Double-click (or enter) to edit

```
# -------------------------------------------------
# R Code for Healthcare Cost Analysis
# -------------------------------------------------

# Install necessary packages
install.packages("dplyr")
install.packages("ggplot2")
install.packages("readxl")  # <-- ADDED THIS PACKAGE

library(dplyr)
library(ggplot2)
library(readxl)   # <-- ADDED THIS LIBRARY

# -------------------------------------------------
# 0. Load and Clean Data
# -------------------------------------------------

file_path <- "/content/1555054100_hospitalcosts.xlsx"

# --- OLD CODE ---
# df <- read.csv(file_path)
# --- NEW CODE ---
df <- read_excel(file_path) # <-- THIS IS THE FIX

# Clean the data (remove rows with NA values)
df_cleaned <- na.omit(df)

cat("--- Initial Data Structure ---\n")
str(df_cleaned)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

--- Initial Data Structure ---
tibble [499 × 6] (S3: tbl_df/tbl/data.frame)
 $ AGE   : num [1:499] 17 17 17 17 17 17 17 16 16 17 ...
 $ FEMALE: num [1:499] 1 0 1 1 1 0 1 1 1 1 ...
 $ LOS   : num [1:499] 2 2 7 1 1 0 4 2 1 2 ...
 $ RACE  : num [1:499] 1 1 1 1 1 1 1 1 1 1 ...
 $ TOTCHG: num [1:499] 2660 1689 20060 736 1194 ...
 $ APRDRG: num [1:499] 560 753 930 758 754 347 754 754 753 758 ...
 - attr(*, "na.action")= 'omit' Named int 277
  ..- attr(*, "names")= chr "277"
```

Double-click (or enter) to edit

```
# -------------------------------------------------
# 1. Task 1: Patient Statistics by Age
# -------------------------------------------------
cat("\n--- Task 1: Age Category Statistics ---\n")

age_analysis <- df_cleaned %>%
  group_by(AGE) %>%
  summarise(
    TotalExpenditure = sum(TOTCHG),
    VisitCount = n()
  ) %>%
  arrange(desc(VisitCount)) # Sort by visit count

# Print the full analysis
print(age_analysis, n=20) # n=20 to make sure all rows are printed

# Find the age with maximum expenditure
max_exp_age <- age_analysis[which.max(age_analysis$TotalExpenditure), ]
cat("\nAge category with maximum expenditure:\n")
print(max_exp_age)
```

```
# Find the age with most frequent visits
max_visit_age <- age_analysis[which.max(age_analysis$VisitCount), ]
cat("\nAge category with most frequent visits:\n")
print(max_visit_age)
```

```
--- Task 1: Age Category Statistics ---
# A tibble: 18 × 3
     AGE TotalExpenditure VisitCount
   <dbl>            <dbl>      <int>
 1     0           676962        306
 2    17           174777         38
 3    15           111747         29
 4    16            69149         29
 5    14            64643         25
 6    13            31135         18
 7    12            54912         15
 8     1            37744         10
 9    11            14250          8
10    10            24469          4
11     3            30550          3
12     7            10087          3
13     4            15992          2
14     5            18507          2
15     6            17928          2
16     8             4741          2
17     9            21147          2
18     2             7298          1

Age category with maximum expenditure:
# A tibble: 1 × 3
    AGE TotalExpenditure VisitCount
  <dbl>            <dbl>      <int>
1     0           676962        306

Age category with most frequent visits:
# A tibble: 1 × 3
    AGE TotalExpenditure VisitCount
  <dbl>            <dbl>      <int>
1     0           676962        306
```

```
# --------------------------------------------------
# 2. Task 2: Diagnosis Group (APRDRG) Analysis
# --------------------------------------------------
cat("\n--- Task 2: Diagnosis Group (APRDRG) Statistics ---\n")

aprdrg_analysis <- df_cleaned %>%
  group_by(APRDRG) %>%
  summarise(
    TotalExpenditure = sum(TOTCHG),
    VisitCount = n()
  )

# Find the APRDRG with maximum expenditure
max_exp_aprdrg <- aprdrg_analysis[which.max(aprdrg_analysis$TotalExpenditure), ]
cat("\nDiagnosis group (APRDRG) with maximum expenditure:\n")
print(max_exp_aprdrg)

# Find the APRDRG with maximum hospitalization (visits)
max_visit_aprdrg <- aprdrg_analysis[which.max(aprdrg_analysis$VisitCount), ]
cat("\nDiagnosis group (APRDRG) with maximum hospitalization:\n")
print(max_visit_aprdrg)
```

```
--- Task 2: Diagnosis Group (APRDRG) Statistics ---

Diagnosis group (APRDRG) with maximum expenditure:
# A tibble: 1 × 3
  APRDRG TotalExpenditure VisitCount
   <dbl>            <dbl>      <int>
1    640           436822        266

Diagnosis group (APRDRG) with maximum hospitalization:
# A tibble: 1 × 3
  APRDRG TotalExpenditure VisitCount
   <dbl>            <dbl>      <int>
1    640           436822        266
```

```
# --------------------------------------------------
# 3. Task 3: Race and Hospitalization Costs
```

```
# -------------------------------------------------
cat("\n--- Task 3: Analysis of Race and Hospitalization Costs (ANOVA) ---\n")

# Convert RACE to a factor (a categorical variable) for ANOVA
# This tells R to treat the race codes (1, 2, 3...) as separate groups
df_cleaned$RACE <- as.factor(df_cleaned$RACE)

# Perform the one-way ANOVA test
# We are testing if TOTCHG changes ~ based on RACE
anova_result <- aov(TOTCHG ~ RACE, data = df_cleaned)

cat("\n--- ANOVA Test Summary ---\n")
print(summary(anova_result))
```

```
--- Task 3: Analysis of Race and Hospitalization Costs (ANOVA) ---

--- ANOVA Test Summary ---
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
```

```
# -------------------------------------------------
# 4. Task 4: Cost Severity by Age and Gender
# -------------------------------------------------
cat("\n--- Task 4: Analysis of Cost Severity by Age and Gender ---\n")

# Calculate mean costs
age_gender_analysis <- df_cleaned %>%
  group_by(AGE, FEMALE) %>%
  summarise(Mean_Cost = mean(TOTCHG), .groups = 'drop')

# Recode FEMALE to 'Male' and 'Female' for the plot
age_gender_analysis$Gender <- ifelse(age_gender_analysis$FEMALE == 0, "Male", "Female")

# Print the table
cat("\n--- Mean Costs by Age and Gender Table ---\n")
print(age_gender_analysis, n = 40) # Print all rows

# Create a grouped bar chart
age_gender_chart <- ggplot(age_gender_analysis, aes(x = factor(AGE), y = Mean_Cost, fill = Gender)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Mean Hospital Costs by Age and Gender",
       x = "Age",
       y = "Mean Hospital Cost ($)",
       fill = "Gender") +
  theme_minimal()

# Save the chart as a PNG file
ggsave("age_gender_cost_analysis_bar_chart.png", plot = age_gender_chart)

cat("\nChart 'age_gender_cost_analysis_bar_chart.png' has been saved to your Colab files.\n")
```

```
--- Task 4: Analysis of Cost Severity by Age and Gender ---

--- Mean Costs by Age and Gender Table ---
# A tibble: 31 × 4
     AGE FEMALE Mean_Cost Gender
   <dbl>  <dbl>     <dbl> <chr>
 1     0      0     2198. Male
 2     0      1     2230. Female
 3     1      0     4328. Male
 4     1      1     1561  Female
 5     2      0     7298  Male
 6     3      0    11164. Male
 7     3      1     8223  Female
 8     4      0     9230  Male
 9     4      1     6762  Female
10     5      0     7923  Male
11     5      1    10584  Female
12     6      0     8964  Male
13     7      0     3362. Male
14     8      0     2370. Male
15     9      0    10574. Male
16    10      0     7770. Male
17    10      1     1160  Female
18    11      0     1468  Male
19    11      1     2721  Female
```

```
20    12      0     2592. Male
21    12      1     4373. Female
22    13      0     1054  Male
23    13      1     1923. Female
24    14      0     5741  Male
25    14      1     1985. Female
26    15      0     7223  Male
27    15      1     2080. Female
28    16      0     4630. Male
29    16      1     1799. Female
30    17      0     3961. Male
31    17      1     4931. Female
Saving 6.67 x 6.67 in image

Chart 'age_gender_cost_analysis_bar_chart.png' has been saved to your Colab files.
```

```
# -------------------------------------------------
# 5. Task 5: Predicting Length of Stay (LOS)
# -------------------------------------------------
cat("\n--- Task 5: Predicting Length of Stay (LOS) ---\n")

# Build a linear regression model
# We are testing: LOS ~ AGE + FEMALE + RACE
# We must treat FEMALE and RACE as categorical factors
# Note: df_cleaned$RACE was already converted to a factor in Task 3
los_model <- lm(LOS ~ AGE + as.factor(FEMALE) + RACE, data = df_cleaned)

# Print the model summary
cat("\n--- OLS Regression Model Summary for Predicting LOS ---\n")
print(summary(los_model))
```

```
--- Task 5: Predicting Length of Stay (LOS) ---

--- OLS Regression Model Summary for Predicting LOS ---

Call:
lm(formula = LOS ~ AGE + as.factor(FEMALE) + RACE, data = df_cleaned)

Residuals:
   Min     1Q Median     3Q    Max 
-3.211 -1.211 -0.857  0.143 37.789 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.85687    0.23160  12.335   <2e-16 ***
AGE                -0.03938    0.02258  -1.744   0.0818 .  
