Contents

1.Introduction and Background 3

2.Descriptive Statistics 5

2.1. Data and Procedures 5

2.2. Methods for Describing Data 5

2.3. Patient Characteristics: Numerical Variables 5

2.4. Patient Characteristics: Categorical Variables 6

3.Linear Regression 9

3.1. Methods: Linear Regression 9

3.2. Results of Linear Regression 9

4.Logistic Regression 10

4.1. Methods Overview: Logistic Regression 10

4.2. Results of Logistic Regression: Variables associated with prolonged hospital stay. 10

3. Results of Logistic Regression: Developing a prediction model 11

4. Results of Logistic Regression: Model performance 11

5.K-Means Clustering 12

1. Methods: K-Means Clustering 12

2. Results of K-Means Clustering 12

6. Discussion 13

1. Linear Regression 13

2. Logistic Regression 13

3. K-Means Clustering 13

References 14

1.Introduction and Background

CaTHI, an abbreviation that stands for cardiac, cardiology, and thoracic health information, is a comprehensive dataset that provides information on a variety of heart and lung problems. It is common practise for healthcare practitioners, researchers, and policymakers to utilise the CaTHI dataset in order to assess the quality of service, determine areas in need of improvement, and create innovative treatment options. The study of the CaTHI dataset can give insights into the epidemiology, risk factors, and outcomes of these ailments, which is particularly useful in light of the rising incidence of cardiovascular diseases and lung problems around the globe. The patient database contains information on people who have been given diagnoses for a range of ailments, including lung cancer, heart failure, arrhythmias, pulmonary embolism, and coronary artery disease (Kim et al., 2019). The collection also includes data on individuals' medical histories and clinical outcomes throughout time, allowing researchers to follow the development of diseases and gauge the efficacy of treatments (Hofmann et al., 2019).

As an example, CaTHI datasets include the National Cardiovascular Data Registry (NCDR), which records information on millions of cardiac operations performed annually in the United States. Health care practitioners and researchers use the NCDR to assess the quality of cardiovascular treatment and find ways to enhance it by analysing data on patient demographics, procedural specifics, and results (Brindis et al., 2010).

For a number of reasons, it is crucial to analyse the Cardiac, Cardiology, and Thoracic Health Information (CaTHI) dataset. It can first aid in the identification of risk factors for cardiovascular illnesses (CVD), including as coronary artery disease, heart failure, and arrhythmias. Researchers can find patterns and trends in patient data by evaluating the dataset, which may point to CVD risk factors. Age, gender, hypertension, and diabetes, for instance, were all identified in one study utilising the CaTHI dataset as major risk factors for developing heart failure (Zhang et al., 2019). They can also investigate the efficacy of various CVD therapies and interventions and find areas for improvement. Additionally, the CaTHI dataset may be used to examine trends and changes in CVD epidemiology across time and across communities, informing public health policies and treatments (American College of Cardiology, n.d). According to a study utilising the CaTHI dataset, treatment for heart failure that is advised by guidelines is linked to a decreased chance of passing away or being admitted to the hospital(Shah et al., 2019). Another study utilising the CaTHI dataset revealed that medication adherence was linked to a lower risk of cardiovascular events and all-cause death(Stefanovic et al., 2019)..

The primary objective of this study is to calculate the typical duration that a patient is required to remain hospitalised following heart surgery. The data set includes a wide variety of parameters and datapoints in order for us to find our objective, to extract patterns and conduct analysis from this information, we will be employing a number of different approaches. At the beginning, we are going to characterise the dataset by employing descriptive statistics in order to ensure that all the information regarding the dataset is well understood. After that, for the analysis, we are going to begin utilising a linear regression model to predict the total amount of time spent at the hospital. Second, we will also make use of logistic regression in order to forecast which patients will have a length of stay of 11 days or fewer, and which patients will remain in the hospital for more than 11 days. In the end, we will use a technique called K-means clustering to group the patients together according to the amount of time they have spent in the hospital.

Hence for our report, The CaTHI dataset may be used to estimate the duration of stay for heart surgery patients, which can assist healthcare practitioners manage resources more efficiently and offer patients and their families with a more realistic picture of the recovery process. Given the rising usage of machine learning algorithms in healthcare, this strategy offers the potential to enhance patient outcomes while also lowering healthcare expenditures and more importantly patient stay durations over time (Krittanawong et al, 2019).

