ELECTRONIC ASSIGNMENT COVERSHEET

Student Name and Number:	Yeo Xiu Juan (34678307) Gabriel Ng (34400923)

Unit Code:	ICT513
Unit name:	Data Analytics
Date:	16th July 2023
Assignment name:	Project Report
Tutor:	Dr Goh Zhang Hao

Table of Contents

Abstract ***3***

1.Introduction ***3***

2.Methodology ***3***

2.1 Data Analysis **3**

3.Results **4**

3.1 Descriptive Statistics **4**

3.1.1 Distribution and variability of response variable (Pre-Surgery Time, Post Surgery Time and Total Treatment Time) 4

3.1.2 Demographic X Pre-Surgery Time 4

3.1.3 Surgical corrections & Hospital Type X Pre-Surgery Time 5

3.1.4 Diagnostics plot for all the predictors. 6

3.2 Result of Application of Statistical Techniques **6**

3.2.1 Statistical Result for R1 6

3.2.2 Set of Predictors evaluated by PCA. 6

3.2.3 Collinearity Test 6

3.2.4 Multiple Linear Regression for R1 - Demographic Profile, Surgical Correction, Facial Type & Hospital Type 6

3.2.4.1 Multiple Linear Regression Assumptions 7

3.2.5 Compare the characteristic of individual profile in relation to treatment time 8

3.2.6 Predict treatment time in relation to the characteristic of individual's dental profile 8

4. Conclusion ***9***

5.Limitations & Recommendations for Future Studies ***10***

6. References ***10***

Influence of Individual's Characteristics on Dental Pre-surgery and Treatment Time

Abstract

This report provides an overview of an extensive analysis aiming to examine the factors that associated to dental pre-surgery time and compare the characteristics of the individuals in relation to short or long treatment duration (ie. 2 years) The report found Individuals who visited public hospital and had one tooth extraction positively related to pre-surgery time. On the other hand, exceeding 2 years treatment time include the patient's younger age (15-20 years), receiving treatment at a public hospital, and undergoing additional medical procedures. Lastly, individuals who had do not extract a tooth is likely to experience shorter estimated treatment duration 22 days out of 81 days, on the other hand, individuals who perform one tooth extraction and Skeletal AP (I) is likely to experience shorter estimated treatment duration 12 days out of 22 days while Skeletal AP (III) is likely to experience slightly longer estimated treatment duration of 15 days out of 4 days.

***Keywords**: Principal Component Analysis, Multiple Linear Regression, Decision Tree, Visualization*

1.Introduction

From a medical perspective, it is valuable to understand the factors that associated with treatment time, such as the severity of underbite and overbite, the specific type of surgery required, and the financial resources available to patients. It is important to determine whether shorter treatment times are associated with a patient's financial means, as this can affect the overall treatment duration inaccurately. Furthermore, this analysis offers the advantage of enabling doctors to identify and proactively plan for future patients who have similar complications. Therefore, our overall research question is: 1) Which dental profile of individuals is associated with pre-surgery time? 2) Characteristics of individuals who had to go through a longer treatment time in contrast to those with a treatment time of less than two years. 3) Predict short or long treatment duration (ie. 2 years) according to the characteristics of the individual's dental profile.

Following this section, the report is organised as follows.The methodology used in Section 2,Descriptive statistics and the results in section 3,conclusion in section 4, followed by limitations and further direction in section 5.

2.Methodology

The report was performed using RStudio.

2.1 Data Analysis

Exploratory Data Analysis

Histogram (Section 3.1.1) was performed to visualize the variability and the shape of the response variable (pre-surgery time, Post Surgery Time, and Treatment Time), while Boxplot (Section 3.1.2) was performed to evaluate the variability of the various predictors' variables against the selected response variable (Pre-Surgery Time). Mosaic plot (Hehman & Xie, 2021) in (section 3.2.4.1) was performed to visualize the characteristics of individuals who experience short or long treatment time (ie.2 years) .

