README: Val di Fassa Tourism Statistical Analysis

Authors:

- Giovanni Vitali (2119998)

- Giacomo Vezzosi (2104369)

- Mandana Goudarzi (2122279)

Overview

This R script, `ValdiFassa_Tourism_Analysis.R`, is the code of our project for "Business Economic and Financial Data" master degree course, from University of Padova. It performs an in-depth analysis of tourism data for the Val di Fassa region. The script includes preprocessing, exploratory data analysis (EDA), time series modeling, and forecasting. It is designed to evaluate trends and seasonal patterns in tourist arrivals, using various statistical techniques.

Prerequisites

The analysis is based on the dataset `ValDiFassa_tourism_data.xlsx`. This file must be placed in the **same directory** as the R script for the code to function correctly.

Required Libraries

The script uses the following R packages:

- `here`
- `lubridate`
- `zoo`
- `Metrics`
- `mgcv`
- `ggplot2`
- `sf`
- `rnaturalearth` (required package: `devtools`)
- `rnaturalearthdata` (required package: `devtools`)
- `readxl`
- `DIMORA`
- `forecast`
- `fpp2`
- `dplyr`
- `lmtest`
- `tidyr`
- `prophet`
- `gam`
- `tseries`
- `scales`
- `gridExtra`
- `corrplot`

Script Sections

1. Data Loading and Preprocessing

- **Input File**: `ValDiFassa_tourism_data.xlsx`
- **Sheets Utilized**:
  - `ValDiFassa`: Total and hotel arrivals data.
  - `Comparison Trentino-VdF`: Comparison between Trentino and Val di Fassa.
  - `Nationality`: Proportion of foreign tourist arrivals.
- **Preprocessing Steps**:
  - Extract and clean relevant columns (e.g., `Total Arrivals`, `Hotel Arrivals`).
  - Convert date strings to `Date` objects.
  - Create time series objects with a monthly frequency.

2. Exploratory Data Analysis (EDA)

- **Goals**:
  - Visualize the trends, seasonal patterns, and geographic distributions.
  - Analyze the proportion of tourists (Italians vs. foreigners).
- **Key Plots**:
  - Geographic maps showing tourist density.
  - Year-over-year comparisons of arrivals.
  - Cumulative sum plots.
  - Seasonal trends and pie charts for arrivals distribution.

3. Time Series Modeling

The script implements several time series models:

a. Linear Models (TSLM)

- **Description**: Fits a linear trend and seasonality model.
- **Metrics**: MAE, RMSE, AIC.

b. Exponential Smoothing (ETS)

- **Description**: Models the series with additive or multiplicative methods.
- **Residual Analysis**: Includes diagnostics using `checkresiduals`.

c. ARIMA/SARIMA

- **Description**: Automatically identifies the best ARIMA model using `auto.arima`.
- **Forecast**: Generates predictions for the next 12 months.

d. Prophet

- **Description**: Models seasonality, trend, and special events (e.g., pandemic impact).
- **Features**:
  - Multiplicative seasonality.
  - Pandemic holidays added as external regressors.

e. Generalized Additive Models (GAM)

- **Description**: Flexible modeling of nonlinear trends and seasonal effects.
- **Splines Used**:
  - `s(time_index, k=24)` for trends.

    - `s(month_num, bs="cc", k=12)` for seasonality.

4. Forecasting

- Generates forecasts for:
    - Total arrivals.
    - Hotel arrivals.
    - Trentino - Val di Fassa hotel arrivals.
    - Proportion of foreign tourists.
- Forecast horizon: 12 months.

5. Residual Diagnostics

- Residual checks performed for all models using:
    - ACF and PACF plots.
    - `checkresiduals()` for statistical diagnostics.

How to Use

1. Ensure the dataset `ValDiFassa_tourism_data.xlsx` is in the same directory as the script.
2. Open the script in R or RStudio.
3. Install the required libraries if not already installed.
4. Run the script section by section to:
    - Load and preprocess the data.
    - Perform EDA to understand the data patterns.
    - Fit models and generate forecasts.
5. Review the output plots and metrics to interpret results.

Outputs

- **Visualizations**:
    - Maps, line charts, pie charts, and bar plots.
    - Seasonal decomposition and year-over-year trends.
- **Forecast Results**:
    - Predictions for the next 12 months for total, hotel, and foreign tourist arrivals.
    - Residual analysis plots.
- **Metrics**:
    - MAE, RMSE, and AIC for each model.

Notes

- The dataset spans December 2018 to the most recent available month.
- All time series have a monthly frequency.
- The pandemic period (November 2020 to May 2021) is explicitly modeled in the Prophet analysis.
