**Healthcare cost Analysis.**

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.**

**Code**

# 1. Find the age category of people who frequently visit the hospital and have the maximum expenditure.

> age\_stats <- hospital\_data %>%

+ group\_by(AGE) %>%

+ summarise(Avg\_Expenditure = mean(TOTCHG), Visits = n()) %>%

+ arrange(desc(Visits), desc(Avg\_Expenditure))

> # Display the results

> print("Age Statistics:")

[1] "Age Statistics:"

> print(age\_stats)

# A tibble: 18 × 3

AGE Avg\_Expenditure Visits

*<dbl>* *<dbl>* *<int>*

1 0 2209. 307

2 17 4599. 38

3 15 3853. 29

4 16 2384. 29

5 14 2586. 25

6 13 1730. 18

7 12 3661. 15

8 1 3774. 10

9 11 1781. 8

10 10 6117. 4

11 3 10183. 3

12 7 3362. 3

13 9 10574. 2

14 5 9254. 2

15 6 8964 2

16 4 7996 2

17 8 2370. 2

18 2 7298 1

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **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.**

**code**

> # 2. Find the diagnosis-related group with maximum hospitalization and expenditure

> max\_aprdrg\_stats <- hospital\_data %>%

+ group\_by(APRDRG) %>%

+ summarise(Total\_Hospitalizations = n(), Total\_Expenditure = sum(TOTCHG)) %>%

+ arrange(desc(Total\_Hospitalizations), desc(Total\_Expenditure))

>

> # Display the results

> print("Diagnosis-Related Group Statistics:")

[1] "Diagnosis-Related Group Statistics:"

> print(max\_aprdrg\_stats)

# A tibble: 63 × 3

APRDRG Total\_Hospitalizations Total\_Expenditure

*<dbl>* *<int>* *<dbl>*

1 640 267 437978

2 754 37 59150

3 753 36 79542

4 758 20 34953

5 751 14 21666

6 755 13 11168

7 53 10 82271

8 626 6 23289

9 249 6 16642

10 139 5 17766

# ℹ 53 more rows

# ℹ Use `print(n = ...)` to see more rows

> # Convert the tibble to a data frame

> max\_aprdrg\_df <- as.data.frame(max\_aprdrg\_stats)

>

> # Display the entire data frame

> print("Diagnosis-Related Group Statistics:")

[1] "Diagnosis-Related Group Statistics:"

> print(max\_aprdrg\_df)

APRDRG Total\_Hospitalizations Total\_Expenditure

1 640 267 437978

2 754 37 59150

3 753 36 79542

4 758 20 34953

5 751 14 21666

6 755 13 11168

7 53 10 82271

8 626 6 23289

9 249 6 16642

10 139 5 17766

11 633 4 17591

12 138 4 13622

13 639 4 12612

14 614 3 27531

15 636 3 23224

16 347 3 12597

17 812 3 9524

18 581 3 7453

19 422 3 5177

20 930 2 26654

21 115 2 25832

22 225 2 25649

23 344 2 14802

24 57 2 14509

25 634 2 9952

26 760 2 8273

27 420 2 6357

28 723 2 5289

29 560 2 4877

30 811 2 3838

31 756 2 1494

32 911 1 48388

33 602 1 29188

34 421 1 26356

35 49 1 20195

36 317 1 17524

37 137 1 15129

38 720 1 14243

39 23 1 14174

40 863 1 13040

41 92 1 12024

42 740 1 11125

43 308 1 10585

44 114 1 10562

45 21 1 10002

46 97 1 9530

47 206 1 9230

48 204 1 8439

49 710 1 8223

50 313 1 8159

51 952 1 4833

52 50 1 3908

53 51 1 3023

54 141 1 2860

55 580 1 2825

56 561 1 2296

57 566 1 2129

58 58 1 2117

59 750 1 1753

60 143 1 1393

61 776 1 1193

62 54 1 851

63 254 1 615

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **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.**

**Code**

|  |
| --- |
| > # Perform ANOVA test  > anova\_result <- aov(TOTCHG ~ RACE, data = hospital\_data)  >  > # Print the summary of the ANOVA test  > print(summary(anova\_result))  Df Sum Sq Mean Sq F value Pr(>F)  RACE 1 2.488e+06 2488459 0.164 0.686  Residuals 497 7.540e+09 15170268  1 observation deleted due to missingness  > |
|  |
| |  | | --- | | > | |

A screenshot of a computer

Description automatically generated

**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.**

**Code**

> # Load necessary libraries

> library(ggplot2)

> library(dplyr)

> # 1. Visualize hospital costs by age and gender

> ggplot(hospital\_data, aes(x = AGE, y = TOTCHG, color = as.factor(FEMALE))) +

+ geom\_point() +

+ geom\_smooth(method = "lm", se = FALSE) +

+ labs(title = "Hospital Costs by Age and Gender",

+ x = "Age",

+ y = "Total Hospital Costs",

+ color = "Gender") +

+ theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'

>

> # 2. Summary statistics by age and gender

> summary\_stats <- hospital\_data %>%

+ group\_by(AGE, FEMALE) %>%

+ summarise(Mean\_Cost = mean(TOTCHG),

+ Median\_Cost = median(TOTCHG),

+ Total\_Cost = sum(TOTCHG),

+ Total\_Hospitalizations = n())

`summarise()` has grouped output by 'AGE'. You can override using the `.groups` argument.

