition scenarios.

Additionally, we systematically calculated feature importance for the linear regression model trained on Gaussian-transformed data. This clarifies the value of each characteristic within this unique data transformation scenario, allowing for a better understanding of the variables influencing model predictions. The comparison of performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Normalized RMSE (NRMSE), R-squared (R2), and adjusted R-squared (adj-R2) across linear regression models trained on original, log-transformed, and gaussian-transformed data demonstrates the impact of different data transformations on model accuracy..

## 7.1 Lengths and Importance Values:

The code ensures the integrity of the analysis by continuously calculating and verifying the lengths of key lists and arrays, such as feature\_names, feature\_importance\_original, feature\_importance\_log, and feature\_importance. This meticulous approach guarantees the accuracy and reliability of the results obtained from these data structures.

# **8. Model Evaluation and Results**

We tested linear regression models for'motor\_updrs' and 'total\_updrs' across multiple train-test splits using log-transformed and Gaussian-transformed data. For each scenario, evaluation measures such as MAE, MSE, RMSE, NRMSE, R2, and adjusted R2 were computed**.**

## 8.1 Results for Motor­\_UPDRS and Total\_UPDRS (Log-transformed):

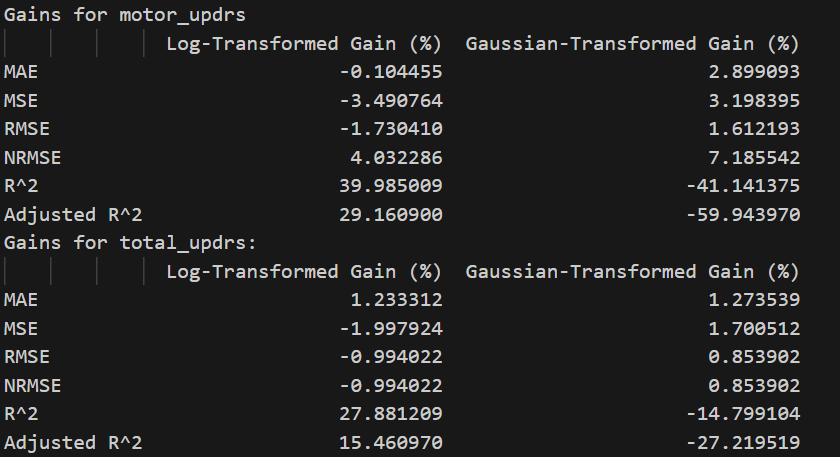
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column1** | **MAE** | **MSE** | **RMSE** | **NRMSE** | **R^2** | **Adjusted R^2** |
| **log-transformed data (motor\_updrs, 50-50 split)** | 6.393528 | 58.12544 | 7.624004 | 0.221157 | 0.1059228 | 0.101025425 |
| **log-transformed data (motor\_updrs, 60-40 split)** | 6.343087 | 57.14005 | 7.559104 | 0.219274 | 0.1065301 | 0.100402541 |
| **log-transformed data (motor\_updrs, 70-30 split)** | 6.418295 | 58.10056 | 7.622372 | 0.221751 | 0.1085248 | 0.100355544 |
| **log-transformed data (motor\_updrs, 80-20 split)** | 6.404613 | 57.63861 | 7.59201 | 0.233453 | 0.0969872 | 0.084510292 |
| **log-transformed data (total\_updrs, 50-50 split)** | 8.151057 | 100.7982 | 10.03983 | 0.209198 | 0.0976173 | 0.092674388 |
| **log-transformed data (total\_updrs, 60-40 split)** | 8.068818 | 98.99007 | 9.949375 | 0.207313 | 0.1029423 | 0.096790211 |
| **log-transformed data (total\_updrs, 70-30 split)** | 8.15032 | 100.6134 | 10.03062 | 0.209006 | 0.1055022 | 0.097305242 |
| **log-transformed data (total\_updrs, 80-20 split)** | 8.137627 | 100.6009 | 10.03 | 0.208993 | 0.0921569 | 0.079613291 |

