Formatting Instructions for NIPS 2015

Yunhao. Yang
Institute of Collaborative Innovation
University of Macau
Macau, China
mc46493@um.edu.mo

Coauthor
Affiliation
Address
email

Abstract

This study involves downloading Macau's tourism data from the Macau Open Data website for data processing and model construction. The primary objective is to predict the tourism industry's data in Macau. Through data cleaning, analysis, and modeling, a predictive model was developed to effectively forecast future trends in Macau's tourism sector. The results of this study provide data support and decision-making references for the development of Macau's tourism industry.

1 Introduction and Motivation

1.1 Introduction

Tourism is a vital sector for Macau's economy, contributing significantly to its GDP and employment. With the increasing availability of open data, there is a growing opportunity to leverage this information for better decision-making and strategic planning in the tourism industry.

This study aims to utilize tourism data from the Macau Open Data website to develop predictive models that can forecast future trends in the tourism sector.

1.2 Motivation

The motivation behind this research is to provide actionable insights and data-driven recommendations to stakeholders in Macau's tourism industry. By accurately predicting tourism trends, businesses and policymakers can make informed decisions to enhance the visitor experience, optimize resource allocation, and improve overall industry performance.

This study not only contributes to the academic field of data science and tourism management but also has practical implications for the sustainable development of Macau's tourism sector.

2 Impact to our society

The predictive models developed in this study have significant implications for Macau's tourism industry and society at large. By providing accurate forecasts of tourism trends, these models can help businesses and policymakers make informed decisions, leading to better resource allocation and enhanced visitor experiences.

This, in turn, can boost the local economy, create more job opportunities, and promote sustainable tourism practices. Additionally, the insights gained from this study can serve as a valuable reference for other regions looking to leverage open data for tourism management and development.

3 Methodology

including data processing, data analysis, and / or other technical contents

3.1 Data downloading

To download the required datasets, we utilized a Python script with Selenium for web automation. The script performs the following steps:

3.1.1 Setup

Import necessary libraries and configure the download path based on the operating system. Determine the default download directory and ensure it exists.

3.1.2 Define download path

Automate the process of accessing the Macau Open Data website, navigating to the datasets section, and performing searches based on specified keywords. The script handles interactions such as clicking on dropdown menus, entering search terms, and initiating downloads.

3.1.3 Execute download

Iterate through a list of keywords related to tourism data, download the corresponding datasets, and store them in the designated download directory. The script also logs network requests to verify successful downloads and extract file names.

3.2 Data processing

We loaded ten datasets related to Macau's tourism industry using the pandas library. These datasets include total consumption, average consumption of staying and non-staying tourists, inbound tourists, staying tourists, non-staying tourists, hotel occupancy rate, mainland individual visitors, average length of stay for tourists, and average length of stay for guests.

3.2.1 Previewing data

For each dataset, we displayed the first few rows to understand the structure and content of the data. This step helps in identifying any immediate issues such as missing values or incorrect data types.

3.2.2 Creating a unified DataFrame

We created a unified DataFrame with a quarterly time index from 2010 to 2024. This involved: Extracting and adding relevant data from each dataset to the DataFrame. Renaming the index to match the quarterly format. Ensuring the index is correctly formatted and named.

3.2.3 Heading missing values

We replaced all occurrences of the '~' symbol with NaN to standardize missing values. We used backward fill (bfill) to handle missing values, ensuring that the data remains consistent and complete for analysis.

3.2.4 Data type conversion and date conversion

We converted relevant columns to integer type (Int64) to facilitate numerical operations and analysis. We converted the quarterly index to a standard date format and then to Unix timestamps. This step is crucial for time series analysis and modeling.

3.2.5 Data overview

We displayed the first few rows of the unified DataFrame, along with data type information and statistical summaries. This provides a comprehensive overview of the data, ensuring that it is ready for further analysis and modeling.

These data processing steps ensure that the datasets are clean, consistent, and ready for subsequent analysis and model development.

