Programming for Data Analysis and AI

Table of contents

Part 1: Descriptive Analysis

In preparation for accurate data analysis, certain attribute types within the dataset underwent a necessary transformation from numeric to categorical. This transformation was imperative, particularly for ordinal data collected from a survey that had been initially represented using numeric values, a representation that lacked inherent meaning. Subsequently, a comprehensive descriptive analysis was conducted, prioritising numeric attributes to unveil crucial insights into the dataset's characteristics and distribution patterns.The descriptive analysis of the numeric attributes provided valuable information about the time-related attributes of the dataset(Figure 1). For 'time_bp' (time spent on work before the pandemic), the average duration is approximately 7.42 hours, with a standard deviation of 2.01 hours. The minimum and maximum durations are 4 and 12 hours, respectively, indicating a range of work hours before the pandemic. 'time_dp' (time spent on work during the pandemic) shows a slightly higher average of around 7.97 hours, with a larger standard deviation of 2.66 hours. The distribution of work hours during the pandemic appears to be more varied with people working more hours. Notably, the 'travel_time' attribute, representing the time spent on travel, has a mean of 1.03 hours and a standard deviation of 0.71 hours. The majority of values fall within the range of 0.5 to 1.5 hours, as indicated by the interquartile range.

	time_bp	time_dp	travel_time
count	1175.000000	1175.000000	1175.000000
mean	7.415319	7.971915	1.027660
std	2.005385	2.657007	0.713314
min	4.000000	4.000000	0.500000
25%	5.000000	5.000000	0.500000
50%	7.000000	9.000000	0.500000
75%	9.000000	9.000000	1.500000
max	12.000000	12.000000	3.000000

Figure 1: Descriptive analysis of Numeric Attributes

	count	unique	top	freq
age	1175	7	19-25	345
gender	1175	3	Male	649
occupation	1175	8	Working Professional	479
line_of_work	479	8	Teaching	217
easeof_online	1175	5	1	329
home_env	1175	5	3	327
prod_inc	1175.0	5.0	0.5	302.0
sleep_bal	1175.0	5.0	-0.5	313.0
new_skill	1175.0	5.0	0.5	366.0
fam_connect	1175.0	5.0	0.5	414.0
relaxed	1175.0	5.0	0.0	306.0
self_time	1175.0	5.0	0.0	417.0
like_hw	1175	15	100	233
dislike_hw	1175	15	1111	264
prefer	1175	2	Complete Physical Attendance	836
certaindays_hw	1175	3	Yes	568

Figure 2: Descriptive Analysis of Categorical Attributes

Next the descriptive analysis of the numeric attributes took place,Figure 2 presents a comprehensive view of the dataset, it's crucial to note that the 'line_of_work' attribute contains missing values in 696 instances, highlighting the need for careful handling in subsequent stages. Despite this, the dataset exhibits a noteworthy level of detail, with 76% of attributes featuring more than two categories which means that there is great variability within each attribute. Notably, some attributes reveal significant imbalances. For example, within the 'line_of_work' attribute, which showcases diversity with eight categories, certain professions exhibit imbalances. 'Teaching' emerges as the predominant category, representing 45% of instances within this attribute. Similarly, 'prefer' demonstrates a binary distribution, with 'Complete Physical Attendance' overwhelmingly represented at 71%. These imbalances underscore the importance of thorough exploration and consideration of potential biases in the dataset. It's worth noting that the attributes, namely 'prod_inc,' 'sleep_bal,' 'new_skill,' 'fam_connect,' 'relaxed,' and 'self_time,' are originally ordinal in nature but were initially represented using numeric values. For the purpose of the analysis, these attributes have been transformed into categorical variables. Originating from survey responses, the numeric values in each case correspond to the degree of change, where a lower numeric value signifies a lower degree of change or difficulty or likelihood, and as the numeric values increase, the degree of change difficulty and likelihood intensifies. This reclassification enhances the interpretability of these attributes and ensures a more meaningful analysis within the context of the survey data. Below in Figures 3-18 more details about each category frequency of the attributes can be found.Attribute occupation didn’t fit within a chart so a table was used instead of a count plot.

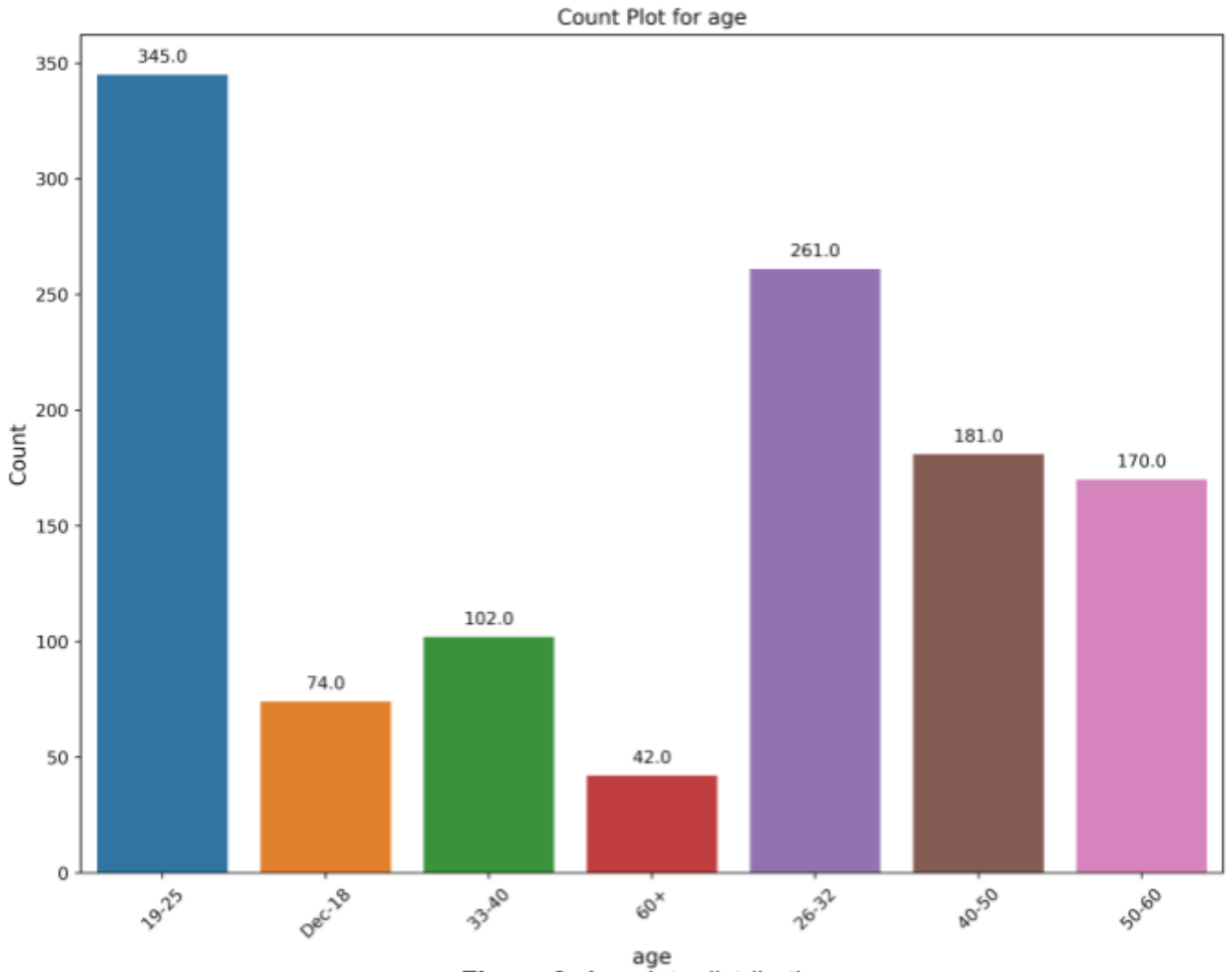

Figure 3: Age data distribution

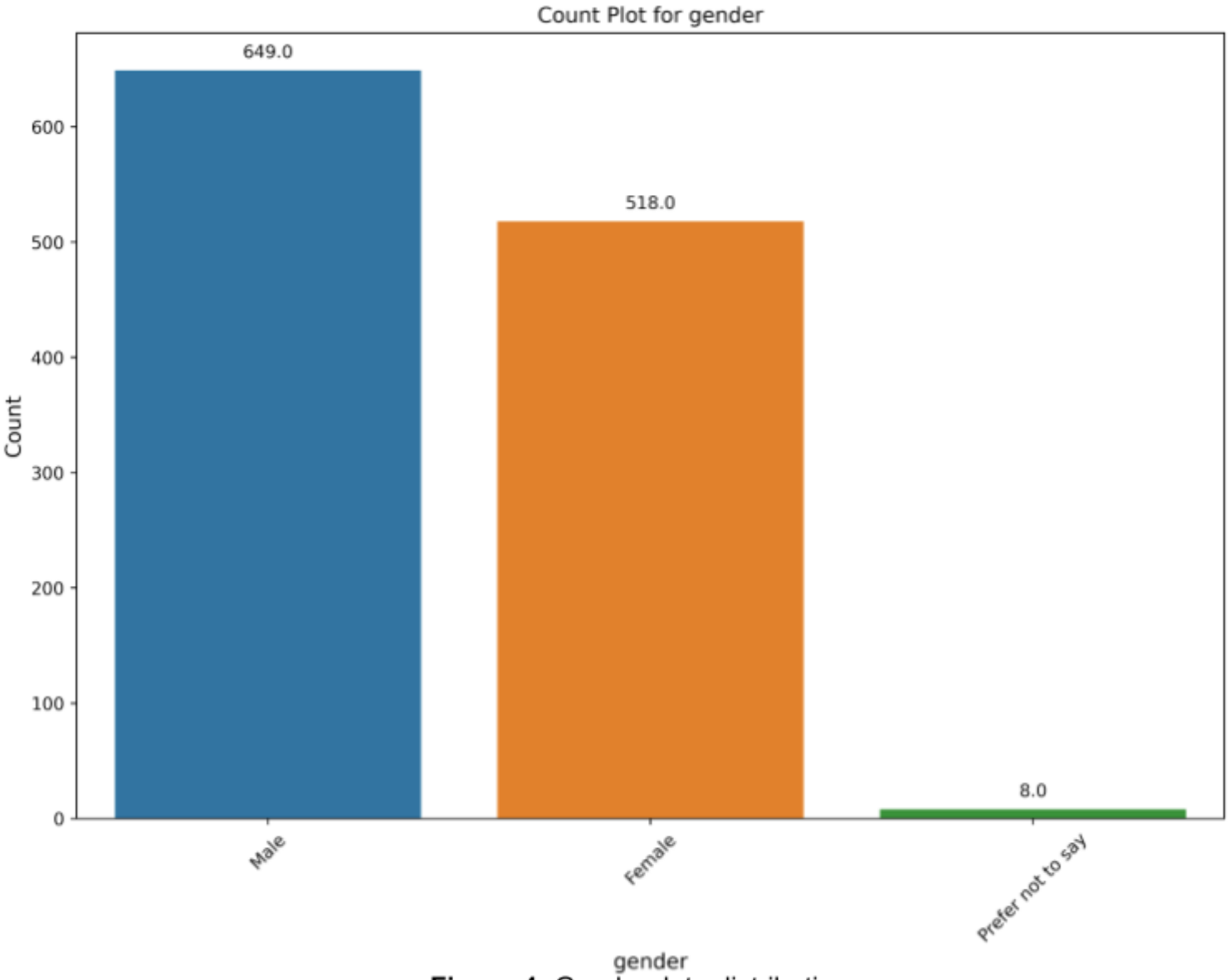

Figure 4: Gender data distribution

	occupation	Frequency
0	Currently Out of Work	44
1	Entrepreneur	119
2	Homemaker	82
3	Medical Professional aiding efforts against COVID-19	73
4	Retired/Senior Citizen	2
5	Student in College	358
6	Student in School	18
7	Working Professional	479

