Package Pricing at Mission Hospital

Case Study Report

Table of Contents

Intro

- The case is used to demonstrate how multiple regression models can be used to determine the relationship between the total cost of treatment and patient's health and demographic parameters at the time of admissions.

  - Objective
    - Develop a simple linear regression model to check if there is an association between total cost and body weight
    - Build a multiple regression predictive model and identify statistically significant predictors that Mission Hospital can use to determine treatment costs.

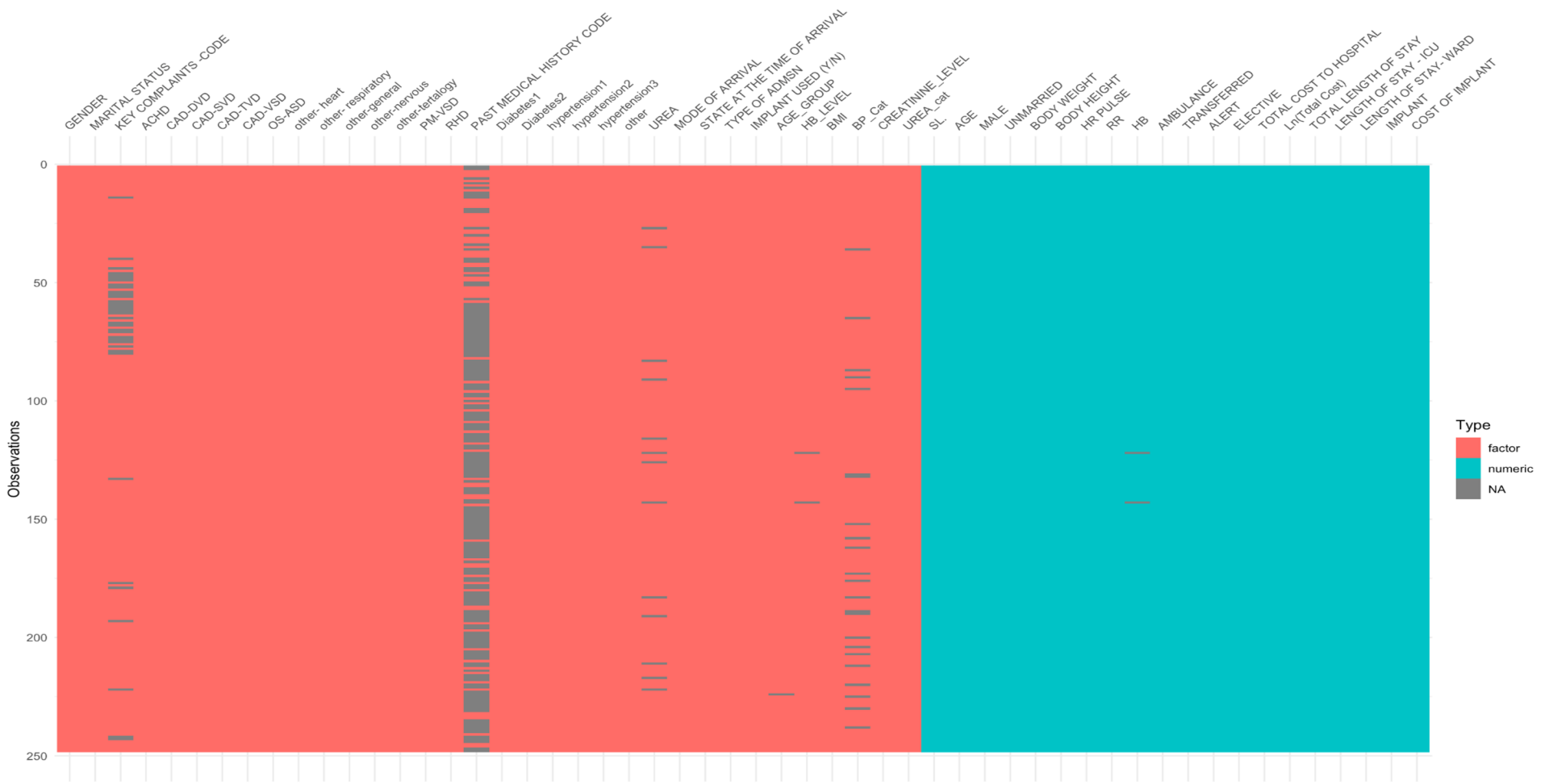
Data Exploration

- Target variable: “Total Cost To Hospital”
- Number of observations: 248
- Explanatory Variables: 19
- Removed “SL” index column