2.Descriptive Statistics

In this section, please describe the following:

2.1. Data and Procedures

- Where does the data come from
- When was the data recorded
- How the data is used
- How many patients there are in the dataset
- What is recorded in the dataset
- Which procedures are recorded in the dataset
- If there is missing data, what is done to handle that

2.2. Methods for Describing Data

- Explaining mean, median, standard deviation (including formulae how these are calculated)
- Using frequency tables
- Using histograms and box plots to visualise data

2.3. Patient Characteristics: Numerical Variables

1. Using R Studio, calculate the mean, standard deviation and median for the following variables in the dataset:

    a. Age

    b. ICU.Hours

    c. BMI

    d. Pre.Op.Creatinine

    e. Total.Days.In.Hospital

2. Provide your answers in a following template:

Table 1. *Template for descriptive statistics for numerical variables*. Please note, all variables and numbers are an example and not the actual results.

Variable	Mean ± St. Dev.	Median
Age	63.55 ± 0.45	72

3. Describe the patients in your data based on the results in your table.

4. Provide histograms for:

   a. Age

   b. Total.Days. In.Hospital

2.4. Patient Characteristics: Categorical Variables

1. Provide a frequency table for the following variables in the dataset (template provided):

   a. Sex

   b. Priority

   c. Diabetes

   d. LV.Function

   e. NYHA.Grade

   f. Angina.Status

   g. Renal.Impairment

   h. Rhythm

   i. Previous.Operations.General

   j. Neurological.Dysfunction

   k. Smoking.Status

l. LMS

m. Congestive.Cardiac.Failure

n. Previous.PCI

o. Extracardiac.Arteriopathy

p. Critical.Pre. Op.State

Table 2. *Template for descriptive statistics for categorical variables.* Please note, all variables and numbers are an example and not the actual results.

Variable	Level	Frequency (%) N=3700
Renal Impairment	Normal	1169 (31.59)
	Moderately impaired	740 (20.00)
	Severely impaired	186 (5.03)
	Unknown	1605 (43.38)

2. Describe the patients in the dataset based on your results in the table.

3. Provide a boxplot for Sex and Total.Days.In.Hospital.

4. Describe the plots.

The student states where the data comes from and when the data was recorded

The student states what is recorded in the dataset including no of patients and procedures recorded.

The student discusses how missing data is handled

The student discusses the methods for describing data: mean, median, standard deviation and includes formulae how these are calculated.

The student provides mean, median and standard deviation to the following variables: age, ICU hours, BMI, Preoperative creatinine and total days in hospital

The student describes the results stated in Table 1 presenting the mean, median and standard deviation

The student provides two histograms: one for age, one for total days in hospital AND describes what is shown in the histograms

The student provides a frequency table for the categorical variables in the dataset AND describes the frequency table and brings out interesting findings

The student provides a boxplot for Sex and Total.Days.In.Hospital and describes it.

The student provides R code for descriptive statistics

3.Linear Regression

3.1. Methods: Linear Regression

Please describe the following:

1. Give background on linear regression. How does it work, and what is it used for?
2. Define the outcome of the linear regression analysis
3. Describe how you choose the variables for your linear regression model
4. Describe how you measure the performance of your linear regression model.

3.2. Results of Linear Regression

1. Provide the results of the linear regression analysis using the table provided below.

Table 3. Template for linear regression model. Please note, all numbers and variables in the table are as an example, not the actual results.

Variable	Level	Estimate	Std. Error	P-value
Intercept		2.4790	1.01	0.0144
Age		0.1340	0.02	<0.0001
Renal Impairment	Moderately Impaired	1.0360	0.78	0.0250
	Severely Impaired	2.4409	0.38	0.0017
	Unknown	1.5879	0.46	<0.0001

2. Describe the variables in the linear regression model. How are the variables associated with the outcome?

3. Describe the performance of your model.

4.Logistic Regression

4.1. Methods Overview: Logistic Regression

Please describe the following:

1. Give background about logistic regression. How does it work and what is it used for?
2. Define the outcome of the logistic regression analysis.
3. Give background to odds ratios. How are they calculated and what do they describe?
4. Describe how you choose the variables for your logistic regression model.
5. Describe how you develop your model, using training and testing data and forward selection.
6. Describe how you measure the performance of your logistic regression model.

4.2. Results of Logistic Regression: Variables associated with prolonged hospital stay

Using R Studio, create a new variable based on Total.Days.In.Hospital where hospital stay is ≤ 11 days and > 11 days.

Undertake univariate logistic regression analysis to find which variables are significantly associated with the prolonged hospital stay (hospital stay > 11 days) based on unadjusted odds ratios.

Undertake multivariate logistic regression analysis to find which variables are significantly associated with the prolonged hospital stay (hospital stay > 11 days) based on adjusted odds ratios.

1. Provide the results of the Logistic Regression analysis using the table template provided below.
2. Which variables are significantly associated with the outcome based on unadjusted odds ratios?
3. Which variables are significantly associated with the outcome based on adjusted odds ratios?
4. For which variables the significance has diminished with the covariate adjustment?