Principal Component Pre-Processing

The graphical display in (Figure2a-g) shows multiple outliers against the response variable (Pre-Surgery Time). In addition, dental dataset consists of *N* =195 failed to meet the rule of thumb of sample size 20:1 suggest by (Burmeister & Aitken, 2012) in a regression model. Applying apply 20:1 rule , sex (2 categories), Skeletal Type (3 categories), Vertical (3 categories), Transverse (2), Max(3), Man(3), Single BiMax(2), Exo(2),other(2),Hospital (2) : $\eta = (2+3+3+2+3+3+2+2+2+2)-1 * 20 =$ **460** participants required for the study. Furthermore, in figure (3a), Normal QQ plot violates the assumption of

normality. Models with more variables are likely to "overfit" the regression model. Hence, data pre-processing, the principal component is introduced to mitigate overfitting and improve the variability of a regression model.

Multiple Linear Regression

Parametric Multiple Linear Regression was assessed in section (3.2.4) to explore the set of predictor variables evaluated by PCA to assesses the significant association to the response variable (Pre-Surgery Time).

Decision Tree

Supervised Learning, Decision Tree resistant to outliers was performed in section (3.2.6) to predict qualitative variables (short or long treatment duration (ie. 2 years)) across the qualitative variables of individual characteristics.

3.Results

In this section, we will review the descriptive statistics and results of our research questions.

3.1 Descriptive Statistics

In this section, we will explore the distribution of response variable (Pre-Surgery Time, Post Surgery Time and Total Treatment Time). Furthermore, we will review the relationships between predictors variable i) Demographic (Sex, Age group); ii) surgical corrections (Skeletal AP, Vertical, Transverse, Maxilla, Mandible, Single_Bi Max, Exo) against response variable Pre-Surgery Time.

3.1.1 Distribution and variability of response variable (Pre-Surgery Time, Post Surgery Time and Total Treatment Time)

Pre-Surgery Time displays a positive right skew distribution while Post-Surgery Time appears to be narrower with no symmetrical shape. Treatment Time, on the other hand, appears to have a slightly symmetrical shape (see Figure 1a-c). Pre-Surgery Time performed better than the two response variables. and as a result, it will be assessed in the following section to examine the relationships between the various predicators' variables. Log transformation to pre-surgery time is also performed to address the skewness (Max Kuhn & Johnson, 2013) (See Figure 1d) .

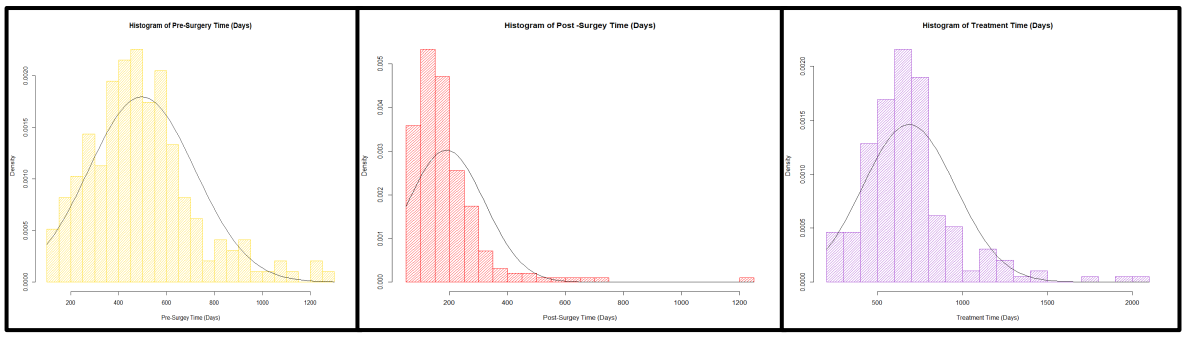


Figure 1 (a) Histogram of Pre-Surgery Time; (b) Histogram of Post-Surgery Time; (c) Histogram of Treatment Time

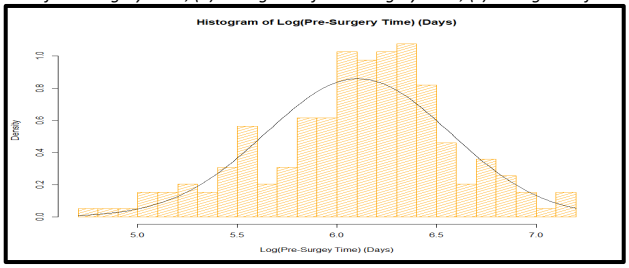


Figure 1 (d) Histogram of Log (Pre-Surgery Time)

