Data-Driven Decision Making: Exploring Distributions, Transformations, and Predictive Modeling in Network Analysis

Brandon Toews

University of Gloucestershire

Table of Contents

1. Description

In this section, we elaborate on the steps taken to clean the dataset obtained from ‘ookla_speed_q4_2022.csv’. The dataset, consisting of 20,000 entries and 7 features related to network performance, underwent a comprehensive cleaning process.

1.1. Initial Data Overview

Upon loading the dataset using the pandas library in Python, we observed that it contained 20,000 entries with 7 columns. An initial assessment revealed the presence of missing values in the 'avg_lat_down_ms' and 'avg_lat_up_ms' columns. (Fig. 1)

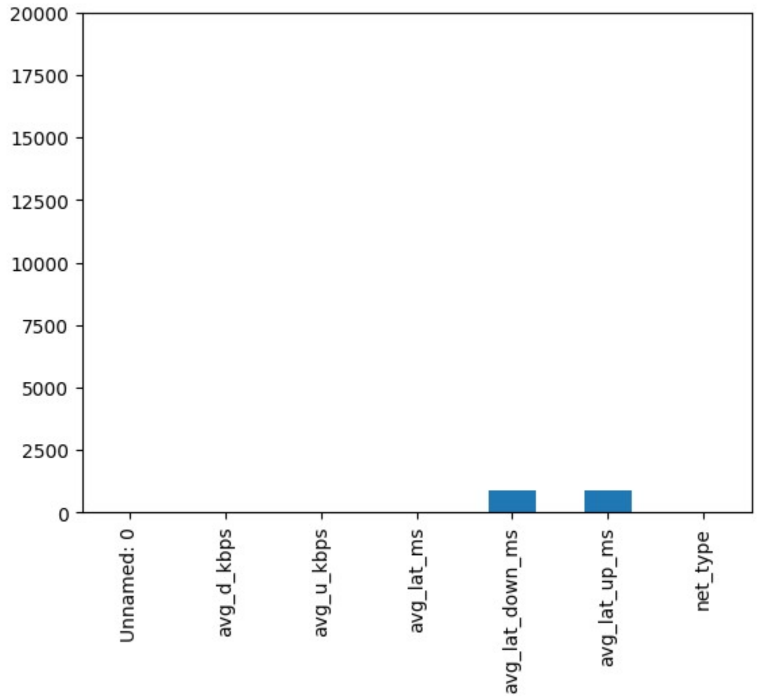

Figure 1. Bar plot of missing values in each column

1.2. Data Cleaning Steps

- **Dropping Rows with Missing Values:** Rows containing missing values were dropped to ensure the reliability of our subsequent analyses. (Fig. 1)
- **Column Removal:** We removed the unnecessary 'Unnamed: 0' column, as it served as an unnamed index and did not contribute to the analysis.

- **Spelling Corrections and Categorization:** We addressed spelling errors in the 'net_type' column, changing 'moblie' to 'Mobile' and capitalizing 'fixed'. The 'net_type' column was then converted to a categorical data type.
- **Duplicate Entry Removal:** Duplicate entries were identified and subsequently dropped to ensure the uniqueness of our data.
- **Conversion of Float Columns to Int:** We verified the 'avg_lat_down_ms' and 'avg_lat_up_ms' columns for floating-point values and converted them to integers if necessary.

1.3. Column Renaming and Unit Conversion

To enhance clarity, we renamed columns related to average download and upload speeds and converted the corresponding values from kilobits per second to megabits per second.

1.4. Resulting Dataset

The resulting cleaned dataset, now saved as 'cleaned_dataset.parquet', comprises 19,030 entries and 6 columns, each with non-null values. The 'net_type' column is categorized into 'Mobile' and 'Fixed'. The dataset is now ready for further analysis and modeling.

2. Comprehensive Data Analysis

In this section, we delve into the exploratory data analysis (EDA) process, aiming to comprehend the underlying distributions, compare fixed and mobile network data, and identify any notable correlations. Recognizing that the initial data exhibited heavily positively skewed distributions, we undertook a series of data transformations to bring the distributions closer to normality. The primary objective was to enhance the suitability of the data for subsequent hypothesis testing.

2.1. Understanding Initial Distributions

The initial step involved an examination of the distributions of both fixed and mobile network data. Histograms, box plots (Figs. 2-6), and summary statistics (Tabs. 1 & 2) were employed to gain insights into the central tendencies, dispersions, and skewness of the datasets. Notably, the distributions were observed to be heavily positively skewed, prompting the need for transformation to meet the assumptions of parametric statistical tests.

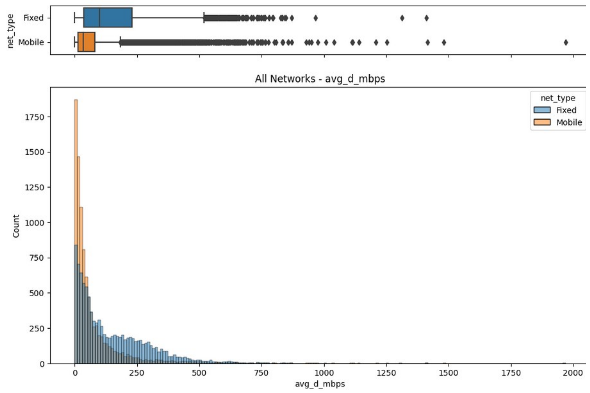

Figure 2. Box/Hist plots of ’avg_d_mbps’

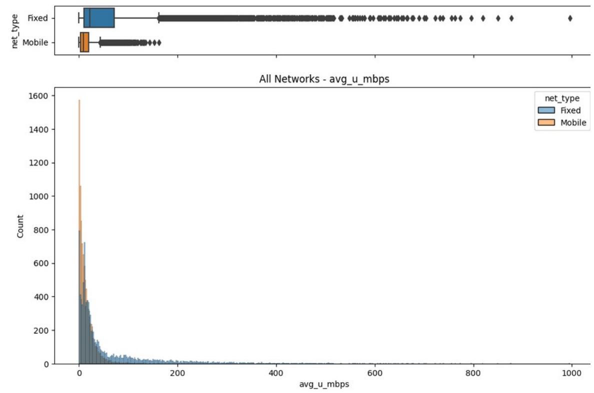

Figure 3. Box/Hist plots of ’avg_u_mbps’

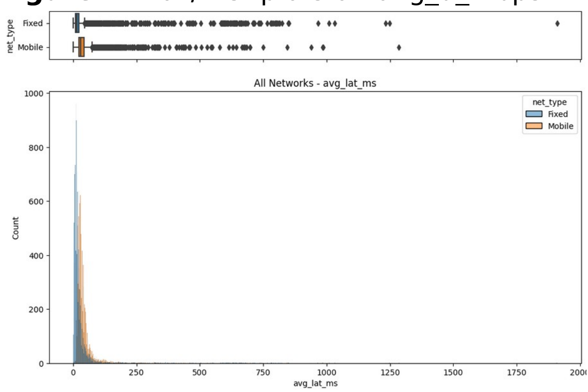

Figure 4. Box/Hist plots of ’avg_lat_ms’

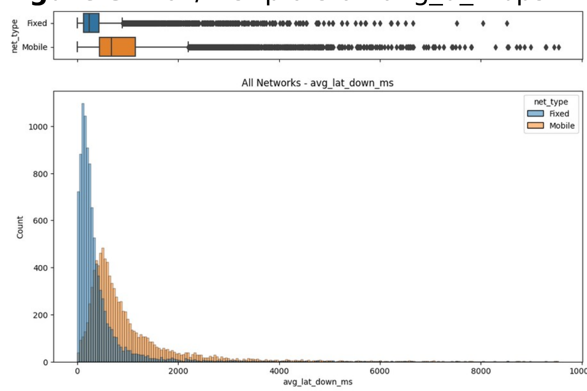

Figure 5. Box/Hist plots of ’avg_lat_down_ms’

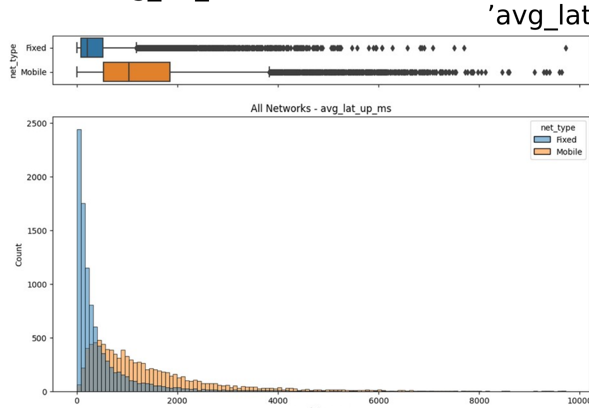

Figure 6. Box/Hist plots of ’avg_lat_up_ms’

	Network Type	Min	Max	Mode	Median	Mean	± Std Dev
avg_d_mbps	Fixed	0.049	1411.955	6.44	100.1625	147.033	139.263
avg_d_mbps	Mobile	0.012	1969.184	0.968	33.405	72.726	113.78
avg_u_mbps	Fixed	0.005	995.535	0.477	22.4605	60.684	93.141
avg_u_mbps	Mobile	0.002	162.581	0.626	9.431	14.18	15.827
avg_lat_ms	Fixed	1.0	1909.0	11.0	15.0	27.525	71.193
avg_lat_ms	Mobile	1.0	1285.0	28.0	31.0	40.786	46.625
avg_lat_down_ms	Fixed	2.0	8518.0	115.0	236.5	386.214	548.102
avg_lat_down_ms	Mobile	5.0	9538.0	469.0	682.0	978.102	949.854
avg_lat_up_ms	Fixed	4.0	9718.0	39.0	209.0	462.82	714.226
avg_lat_up_ms	Mobile	3.0	9641.0	365.0	1030.5	1411.871	1274.976

Table 1. Summary statistics of each network type

	Network Type	Skew	Kurtosis
avg_d_mbps	Everything	2.21999	8.5182
avg_d_mbps	Fixed	1.37781	2.69248
avg_d_mbps	Mobile	4.10912	28.08801

	Network Type	Skew	Kurtosis
avg_u_mbps	Everything	4.29384	25.16311
avg_u_mbps	Fixed	3.07749	12.79495
avg_u_mbps	Mobile	2.64466	11.07589

	Network Type	Skew	Kurtosis
avg_lat_ms	Everything	10.3845	153.31423
avg_lat_ms	Fixed	10.00117	132.10633
avg_lat_ms	Mobile	10.49187	169.85494

	Network Type	Skew	Kurtosis
avg_lat_down_ms	Everything	3.65544	19.98137
avg_lat_down_ms	Fixed	5.22333	40.61823
avg_lat_down_ms	Mobile	3.21657	15.08425

	Network Type	Skew	Kurtosis
avg_lat_up_ms	Everything	2.43104	7.82215
avg_lat_up_ms	Fixed	3.75387	20.18589
avg_lat_up_ms	Mobile	1.98862	5.09926

Table 2. Skew and kurtosis values for everything and each network type

2.2. Comparative Analysis

To assess the disparities between fixed and mobile networks, we conducted thorough comparative analyses. Kernel density plots (Figs. 7 - 16) and statistical tests (Table 2) were leveraged to highlight variations in central tendencies. These comparisons served as a foundation for subsequent transformations and allowed us to pinpoint differences between the two networks.