Table 1: Summary Statistics of Tourism Data

Summary	Total_ consumption	Average consumption_staying	Average consumption_non_staying	Inbound tourists	staying_ tourists	non_staying_ tourists	Hotel occupancy_rate	Mainland individual_vistors	Average_length_ of_staying_touristst	Average_length of_tourists
Count	58.0	58.0	58.0	5.800000e+01	5.800000e+01	5.800000e+01	58.000000	5.800000e+01	58.000000	58.000000
Mean	12402337490.362068	3508.482759	711.793103	6.476239e+06	3.150274e+06	3.325965e+06	75.172414	1.973558e+06	4.508621	3.920690
Std	4882264884.75848	1160.763032	143.312902	2.715236e+06	1.354667e+06	1.389243e+06	20.036450	9.468059e+05	0.440607	1.978371
min	85374779.0	2294.0	534.0	4.973000e+04	2.484900e+04	2.488100e+04	12.066667	2.340000e+02	3.800000	2.700000
25%	9728416140.75	2733.0	611.0	6.176199e+06	2.937190e+06	3.202757e+06	78.358333	1.420776e+06	4.200000	3.100000
50%	13408510512.5	3081.5	672.0	7.378250e+06	3.522882e+06	3.783353e+06	83.383333	2.093024e+06	4.300000	3.500000
75%	15942551343.5	3880.25	745.5	8.102742e+06	4.060634e+06	4.119342e+06	86.866667	2.584618e+06	4.800000	3.800000
Max%	20348996921	7833	1188	1.035976e+07	4,880,404	5,624,732	92,866,667	3,826,241	5,900,000	14

3.3 Data visualization

The data analysis and visualization steps involved in this study are as follows.

3.3.1 Box plot analysis

We created box plots for each numerical variable to identify outliers and understand the distribution of the data. This helps in detecting any anomalies and understanding the spread of the data.