3.1.2 Demographic X Pre-Surgery Time

Figure 2 (a) shows a comparison of male and female pre-surgery time distribution, with male having median value of 517 days and female having a median of 467 days. Females have a positively right skewed distribution, while male have a negatively left skewed distribution. Outliers are represented by bubble shape. Figure 2 (b) shows differences in pre-surgery time distributions between age groups. Children

have a higher median value of 587 days, while middle-aged and young adults have lower medians of 439 and 482 days, respectively. Young and middle-aged adults have a positively right-skewed distribution, while children have a negatively left-skewed distribution.

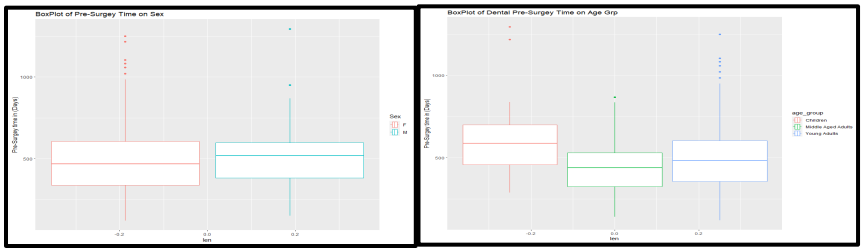


Figure 2: (a) Boxplot of Dental Pre-Surgery Time on Sex; (b) Boxplot of Dental Pre-Surgery Time on AgeGrp

3.1.3 Surgical corrections & Hospital Type X Pre-Surgery Time

Figure (2c) provides comparisons of the distributions of differences between (Skeletal AP) on Pre-Surgery Time. It appears that (C1 II & III) has a roughly higher median value of 462 and 498 days while (C1 I) has a lower median value of 378 days. (C1 I) appears to be more symmetrical, (C1 II & III) display a positively right-skewed distribution. Figure (2d) provides comparisons of the distributions of differences between (Vertical) on Pre-Surgery Time. It appears that (normo) has a higher median value of 476 days than (hyper and hypo) has a lower median value of 455 and 472 days. (hyper and normo) displays a positively right skewed where the median is closer to the bottom of the box and the whiskers is shorter on the lower end of the box whereas (hypo) displays a negatively left-skewed distribution.

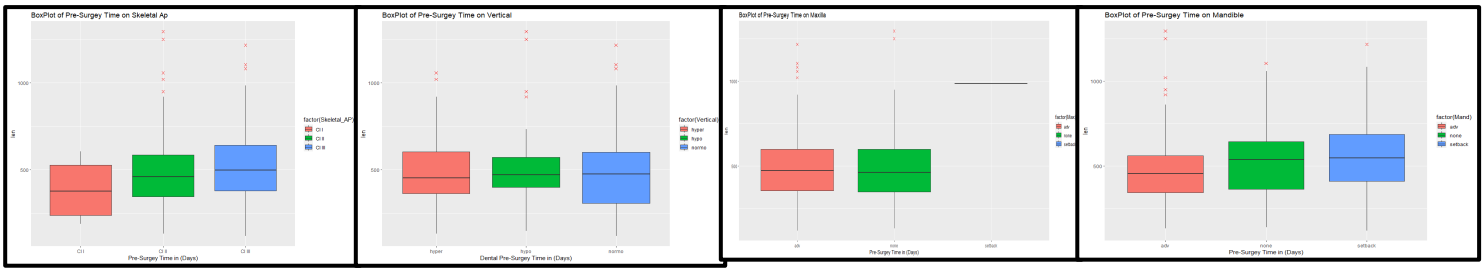


Figure 2: (c) Boxplot of Pre-Surgery Time on Skeletal AP; (d) Boxplot of Pre-Surgery Time time on Vertical ,(e) Boxplot of Pre-Surgery Time on Maxilla, (f) Boxplot of Pre-Surgery Time on Mandible