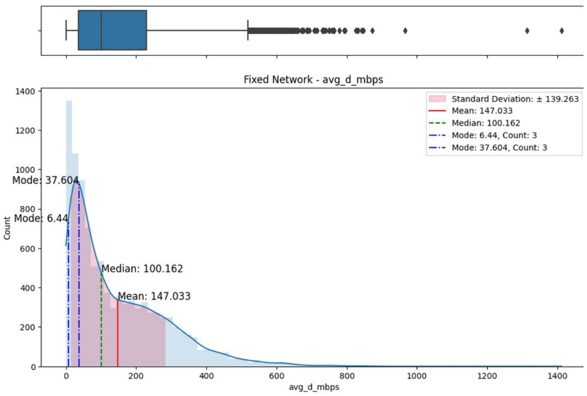

Figure 7. Plots of 'avg_d_mbps' for fixed network

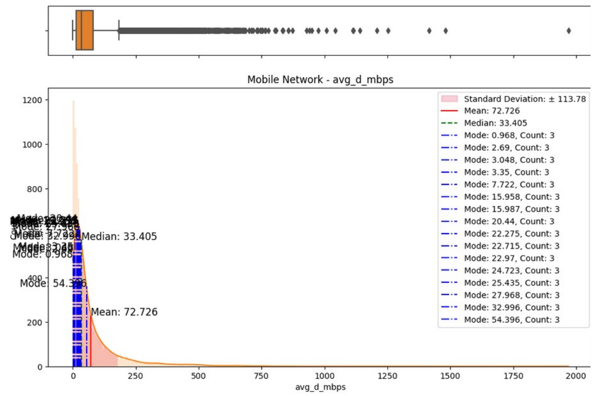

Figure 8. Plots of 'avg_d_mbps' for mobile network

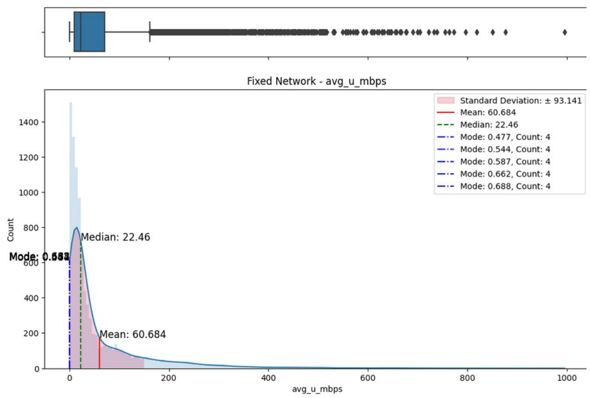

Figure 9. Plots of 'avg_u_mbps' for fixed network

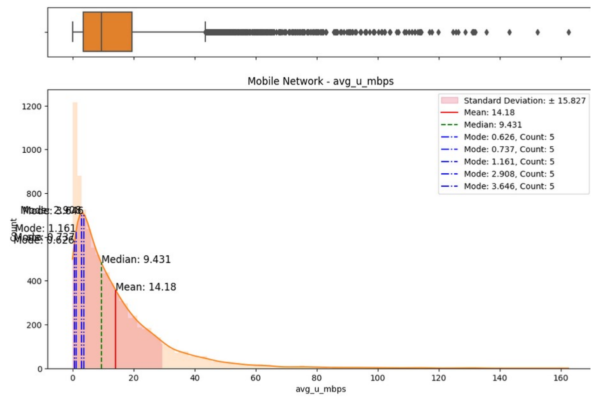

Figure 10. Plots of 'avg_u_mbps' for mobile network

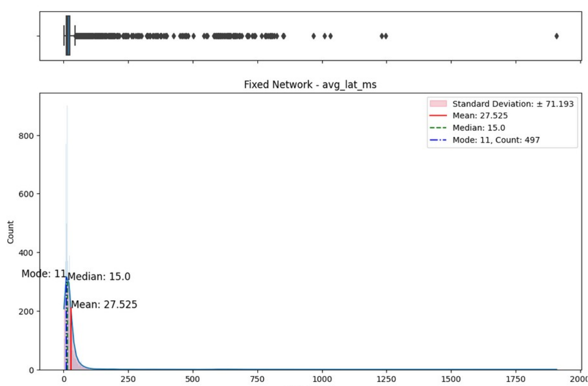

Figure 11. Plots of 'avg_lat_ms' for fixed network

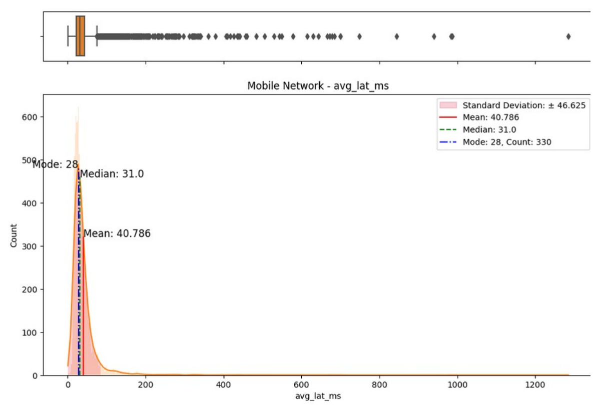

Figure 12. Plots of 'avg_lat_ms' for mobile network

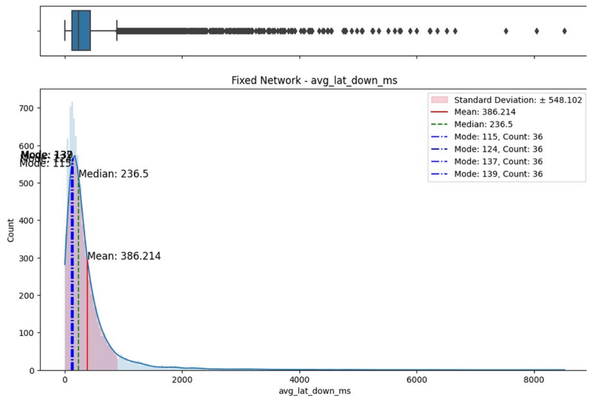

Figure 13. Plots of 'avg_lat_down_ms' for fixed network

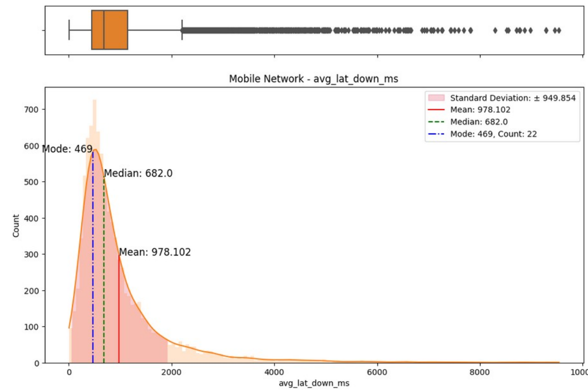

Figure 14. Plots of 'avg_lat_down_ms' for mobile network

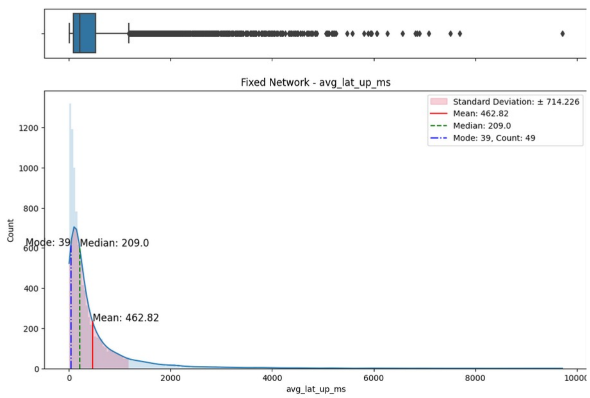

Figure 15. Plots of 'avg_lat_up_ms' for fixed network

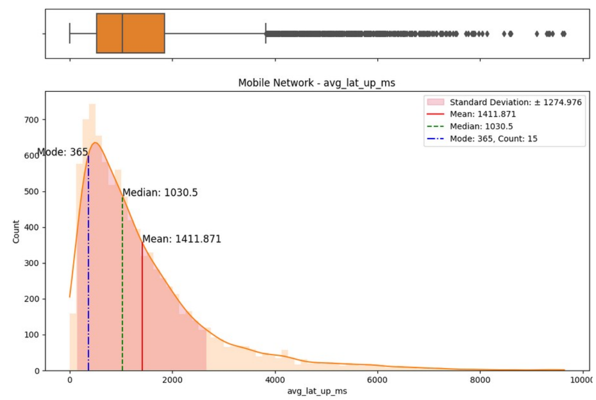

Figure 16. Plots of 'avg_lat_up_ms' for mobile network

2.3. Data Transformations

Several data transformations were applied, including but not limited to logarithmic, Box-Cox, and Yeo-Johnson transformations. Each transformation was carefully chosen based on its appropriateness for the given context and the nature of the initial distributions. Log transformations, for instance, are effective in addressing exponential growth patterns, while Box-Cox transformations are versatile in handling skewed data. (Lee, S. X. and McLachlan, G. J., 2022) (West, R. M., 2022)

2.4. Comparative Assessment of Transformations

A meticulous examination of the transformed datasets ensued, involving comparative analyses with the original data. Visualizations (Figs. 17 - 24) and statistical measures, including skewness and kurtosis tests (See Table 3), were employed to quantify the improvements brought about by each transformation. The Yeo-Johnson transformation consistently demonstrated superior results in terms of bringing the data closer to a normal distribution. (Figs. 25 - 34)

