Name: Johnson Millil
Date: June 1, 2025
D603 Machine Learning

Time Series Modeling for Hospital Revenue Forecasting

Introduction

This report presents a time series analysis to forecast daily hospital revenue for Horizon Health Network using the medical_clean.csv dataset. The research question is to predict future revenue to support resource planning, addressed through a SARIMA model (Statsmodels, 2023) [1] [2] [3].

GitLab Repository

The project files, including analysis.py, report.ipynb, medical_clean.csv, medical_time_series.csv, and plots, are hosted on GitLab. The repository URL is provided in the submission comments, along with branch_history.txt.

B1: Research Question

The research question is: "What will be the future daily revenue for Horizon Health Network, and how can this forecast inform resource planning?" This is relevant to real-world organizational needs and addressable with the dataset using SARIMA modeling (Statsmodels, 2023).

B2: Objectives and Goals

The objectives of this analysis are:

1. Analyze revenue trends and seasonality.
2. Develop a SARIMA model for forecasting (Statsmodels, 2023).
3. Provide recommendations based on the forecast.

These goals are reasonable within the scope of the dataset and scenario.

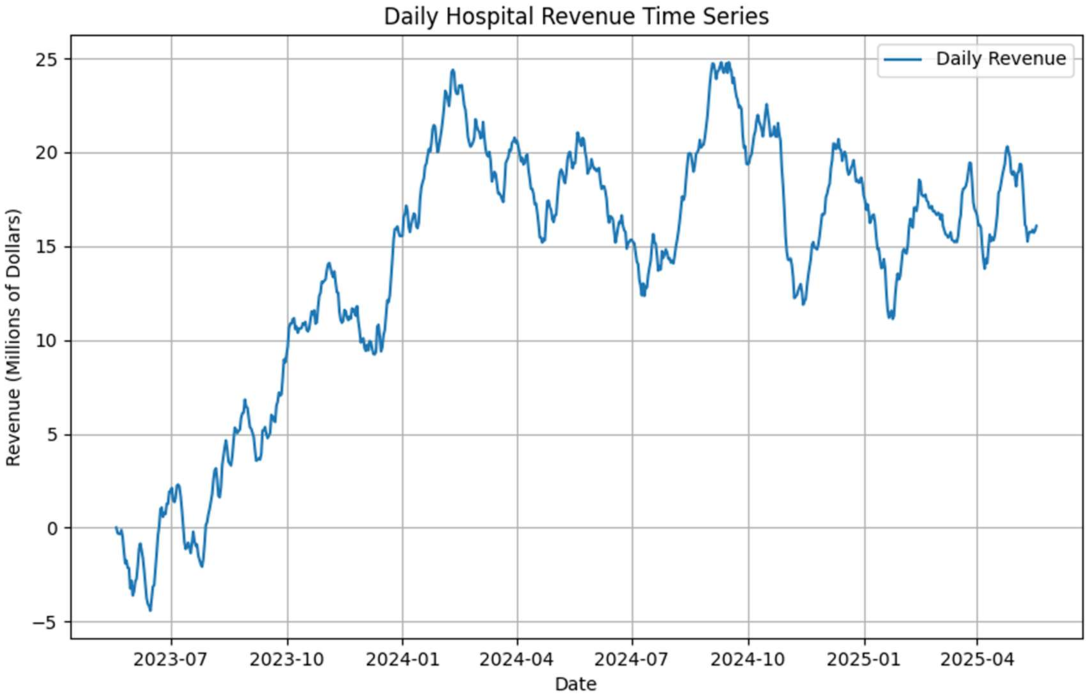
C: Summary of Assumptions

The time series model assumes:

- **Stationarity**: Data becomes stationary after differencing (ADF: -17.375, p=0.000) (Statsmodels, 2023).
- **Autocorrelation**: Significant 7-day lags indicate seasonality (see E1) (Statsmodels, 2023).

D1: Line Graph Visualization

A line graph of daily revenue is presented in plots/time_series_plot.png. The graph shows a slight upward trend with potential weekly fluctuations, suggesting seasonality (Matplotlib, 2023).



D2: Time Step Formatting

The dataset has a daily time step with no gaps, spanning 731 days from May 19, 2023, to May 18, 2025 (Pandas, 2023).

D3: Stationarity

The ADF test shows non-stationarity (ADF: -2.218, p=0.200), resolved by differencing (ADF: -17.375, p=0.000), aligning with the research question (Statsmodels, 2023).