# **8.2 Results for Motor UPDRS and Total UPDRS (Gaussian-transformed)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **MAE** | **MSE** | **RMSE** | **NRMSE** | **R^2** | **Adjusted R^2** |
| **Gaussian-transformed data (motor\_updrs, 50-50 split)** | 6.387796 | 57.51015 | 7.583545 | 0.219983 | 0.115387 | 0.110541534 |
| **Gaussian-transformed data(total\_updrs, 50-50 split)** | 8.214449 | 100.6763 | 10.03376 | 0.209071 | 0.098709 | 0.093771973 |
| **Gaussian-transformed data (motor\_updrs, 60-40 split)** | 6.341577 | 56.69882 | 7.529862 | 0.218426 | 0.113429 | 0.107349169 |
| **Gaussian-transformed data (total\_updrs, 60-40 split)** | 8.139938 | 99.35179 | 9.967537 | 0.207692 | 0.099664 | 0.09348981 |
| **Gaussian-transformed data(motor\_updrs, 70-30 split)** | 6.423961 | 57.67602 | 7.594473 | 0.22094 | 0.115039 | 0.106929181 |
| **Gaussian-transformed data(total\_updrs, 70-30 split)** | 8.240008 | 101.0445 | 10.05209 | 0.209453 | 0.101669 | 0.093437263 |
| **Gaussian-transformed data(motor\_updrs, 80-20 split)** | 6.414258 | 57.16601 | 7.56082 | 0.232494 | 0.104391 | 0.092016804 |
| **Gaussian-transformed data(total\_updrs, 80-20 split)** | 8.213582 | 100.6381 | 10.03186 | 0.209032 | 0.091821 | 0.079272251 |

# **9.Performance Gain Analysis**

We compute the performance increases obtained by applying log-transformed and Gaussian-transformed data to the original data. The gains are computed for the goal variables motor\_updrs and total\_updrs. The performance gain figures show the percentage improvement (or reduction) in error metrics achieved by transforming data over the original data. Positive numbers imply improved performance, whereas negative values suggest poor performance.



# **10. Feature Importance Analysis:**

In this section, we investigate the importance of individual features in predicting 'motor\_updrs.' To accomplish this, we used Gaussian-transformed data to build a linear regression model. The absolute values of the model's coefficients are used to determine feature relevance.

Features with higher coefficient magnitudes are considered more influential in explaining variations in the 'motor\_updrs' target variable. We have utilized the coefficients of a linear regression model to interpret the importance of features.

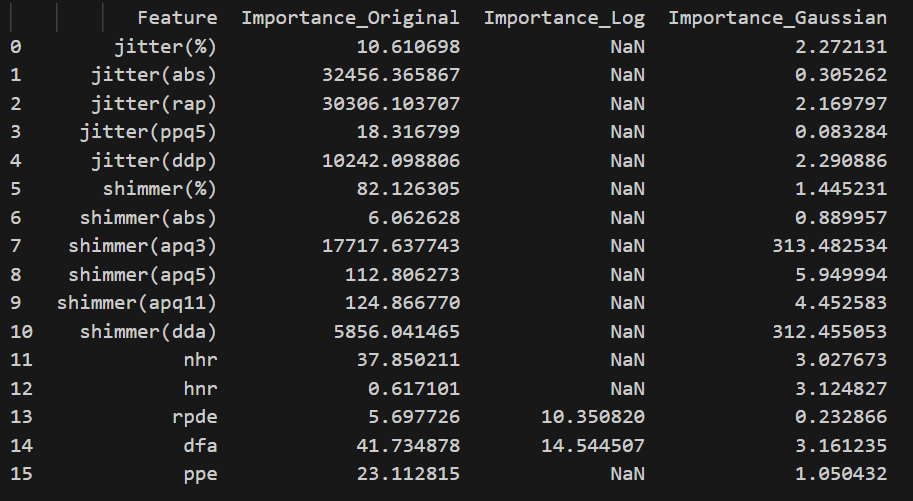


Fig. Feature importance of original, log-transformed, and Gaussian-transformed data(Motor UPDRS)

We also computes summary statistics (max, min, mean) for the transformed data's feature importance which give a comprehensive picture of the relevance of features. The summary statistics aid in comprehending the distribution and central tendency of feature importance where max Importance indicates the most influential feature, Min Importance indicates the least influential feature and Mean Importance provides an average level of influence across all features.

