Final-Project–3-.R

tarunbatchu

2024-05-15

```
# Load the RMySQL library for database operations
library(RMySQL)
```

```
## Loading required package: DBI
```

```
# Establish a connection to the MySQL database named 'Final'
mysqlconnection <- dbConnect(RMySQL::MySQL(),
                             dbname='Final',
                             host='localhost',
                             port=3306,
                             user='root',
                             password='cmsc2024')

# Prepare an SQL query to clean data, categorize players, and rank them based on average points
analysis_query <- "
WITH CleanedData AS (
    SELECT player_name,
           COALESCE(pts, 0) as pts,  # Replace NULL points with 0
           COALESCE(usg_pct, 0) as usg_pct  # Replace NULL usage percentage with 0
    FROM all_seasons
),
PlayerAveragePoints AS (
    SELECT player_name, AVG(pts) as avg_pts  # Calculate average points for each player
    FROM CleanedData
    GROUP BY player_name
),
PlayerCategories AS (
    SELECT player_name,
           avg_pts,
           CASE  # Categorize players based on their average points
               WHEN avg_pts > 24 THEN 'Superstar'
               WHEN avg_pts > 18 AND avg_pts <= 24 THEN 'Good Player'
               WHEN avg_pts > 8 AND avg_pts <= 18 THEN 'Role Player'
               ELSE 'Bench Player'
           END AS player_category
    FROM PlayerAveragePoints
),
PlayerRanks AS (
    SELECT pc.player_name,
           pc.avg_pts,
           cd.usg_pct,
           RANK() OVER(PARTITION BY pc.player_category ORDER BY pc.avg_pts DESC) as player_rank,  # Rank players within each category
           pc.player_category
    FROM PlayerCategories pc
    JOIN CleanedData cd ON pc.player_name = cd.player_name
)
SELECT player_name,
       avg_pts,
       usg_pct,
       player_rank,
       player_category,
       NTILE(4) OVER(ORDER BY avg_pts DESC) as performance_quartile  # Divide players into quartiles based on average points
FROM PlayerRanks
"
```

```
# Execute the SQL query and store the results in a dataframe
analysis_result <- dbSendQuery(mysqlconnection, analysis_query)
analysis_df <- fetch(analysis_result, n = -1)  # Fetch all rows from the result set
dbClearResult(analysis_result)  # Clear the result set
```

```
## [1] TRUE
```

```
# Convert the player_category column to a factor with ordered levels
analysis_df$player_category <- factor(analysis_df$player_category,
                                      levels = c('Superstar', 'Good Player', 'Role Playe
r', 'Bench Player'))

# Perform a linear regression analysis to study the relationship between average points
and usage percentage
lm_model <- lm(avg_pts ~ usg_pct + player_rank, data = analysis_df)
summary(lm_model)  # Display the summary of the linear model
```

