**Project Report**

**Rushikesh Khankar**

**Healthcare cost analysis**

Project 7

DESCRIPTION

**Background and Objective:**

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on healthcare costs and their utilization.

**Domain:** Healthcare

**Dataset Description:**

Here is a detailed description of the given dataset:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Age | Age of the patient discharged |
| Female | A binary variable that indicates if the patient is female |
| Los | Length of stay in days |
| Race | Race of the patient (specified numerically) |
| Totchg | Hospital discharge costs |
| Aprdrg | All Patient Refined Diagnosis Related Groups |

**Analysis to be done:**

1. To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

3. To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.

5. Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

6. To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

My Code :{Solution}

library(dplyr)

> setwd("/Users/rushikeshkhankar/Desktop/R")

> getwd()

[1] "/Users/rushikeshkhankar/Desktop/R"

> #Loading Dataset

> data <- read.csv("~/Desktop/R/Project/Projects for Submission/Healthcare/Healthcare/HospitalCosts.csv")

> data

AGE FEMALE LOS RACE TOTCHG APRDRG

1 17 1 2 1 2660 560

2 17 0 2 1 1689 753

3 17 1 7 1 20060 930

4 17 1 1 1 736 758

5 17 1 1 1 1194 754

6 17 0 0 1 3305 347

7 17 1 4 1 2205 754

8 16 1 2 1 1167 754

9 16 1 1 1 532 753

10 17 1 2 1 1363 758

11 17 1 2 1 1245 758

12 15 0 2 1 1656 753

13 15 1 2 1 1379 751

14 15 1 4 1 2346 758

15 15 1 7 1 4006 753

16 15 1 4 1 2181 758

17 14 1 1 1 628 754

18 14 1 4 1 2463 758

19 15 1 3 1 1956 753

20 14 1 3 1 1802 758

21 13 1 1 1 3188 812

22 17 1 2 1 2129 566

23 12 0 1 1 7421 249

24 15 1 1 1 1122 422

25 13 1 2 4 1173 754

26 12 0 2 1 3625 812

27 11 1 2 1 3908 50

28 15 0 1 1 3994 139

29 11 0 0 1 1033 753

30 10 0 2 1 2860 141

31 11 0 2 1 3814 420

32 7 0 0 1 1132 139

33 16 1 2 6 1163 751

34 17 1 1 1 610 751

35 6 0 3 1 9530 97

36 15 1 1 1 1268 811

37 17 1 4 1 2582 753

38 16 1 2 1 1287 755

39 17 1 3 1 6594 930

40 13 1 0 1 909 755

41 7 0 0 1 2530 347

42 11 1 2 2 1534 753

43 3 0 5 1 14243 720

44 16 1 3 1 1699 754

45 2 0 2 1 7298 53

46 16 1 1 1 636 754

47 15 1 1 1 626 754

48 1 0 2 1 3782 53

49 14 1 2 1 1444 753

50 14 1 2 1 1183 754

51 14 1 5 1 3045 754

52 14 1 5 1 3624 754

53 14 1 12 1 6810 760

54 1 0 1 1 1409 249

55 13 0 2 1 1211 754

56 1 0 4 1 9606 53

57 1 1 1 1 1411 249

58 15 1 0 1 607 754

59 1 0 1 1 2932 249

60 1 0 3 1 5075 139

61 14 1 1 1 762 753

62 16 1 6 1 6329 753

63 17 1 1 1 1226 753

64 3 1 4 1 8223 710

65 17 0 2 1 1193 776

66 13 1 2 1 1076 754

67 12 1 6 1 17434 115

68 12 1 2 1 1647 753

69 14 1 7 1 3865 754

70 13 1 1 1 628 754

71 15 1 1 1 806 755

72 0 1 41 1 29188 602

73 0 0 2 1 4717 138

74 0 0 12 1 15129 137

75 0 1 2 1 1085 640

76 0 0 3 1 1607 640

77 0 1 3 1 1499 640

78 0 1 3 1 7648 53

79 0 1 2 1 1527 640

80 0 0 2 1 1483 640

81 0 1 4 1 2844 640

82 0 1 3 1 3124 640

83 0 0 3 1 1760 640

84 0 1 2 1 1278 640

85 0 1 2 1 1620 640

86 0 1 2 1 1220 640

87 0 1 2 1 1134 640

88 16 1 0 1 1235 754

89 0 0 3 1 1656 640