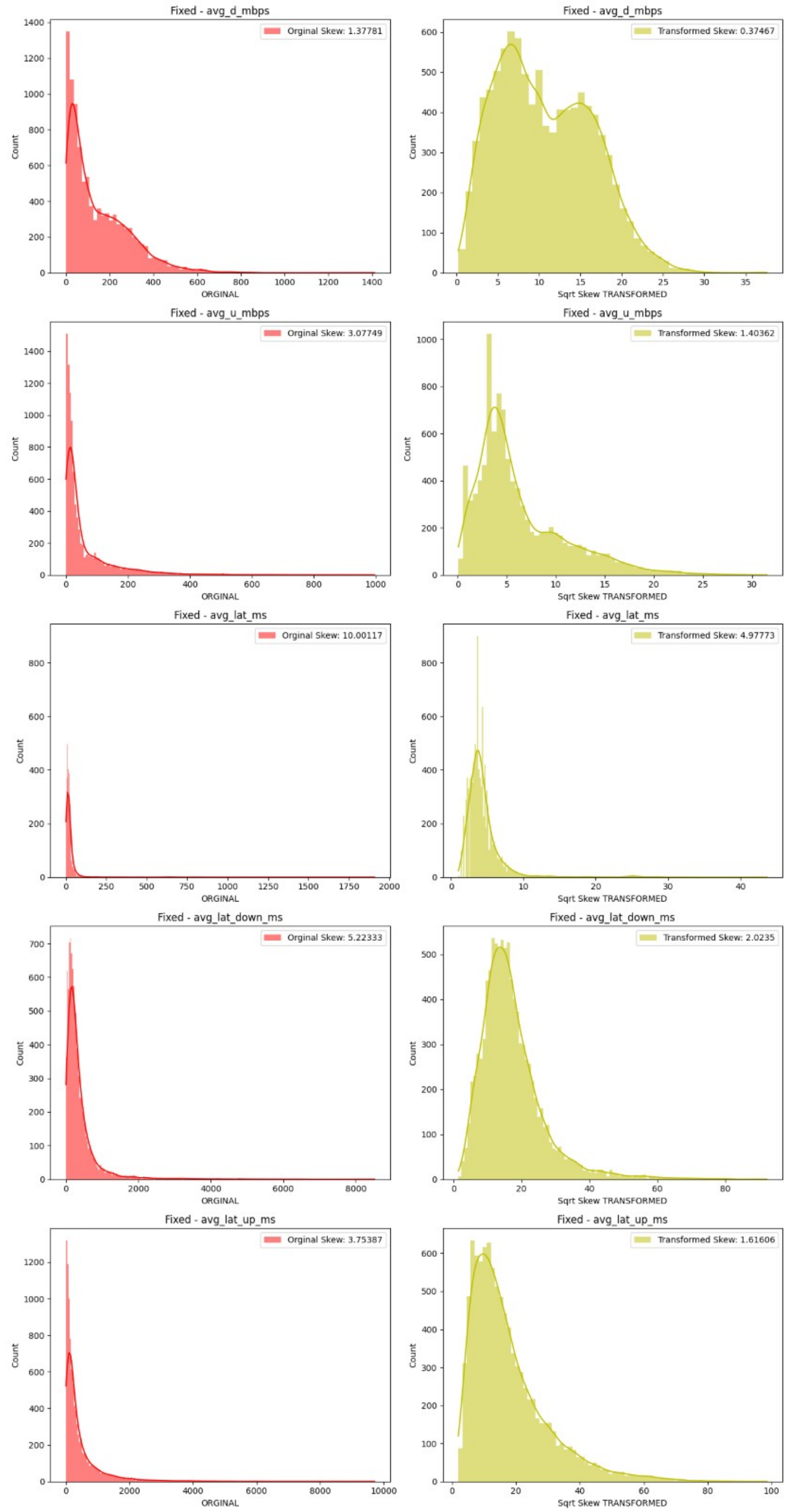

Figure 17. Comparison of original & Sqrt transformed on fixed network

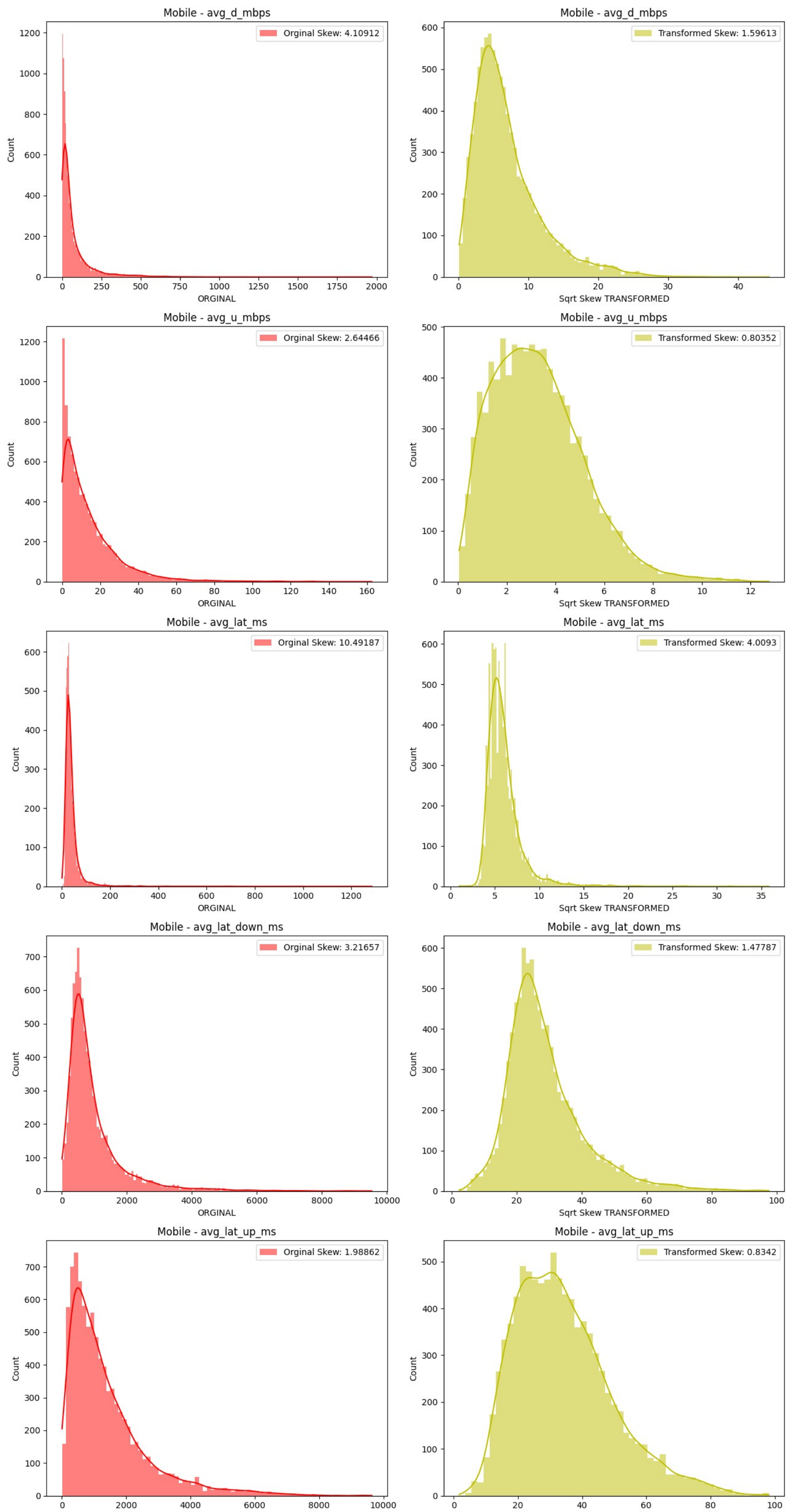

Figure 18. Comparison of original & Sqrt transformed on mobile network

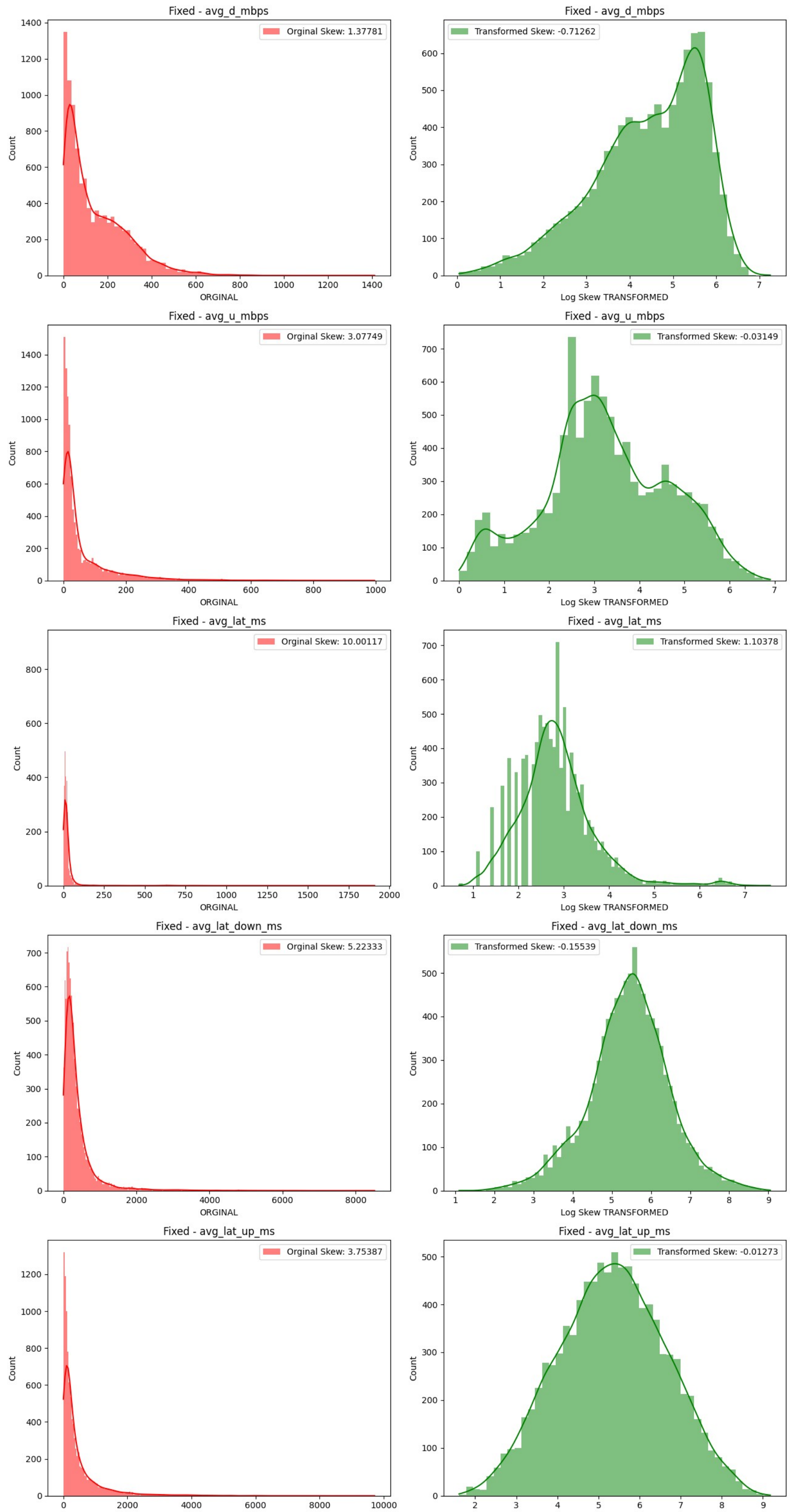

Figure 19. Comparison of original & Log transformed on fixed network

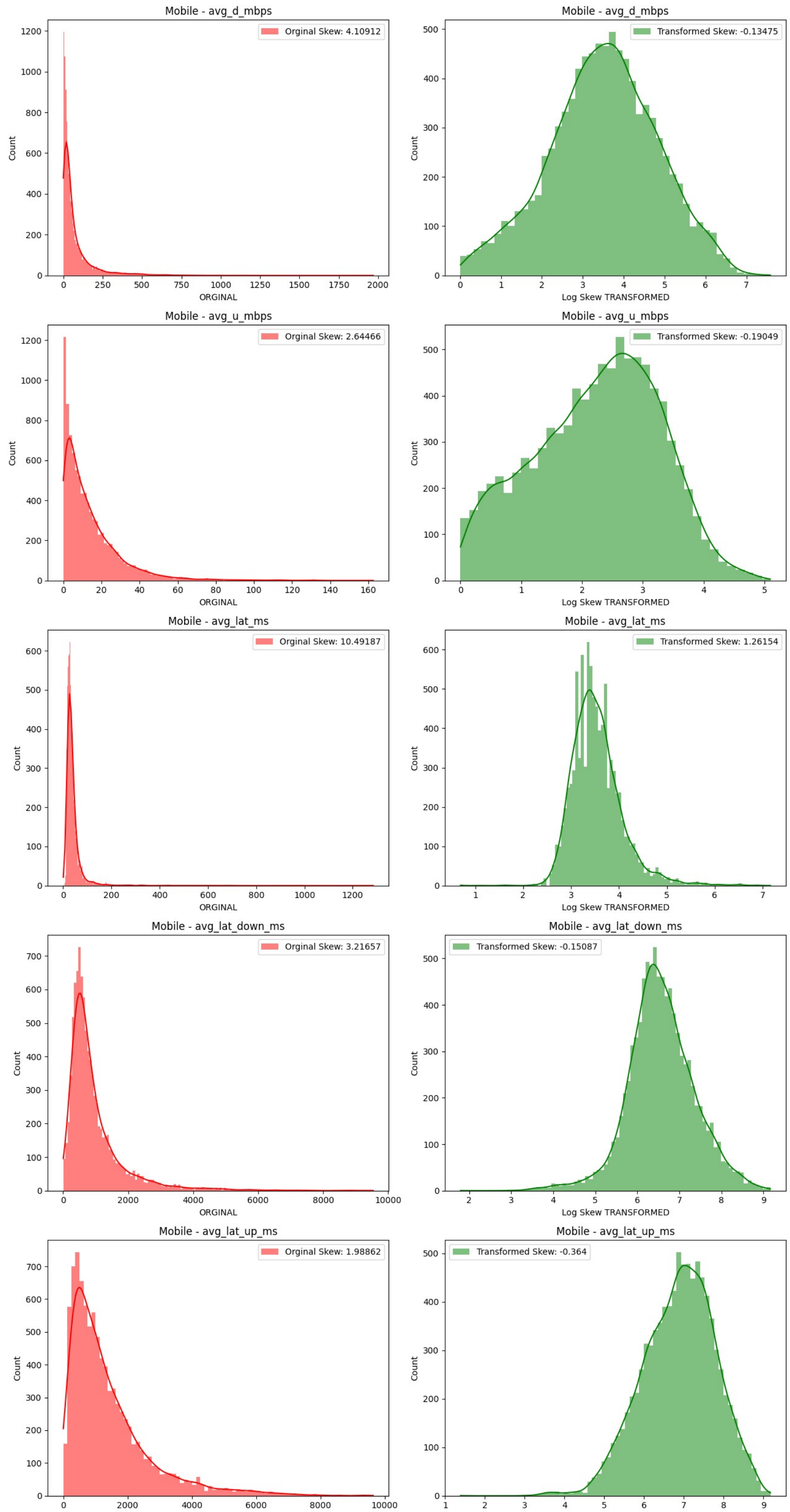

Figure 20. Comparison of original & Log transformed on mobile network

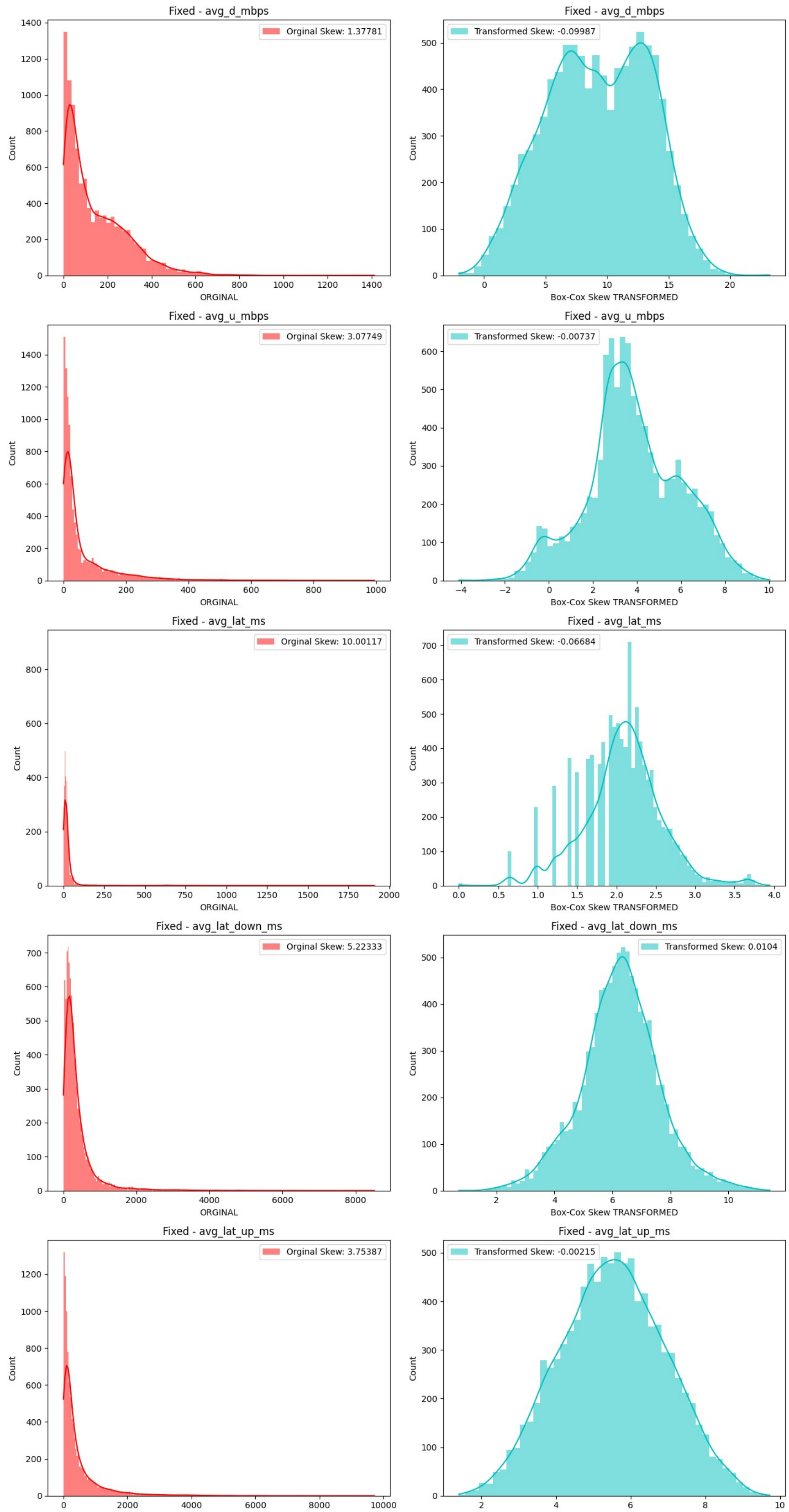

Figure 21. Comparison of original & Box-Cox transformed on fixed network

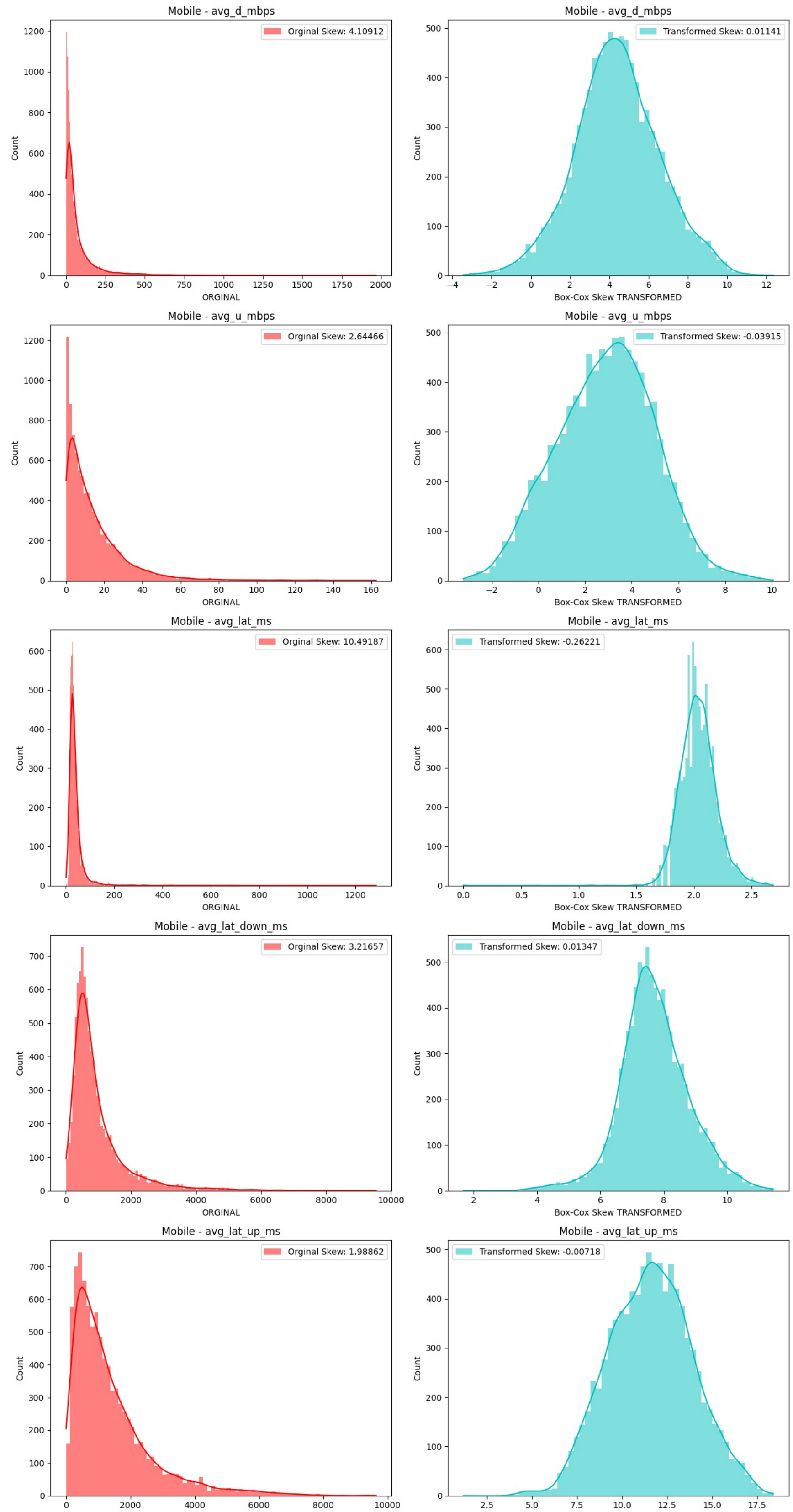

Figure 22. Comparison of original & Box-Cox transformed on mobile network

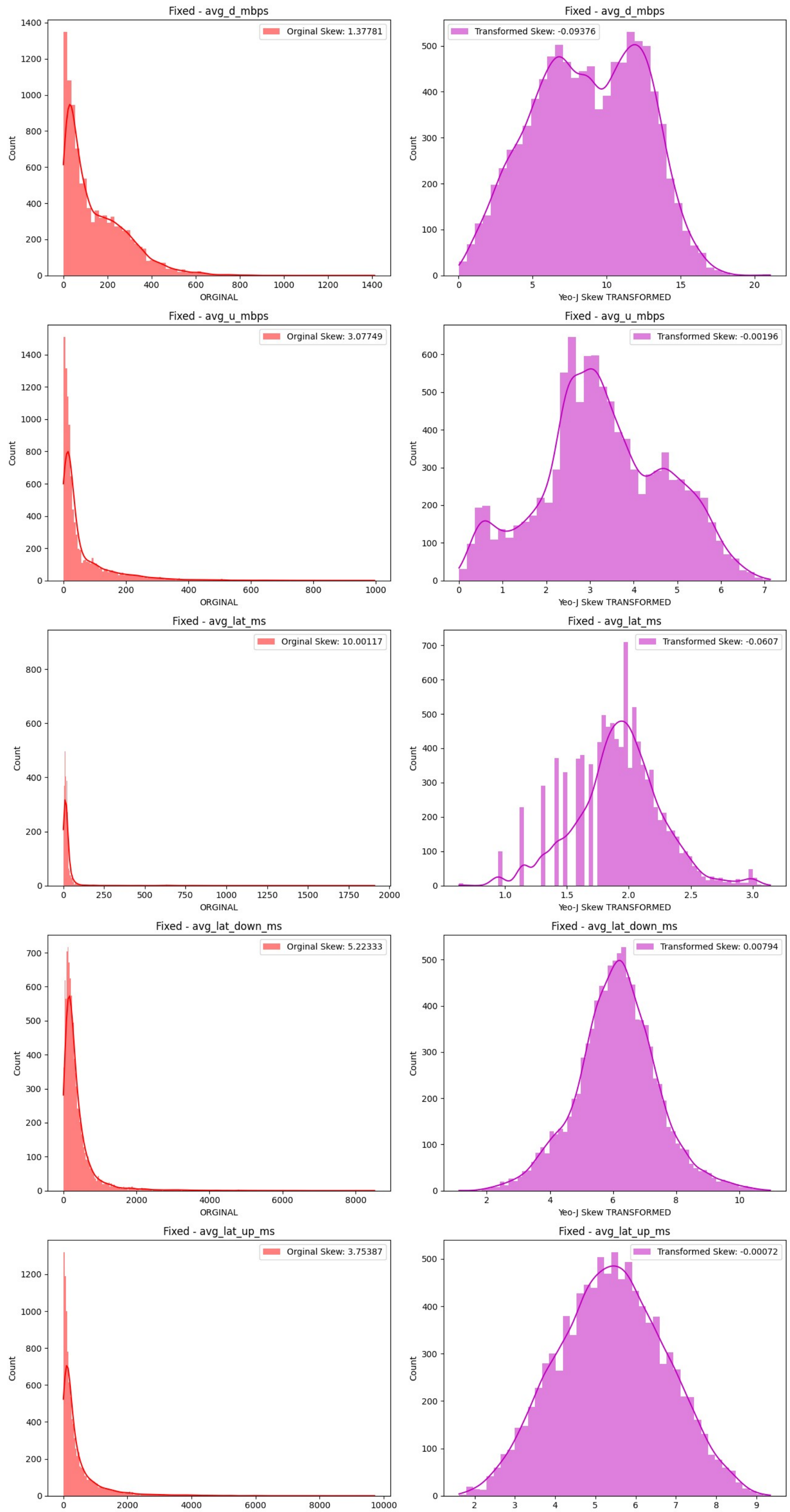

Figure 23. Comparison of original & Yeo-Johnson transformed on fixed network

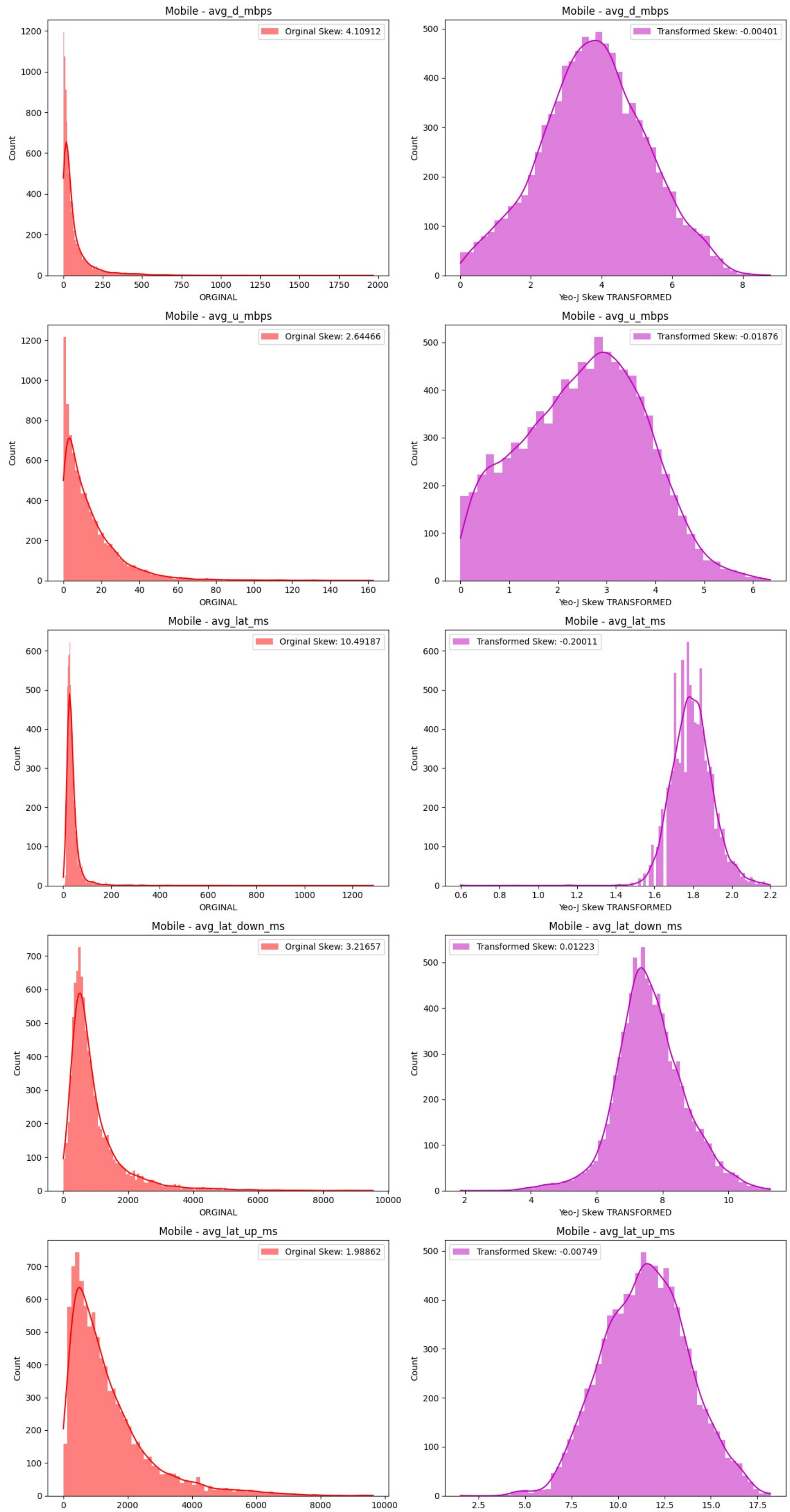

Figure 24. Comparison of original & Yeo-Johnson transformed on mobile network

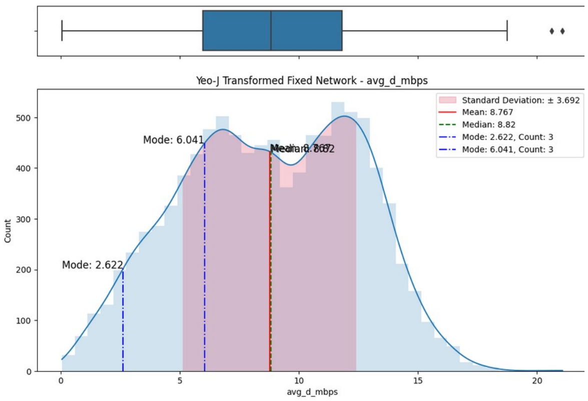

Figure 25. Plots of 'avg_d_mbps' for Yeo-Johnson transformed fixed network

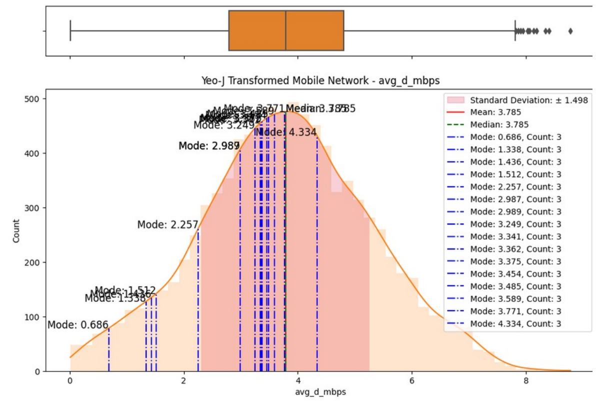

Figure 26. Plots of 'avg_d_mbps' for Yeo-Johnson transformed mobile network

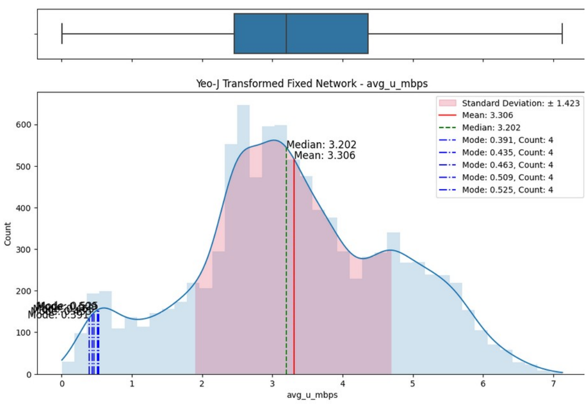

