Task 2: Business Understanding

Identifying Your Business Goals

Background

Theater plays a significant role in cultural enrichment and community engagement. However, with evolving audience preferences, economic constraints, and regional disparities, understanding theater performance trends is essential. This project focuses on analyzing demographic engagement, financial health, and regional trends in theater using data from Statistikaamet (2007–2023).

Business Goals

1. Identify trends in audience engagement across demographic groups and genres.
2. Evaluate the financial health and revenue sources of theaters, comparing state-funded and independent entities.
3. Understand regional differences in attendance and genre-specific popularity to inform targeted marketing and funding decisions.

Business Success Criteria

- Provide actionable insights on audience preferences to theater management for programming decisions.
- Identify financially vulnerable theaters or genres to aid policymakers in resource allocation.
- Deliver a report detailing regional and demographic-specific attendance trends for cultural planners.

Inventory of Resources

- **Data Sources:**
  Statistikaamet datasets (KU086, KU091, KU94, KU109), spanning 17 years of theater performance data.
- **Tools and Technologies:**
  Python (for data analysis, visualization, and machine learning), pandas, matplotlib, and clustering/classification algorithms.
- **Expertise:**
  Team members’ proficiency in data mining, machine learning, and statistical analysis.
- **Infrastructure:**
  Computing resources for data processing and analysis.

Requirements, Assumptions, and Constraints

- **Requirements:**
  - Ensure data integrity and completeness for robust analysis.

    - Generate interpretable results that align with stakeholders’ needs.
- **Assumptions:**
    - Audience demographics influence attendance preferences.
    - Financial data is accurate and representative of the theaters analyzed.
- **Constraints:**
    - Limited data granularity on sub-genres or performance-specific features.
    - Potential missing data points in specific years or categories.

Risks and Contingencies

- **Risks:**
    - Inconsistent data formatting due to JSON-STAT structure.
    - Insufficient coverage of independent theaters or niche genres.
- **Contingencies:**
    - Perform data cleaning and imputation where necessary.
    - Highlight limitations in the final report to manage stakeholder expectations.

Terminology

- Audience Categories: Children (up to 12 years), youth (13–24 years), adults (25+ years).
- Performance Genres: Drama, comedy, musicals, experimental theater, etc.
- Financial Metrics: Revenue from ticket sales, public funding, and operational expenses.

Costs and Benefits

- **Costs:**
    - Time investment for data preprocessing and analysis.
    - Computational resources for running machine learning models.
- **Benefits:**
    - Improved decision-making for theater programming and marketing.
    - Enhanced allocation of cultural funding based on regional or demographic needs.

Defining Your Data-Mining Goals

Data-Mining Goals

1. Clustering: Segment demographic groups by their attendance patterns to identify core audience bases.
2. Classification: Predict demographic preferences for genres and performance timings.
3. Regression Analysis: Analyze how ticket pricing and funding impact attendance.
4. Anomaly Detection: Highlight theaters with unusual financial trends or attendance patterns.

Data-Mining Success Criteria

- Successful identification of demographic clusters with distinct attendance patterns.

- Accurate prediction models (e.g., 85% accuracy or higher) for audience preferences.
- Reliable regression models showing correlations between financial inputs and attendance.
- Clear identification of outliers in financial or attendance data for further investigation.

Summary

This project aims to leverage data analysis and machine learning techniques to provide actionable insights into audience engagement, financial trends, and regional disparities in theater performances. The findings will benefit theater managers, policymakers, and cultural planners by enabling informed decisions on programming, funding, and marketing strategies.

Task 3: Data Understanding

1. Gathering Data

a. Outline Data Requirements

The objective of this project is to analyze theater data from Statistikaamet to identify patterns and trends, such as attendance, new productions, and performance categories. For this, we require data spanning multiple years and encompassing categories like total performances, audience demographics, genres, and other relevant metrics.

b. Verify Data Availability

The data has been sourced from Statistikaamet's API endpoints, specifically datasets `KU086`, `KU091, KU94, KU109`. These datasets cover key metrics, including annual performance statistics categorized by audience type (children, youth, adults) and genre. The data is available in JSON-STAT format and was successfully retrieved and converted to CSV files.

c. Define Selection Criteria

- **Temporal Range**: Data from 2007 to 2023 was selected to ensure long-term trends can be analyzed.
- **Categories**: Only categories relevant to theater performances, audience demographics, and genres were included.
- **Exclusion**: Years prior to 2007 were excluded due to the data’s incompleteness.