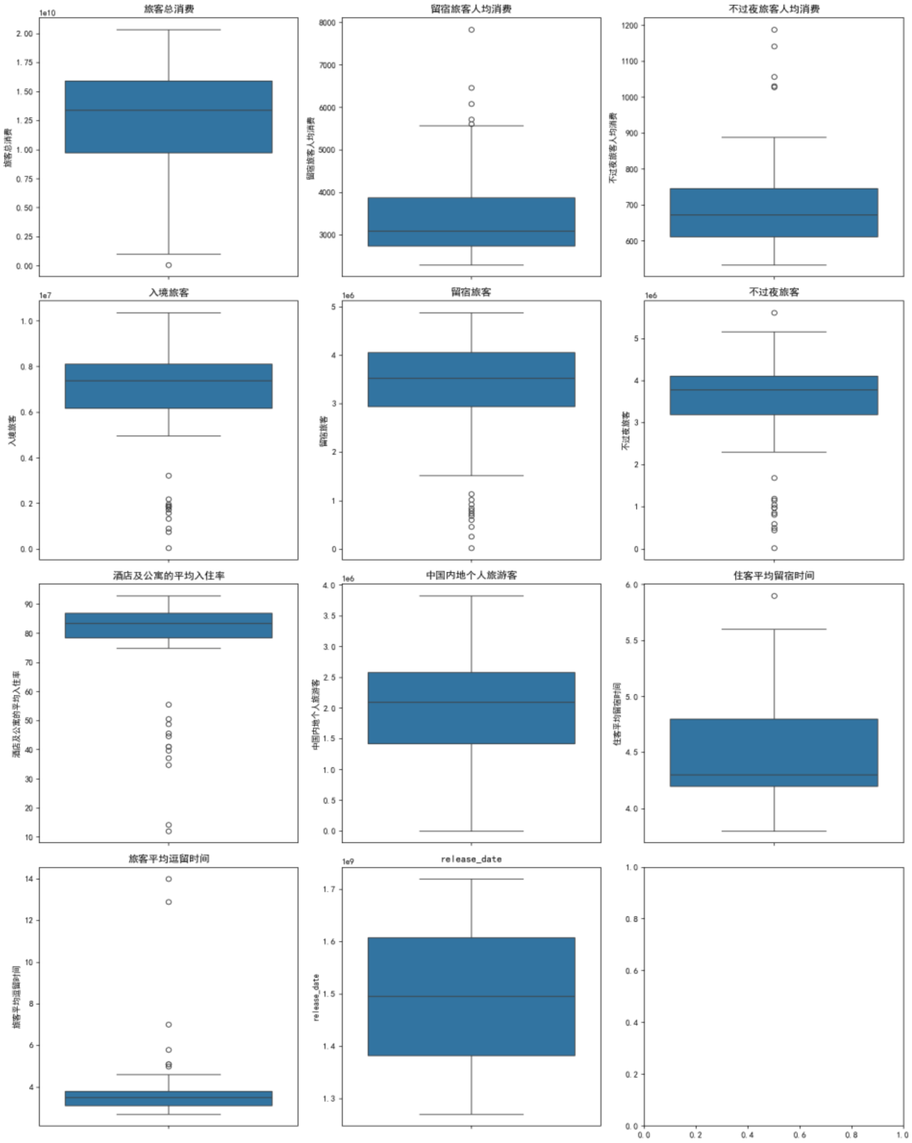

Figure 1: Box Plot

3.3.2 Time series analysis

We plotted time series graphs for key variables such as total consumption, average consumption of staying and non-staying tourists, hotel occupancy rate, average length of stay for tourists, and average length of stay for guests. These plots help in visualizing trends and patterns over time.

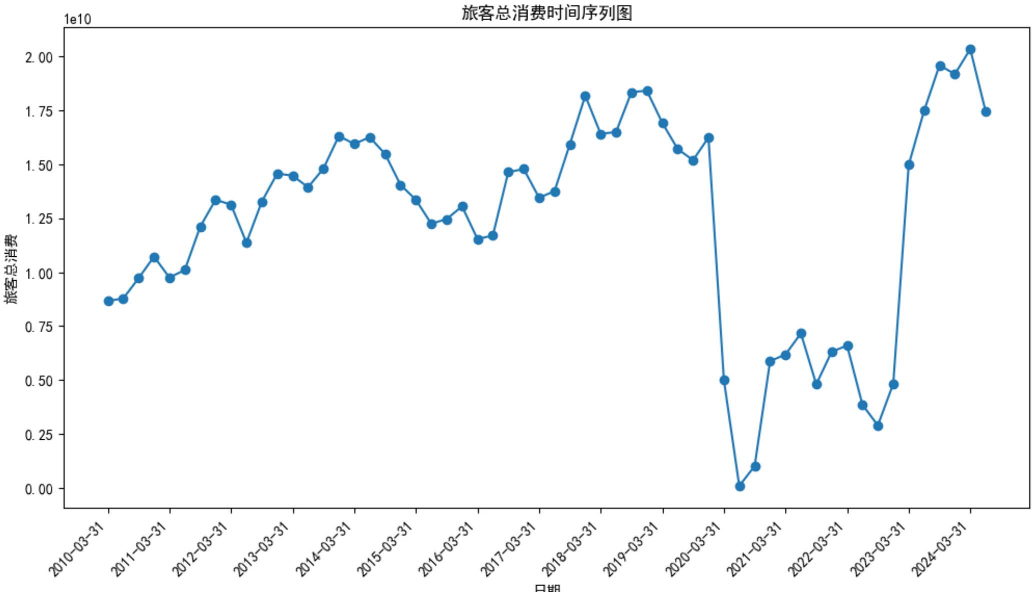

Figure 2: Time Series Analysis

3.3.3 Principal component analysis (PCA)

We performed PCA to reduce the dimensionality of the dataset and visualize the data in a two-dimensional space. This helps in identifying underlying patterns and relationships between variables. We used Bokeh to create interactive PCA plots with hover tools for better visualization.

