*This is the main submission document**. Save and rename this document filename with your registered full name as Prefix before submission.***

Full Name	**ADAM SOH SHI JIE**
Email Address	**ADAM0025@e.ntu.edu.sg**

** : Delete and replace as appropriate.*

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square brackets below to indicate your selection.

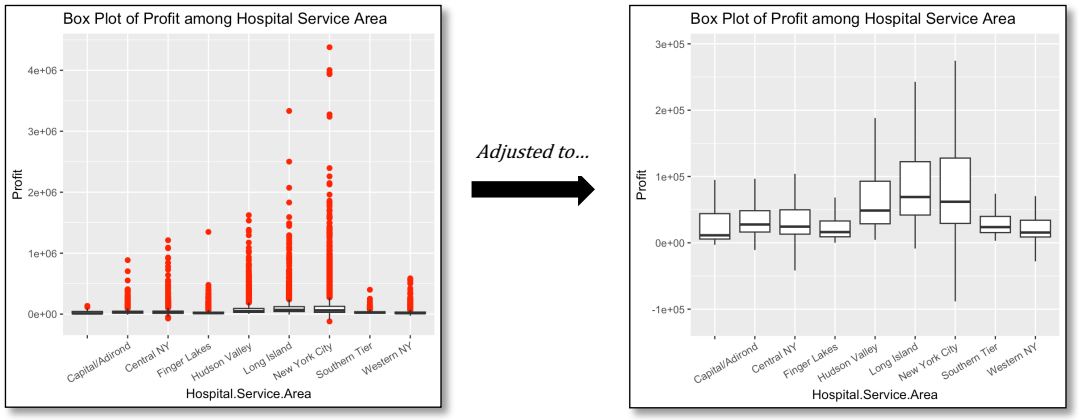
[X] I have read and accept the above.

Table of Contents

For each question, please start your answer in a new page.

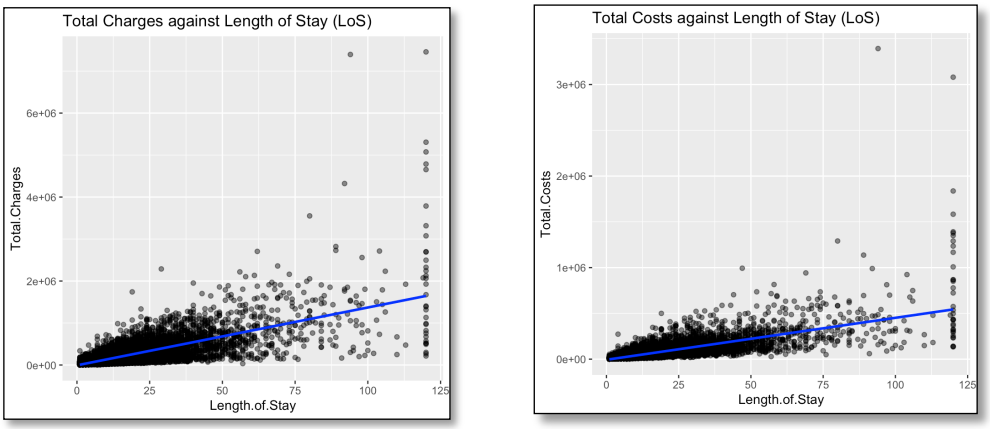
Answer to Q1:

1. Among all hospital service areas, **Hudson Valley**, **Long Island** and **New York City** generate *higher profits*, while **Capital/Adirondack** and **Southern Tier** show *lower profitability*.



I evaluate the distribution of Profits ($Profits = Total.Charges - Total.Costs$) across different Hospital Service Area using a boxplot. However, there are some significant outliers affect the clarity of the visualization. To observe the general case, I adjusted the boxplot to focus on the 'box-and-whisker' part to analyse the overall profits across different hospital service areas by hiding outliers. The finding suggests that Hudson Valley, Long Island and New York City being highly urbanised, tend to have higher profits, while Capital/Adirondack and Southern Tier, which are predominantly rural, show lower profit level.