Figure 5: Occupation data distribution

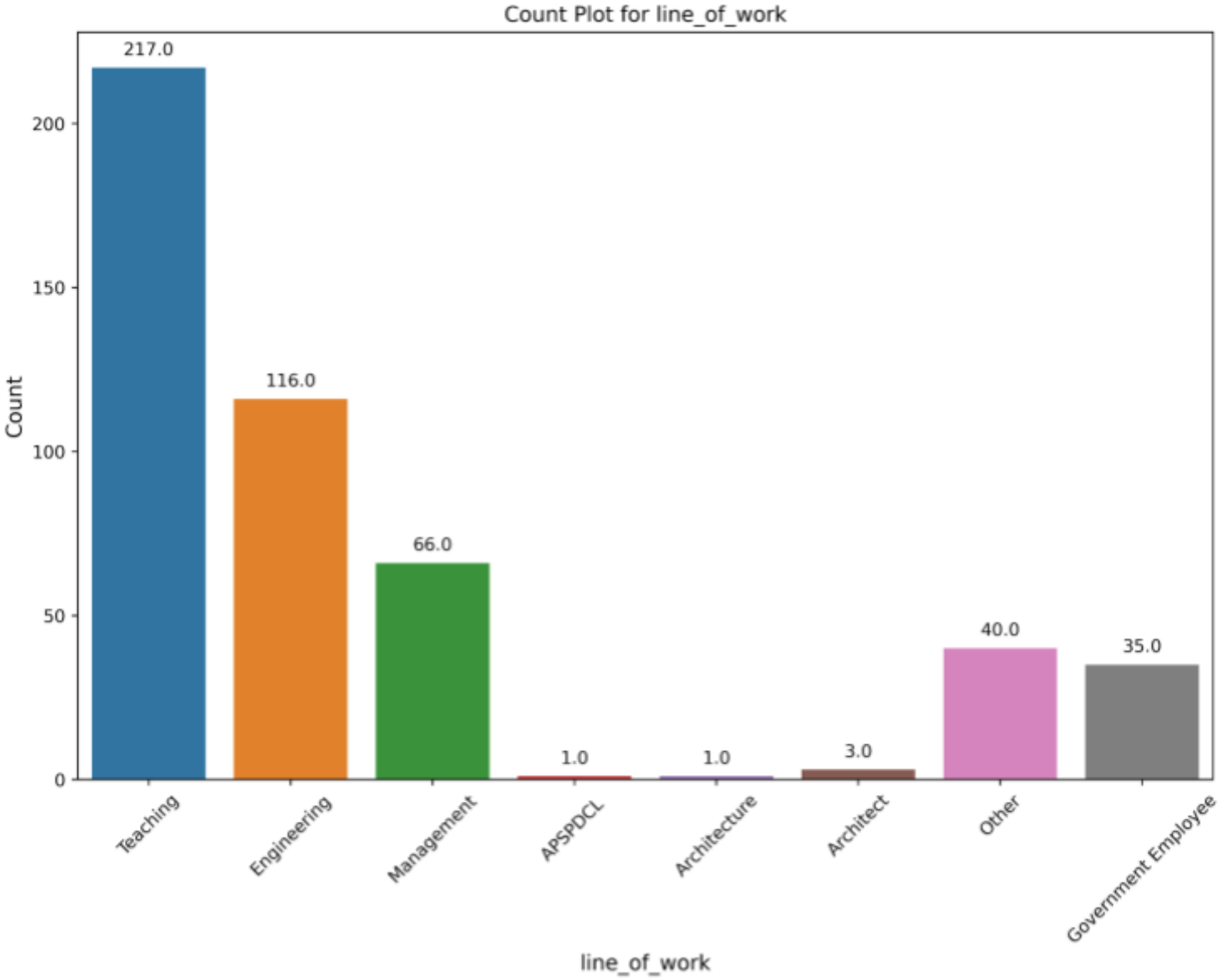

Figure 6: Line of work data distribution

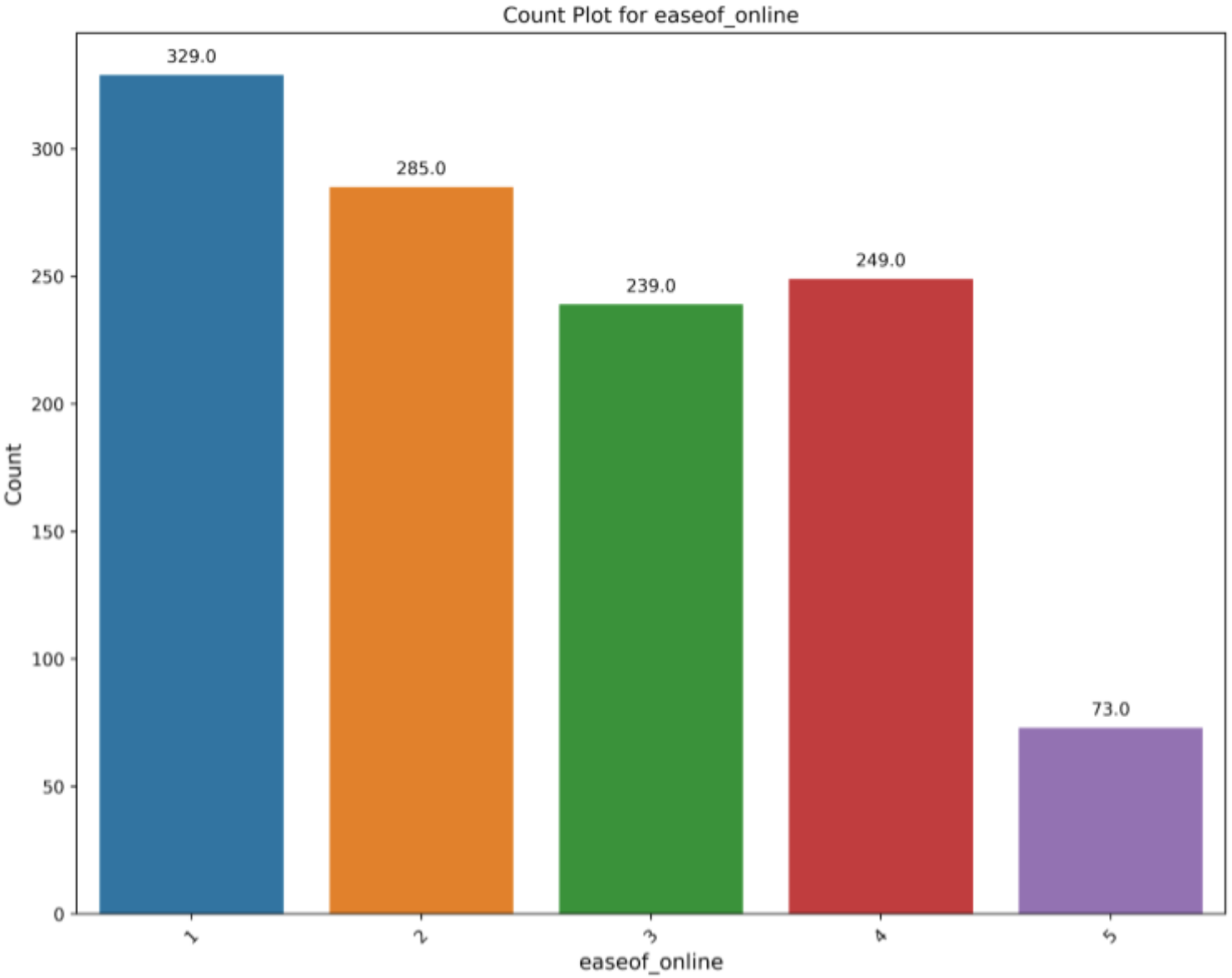

Figure 7: EaseOf Online data distribution

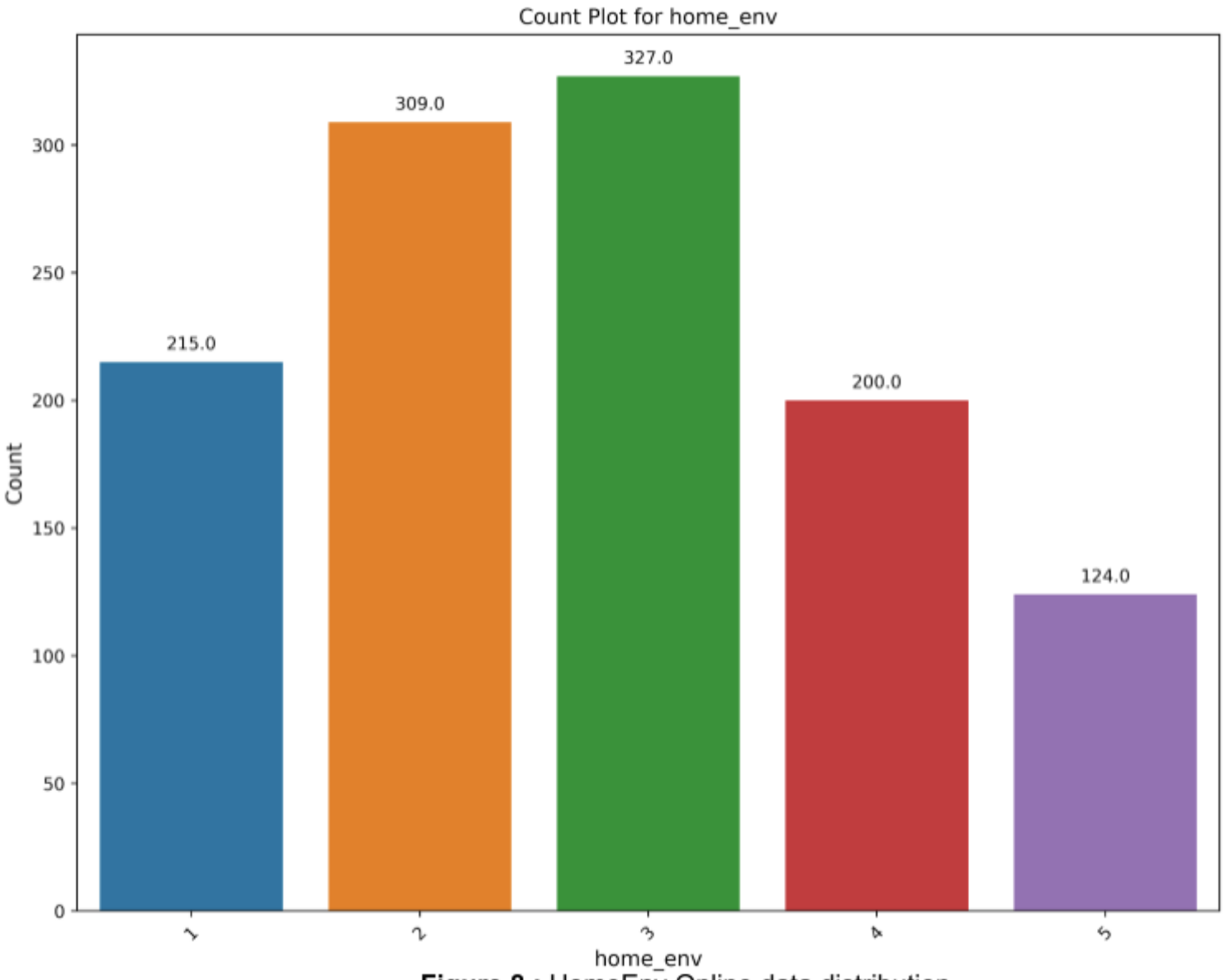

Figure 8 : HomeEnv Online data distribution

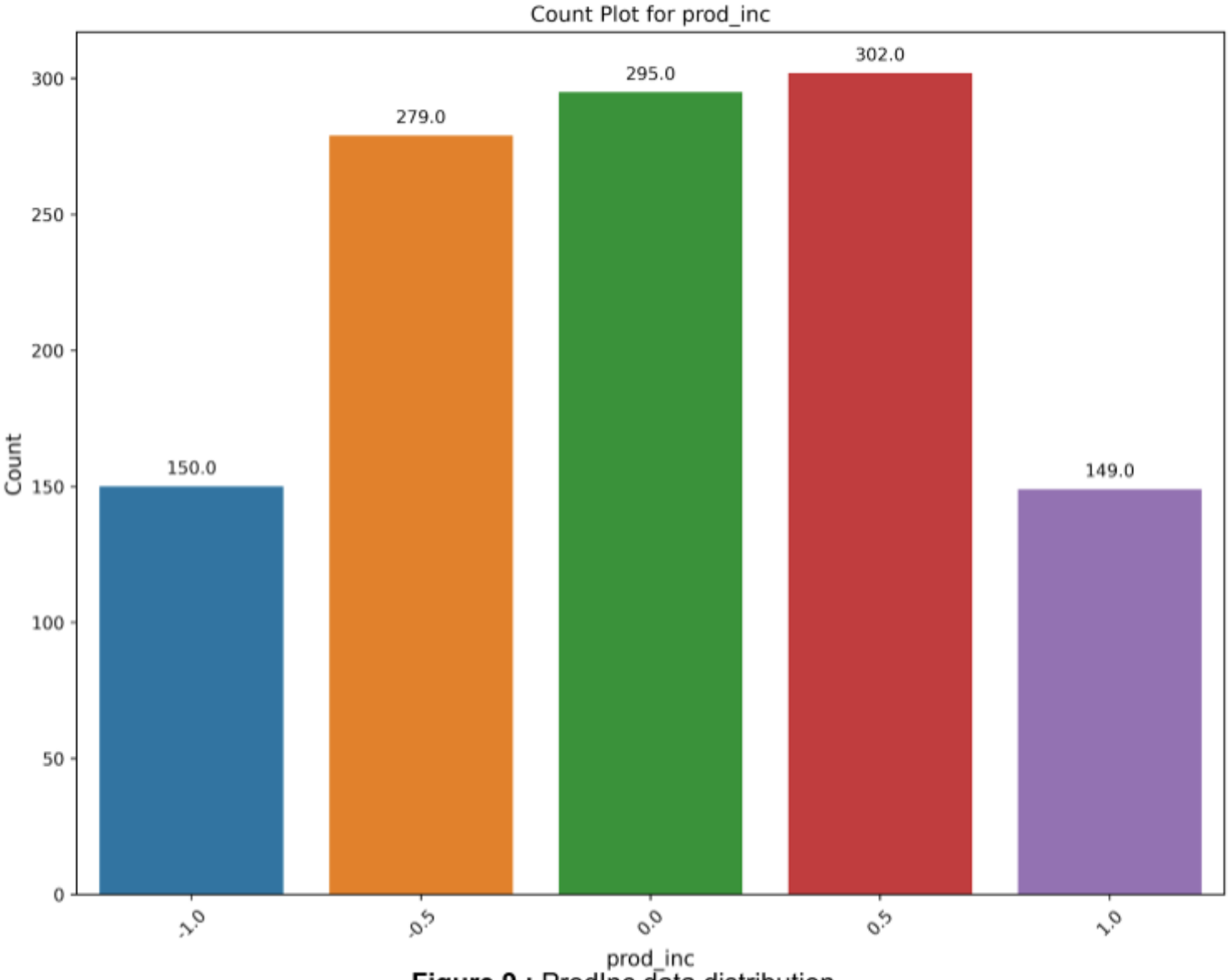

Figure 9 : ProdInc data distribution

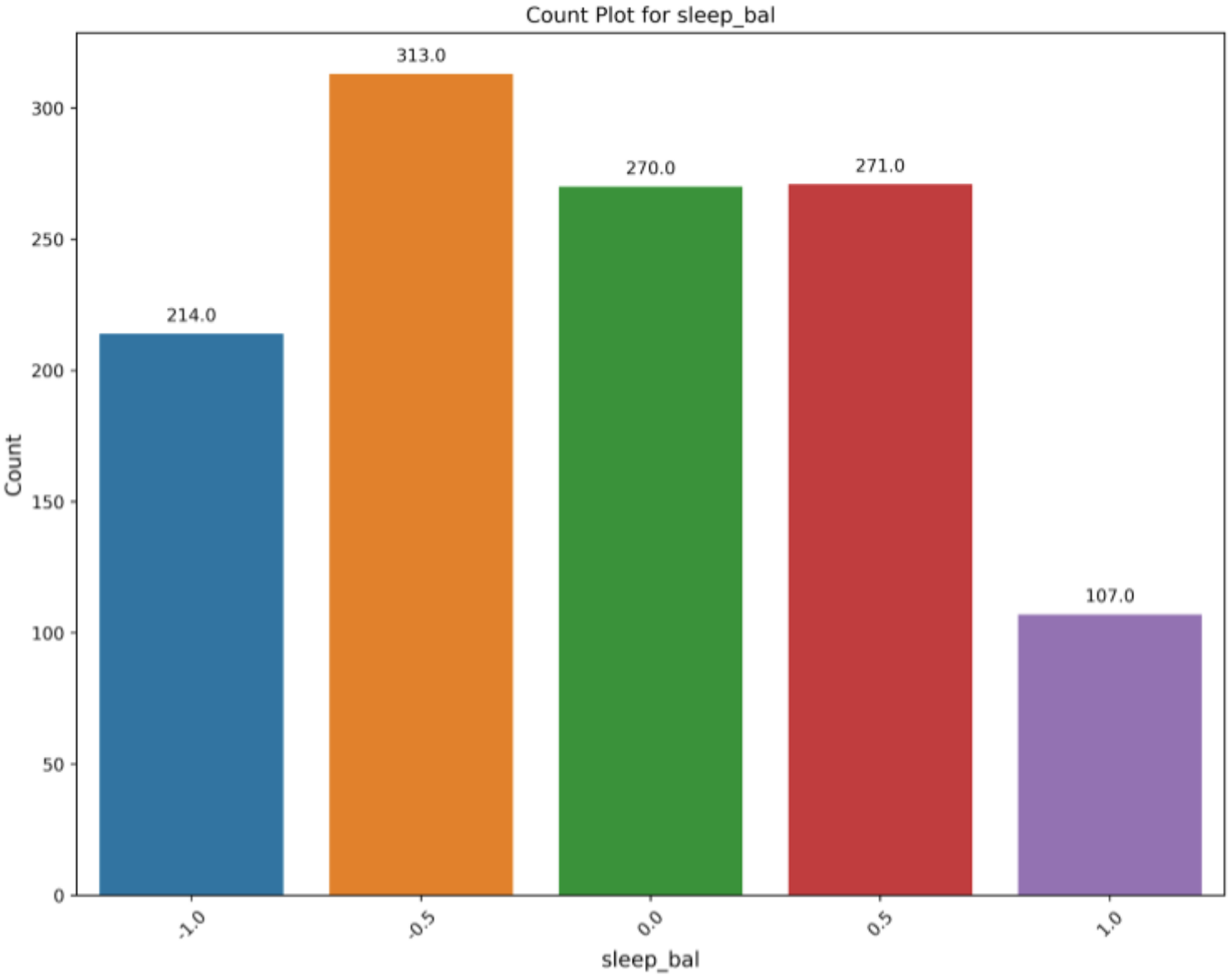

Figure 10 : SleepBal data distribution

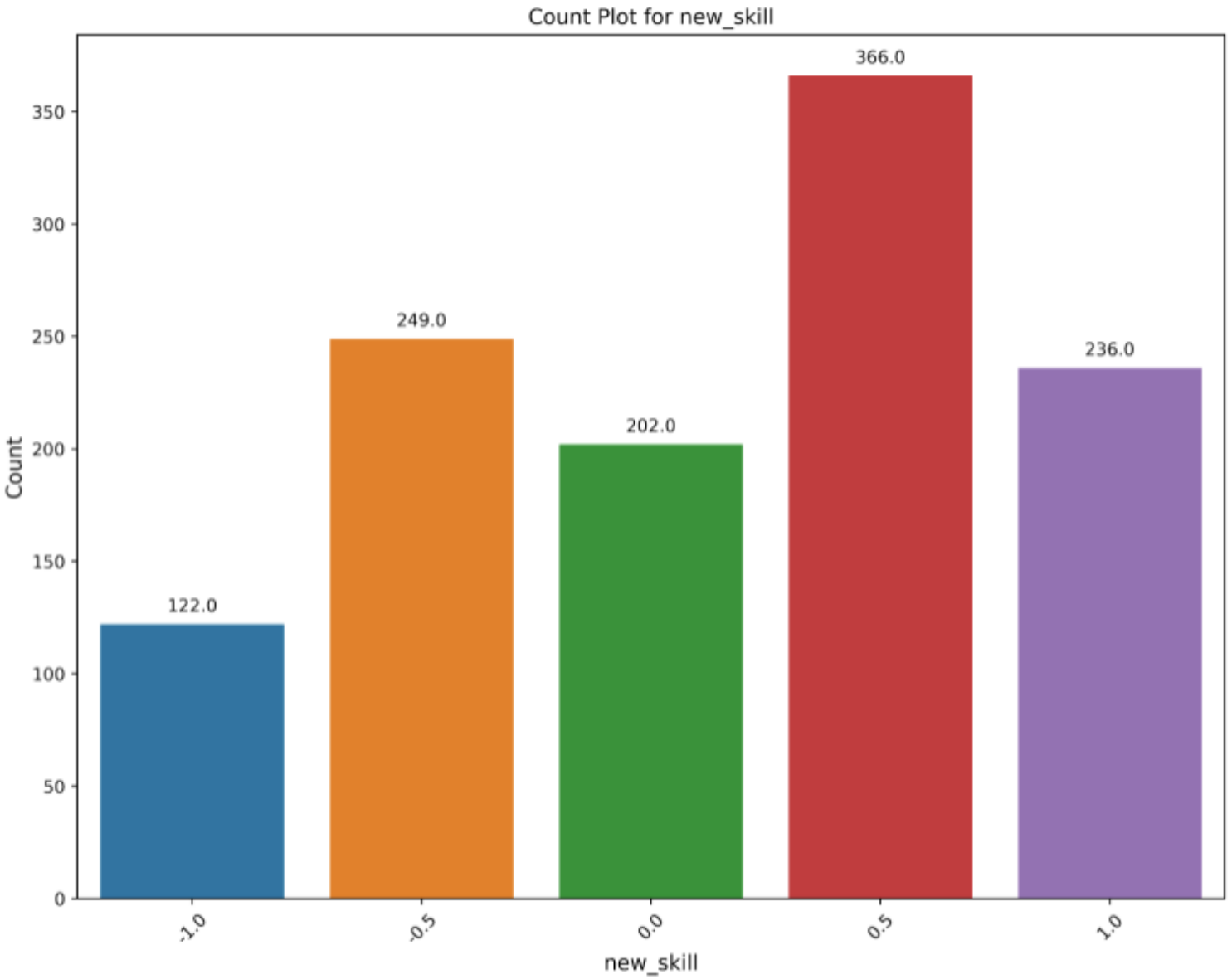

Figure 11 : NewSkill data distribution

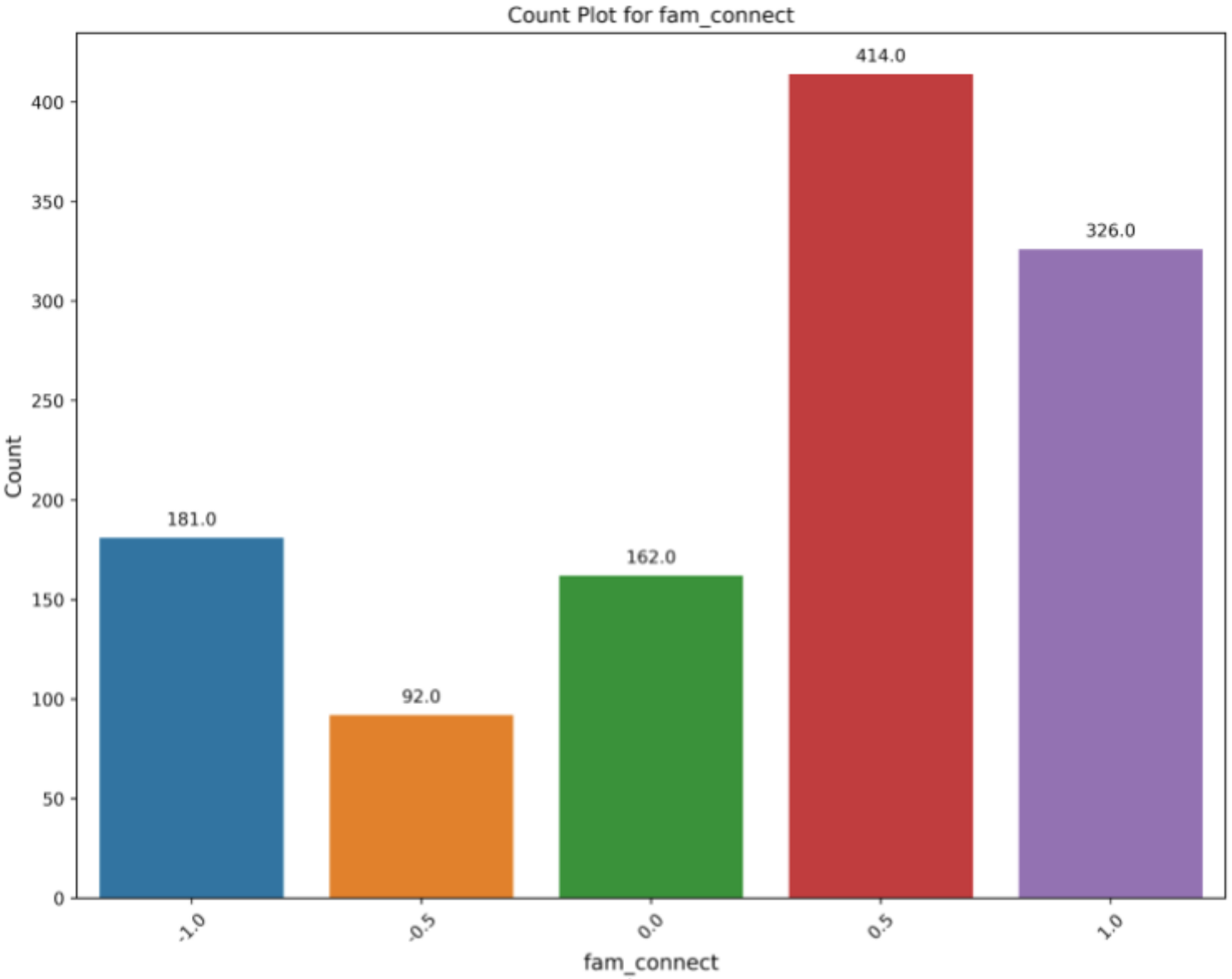

Figure 12 : FamConnect data distribution

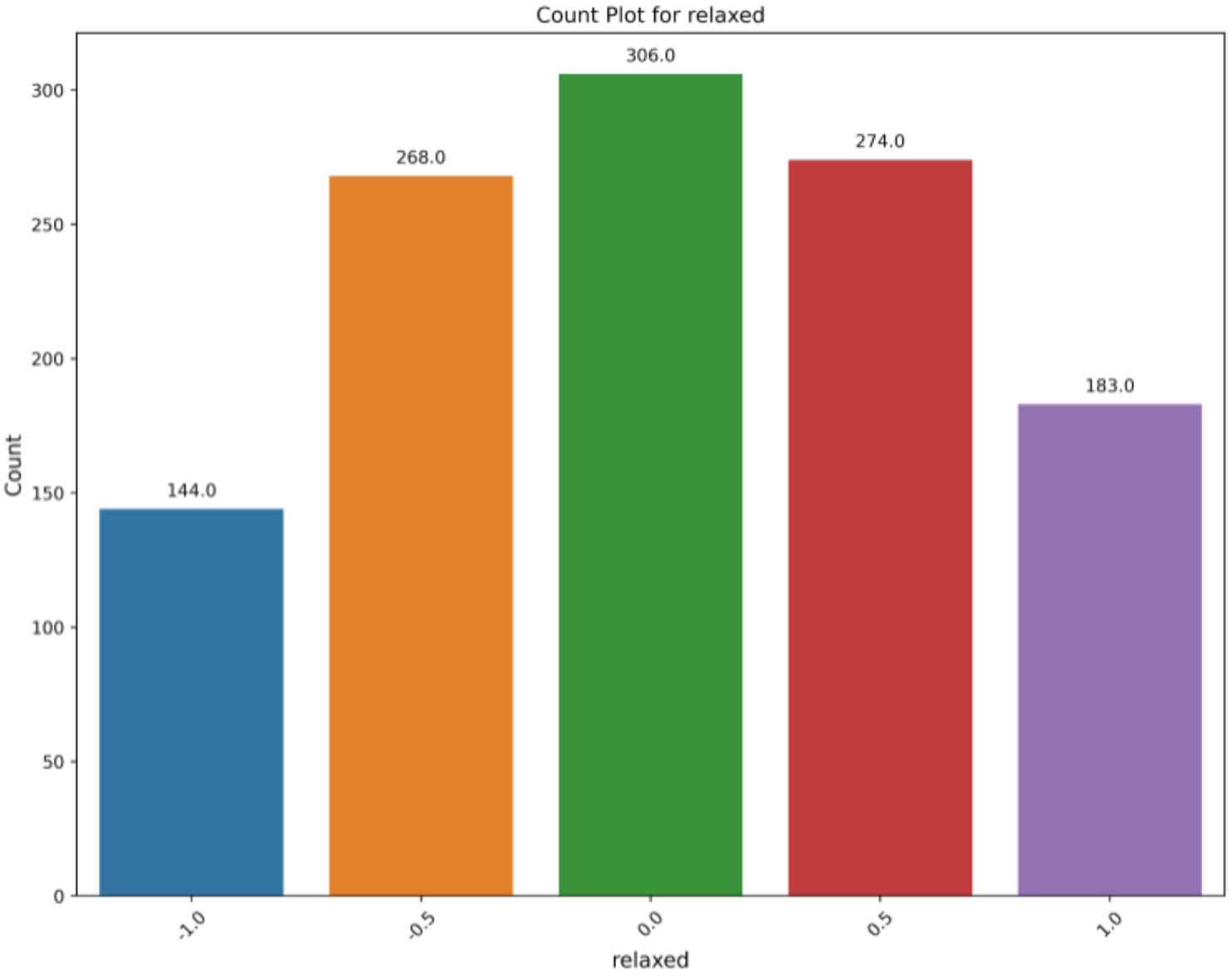

Figure 13 : Relaxed data distribution

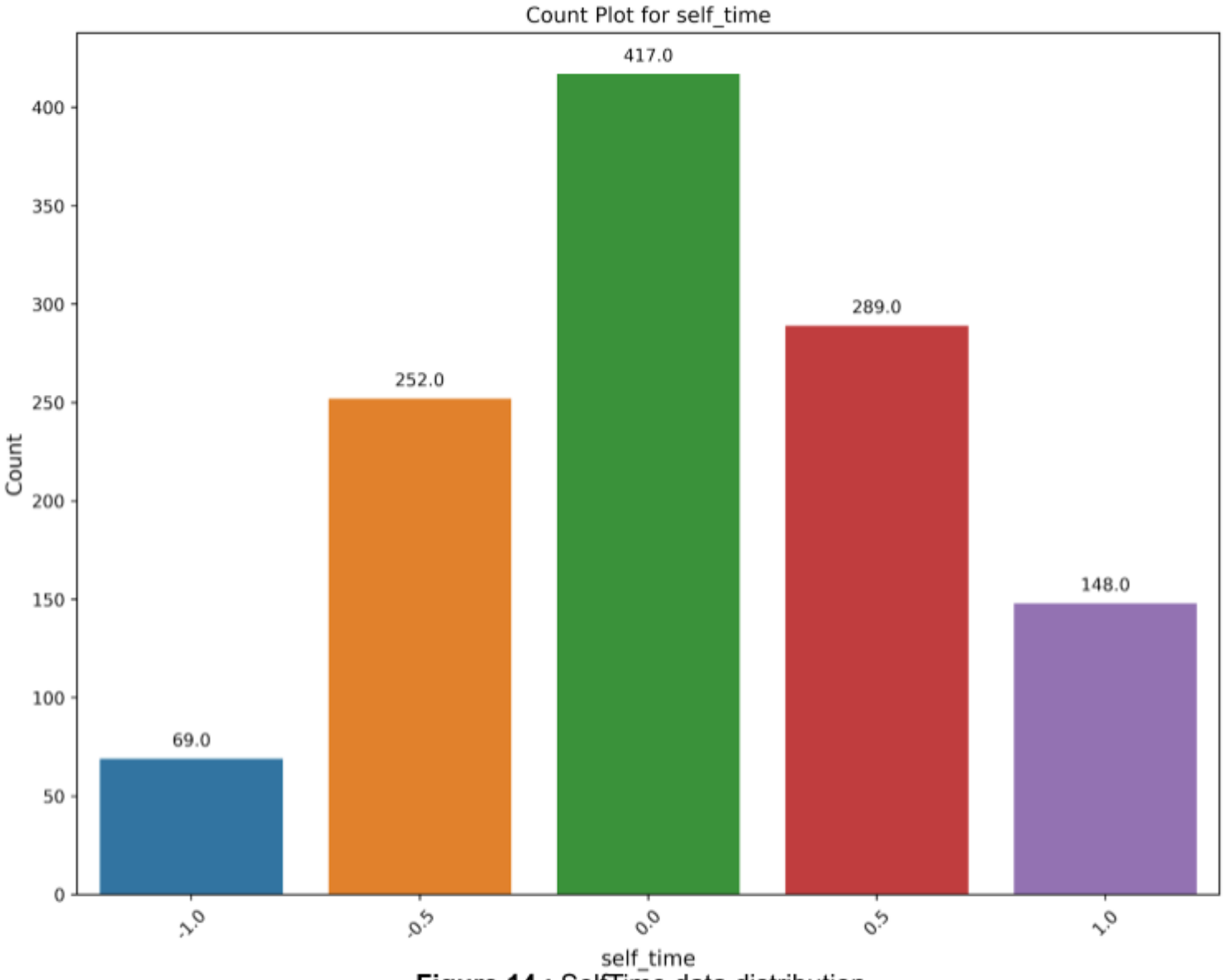

Figure 14 : SelfTime data distribution

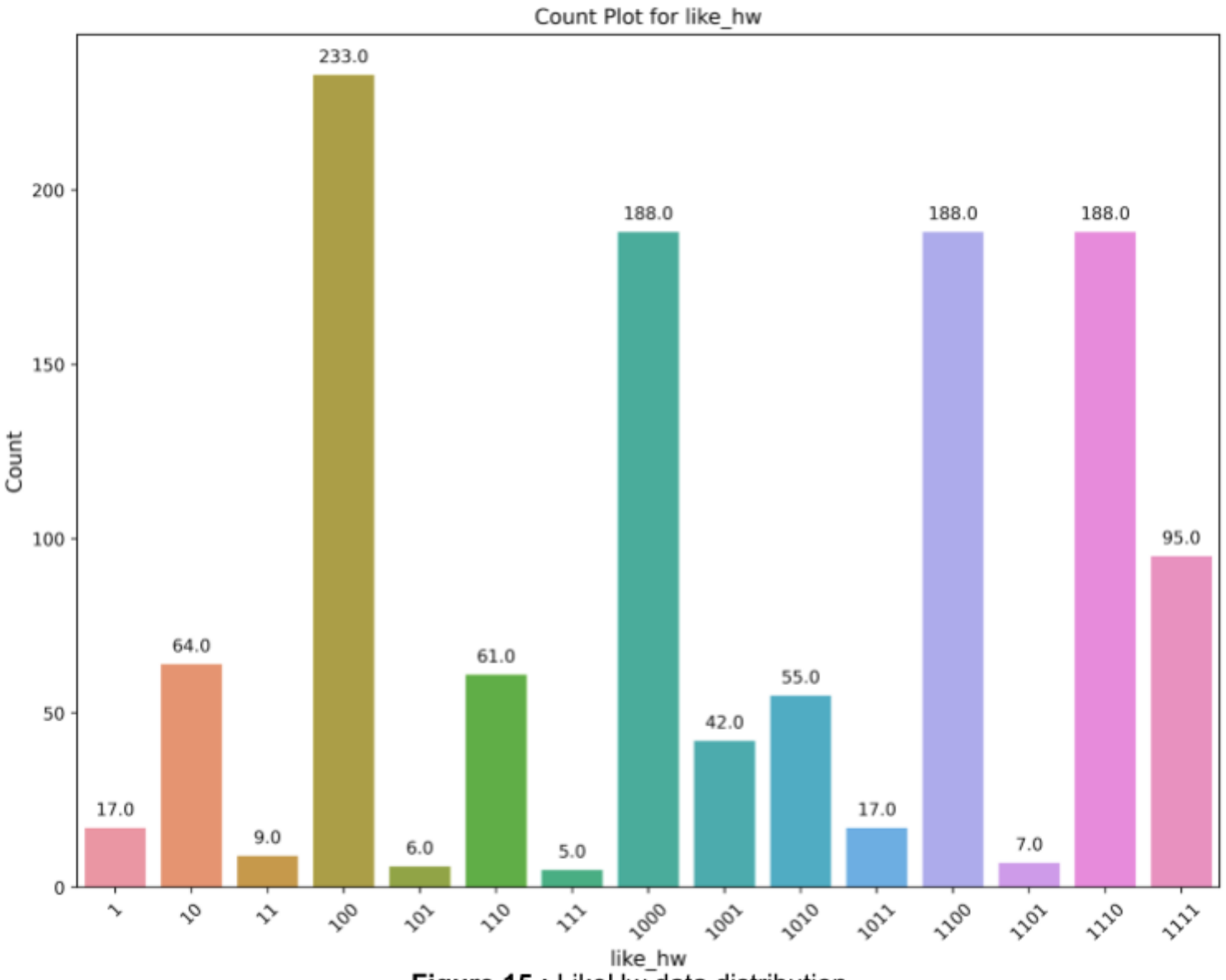

Figure 15 : LikeHw data distribution

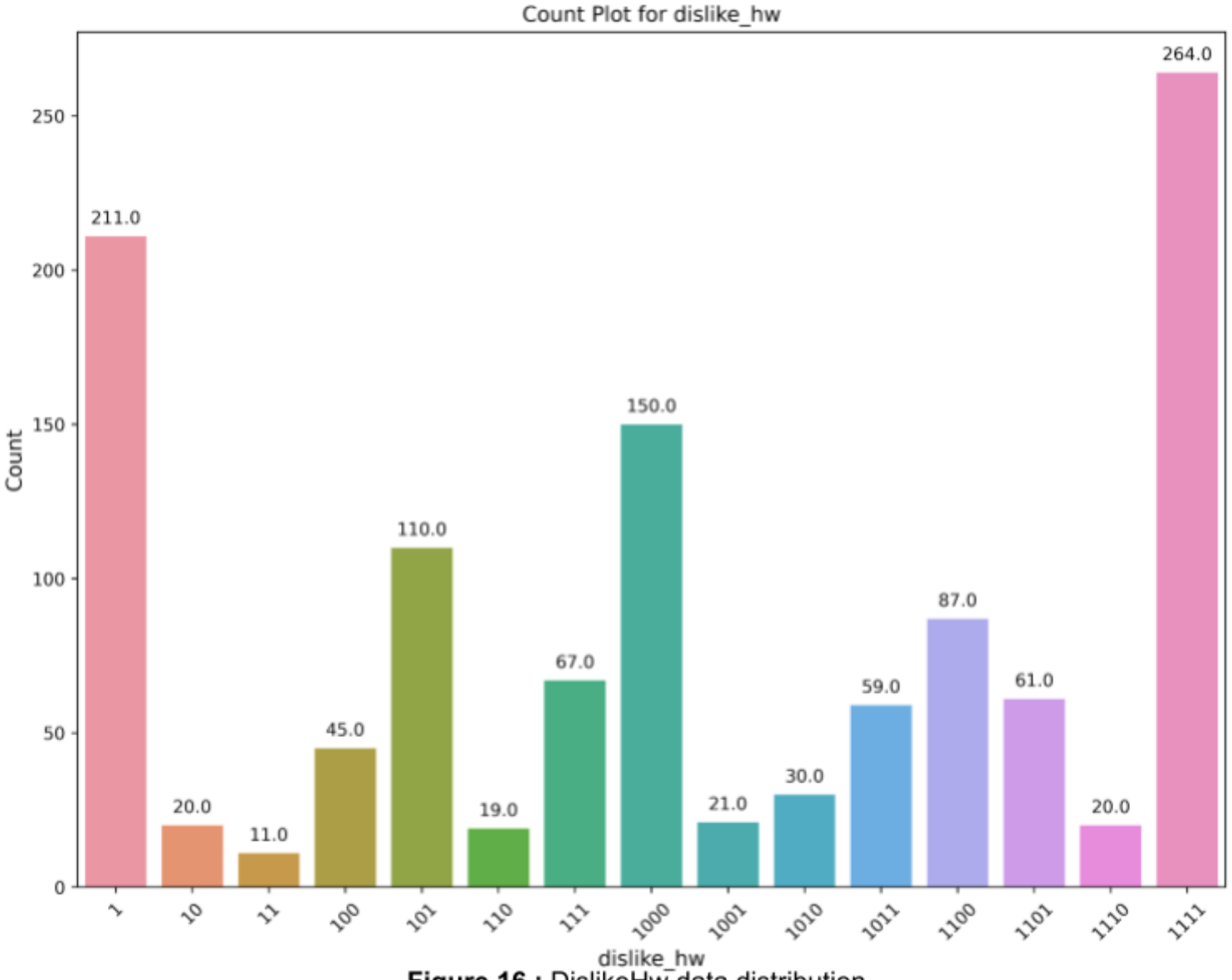

Figure 16 : DislikeHw data distribution

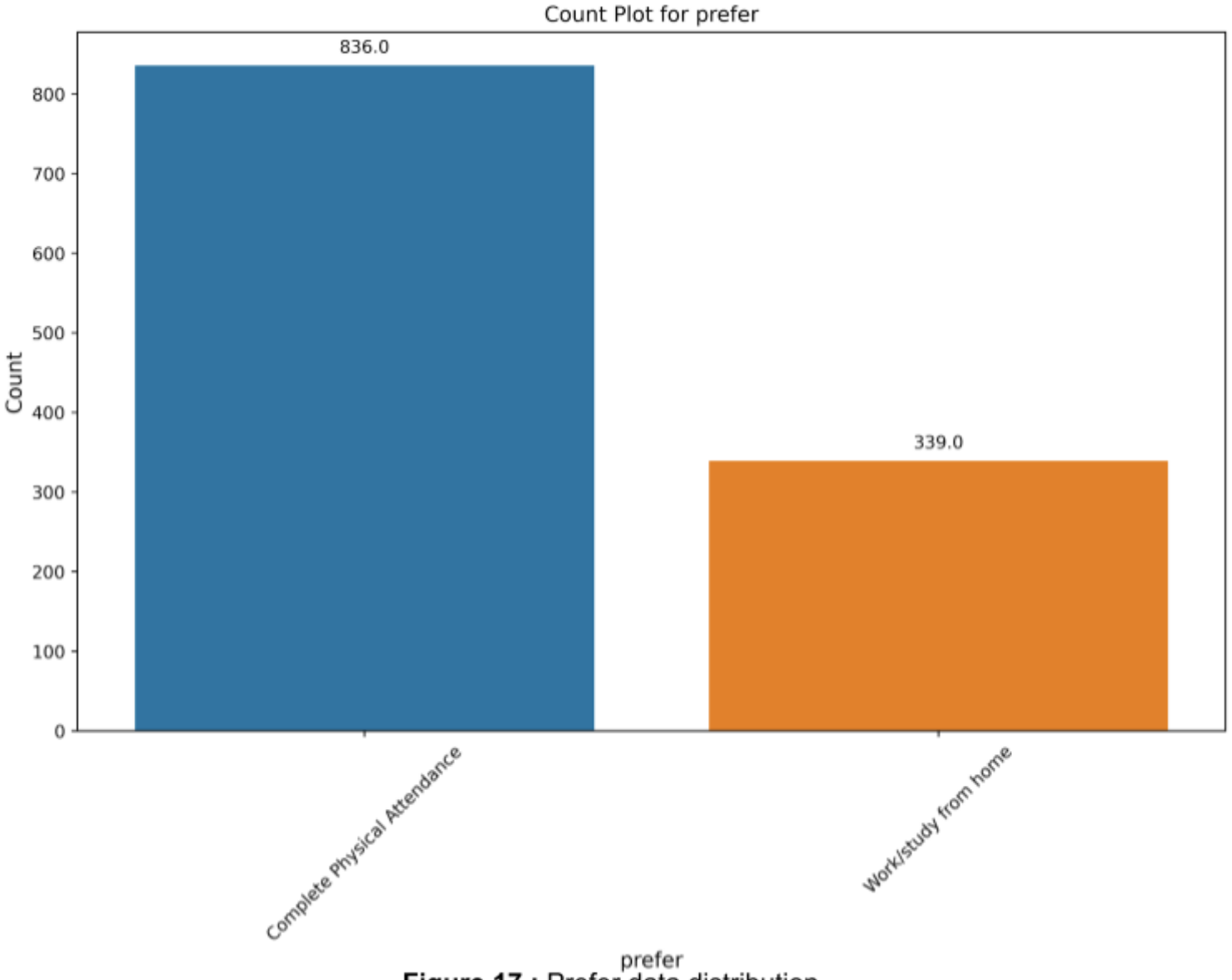

Figure 17 : Prefer data distribution

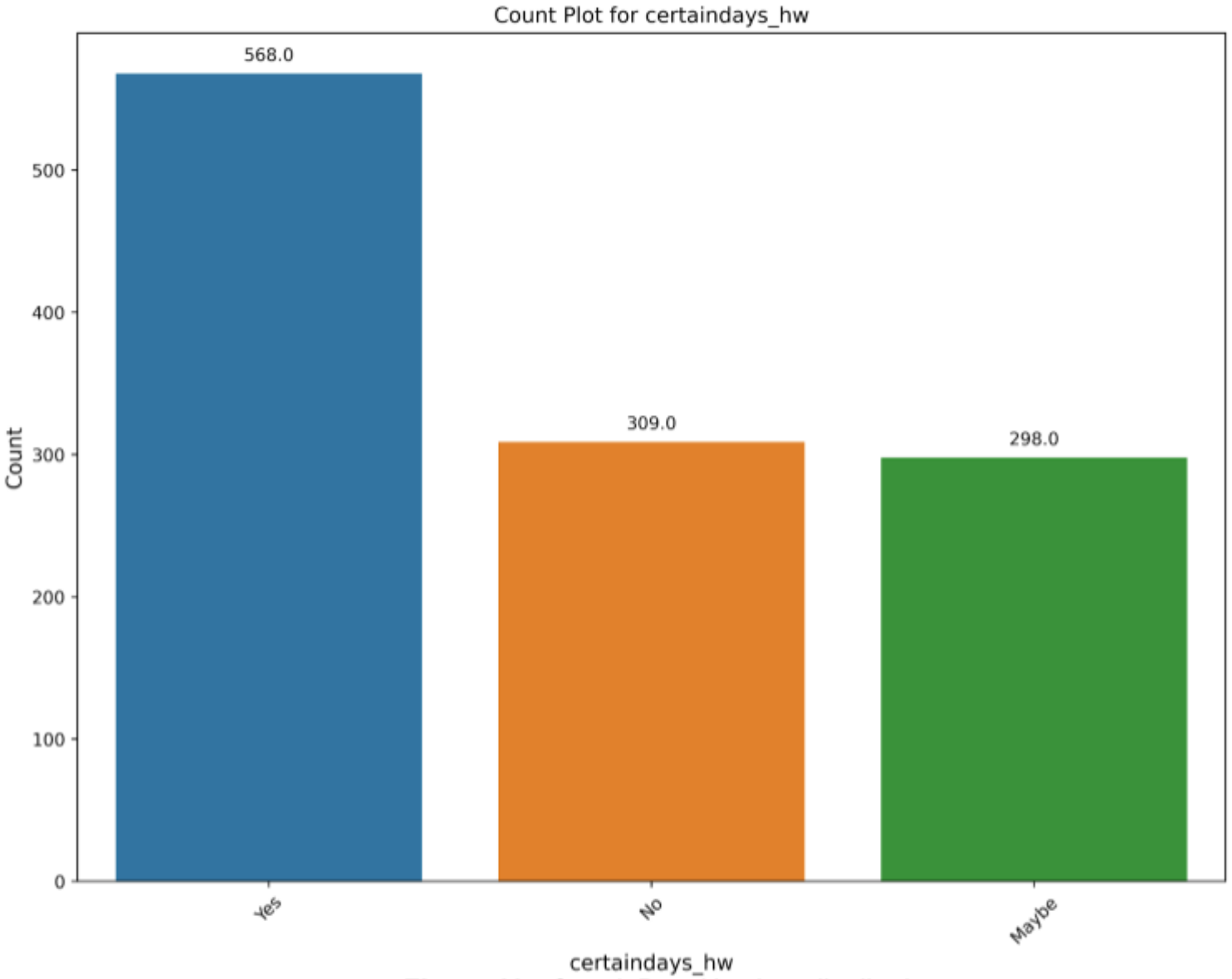

Figure 18 : CertainDaysHw data distribution

Data Visualisations

Visualisation 1:

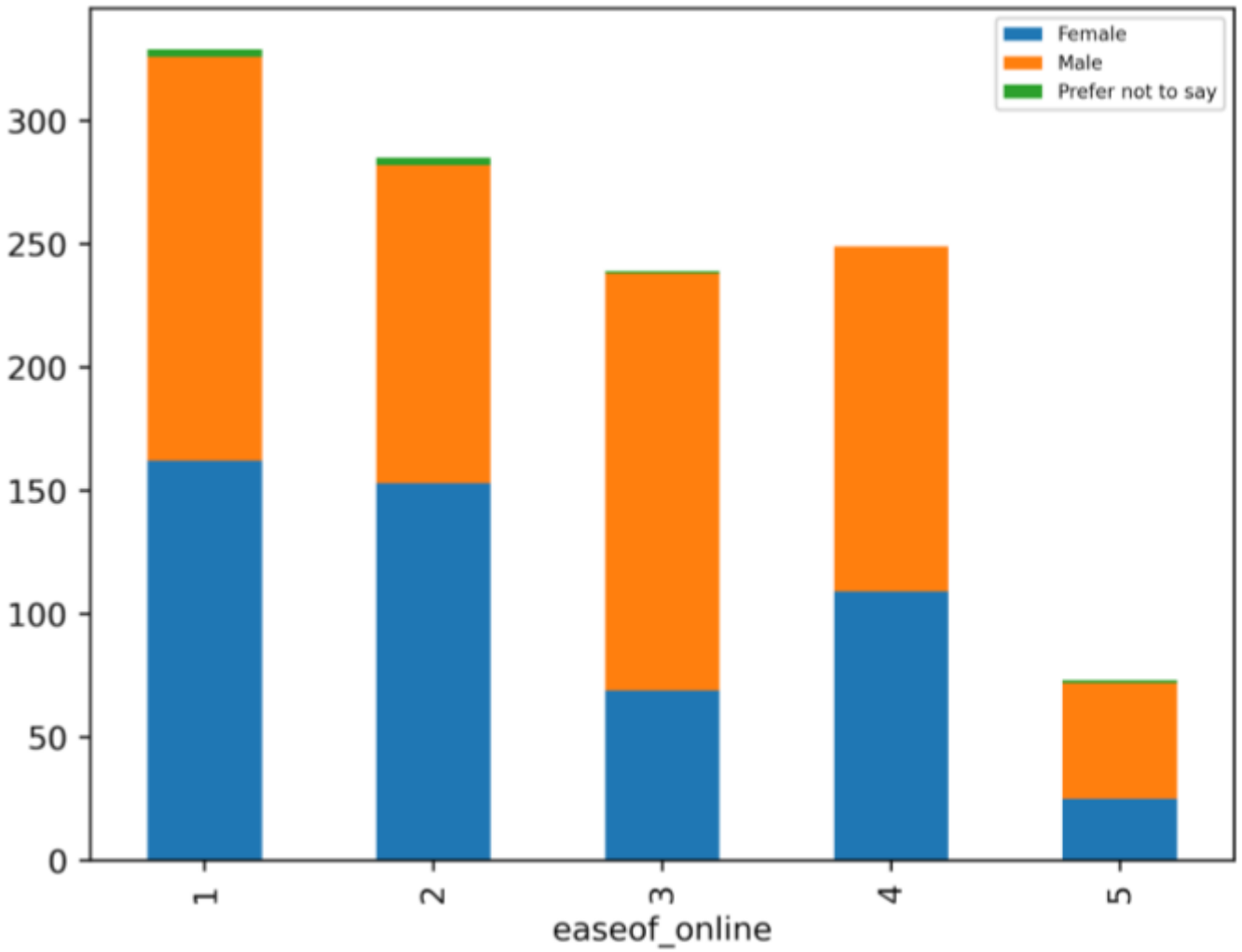

Figure 19: Gender and EaseOf_Online comparison

Figure 19 reveals the relationship between gender and ease of online work:

- The 'easeof_online' variable spans a scale of 1 to 5, denoting varying levels of ease with online work, where 1 signifies the lowest and 5 the highest ease.
- Males exhibit dominance in ratings 3 to 5, encompassing neutral, easy, and very easy perceptions, collectively constituting 63.6% of the entire population within those ratings. This suggests that, on the whole, males found the transition to online work relatively easier than females.
- Conversely, females account for 52% of the population in ratings 1 and 2 which translates to hard and very hard, indicating a higher prevalence of perceiving online work as challenging. Notably, this is observed even though the male population within the dataset is 6% larger. The findings imply that females generally found working online more challenging than their males.
- In conclusion the majority of individuals in this dataset found working online hard as 65% of the population voted for 1 and 2 and the other 35% 4 and 5(not using rating 3 in this comparison as it means neutral). But out of the minority of individuals who found it easy it is dominated by males.

