Project Proposal: Data Analytics Guidance

Step 1: Project Preparation

Client and Dataset Selection:

For this project, I have selected the SportStats dataset. SportStats is a sports analysis firm partnering with local news and elite personal trainers to provide insightful analytics. The dataset includes two CSV files:

noc_regions.csv – Contains columns: NOC, region, and notes.
athlete_events.csv – Contains columns: ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal.

The reason for selecting this dataset is that it allows for uncovering trends in sports performance, athlete demographics, and country-based performance insights, which can be leveraged for news stories or health-related insights.

Data Import and Cleaning Steps:

Data Acquisition: Imported the datasets into a SQL database using SQL scripts and DBeaver for query execution.

Data Cleaning:

- Handled missing values using SQL queries such as COALESCE() for default values.
- Removed duplicate entries using DELETE with ROW_NUMBER() OVER() partitions.
- Standardized categorical values using CASE statements for consistency (e.g., country names, event types).
- Converted date columns using CAST() and FORMAT() for proper analysis.

Initial Exploration:

- Used SQL aggregate functions (AVG, COUNT, MAX, MIN) to generate descriptive statistics.
- Identified outliers using SQL queries with standard deviation calculations.
- Conducted correlation analysis using JOINS and GROUP BY clauses.
- Entity Relationship Diagram (ERD): The proposed ERD includes:
- Athlete Events Table: Contains athlete ID, name, sex, age, height, weight, team, NOC (links to noc_regions), games, year, season, city, sport, event, and medal.
- NOC Regions Table: Contains NOC, region, and additional notes.

Step 2: Analysis Plan

Project Description:

This project aims to analyze sports performance trends and athlete statistics to provide actionable insights for sports analysts, journalists, and personal trainers. These findings can be used to develop compelling news stories, improve training programs, and identify key health trends in sports.

Questions to Answer:

What are the key performance trends across different sports and events?
How do athlete demographics (age, gender, height, weight, country) influence performance outcomes?
Are there seasonal or event-based patterns that affect athlete performance?

Initial Hypotheses:

Younger athletes tend to outperform older athletes in endurance-based sports.
Certain countries excel in specific sports due to training culture and infrastructure.
Athlete performance fluctuates based on seasonal training cycles and event schedules.

Approach:

Feature Selection:

Key fields: athlete demographics (age, gender, height, weight, country), event details (sport, season, year, city), and performance metrics (medals won).

Exploratory Data Analysis (EDA):

- Used SQL queries in DBeaver to generate frequency distributions, trend analysis, and aggregate statistics.
- Conducted segmentation analysis using GROUP BY and HAVING clauses.
- Statistical Testing & Modeling:
- Used SQL correlation techniques (e.g., JOIN and GROUP BY for trend detection).
- Applied time-series analysis using SQL window functions (e.g., LAG(), LEAD()).
- Considered predictive modeling using stored procedures for deeper insights.

Validation:

Evaluated findings using SQL hypothesis testing methods.
Cross-validated any predictive insights using sample dataset partitions.

Step 3: Initial Data Exploration & Findings

Descriptive Statistics & Analysis:

Summary Statistics:

- Used COUNT(), AVG(), MIN(), and MAX() functions to understand distributions of age, height, weight, and medal counts.
- Identified missing values in key columns like Age, Height, and Weight using WHERE clauses.
- Analyzed the most frequent sports and events using GROUP BY Sport, Event.

Key Findings:

- The average age of athletes varies significantly by sport, with endurance-based sports having younger participants.
- Male athletes have a higher participation rate overall, but some sports exhibit near gender parity.
- Certain countries dominate specific sports, suggesting specialized training programs and sports culture.

Hypothesis Validation:

- Proved: Younger athletes tend to perform better in endurance sports.
- Partially Proved: Certain countries dominate specific sports, but there are exceptions.
- Disproved: Seasonal trends in performance were less significant than expected.

Next Steps & Additional Questions:

Investigate the impact of height and weight on medal-winning probabilities.
Explore how host country advantage influences athlete performance.

Analyze participation trends over different Olympic years to identify long-term shifts.

Step 4: Deeper Analysis (Milestone 3)

Correlations & Deeper Insights:

Correlations Discovered:

There is a strong correlation between athlete age and performance in endurance-based sports (negative correlation: younger athletes perform better in endurance events like swimming and gymnastics).

Certain countries have a higher medal efficiency (total medals per athlete), indicating strong training programs (e.g., USA, Russia, and China).

Expanded Features:

- Looking at medal efficiency (total medals per athlete per country) to assess training success.
- Examining the relationship between athlete height and success in certain sports, such as basketball and rowing.

New Metrics Introduced:

- Medal Efficiency Score = Total Medals / Total Athletes per Country
- This metric helps identify which countries have the most successful training programs.
- Performance Consistency Index = Number of medals won per Olympic year for top-performing countries.

This will show how consistent certain countries are in securing medals over time.

Visualization with R Console:

- Used R Console for data visualization, generating bar charts and plots to illustrate trends.
- Created medal distribution visualizations to compare top-performing countries.
- Generated age distribution graphs to analyze athlete demographics in different sports.

Next Steps:

- Compare male and female athlete performance across different sports.
- Explore how geopolitical factors (e.g., country size, GDP) impact sports success.
- Refine findings and prepare final reporting.

This document will continue evolving as further insights emerge, with a focus on refining conclusions and recommendations for the final presentation.