Figure 27. Plots of 'avg_u_mbps' for Yeo-Johnson transformed fixed network

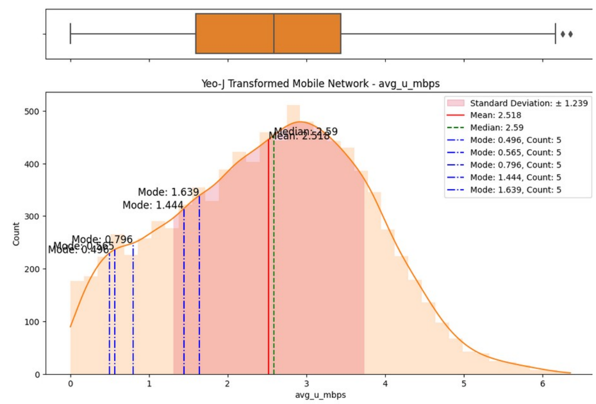

Figure 28. Plots of 'avg_u_mbps' for Yeo-Johnson transformed mobile network

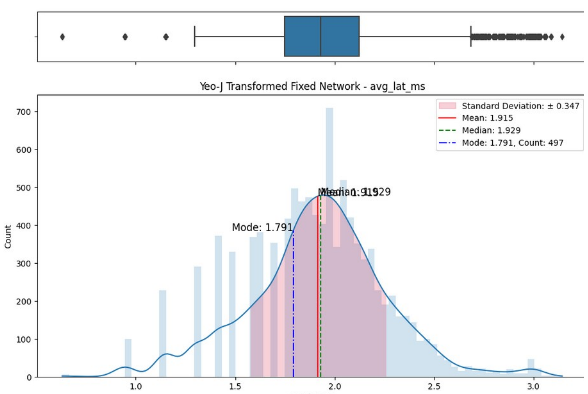

Figure 29. Plots of 'avg_lat_ms' for Yeo-Johnson transformed fixed network

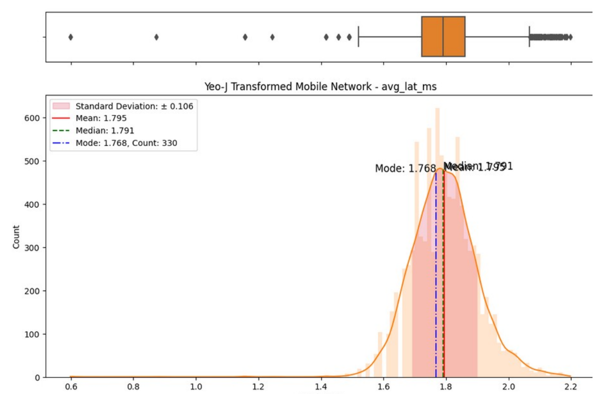

Figure 30. Plots of 'avg_lat_ms' for Yeo-Johnson transformed mobile network

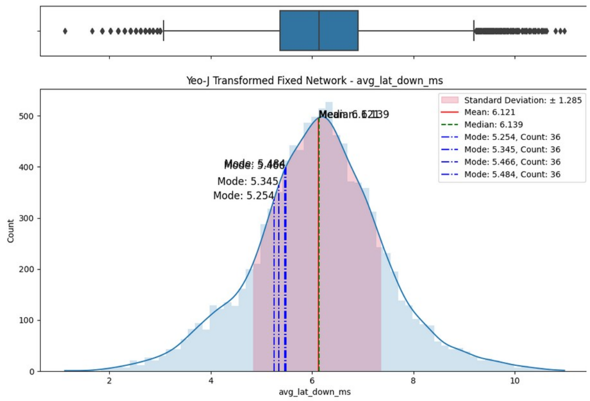

Figure 31. Plots of ’avg_lat_down_ms’ for Yeo-Johnson transformed fixed network

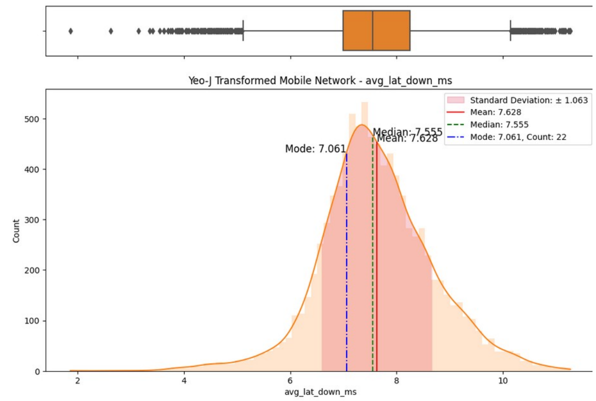

Figure 32. Plots of ’avg_lat_down_ms’ for Yeo-Johnson transformed mobile network

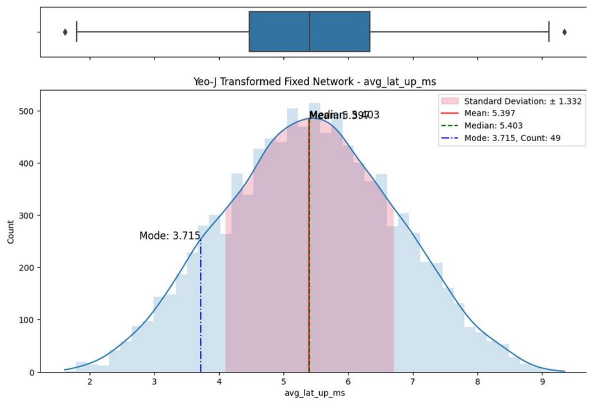

Figure 33. Plots of ’avg_lat_up_ms’ for Yeo-Johnson transformed fixed network

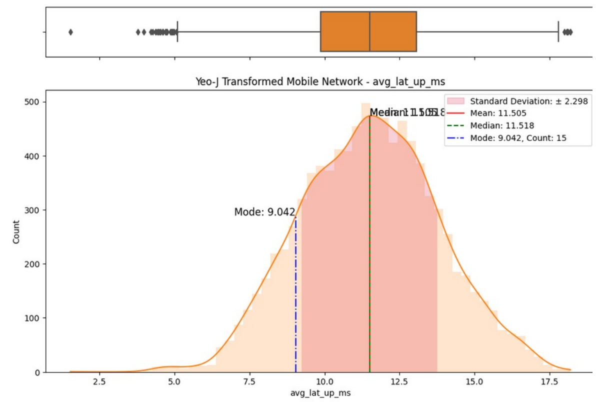

Figure 34. Plots of ’avg_lat_up_ms’ for Yeo-Johnson transformed mobile network

	Network Type	Org Skew	Sqrt Skew	Log Skew	Box-Cox Skew	Yeo-J Skew
avg_d_mbps	Fixed	1.37781	0.37467	-0.71262	-0.09987	-0.09376
avg_d_mbps	Mobile	4.10912	1.59613	-0.13475	0.01141	-0.00401
	Network Type	Org Skew	Sqrt Skew	Log Skew	Box-Cox Skew	Yeo-J Skew
