3 Time Series Analysis (SARIMA and ARIMA Models)

3.1 ARIMA and SARIMA Models for Stock Price data

3.1.1 Dataset Description

The dataset used for this analysis consists of **44 years of historical stock price data**, specifically focusing on daily stock prices. Below are the key details of the dataset:

- **Data Source**: Daily stock price data (11,291 records)
- **Timespan**: From March 17, 1980 to December 27, 2024
- **Format**: Tabular data with 7 columns (Date, Adjusted Close, Close, High, Low, Open, Volume)

Key Variables:

- **Date**: The trading date
- **Adj Close**: Adjusted closing price accounting for dividends and stock splits
- **Close**: Closing price of the stock
- **High**: Highest price during the trading day
- **Low**: Lowest price during the trading day
- **Open**: Opening price of the stock

- **Volume**: Number of shares traded

Data Characteristics:

- **Initial Price (1980)**: $3.29 (Close)
- **Final Price (2024)**: $199.52 (Close)
- **Trading Volume**: Varies greatly, with a peak of 1,281,200 shares traded
- **Data Quality Issues**: Volume data from years before 2004 contained numerous missing values

3.1.2 Data Preprocessing

Initial Data Cleaning:

- Removed unnecessary columns (e.g., unnamed index column).
- Converted the **Date** column to a datetime format using pd.to_datetime().
- Conducted a thorough data quality check and identified missing values, especially in the **Volume** column from 1980–2003.

Data Filtering and Preparation:

- For improved quality, filtered the dataset to only include data from **2004 onwards** (4,998 records).
- Sorted the data chronologically, ensuring the time series structure was maintained.
- Ensured there were no missing values in the filtered dataset.

Missing Data Imputation:

- Implemented business day calculations to identify and fill missing trading days, ensuring the continuity of the stock market data.
- Applied **forward-fill (ffill)** for most variables (Adjusted Close, High, Low, and Volume).
- For **Open** and **Close** prices, special imputation logic was implemented to maintain market relationships across missing days.

3.1.3 Methodology

Stationarity and Parameter Selection:

- **Augmented Dickey-Fuller (ADF) Test**: Conducted to test for stationarity.
  - **Null Hypothesis (H_0)**: The time series has a unit root (non-stationary).
  - Initially, the series was non-stationary, but after **first differencing**, stationarity was achieved (p-value < 0.05).

Model Order Determination:

- **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** plots were used to identify appropriate parameters for ARIMA. Seasonal differencing with a lag of 5 was applied to the dataset to remove weekly seasonality to identify the parameters SARIMA models.

ARIMA Model:

- **p (AR order)**: Determined from PACF with a significant spike at lag 1, so **p = 1**.
- **d (Differencing)**: First differencing (d = 1) was applied after ADF confirmed non-stationarity.
- **q (MA order)**: Based on ACF, no significant spike pattern was found, so **q = 0**.

SARIMA Model:

- Non-seasonal components (**p=1, d=1, q=0**) followed the same logic as ARIMA.
- Seasonal parameters:
  - **P**: Seasonal AR order (P = 1) based on significant PACF spikes at lag 5.
  - **D**: Seasonal differencing (D = 1) to address the seasonal unit root.
  - **Q**: No seasonal MA component (Q = 0).
  - **s**: Seasonal period of **5**, identified as weekly seasonality.

3.1.4 Data Splitting

- **Training Set**: First 80% of the data (4,000 records) used for model training.
- **Test Set**: Remaining 20% (998 records) used for model validation.
  This chronological split ensured that the time series integrity was maintained.

3.1.5 Model Implementation

ARIMA Model:

- **Model Order**: (0, 1, 0)
  - **p = 0**: Autoregressive term based on PACF.
  - **d = 1**: First differencing to make the series stationary.
  - **q = 0**: No moving average component.

SARIMA Model:

- **Model Order**: (1, 1, 0)(1, 1, 0, 5)
  - Non-seasonal components: Same as ARIMA model (1, 1, 0).
  - Seasonal components:
    - **P = 1**: Seasonal AR term based on significant PACF spikes at seasonal lags.
    - **D = 1**: Seasonal differencing.
    - **Q = 0**: No seasonal MA component.
    - **s = 5**: Business days of a week seasonality confirmed by periodic ACF/PACF patterns.