2. Describing Data

The gathered data is structured as follows:

- **Dataset Dimensions**: The main dataset has 1 row and 4,131 columns, indicating that the data was initially flattened. After preprocessing, the data was transposed and restructured into a format suitable for analysis.
- **Key Fields**:
  - **Year**: Indicates the reporting year.

    - **Category**: Specifies the audience type (e.g., children, youth, adults).
    - **Genre**: Details the type of performances, such as drama, comedy, musicals, etc.
    - **Metrics**: Includes total performances, new productions, audience size (in thousands), and guest performances.

The dataset was preprocessed to convert the JSON-STAT structure into a tabular format with rows representing records and columns representing attributes.

3. Exploring Data

a. Initial Observations

- The dataset spans 17 years and contains metrics broken down by category and genre.
- Total performance counts and audience metrics vary significantly across years, reflecting potential trends or external factors (e.g., pandemic impacts).
- Audience demographics provide insights into which groups engage most with theater performances.

b. Statistical Summary

Using pandas, descriptive statistics were generated:

- **Performance Metrics**:
    - Average total performances per year: ~420
    - Average audience size per year: ~1,000,000 (thousand visitors)
- **Category-Specific Observations**:
    - Adults consistently account for the largest share of the audience.
    - Children's and youth-specific performances show variability across years.

c. Visual Exploration

- Line plots of audience size over time show upward trends until 2019, followed by a decline during 2020-2021, possibly due to COVID-19 disruptions.
- Bar charts comparing new productions across genres reveal a consistent dominance of drama over other genres.

4. Verifying Data Quality

a. Completeness

- Missing Values: Some years have incomplete records for specific categories or genres (e.g., ".." placeholders). These were replaced with NaN for numerical consistency.
- Temporal Gaps: Data is available for all required years, but specific fields (like new productions by genre) require interpolation or exclusion due to missing values.

b. Consistency

- Column names and metrics were standardized during preprocessing. The JSON-STAT structure occasionally caused misalignments, which were resolved by reshaping and reassigning headers.

c. Accuracy

- The data comes directly from Statistikaamet, ensuring a high degree of reliability.
- Spot checks on sample data validated logical consistency (e.g., total audience equals the sum of audience subcategories).

d. Noise Handling

- Irrelevant fields, such as metadata unrelated to the analysis (e.g., `updated`, `source`), were removed.
- Placeholder values (`..`) were systematically converted to `NaN` to ensure numerical fields could be analyzed correctly.

Task 4: Planning your project

1. Data Preprocessing (15 hours)

- **Tasks**:
  - Clean, format, and standardize the data.
  - Handle missing values and split data for analysis.
  - Perform exploratory analysis to identify trends.
- **Team Member Allocation**:
  - Student A: 5 hours
  - Student B: 5 hours
  - Student C: 5 hours

2. Audience Engagement Analysis (15 hours)

- **Tasks**:
  - Use clustering and classification to analyze audience demographics and preferences.
  - Create visualizations to summarize findings.
- **Team Member Allocation**:
  - Student A: 5 hours
  - Student B: 5 hours
  - Student C: 5 hours

3. Financial Trend Analysis (12 hours)

- **Tasks**:

    - Analyze revenue and expenses using regression models.
    - Detect financial anomalies and visualize key metrics.
- **Team Member Allocation**:
    - Student A: 4 hours
    - Student B: 4 hours
    - Student C: 4 hours

4. Regional and Genre Trends (12 hours)

- **Tasks**:
    - Study regional attendance differences and genre-specific trends.
    - Create maps and charts to illustrate findings.
- **Team Member Allocation**:
    - Student A: 4 hours
    - Student B: 4 hours
    - Student C: 4 hours

5. Report and Presentation (6 hours)

- **Tasks**:
    - Compile findings into a report and prepare a presentation.
- **Team Member Allocation**:
    - Student A: 2 hours
    - Student B: 2 hours
    - Student C: 2 hours

Methods and Tools

- **Methods**: Clustering, classification, regression, anomaly detection.
- **Tools**: Python, Jupyter Notebook, Excel

Comments:
Tasks are evenly distributed to ensure balanced workloads. Regular team check-ins will maintain progress and address challenges.