avg_u_mbps	Fixed	3.07749	1.40362	-0.03149	-0.00737	-0.00196
avg_u_mbps	Mobile	2.64466	0.80352	-0.19049	-0.03915	-0.01876
	Network Type	Org Skew	Sqrt Skew	Log Skew	Box-Cox Skew	Yeo-J Skew
avg_lat_ms	Fixed	10.00117	4.97773	1.10378	-0.06684	-0.0607
avg_lat_ms	Mobile	10.49187	4.0093	1.26154	-0.26221	-0.20011
	Network Type	Org Skew	Sqrt Skew	Log Skew	Box-Cox Skew	Yeo-J Skew
avg_lat_down_ms	Fixed	5.22333	2.0235	-0.15539	0.0104	0.00794
avg_lat_down_ms	Mobile	3.21657	1.47787	-0.15087	0.01347	0.01223
	Network Type	Org Skew	Sqrt Skew	Log Skew	Box-Cox Skew	Yeo-J Skew
avg_lat_up_ms	Fixed	3.75387	1.61606	-0.01273	-0.00215	-0.00072
avg_lat_up_ms	Mobile	1.98862	0.8342	-0.364	-0.00718	-0.00749

Table 3. Comparison of original skew with data transformation skews on both networks

2.5. Correlation Analysis

In addition to distribution improvements, we investigated the impact of transformations on correlation structures within the data. Scatter plots (Figs. 35 - 39) and correlation matrices (Figs. 40 - 41) were employed to evaluate changes in relationships between variables. This step aimed to ensure that the transformations not only enhanced distributions but also preserved or revealed meaningful associations.