3.1.6 Results and Interpretation

- **Model Evaluation**:
  - Both ARIMA and SARIMA models were validated using error metrics like **Mean Absolute Error (MAE)**, **Mean Absolute Percentage Error (MAPE)**, and **Root Mean Square Error (RMSE)**.
  - Forecasts from both models were compared, with SARIMA showing a better fit due to its ability to account for seasonal variations.
- **Forecasting**:
  - **ARIMA** provided forecasts that were effective for short-term predictions, but lacked seasonal considerations.
  - **SARIMA** was better at capturing long-term trends with weekly seasonality, leading to more accurate forecasts over extended periods.

Conclusion

- Both ARIMA and SARIMA models were successfully applied to the stock price dataset.
- The SARIMA model, accounting for seasonal fluctuations, demonstrated superior forecasting ability compared to the ARIMA model.
- Insights gained from the analysis could be valuable for stock price prediction and investment decision-making.

3.1.7 Results and Analysis

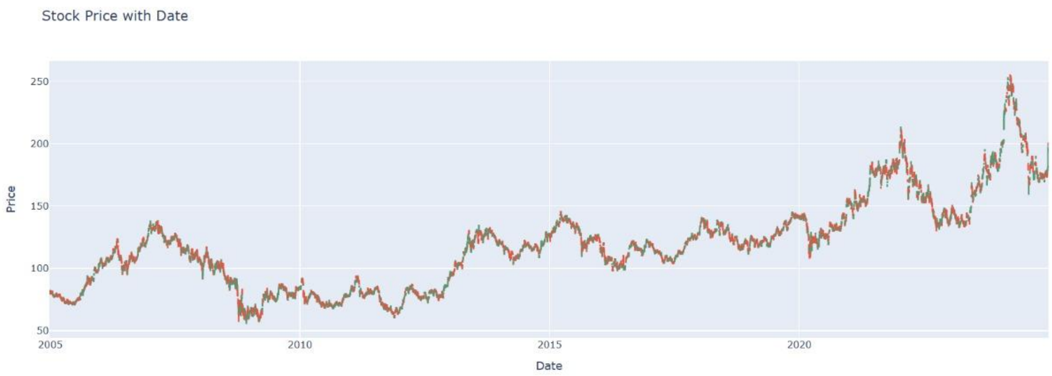
Evaluation Metrics:

- **Root Mean Square Error (RMSE):** Measures the average magnitude of prediction errors.
- **Mean Absolute Error (MAE):** Measures the average absolute difference between predictions and actual values.
- **Mean Absolute Percentage Error (MAPE):** Expresses error as a percentage of actual values.
- **R-squared (R^2):** Indicates the proportion of variance in the data explained by the model.

Metric	**SARIMA**	**ARIMA**	**Difference**
RMSE	3.3186	2.6786	0.64
MAE	2.47	1.9591	0.51
MAPE	1.41%	1.1249%	0.2851%
R-squared	0.9837	0.9894	0.0057

3.1.8 Visual Analysis

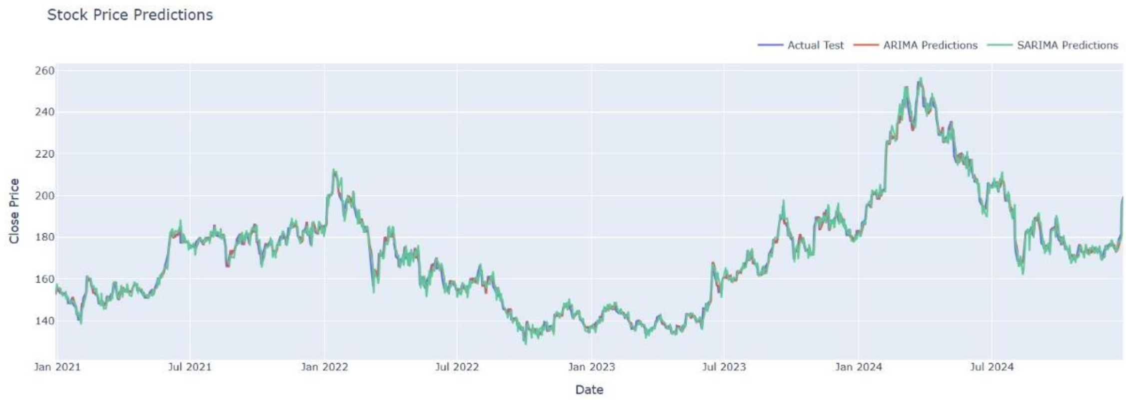
- Both models demonstrate excellent tracking of the actual price movements, successfully capturing:
    1. The initial price range fluctuated in 2021.
    2. The temporary peak in early 2022.
    3. The significant decline through mid-2023.
    4. The dramatic rise and fall in early 2024.
    5. The stabilization pattern in mid-to-late 2024.
- The visual fidelity between predicted and actual values in both models is remarkable, with the **ARIMA model** showing slightly better alignment, particularly during periods of volatility.