Figure (2e) provides comparisons of the distributions of differences between (Maxilla) on Pre-Surgery Time. It appears that (Setback) has a higher median value of 985 days than (adv and none) has a lower median value of 475 and 464 days. (adv & none) displays a positively right skewed whereas(setback) appears to have few individuals where we can't analyse the dispersion of the distribution. Figure (2f) provides comparisons of the distributions of differences between (Mandible) on dental treatment time. It appears that (Setback) took longer treatment time with a median value of 549 days than (adv and none) with a median value of 457 and 539 days.(Setback and none) displays a positively right skewed; (adv) displays a negatively left-skewed distribution with some outliers that were represented by the asterisks. Figure (2g) provides comparisons of the distributions of differences between (Single Bimax) on Pre-Surgery Time. It appears that (Single Jaw) took longer treatment time with a median value of 481.5 days than (double Jaw) with a median value of 464 days. (Double Jaw) displays a positively right skewed whereas (Single Jaw) appears to be more appears more symmetrical, with some outliers that were represented by asterisks. Figure (2h) provides comparisons of the distributions of differences between (Transverse) on pre-surgey time. It appears that (Narrow) took longer treatment time with a median value of 518 days than (Normal) 459 days. (Normal) displays a positive right skewed whereas (Narrow) displays a negatively left skewed values with some outliers that were represented by the asterisks. Figure (2I) provides comparisons of the distributions of differences between (Exo) on dental treatment time. It appears that Individuals who took Exo Treatment (yes) took longer treatment time with a median value of 585 days than Individuals who did not take Exo Treatment (no) 438 days. Individuals who took Exo Treatment (yes) display a positive right skew whereas Individuals who did not take Exo Treatment (no) appear to be more symmetrical, (yes and no) had some outliers that were represented by asterisks. Figure (2J) provides comparisons of the distributions of differences between (Hospital) on Pre-Surgery Time. It appears that individuals that visited private hospital has a shorter median of 339 days whereas individuals Visited public hospital longer median of 539 days.

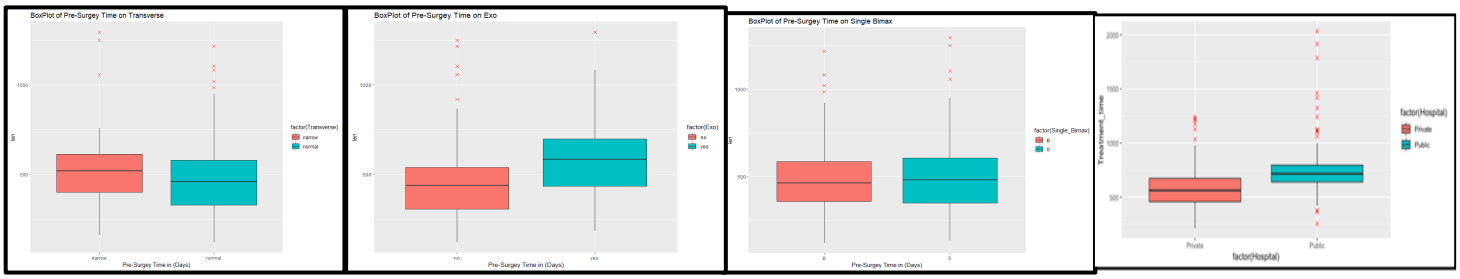


Figure 2: (g) Boxplot of Pre-Surgery Time on Single Bimax ; (h) Boxplot of Pre-Surgery Time on Transverse ; (I) Boxplot of Pre-Surgery Time on Exo (J) Boxplot of Pre-Surgery Time on Hospital.

3.1.4 Diagnostics plot for all the predictors.

Normal Q-Q plot, display that there are some deviations between empirical quantiles of the residuals and theoretical quantiles in the bottom left tail and top right tail of the distribution, these deviations suggest significant issues with the normality assumptions.

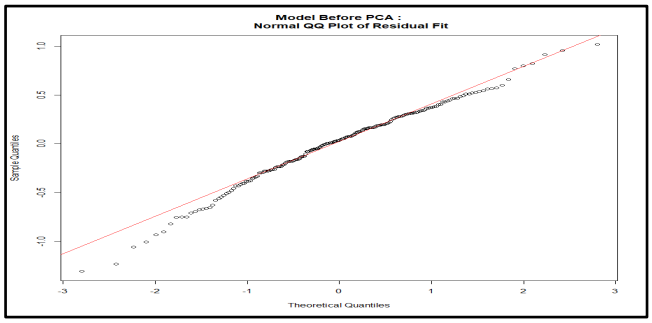


Figure 3 (a) : Diagnostic plot for all the predictors variables

3.2 Result of Application of Statistical Techniques

In this section, we will review the Statistical result for research questions 1 and 2.