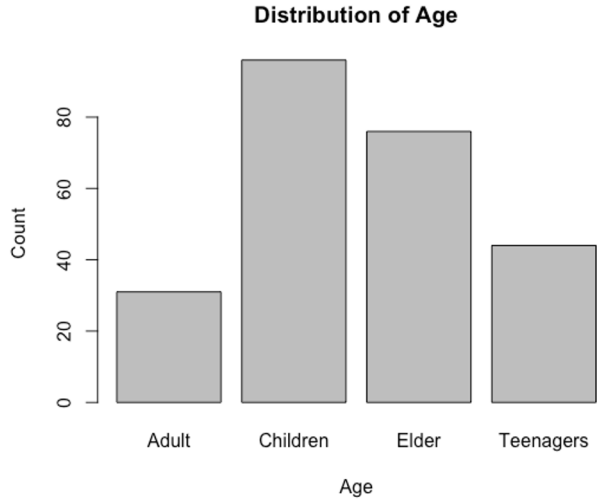
Demographical Data	Medical Data	Admissions Data	Symptoms Data
-Age -Gender -BMI = Height/Weight -Marital Status	-Key complaint code -Past medical history code -Implant (Y/N)	-Total Length of stay -Length of stay - ICU -Length of stay - Ward -Mode of arrival -State at arrival -Type of admission	-HR Pulse -BP - High/Low -Respiratory rate -HB -Urea -Creatinine

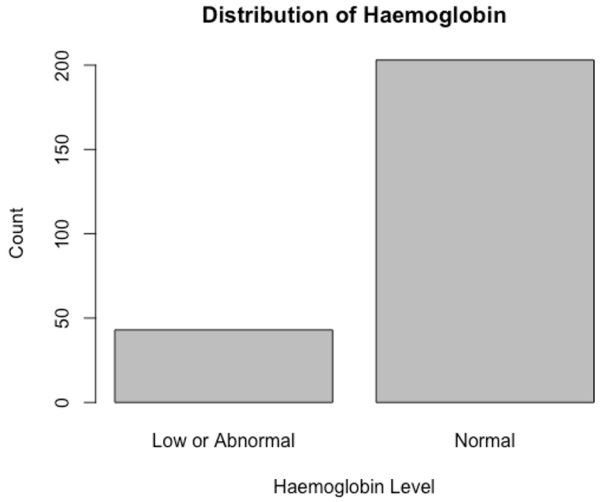
Feature Engineering

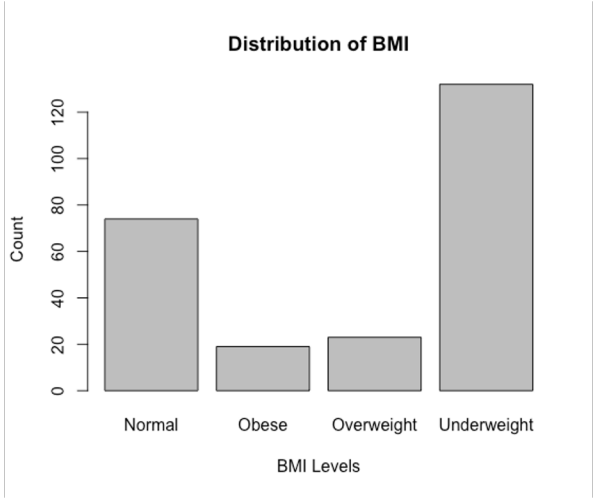
- Age groups
  - less than 11 -> “Children”
  - 11-25 -> “Teenager”
  - 26-49 -> “Adult”
  - 50 and above -> “Elder”
- BP ranges
  - Normal
  - Elevated
  - Hypertension - Stage I
  - Hypertension - Stage 2
  - Critical
- BMI
  - Underweight
  - Normal
  - Overweight
  - Obese
- Hemoglobin
  - Female: 12 -15.5 normal
  - Male: 13-17.5 -> normal
  - Thus, HB level >= 11 -> normal
  - Otherwise -> abnormal
- Urea
  - 7 to 20 mg/dl -> normal
  - Otherwise -> abnormal
- Creatinine levels
  - Age < 3 and creatinine 0.3-0.7 -> normal
  - Age: 3-18 and creatinine: 0.5-1.0 -> normal
  - Age >18 and Female and creatinine: 0.6 - 1.1 -> normal
  - Age > 18 and Male and creatinine: 0.9 - 1.3 -> normal
  - Otherwise -> abnormal

