|  |
| --- |
| > HospitalCosts = read.csv(file= choose.files())  >  > library(tidyverse)  >  > hosp = as\_tibble(HospitalCosts)  >  > #Visually inspecting the data  > hosp  # A tibble: 500 x 6  AGE FEMALE LOS RACE TOTCHG APRDRG  *<int>* *<int>* *<int>* *<int>* *<int>* *<int>*  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  # ... with 490 more rows  >  > summary(hosp)  AGE FEMALE LOS RACE TOTCHG APRDRG  Min. : 0.000 Min. :0.000 Min. : 0.000 Min. :1.000 Min. : 532 Min. : 21.0  1st Qu.: 0.000 1st Qu.:0.000 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.: 1216 1st Qu.:640.0  Median : 0.000 Median :1.000 Median : 2.000 Median :1.000 Median : 1536 Median :640.0  Mean : 5.086 Mean :0.512 Mean : 2.828 Mean :1.078 Mean : 2774 Mean :616.4  3rd Qu.:13.000 3rd Qu.:1.000 3rd Qu.: 3.000 3rd Qu.:1.000 3rd Qu.: 2530 3rd Qu.:751.0  Max. :17.000 Max. :1.000 Max. :41.000 Max. :6.000 Max. :48388 Max. :952.0  NA's :1  > #glimpse(hosp)  >  > #There is one NA value available in RACE column, which needs to be imputed  > #with the mean value  >  > hosp$RACE[is.na(hosp$RACE)] = round(mean(hosp$RACE, na.rm = T), digits = 0)  >  > #Plot the corr plot to visualize inter-relations if any  > library(MASS)  > library(corrplot)  > corrplot(cor(hosp))  >  > #Below are some of the findings for TOTCHG  > #1 Total Discharge has below correlations  > #1 high +ve co-relation with Length Of Stay  > #2 low -ve co-relation with Diagnosis Groups  > #3 low +ve co-relation with Age  > #There are also co-relations exists with  > #High +ve co-relation between Age and Gender  > #Low -ve co-relation between Age and Length of Stay  > #High +ve co-relation between Age and Diagnosis Groups  > #There is +ve co-relation between Gender and Diagnosis Groups  >  > #Visualize the data and convert non-continuous data to factor variables  > table(hosp$AGE)  0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17  307 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38  > table(hosp$FEMALE)  0 1  244 256  > table(hosp$LOS)  0 1 2 3 4 5 6 7 8 9 10 12 15 17 18 23 24 39 41  15 79 224 98 38 14 8 11 1 1 1 2 1 1 2 1 1 1 1  > table(hosp$RACE)  1 2 3 4 5 6  485 6 1 3 3 2  > table(hosp$TOTCHG)  532 547 548 550 607 610 615 616 622 625 626 627 628 629 634 636 651  2 1 1 1 3 1 1 1 4 2 1 1 2 1 1 1 1  678 707 715 735 736 743 762 763 767 779 783 806 825 833 840 851 869  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  871 874 877 882 887 888 909 921 935 972 1033 1039 1051 1065 1076 1083 1085  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1  1086 1093 1096 1097 1099 1102 1104 1105 1106 1111 1115 1118 1120 1122 1125 1128 1132  1 1 8 1 3 1 1 1 2 1 1 1 1 1 1 2 1  1134 1135 1136 1139 1141 1153 1156 1157 1160 1161 1163 1164 1167 1171 1173 1174 1175  1 1 1 1 1 1 5 1 1 1 1 1 1 2 2 1 1  1176 1178 1183 1191 1193 1194 1199 1200 1211 1214 1217 1220 1224 1226 1231 1235 1236  1 2 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1  1237 1245 1246 1250 1251 1263 1264 1265 1266 1267 1268 1269 1270 1272 1273 1275 1277  1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1  1278 1280 1281 1282 1283 1284 1285 1286 1287 1290 1291 1293 1298 1299 1302 1303 1305  2 4 3 1 2 1 1 2 1 1 1 1 1 1 1 1 1  1306 1307 1309 1310 1317 1320 1321 1322 1324 1330 1335 1337 1340 1348 1351 1362 1363  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2  1368 1369 1375 1379 1381 1382 1383 1385 1393 1395 1396 1404 1407 1408 1409 1411 1413  1 1 1 1 1 1 1 2 2 2 1 1 1 2 1 2 1  1416 1418 1420 1422 1427 1428 1431 1436 1437 1439 1444 1446 1454 1458 1463 1465 1475  1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1  1477 1483 1491 1499 1501 1505 1513 1517 1526 