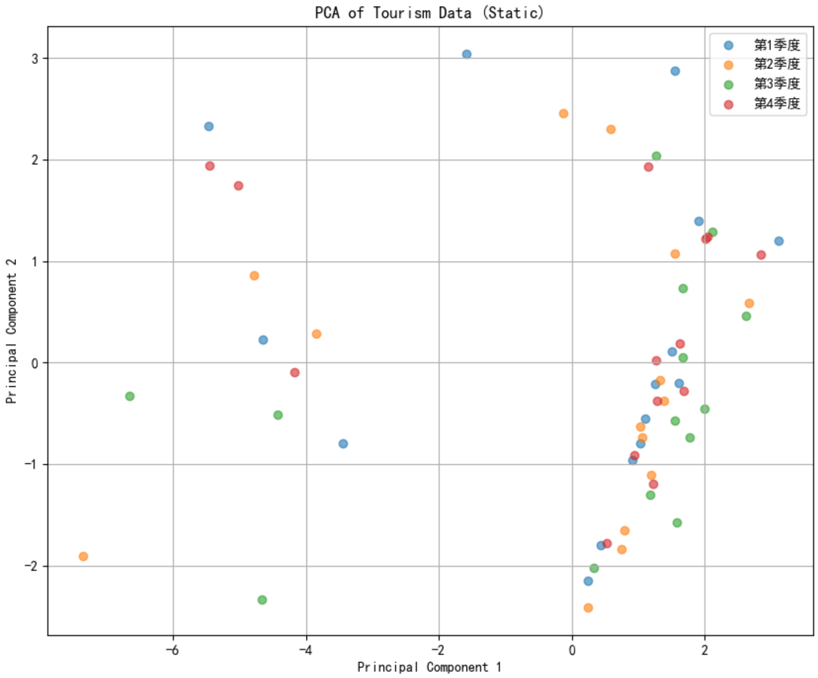

Figure 3: PCA Plot

3.3.4 Uniform manifold approximation and projection (UMAP)

We applied UMAP to further reduce the dimensionality of the dataset and visualize the data in a two-dimensional space. UMAP is particularly useful for capturing non-linear relationships between variables. We used Bokeh to create interactive UMAP plots with hover tools for better visualization.