3.2.1 Statistical Result for R1

In this section, we will review the PCA result for research question 1.

3.2.2 Set of Predictors evaluated by PCA.

The set of predictors evaluated by PCA from PC 1 and PC2 accordingly to the contribution percentage are (Single-BiMax,Max,Hospital,Age Start Treatment,Skeletal AP,Exo,Mand,Transverse,Vertical and Hospital) .Furthermore, ("others") variable has no correlation and ("sex") variable was not evaluated by PCA. Therefore, these variables will not be considered in regression model.

3.2.3 Collinearity Test

Tables (1) ,VIFs with values greater than 10 indicates violation of collinearity (James et al., 2021). Therefore, 3 variables are removed: [Mand , Max and Single Bi Max].

Variable	VIF
Age group	1.35
Vertical	1.55
Transverse	1.16
Max	60.63
Man	90.40
Single Bi Max	54.30
Exo	1.11
Skeletal AP	4.80
Hospital	1.42

Table 1: Collinearity Table for selected Predicators Variables

3.2.4 Multiple Linear Regression for R1 - Demographic Profile, Surgical Correction, Facial Type & Hospital Type

After elimination, the multiple linear regression model is left variables: age.grp, vertical, Transverse ,Exo ,Skeletal AP and Hospital.

Variables that have significant effect on log(pre-surgery time) are as follows:

- Reporting level of Exo (at least one tooth extracted) have an estimated of pre-surgery days of 0.299 (0.169,0.430) *longer* than those reporting at (no extraction of tooth level) of Exo after adjusting for age start treatment Vertical,Transverse,Exo ,Skeletal AP and the type of hospital visited by the individuals.
- Reporting level of Hospital (public) have an estimated of pre-surgery days of 0.236 (0.100,0.372) longer than those reporting at (Private) level of hospital after adjusting for treatment Vertical,Transverse,Exo ,Skeletal AP type of indivuduals.

This also suggests that there is no significant relationship between age start treatment group(Middle ,young and children),Vertical (hypo and normo), Transverse (normal), Skeletal AP (C1 II and III)) against pre-surgery time.

Multiple Linear Regression Equation

Log(pre-surgery time) = $\beta_{5.754} + \beta_{-0.134} * Middle\ Age\ Adults + \beta_{-0.179} * young\ Adults + \beta_{0.032} * Vertical(hypo) + \beta_{0.010} * vertical(normo) + \beta_{-0.022} * Transverse(normal) + \beta_{0.299} * (Exo)Yes + \beta_{0.240} * (Skeletal\ AP\ )C1\ II + \beta_{0.340} * (Skeletal\ AP\ )C1\ III + \beta_{0.236} * Hospital$(Public)

β	Variables	Coefficients	2.5%	97.5%	P-Value
$\boldsymbol{\beta_0}$	Intercept	5.754	5.249	6.30	0.000
$\boldsymbol{\beta_1}$	Factor (age_grp) Middle Aged Adults	-0.134	-0.368	-0.100	0.261
$\boldsymbol{\beta_2}$	Factor (age_grp) Young Adults	0.179	-0.390	0.031	0.094
$\boldsymbol{\beta_3}$	Factor (vertical) hypo	0.032	-0.129	0.194	0.693
$\boldsymbol{\beta_4}$	Factor (vertical) normo	0.010	-0.130	0.150	0.891
$\boldsymbol{\beta_5}$	Factor (Transverse)normal	-0.022	-0.170	-0.126	-0.769
$\boldsymbol{\beta_6}$	(Exo) Yes	0.299	0.169	0.430	0.000
$\boldsymbol{\beta_7}$	(Skeletal AP)C1 II	0.240	-0.193	0.674	0.276
$\boldsymbol{\beta_8}$	(Skeletal AP)C1 III	0.340	-0.097	0.776	0.126
$\boldsymbol{\beta_9}$	Factor (Hosptial) Public	0.236	0.100	0.372	0.001

Table 1 : Multiple Linear Regression Model Summary, R^2 = 21.3%

3.2.4.1 Multiple Linear Regression Assumptions

Scatterplot display the residuals against the fitted values of the observations seems to be scattered along the 0 points. We may conclude that it is unlikely indication of violation of homoscedasticity. Normal Q-Q plot, display that there are some deviations between empirical quantiles of the residuals and theoretical quantiles in the left tail of the distribution, these deviations are not large enough to suggest significant issues with the normality assumptions.