Visualisation 2:

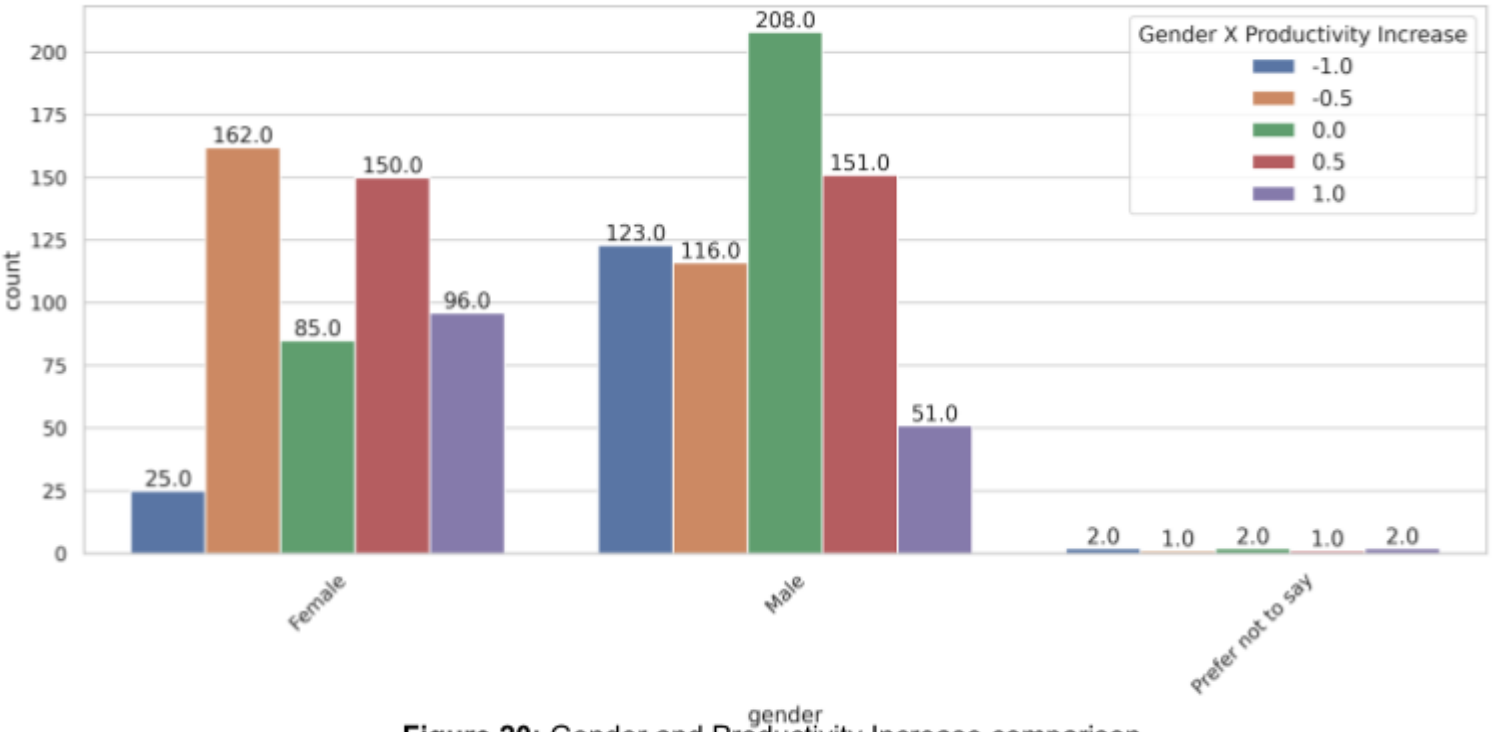

Figure 20: Gender and Productivity Increase comparison

Figure 20 delves into the interplay between gender and perceived productivity increase while working from home:

- The 'prod_inc' variable serves as a metric capturing individuals' perspectives on the impact of remote work on productivity, ranging from -1 (strongly disagree) to 1 (strongly agree), with 0 indicating a neutral standpoint.
- Notably, a significant portion of males, constituting the majority, expressed a belief that productivity remained unchanged (voted for 0), suggesting a perception that remote work did not alter productivity compared to office settings. However, upon closer examination, focusing on categories -1, -0.5, 0.5, and 1 while excluding the neutral stance (0), a nuanced picture emerges. A substantial 54% of males leaned towards the negative categories, implying a prevailing sentiment that productivity decreased while working remotely.
- Conversely, the majority of females favoured the -0.5 category, indicating a belief that productivity decreased when working from home compared to the office. The distribution between categories -0.5 and 0.5 is closely matched, with 162 and 150 votes, respectively. To gain a comprehensive view, the analysis compares categories -1, -0.5, 0.5, and 1, revealing that 57% of females believe that productivity increases with remote work.
- In summary, the data suggests a contrast in opinions between genders. The majority of females lean towards the perspective that productivity increases with remote work, while males, in contrast, exhibit a tendency to disagree with this notion. These nuanced insights contribute to a richer understanding of how different genders perceive the impact of remote work on productivity.

Visualisation 3:

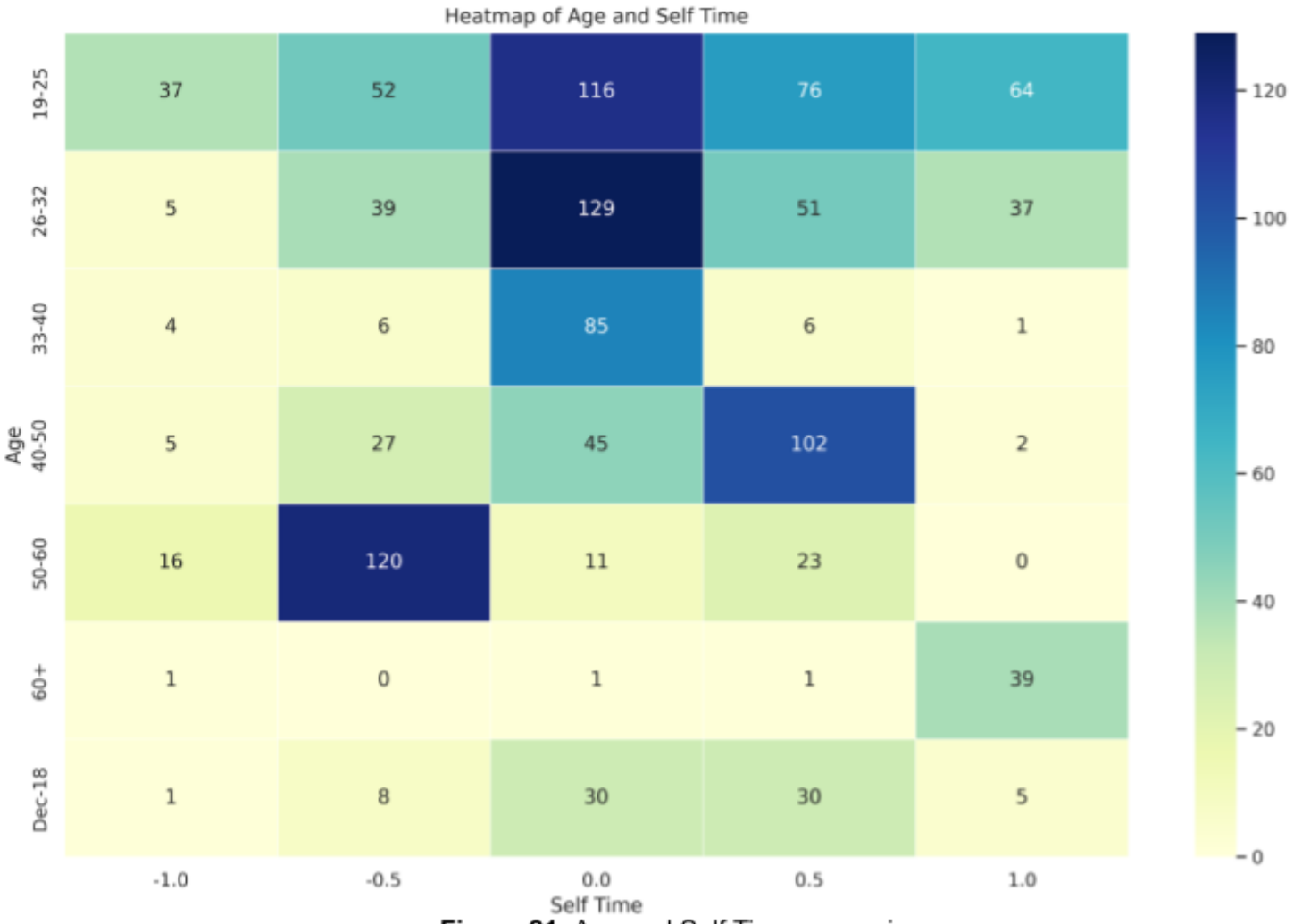

Figure 21: Age and Self Time comparison

Figure 21 provides insights into the relationship between age groups and self time (rating of how much self time was procured):

- The most common category for individuals in the age groups 19-25 and 26-32 is the 0 category which means that they believe that there was no change in the self time while working from home. But taking into consideration only the negative (-1 and -0.5) and positive(0.5 and 1) categories the majority of people 61% and 67% respectively in those age groups tend more to the increased self time than decreased while working from home.
- Age group 33-40 strongly believe that there is no change in self time as 83% voted for the 0 category.
- Age group 40-50 are the only group whose most dominant category was 0.5 category advocating that self time increased slightly by working from home with a total of 56% of the individuals in that group voting for that.

- The 50-60 strongly believed that self time slightly decreased by working from home with a total of 71% of them voting for category -0.5 , making that the only age group that believes that self time decreased instead of increased by working from home.
- The 60+ strongly believe that self time increased by a lot by working from home as 93% of all the individuals in that group voted for category 1.
- Age group Dec-18 had the same votes for category 0 and 0.5 indicating their inclination to the increase of self time.
- Overall except from the 50-60 age group all other groups believe that they had more self time while working remotely.

Visualisation 4:

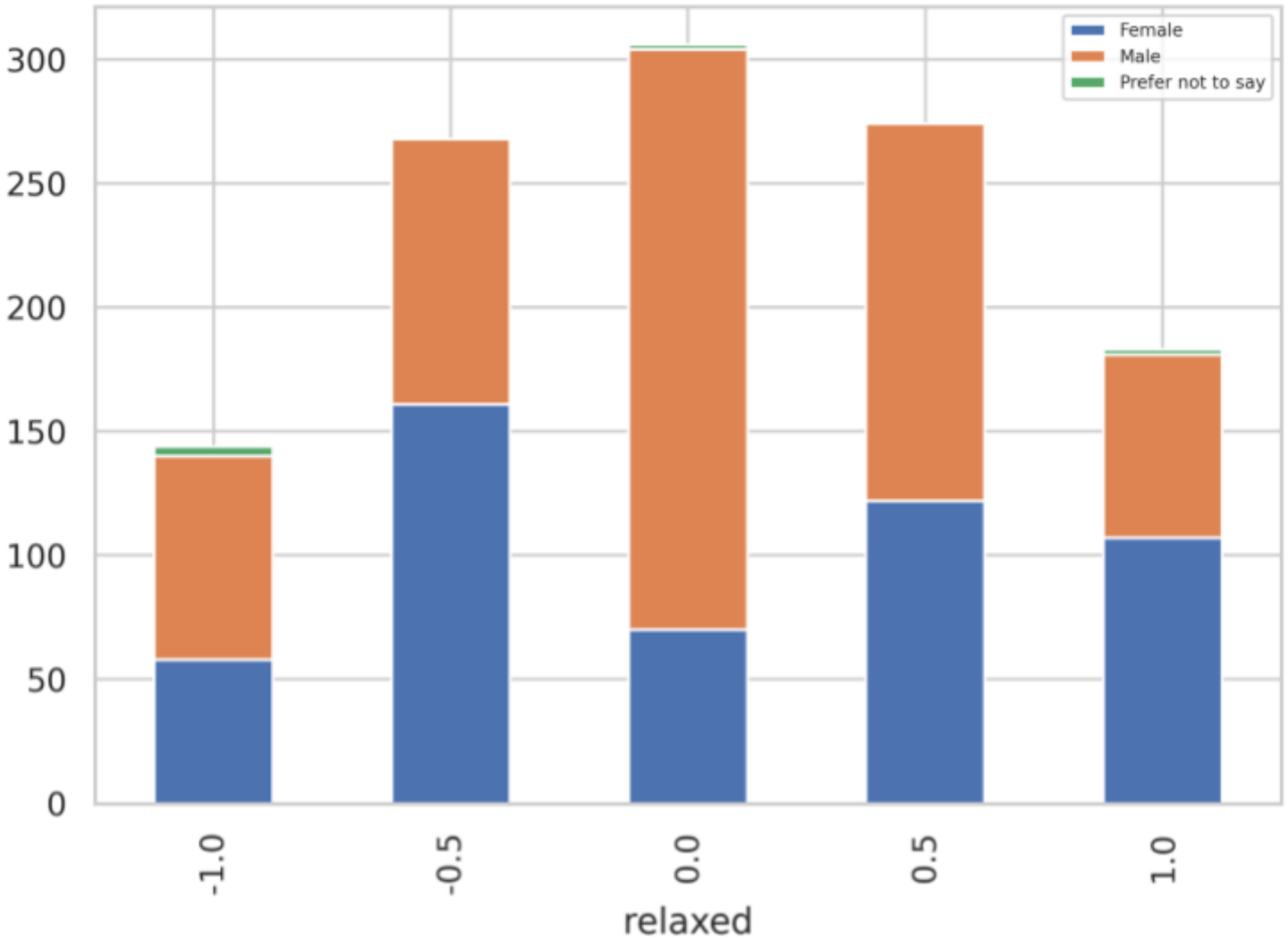

Figure 22: Gender and Relaxed comparison

Figure 22 provides insights into the perceived levels of relaxation among different genders while working from home. The ratings for relaxation range from -1, indicating not relaxed, to 1, signifying very relaxed:

- The prevalent choice among males was the neutral category (0), selected by 36% of the respondents. However, to discern the predominant direction of their sentiment, a comparison between negative and positive categories was conducted. The results revealed that 54% of males leaned towards feeling more relaxed while working from home.
- The predominant choice among females was the -0.5 category, selected by 31% of the population. This indicates that the majority of women experienced increased stress while working remotely.
- Once again, a disparity in relaxation levels between the two genders is apparent. Males tended to find working from home more relaxing, while females perceived it as more stressful. This observation contrasts with the information gleaned from Figure 20, where females expressed a perception of increased productivity while working from home, whereas males indicated the opposite trend.

Visualisation 5:

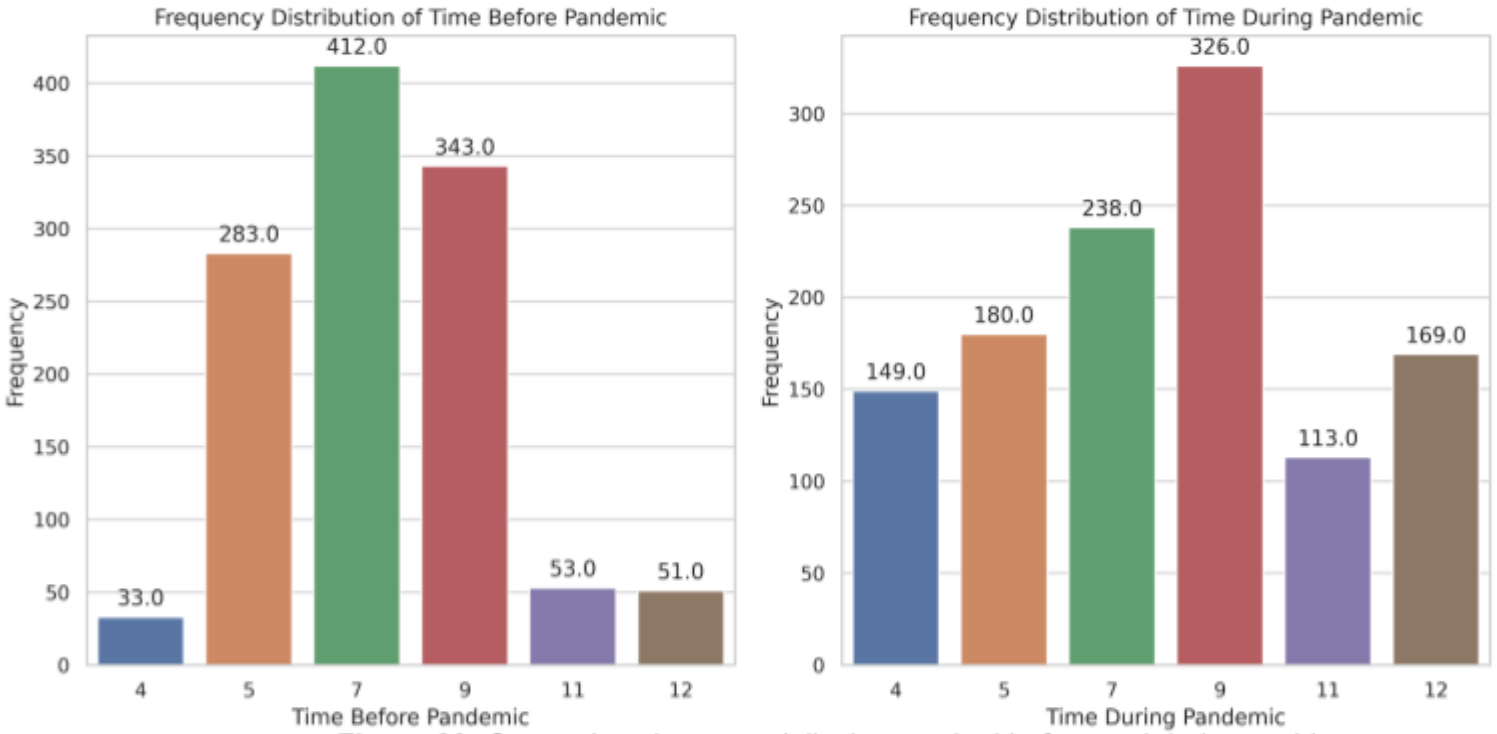

Figure 23: Comparison between daily time worked before and during covid

Figure 23 shows the comparison between the daily time worked for people before and during covid:

- The average daily work time has increased from 7.42 hours before COVID-19 to 7.97 hours during COVID-19. This suggests a general trend of longer daily work hours during the pandemic.
- There has been a substantial surge in the number of individuals working 11 hours, with a notable increase of 113%. Similarly, for those working 12 hours, the surge is even more significant, reaching a remarkable increase of 231%.
- A noteworthy observation is the substantial change in the total hours worked by individuals before and during the COVID-19 pandemic. The dataset indicates that the cumulative hours worked before the pandemic amounted to 8713, and during the pandemic, this figure experienced a considerable increase to 9367. This represents a percentage increase of 7.5%, signifying an additional 654 hours worked per day across the entire population.