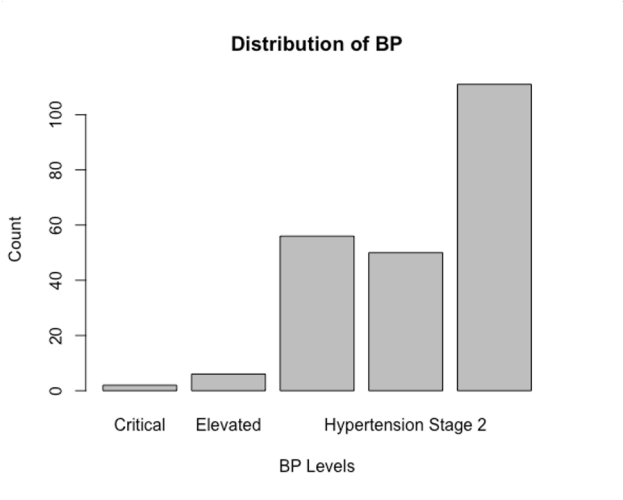


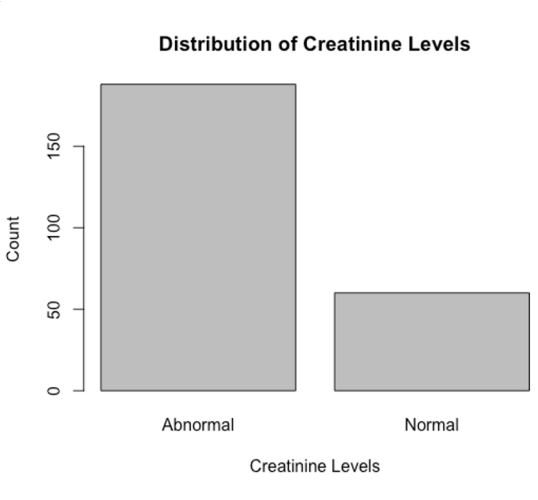
- Removed null values in the following columns “UREA”, “HB_LEVEL”, “BP_Cat”, and “HB”


Distribution of Age
Count
0
20
40
60
80
Adult
Children
Elder
Teenagers
Age

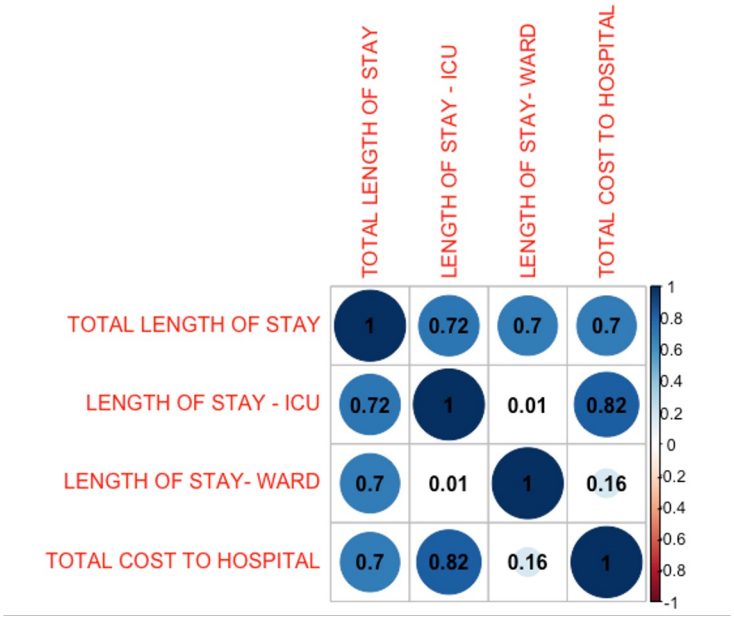

Distribution of Haemoglobin
Count
0
50
100
150
200
Low or Abnormal
Normal
Haemoglobin Level