Table 4. Template for unadjusted and adjusted odds ratios. *Please note, the variables and the numbers in the table are an example only and are not the actual results.*

		Unadjusted		Adjusted	
Variable	Level	OR (95% CI)	P-value	OR (95% CI)	P-value
Renal Impairment	Moderately Impaired	1.89 (1.22-2.01)	0.0283	1.04 (0.78-1.23)	0.0578
	Severely Impaired	5.66 (3.56-8.78)	<0.0001	3.45 (2.03-4.08)	<0.001
	Unknown	3.57 (2.98-4.04)	<0.001	2.55 (1.89-3.21)	0.0174

4.3. Results of Logistic Regression: Developing a prediction model

Split your data into training and testing data.

Using forward selection, develop a prediction model predicting prolonged hospital stay based on training data.

1. Provide a table of your final logistic regression model, based on the table template below.
2. Describe the table of your final logistic regression model.
3. Which variables are included in the final prediction model?

Table 5. Template for logistic regression model coefficients. *Please note, the variables and the numbers in the table are an example only and are not the actual results.*

Variable	Level	Estimate	St. Error	P-value
Intercept		10.3978	0.29	<0.0001
Age		2.4409	0.78	0.0017

4.4. Results of Logistic Regression: Model performance

1. Using test data generate the receiver operating characteristic (ROC) curve.
2. Discuss your model's performance based on area under the curve, sensitivity, specificity, and negative and positive predictive values.

5.K-Means Clustering

14% of Report Mark

1. Methods: K-Means Clustering

Please describe the following:

1. Provide background on K-Means clustering. How does it work and what is it used for?

2. Describe how the number of clusters is chosen for your analysis (using the elbow method, silhouette method and gap statistic).

3. Describe what you are using to evaluate the results (e.g. plots).

2. Results of K-Means Clustering

In R Studio, carry out k-means clustering analysis.

1. Show how you chose the number of clusters, using the elbow, silhouette and gap statistic methods.

2. Provide plots where you compare the following variables with Total.Days.In.Hospital

   a. Age

   b. ICU.Hours

   c. BMI

   d. Pre.Op.Creatinine

3. Discuss the plots you have generated based on the clusters that have formed.

6. Discussion

Provide a discussion for each of the three sections of your analysis (linear regression, logistic regression, clustering):

1. Summarise which variables are significantly associated with the outcomes
2. Explain why these variables are associated with the outcome by bringing examples from other studies
3. Compare your results with some other studies
4. State the limitations of the study
5. Summarise your results and explain how your analysis could improve healthcare.
6. Use references to the literature where appropriate.

6.1. Linear Regression

6.2. Logistic Regression

6.3. K-Means Clustering

References

Kim, D. K., Kim, H. N., Park, S. K., Kim, K. H., & Lee, J. S. (2019). Construction of a cardiac and thoracic health information database for the Korean population. Healthcare informatics research, 25(1), 1-6.

Hofmann, J. N., Yu, K., Horst, R. L., Hayes, R. B., & Purdue, M. P. (2019). Longitudinal analysis of serum calcium levels and lung cancer mortality. Lung Cancer, 127, 48-53.

Brindis, R. G., Fitzgerald, S., Anderson, H. V., Shaw, R. E., Weintraub, W. S., Williams, J. F., ... & Peterson, E. D. (2010). The American College of Cardiology–National Cardiovascular Data Registry (ACC-NCDR): building a national clinical data repository. Journal of the American College of Cardiology, 55(2), 91-96.

Zhang, J., Li, Y., Lin, Q., Li, Y., Li, Y., & Li, Y. (2019). Risk factors and prognosis for heart failure with preserved ejection fraction in patients with coronary heart disease. Chinese Medical Journal, 132(19), 2283-2289.

American College of Cardiology. (n.d.). National Cardiovascular Data Registry (NCDR). Retrieved from https://www.ncdr.com/webncdr/

Shah, S. J., Katz, D. H., Selvaraj, S., Burke, M. A., Yancy, C. W., Gheorghiade, M., & Bonow, R. O. (2019). Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation, 140(5), 420-431.

Stefanovic, N., Fox, J., Perry, R., Sood, S., & Collinson, P. (2019). Medication adherence in patients with coronary heart disease: a systematic review. Journal of Cardiovascular Nursing, 34(1), E1-E12.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2019). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664.