90 0 0 4 5 4072 639

91 0 0 2 5 1393 143

92 0 0 0 5 615 254

93 16 1 1 1 779 755

94 0 0 2 1 1385 640

95 0 0 2 1 1224 640

96 0 1 3 1 1779 640

97 0 0 2 1 1526 640

98 15 1 1 1 882 754

99 0 0 1 1 2075 581

100 0 0 17 1 12042 633

101 0 0 2 1 1309 640

102 0 0 2 1 1290 640

103 0 0 2 1 1280 640

104 0 0 3 1 1719 640

105 0 1 2 1 1102 640

106 0 1 3 1 1543 640

107 0 1 2 1 1174 640

108 0 1 2 1 1105 640

109 0 0 2 1 1335 640

110 0 0 2 1 1550 640

111 0 0 4 1 2473 640

112 0 0 2 1 1322 640

113 0 0 4 1 2553 640

114 15 0 5 1 2835 753

115 0 1 2 1 1191 640

116 0 0 2 1 1439 640

117 0 1 2 1 1237 640

118 0 0 2 1 1265 640

119 0 1 4 1 2280 640

120 0 0 2 1 1096 640

121 0 1 2 1 1156 640

122 0 0 2 1 1199 640

123 13 1 10 1 5615 754

124 0 1 4 1 2518 640

125 15 0 0 1 625 754

126 0 1 2 1 1246 640

127 0 1 3 1 1821 640

128 0 0 5 1 3101 626

129 12 1 2 1 1293 754

130 0 1 2 1 1176 640

131 0 0 3 1 1891 640

132 5 1 2 1 10584 53

133 13 1 3 1 2373 754

134 0 0 1 1 935 640

135 0 0 2 1 1395 640

136 0 0 2 1 1561 640

137 0 1 7 1 6912 636

138 12 1 2 1 1157 754

139 0 0 3 1 2197 640

140 0 0 4 1 2288 640

141 16 1 4 1 2348 754

142 0 0 2 1 1320 640

143 0 1 2 1 1139 640

144 0 1 4 1 2134 639

145 0 0 2 1 1407 640

146 0 0 2 1 1982 640

147 0 0 4 1 2539 640

148 0 0 2 1 1528 640

149 0 1 2 1 1513 640

150 0 1 2 1 1191 640

151 0 0 2 1 1280 640

152 0 0 2 1 3977 139

153 0 1 2 1 1269 640

154 0 0 2 1 1501 640

155 0 1 2 1 1396 640

156 0 0 3 1 1777 640

157 0 1 1 1 833 640

158 0 1 1 1 715 640

159 17 1 5 1 2936 751

160 0 0 2 1 1375 640

161 0 0 2 1 1330 640

162 0 0 2 1 1628 640

163 0 0 2 1 1368 640

164 12 1 1 1 622 755

165 17 0 2 1 14174 23

166 7 0 1 1 6425 57

[ reached 'max' / getOption("max.print") -- omitted 334 rows ]

> #Descriptive Analysis

> View(data)

> str(data)

'data.frame': 500 obs. of 6 variables:

$ AGE : int 17 17 17 17 17 17 17 16 16 17 ...

$ FEMALE: int 1 0 1 1 1 0 1 1 1 1 ...

$ LOS : int 2 2 7 1 1 0 4 2 1 2 ...

$ RACE : int 1 1 1 1 1 1 1 1 1 1 ...

$ TOTCHG: int 2660 1689 20060 736 1194 3305 2205 1167 532 1363 ...

$ APRDRG: int 560 753 930 758 754 347 754 754 753 758 ...

> summary(data)

AGE FEMALE

Min. : 0.000 Min. :0.000

1st Qu.: 0.000 1st Qu.:0.000

Median : 0.000 Median :1.000

Mean : 5.086 Mean :0.512

3rd Qu.:13.000 3rd Qu.:1.000

Max. :17.000 Max. :1.000

LOS RACE

Min. : 0.000 Min. :1.000

1st Qu.: 2.000 1st Qu.:1.000

Median : 2.000 Median :1.000

Mean : 2.828 Mean :1.078

3rd Qu.: 3.000 3rd Qu.:1.000

Max. :41.000 Max. :6.000

NA's :1

TOTCHG APRDRG

Min. : 532 Min. : 21.0

1st Qu.: 1216 1st Qu.:640.0

Median : 1536 Median :640.0

Mean : 2774 Mean :616.4

3rd Qu.: 2530 3rd Qu.:751.0

Max. :48388 Max. :952.0

> #Handling Missing Values

> data <- na.omit(data)

> hist(data$AGE, main = "Frequency of Patients", breaks = 8, col = "darkorange")

> attach(data)

> AGE = as.factor(AGE)

> summary(AGE)

0 1 2 3 4 5 6 7 8 9

306 10 1 3 2 2 2 3 2 2

10 11 12 13 14 15 16 17

4 8 15 18 25 29 29 38

> max\_value = max(aggregate(TOTCHG~AGE, FUN = sum,data = data))

> max\_value

[1] 676962

> hist(data$APRDRG, main = "Diagnosis Related Group", breaks = 8, col = "darkorange")