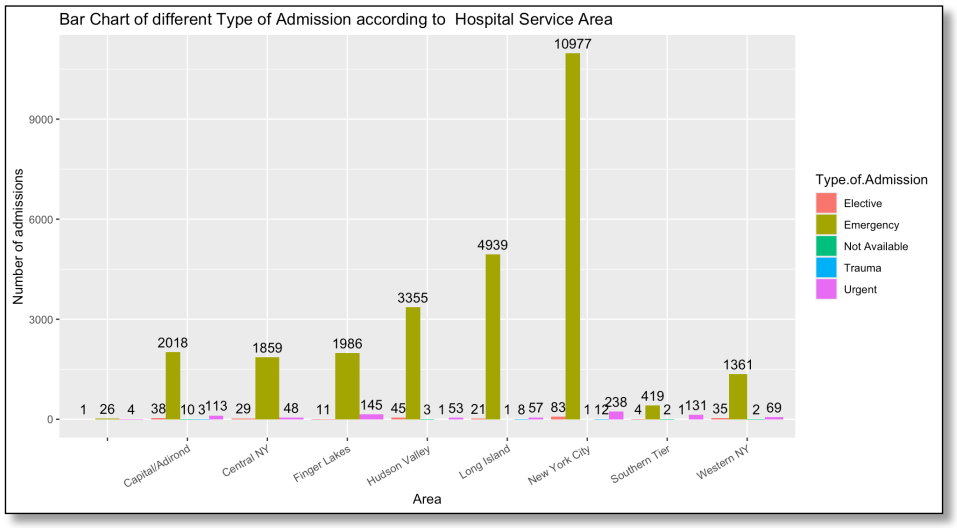
2. **Total.Charges** and **Total.Costs** have *strong positive correlations* against **Length.of.Stay (LoS)**



	Length.of.Stay	Total.Charges	Total.Costs
Length.of.Stay	1.0000000	0.7586303	0.7549333
Total.Charges	0.7586303	1.0000000	0.8846609
Total.Costs	0.7549333	0.8846609	1.0000000

As observed in the diagram above, I noticed a strong correlation between Total Charges and Length of Stay (LoS) with a correlation coefficient of 0.7586 and between Total Costs and Length of Stay (LoS) with a correlation coefficient of 0.7549.

3. The number of **Emergency** cases (Type.of.Admission) are the ***highest*** across all different Hospital Service Area.



The diagram clearly demonstrates that emergency cases make up the majority of admissions across all hospital service areas, regardless of location. This suggests that emergency care is a significant importance of hospital demand in every region, overshadowing other types of admissions such as elective or urgent cases. This trend may highlight the critical role of emergency services in each area, emphasizing the need for hospitals to allocate their resources, staff, and infrastructure to handle the high volume of emergency patients more efficiently.

Optional: I discovered 3 extra findings which are included in the R script. Feel free to run the script to find out more about it if interested!

Answer to Q2:

List of Potential Predictor X variables:

1. # Hospital.Service.Area
2. # Age.Group
3. # Gender
4. # Race
5. # Ethnicity
6. # Length.of.Stay
7. # Type.of.Admission
8. # APR.Severity.of.Illness.Description
9. # APR.Risk.of.Mortality
10. # APR.Medical.Surgical.Description
11. # Emergency.Department.Indicator
12. # Total.Charges
13. # Total.Costs

Final Dimensions of Dataset: 27777 observations x 13 variables

(Optional) Rationale in data-cleaning:

1. I removed negative values under the "Profits" columns I created, calculated by $(Profits = Total.Charges - Total.Costs)$ These negative values are likely to represent invalid entries, as they show cases which costs exceeded the charges. Since they consist only small fraction of dataset (i.e. 328), I decided to remove them to maintain the integrity of data analysis.