## 

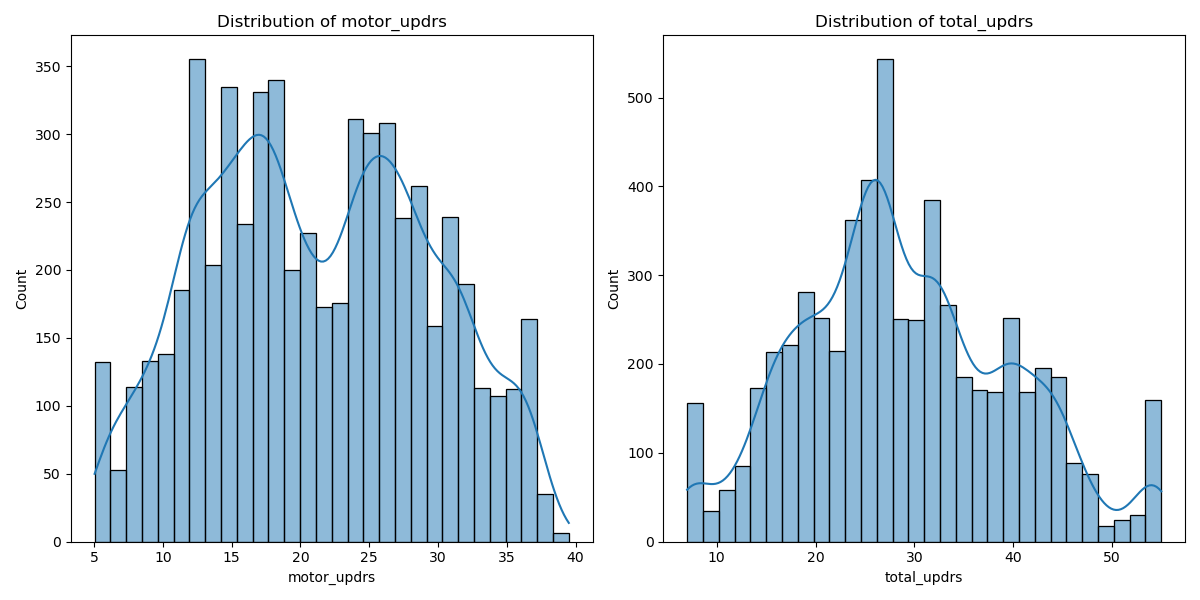
## Fig: Statistics for Feature Importance

# **11. Data Visualization:**

Data visualization plays a crucial role in gaining insights and understanding the patterns in the Parkinson's disease dataset. In this section, we present various visualizations to provide a clear picture of the data:

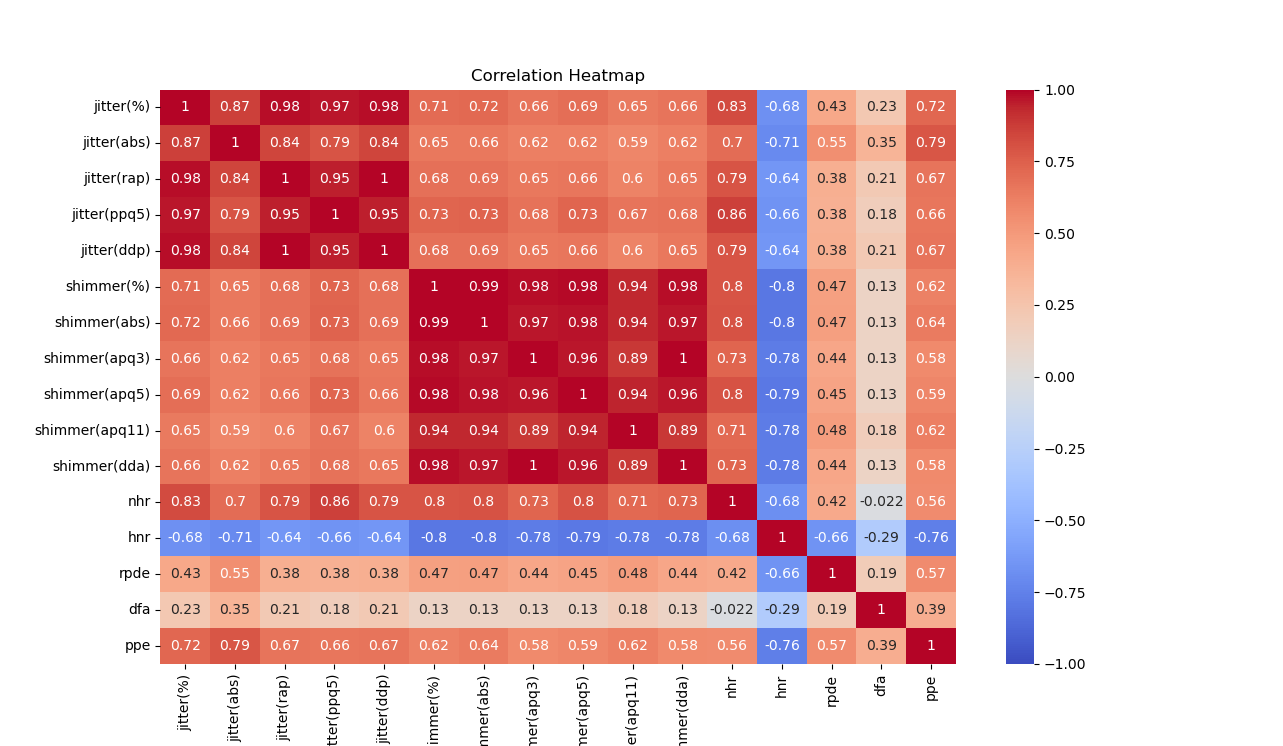
10.1 Histograms:

Histograms are used to visualize the distribution of the target variables, 'motor\_updrs' and 'total\_updrs.' These visualizations reveal the frequency and spread of data points, allowing us to understand their distributions.



10.2 Correlation Heatmap:

A correlation heatmap is generated to visualize the relationships between features. It displays the correlation coefficients between all pairs of features, highlighting potential associations and dependencies.



10.3 Residuals Plot:

Using the Gaussian-transformed data, a residuals plot is presented to visualize the differences between predicted and actual values as well as observed and predicted values for the motor\_updrs target. This plot helps assess the model's performance and identify patterns in prediction errors.