Part 2: Data Preparation

Data preparation involves the process of cleaning, transforming, and organising the data to ensure its suitability for use in building machine learning models. The goal is to enhance the quality of the data and optimise its format, making it conducive for accurate and effective model training.

Missing values handling:

Initially, the dataset was examined for missing values, revealing that the 'line_of_work' attribute was the only one with such values. Figure 24 illustrates that this attribute, indicating the individual's line of work, had a substantial 59% of its entries missing, totaling 696 instances. Given this high percentage, discarding the missing values was not a viable option, as it would involve eliminating over half of the dataset. To address this, a closer examination of the 'line_of_work' attribute was undertaken. Figure 25 demonstrates a breakdown of the missing values across different age groups. Notably, 62% of the missing records fall into categories such as Dec-18 (indicating children), 19-25 (representing individuals likely to be in university), and 60+ (reflecting retirement age). Considering these age groups are typically associated with non-working individuals, it was assumed that the remaining 38% of the missing values are also likely unemployed. As a result, all missing records in the 'line_of_work' attribute were assigned the 'other' category, representing the assumption that this category encompasses unemployed individuals.

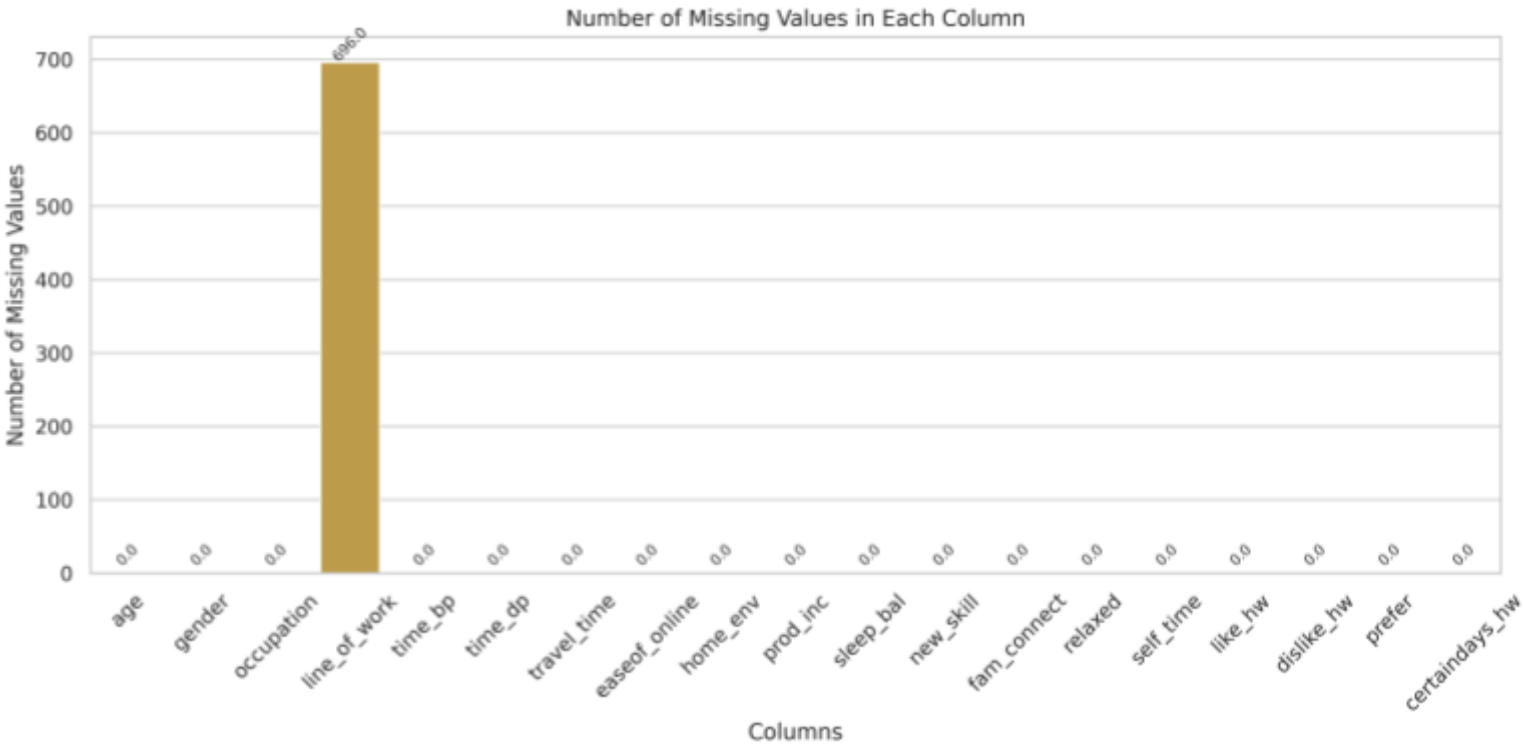

Figure 24: Missing values

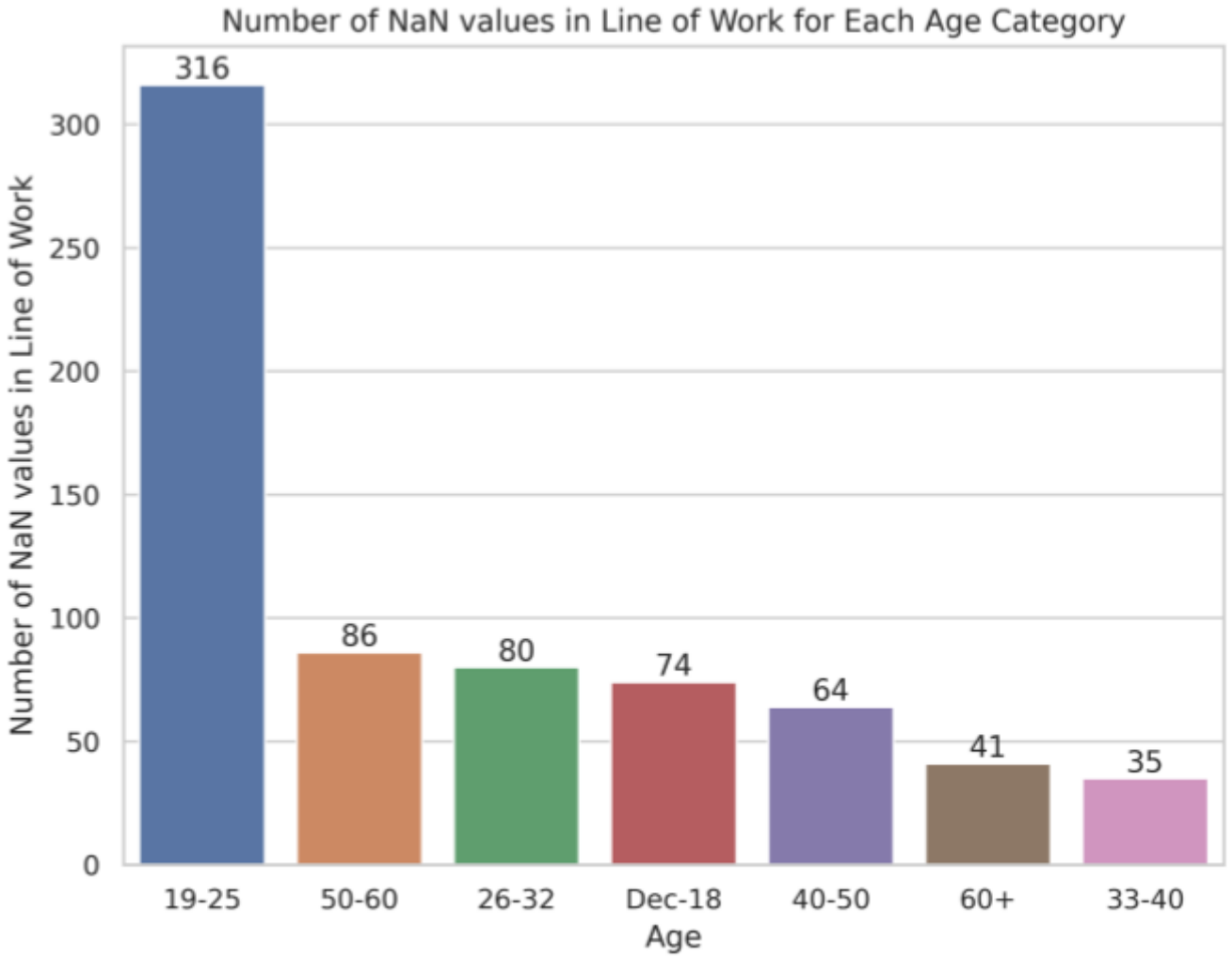

Figure 25: Age group of Nan values in line of work

Outlier Detection:

The next step in data preparation was outlier handling, outlier handling involves identifying and addressing data points that significantly deviate from the majority of the dataset, aiming to prevent them from disproportionately influencing machine learning models outcome.To identify outliers, boxplot visualisations were employed for all numeric attributes in the dataset. Upon careful examination, it was noted that none of the attributes exhibited outliers.

Data Transformation:

Given that machine learning algorithms generally perform more effectively when working with numeric data rather than categorical data, the transformation of certain attributes was an essential step in data preparation. This transformation specifically impacted the attributes of age, gender, occupation, line of work, like_hw, dislike_hw, prefer, and certaindays_hw. This conversion from categorical to numeric representation facilitates the algorithms' ability to derive meaningful patterns and insights from the dataset.

Normalisation:

Normalisation is the process of scaling numerical attributes to a standard range, to ensure consistency and prevent features with larger scales from dominating in machine learning models, thereby enhancing the model's performance and convergence during training. To achieve normalisation, the Min-Max Scaler(1) was employed on all attributes after they were transformed to numeric except from the target attribute of the classification which is the gender, because distance based algorithms will be used where normalising the target isn't ideal.The Min-Max Scaler was chosen because it scales the data to a specific range (usually 0 to 1), preserving the relationships between data points. This ensures that the transformed features maintain their relative proportions while being standardised, which then leads to better performing models.

Dimensionality reduction:

Dimensionality reduction is fundamental in machine learning as it seeks to simplify intricate datasets by decreasing the number of features while preserving vital information, thereby enhancing the efficiency of models. The reduction process involved assessing the Pearson correlation of all attributes with the target of the classification, guiding the selection of attributes for removal. As depicted in Figure 26, the attributes with the highest positive positive or negative correlation were retained. These attributes include occupation, time_bp, like_hw, prod_inc, age, time_dp, prefer, and line_of_work.

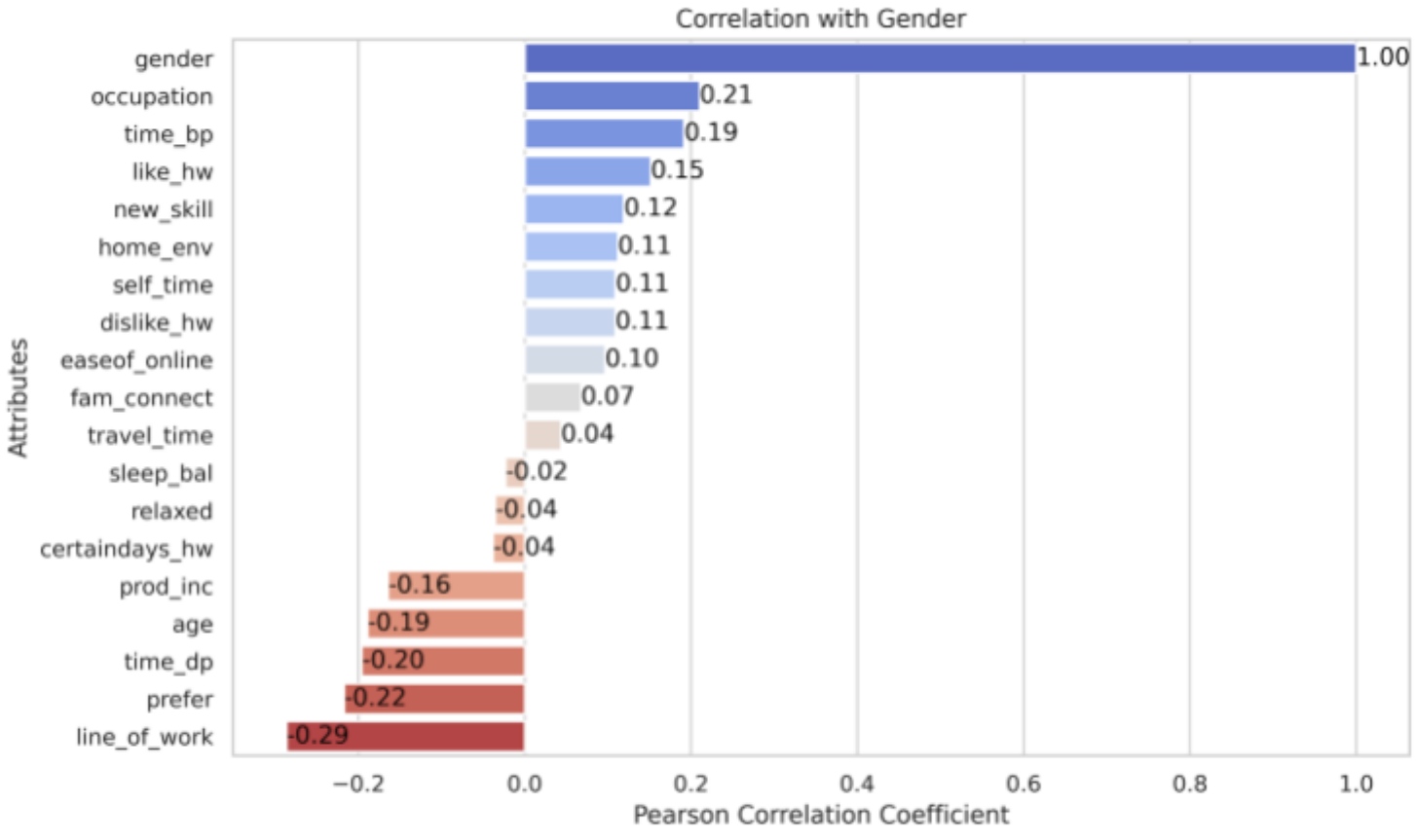

Figure 26: Gender Correlation with all attributes

Part 3: Classification

To facilitate binary classification in building the classification models, the target attribute underwent modification. Given the limited instances of the third category within the gender attribute (Figure 4), it was necessary to drop those instances, allowing for a more balanced binary classification approach. Additionally, given the dataset's modest size, all models were trained using cross-validation. This approach optimally utilises the limited data by employing it for both training and testing across multiple folds, resulting in a more dependable assessment of the model's performance.
Figure 27 shows the comparison of the models trained using different performance metrics where higher value shows higher performance.

Algorithms	Precision	Recall	F1 Score	Accuracy Score	AUC Score
Random Forest	0.9083	0.9066	0.906	0.9066	0.901
MLP	0.8575	0.856	0.855	0.856	0.8491
ADA Boost	0.8387	0.8389	0.8385	0.8389	0.8347
Logistic Regression	0.7117	0.7129	0.7109	0.7129	0.7043

Figure 27: Model Comparison Using Different Metrics

MultiLayer Perceptron:

The Multilayer Perceptron (MLP) model exhibited strong performance with an overall accuracy of 86%.For the classification of males (coded as 1), the model achieved a precision of 84% and a recall of 91%. This suggests that the model correctly identified 84% of instances where the actual gender was male, and among all actual male instances, it successfully captured 91%. On the other hand, for the classification of females (coded as 0), the model demonstrated a precision of 88%, indicating that it correctly classified 88% of instances as female. The recall for females was 79%, signifying that the model effectively captured 79% of all actual female instances. The F1-score, which takes into account the class imbalance, was 86%, indicating a balanced performance across both gender classes. The AUC/ROC curve, visually depicted in Figure 28, further supports the model's discriminative ability. With an AUC score of 0.85, the model's curve exhibits a strong ability to distinguish between male and female instances. The closer the AUC score is to 1, the better the model's overall performance, and in this case, the MLP model demonstrates a commendable capability in gender classification tasks.Furthermore, the confusion matrix, available in Figure 29, provides a detailed breakdown of the model's classification results, offering insights into specific instances of correct and incorrect predictions for each gender class mentioned above.