Distribution of BMI
Count
0
20
40
60
80
100
120
Normal
Obese
Overweight
Underweight
BMI Levels

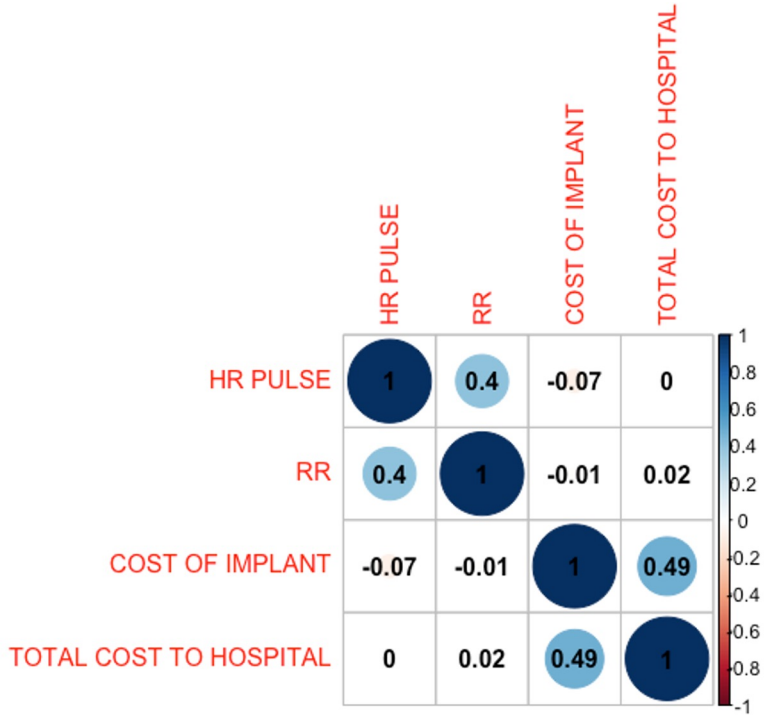

Distribution of BP
Count
0
20
40
60
80
100
Critical
Elevated
Hypertension Stage 2
BP Levels


Distribution of Creatinine Levels
Count
0
50
100
150
Abnormal
Normal
Creatinine Levels

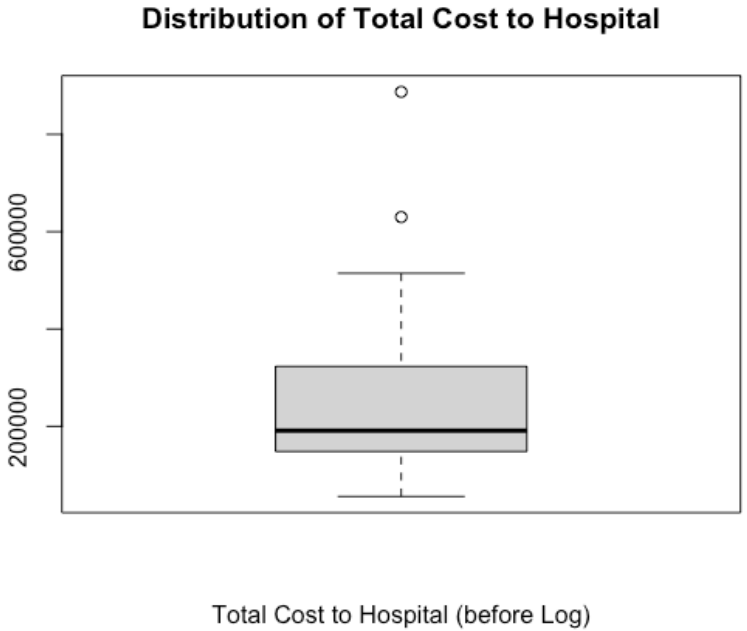
EDA (Correlation)