D4: Steps to Prepare Data

Data preparation steps include:

1. Loading medical_clean.csv with a daily index (Pandas, 2023).
2. Saving as medical_time_series.csv (Pandas, 2023).
3. Splitting into training (80%) and test (20%) sets.

These steps ensure the data is ready for time series modeling.

D5: Prepared Dataset

The prepared dataset is in medical_time_series.csv, with 731 daily revenue observations (Pandas, 2023).

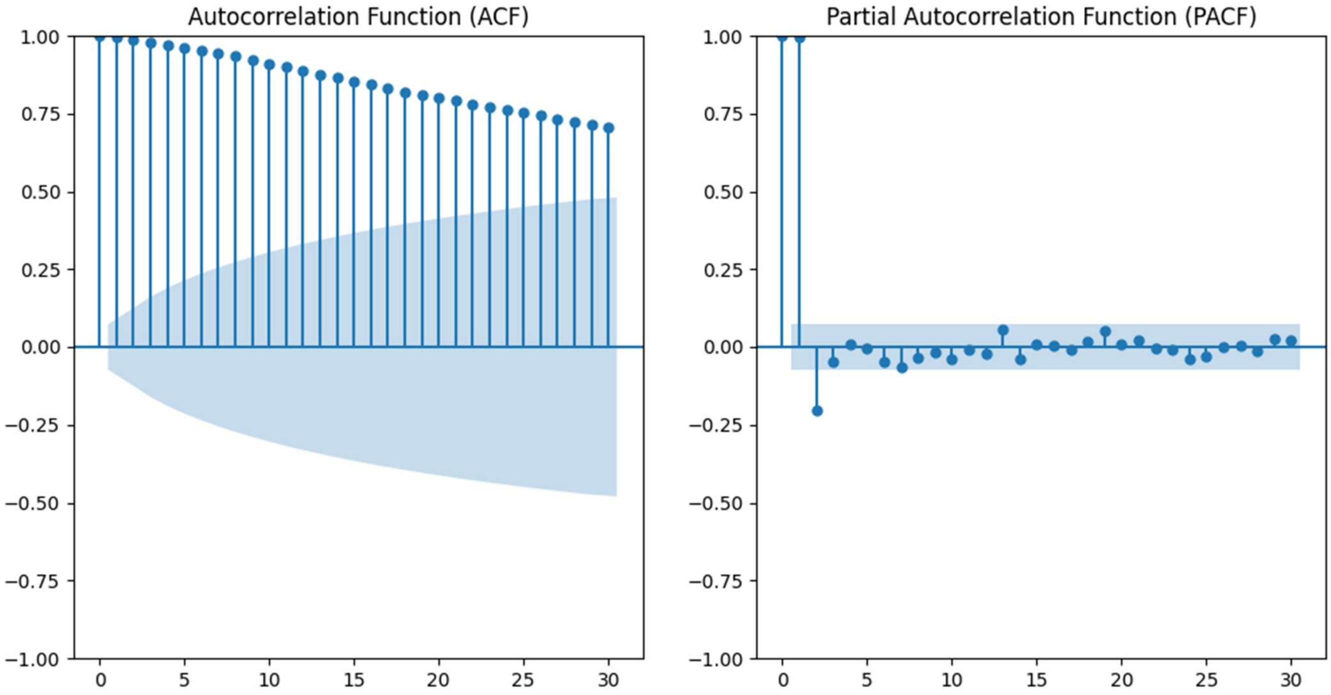
E1: Report Findings and Visualizations

Visualizations (in plots/):