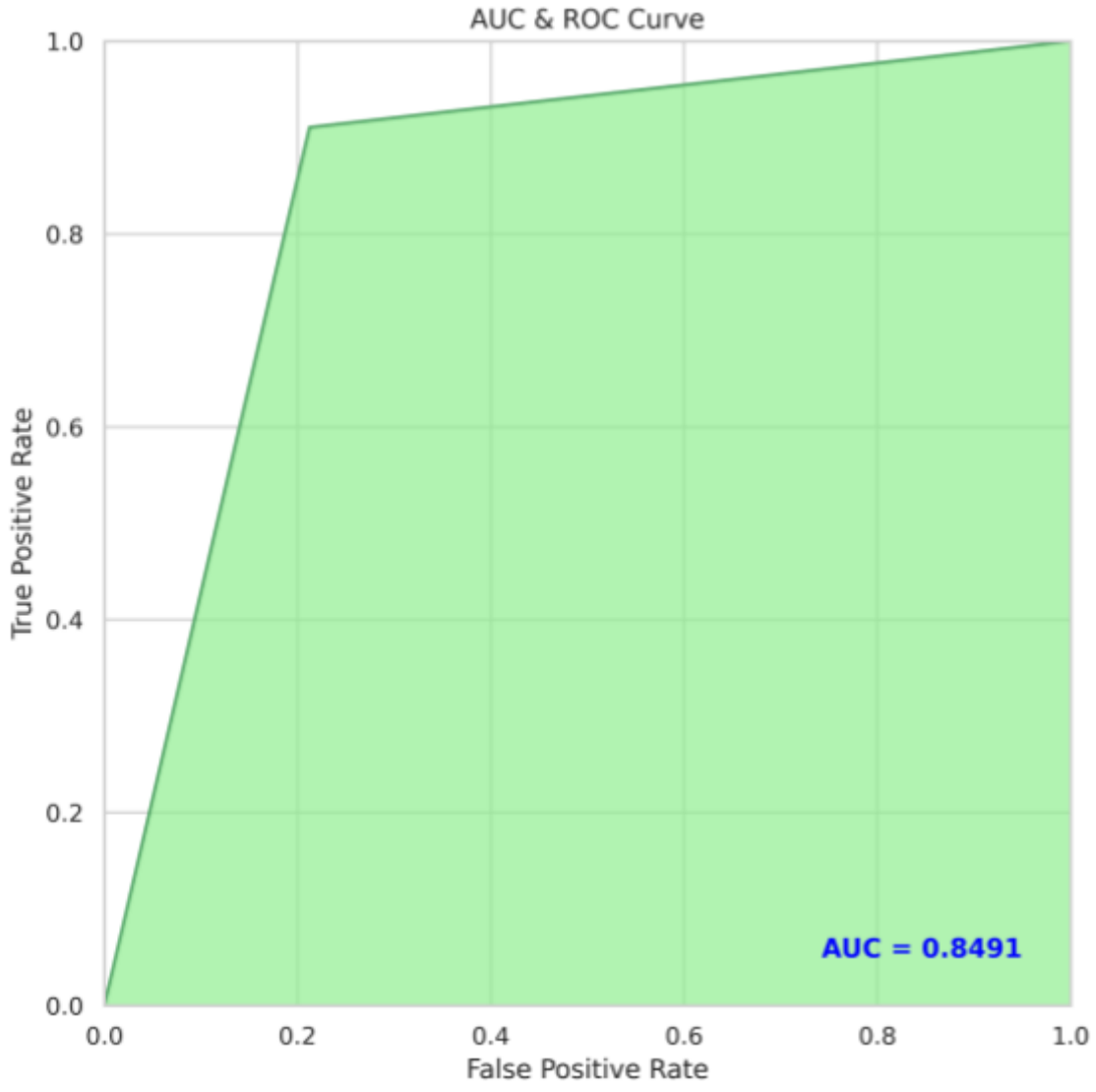

Figure 28: MLP AUC/ROC Curve

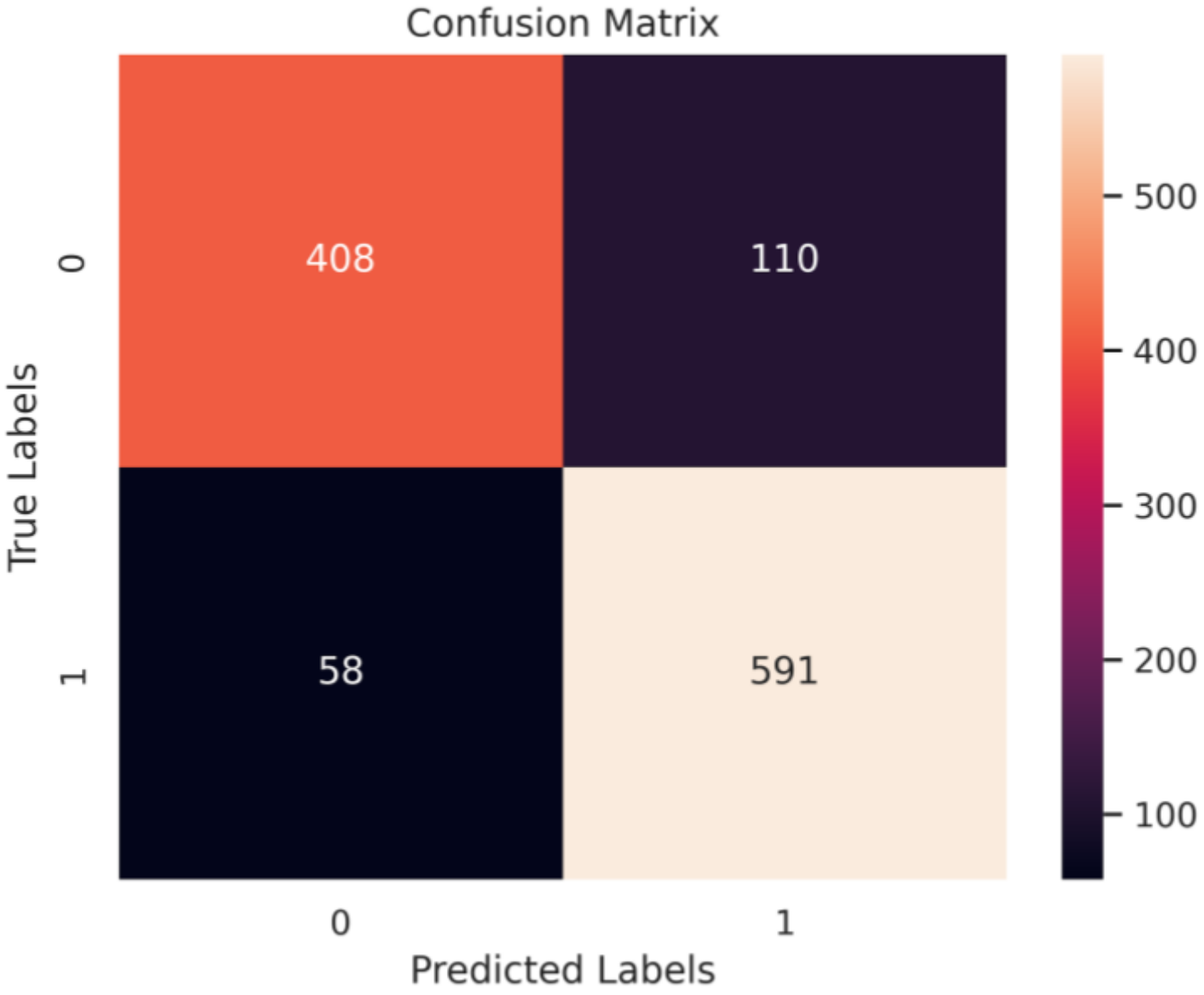

Figure 29: MLP Confusion Matrix

Logistic Regression:

The Logistic Regression model exhibited relatively poorer performance compared to other models, with an overall accuracy of 71%. In predicting males (coded as 1), the model achieved a precision of 72% and a recall of 78%, indicating it correctly identified 72% of instances where the actual gender was male, and among all actual male instances, it captured 78%. However, for the classification of females (coded as 0), the model showed a precision of 70%, correctly classifying 70% of instances as female, with a recall of 63%, signifying that it captured only 63% of all actual female instances. The F1-score, accounting for class imbalance, was 71%, suggesting a somewhat balanced performance across both gender classes. The AUC/ROC curve, depicted in Figure 30, demonstrated discriminative capabilities with an AUC score of 0.7. While an AUC score closer to 1 indicates better overall performance, the Logistic Regression model's score of 0.7 suggests suboptimal discriminative ability. Figure 31 presents the confusion matrix, offering insights into specific instances of correct and incorrect predictions for each gender class.

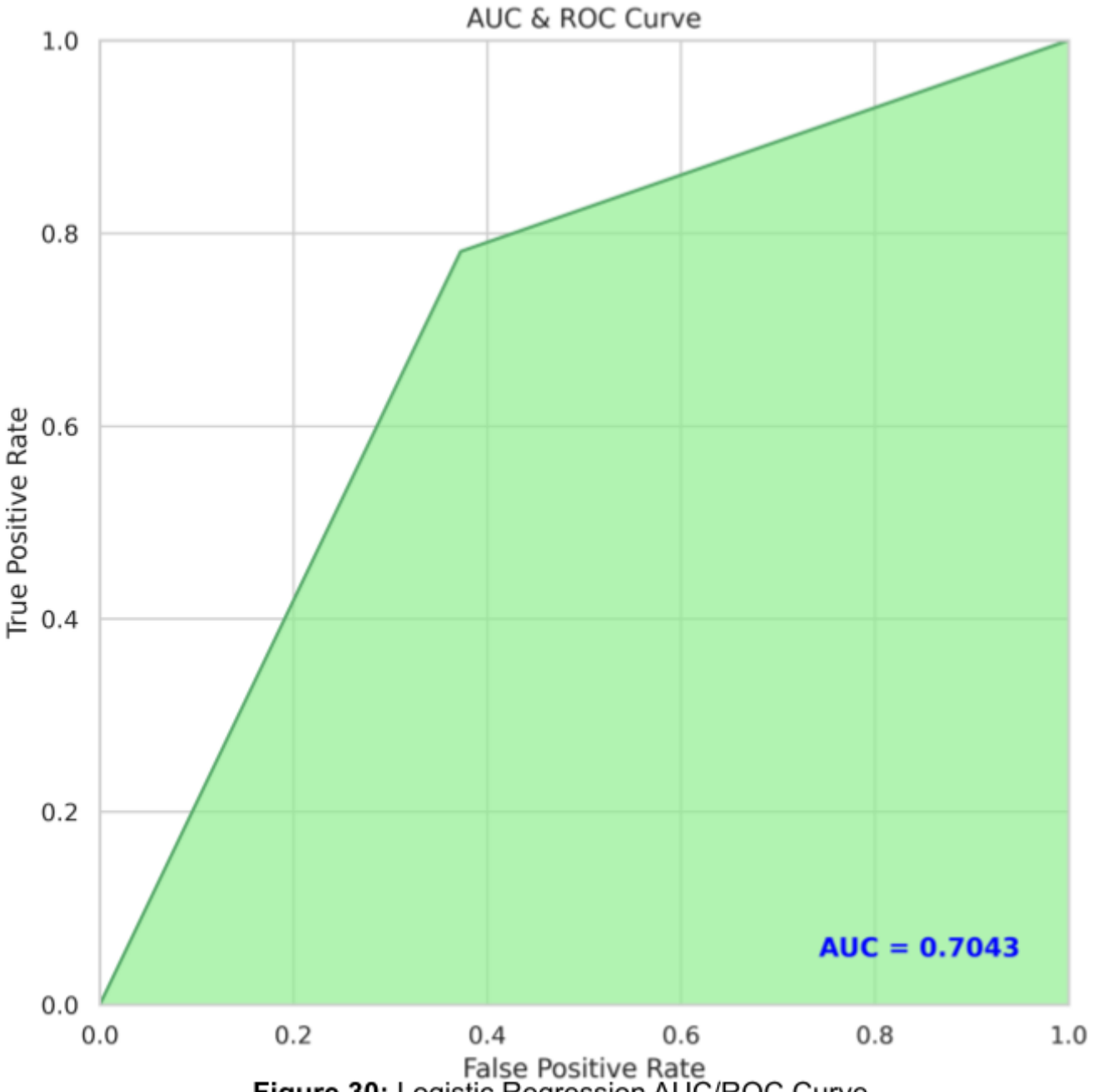

Figure 30: Logistic Regression AUC/ROC Curve

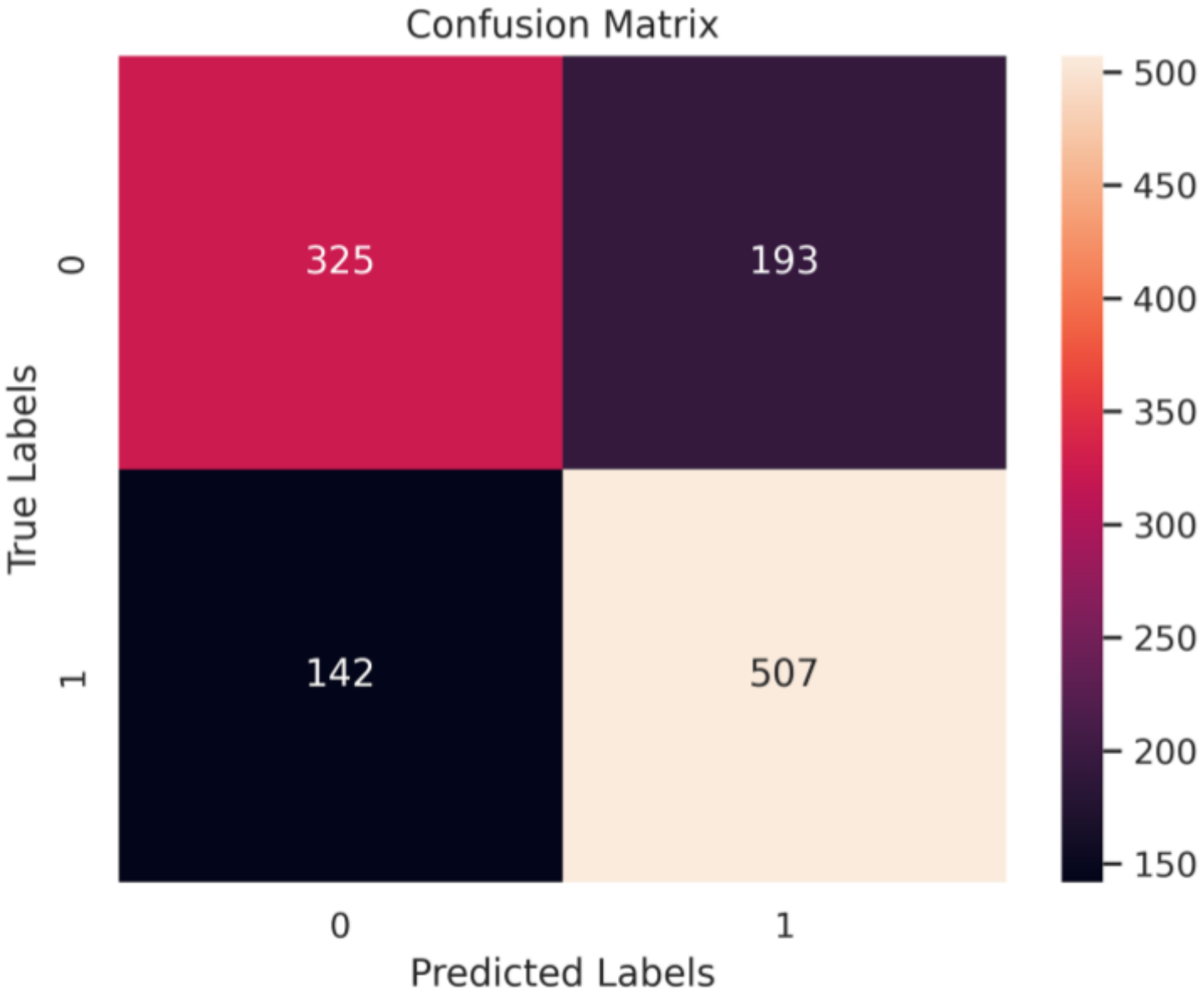

Figure 31: Logistic Regression Confusion Matrix

Random Forest:

The Random Forest model demonstrated outstanding performance, emerging as the top-performing model among all the classifiers. With an impressive overall accuracy of 91%, the model showcased robust capabilities in gender classification tasks. For the prediction of males (coded as 1), the Random Forest model achieved a precision of 89% and a recall of 95%, indicating that it accurately identified 89% of instances where the actual gender was male, and among all actual male instances, it successfully captured 95%. On the flip side, for the classification of females (coded as 0), the model displayed a precision of 93%, correctly classifying 93% of instances as female, with a recall of 85%, signifying that it captured 85% of all actual female instances.The F1-score, accounting for class imbalance, was an exceptional 91%, indicating a harmonious performance across both gender classes. The AUC/ROC curve, visually depicted in Figure 32, further underscored the model's discriminatory prowess, boasting an AUC score of 0.9. The closer the AUC score is to 1, the better the model's overall performance, and in this case, the Random Forest model demonstrated an outstanding capability in gender classification tasks. Figure 33 provides a detailed confusion matrix, offering insights into specific instances of correct and incorrect predictions for each gender class mentioned above. Overall, the Random Forest model's superior accuracy, precision, recall, F1-score, and AUC score collectively position it as the most effective classifier for this gender classification task.

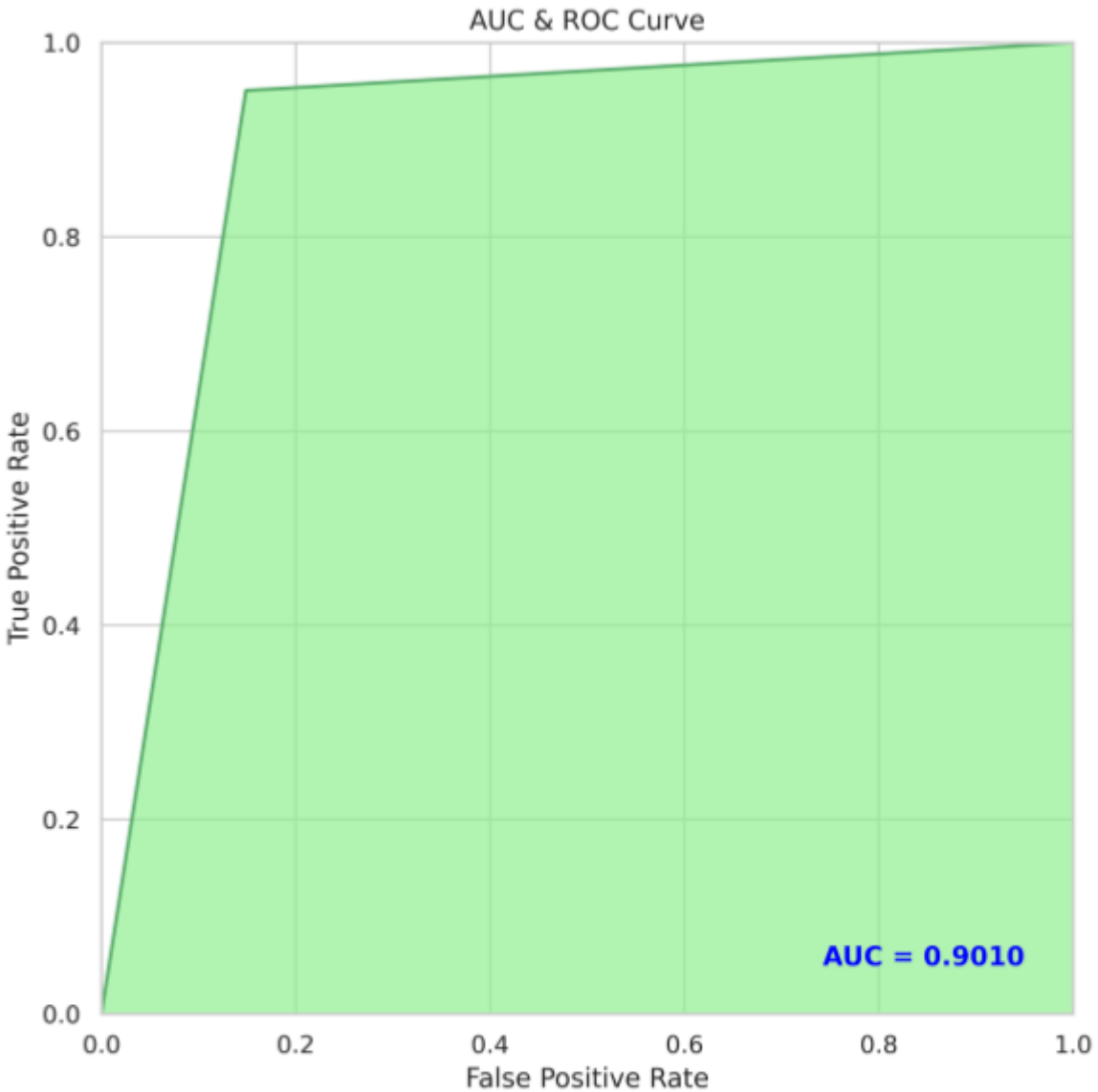

Figure 32: Random Forest AUC/ROC Curve

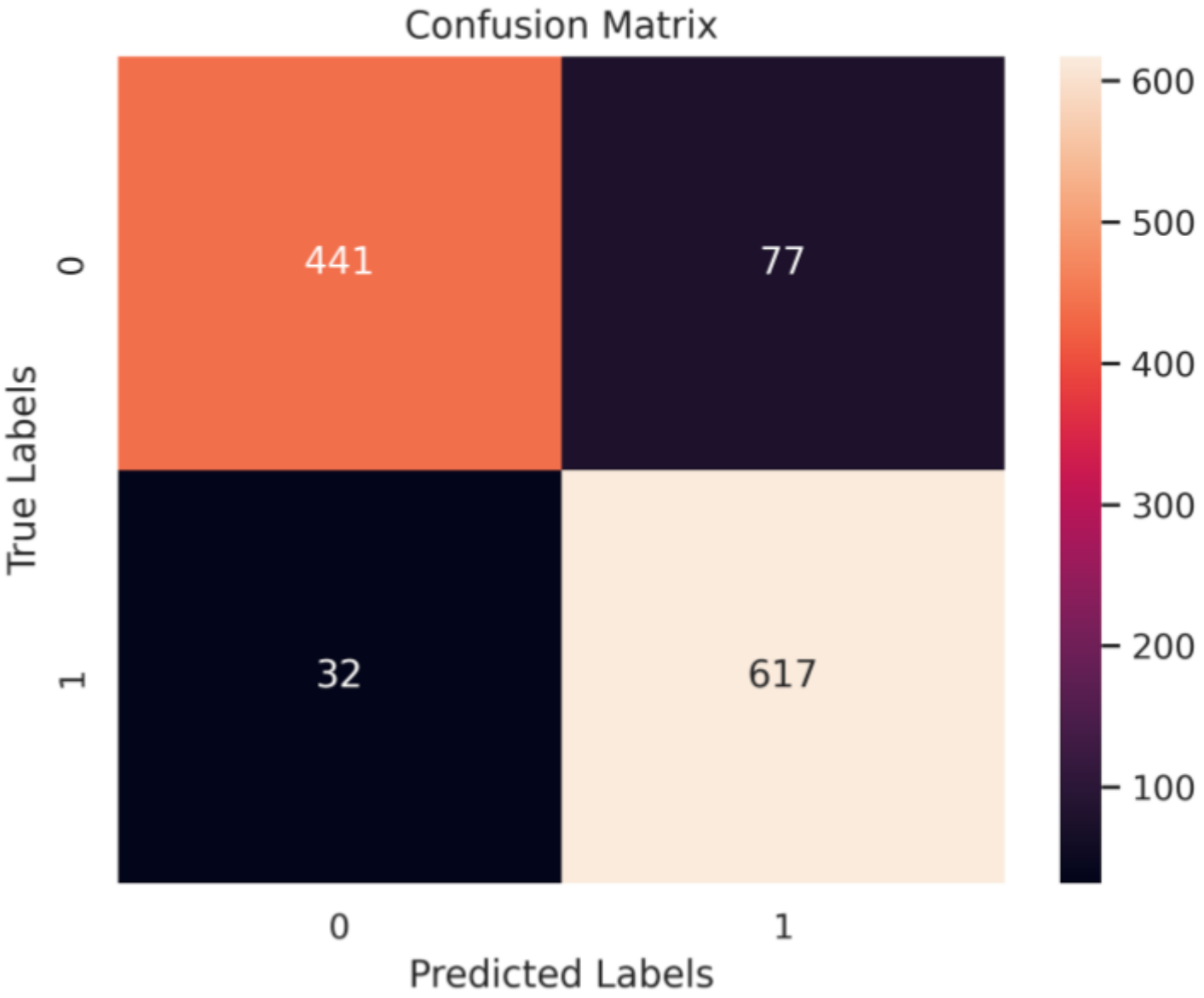

Figure 33: Random Forest Confusion Matrix

ADA Boost(3):

The AdaBoost model demonstrated a decent performance in gender classification, achieving an overall accuracy of 84%. The AdaBoost model demonstrated a decent performance in gender classification, achieving an overall accuracy of 84%. This algorithm works by combining multiple weak learners to create a strong learner, with a focus on improving the classification of instances that previous models found challenging.For the identification of males (coded as 1), the AdaBoost model achieved a precision of 84% and a recall of 87%, suggesting that it accurately identified 84% of instances where the actual gender was male, and among all actual male instances, it successfully captured 87%. Conversely, for the classification of females (coded as 0), the model exhibited a precision of 83%, correctly classifying 83% of instances as female, with a recall of 80%, signifying that it captured 80% of all actual female instances.The F1-score, considering class imbalance, was 84%, indicating a balanced performance across both gender classes. The AUC/ROC curve, visually depicted in Figure 34, reinforced the model's discriminatory capability, boasting an AUC score of 0.84. This score indicates a solid ability to distinguish between male and female instances.While not surpassing the performance of the Random Forest model, AdaBoost's accuracy, precision, recall, F1-score, and AUC score collectively demonstrate its effectiveness in gender classification tasks. The algorithm's utilisation of an ensemble approach, combining weak learners to create a strong model, contributes to its ability to handle complex relationships within the data.The confusion matrix of this classifier can be seen in figure 35.