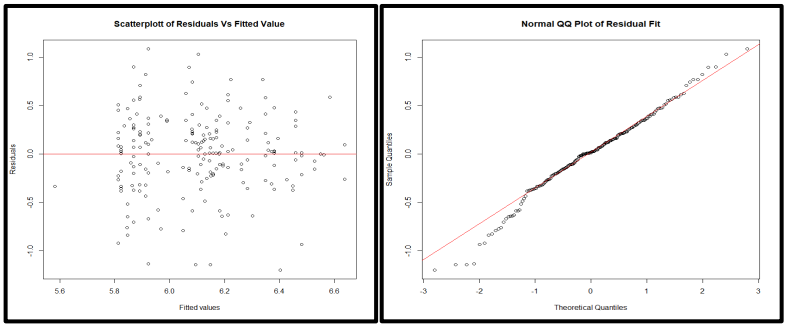


Figure 4(a) Diagnostics plot of Multiple Linear Regression Assumption

3.2.5 Compare the characteristic of individual profile in relation to treatment time.

To compare characteristics between patient groups with short and long treatment time, the factorized "Treatment time" variable was plotted against the rest of the variables. The plots (Figure 5) above presented is to indicate that there is a difference in proportion between the "short" and "long" groups and they may be factors in explaining the difference.

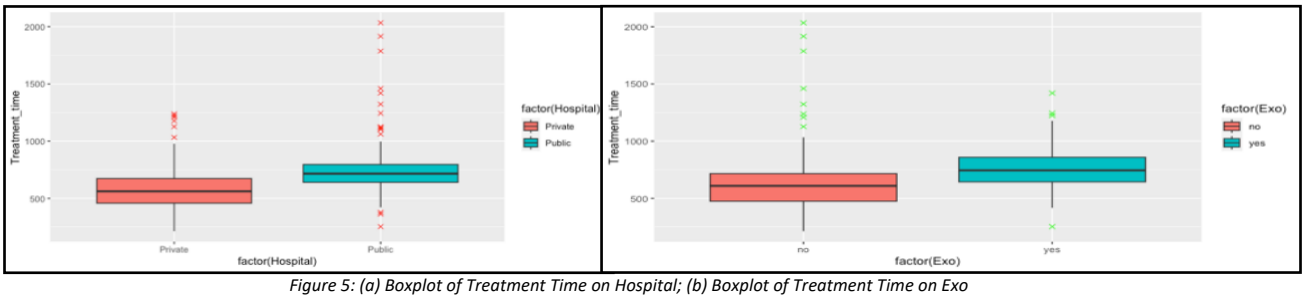


Figure 5: (a) Boxplot of Treatment Time on Hospital; (b) Boxplot of Treatment Time on Exo

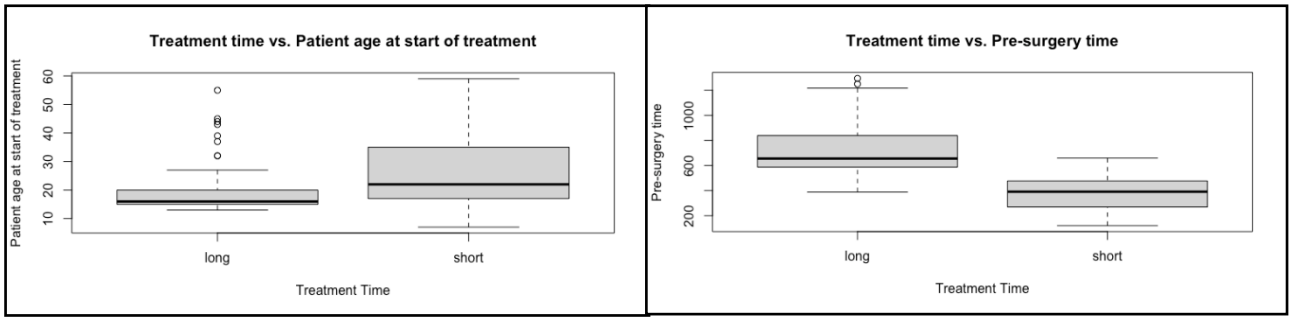


Figure 6: (a) BoxPlot of Patient age at the start of treatment time & (b) BoxPlot of Pre-survey time at the start of treatment time