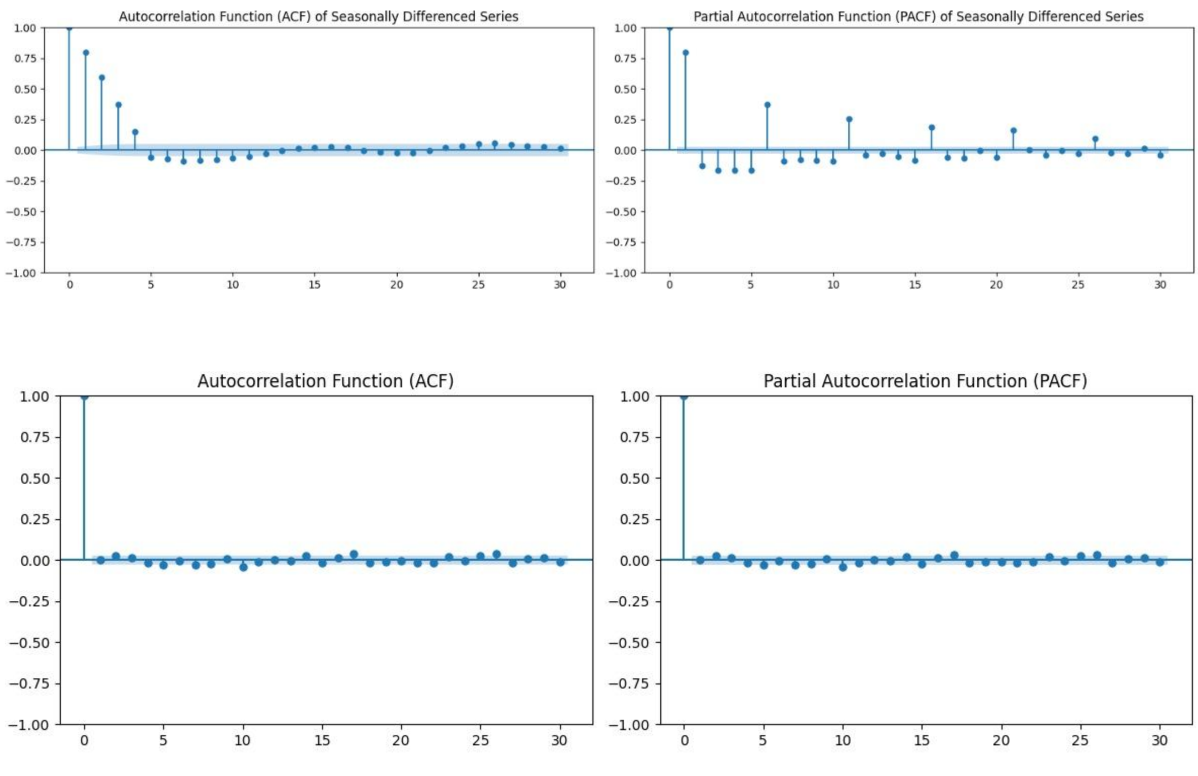
Evaluation Plot

3.1.9 Key Insights

- Superior ARIMA Performance:
  - The ARIMA model outperforms SARIMA across all four evaluation metrics, with particularly notable improvements in error measurements.
  - ARIMA reduces **RMSE by** 0.64 , **MAE by** 0.51, and **MAPE by** 0.2851% compared to SARIMA.
- High Explanatory Power:
  - Both models demonstrate exceptional fit to the data, with **R-squared values** above 0.98, indicating they explain over 98% of the variance.
  - The ARIMA model achieves **marginally better explanatory power**, with an **R-squared** of **0.9894**.
- Seasonal Component Impact:
  - The superior performance of the non-seasonal ARIMA model suggests that incorporating seasonal components (as in SARIMA) introduces unnecessary complexity that may obscure rather than enhance the forecasting accuracy for this particular time series



Line chart: Denoting all features



ACF & PACF plots

Conclusions:

The comparative analysis reveals that the **ARIMA model** offers superior forecasting performance for this financial time series. Despite the common assumption that seasonal models (like SARIMA) provide better predictions for financial data, our analysis demonstrates that the simpler **ARIMA approach** yields more accurate results for this dataset.