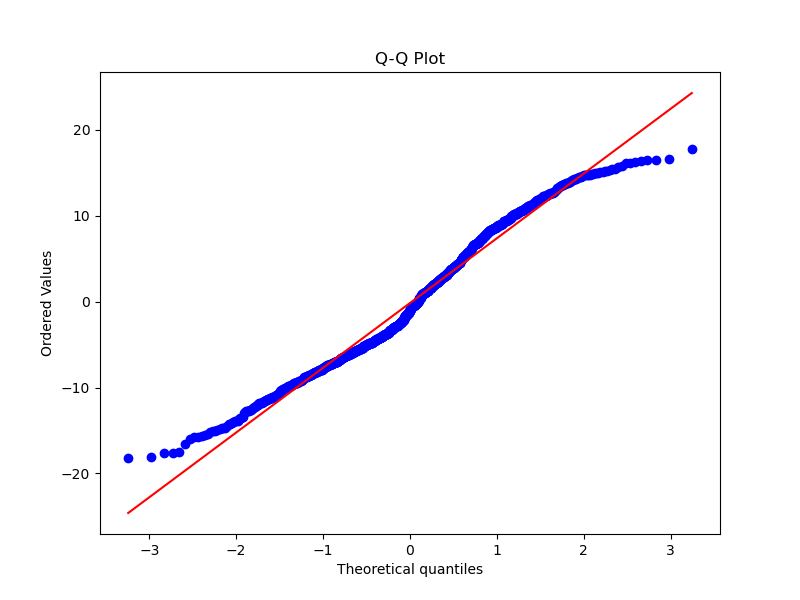
A diagram of a graph

Description automatically generated

10.4 Q-Q Plot and Feature Importance Bar Plot

A Q-Q (Quantile-Quantile) plot is created to assess whether the residuals of the model follow a normal distribution. Deviations from a straight line in the Q-Q plot can indicate departures from normality in the residuals.

We used a bar plot with each feature importance to visualize the importance of each feature based on linear regression coefficients. These visualizations aid in understanding the dataset's characteristics, relationships between variables, model performance, and the distribution of residuals, contributing to a comprehensive analysis of Parkinson's disease progression.