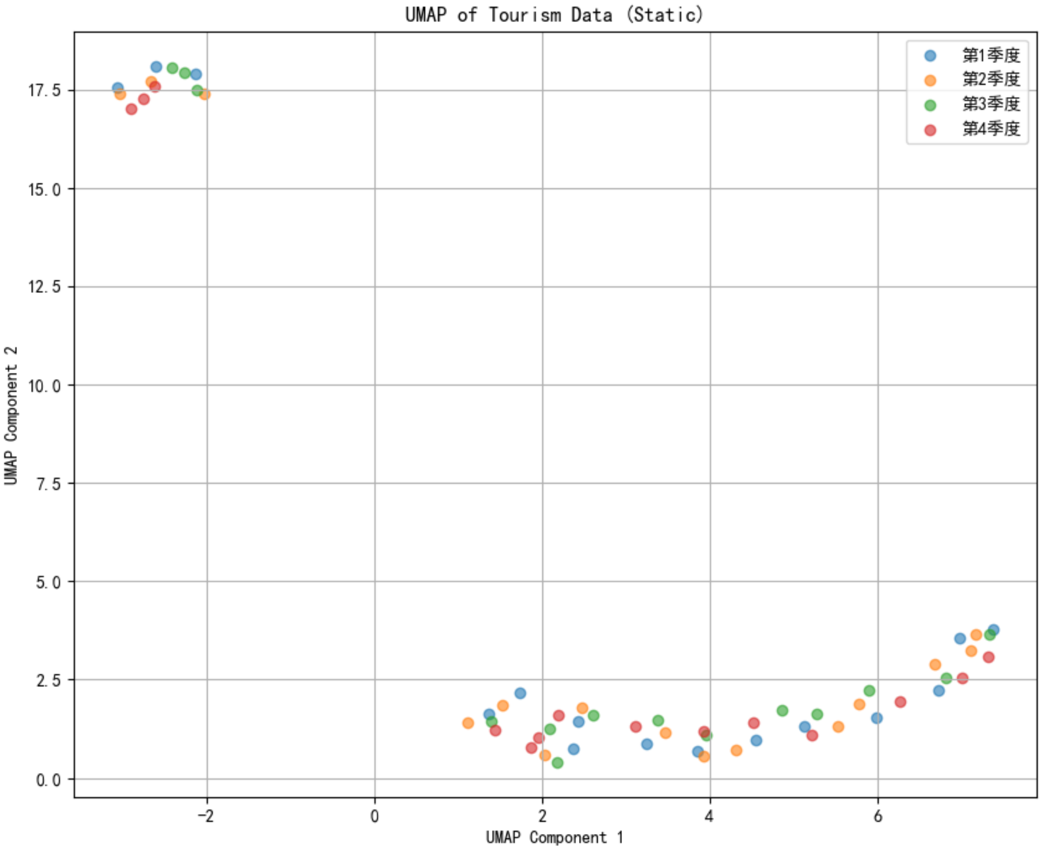

Figure 4: UMAP Plot

3.3.5 Correlation analysis

We calculated the correlation matrix to understand the relationships between different variables. A heatmap was created to visualize the correlation coefficients, highlighting strong positive or negative correlations. This helps in identifying which variables are most closely related to each other

.

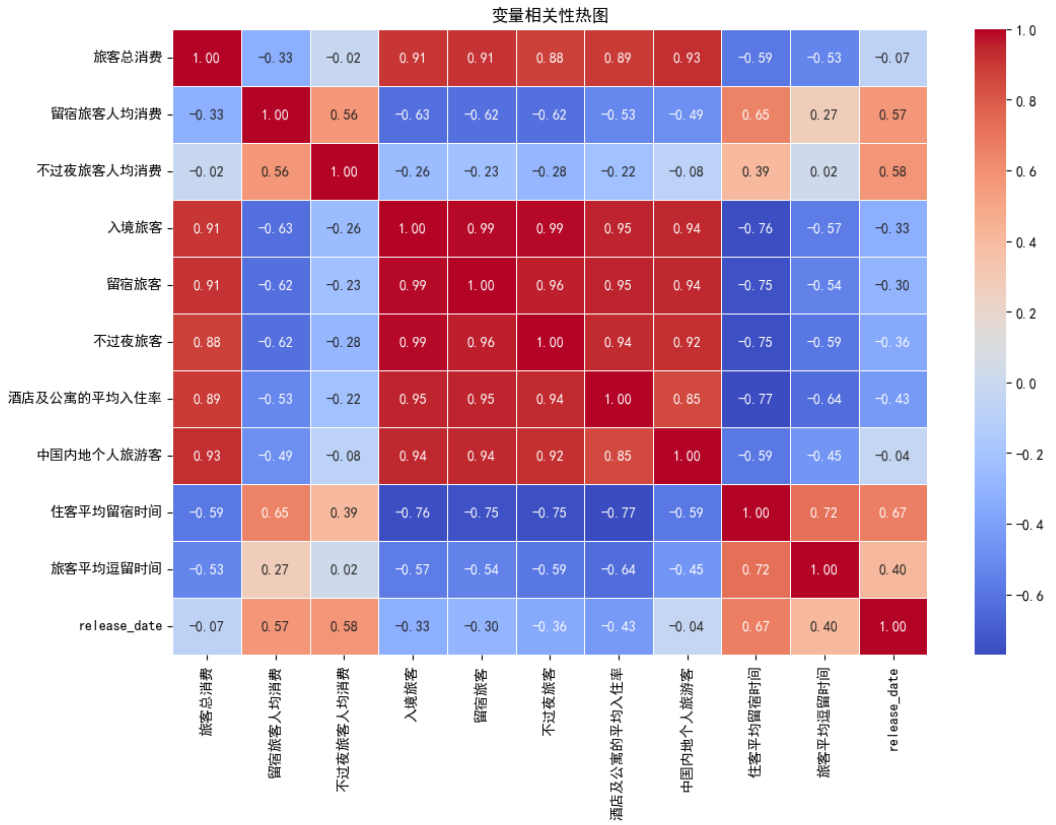

Figure 5: Correlation Analysis

These data analysis and visualization steps provide a comprehensive understanding of the dataset, revealing important trends, patterns, and relationships that are crucial for further analysis and model development.