3.1.10 Practical Applications

- ➢ Investment Decision Support:
  - o More accurate forecasts enable better entry and exit point identification.
  - o Reduced prediction errors translate to lower investment risk.
  - o Seasonal insights provide additional market timing signals.
- ➢ Risk Management:
  - o Improved forecasting precision supports more effective hedging strategies.
  - o A better understanding of expected price movements aids portfolio allocation decisions.
  - o Seasonal patterns can be utilized for diversification purposes.
- ➢ Trading Strategy Development:
  - o **SARIMA forecasts** can serve as the foundation for algorithmic trading systems.
  - o Seasonal components can be isolated for specialized seasonal trading strategies.
  - o Combined with technical indicators, these forecasts may enhance trading performance.
- ➢ Financial Planning:
  - o More accurate stock price predictions support better financial planning.
  - o Institutional investors can better anticipate market movements for capital allocation.
  - o Individual investors gain improved tools for retirement and investment planning.

3.2 ARIMA and SARIMA Models for Air Passenger Forecasting

3.2.1 Dataset Description

This analysis utilizes the well-known **AirPassengers** dataset, which contains monthly totals of international airline passengers over a 12-year period.

- **Data Source**: AirPassengers.csv
- **Timespan**: January 1949 to December 1960
- **Format**: Monthly time series data

Key Variables

- **Date**: Monthly timestamp
- **Passengers**: Number of airline passengers (in thousands)

Data Characteristics

- **Starting value (Jan 1949)**: 112 passengers (in thousands)
- **Ending value (Dec 1960)**: 432 passengers (in thousands)
- The dataset exhibits a **strong upward trend** and **clear seasonal patterns**, making it ideal for seasonal time series modeling.

3.2.2 Data Preprocessing

Initial Data Cleaning

- Renamed columns:
  - 'Month' → 'Date'
  - '#Passengers' → 'Passengers'
- Converted 'Date' to datetime format using pd.to_datetime().
- Extracted **Year** and **Month** for exploratory data analysis.
- Confirmed **no missing or duplicate entries**.

Data Verification

- Ensured **chronological sorting** of observations.
- Performed exploratory analysis using:
  - Yearly and monthly groupings
  - Trend plots to observe growth and seasonality

3.2.3 Methodology

Stationarity and Parameter Selection

ADF Test (Augmented Dickey-Fuller)

- **Null Hypothesis (H_0)**: The time series has a unit root (i.e., it is non-stationary).
- **ADF test on original series**: Non-stationary ($p > 0.05$)
- Applied **first differencing**.
- **ADF test on differenced series**: Stationary ($p < 0.05$)

