**Topic: Predicting the Popularity of Chinese Pop Songs in the International Market**

**Step 1: Data Generation** Primarily sourced from the internet, other forms can also be considered.

**Step 2: Data Collection** Utilize web scraping to obtain the following data on Chinese pop songs from the past 30 years:

* Song title
* Singer
* Composer
* Lyricist
* Arranger
* Year
* Genre (from major music platforms)
* BPM
* Key
* Data from the sheet music (mostly images)

**Step 3: Data Storage**

1. Store as song\_data.csv file
2. Store as song\_data.db file

**Step 4: Data Preprocessing**

* Handle missing values (manually fill if the number is small)
* Standardize or normalize the data
* Encode categorical variables (e.g., genre, key) using methods like One-Hot Encoding Preprocessing languages: Python, SQLite

**Step 5: Data Analysis** Consider two types of models: supervised learning and unsupervised learning:

1. **Regression to Predict:**
   * **X:** Obtain a score (1-100) based on factors affecting popularity
   * **Y:** Obtain a score (1-100) based on popularity criteria
   * Each song will be a point on a coordinate system, using regression to fit a function to predict the popularity of Chinese pop songs in the international market
2. **Machine Learning Prediction:** Possible algorithms:
   * Decision Tree
   * Random Forest
   * Gradient Boosting Machine (GBM)
   * Neural Networks
   * Split the dataset into training and test sets, train the model with the training set. Common methods include cross-validation. Evaluate the model's performance using the test set. Common evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R², etc. Optimize the model's performance by tuning hyperparameters, increasing data volume, or trying different models.

**Step 6: Data Visualization** Tools: Matlab, R, Excel, Python, Tableau

**Project Objectives: Predicting the Finalists for the Men's 100m Sprint at the Paris Olympics**

**Assumptions:**

* Ignore differences in athletes' health status.
* Ignore external factors such as competition venues and weather.

**Model Construction:**

* **Target Variable (y):** Whether the athlete reaches the final (0/1).
* **Predictors (X):** (Time frame: the past year)
  + Personal best time (max)
  + Recent competition results (mean/sum)
  + Performance consistency (variance of results) - Standard Deviation (SD)
  + Error rate (false starts): number of errors/total number of competitions

**General Process:**

1. Identify the semifinalists (24 athletes) for the 100m sprint in the 2021, 2016, 2012, and 2008 Olympics, and find the corresponding feature data (personal best time, recent competition results, recent performance variance, error rate).
2. Obtain weight predictions (choose model method):
   * (1) Analytic Hierarchy Process/Entropy Method for calculating weights
   * (2) Logistic Regression Analysis
   * (3) Machine Learning Models: Support Vector Machine (SVM)/Gradient Boosting Machine (GBM)

**Model Validation:**

* Compare the results predicted by the model with the actual finalists in the 2021, 2016, 2012, and 2008 Olympics. If the accuracy is low, change the model method.

**Prediction:**

* Obtain the recent competition records of the top 50 ranked athletes in the world.
* Input the data into the model and rank the athletes based on the likelihood of reaching the final.

**Potential Data for Analysis:**

* **Historical Competition Results:** Collect past Olympic and major event 100m sprint results.
* **Athlete Profiles:** Filter athletes' competition results and error records from the past year.

**Databases:**

https://worldathletics.org/competition/calendar-results/results/7211616?day=2

https://worldathletics.org/competition/calendar-results/results/7209387?eventId=10229630  
<https://worldathletics.org/athletes/kenya/ferdinand-omanyala-14747153>  
<https://worldathletics.org/athletes/south-africa/akani-simbine-14417763>  
https://worldathletics.org/athletes/italy/lamont-marcell-jacobs-14453864

**Topic: Predicting Fossil Discovery Locations for New Paleontological Species**

**Overview:** Finding new species fossils is a challenging task for paleontologists. This project aims to use existing fossil records, data analysis, and machine learning methods to predict the most likely locations for discovering new species fossils, thereby improving excavation efficiency and success rates.

**Data Source:**

* **Paleobiology Database (PBDB):** Contains detailed records of global fossil discoveries, including species name, geographic location (latitude and longitude), stratigraphic information, discovery time, environmental data, etc.
  + Website: [PBDB](https://paleobiodb.org/)
  + API documentation: [PBDB API](https://paleobiodb.org/data1.2/)
* **Geobiodiversity Database (GBDB):** Chinese biodiversity database.

**Model Building:** Using PBDB as an example, the data features might include species name, record time, fossil discovery latitude and longitude, environment description, geological age of the fossil, etc. Add a column "first discovery" to the data.

1. **General Analysis of Potential New Species Fossil Discovery Locations:**
   * Use latitude, longitude, and environmental data as features, and "first discovery" as the target variable.
   * Attempt multiple regression models for analysis.
   * Predict and identify the most important features. Finally, use libraries like Folium for geographical visualization.
2. **Predicting the Next Possible Discovery Location for New Species Fossils:**
   * In addition to latitude, longitude, and environmental information, include data entry timestamps.
   * Consider the changing likelihood of new species fossil discoveries over time. Incorporate time progression into the analysis to accurately reflect changes and shifts in optimal locations over time.
   * Use time series analysis methods for predictions.
3. **Identifying Geographic and Environmental Characteristics Prone to New Species Discoveries:**
   * Conduct cluster analysis and correlation analysis on the geographic information of discovered new species fossils.

Finally, compare the analysis results with actual research theories to verify the effectiveness of the analysis.

**Research Proposal on Beverage Store Selection**

## Reality Background

Since 2023, ultra-low-price brands such as Pinduoduo, Luckin Coffee, and Mixue Bingcheng have occupied a significant market share in China and are striving for international expansion. Despite Pinduoduo frequently encountering product quality issues and Mixue Bingcheng using cheap ingredients like non-dairy creamer and saccharin, consumer enthusiasm remains high.

## Research Question

In the current context (post-pandemic, high inflation, the majority of Chinese people being in the 'lower-tier market'), is the low price of beverage stores becoming more attractive than high quality?

## Data Sources

Obtain data from the years 2020, 2022, and 2024 for comparison.

## Companies to Study

Mixue Bingcheng (low price)  
Heytea  
Bubble Tea King  
Gong Cha  
Tea Master (relatively high price)  
Etc.

## User Review Sources (Apps)

Dianping  
Meituan  
Ele.me  
Etc.

## Data Retrieval (Scraper)

https://github.com/Sniper970119/dianping\_spider  
https://github.com/zsflalala/CommentRepile

## Evaluation Dimensions (Keywords)

Environment  
Atmosphere  
Store Foot Traffic  
Service  
Price  
Taste  
Repurchase Rate

## Optional Consideration Dimensions

Local geographic location  
Economic situation  
Popularity on social media platforms like Weibo

## Data Storage

PostgreSQL-based database

## Data Analysis

Perform keyword analysis and summarization of user reviews  
Method: Unsupervised learning  
Tabulate each store by year and analyze in conjunction with the total sales for each year

## Data Visualization

Omitted for brevity