2. I removed the "U", which is unknown under "Gender" column because these are invalid data, which consists of 3 entries only.

3. I also removed "NA" values in the dataset since they are relatively few of them only compared to the dataset.

4. I dropped some columns due to the reasons below:
    a. "Discharge.Year", "CCSR.Diagnosis.Code", "CCSR.Diagnosis.Description" is 1-level factor.
    b. "Payment Topology" has nothing to do with Length of Stay.
    c. "Patient.Disposition" & "APR.DRG.Description" is description / statement which has no meaning / related to Length of Stay.
    d. "APR.DRG.Code" mostly are 720 which has no meaning to Length of Stay as well.
    e. "APR.Severity.of.Illness.Code" is the same as "APR.Severity.of.Illness.Description".

5. I decided not to remove outliers for "Length of Stay", "Total Costs", and "Total Charges" because they reflect critical cases that impact hospital resource allocation and finances. Removing them could result in biased analysis by excluding important scenarios.

Answer to Q3:

Model	Complexity	Testset RMSE
Linear Regression	4 predictor X variables	5.930 days
CART	22 terminal nodes	5.908 days

Linear Regression:

1. Since I did not remove outliers, we need to do log transformation on those variables to dampen the effects of outliers which can significantly influence my regression analysis.
2. First, I checked the skewness of those 3 variables. I found out that all 3 of them are right-skewed. Therefore, I can perform log transformation.

```
skewness(test[,Length.of.Stay])   # skewness = 4.026
skewness(test[,Total.Charges])    # skewness = 9.604
skewness(test[,Total.Costs])    # skewness = 13.971
```

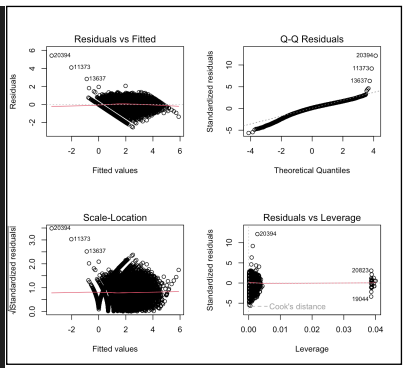
3. I constructed an initial linear regression model without adjustment to check the significance level of each variable and conducted a VIF test to check on multicollinearity.
4. There are a few concerns to address:
   a. GVIF of "APR.Severity.of.Illness.Description" and "APR.Risk.of.Mortality" are around 5, which indicates strong multicollinearity.
   b. GVIF of "Total.Charges" and "Total.Costs" are also around 5, which indicates strong multicollinearity.
5. Considering all conditions above, I finally constructed the adjusted linear regression model with 4 most significant predictor variables (i.e. Hospital.Service.Area, Age.Group, Total.Costs, APR.Severity.of.Illness.Description)

```
m1 <-  lm(log(Length.of.Stay) ~ . + log(Total.Costs) - Total.Costs - Total
.Charges - Type.of.Admission - APR.Risk.of.Mortality - Race - Ethnicity -
Emergency.Department.Indicator - Gender, data = test)
m1 <- step(m1)
summary(m1)
vif(m1)
par(mfrow = c(2,2))
plot(m1)
par(mfrow = c(1,1))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5214 -0.2467  0.0545  0.3009  5.4207

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              -6.766521   0.095469 -70.876  < 2e-16 ***
Hospital.Service.AreaCapital/Adirond       0.156921   0.088233   1.778 0.075336 .
Hospital.Service.AreaCentral NY            0.106769   0.088295   1.209 0.226585
Hospital.Service.AreaFinger Lakes          0.322224   0.088235   3.652 0.000261 ***
Hospital.Service.AreaHudson Valley        -0.140236   0.088052  -1.593 0.111251
Hospital.Service.AreaLong Island          -0.144902   0.087950  -1.648 0.099454 .
Hospital.Service.AreaNew York City        -0.270898   0.087854  -3.084 0.002048 **
Hospital.Service.AreaSouthern Tier         0.208920   0.089733   2.328 0.019907 *
Hospital.Service.AreaWestern NY            0.220977   0.088497   2.497 0.012530 *
Age.Group18 to 29                          0.227706   0.025328   8.990  < 2e-16 ***
Age.Group30 to 49                          0.257897   0.021961  11.743  < 2e-16 ***
Age.Group50 to 69                          0.297939   0.021144  14.091  < 2e-16 ***
Age.Group70 or Older                       0.354694   0.021005  16.886  < 2e-16 ***
APR.Severity.of.Illness.DescriptionModerate  0.021938   0.016399   1.338 0.180997
APR.Severity.of.Illness.DescriptionMajor   0.017121   0.016187   1.058 0.290198
APR.Severity.of.Illness.DescriptionExtreme -0.106396   0.016500  -6.448 1.15e-10 ***
APR.Medical.Surgical.DescriptionSurgical  -0.057664   0.009092  -6.342 2.30e-10 ***
log(Total.Costs)                           0.847547   0.003399 249.341  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.447 on 27759 degrees of freedom
Multiple R-squared:  0.7584,    Adjusted R-squared:  0.7583
F-statistic:  5126 on 17 and 27759 DF,  p-value: < 2.2e-16
```