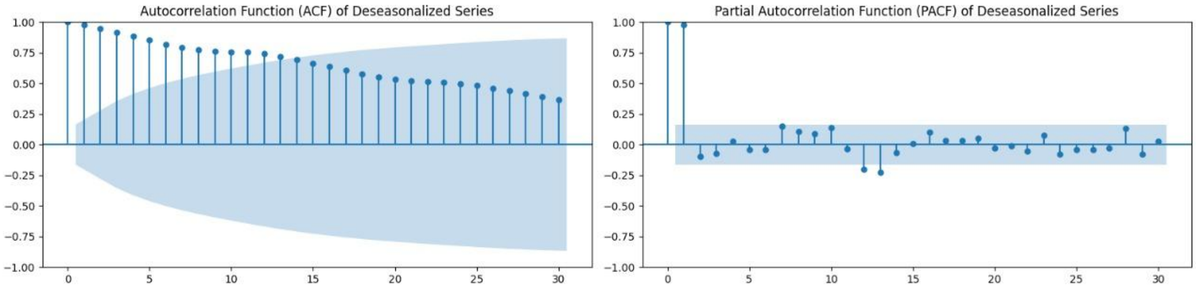
3.2.4 Model Order Determination

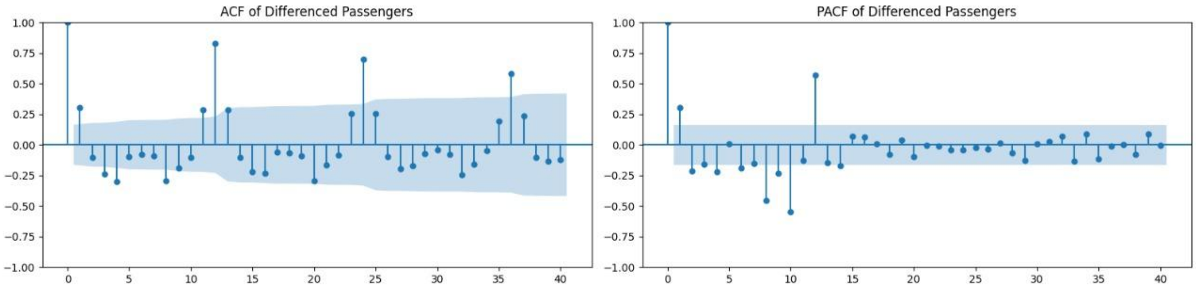
ARIMA (p, d, q) Selection

- p = 1: PACF showed a spike at lag 1
- d = 1: First differencing confirmed by ADF
- q = 1: ACF indicated MA component at lag 1
- **Final ARIMA model**: ARIMA(1,1,1)

SARIMA (p, d, q)(P, D, Q, s) Selection

- Seasonal PACF: Spike at lag 12 → P = 1
- Seasonal ACF: Gradual decay → Q = 1
- Seasonal frequency s = 12 (monthly)
- Seasonal differencing D = 1
- **Final SARIMA model**: SARIMA(1,1,1)(1,1,1,12)





3.2.5 Data Splitting for Model Evaluation

- **Training set**: First 80% (January 1949 – October 1958)
- **Testing set**: Remaining 20% (November 1958 – December 1960)

3.2.6 Results and Analysis

Evaluation Metrics

Metric	SARIMA	ARIMA	Improvement (SARIMA vs ARIMA)
RMSE	30.21	97.44	67.23 lower (≈69% reduction)
MAE	23.62	85.20	61.58 lower (≈72% reduction)

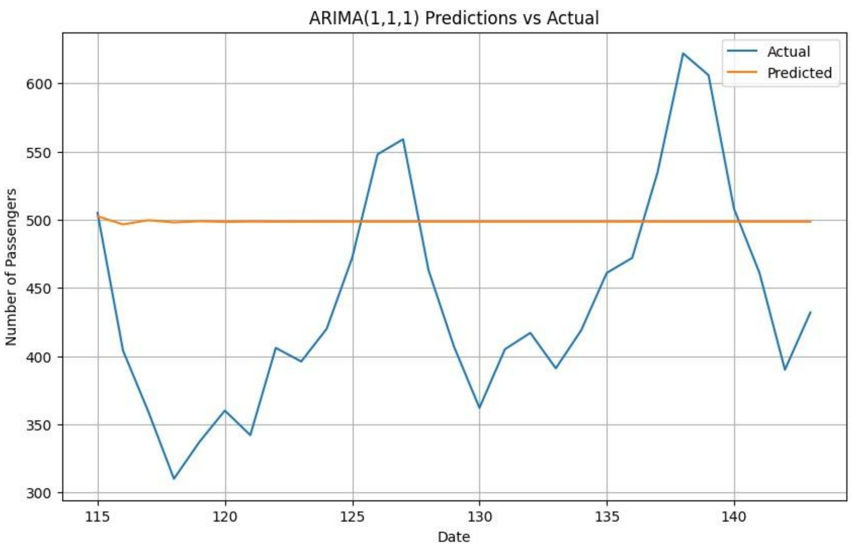
- **SARIMA outperformed ARIMA** significantly in both **Root Mean Square Error (RMSE)** and **Mean Absolute Error (MAE)**.
- The RMSE for SARIMA is **over three times lower** than that of ARIMA.
- The MAE shows a **72% error reduction** for SARIMA.