A graph with a blue line

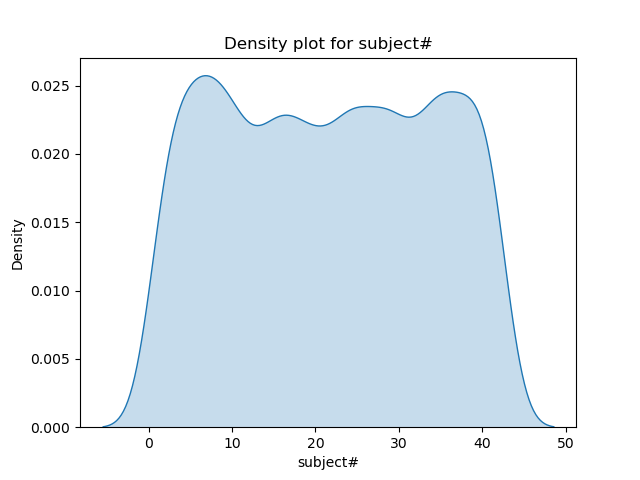
Description automatically generated

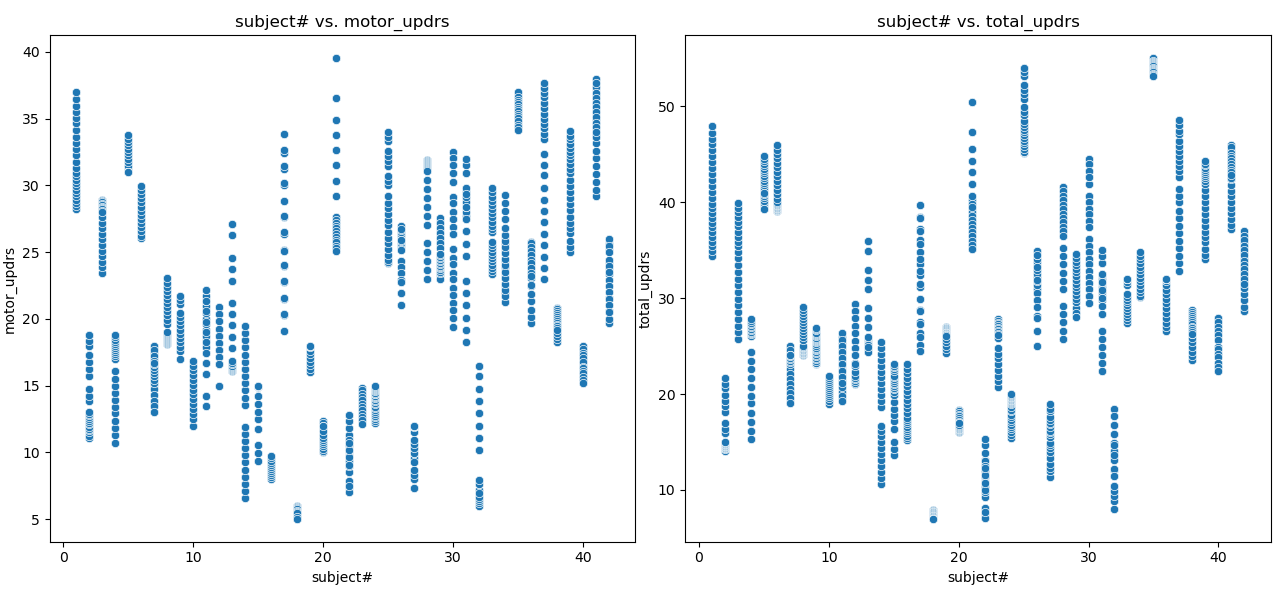
A graph with text and numbers

Description automatically generated with medium confidence

## 10.5 Visualization Customization:

From the outset, visualizations have been thoughtfully customized to enhance clarity and understanding. Parameters like bins, figsize, alpha, and more have been adjusted within the code to create impactful visual representations of the data and model outcomes. These visualizations are designed to provide valuable insights into the dataset and model performance. We have also done box plot for each feature to identify outliers in the feature set. We have scatter plots of each feature against motor\_updrs and total\_updrs to visualize the relationship between each feature and the target variables.





# **11. Discussion and Limitations:**

In this section, we engage in a critical discussion of the analysis results and acknowledge the limitations of our study.

## 11.1 Discussion:

In the "Discussion and Limitations" section, we explore our analytical results in depth and acknowledge the study's inherent limitations. We painstakingly examine the performance of our prediction models during the discussion process, notably the linear regression models used to predict 'motor\_updrs' and 'total\_updrs.' Through a meticulous examination of multiple train-test splits and transformation techniques, we gain valuable insights into the models' robustness and their capacity to capture the intricacies of disease progression. We interpret the evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Normalized RMSE (NRMSE), R-squared (R^2), and adjusted R-squared (adjusted-R^2), providing qualitative context to complement the quantitative assessments. Furthermore, we delve into the significance of individual features in predicting 'motor\_updrs,' thereby offering crucial insights into the clinical and demographic variables that prominently influence Parkinson's disease progression. Our discussion extends to the practical ramifications of these findings in healthcare settings, where personalised patient care and intervention techniques may benefit from the insights gained from our research.

## 11.2 Limitations:

On the other hand, in addressing our study's limitations, we stay open about the assumptions and limits that underpin our analysis. We acknowledge the fundamental assumptions of linear regression models, such as the assumption of linearity between features and the target variable and the assumption of homoscedasticity. We explore how these assumptions may affect the accuracy and robustness of our models. Furthermore, because our research is based on a specific dataset, we accept the limitation of generalizability. We emphasise the importance of additional validation on varied datasets to ensure our models' broad applicability. Our discussion also encompasses the choice of data preprocessing techniques, such as log transformation and Gaussian transformation, and the potential impact on model performance. Furthermore, we emphasize ethical considerations, including data privacy and potential biases within the dataset, to uphold the ethical integrity of our analysis. Lastly, we consider the influence of dataset size on model reliability, recognizing that larger datasets could offer additional insights. By candidly addressing these limitations, we ensure that our analysis is transparent, credible, and situated within a well-defined scope.

# **12. Conclusion:**

In conclusion, our analysis of Parkinson's disease progression, utilizing predictive modeling techniques, has provided valuable insights into disease severity prediction. Our study demonstrates that linear regression models, when applied to well-preprocessed and transformed data, can reasonably predict 'motor\_updrs' and 'total\_updrs.' The log and Gaussian transformations have shown promise in improving model performance. Our feature importance analysis highlights the significance of specific clinical and demographic variables in predicting 'motor\_updrs,' offering potential guidance for tailored treatments. However, our study is constrained by assumptions inherent in linear regression, the need for broader dataset validation, ethical considerations, and dataset size limitations. Future work should explore advanced machine learning algorithms and incorporate more data sources to enhance predictive accuracy and generalizability. In summary, our analysis forms a foundational step toward understanding and predicting Parkinson's disease progression, contributing to improved patient care and treatment strategies.