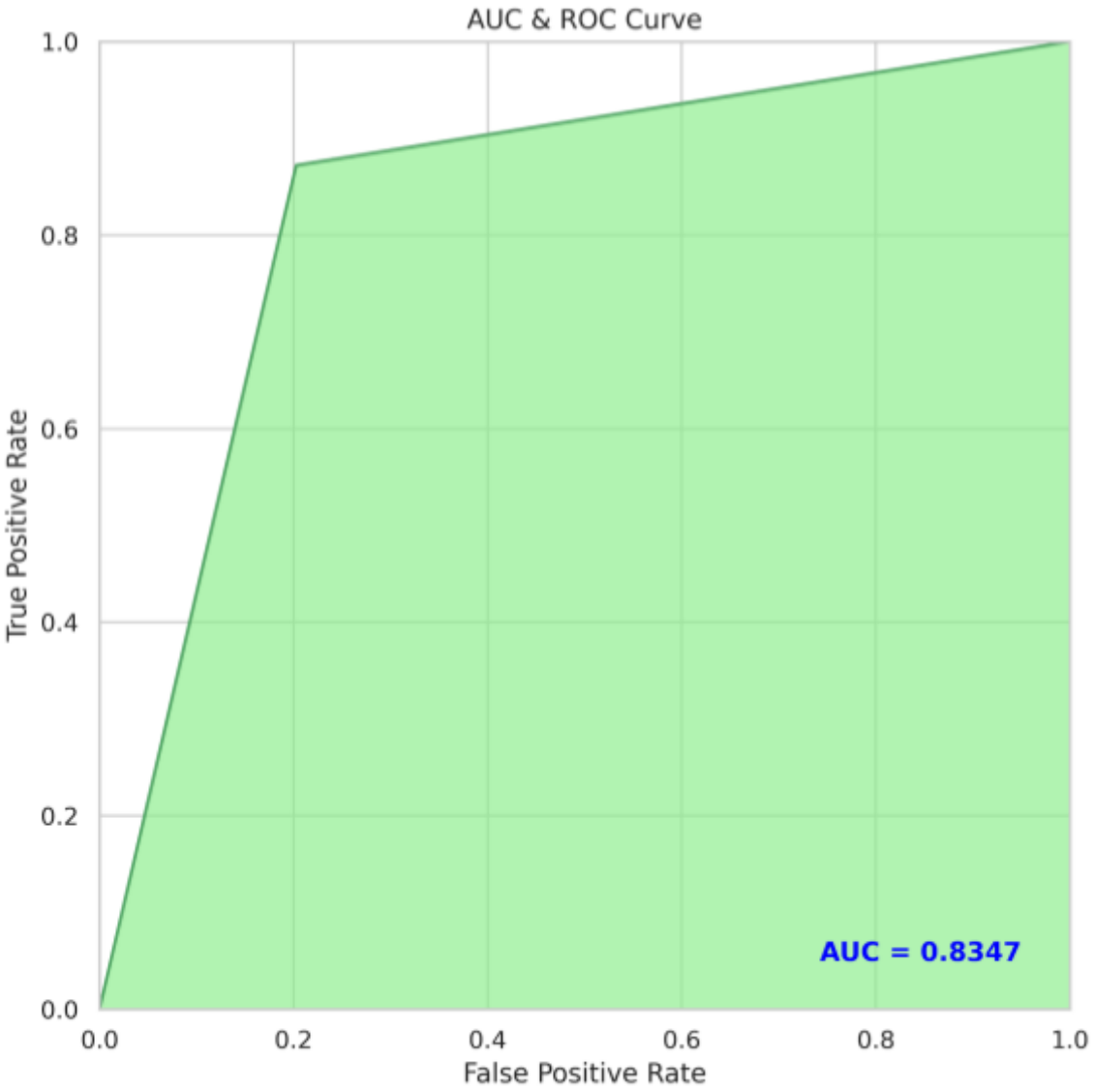

Figure 34: ADA Boost AUC/ROC Curve

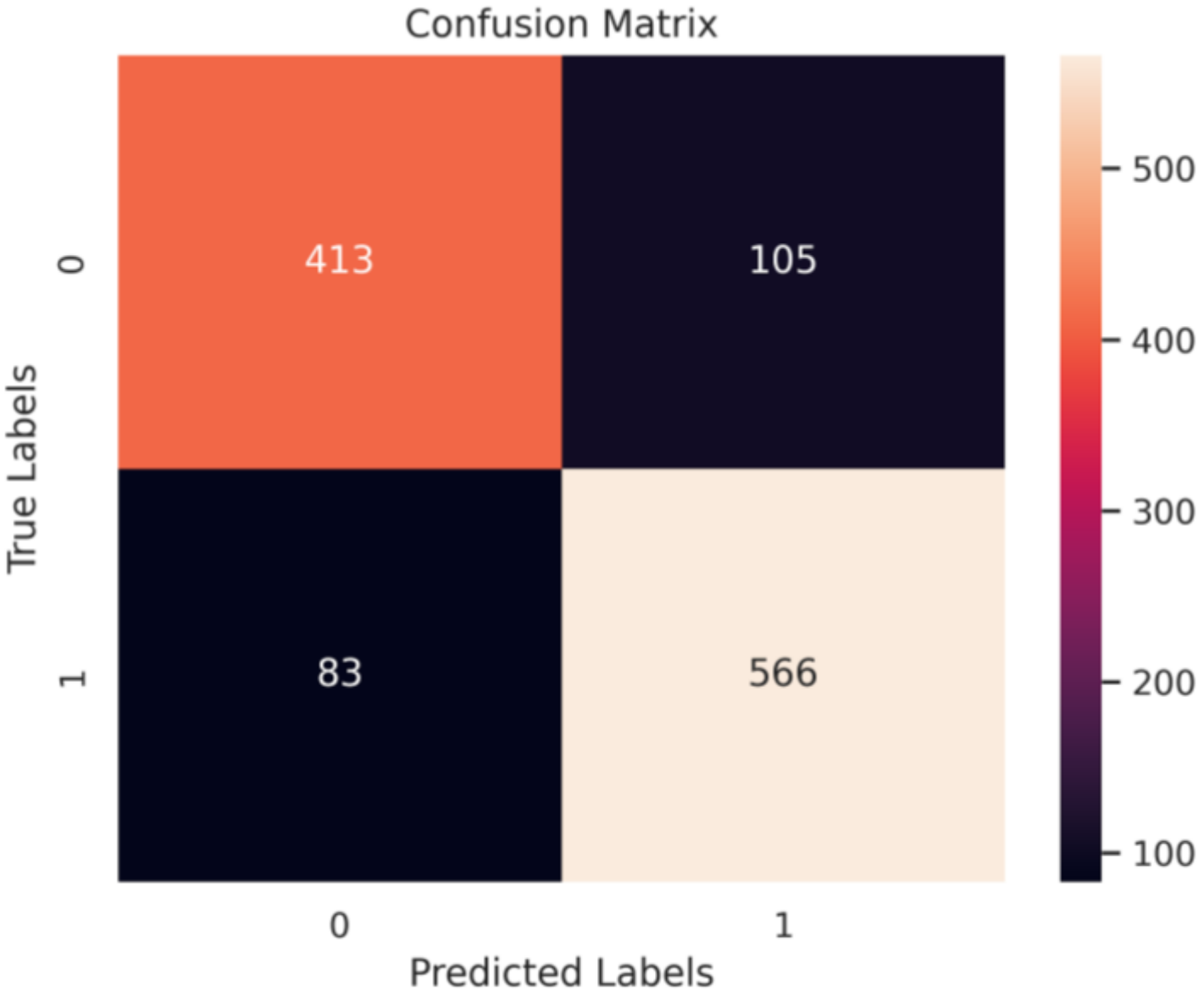

Figure 35: ADA Boost Confusion Matrix

Overall the models performed decently on the dataset with the exception of the Logistic Regression model which performed poorly. Figure 36 shows the accuracy of each model in order from best performing model to worst.

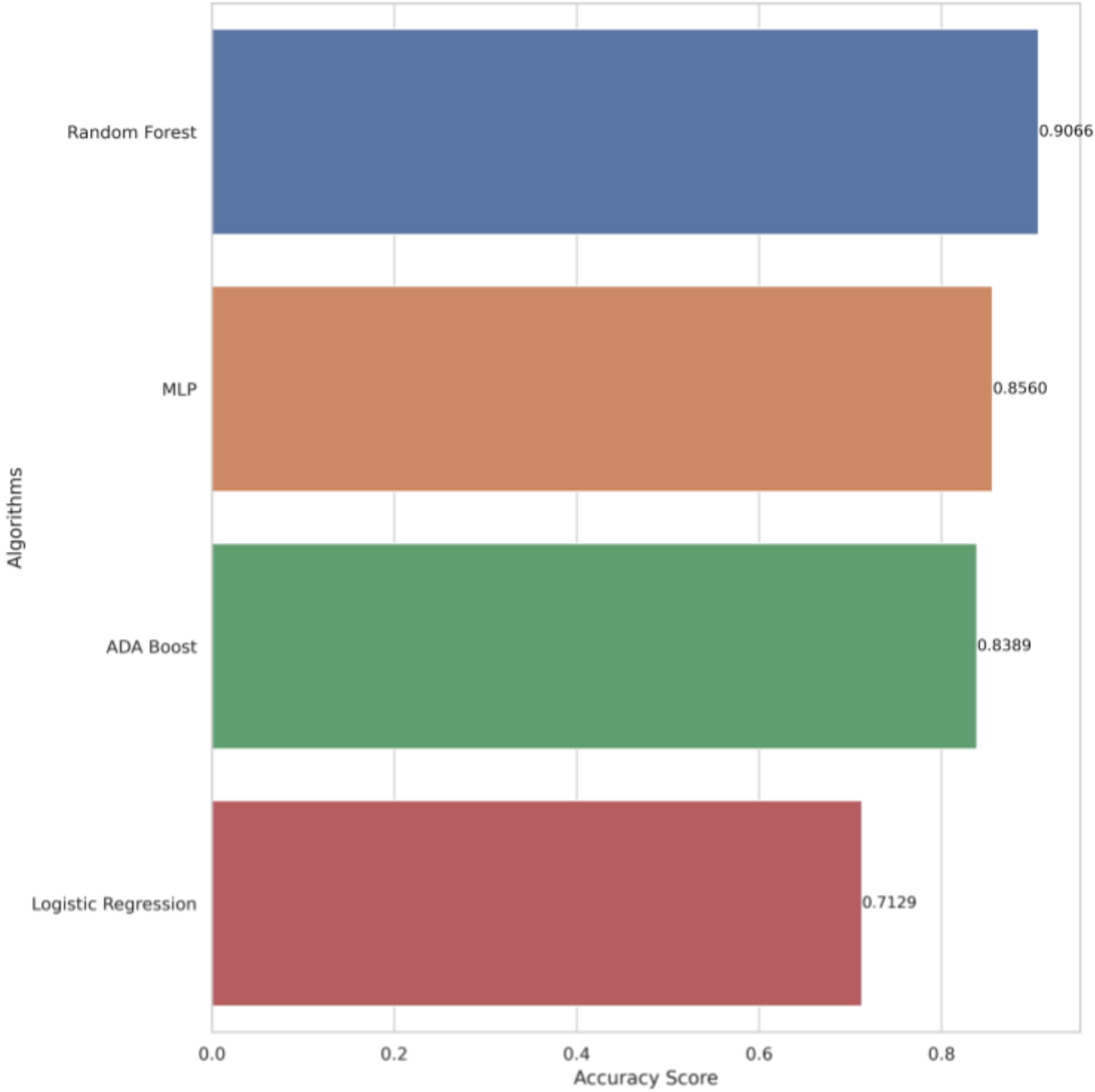

Figure 36: Accuracy Ranking of Models

Part 4: Regression

In this phase, the data underwent similar preparatory steps as previously outlined, with a notable distinction in the dimensionality reduction process. The target variable for regression shifted from gender, as in the classification task, to the attribute "self_time." The selection of attributes for regression involved evaluating their Pearson correlation with the target attribute, as illustrated in Figure 37. This correlation analysis aided in identifying the attributes most relevant to predicting "self_time" in the regression model.

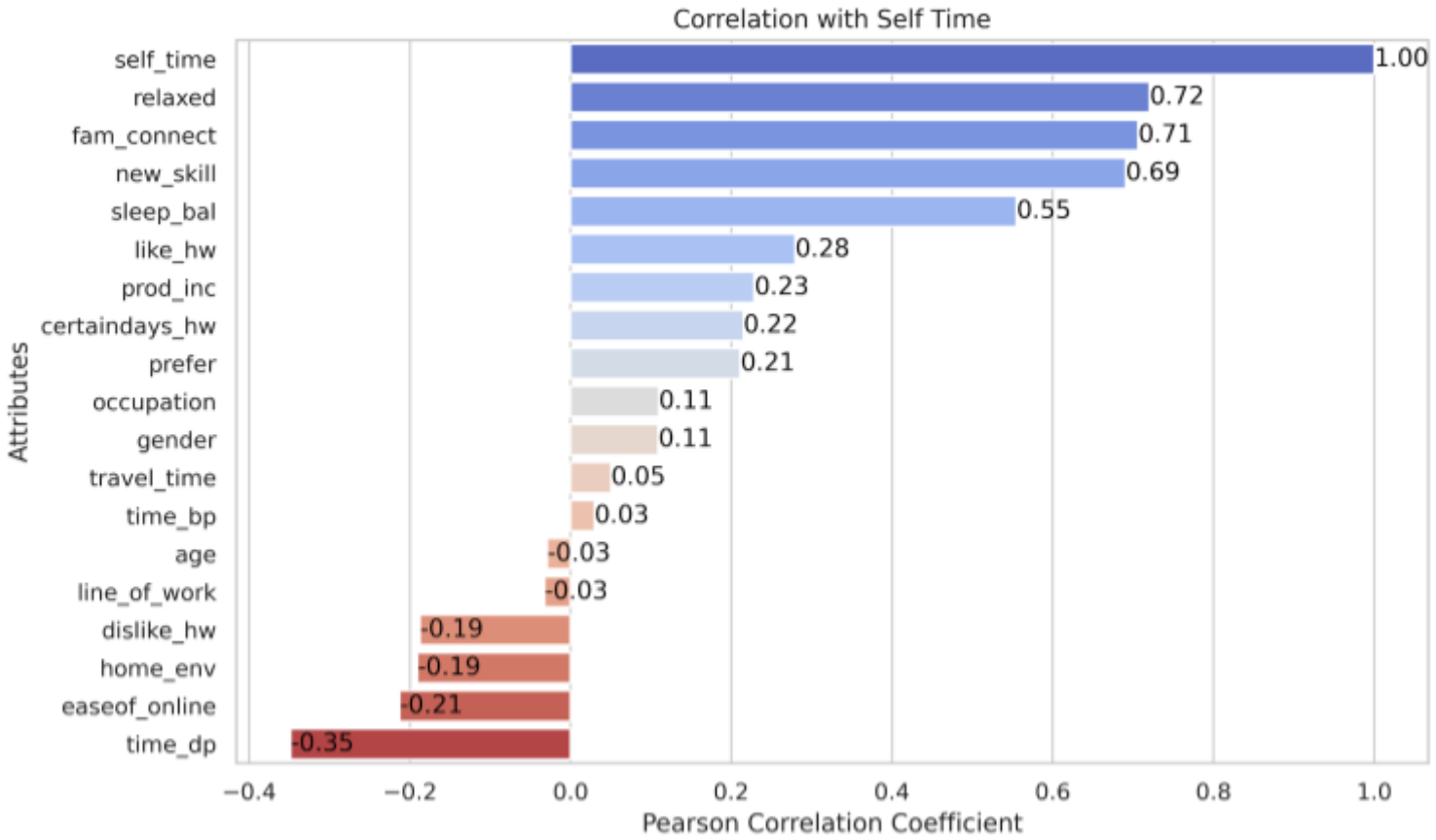

Figure 37: Attributes correlation to Self Time

The models were trained using cross validation as the size of the dataset is very limited and the attributes used to train the regression models were relaxed, fam_connect, new_skill, sleep_bal, time_dp and like_hw as these are the attributes with the highest correlation to the target attribute.

Figure 38 shows the different metrics acquired for the 3 regression models train:

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R2_Score	Adjusted R2_Score
RegressionTree	0.208530	0.099773	0.315868	0.659365	0.657615
LinearRegression	0.254252	0.111152	0.333394	0.620515	0.618566
SVR	0.279447	0.177672	0.421512	0.393406	0.390290

Figure 38: Regression Models Metric Comparison

- Mean Absolute Error (MAE)(6): For MAE, a lower value is better. The MAE indicates that, on average, the model's predictions deviate by approximately 0.21 hours for the regression tree(4), 0.25 hours for linear regression and 0.28 hours for svr(5) from the actual values. Lower MAE values signify more accurate predictions which means that the regression tree had the best score for this metric, then second best was linear regression and last svr.
- Mean Squared Error (MSE)(7): Similar to MAE, for MSE, a lower value is better. The MSE means that, on average, the squared deviations from the actual values are around 0.10 hours for the regression tree,0.11 hours for the linear regression and 0.17 hours for the svr. Lower MSE values indicate better accuracy, with smaller squared errors which means that again the best performing model was the regression tree then closely behind was linear regression and last the svr model.
- Root Mean Squared Error (RMSE)(8): Again, a lower value is better for RMSE. The RMSE provides an interpretable scale for the average prediction error, and a lower RMSE suggests more accurate predictions with smaller deviations. As it is evident once again the best performing model was the regression tree, then closely behind was linear regression and at last again was the svr model.
- R2 Score(9): For the R2 score, a higher value is better. The R2 score indicates how good the model explains the variance in the target attribute. Higher R2 scores signify a better fit, capturing a larger proportion of the variability in the target attribute and from what it is shown in figure 38 it can be stated that again that the regression tree model performed the best then the linear regression model is second and again last was the svr model.
- Adjusted R2 Score: Similar to R2, a higher value is better for the adjusted R2 score. The adjusted R2 score considers the number of predictors and reinforces the model's ability to explain variability in self_time while avoiding overfitting.Again as in all of the metrics the regression tree performed the best then second place is linear regression model and last was the svr model.

In conclusion, the regression tree emerged as the top-performing model, surpassing the other models across all metrics. While linear regression closely followed, demonstrating robust performance, the SVR model exhibited the poorest results, rendering it unsuitable for the dataset. It's worth noting that both the regression tree and linear regression models prove effective for the specific task at hand.

Part 5: Clustering

The data preparation procedures in this section mirrored those in previous sections, with a notable exception in the dimensionality reduction phase. To determine the attributes to be excluded, a Pearson correlation analysis was conducted for all attributes, as depicted in Figure 39.