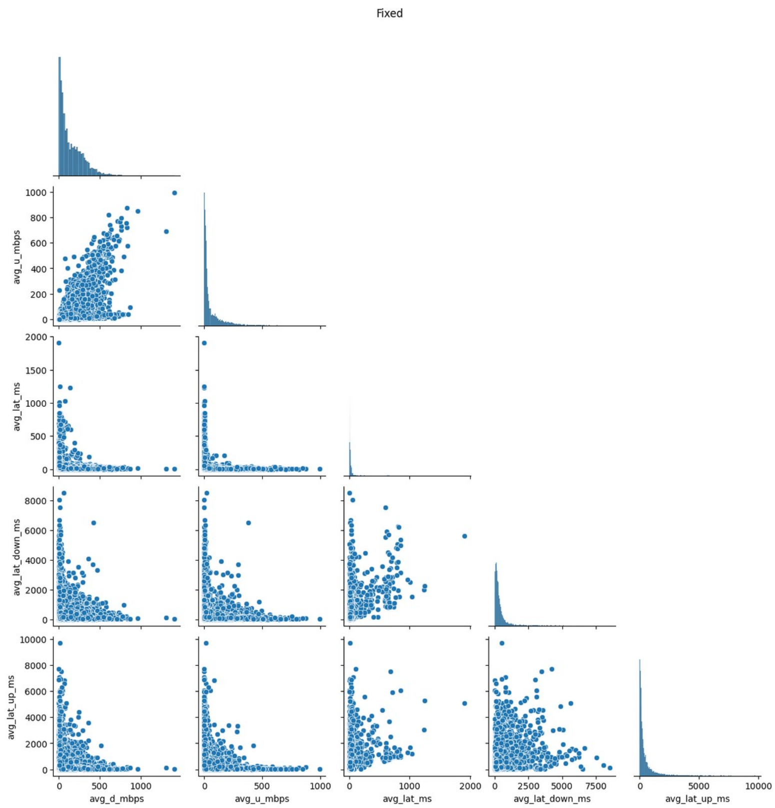

Figure 35. Pairplot of original fixed network

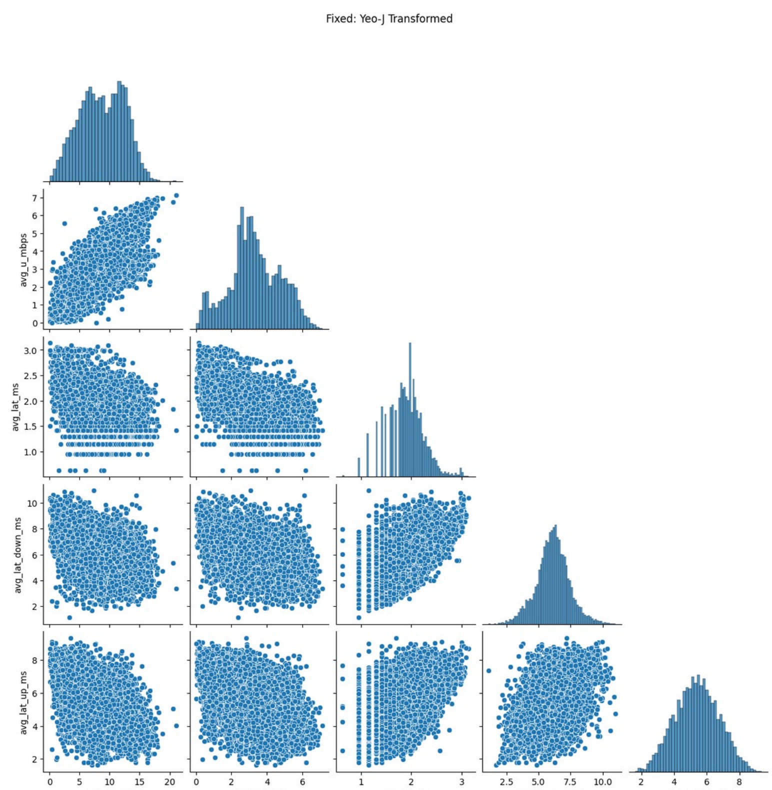

Figure 36. Pairplot of Yeo-Johnson transformed fixed network

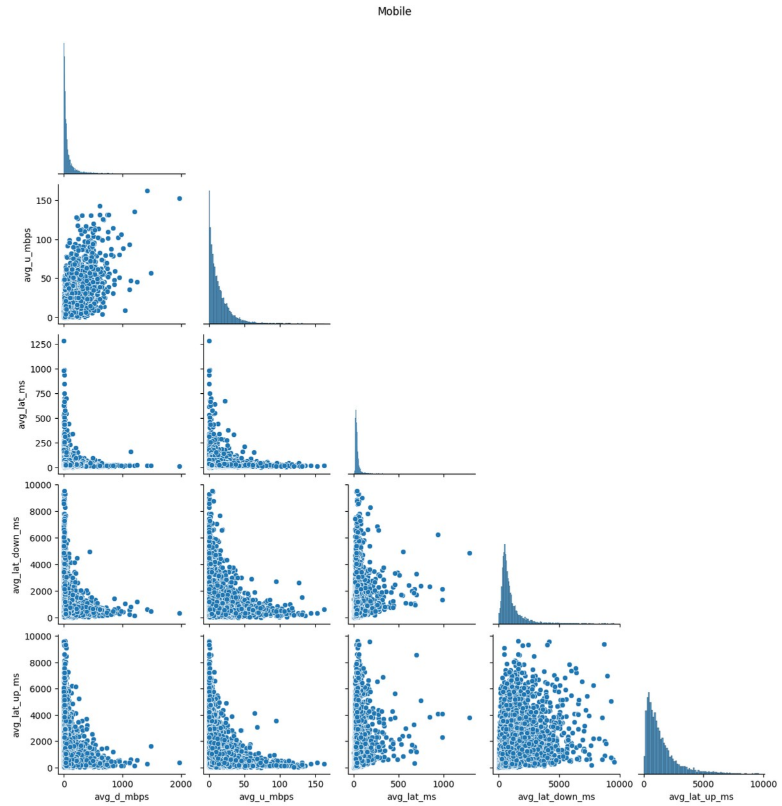

Figure 37. Pairplot of original mobile network

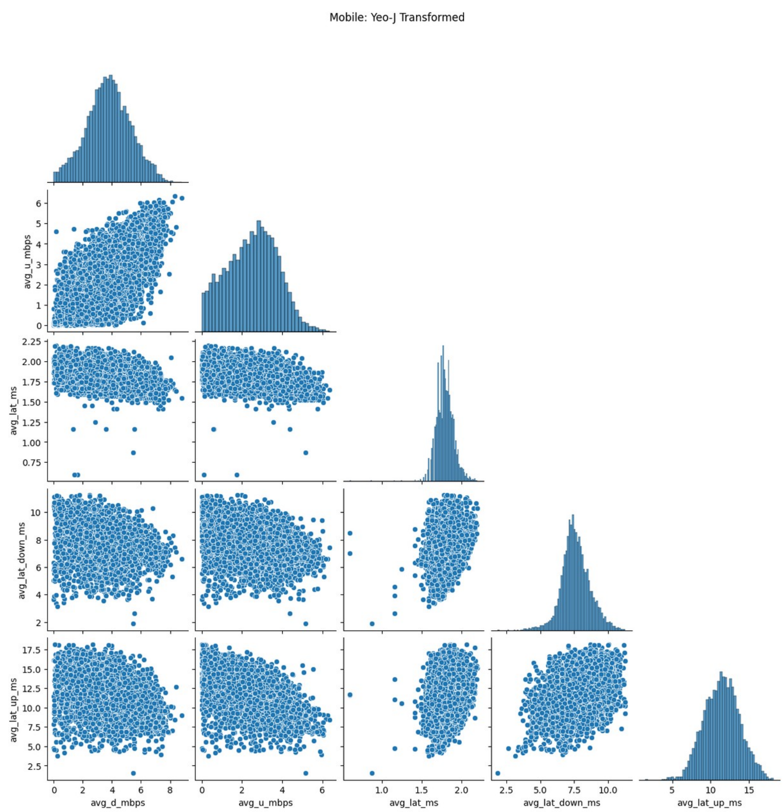

Figure 38. Pairplot of Yeo-Johnson transformed mobile network

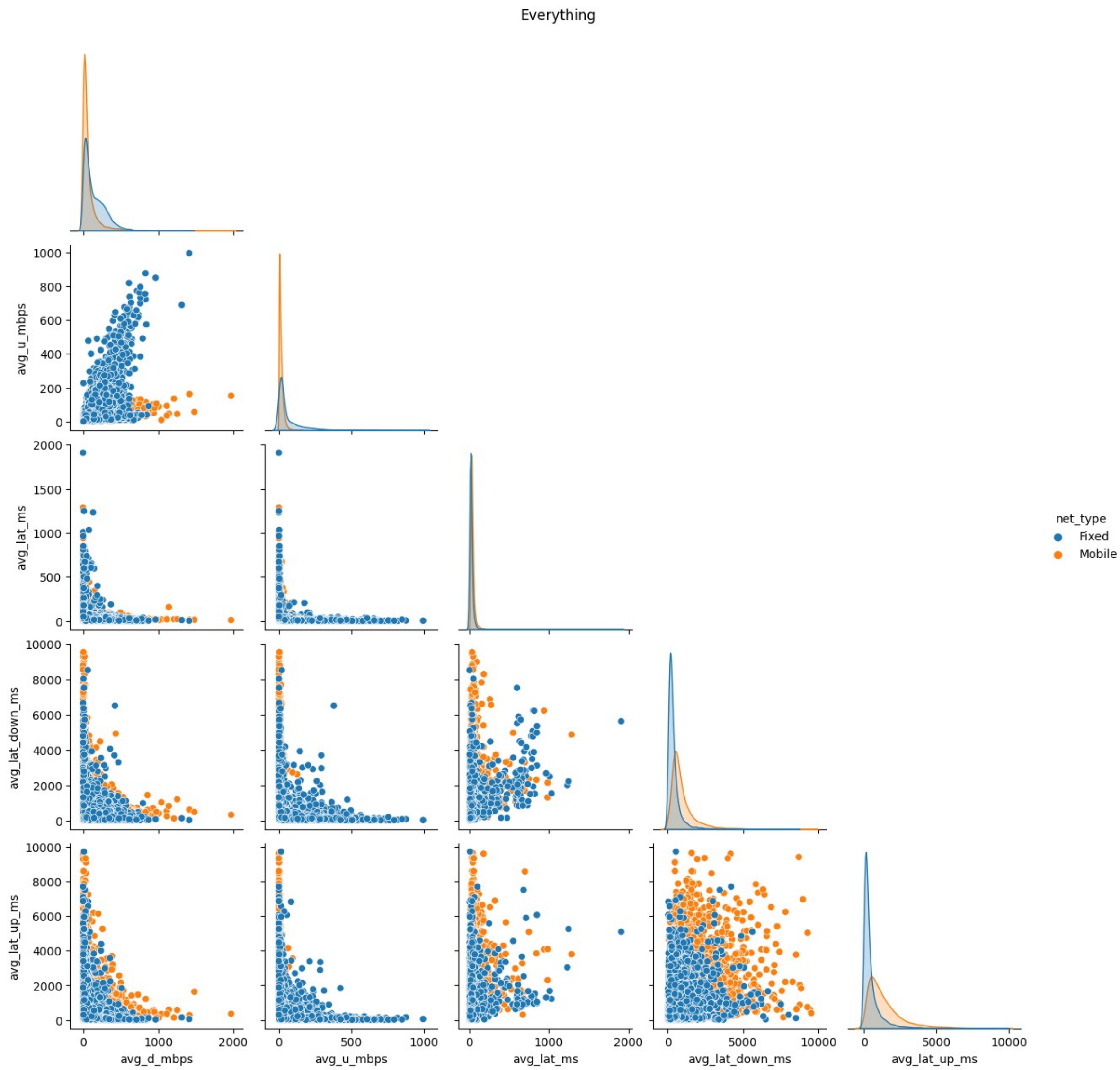

Figure 39. Pairplot of original with both networks

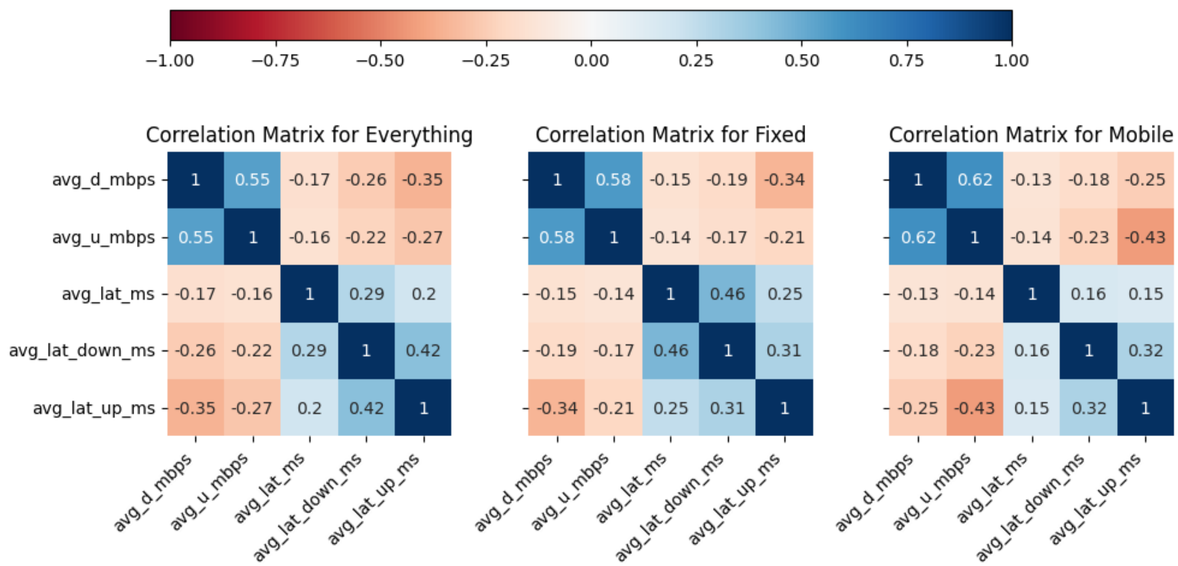

Figure 40. Correlation heatmap matrices for everything and both networks

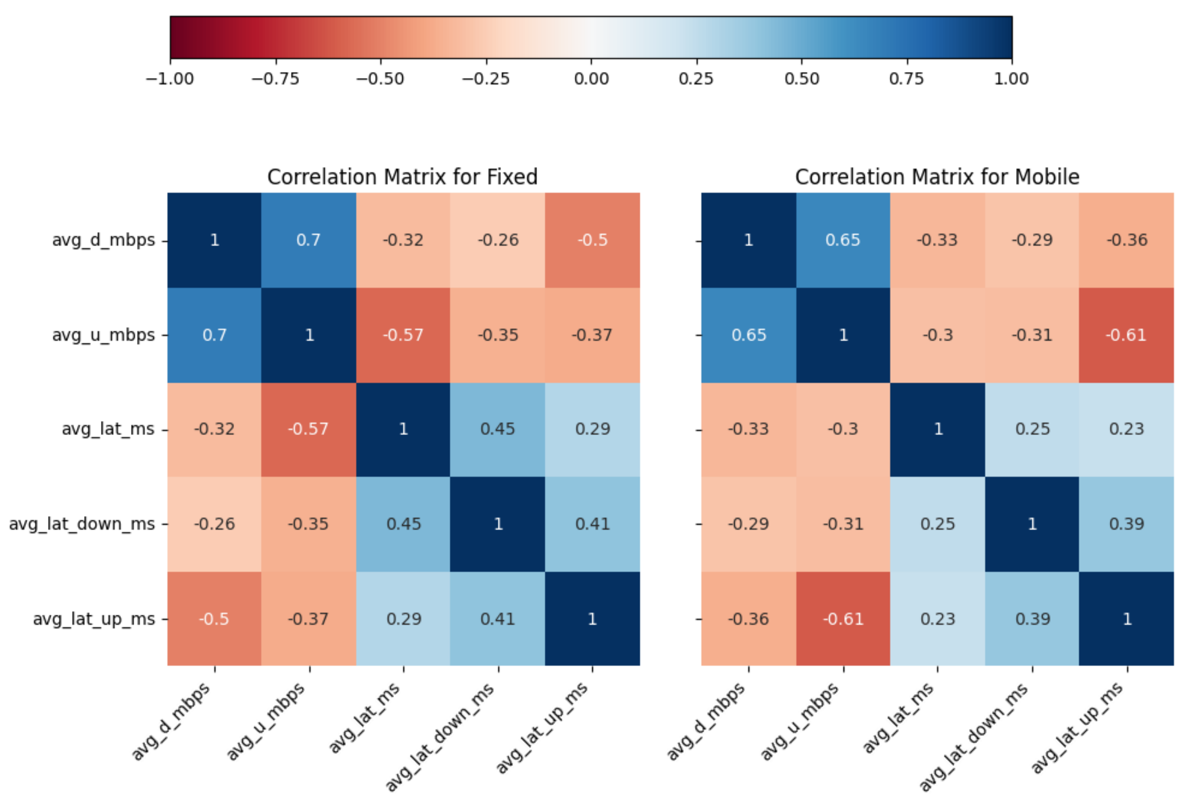

Figure 41. Correlation heatmap matrices for both networks after Yeo-Johnson transformation

2.6. Conclusion

The described EDA and distribution transformations constitute a critical phase in preparing the data for hypothesis testing. The chosen transformations were justified through a systematic exploration of initial distributions, comparative analyses, and a thorough assessment of the impact on correlations. The Yeo-Johnson transformation demonstrated a remarkable ability to normalize skewed data, effectively mitigating the positive skewness observed in the initial distributions. This methodical approach ensures that subsequent analyses are conducted on data that aligns more closely with parametric assumptions, enhancing the robustness and reliability of the findings.

3. Hypothesis Definition and Testing