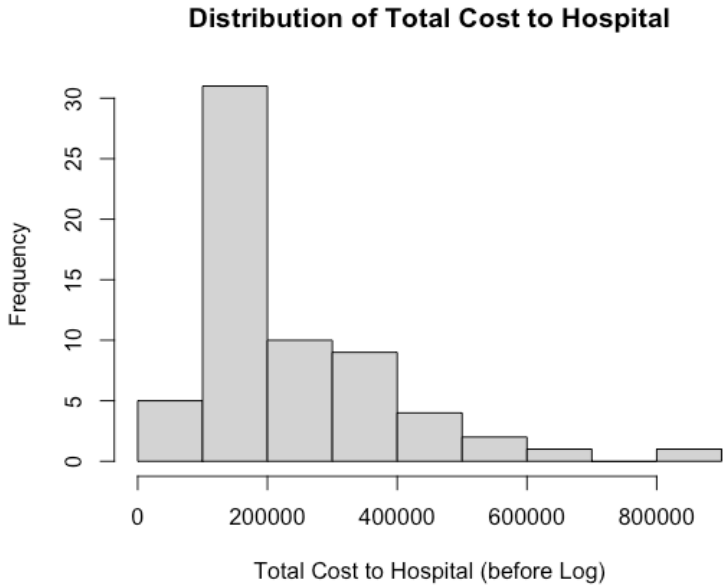


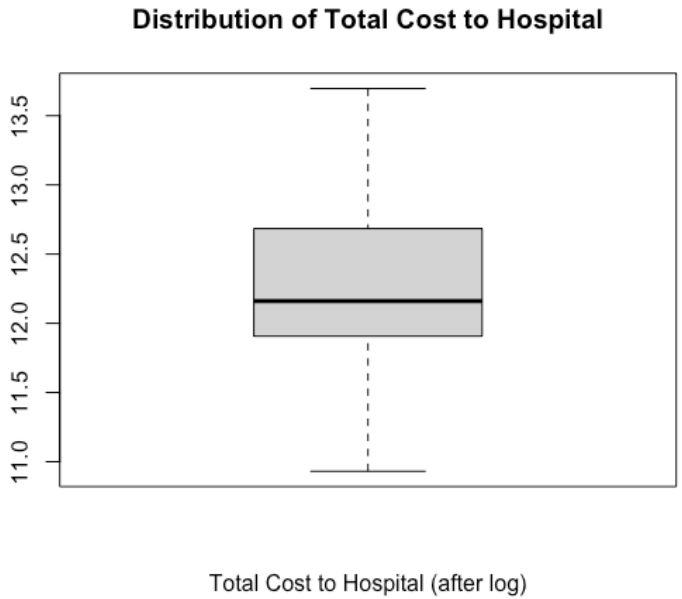
- Only consider one “STAY” variable
- “Total length of stay - ICU” is significantly correlated with the target variable “ Total Cost To Hospital”

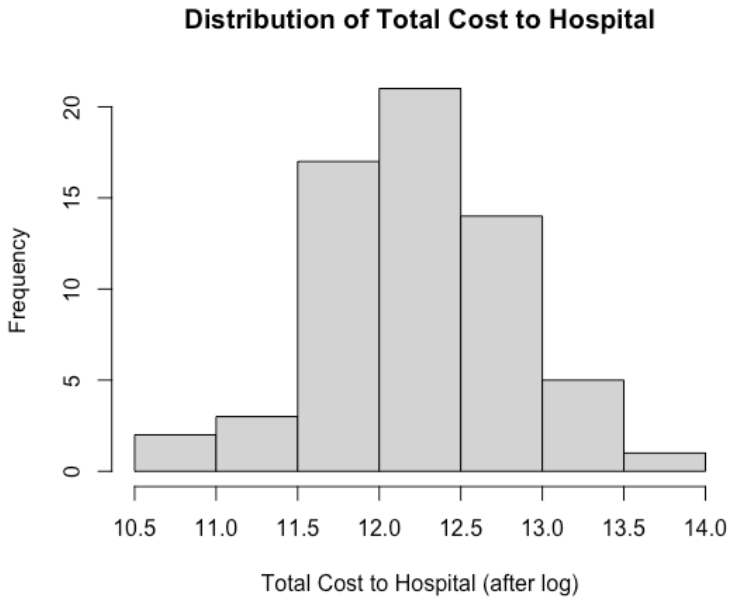


Distribution of Target Variable









Log Transformation (Target Variable)

Taking the log would make the distribution of our transformed variable appear more symmetric.