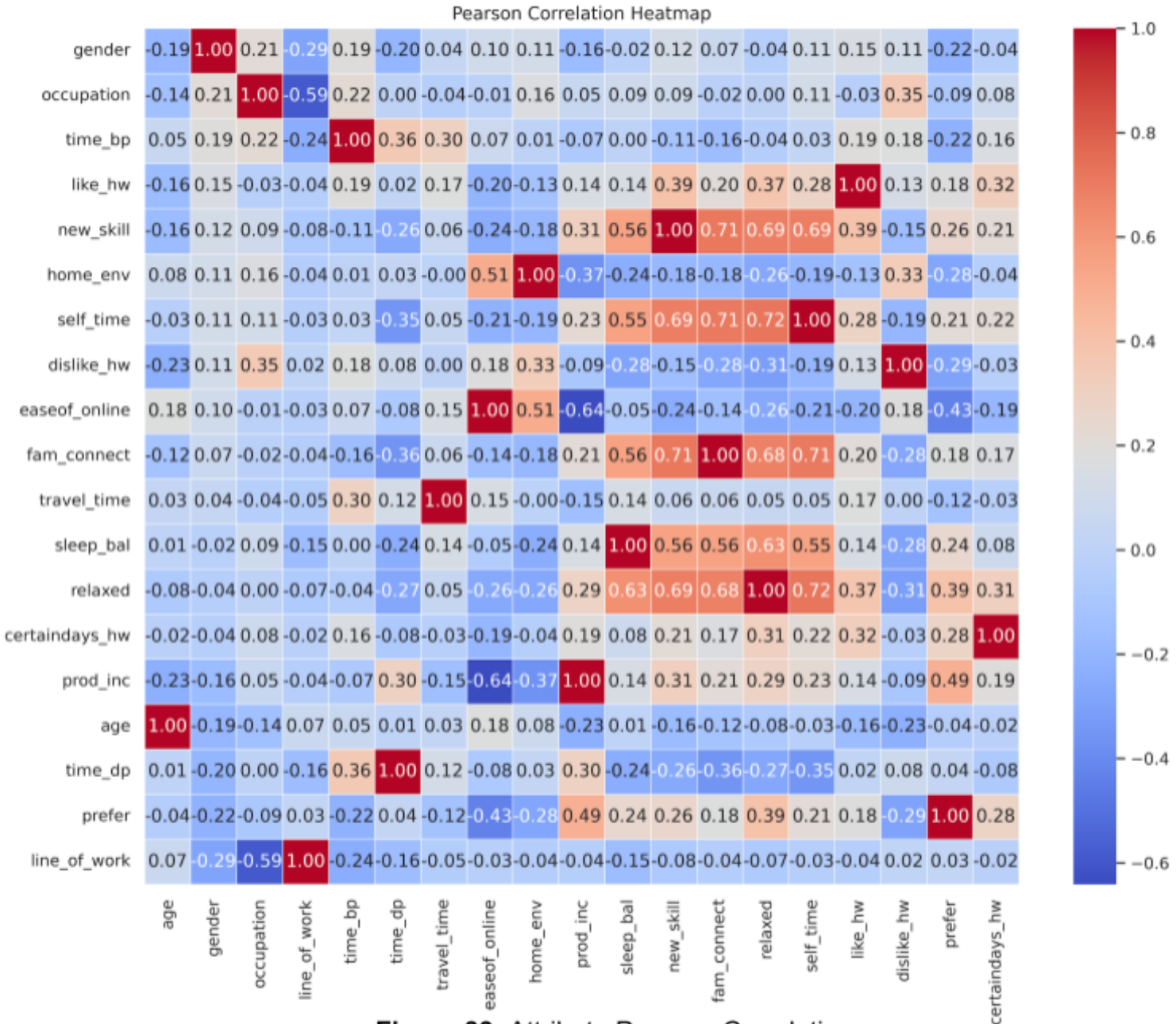

Figure 39: Attribute Pearson Correlation

As it can be seen from Figure 39 the attributes with the lowest values which means they have the lowest correlation to all of the other attributes will be dropped. Attributes gender, travel_time, age, time_bp, dislike_hw, certaindays_hw, time_dp will be dropped as they have the least correlation with the other attributes. Additionally, line_of_work will be excluded, considering its similarity to occupation, where occupation exhibits higher correlation with other attributes and will be retained.

Models comparison:

Algorithms	Cluster 1 Coverage	Cluster 2 Coverage	Silhouette Score
Agglomerative	405	8	0.17
K-Means	629	546	0.16

Figure 40: Clustering Models Comparison

Kmeans:

Before initiating the training of the KMeans clustering model, the optimal number of clusters was determined through the elbow method, as depicted in Figure 41. This method involves visualising the within-cluster sum of squares across a range of cluster numbers to identify the point where the rate of reduction in variance slows down, signifying an optimal choice for the number of clusters. The elbow visualisation indicated that the optimal number of clusters was 2, and that was the chosen configuration for the model.Figure 42 shows the silhouette plot of the 2 clusters and figure 43 visualises the two clusters for better understanding.

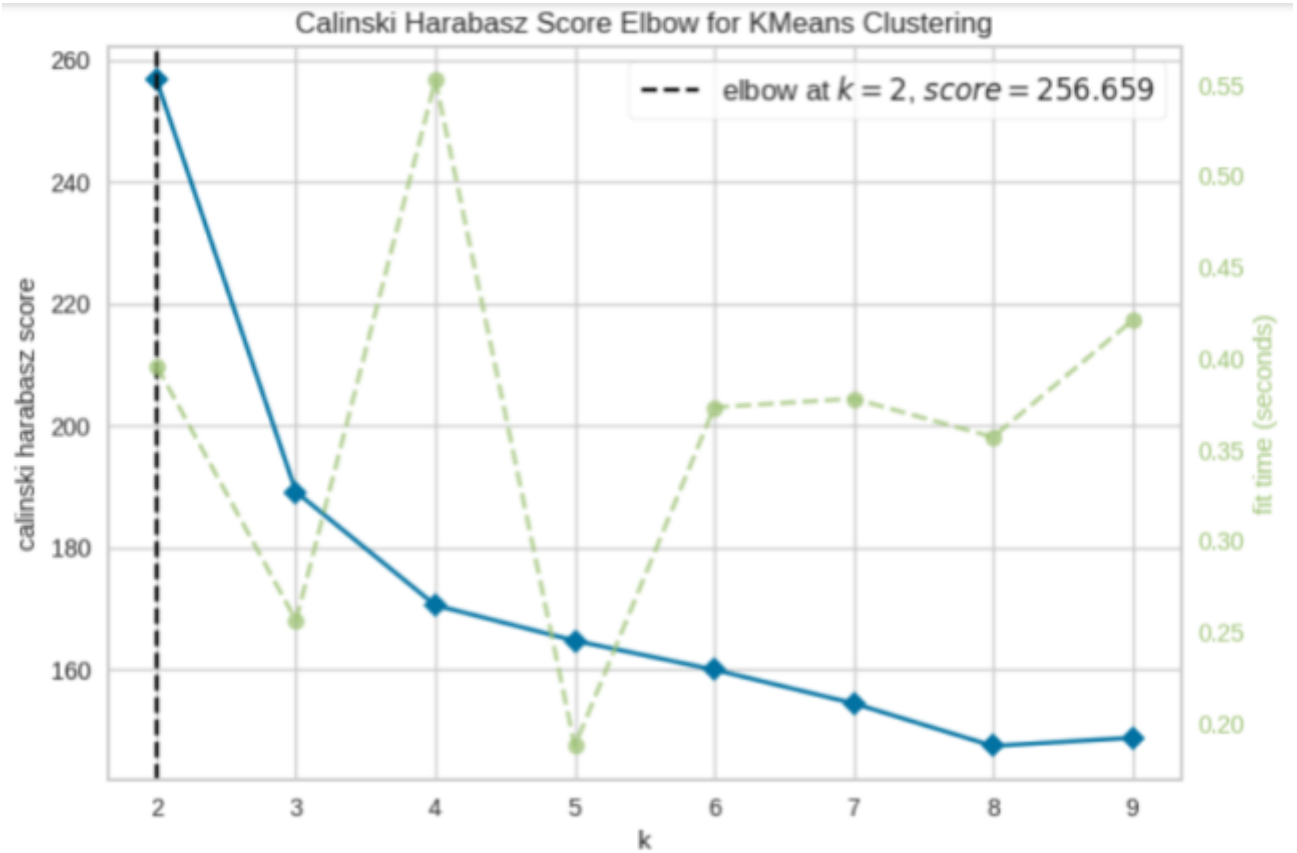

Figure 41: Elbow method visualisation

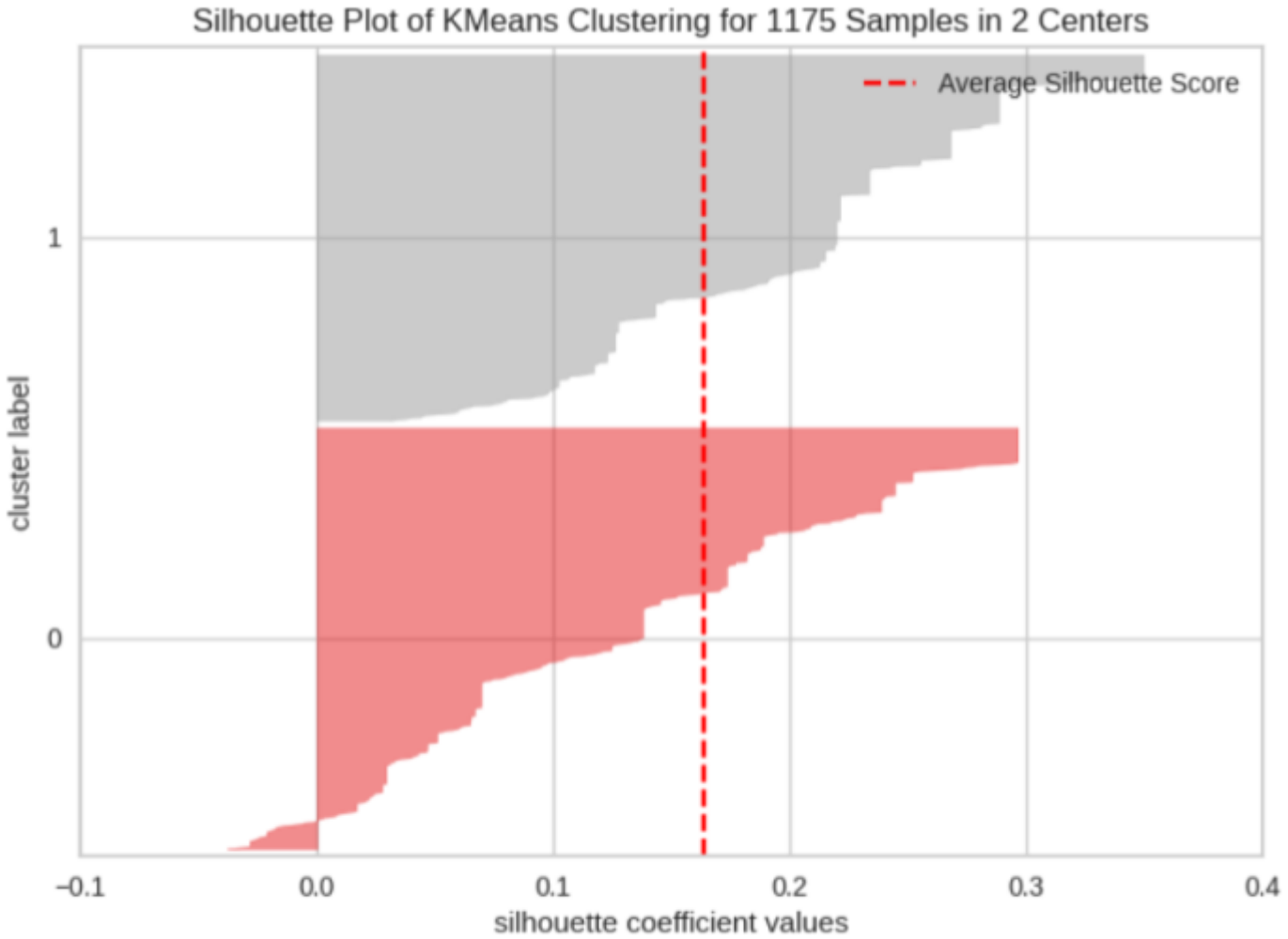

Figure 42: Silhouette score for clusters

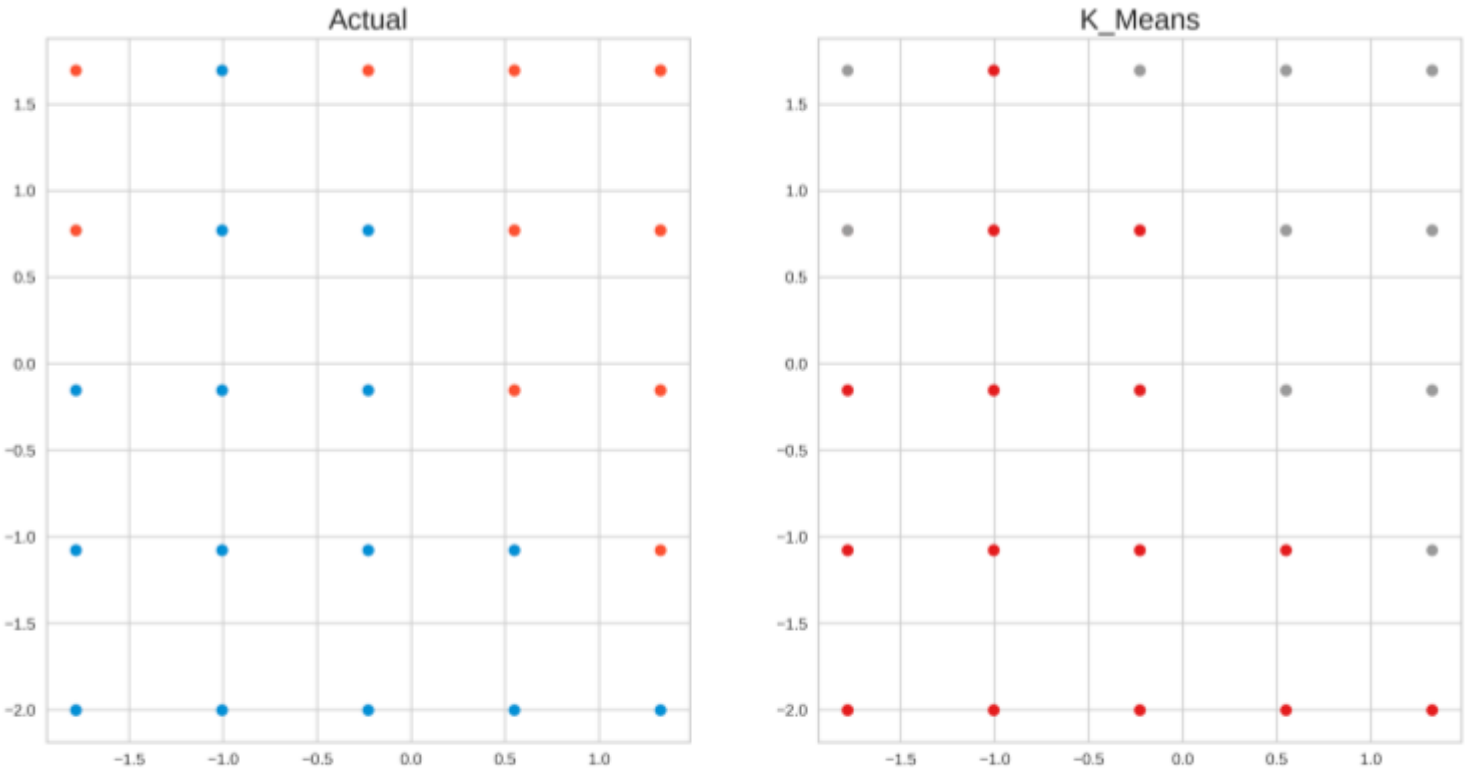

Figure 43: Cluster Visualisation

Agglomerative(10):

Agglomerative clustering is a hierarchical clustering technique that starts with individual data points and merges them step by step based on similarity until a single cluster is formed. It builds a tree-like structure, known as a dendrogram(Figure 44), where the leaves represent individual data points, and the branches illustrate the merging process. The similarity between clusters is measured using a linkage criterion, in this case average linkage, to determine which clusters to merge at each step.

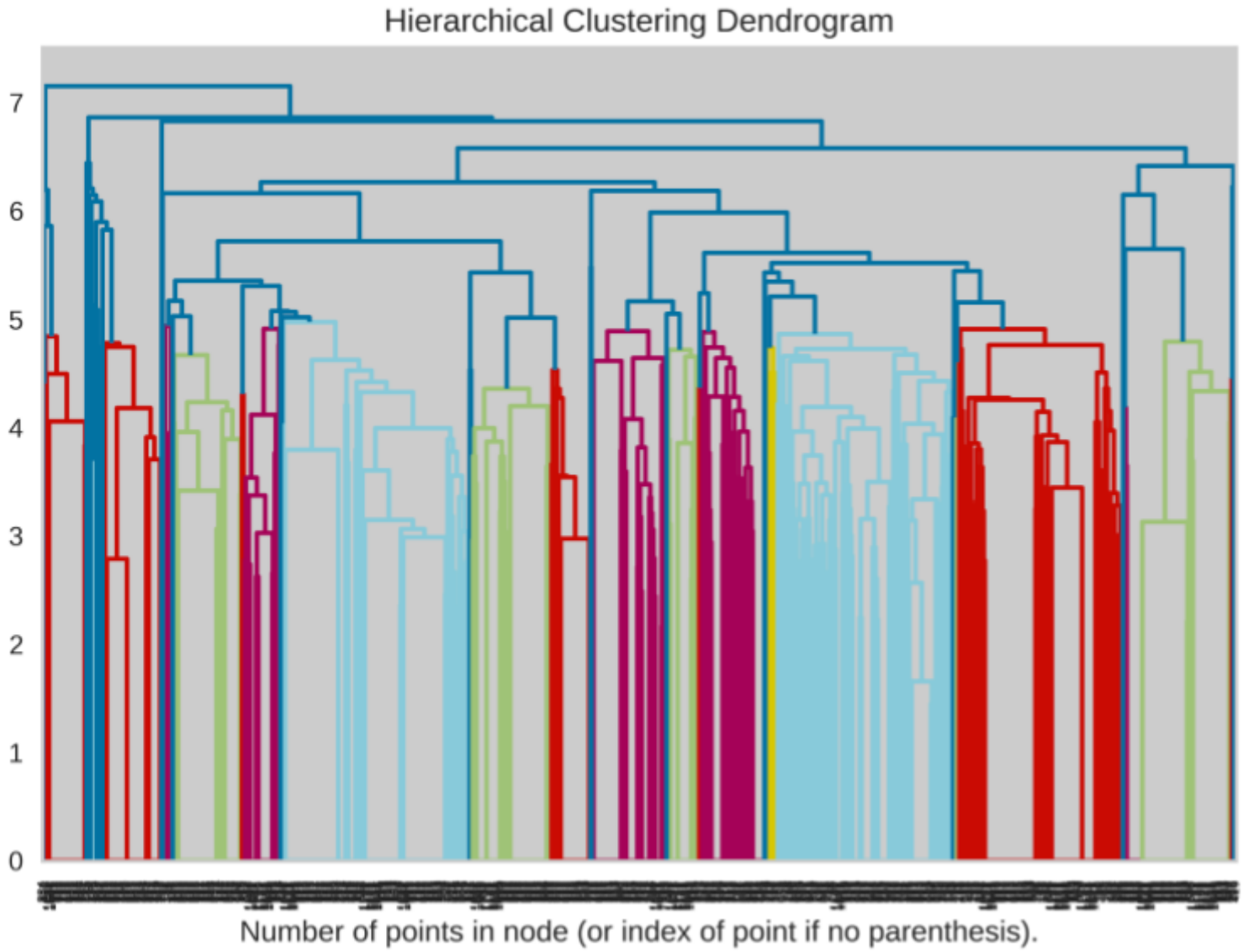

Figure 44: Dendrogram of Agglomerative model

In comparing the clustering models, the Agglomerative and K-Means algorithms were assessed based on their coverage and silhouette score. For Agglomerative, Cluster 1 has a coverage of 405, while Cluster 2 has a coverage of 8, resulting in a silhouette score of 0.17. On the other hand, K-Means exhibits a distribution of 629 instances for Cluster 1 and 546 instances for Cluster 2, with a silhouette score of 0.16. In terms of coverage, both algorithms show notable differences, with Agglomerative having a more balanced representation across clusters. However, the silhouette scores for both models indicate that the clusters are not well-separated. The silhouette score reflects both intra-class similarity and inter-class dissimilarity, and a low value, as seen in both cases, suggests that the clusters may not be distinct, impacting the appropriateness of the assigned clusters for each data point making them unfit for the task.

References

1. Sklearn.preprocessing.MinMaxScaler.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.https://scikit-learn.org/stable/about.html#citing-scikit-learn
3. Sklearn.ensemble.adaboostclassifier.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
4. Sklearn.tree.decisiontreeregressor.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
5. Sklearn.svm.SVR.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
6. SKLEARN.METRICS.MEAN_ABSOLUTE_ERROR.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html
7. SKLEARN.METRICS.MEAN_SQUARED_ERROR.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
8. SKLEARN.METRICS.ROOT_MEAN_SQUARED_ERROR.scikit. https://scikit-learn.org/dev/modules/generated/sklearn.metrics.root_mean_squared_error.html
9. Sklearn.metrics.r2_score.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
10. Sklearn.cluster.AgglomerativeClustering.scikit. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html