> # Print the summary statistics

> print("Summary Statistics by Age and Gender:")

[1] "Summary Statistics by Age and Gender:"

> print(summary\_stats)

# A tibble: 31 × 6

# Groups: AGE [18]

AGE FEMALE Mean\_Cost Median\_Cost Total\_Cost Total\_Hospitalizations

*<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<int>*

1 0 0 2198. 1496 373734 170

2 0 1 2222. 1436 304384 137

3 1 0 4328. 3357 34622 8

4 1 1 1561 1561 3122 2

5 2 0 7298 7298 7298 1

6 3 0 11164. 11164. 22327 2

7 3 1 8223 8223 8223 1

8 4 0 9230 9230 9230 1

9 4 1 6762 6762 6762 1

10 5 0 7923 7923 7923 1

# ℹ 21 more rows

# ℹ Use `print(n = ...)` to see more rows

> # Print the entire summary statistics table

> print("Summary Statistics by Age and Gender:")

[1] "Summary Statistics by Age and Gender:"

> print(summary\_stats, n = Inf) # Setting n to Inf shows all rows

# A tibble: 31 × 6

# Groups: AGE [18]

AGE FEMALE Mean\_Cost Median\_Cost Total\_Cost Total\_Hospitalizations

*<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<int>*

1 0 0 2198. 1496 373734 170

2 0 1 2222. 1436 304384 137

3 1 0 4328. 3357 34622 8

4 1 1 1561 1561 3122 2

5 2 0 7298 7298 7298 1

6 3 0 11164. 11164. 22327 2

7 3 1 8223 8223 8223 1

8 4 0 9230 9230 9230 1

9 4 1 6762 6762 6762 1

10 5 0 7923 7923 7923 1

11 5 1 10584 10584 10584 1

12 6 0 8964 8964 17928 2

13 7 0 3362. 2530 10087 3

14 8 0 2370. 2370. 4741 2

15 9 0 10574. 10574. 21147 2

16 10 0 7770. 2925 23309 3

17 10 1 1160 1160 1160 1

18 11 0 1468 1106. 8808 6

19 11 1 2721 2721 5442 2

20 12 0 2592. 1559 15553 6

21 12 1 4373. 1647 39359 9

22 13 0 1054 1168 4216 4

23 13 1 1923. 1304 26919 14

24 14 0 5741 5765 22964 4

25 14 1 1985. 1444 41679 21

26 15 0 7223 3414. 72230 10

27 15 1 2080. 1268 39517 19

28 16 0 4630. 3060. 27779 6

29 16 1 1799. 1275 41370 23

30 17 0 3961. 3285 51495 13

31 17 1 4931. 2129 123282 25

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **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.**