Statistical Significance Tests (at p=0.05)

- T-Test and Anova tests were performed on specific variables to understand their effect on target variable
- T test was performed on variables with 2 categories.
- ANOVA test was done to test for variables with more than two levels
- The following variables were tested insignificant:
  - Other-heart
  - Other-nervous
  - Other-respiratory
  - Diabetes
  - Hypertension
  - Hemoglobin
  - PM-VSD
  - Gender
  - CAD-SVD
  - CAS-VSD
  - Creatinine
  - Urea

T-Test (TOTAL LENGTH vs. TOTAL COST)

```
data:  data$`TOTAL COST TO HOSPITAL` and data$`TOTAL LENGTH OF STAY`
t = 25.249, df = 242, p-value < 0.0000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 184308.5 215499.3
sample estimates:
   mean of x    mean of y
199915.55930     11.63374
```

ANOVA (AGE GROUPS vs TOTAL COST)
SIGNIFICANT

```
Fit: aov(formula = data$`TOTAL COST TO HOSPITAL` ~ data$AGE_GROUP, data = data)

$`data$AGE_GROUP`
                          diff        lwr       upr     p adj
Children-Adult       -75467.238 -137231.14 -13703.33 0.0095454
Elder-Adult           34600.928  -28741.52  97943.38 0.4922724
Teenagers-Adult      -73478.337 -143034.92  -3921.75 0.0338655
Elder-Children       110068.166   64531.28 155605.05 0.0000000
Teenagers-Children     1988.901  -51857.73  55835.54 0.9996866
Teenagers-Elder     -108079.265 -163729.48 -52429.05 0.0000059
```

ANOVA
‘BMI’σ vs. ‘TOTAL COST’
SIGNIFICANT

```
    95% family-wise confidence level

Fit: aov(formula = data$`TOTAL COST TO HOSPITAL` ~ data$BMI, data = data)

$`data$BMI`
                           diff        lwr        upr     p adj
Obese-Normal           -44656.56 -133554.13  44240.999 0.5640494
Overweight-Normal      -21067.62  -93884.91  51749.666 0.8772271
Underweight-Normal     -84765.93 -129061.39 -40470.475 0.0000083
Overweight-Obese        23588.94  -79806.48 126984.370 0.9349608
Underweight-Obese      -40109.37 -125843.52  45624.783 0.6208650
Underweight-Overweight -63698.31 -132618.01   5221.386 0.0815071
```

```
> anova_BMI <- aov(data$`TOTAL COST TO HOSPITAL`~data$BMI, data = data)
> summary(anova_BMI)
             Df        Sum Sq      Mean Sq F value    Pr(>F)
data$BMI      3  364105890849 121368630283   8.732 0.0000161 ***
Residuals   239 3321969833192  13899455369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA
'HYPERTENSION'
INSIGNIFICANT

```
> TukeyHSD(anova_BP)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = data$`TOTAL COST TO HOSPITAL` ~ data$BP_Cat, data = data)

$`data$BP_Cat`
                                           diff        lwr       upr     p adj
Elevated-Critical                     -89085.273 -367137.47 188966.92 0.9035242
Hypertension Stage 1-Critical         -48554.265 -293693.62 196585.09 0.9824685
Hypertension Stage 2-Critical         -21668.682 -267334.09 223996.73 0.9992292
Normal-Critical                       -79191.389 -322190.79 163808.01 0.8979568
Hypertension Stage 1-Elevated          40531.009 -105882.08 186944.10 0.9412518
Hypertension Stage 2-Elevated          67416.591  -79875.57 214708.75 0.7164713
Normal-Elevated                         9893.885 -132907.36 152695.13 0.9997034
Hypertension Stage 2-Hypertension Stage 1  26885.583  -40011.82  93782.98 0.8034783
Normal-Hypertension Stage 1           -30637.124  -86961.88  25687.63 0.5659740
Normal-Hypertension Stage 2           -57522.707 -116094.59   1049.18 0.0569566
```

```
> anova_BP <- aov(data$`TOTAL COST TO HOSPITAL`~data$BP_Cat, data = data)
> summary(anova_BP)
             Df        Sum Sq     Mean Sq F value Pr(>F)
data$BP_Cat   4  131030997078 32757749270   2.138 0.0772 .
Residuals   216 3310010777237 15324123969
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
22 observations deleted due to missingness
```

Linear Regression Model ("Body Weight" and "Total Cost")

```
lm(formula = data$`Ln(Total Cost)` ~ data$`BODY WEIGHT`)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44260 -0.36255 -0.00179  0.33706  1.36488

Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)        11.692802   0.165290  70.741 < 0.0000000000000002 ***
data$`BODY WEIGHT`  0.010638   0.002919   3.645             0.000554 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5119 on 61 degrees of freedom
Multiple R-squared:  0.1788,    Adjusted R-squared:  0.1654
F-statistic: 13.29 on 1 and 61 DF,  p-value: 0.0005543
```

- Interpretations:
  - Positive relationship
  - Equation:
- **ln(Total Cost) = 0.010638*(Body Weight) + 11.692802**
  - **Body Weight Coefficient = 0.010638**
  - If there is 1 unit increase in weight, there will be 0.010638 times log value of total cost.
  - e^0.010638 = 1.01069 -> 0.01069 unit increase from total cost
  - The average cost of a person weight 50 kg is 198,723 INR.
  - A patient weighing 51 kg is likely to spend more than 198,723*0.01069 = 2124 INR than a patient weighing 50 kg.

Model 1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.3459691719	0.3515111128	32.278	< 0.0000000000000002	***
AGE	0.0006479066	0.0028088900	0.231	0.81797	
GENDERM	0.0191263542	0.0318898014	0.600	0.54979	
MALE	NA	NA	NA	NA	
`MARITAL STATUS`UNMARRIED	0.0108290423	0.0782212895	0.138	0.89012	
UNMARRIED	NA	NA	NA	NA	
ACHD1	0.0225156523	0.0673374970	0.334	0.73868	
`CAD-DVD`1	0.1213256701	0.0711292841	1.706	0.09065	.
`CAD-SVD`1	-0.0204593580	0.1479139545	-0.138	0.89022	
`CAD-TVD`1	0.1757910838	0.0709877229	2.476	0.01467	*
`CAD-VSD`1	0.0184896812	0.1762945620	0.105	0.91665	
`OS-ASD`1	0.1090288546	0.0695670699	1.567	0.11969	
`other- heart`1	0.0707631128	0.0611290777	1.158	0.24933	
`other- respiratory`1	0.1272839125	0.0762611525	1.669	0.09771	.
`other-general`1	-0.5792406993	0.2014064873	-2.876	0.00477	**
`other-nervous`1	0.1289950111	0.1822854565	0.708	0.48053	
`other-tertalogy`1	0.1965230785	0.0829175124	2.370	0.01938	*
`PM-VSD`1	0.0915104672	0.1139371654	0.803	0.42347	
RHD1	0.0855084877	0.0854468792	1.001	0.31898	
`HR PULSE`	0.0009688156	0.0009336539	1.038	0.30151	
RR	-0.0049016965	0.0046182907	-1.061	0.29066	
Diabetes11	0.0302009370	0.0771186545	0.392	0.69604	
Diabetes21	0.0915751132	0.0895343096	1.023	0.30846	
hypertension11	0.0639210298	0.0608633790	1.050	0.29572	
hypertension21	-0.0397867185	0.0670191029	-0.594	0.55386	
hypertension31	0.0668182474	0.1160641795	0.576	0.56590	
other1	-0.1087669149	0.0618083872	-1.760	0.08100	.
HB	-0.0001832321	0.0062122328	-0.029	0.97652	
UREA143	0.0211570076	0.1898770841	0.111	0.91147	
UREA21	0.0685489774	0.0744295949	0.921	0.35890	
UREAAbormal	NA	NA	NA	NA	
`MODE OF ARRIVAL`TRANSFERRED	0.1794924526	0.2375683086	0.756	0.45141	
`MODE OF ARRIVAL`WALKED IN	0.1345863322	0.1979587779	0.680	0.49790	
AMBULANCE	NA	NA	NA	NA	
TRANSFERRED	NA	NA	NA	NA	
ALERT	NA	NA	NA	NA	
`TYPE OF ADMSN`EMERGENCY	0.1093134954	0.1961279290	0.557	0.57832	
ELECTIVE	NA	NA	NA	NA	
`TOTAL COST TO HOSPITAL`	0.0000025410	0.0000003035	8.372	0.000000000000122	***
`TOTAL LENGTH OF STAY`	-0.0017775500	0.0865562139	-0.021	0.98365	
`LENGTH OF STAY - ICU`	0.0151246542	0.0874774587	0.173	0.86302	
`LENGTH OF STAY- WARD`	0.0139276053	0.0861629632	0.162	0.87186	
`IMPLANT USED (Y/N)`Y	0.2133593819	0.0669268743	3.188	0.00183	**
IMPLANT	NA	NA	NA	NA	
`COST OF IMPLANT`	-0.0000004358	0.0000011755	-0.371	0.71146	
AGE_GROUPChildren	-0.0507381868	0.1323146437	-0.383	0.70205	
AGE_GROUPElder	-0.0057103334	0.0775503047	-0.074	0.94142	
AGE_GROUPTeenagers	-0.0378442029	0.1107566488	-0.342	0.73318	
HB_LEVELNormal	-0.0405499256	0.0550249430	-0.737	0.46260	
BMIObese	-0.0171606656	0.0652594748	-0.263	0.79303	
BMIOverweight	0.0040613650	0.0577353741	0.070	0.94404	
BMIUnderweight	-0.0593045202	0.0498635834	-1.189	0.23666	
BP_CatElevated	-0.0868526084	0.1834336762	-0.473	0.63673	
BP_CatHypertension Stage 1	-0.1225821729	0.1382819414	-0.886	0.37714	
BP_CatHypertension Stage 2	-0.1361084904	0.1426770484	-0.954	0.34202	
BP_CatNormal	-0.1332835001	0.1372259514	-0.971	0.33337	
CREATININE_LEVELNormal	0.1111828053	0.0434733649	2.557	0.01179	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1648 on 120 degrees of freedom
Multiple R-squared: 0.9237, Adjusted R-squared: 0.8932
F-statistic: 30.27 on 48 and 120 DF, p-value: < 0.00000000000000022