4 Experiments

We created a unified DataFrame with a quarterly time index from 2010 to 2024, ensuring the data was clean and consistent for analysis.

4.1 Initial model

We initially constructed a polynomial regression model to explore the relationship between various features and total consumption. The model was trained using data from 2010 to the present, and the results showed a strong correlation between the total consumption and the selected features.

4.1.1 Improved model

To improve the accuracy of the model, we reduced the features and adjusted the polynomial degree to 1. This model aims to explore the relationship between other features and consumption.

4.2 Pre-COVID model training

We trained the model using data from before the COVID-19 pandemic to predict the tourism trends in Macau under normal conditions.

4.2.1 Feature selection

We initially selected four features to explore their relationship with total consumption.

The results indicated a high coefficient of determination (R^2) between these features and total consumption, suggesting a strong correlation.

Further analysis revealed a significant correlation between inbound tourists and the other three features (release date, average consumption of staying tourists, and average consumption of non-staying tourists). This finding prompted us to focus on predicting inbound tourists as an intermediary step.

4.3 Post-COVID predictions

For the post-COVID predictions, we assumed that the average consumption values during the pandemic would be equal to the average values from the previous quarters.

Using the trained model, we predicted the inbound tourists for the period after 2020. These predictions were then used to forecast the total consumption for the same period.

4.4 Results and visualization

The predicted values were combined with the actual data to visualize the impact of the pandemic on Macau's tourism industry. The results were plotted to compare the actual total consumption with the predicted total consumption under the assumption of no pandemic. The visualization clearly showed the differences and provided insights into the potential impact of the pandemic on the tourism sector.

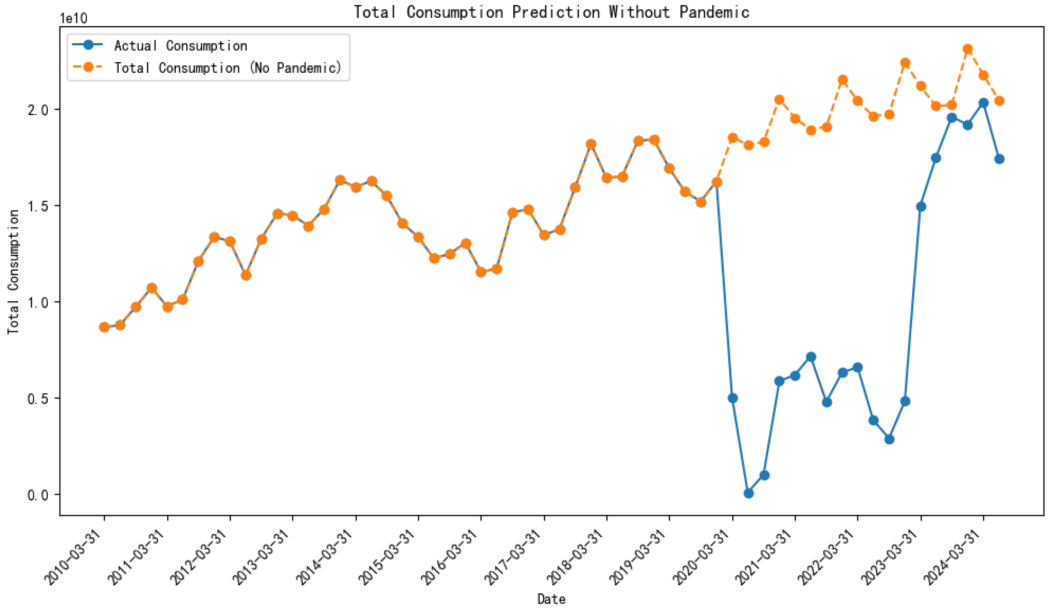

Figure 6: Total Consumption prediction

Acknowledgments

I would like to express my gratitude to the AI tools and technologies that assisted me in completing this research.

References