1527 1528 1530 1534 1535 1538 1543 1550  1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1  1561 1576 1578 1581 1603 1607 1620 1628 1647 1651 1656 1660 1669 1674 1685 1688 1689  1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 2 2  1699 1701 1711 1719 1725 1739 1743 1748 1749 1753 1758 1760 1768 1769 1772 1777 1779  1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1  1782 1786 1794 1795 1802 1805 1807 1808 1818 1821 1828 1832 1843 1845 1874 1875 1883  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  1886 1887 1888 1890 1891 1908 1915 1933 1946 1947 1952 1956 1964 1976 1982 1997 2008  1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1  2023 2028 2029 2049 2051 2057 2075 2080 2082 2089 2090 2100 2112 2117 2118 2129 2134  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1  2144 2149 2181 2187 2195 2197 2201 2204 2205 2217 2218 2220 2238 2255 2280 2288 2296  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  2330 2346 2348 2354 2373 2401 2463 2473 2516 2518 2530 2531 2539 2540 2543 2553 2570  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  2582 2592 2595 2632 2660 2711 2724 2735 2777 2805 2825 2835 2840 2844 2847 2860 2877  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  2925 2932 2936 2977 3023 3041 3045 3074 3101 3102 3124 3126 3188 3257 3285 3305 3315  1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1  3342 3497 3578 3588 3609 3624 3625 3656 3674 3782 3800 3814 3865 3908 3969 3977 3994  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  4006 4072 4227 4288 4296 4304 4412 4677 4717 4833 4931 5014 5075 5568 5615 5788 5881  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  5934 6329 6425 6594 6692 6762 6810 6912 7298 7421 7648 7923 8050 8084 8159 8184 8223  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  8398 8439 8631 9129 9230 9530 9606 9932 10002 10431 10562 10584 10585 10649 10756 11125 11685  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  11877 12024 12042 13040 13112 14174 14243 15129 16520 17434 17524 20060 20195 26356 29188 48388  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  > summary(hosp$TOTCHG)  Min. 1st Qu. Median Mean 3rd Qu. Max.  532 1216 1536 2774 2530 48388  > table(hosp$APRDRG)  21 23 49 50 51 53 54 57 58 92 97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317  1 1 1 1 1 10 1 2 1 1 1 1 2 1 4 5 1 1 1 1 2 6 1 1 1 1  344 347 420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754  2 3 2 1 3 2 1 1 1 3 1 3 6 4 2 3 4 267 1 1 2 1 1 14 36 37  755 756 758 760 776 811 812 863 911 930 952  13 2 20 2 1 2 3 1 1 2 1  >  >  > #Age is ordinal categorical variable and needs to be converted to factor  > #Gender is a bivariate categorical variable and to be converted to factor  > #Length of stay ordinal categorical variable and to be converted to factor  > #Race is a categorical variable  > #Discharge cost is continuous variable  > #Diagnosis Group is categorical variable  >  >  >  >  > #1 To record the patient statistics,  > #the agency wants to find the age category of people who frequent the hospital and  > #has the maximum expenditure.  >  > tapply(X = hosp$TOTCHG, INDEX = list(hosp$AGE), FUN = sum)  0 1 2 3 4 5 6 7 8 9 10 11 12 13  678118 37744 7298 30550 15992 18507 17928 10087 4741 21147 24469 14250 54912 31135  14 15 16 17  64643 111747 69149 174777  > names(which.max(tapply(X = hosp$TOTCHG, INDEX = list(hosp$AGE), FUN = sum)))  [1] "0"  >  > #-------------------------------------------------------------------------------------  > #Conclusion  > #Infants spend the max amount of 678118 for the hospitalization  > #-------------------------------------------------------------------------------------  >  > #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.  >  > tapply(X = hosp$TOTCHG, INDEX=list(hosp$APRDRG), FUN = sum)  21 23 49 50 51 53 54 57 58 92 97 114 115 137  10002 14174 20195 3908 3023 82271 851 14509 2117 12024 9530 10562 25832 15129  138 139 141 143 204 206 225 249 254 308 313 317 344 347  13622 17766 2860 1393 8439 9230 25649 16642 615 10585 8159 17524 14802 12597  420 421 422 560 561 566 580 581 602 614 626 633 634 636  6357 26356 5177 4877 2296 2129 2825 7453 29188 27531 23289 17591 9952 23224  639 640 710 720 723 740 750 751 753 754 755 756 758 760  12612 437978 8223 14243 5289 11125 1753 21666 79542 59150 11168 1494 34953 8273  776 811 812 863 911 930 952  1193 3838 9524 13040 48388 26654 4833  >  > names(which.max(tapply(X = hosp$TOTCHG, INDEX=list(hosp$APRDRG), FUN = sum)))  [1] "640"  >  > #-------------------------------------------------------------------------------------  > #Conclusion  > #Diagnosis group#640 spend the max amount of 437978 on the hospitalization  >  > #-------------------------------------------------------------------------------------  >  >  > #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.  >  > aov(TOTCHG~RACE, data = hosp) -> aov\_model  > summary(aov\_model)  Df Sum Sq Mean Sq F value Pr(>F)  RACE 1 2.454e+06 2453726 0.162 0.687  Residuals 498 7.542e+09 15145146  >  > #-------------------------------------------------------------------------------------  > #Conclusion  >  > #pvalue comes out to be very high 68.7% this means we can reject the null hypothesis.  > #this means there is no relation between the race of patient and the hospital cost  > #-------------------------------------------------------------------------------------  >  > #4 To properly utilize the costs, the agency has to analyze the severity of the hospital  > #costs by age and gender for proper allocation of resources.  >  > #Note : Data has not been splitted into Test and Train due to small sample size  > lm\_age\_gender = lm(formula = TOTCHG~AGE + FEMALE ,data = hosp)  > summary(lm\_age\_gender)  Call:  lm(formula = TOTCHG ~ AGE + FEMALE, data = hosp)  Residuals:  Min 1Q Median 3Q Max  -3406 -1443 -869 -152 44951  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2718.63 261.14 10.411 < 2e-16 \*\*\*  AGE 86.28 25.48 3.387 0.000763 \*\*\*  FEMALE -748.19 353.83 -2.115 0.034967 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3845 on 497 degrees of freedom  Multiple R-squared: 0.0261, Adjusted R-squared: 0.02218  F-statistic: 6.66 on 2 and 497 DF, p-value: 0.001399  >  > #-------------------------------------------------------------------------------------  > #Conclusion  > #Null hypothesis is rejected as the pvalue is 0.001  > #Age is significant factors for calculating Total Discharge cost and null hypothesis  > #is rejected at 0.1% significance level  > #Gender is also a significant factor for calculating Total Discharge cost and  > #null hypothesis is rejected at 5% significance level  >  > #linear regression equation will be as given below  >  > #Total Charge = 86.28 \* Age - 748.19 \* Female + 2718.63  > #-------------------------------------------------------------------------------------  >  >  > #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.  >  > #Note : Data has not been splitted into Test and Train due to small sample size  > lm\_age\_gender\_race = lm(formula = LOS~AGE + FEMALE + RACE ,data = hosp)  > summary(lm\_age\_gender\_race)  Call:  lm(formula = LOS ~ AGE + FEMALE + RACE, data = hosp)  Residuals:  Min 1Q Median 3Q Max  -3.213 -1.213 -0.848 0.152 37.787  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.94209 0.39280 7.490 3.18e-13 \*\*\*  AGE -0.03925 0.02227 -1.762 0.0787 .  