as.factor(FEMALE)1  0.35391    0.31292   1.131   0.2586    
RACE2              -0.37501    1.39568  -0.269   0.7883    
RACE3               0.78922    3.38581   0.233   0.8158    
RACE4               0.59493    1.95716   0.304   0.7613    
RACE5              -0.85687    1.96273  -0.437   0.6626    
RACE6              -0.71879    2.39295  -0.300   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
```

```
# -------------------------------------------------
# 6. Task 6: Main Variable Affecting Hospital Costs
# -------------------------------------------------
cat("\n--- Task 6: Finding Variable that Mainly Affects Hospital Costs ---\n")

# Build a comprehensive linear regression model
# We are testing: TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG
# We must treat FEMALE, RACE, and APRDRG as categorical factors
# Note: df_cleaned$RACE was already a factor from Task 3

cost_model <- lm(TOTCHG ~ AGE + as.factor(FEMALE) + LOS + RACE + as.factor(APRDRG),
                 data = df_cleaned)

# Print the model summary
cat("\n--- OLS Regression Model Summary for Predicting Hospital Costs (TOTCHG) ---\n")
print(summary(cost_model))
```

```
--- Task 6: Finding Variable that Mainly Affects Hospital Costs ---
```

```
--- OLS Regression Model Summary for Predicting Hospital Costs (TOTCHG) ---

Call:
lm(formula = TOTCHG ~ AGE + as.factor(FEMALE) + LOS + RACE + 
    as.factor(APRDRG), data = df_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-5403.7  -188.8   -52.0   113.5  5403.7 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7017.4364   966.0317   7.264 1.79e-12 ***
AGE                     86.5944    20.7881   4.166 3.76e-05 ***
as.factor(FEMALE)1    -136.8780    78.7821  -1.737 0.083032 .  
LOS                    664.6593    21.2924  31.216  < 2e-16 ***
RACE2                  269.7343   408.6436   0.660 0.509563    
RACE3                  641.3334   862.2531   0.744 0.457413    
RACE4                  106.4079   458.4198   0.232 0.816557    
RACE5                 1577.1875   908.2736   1.736 0.083201 .  
RACE6                  -73.8266   566.3145  -0.130 0.896340    
as.factor(APRDRG)23   4355.1399  1182.4224   3.683 0.000260 ***
as.factor(APRDRG)49   7890.6917  1187.2479   6.646 9.18e-11 ***
as.factor(APRDRG)50  -5254.4156  1194.6819  -4.398 1.38e-05 ***
as.factor(APRDRG)51  -7323.6414  1184.2871  -6.184 1.46e-09 ***
as.factor(APRDRG)53  -1199.9825   954.2018  -1.258 0.209230    
as.factor(APRDRG)54  -8166.3229  1184.4591  -6.895 1.95e-11 ***
as.factor(APRDRG)57   -860.5678  1081.7666  -0.796 0.426752    
as.factor(APRDRG)58  -5651.6901  1238.0309  -4.565 6.54e-06 ***
as.factor(APRDRG)92   3042.9880  1184.6409   2.569 0.010546 *  
as.factor(APRDRG)97     -0.9807  1211.2219  -0.001 0.999354    
as.factor(APRDRG)114   771.2360  1199.1537   0.643 0.520471    
as.factor(APRDRG)115  2529.0158  1063.9012   2.377 0.017887 *  
as.factor(APRDRG)137   135.6525  1262.5545   0.107 0.914488    
as.factor(APRDRG)138 -4574.7058  1042.1335  -4.390 1.43e-05 ***
as.factor(APRDRG)139 -4931.6448   985.5923  -5.004 8.23e-07 ***
as.factor(APRDRG)141 -6352.6992  1195.6625  -5.313 1.74e-07 ***
as.factor(APRDRG)143 -8530.9425  1540.3888  -5.538 5.34e-08 ***
as.factor(APRDRG)204 -2044.5193  1182.5513  -1.729 0.084547 .  
as.factor(APRDRG)206  -127.7919  1220.9720  -0.105 0.916691    
as.factor(APRDRG)225   895.8186  1049.4715   0.854 0.393810    
as.factor(APRDRG)249 -5315.2746   997.4554  -5.329 1.60e-07 ***
as.factor(APRDRG)254 -7979.6240  1540.5665  -5.180 3.43e-07 ***
as.factor(APRDRG)308  2123.5545  1199.1936   1.771 0.077303 .  
as.factor(APRDRG)313 -1178.3407  1110.6267  -1.061 0.289302    
as.factor(APRDRG)317  4988.0046  1200.2669   4.156 3.92e-05 ***
as.factor(APRDRG)344 -2162.1842  1056.0034  -2.048 0.041217 *  
as.factor(APRDRG)347 -3802.5781  1012.6388  -3.755 0.000197 ***
as.factor(APRDRG)420 -6004.9500  1049.0367  -5.724 1.96e-08 ***
as.factor(APRDRG)421 -6583.1473  1473.4757  -4.468 1.01e-05 ***
as.factor(APRDRG)422 -7058.7682  1015.5830  -6.950 1.37e-11 ***
as.factor(APRDRG)560 -7243.4821  1045.9573  -6.925 1.60e-11 ***
as.factor(APRDRG)561 -8455.5174  1188.4307  -7.115 4.75e-12 ***
as.factor(APRDRG)566 -7552.9821  1184.1817  -6.378 4.65e-10 ***
as.factor(APRDRG)580 -4857.0957  1244.3640  -3.903 0.000110 ***
```