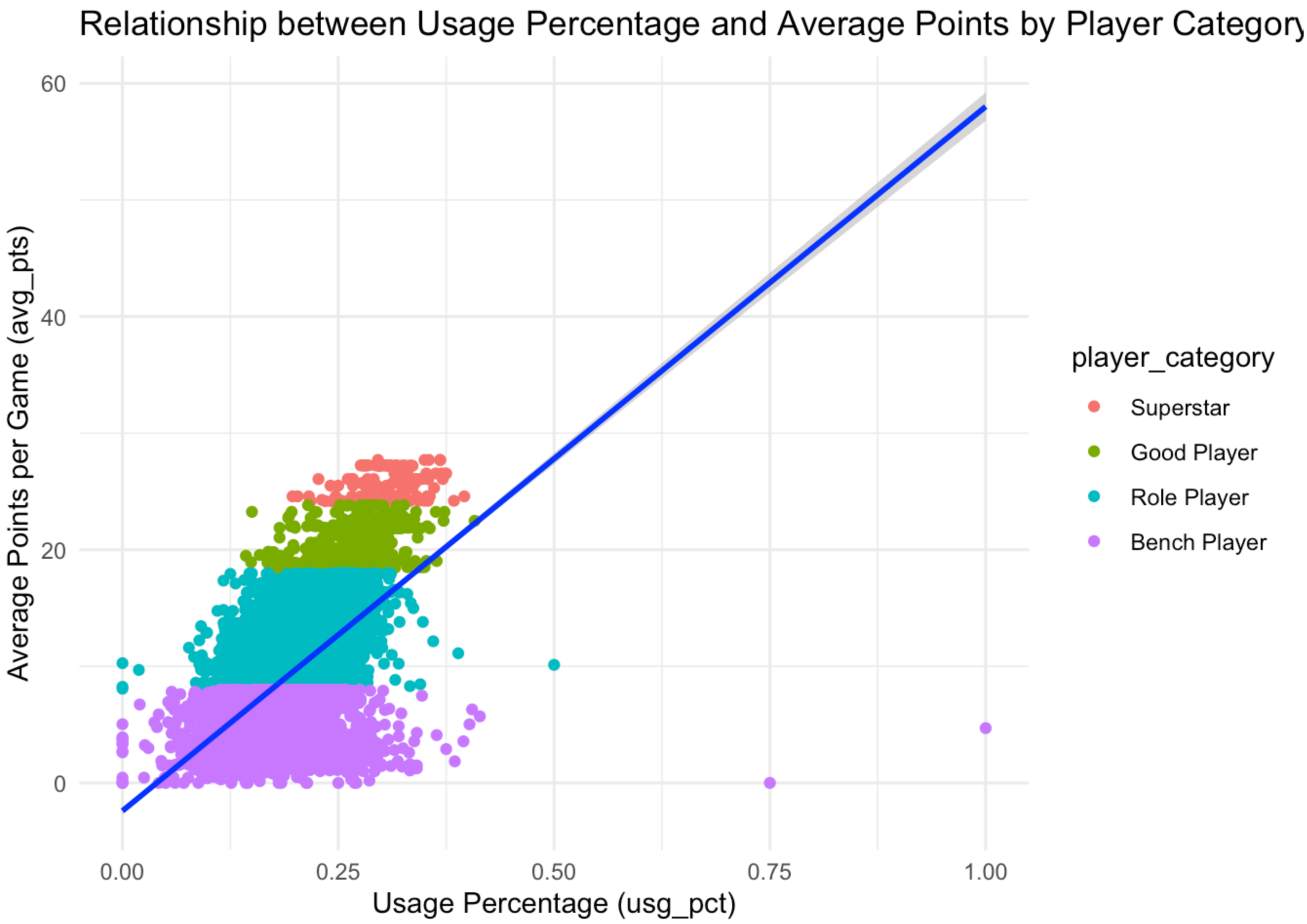
```
##
## Call:
## lm(formula = avg_pts ~ usg_pct + player_rank, data = analysis_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.887  -2.499   0.200   2.379  11.571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.823e+00  1.574e-01   37.01   <2e-16 ***
## usg_pct      3.933e+01  6.584e-01   59.74   <2e-16 ***
## player_rank -1.822e-03  2.305e-05  -79.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.232 on 10427 degrees of freedom
## Multiple R-squared:  0.6101, Adjusted R-squared:   0.61
## F-statistic:  8157 on 2 and 10427 DF,  p-value: < 2.2e-16
```

```
print("Hypothesis: I believe that as the usage percentage increases, we find that player
s tend to average more points.")
```

```
## [1] "Hypothesis: I believe that as the usage percentage increases, we find that playe
rs tend to average more points."
```

```
# Create a scatter plot to visualize the relationship between usage percentage and average points
library(ggplot2)  # Load the ggplot2 library for plotting
ggplot(analysis_df, aes(x = usg_pct, y = avg_pts, color = player_category)) +
  geom_point() +  # Add points to the plot
  geom_smooth(method = "lm", col = "blue") +  # Add a linear regression line
  theme_minimal() +  # Use a minimal theme for the plot
  labs(title = "Relationship between Usage Percentage and Average Points by Player Category",
       x = "Usage Percentage (usg_pct)",
       y = "Average Points per Game (avg_pts)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Disconnect from the MySQL database
dbDisconnect(mysqlconnection)
```

```
## [1] TRUE
```

```
print("In general, I saw that as usage rate increases, the number of points typically in
crease as well. I also noticed that superstars had higher usage rates than the other typ
es of players. This makes sense as superstars, who tend to be much better players, have
more control of the ball being the better players on the court.")
```

```
## [1] "In general, I saw that as usage rate increases, the number of points typically i
ncrease as well. I also noticed that superstars had higher usage rates than the other ty
pes of players. This makes sense as superstars, who tend to be much better players, have
more control of the ball being the better players on the court."
```