This section explores the variability and average download speed differences between fixed and mobile networks. Our goal is to determine if the standard deviation of 'avg_d_mbps' varies significantly between the networks, providing insights into their consistency, and to establish whether one network has significantly higher average download speeds.

3.1. Methodology

We employed a comprehensive set of statistical tests, considering the positively skewed nature of the original 'avg_d_mbps' dataset.

3.1.1. Levene's Test: Untransformed Data

Levene's test was conducted on the untransformed 'avg_d_mbps' data to assess whether the standard deviation of download speeds differs significantly between fixed and mobile networks.

- *Decision Justification*: Levene's test is robust for assessing equality of variances, and its non-parametric nature aligns well with the skewed distribution of the original data. (Yuhang Zhou, Yiyang Zhu and Weng Kee Wong, 2023) (Hosken, D. J., Buss, D. L. and Hodgson, D. J., 2018)

3.1.2. F-Test: Yeo-Johnson Transformed Data

An F-test was performed on Yeo-Johnson transformed data to compare variances between fixed and mobile networks after addressing the skewness.

- *Decision Justification*: F-test is suitable for comparing variances, and using the transformed data allows us to make robust comparisons while accounting for skewness.

3.1.3. T-Tests: Untransformed and Yeo-Johnson Transformed Data

Independent sample t-tests were conducted on both untransformed and transformed 'avg_d_mbps' data to assess whether one network has significantly higher average download speeds than the other.

- *Decision Justification*: T-tests are appropriate for comparing means, and conducting them on both datasets ensures a comprehensive evaluation of average download speeds.