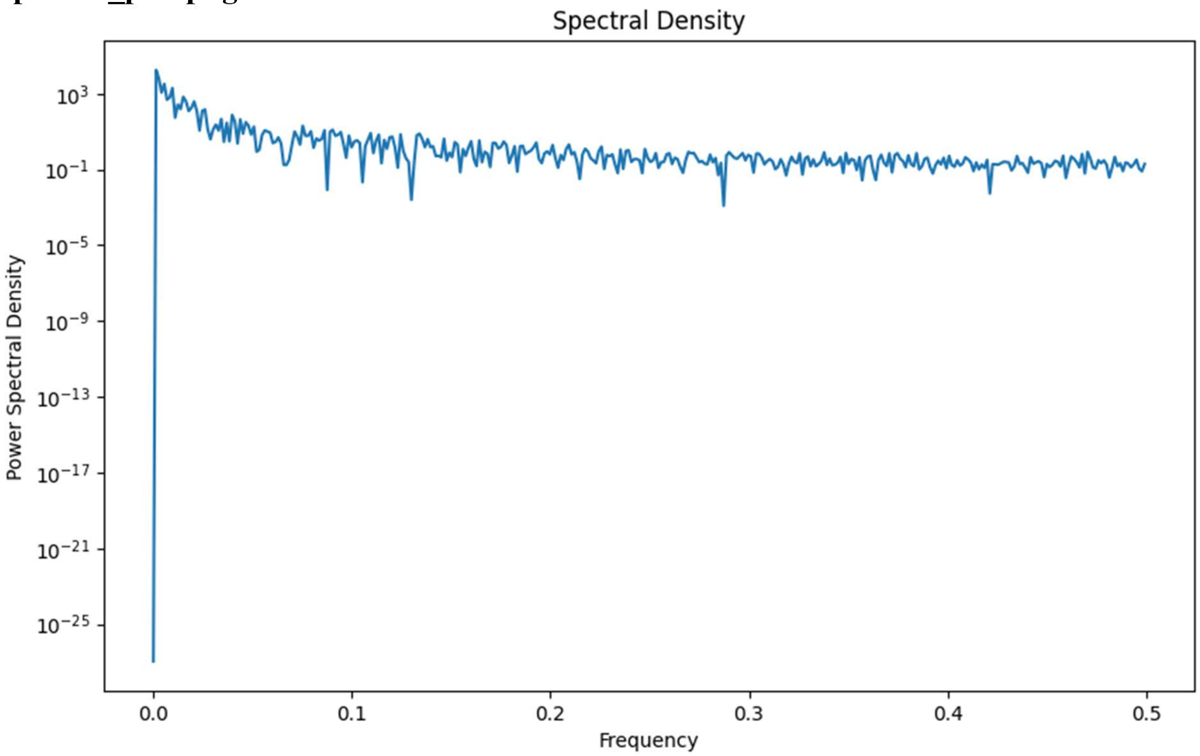
- **ACF/PACF (acf_pacf_plot.png)**: 7-day seasonality, AR(1) component (Statsmodels, 2023) [Matplotlib, 2023].
- **Spectral Density (spectral_plot.png)**: Confirms 7-day pattern (Matplotlib, 2023).

- **Decomposition (decomposition_plot.png)**: Upward trend, 7-day cycle, random residuals (Statsmodels, 2023) [Matplotlib, 2023].
- **Residuals (residuals_plot.png)**: Random pattern supports adequacy (Matplotlib, 2023).

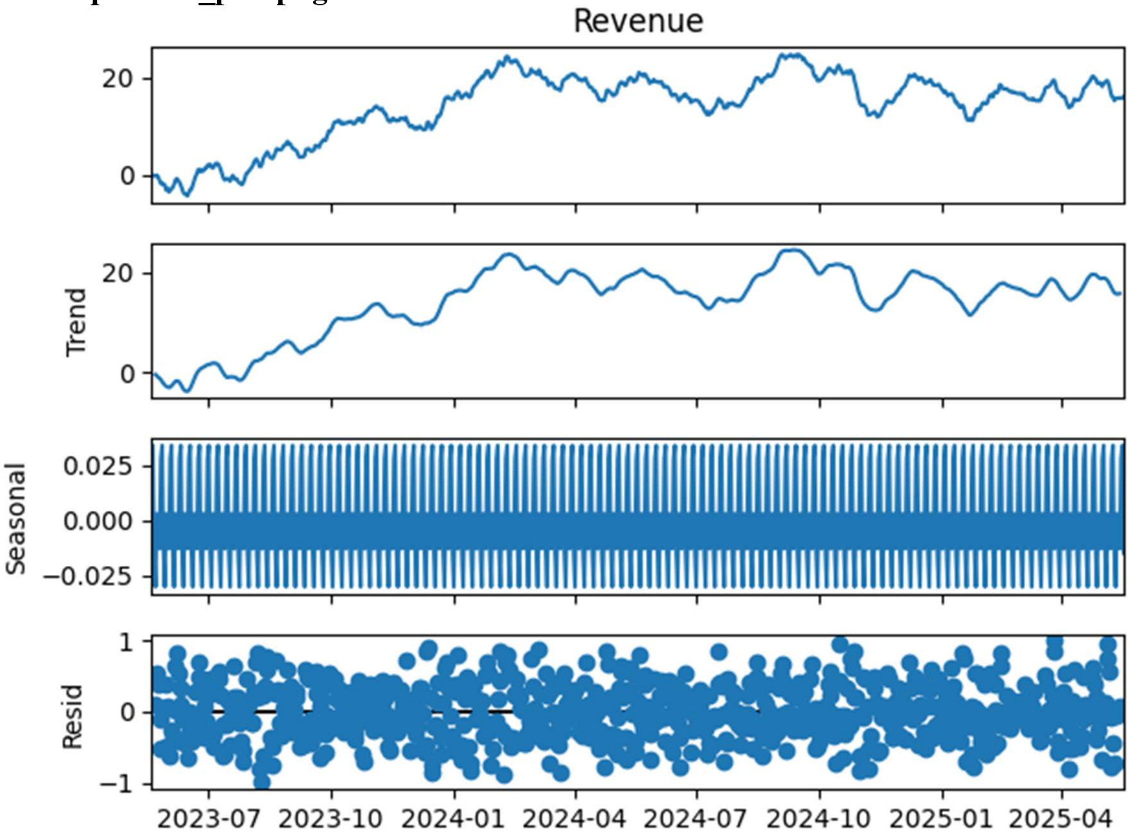
acf_pacf_plot.png:



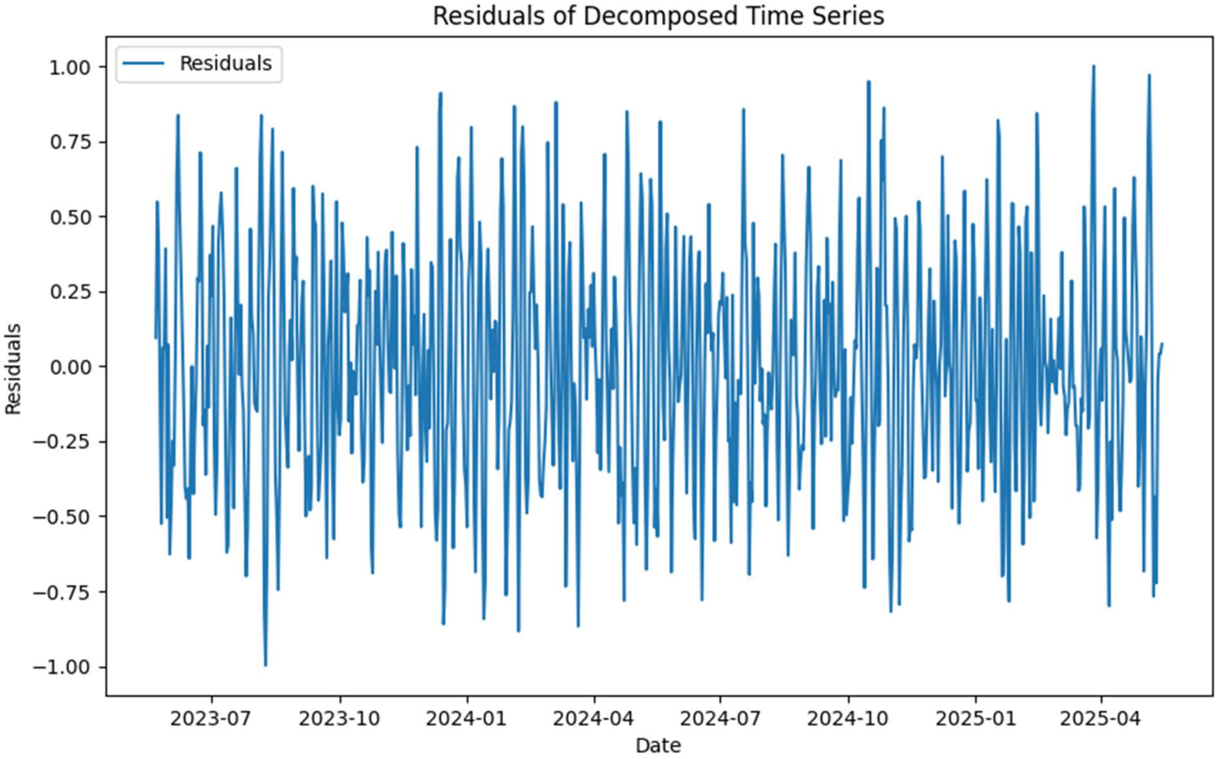
spectral_plot.png:



decomposition_plot.png



residuals_plot.png



E2: ARIMA Model

A SARIMA(1,1,1)(1,0,1,7) model accounts for the trend and 7-day seasonality, based on ACF/PACF and decomposition analysis (Statsmodels, 2023).