R-squared: 0.7584
Adjusted R-squared: 0.7583

6. Lastly, I did a 70-30 train-test split to check the performance of the model by calculating RMSE. Since I log-transformed the variable, I make sure that it is transformed back to the original scale to evaluate the model. The RMSE value is 5.930 days.

CART:

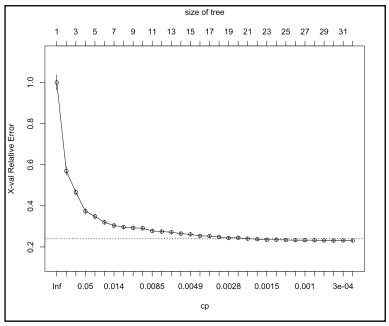
1. First, I did a 70-30 train-test split before constructing CART model.

```
# Split into 70% trainset and 30% testset
train <- sample.split(Y = test$Length.of.Stay, SplitRatio = 0.7)
trainset <- subset(test, train == T)
testset <- subset(test, train == F)
```

2. I construct the CART model by considering all X predictor variables.
3. This is because CART will select the best variable (that produces largest reduction in variance) recursively at each level and split accordingly.
4. I also set the max depth of 5 to prevent tree grow too deep and overfitting.

```
cart1 <- rpart(Length.of.Stay ~ . ,data = trainset, method = 'anova', control =
rpart.control(minsplit = 2,cp = 0, maxdepth = 5))
```

5. I noticed there are 32 terminal nodes by performing rpart.plot() function on cart1 model.
6. Next, I want to observe the cp value by performing printcp() function and plotcp() function to visualise the cp trend.



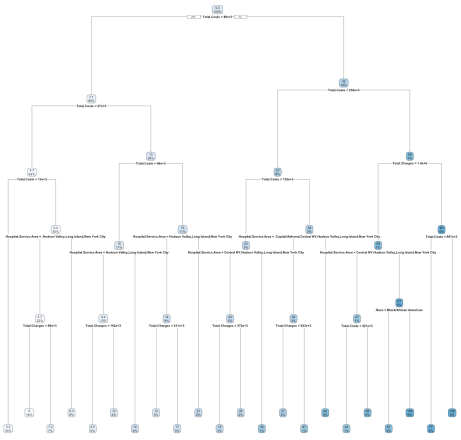
7. To obtain the optimal value of cp, I run the automation code learnt from tutorial as below.

```
# Automation to find the optimal cp.
# Compute min CV error + 1SE in maximal tree cart1.
CVerror.cap <- cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xerror"] +
cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xstd"]

# Find the optimal CP region whose CV error is just below CVerror.cap in maximal tree cart.
i <- 1; j<- 4
while (cart1$cptable[i,j] > CVerror.cap) {
  i <- i + 1
}

# Get geometric mean of the two identified CP values in the optimal region if optimal tree
has at least one split.
cp.opt = ifelse(i > 1, sqrt(cart1$cptable[i,1] * cart1$cptable[i-1,1]), 1)
cp.opt  # cp_opt = 0.0018
```