3.1.4. Mann-Whitney U Test: Untransformed Data

A non-parametric Mann-Whitney U test was performed on the untransformed data to corroborate findings from the t-tests and provide additional robustness.

- *Decision Justification*: The non-parametric nature of the Mann-Whitney U test suits skewed data, offering an alternative perspective on average download speed differences. (Mori, M. *et al.,* 2024) (María Teresa Politi, Juliana Carvalho Ferreira and Cecilia María Patino, 2021)

3.2. Results and Interpretation

3.2.1. Levene's Test: Untransformed Data

- **Statistic:** 1046.03, **p-value:** 0.0
- *Conclusion*: The standard deviation of 'avg_d_mbps' significantly differs between fixed and mobile networks.

3.2.2. F-Test: Yeo-Johnson Transformed Data

- **F statistic:** 6.07, **p-value:** 0.0
- *Conclusion*: The F-test on transformed data reinforces the conclusion that the standard deviation of 'avg_d_mbps' varies significantly

between networks. Also, it indicates that the fixed network has significantly higher average download speeds and a higher standard deviation than the mobile network.

3.2.3. T-Tests: Untransformed and Transformed Data

- *Untransformed Data*:
  - **t statistic:** 40.16, **p-value:** 0.0
  - *Conclusion*: The fixed network has significantly higher average download speeds than the mobile network, and it also exhibits a higher standard deviation.

- *Yeo-Johnson Transformed Data*:
  - **t statistic:** 120.57, **p-value:** 0.0
  - *Conclusion*: The transformed data supports the initial conclusion of the fixed network outperforming the mobile network in both average download speeds and standard deviation.

3.2.4. Mann-Whitney U Test: Untransformed Data

- **U statistic:** 63199341.5, **p-value:** 0.0
- *Conclusion*: The Mann-Whitney U test aligns with t-test results, indicating that the fixed network tends to have significantly higher average download speeds and a higher standard deviation.

3.3. Summary

Our multifaceted analysis, incorporating Levene's test, F-test, t-tests on both original and transformed data, and the Mann-Whitney U test, consistently suggests that the fixed network exhibits significantly higher average download speeds compared to the mobile network. However, it's important to note that this superior performance is accompanied by a higher standard deviation, indicating a greater degree of variability in download speeds. While the fixed network showcases higher speeds on average, the increased standard deviation suggests a higher level of variability, implying that the consistency of download speeds in the fixed network may be more variable than that of the mobile network. This thorough approach provides a nuanced understanding of the network performance, acknowledging the strengths and potential areas of variability.

4. Implementation

4.1. Regression Models for Average Download Speed

4.1.1. Linear Regression

Uni-variate and Multivariate linear regression models were employed to predict average download speed (avg_d_mbps). The initial models were trained on the original data, and the others were trained on Yeo-Johnson transformed data. The Yeo-Johnson transformed data exhibited a marginal improvement in performance, suggesting that addressing skewness contributed to better predictions (Pan, P., Li, R. and Zhang, Y., 2023). The mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) were used to evaluate model performance. (See Tables 4 & 5) (Subasi, A. *et al.,* 2020)

	Original data	Transformed data
Mean Abs Error	88.74945022250473	83.58067409589175
Mean Sq2 Error	12988.151119107111	12465.236420152909
Root Mean Sq2 Error	113.96556988453624	111.64782317695634
R2	0.3061458386523561	0.3340810340910251

Table 4. Comparison of uni-variate linear regression models trained on original and transformed data

	Original data	Transformed data
Mean Abs Error	84.07409357832987	78.7982326741036
Mean Sq2 Error	11856.27551016041	10688.05900893568
Root Mean Sq2 Error	108.88652584300966	103.38306925670024
R2	0.3666129978494965	0.42902156341794895

Table 5. Comparison of multivariate linear regression models trained on original and transformed data

4.1.2. Gradient Boosting Regression

A multivariate Gradient Boosting Regressor was employed as a more sophisticated regression model (Subasi, A. *et al.,* 2020). The model was trained on the original data, and its performance was evaluated using the same metrics. The Gradient Boosting model outperformed the linear regression models, achieving an R2 of 0.54. Gradient Boosting Regression demonstrated superior predictive power compared to linear regression.

	Original data
Mean Abs Error	66.39803957696711
Mean Sq2 Error	8591.94700117001
R2	0.5410002450564297
Root Mean Sq2 Error	92.69275592607013

Table 6. Gradient boosting results

4.2. Classification Models for Network Type

4.2.1. Support Vector Machine (SVM)

An SVM classification model was trained on original, and Yeo-Johnson transformed data to predict the network type (Fixed or Mobile). Again, the transformed data trained model performed better than the other, achieving an accuracy of approximately 87%. The confusion matrix (Figs. 42 & 43) and classification report provided insights into precision, recall, and F1-score for each class.

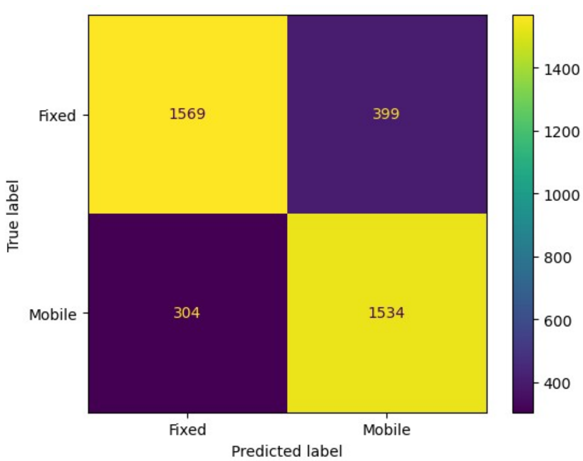

Figure 42. SVM original data confusion matrix

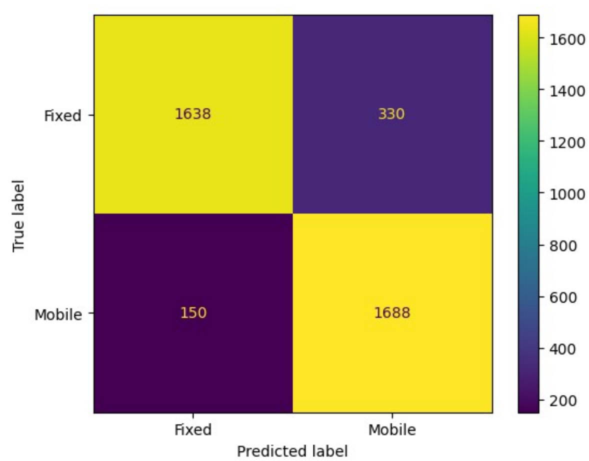

Figure 43. SVM transformed data confusion matrix

	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Fixed	0.84	0.80	0.82	0.92	0.83	0.87
Mobile	0.79	0.83	0.81	0.84	0.92	0.88

4.2.2. Random Forest Classifier

A Random Forest Classifier was also employed for classification, achieving an accuracy of approximately 87%. (Figs. 44 & 45) A grid search was conducted to fine-tune hyperparameters, resulting in optimal values for max_depth, max_leaf_nodes, min_samples_leaf, and min_samples_split. (Behera, G. and Nain, N., 2022)

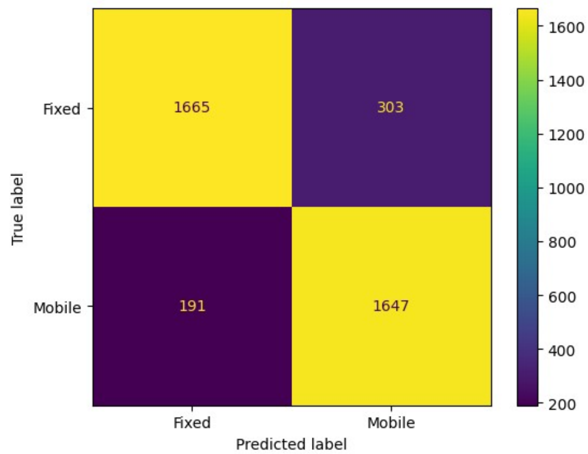

Figure 44. Random forest confusion matrix

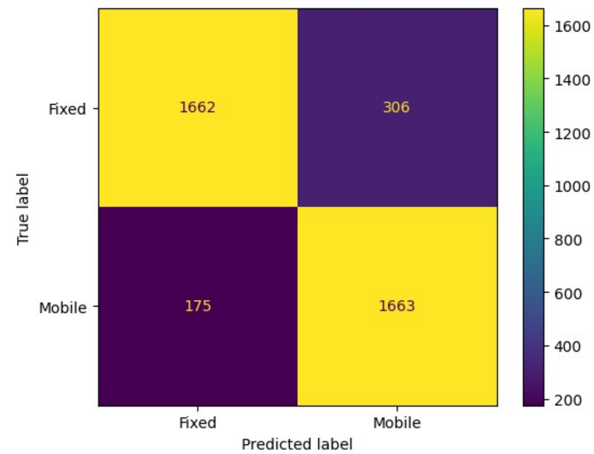

Figure 45. Random forest confusion matrix trained with more optimal parameters

	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Fixed	0.90	0.85	0.87	0.90	0.84	0.87
Mobile	0.84	0.90	0.87	0.84	0.90	0.87

4.3. Model Comparison and Analysis

The choice of models depended on the nature of the prediction task. Gradient Boosting Regression demonstrated superior performance in predicting average download speed, while Random Forest Classification excelled in predicting network types. The decision to employ Yeo-Johnson transformation in regression was justified by the slight improvement in predictive accuracy (Pan, P., Li, R. and Zhang, Y., 2023). Both SVM and Random Forest Classifier provided competitive results for network classification, with the latter outperforming SVM.

5. References

Behera, G. and Nain, N. (2022) “Gso-Crs: Grid Search Optimization for Collaborative Recommendation System,” *Sādhanā : Published by the Indian Academy of Sciences*, 47(3). doi: 10.1007/s12046-022-01924-0.

Hosken, D. J., Buss, D. L. and Hodgson, D. J. (2018) ‘Beware the F Test (or, How to Compare Variances)’, *Animal behaviour*, 136, pp. 119–126.

Lee, S. X. and McLachlan, G. J. (2022) ‘An Overview of Skew Distributions in Model-Based Clustering’, *Journal of Multivariate Analysis*, 188. doi: 10.1016/j.jmva.2021.104853.

María Teresa Politi, Juliana Carvalho Ferreira and Cecilia María Patino (2021) *Nonparametric Statistical Tests: Friend or Foe?*, 47(4). doi: 10.36416/1806-3756/e20210292.

Mori, M. *et al.* (2024) “An Analytical Investigation of Body Parts More Susceptible to Aging and Composition Changes Using Statistical Hypothesis Testing,” *Healthcare Analytics*, 5. doi: 10.1016/j.health.2023.100284.

Pan, P., Li, R. and Zhang, Y. (2023) “Predicting Punching Shear in Rc Interior Flat Slabs with Steel and Frp Reinforcements Using Box-Cox and Yeo-Johnson Transformations,” *Case Studies in Construction Materials*, 19. doi: 10.1016/j.cscm.2023.e02409.

Subasi, A. *et al.* (2020) “Permeability Prediction of Petroleum Reservoirs Using Stochastic Gradient Boosting Regression,” *Journal of Ambient Intelligence and Humanized Computing*, 13(7), pp. 3555–3564. doi: 10.1007/s12652-020-01986-0.

West, R. M. (2022) “Best Practice in Statistics: The Use of Log Transformation,” *Annals of Clinical Biochemistry*, 59(3), pp. 162–165. doi: 10.1177/00045632211050531.

Yuhang Zhou, Yiyang Zhu and Weng Kee Wong (2023) ‘Statistical Tests for Homogeneity of Variance for Clinical Trials and Recommendations’, *Contemporary Clinical Trials Communications*, 33, p. 101119. doi: 10.1016/j.conctc.2023.101119.