> APRDRG=as.factor(APRDRG)

> summary(APRDRG)

21 23 49 50 51 53 54 57 58 92

1 1 1 1 1 10 1 2 1 1

97 114 115 137 138 139 141 143 204 206

1 1 2 1 4 5 1 1 1 1

225 249 254 308 313 317 344 347 420 421

2 6 1 1 1 1 2 3 2 1

422 560 561 566 580 581 602 614 626 633

3 2 1 1 1 3 1 3 6 4

634 636 639 640 710 720 723 740 750 751

2 3 4 266 1 1 2 1 1 14

753 754 755 756 758 760 776 811 812 863

36 37 13 2 20 2 1 2 3 1

911 930 952

1 2 1

> df=aggregate(TOTCHG~APRDRG,FUN = sum, data = data)

> df[which.max(df$TOTCHG),]

APRDRG TOTCHG

44 640 436822

> aovt <-aov(TOTCHG~factor(RACE), data = data)

> summary(aovt)

Df Sum Sq Mean Sq

factor(RACE) 5 1.859e+07 3718656

Residuals 493 7.524e+09 15260687

F value Pr(>F)

factor(RACE) 0.244 0.943

Residuals

> model1=lm(TOTCHG~AGE+factor(FEMALE),data = data)

> summary(model1)

Call:

lm(formula = TOTCHG ~ AGE + factor(FEMALE), data = data)

Residuals:

Min 1Q Median 3Q Max

-3403 -1444 -873 -156 44950

Coefficients:

Estimate Std. Error

(Intercept) 2719.45 261.42

AGE 86.04 25.53

factor(FEMALE)1 -744.21 354.67

t value Pr(>|t|)

(Intercept) 10.403 < 2e-16 \*\*\*

AGE 3.371 0.000808 \*\*\*

factor(FEMALE)1 -2.098 0.036382 \*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’

0.1 ‘ ’ 1

Residual standard error: 3849 on 496 degrees of freedom

Multiple R-squared: 0.02585, Adjusted R-squared: 0.02192

F-statistic: 6.581 on 2 and 496 DF, p-value: 0.001511

> model2=lm(LOS~AGE+FEMALE+factor(RACE), data = data)

> summary(model2)

Call:

lm(formula = LOS ~ AGE + FEMALE + factor(RACE), data = data)

Residuals:

Min 1Q Median 3Q Max

-3.211 -1.211 -0.857 0.143 37.789

Coefficients:

Estimate Std. Error t value

(Intercept) 2.85687 0.23160 12.335

AGE -0.03938 0.02258 -1.744

FEMALE 0.35391 0.31292 1.131

factor(RACE)2 -0.37501 1.39568 -0.269

factor(RACE)3 0.78922 3.38581 0.233

factor(RACE)4 0.59493 1.95716 0.304

factor(RACE)5 -0.85687 1.96273 -0.437

factor(RACE)6 -0.71879 2.39295 -0.300

Pr(>|t|)

(Intercept) <2e-16 \*\*\*

AGE 0.0818 .

FEMALE 0.2586

factor(RACE)2 0.7883

factor(RACE)3 0.8158

factor(RACE)4 0.7613

factor(RACE)5 0.6626

factor(RACE)6 0.7640

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’

0.1 ‘ ’ 1

Residual standard error: 3.376 on 491 degrees of freedom

Multiple R-squared: 0.008699, Adjusted R-squared: -0.005433

F-statistic: 0.6156 on 7 and 491 DF, p-value: 0.7432

> model3 = lm(TOTCHG~AGE+FEMALE+LOS+RACE+APRDRG,data = data)

> summary(model3)

Call:

lm(formula = TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG, data = data)

Residuals:

Min 1Q Median 3Q Max

-6377 -700 -174 122 43378

Coefficients:

Estimate Std. Error t value

(Intercept) 5218.6769 507.6475 10.280

AGE 134.6949 17.4711 7.710

FEMALE -390.6924 247.7390 -1.577

LOS 743.1521 34.9225 21.280

RACE -212.4291 227.9326 -0.932

APRDRG -7.7909 0.6816 -11.430

Pr(>|t|)

(Intercept) < 2e-16 \*\*\*

AGE 7.02e-14 \*\*\*

FEMALE 0.115

LOS < 2e-16 \*\*\*

RACE 0.352

APRDRG < 2e-16 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’

0.1 ‘ ’ 1

Residual standard error: 2613 on 493 degrees of freedom

Multiple R-squared: 0.5536, Adjusted R-squared: 0.5491

F-statistic: 122.3 on 5 and 493 DF, p-value: < 2.2e-16

**Chart

Description automatically generated**

**Chart, histogram

Description automatically generated**