The boxplot (Figure 6a) of treatment time groups is plotted against the patient age at the start of the treatment. The interquartile range of patient with long treatment time is around 15 to 20 years old, whereas short treatment time roughly between 15 to 35 years old. The median age for patients with long treatment time is at about 15 years old; comparing to patients with short treatment time is about 22 years old. The overall age range of patients with long treatment time is approximately 12-25 years old and those with short treatment time is between 0-60 years old. The boxplot (Figure 6b) of patients with long treatment time is about 600-800 days, while shorter treatment time is 300-500 days. The median is about 600 days of pre-surgery time for long treatment time: comparing to short treatment time of 400 pre-surgery days. As for the range, longer treatment time has a larger range, which is from 400-1200 days. Shorter treatment time is about 100-600 days. In the mosaic plot of treatment time plotted against the type of hospital (Figure 7a), about 70% of patients visited a public hospital has longer treatment, comparing to 40% of patients with shorter treatment. In the mosaic plot (Figure 7c), 50% of patients with long treatment time also have a tooth extraction performed and it is more than patients with short treatment time, with only about 22% of patients having a tooth extraction.

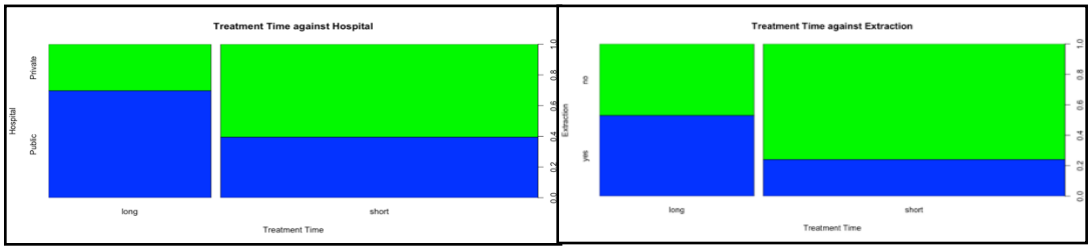
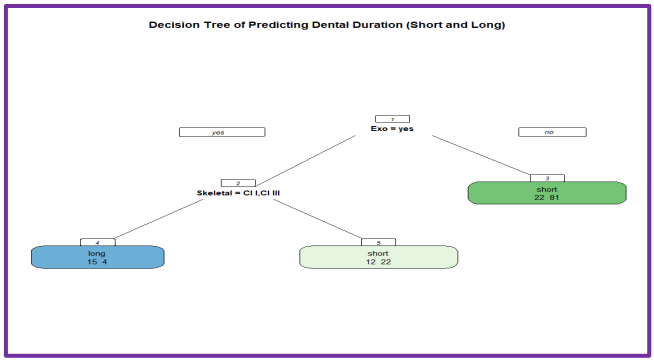


Figure 7: (a) Mosaic Plot of Treatment Time against Hospital; (b) Mosaic Plot of Treatment Time against Extraction

3.2.6 Predict treatment time in relation to the characteristic of individual's dental profile

The result of decision tree displays the prediction of treatment time (short or long) indicating the top internal node correspond to splitting Exo. The left-hand branch coming out from that node consists of observations with the first value of the Exo variable (yes) and the right-hand node consist of (no). Exo: yes, further split 1 node with short and long treatment time that consist of Skeletal Type variable. Hence, individuals who had do not extract a tooth is likely to experience shorter

treatment time probability (22 days out of the predicted 81 days) .Interestingly, individuals who perform one tooth extraction and Skeletal AP (I) is likely to experience even shorter treatment time (12 days out of the predicted 22 days) while Skeletal AP (III) is likely to experience longer (15 days out of predicted 4 days) treatment duration. Confusion Matrix suggest 60% accuracy of the model.