8. Then, I found out the optimal value of cp = 0.0018.
9. Next, I constructed cart2 model by pruning the cart1 model using the optimal value of cp to avoid overfitting, and I noticed there are 22 terminal nodes now.



10. Finally, I tested the performance of cart2 model using the testset data by calculating RSME value, which is 5.908 days.

Answer to Q4:

1. Both linear regression and CART models suggest similar RMSE values, that is 5.930 and 5.908 respectively with CART model having a slightly better accuracy.

2. Comparing to mean and median of the dataset (mean = 9.54 and median = 6), RMSE shows that the predictions are deviated from the original value by around 5.9 days. While both models may not predict short stays accurately, but I believe precise predictions for short stays are less crucial. Shorter length of stay usually release resources more quickly, resulting in higher turnover rate and allowing for quick reallocation.

3. Alternatively, the model's utility is more beneficial for predicting longer stays. By incorporating information from all outliers, the RMSE of 5.9 demonstrates a relatively minor deviation primarily for longer-stay cases. This proves that both model is capable of provide more precise predictions for extended hospitalization. This is very crucial for effective resources allocation in the field of healthcare. Longer stays usually require more intensive care and greater allocation of resources, including specialised staffs and equipment. This can cause a great toll on hospital management and the availability of facilities if resources are not allocated carefully especially when I realised the high number of emergency cases in all regions.

4. In addition, accurate prediction for longer stays are invaluable for financial planning for both hospital management and patients as well. Hospital management can better calculate the costs and potential revenues associated with extended hospitalization. On the other hand, patients have clearer insights into potential expenses, making informed decisions and enable to obtain necessary approvals from their insurance providers.

5. In conclusion, the self-adjusting nature of short stays supports a high resources utilization rate, allowing hospitals to manage capacity effectively without needing predictive models. By accurately forecasting longer stays, the hospital management can better prepare for the necessary staffing levels, ensure the availability of the specially curated machines and medical facilities, and allocate the beds more efficiently. Patients are also able to reduce unexpected costs and better prepare for their financial obligations.

Answer to Q5:

Improvement on Models:

1. We can perform feature engineering. For example, we can introduce some relevant features to aid in our analysis such as seasonality which may impact the length of stay. This could help us to identify the trend of patient behaviour and resource utilization since different seasons could have varying disease outbreak.

2. We can explore more complex models, such as Random Forest. A more complex model like Random Forest can capture a non-linear relationship better and less affected by the impacts of outliers, resulting in a more accurate and precise prediction model.

Improvements on Allocation of Resources

1. Hospital management can take necessary actions to engage with caregivers. We can involve family members or caregivers in the care process by providing them with resources and educations to assist them better manage the older patients' need at home. This is mainly because I discovered older age group contributed to a higher length of stay.

2. The state government can implement strategic approach such as segmenting areas for hospital allocation. It involves defining the geographic boundaries and assigning specific hospitals to serve those areas. Therefore, we can utilise the healthcare resources more efficiently and allocate the hospitalization demands accordingly. This can be seen by observing some areas have relatively higher overall length of stay compared to others.

3. Hospital management can encourage cross-department collaboration and research. It allows them to gather insights on various factors that influence the length of stay. By conducting research, they can come up with a better treatment measure and potentially improve the care quality and promote quicker recovery to reduce the length of stay in hospital.

4. Hospital management can utilise the real-time data integration system. It enables hospital management to better manage their resource allocations by having a real-time update on the availability of equipment and human resources. It can facilitate the allocation process more effectively and fully utilise the resources with this real-time information.