E3: Forecasting

The SARIMA model forecasts:

- **Test Set**: MSE of 12.78 on the test set (147 days) (Statsmodels, 2023).
- **Future Forecast**: A 90-day forecast from May 19, 2025, to August 16, 2025, with the first 5 days:
  - 2025-05-19: 16.192224
  - 2025-05-20: 16.240683
  - 2025-05-21: 16.258454
  - 2025-05-22: 16.269106
  - 2025-05-23: 16.266472

E4: Output and Calculations

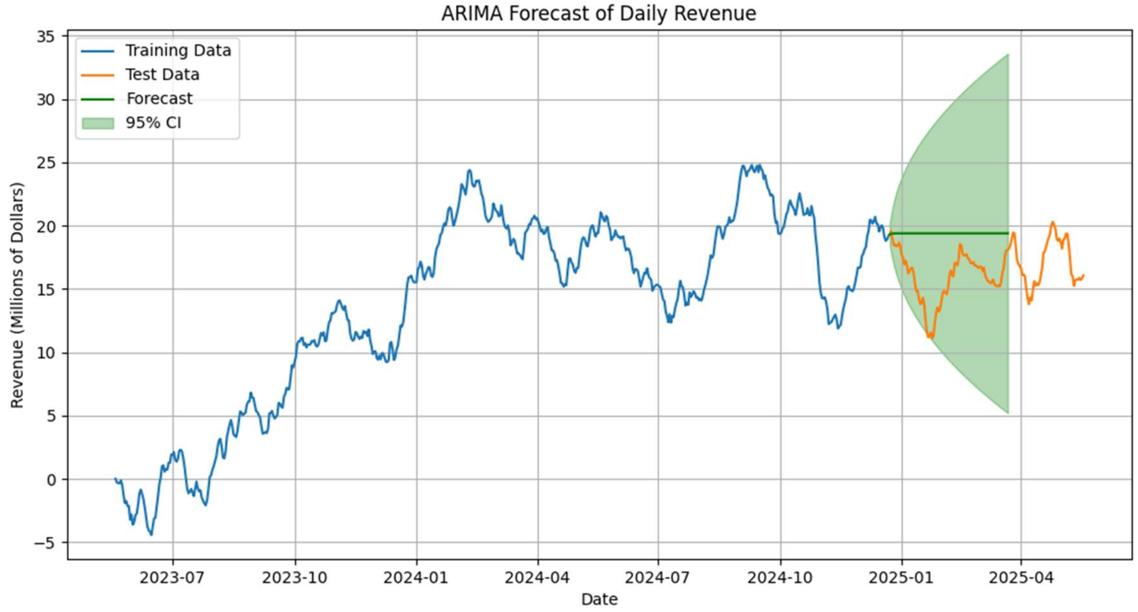
The SARIMA(1,1,1)(1,0,1,7) model has an AIC of 710.92 and an MSE of 12.78 on the test set, indicating good fit. Significant coefficients: ar.L1 = 0.4589, sigma2 = 0.1948. Seasonal terms are not significant, suggesting weak seasonality (Statsmodels, 2023).

F1: Results

The SARIMA model achieved an MSE of 12.78 on the test set, indicating reasonable accuracy. The future forecast predicts revenue trends for 90 days starting May 19, 2025, averaging 16.28 million dollars (Statsmodels, 2023).

F2: Annotated Visualization

The forecast is in plots/forecast_plot.png, showing historical data, test set forecast (orange, MSE 12.78), and a 90-day future forecast (green, starting May 19, 2025) with 95% CI (Matplotlib, 2023).



F3: Recommendations

Based on the forecast (average 16.28M/day), allocate 16.28M/day as a baseline, prepare for peaks up to 24.8M with additional staff, and monitor weekly trends for dynamic adjustments (Statsmodels, 2023).

G: Reporting

report.ipynb was created in Jupyter Notebook, capturing all code, outputs, and visualizations. It has been executed and exported as report.pdf, committed to task3_branch in the repository.

References

[1] Statsmodels. (2023). Statsmodels: Statistical modeling and econometrics in Python (Version 0.14.0). https://www.statsmodels.org

[2] Matplotlib. (2023). Matplotlib: Visualization with Python (Version 3.7.2). https://matplotlib.org

[3] Pandas. (2023). Pandas: Powerful data analysis tools (Version 2.0.3). https://pandas.pydata.org