Prediction	Long	Short
Long	2	1
Short	15	21

Figure 8 (a) Decision Tree of predicting Treatment Time according to individual's profile, (b) Confusion Matrix of Decision Tree

4. Conclusion

In the initial phase of the analysis, a total of 13 variables were examined, consisting of 10 explanatory variables and 3 response variables, no missing values were found. The first step involved exploring on the distribution of response variable (Pre-Surgery Time, Post-Surgery Time, Total Surgery time) and log transformation pre-surgery was computed to address the skewness, and exploratory data analysis was conducted to assess relationship between pre surgery and the original set of predicators variables. Original set of predicators variables was fitted to regression model, which has failed to met the assumption of normality. Furthermore, *N* =195 failed to meet the rule of thumb of sample size 20:1.

Hence, data pre-processing, principal component was introduced to improve the variability of regression model. Multiple Linear Regression is conducted (8 explanatory variables and 1 response variable) was assessed to find the significance effect of pre-surgery time and has met regression assumptions. Regression Model result shows that Individuals who visited public hospital and had one tooth extraction positively related to pre-surgery time.

Analysing patient age, younger individuals in the 15–20 years old age bracket tend to experience lengthier treatment times. Moreover, additional medical procedures like tooth extraction or other surgeries appear to contribute to extended treatment durations. The type of hospital attended is also a crucial factor, with patients visiting public hospitals experiencing longer treatment times. In summary, the factors associated with treatment durations exceeding 2 years include the patient's younger age (15-20 years), receiving treatment at a public hospital, and undergoing additional medical procedures. These factors seem to contribute to longer treatment times in the study.

In terms of prediction of treatment time durations (ie.2 years), Decision Tree result displays individuals who had do not extract a tooth is likely to experience shorter estimated treatment duration 22 days out of 81 days, on the other hand, individuals who perform one tooth extraction and Skeletal AP (I) is likely to experience shorter estimated treatment duration 12 days out of 22 days while Skeletal AP (III) is likely to experience slightly longer estimated treatment duration of 15 days out of 4 days.

5.Limitations & Recommendations for Future Studies

Limitations:

- The current study might have a limited sample size, which could impact the generalizability of the findings. Future studies should aim to include larger and more diverse samples to enhance the robustness of the results.
- If the study was conducted from 2 hospitals, the results may not represent the broader population adequately. Future studies should consider multi-centre or population-based designs to increase the external validity of the findings.
- The current study might not have accounted for all potential confounding variables that could influence treatment duration. Future research should carefully identify and control for these factors to obtain more accurate results.

Recommendations:

- Incorporate patient perspectives, experiences, and dental hygiene in understanding the factors affecting longer dental treatment. Surveys, interviews, or focus groups can provide valuable insights into the patients' perceptions and challenges during the treatment process.
- Explore the impact of psychosocial factors, such as patient anxiety, coping mechanisms, and adherence to treatment plans, on the duration of dental treatment.
- Conduct follow-up studies on patients who underwent longer dental treatments to assess the long-term outcomes and identify potential complications or treatment-related issues.
- Perform Interaction Effect model on Hospital type and Exo to further explore relationship between treatment time.
- Perform Variable Selection Technique (ie., Subset Selection, Shrinkage) (James et al., 2021) to find important variables that are associated to the response variable.

Addressing these limitations and implementing the recommended approaches, future studies can provide a more comprehensive understanding of the factors influencing longer dental treatment and contribute to improving patient care and treatment efficiency.

6. References

Burmeister, E., & Aitken, L. M. (2012). Sample size: How many is enough? Australian Critical Care, 25(4), 271-274. https://doi.org/https://doi.org/10.1016/j.aucc.2012.07.002

Hehman, E., & Xie, S. Y. (2021). Doing Better Data Visualization. Advances in Methods and Practices in Psychological Science, 4(4), 25152459211045334. https://doi.org/10.1177/25152459211045334

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R. In. Springer US. https://doi.org/10.1007/978-1-0716-1418-1_12

Max Kuhn, & Johnson, K. (2013). Applied Predictive Modelling. https://doi.org/10.1007/978-1-4614-6849-3