**Code**

|  |
| --- |
| > # Load necessary libraries  > library(dplyr)  > library(tidyr)  Error in library(tidyr) : there is no package called ‘tidyr’  > install.packages("tidyr")  WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:  https://cran.rstudio.com/bin/windows/Rtools/  Installing package into ‘C:/Users/RONIT DEKA/AppData/Local/R/win-library/4.3’  (as ‘lib’ is unspecified)  also installing the dependencies ‘purrr’, ‘stringr’  trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/purrr\_1.0.2.zip'  Content type 'application/zip' length 498865 bytes (487 KB)  downloaded 487 KB  trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/stringr\_1.5.1.zip'  Content type 'application/zip' length 319072 bytes (311 KB)  downloaded 311 KB  trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/tidyr\_1.3.0.zip'  Content type 'application/zip' length 1281275 bytes (1.2 MB)  downloaded 1.2 MB  package ‘purrr’ successfully unpacked and MD5 sums checked  package ‘stringr’ successfully unpacked and MD5 sums checked  package ‘tidyr’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\RONIT DEKA\AppData\Local\Temp\RtmpOEqKCv\downloaded\_packages  > # Load necessary libraries  > library(dplyr)  > library(tidyr)  > library(ggplot2)  > # Fit a linear regression model  > lm\_model <- lm(LOS ~ AGE + FEMALE + RACE, data = hospital\_data)  >  > # Print the summary of the linear regression model  > print(summary(lm\_model))  Call:  lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital\_data)  Residuals:  Min 1Q Median 3Q Max  -3.22 -1.22 -0.85 0.15 37.78  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.94377 0.39318 7.487 3.25e-13 \*\*\*  AGE -0.03960 0.02231 -1.775 0.0766 .  FEMALE 0.37011 0.31024 1.193 0.2334  RACE -0.09408 0.29312 -0.321 0.7484  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.363 on 495 degrees of freedom  (1 observation deleted due to missingness)  Multiple R-squared: 0.007898, Adjusted R-squared: 0.001886  F-statistic: 1.314 on 3 and 495 DF, p-value: 0.2692  >  > # Visualize the relationship between actual and predicted length of stay  > predicted\_values <- predict(lm\_model, newdata = hospital\_data)  > actual\_vs\_predicted <- data.frame(Actual = hospital\_data$LOS, Predicted = predicted\_values)  >  > ggplot(actual\_vs\_predicted, aes(x = Actual, y = Predicted)) +  + geom\_point() +  + geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +  + labs(title = "Actual vs. Predicted Length of Stay",  + x = "Actual Length of Stay",  + y = "Predicted Length of Stay") +  + theme\_minimal()  Warning message:  Removed 1 rows containing missing values (`geom\_point()`).  > # Load necessary libraries  > library(dplyr)  > library(tidyr)  > library(ggplot2)  >  > # Fit a linear regression model  > lm\_model <- lm(LOS ~ AGE + FEMALE + RACE, data = hospital\_data)  >  > # Print the summary of the linear regression model  > print(summary(lm\_model))  Call:  lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital\_data)  Residuals:  Min 1Q Median 3Q Max  -3.22 -1.22 -0.85 0.15 37.78  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.94377 0.39318 7.487 3.25e-13 \*\*\*  AGE -0.03960 0.02231 -1.775 0.0766 .  FEMALE 0.37011 0.31024 1.193 0.2334  RACE -0.09408 0.29312 -0.321 0.7484  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.363 on 495 degrees of freedom  (1 observation deleted due to missingness)  Multiple R-squared: 0.007898, Adjusted R-squared: 0.001886  F-statistic: 1.314 on 3 and 495 DF, p-value: 0.2692  >  > # Generate predicted values  > predicted\_values <- predict(lm\_model, newdata = hospital\_data)  >  > # Create a data frame with Actual and Predicted values  > actual\_vs\_predicted <- data.frame(Actual = hospital\_data$LOS, Predicted = predicted\_values)  >  > # Remove rows with missing values  > actual\_vs\_predicted <- na.omit(actual\_vs\_predicted)  >  > # Visualize the relationship between actual and predicted length of stay  > ggplot(actual\_vs\_predicted, aes(x = Actual, y = Predicted)) +  + geom\_point() +  + geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +  + labs(title = "Actual vs. Predicted Length of Stay",  + x = "Actual Length of Stay",  + y = "Predicted Length of Stay") +  + theme\_minimal()  > |
|  |
| |  | | --- | | > | |

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

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

**Code**

# Fit a linear regression model

> lm\_model <- lm(TOTCHG ~ AGE + FEMALE + LOS + RACE, data = hospital\_data)

>

> # Print the summary of the linear regression model

> print(summary(lm\_model))

Call:

lm(formula = TOTCHG ~ AGE + FEMALE + LOS + RACE, data = hospital\_data)

Residuals:

Min 1Q Median 3Q Max

-4363 -1114 -644 137 41631

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 735.86 362.16 2.032 0.042699 \*

AGE 115.79 19.54 5.925 5.86e-09 \*\*\*

FEMALE -1027.30 271.23 -3.788 0.000171 \*\*\*

LOS 742.04 39.24 18.911 < 2e-16 \*\*\*

RACE -114.14 255.92 -0.446 0.655802

---

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

Residual standard error: 2936 on 494 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.4353, Adjusted R-squared: 0.4307

F-statistic: 95.19 on 4 and 494 DF, p-value: < 2.2e-16

Result -

The output from the linear regression model summary provides information about the coefficients, their standard errors, t-values, and p-values. Here's an interpretation of the results:

1. **Intercept (Intercept):**
   * Estimate: 735.86
   * This is the estimated intercept when all other predictors are zero. In this context, it may represent the baseline cost when the age, gender, length of stay, and race are zero (though this might not have a practical interpretation).
2. **AGE:**
   * Estimate: 115.79
   * The coefficient indicates that, on average, for each one-unit increase in age, the hospital costs increase by $115.79.
3. **FEMALE:**
   * Estimate: -1027.30
   * The coefficient indicates that, on average, being female is associated with a decrease in hospital costs by $1027.30 compared to being male.
4. **LOS (Length of Stay):**
   * Estimate: 742.04
   * The coefficient indicates that, on average, for each one-unit increase in the length of stay, the hospital costs increase by $742.04.
5. **RACE:**
   * Estimate: -114.14
   * The coefficient for race suggests that, on average, there is a decrease in hospital costs by $114.14 for each one-unit increase in the race variable.

In terms of significance:

* AGE, FEMALE, and LOS have low p-values (indicated by '\*\*\*'), suggesting they are statistically significant predictors of hospital costs.
* RACE has a higher p-value (0.655802), indicating it is not a statistically significant predictor.

The adjusted R-squared value is 0.4307, which suggests that the model explains about 43.07% of the variability in hospital costs.

Overall, based on the p-values, AGE, FEMALE, and LOS seem to be important predictors of hospital costs in this model. The RACE variable, however, does not appear to have a statistically significant impact on hospital costs in this analysis.

A screenshot of a computer

Description automatically generated