FEMALE 0.36421 0.30954 1.177 0.2399  RACE -0.09365 0.29286 -0.320 0.7493  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.361 on 496 degrees of freedom  Multiple R-squared: 0.007757, Adjusted R-squared: 0.001756  F-statistic: 1.293 on 3 and 496 DF, p-value: 0.2763  >  > #Model shows the p\_value is 0.2763 which is very high and relation with the Age at 10%  > #significance, but as there is no direct corelation model needsd to be verified by  > #removing Gender and Race parameters  >  > #Note : Data has not been splitted into Test and Train due to small sample size  > lm\_age = lm(formula = LOS~AGE, data = hosp)  > summary(lm\_age)  Call:  lm(formula = LOS ~ AGE, data = hosp)  Residuals:  Min 1Q Median 3Q Max  -2.997 -0.997 -0.997 0.003 38.003  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.99696 0.18621 16.094 <2e-16 \*\*\*  AGE -0.03322 0.02164 -1.535 0.125  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.359 on 498 degrees of freedom  Multiple R-squared: 0.004711, Adjusted R-squared: 0.002713  F-statistic: 2.357 on 1 and 498 DF, p-value: 0.1253  >  > #-------------------------------------------------------------------------------------  > #Conclusion  > #Null hypothesis is retained as the pvalue is 0.2763  > #Age, Gender are Race are not significant factors for calculating Total Discharge charges  > #-------------------------------------------------------------------------------------  >  > #6 To perform a complete analysis, the agency wants to find the variable that mainly affects  > #the hospital costs  >  > #Note : Data has not been splitted into Test and Train due to small sample size  > lm\_tot\_chg = lm(formula = TOTCHG~., data = hosp)  > summary(lm\_tot\_chg)  Call:  lm(formula = TOTCHG ~ ., data = hosp)  Residuals:  Min 1Q Median 3Q Max  -6377 -700 -173 122 43378  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 5218.7658 507.1007 10.291 < 2e-16 \*\*\*  AGE 134.6828 17.4358 7.725 6.3e-14 \*\*\*  FEMALE -390.4928 247.1489 -1.580 0.115  LOS 743.1433 34.8826 21.304 < 2e-16 \*\*\*  RACE -212.4440 227.6998 -0.933 0.351  APRDRG -7.7910 0.6809 -11.442 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2611 on 494 degrees of freedom  Multiple R-squared: 0.5537, Adjusted R-squared: 0.5492  F-statistic: 122.6 on 5 and 494 DF, p-value: < 2.2e-16  >  >  > lm\_age\_los\_aprdrg = lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = hosp)  > summary(lm\_age\_los\_aprdrg)  Call:  lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = hosp)  Residuals:  Min 1Q Median 3Q Max  -6603 -718 -169 123 43350  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 4959.8572 433.1927 11.450 < 2e-16 \*\*\*  AGE 128.5889 17.0670 7.534 2.34e-13 \*\*\*  LOS 740.8349 34.8778 21.241 < 2e-16 \*\*\*  APRDRG -8.0060 0.6636 -12.065 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2614 on 496 degrees of freedom  Multiple R-squared: 0.5508, Adjusted R-squared: 0.5481  F-statistic: 202.7 on 3 and 496 DF, p-value: < 2.2e-16  >  > #-------------------------------------------------------------------------------------  > #Conclusion  > #Null hypothesis is rejected as the pvalue is very small - 2.2 \* e^-16  > #Age, Length of stay and diagnosis group are significant factors for calculating  > #Total Discharge cost at the significance level of 0.1%  >  > #The equation is as given below  >  > #Total cost = 128.59 \* Age + 740.83 \* Length of Stay - 8.01 \* Diag Grp fact + 4959.86  > #------------------------------------------------------------------------------------- |
|  |
| |  | | --- | | > | |