```
> # Model performance
> # (a) Prediction error, RMSE
> RMSE(predictions, test.data$`Ln(Total Cost)`)
[1] 0.2601287
> # (b) R-square
> R2(predictions, test.data$`Ln(Total Cost)`)
[1] 0.7743484
```

```
Call:
lm(formula = `Ln(Total Cost)` ~ BMI + AGE_GROUP + AMBULANCE +
    `COST OF IMPLANT` + `LENGTH OF STAY - ICU` + `LENGTH OF STAY- WARD` +
    `IMPLANT USED (Y/N)` + `MODE OF ARRIVAL`, data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82594 -0.07037  0.03380  0.11837  0.60401

Coefficients: (1 not defined because of singularities)
                               Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                11.524143022  0.070796479 162.778 < 0.0000000000000002 ***
BMIObese                   -0.032314203  0.079150994  -0.408              0.68364
BMIOverweight              -0.019409670  0.068505820  -0.283              0.77730
BMIUnderweight             -0.059448637  0.060333452  -0.985              0.32599
AGE_GROUPChildren          -0.057086975  0.076441549  -0.747              0.45630
AGE_GROUPElder              0.045401901  0.062000304   0.732              0.46509
AGE_GROUPTeenagers         -0.039902724  0.077246392  -0.517              0.60619
AMBULANCE                  -0.136787327  0.059002910  -2.318              0.02173 *
`COST OF IMPLANT`           0.000003766  0.000001325   2.841              0.00509 **
`LENGTH OF STAY - ICU`      0.101495916  0.006313508  16.076 < 0.0000000000000002 ***
`LENGTH OF STAY- WARD`      0.023444767  0.004971852   4.715           0.00000531 ***
`IMPLANT USED (Y/N)`Y       0.229483710  0.076553476   2.998              0.00317 **
`MODE OF ARRIVAL`TRANSFERRED -0.012332758  0.137274033  -0.090              0.92853
`MODE OF ARRIVAL`WALKED IN             NA           NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2272 on 156 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.7971
F-statistic: 55.98 on 12 and 156 DF,  p-value: < 0.00000000000000022
```

Model 2

```
> # Model performance
> # (a) Prediction error, RMSE
> RMSE(predictions, test.data$`Ln(Total Cost)`)
[1] 0.306256
> # (b) R-square
> R2(predictions, test.data$`Ln(Total Cost)`)
[1] 0.6650867
>
```

Conclusion

The statistically significant predictors that Mission Hospital can use to determine treatment costs are based on the model:

- ‘Cost of Implant’
- ‘Length of Stay’
- Implant used
- Ambulance