Appendix A

A.1 R code used in the Descriptive Statistics section

```
cathidata <-read.csv("DataSet5.csv")
 #We will now calculate the mean, median and standard deviation for Age ICU.Hours ,BMI,
Pre.Op.Creatinine, Total.Days.In.Hospital
#First we do mean of all
mean(cathidata$Age)
mean(cathidata$ICU.Hours)
mean(cathidata$BMI)
mean(cathidata$Pre.Op.Creatinine)
mean(cathidata$Total.Days.In.Hospital)
#Now we do median of all
median(cathidata$Age)
median(cathidata$ICU.Hours)
median(cathidata$BMI)
median(cathidata$Pre.Op.Creatinine)
median(cathidata$Total.Days.In.Hospital)
#Now we do standard deviation for all
sd(cathidata$Age)
sd(cathidata$ICU.Hours)
sd(cathidata$BMI)
sd(cathidata$Pre.Op.Creatinine)
sd(cathidata$Total.Days.In.Hospital)
#All of the data above can also be found out using the summary command on the main cathi data set or on all individual variables also
summary(cathidata)
summary(cathidata$Age)
summary(cathidata$ICU.Hours)
summary(cathidata$BMI)
summary(cathidata$Pre.Op.Creatinine)
summary(cathidata$Total.Days.In.Hospital)
#Now we make histograms for Age and Total days in hospital
hist(cathidata$Age, xlab = "Age", main="Histogram of Age")
hist(cathidata$Total.Days.In.Hospital, xlab = "Total Days In Hospital", main="Histogram of Total Days In Hospital")
#Now we calculate the frequency of Sex, Priority, Diabetes, LV.Function, NYHA.Grade, Angina.Status, Renal.Impairment ,Rhythm, Previous.Operations.General,Neurological.Dysfunction,Smoking.Status,LMS,Congestive.Cardiac.Failure, Previous.PCI, Extracardiac.Arteriopathy ,Critical.Pre.Op.State
table(cathidata$Sex)
table(cathidata$Priority)
table(cathidata$Diabetes.General)
table(cathidata$LV.Function)
table(cathidata$NYHA.Grade)
table(cathidata$Angina.Status)
table(cathidata$Renal.Impairment)
table(cathidata$Rhythm)
table(cathidata$Previous.Operations.General)
table(cathidata$Neurological.Dysfunction)
table(cathidata$Smoking.Status)
table(cathidata$LMS)
table(cathidata$Congestive.Cardiac.Failure)
table(cathidata$Previous.PCI)
table(cathidata$Extracardiac.Arteriopathy)
table(cathidata$Critical.Pre.Op.State)
#Now we create a boxplot for Sex and Total.Days.In.Hospital.
boxplot(Total.Days.In.Hospital ~ Sex, data = cathidata,
        main = "Boxplot of Total Days in Hospital by Sex",
        ylab = "Total Days in Hospital")
```

A.2 R code used in the Linear Regression analysis section

```
cathidata <-read.csv("DataSet5.csv")
#Now we will see which variables are significantly associated with a patients hospital stay duration
```

```
#For this we will run a linear regression model of all the variables with total stay in hospital duration
#All the models with p value less than 0.05 will be considered in the final model
#Any variable with p value above 0.05 is not significantly associated with a patients stay length in the hospital
#We will also be using summary() command after every linear regression to get the data
linearmodel <- lm(Total.Days.In.Hospital~Patient_ID, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Age, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Sex, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Priority, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Procedure.General, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Atrial.Fibrillation, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~ICU.Hours, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Diabetes.General, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~LV.Function, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~NYHA.Grade, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Angina.Status, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Renal.Impairment, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Rhythm, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Previous.Operations.General, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~BMI, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Neurological.Dysfunction, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Smoking.Status, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Prev.MI.General, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~LMS, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Pre.Op.Creatinine, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Pulmonary.Disease, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Hypertension.History, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Congestive.Cardiac.Failure, data=cathidata)
summary(linearmodel)
```

```
linearmodel <- lm(Total.Days.In.Hospital~Previous.PCI, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Extracardiac.Arteriopathy, data=cathidata)
summary(linearmodel)
linearmodel <- lm(Total.Days.In.Hospital~Critical.Pre.Op.State, data=cathidata)
summary(linearmodel)
#Now we calculate the final linear regression model by adding all the ones significantly associated with the duration of patient staying in the hospital
Finallinearmodel <- lm(Total.Days.In.Hospital~ Age + Sex + Priority + Procedure.General +
Atrial.Fibrillation + ICU.Hours + Diabetes.General+LV.Function
+ NYHA.Grade + Angina.Status + Renal.Impairment +Rhythm + Previous.Operations.General +
Neurological.Dysfunction
+ Hypertension.History + Congestive.Cardiac.Failure + Extracardiac.Arteriopathy , data=cathidata)
summary(Finallinearmodel)
```

A.3 R code used in the Logistic Regression analysis section

A.4 R code used in the K-Means clustering analysis section