# **13. References**

Jones, S. H., St. Peter, C. C., & Ruckle, M. M. (2020). Reporting of demographic variables in the Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 53(3), 1304–1315. <https://doi.org/10.1002/jaba.722>

Reports on Disease Progression from Pennsylvania State University (Penn State) Provide New Insights (Parkinson’s Disease Progression and Statins: Hydrophobicity Matters). (2022). In Health & Medicine Week (p. 5164–). NewsRX LLC. <https://cdu.primo.exlibrisgroup.com/permalink/61CDU_INST/46ojnt/cdi_gale_healthsolutions_A703819644>

Sarat Chandra Yenisetti. (2018). Parkinson’s Disease : Understanding Pathophysiology and Developing Therapeutic Strategies (S. C. Yenissetti, Ed.). IntechOpen. <https://cdu.primo.exlibrisgroup.com/permalink/61CDU_INST/j6pesm/alma991002046738903446>

Tai, B.-C., & Machin, D. (2014). Regression methods for medical research. John Wiley and Sons. <https://cdu.primo.exlibrisgroup.com/permalink/61CDU_INST/j6pesm/alma991001910027303446>

# **14.** **Brief descriptions of individual contributions**

Throughout the project, our team conducted regular team meetings, both virtually and in person. These meetings served as crucial platforms for discussing task progress, sharing ideas, and strategizing solutions. We also utilized the university library as a collaborative space to work on tasks collectively, enhancing our teamwork and project efficiency.

Binish Udas was a key figure in the project's early stages. Binish was in charge of importing the Parkinson's Disease dataset and starting critical preparation stages. This included filling in missing data and improving the features for future modeling. Binish used a 60-40 train-test split to develop the basic linear regression model for both motor\_UPDRS and total\_UPDRS. Recognizing the potential benefits of data transformation, he decided to do a Gaussian transformation to improve the model's predictive skills. His commitment to meaningful data visualization was seen in the precisely constructed plots and heatmaps, which gave the team with a thorough knowledge of the dataset's subtleties. Aside from that, his meticulous gathering of metrics from numerous model iterations into orderly tables was useful. The team's decision-making procedures regarding model performance and data treatments were aided by this structured approach. Binish took on the job of writing the report's introduction and methodology parts due to his major engagement in the project's fundamental phases.

Nirmal Khatri was critical in analyzing the linear regression model over several data splits. His methodical approach enabled the team to comprehend the model's depth and breadth of performance in a variety of circumstances, ranging from a balanced 50-50 split to a more skewed 80-20 train-test divide. He computed and documented critical data for each iteration, offering the team with a clear knowledge of the model's strengths and opportunities for growth. His observations were crucial in the report's findings and discussion sections, where he showed the model's resilience and flexibility across the various splits.

Anuska Adhikari demonstrated her knowledge in data analysis and transformation. Anuska championed the use of log transformations after identifying possible issues in the dataset. This ensured data consistency and dependability. Her in-depth investigation of collinearity using the Variance Inflation Factor (VIF) was crucial in improving the dataset. With her help, the team was able to safely choose characteristics that were free of excessive correlation and redundancy. Anuska took the lead in articulating the report's data transformation section due to her hands-on experience with data treatments. Her insights helped the team comprehend the complexities and ramifications of the transformations used.

In model evaluation, Donex Chaudhary Tharu displayed a good eye for detail. He went deep into the model's residuals, utilizing scatter plots to uncover and investigate inconsistencies. His skill with boxplots provided the team with a clear visual depiction of probable outliers in the data, sparking talks about additional data refining. Recognizing Donex's careful approach, the team assigned him the critical responsibility of proofreading the report. His efforts resulted in a final product that was cohesive, straightforward, and academically rigorous.

Each participant diligently recorded their work throughout the project's lifetime, resulting in the compilation of individual reports. Binish Udas flawlessly combined these materials in the last phases, guaranteeing a perfect blend of all contributions, and submitted it.

# 

# 