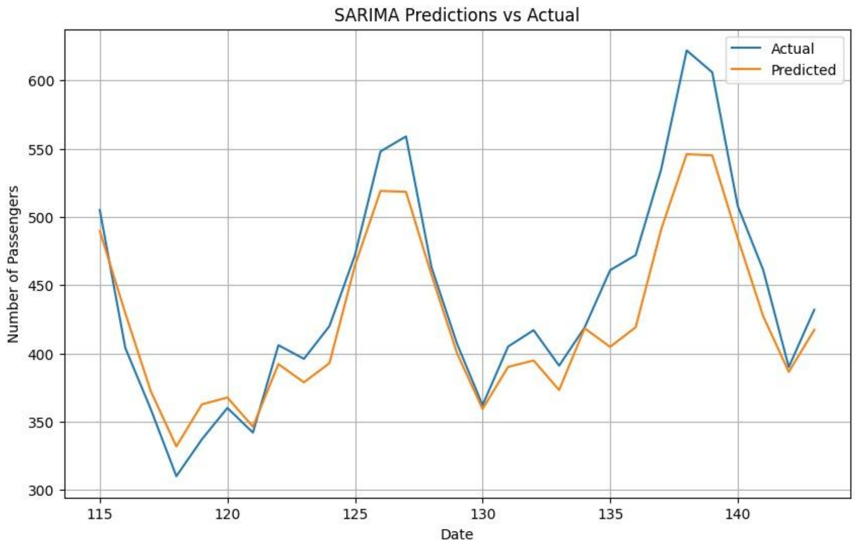
3.2.7 Performance Comparison

- **SARIMA** more accurately captured **recurring seasonal trends** in the dataset (e.g., consistent yearly peaks and troughs).
- **ARIMA** lacked the ability to model seasonality, resulting in **much higher prediction errors**.
- These results confirm the importance of **including seasonality** in forecasting models when working with periodic data.

3.2.8 Visual and Statistical Insights

- Forecast plots showed SARIMA closely tracking actual data.
- SARIMA exhibited **less volatility and more realistic seasonal peaks**, reinforcing its superiority in this context.





Conclusions

This time series analysis demonstrates that the **SARIMA model is significantly more effective** than the ARIMA model for datasets with **strong seasonal patterns**, such as the AirPassengers dataset.

3.2.9 Key Takeaways

- **SARIMA reduced RMSE by 69%** and **MAE by 72%** compared to ARIMA.
- The inclusion of seasonal parameters in SARIMA enables it to model **repeating annual trends**, which ARIMA fails to capture.
- For data with noticeable seasonality, **ARIMA alone is insufficient**, and seasonal models like SARIMA are essential for **accurate and reliable forecasting**.

3.2.10 Practical Applications

- **Airline Demand Forecasting**:

  Accurate predictions help with **staffing, inventory, and capacity planning**.

- **Seasonal Marketing and Pricing**:

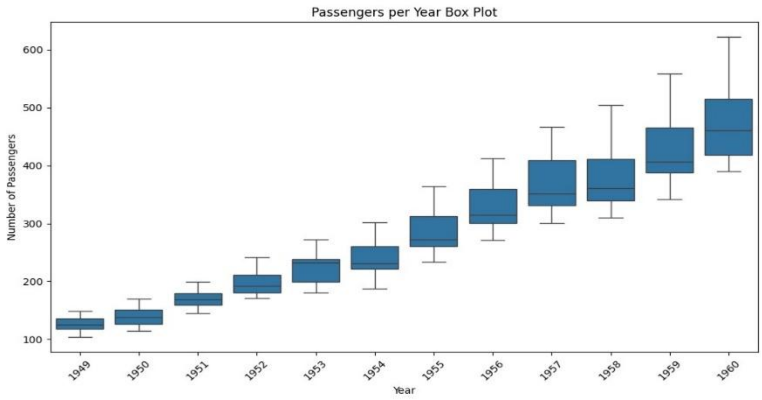
  Identifying **peak travel seasons** allows for better **pricing strategies** and **promotional planning**.

- **Infrastructure Planning**:

  Long-term demand forecasting supports **investment decisions** and **network expansion**.

3.3 Visualization

3.3.1 Box Plot: Passengers per Year



- **Growth Over Time:**

  Each year shows a higher range and median of passenger numbers, confirming the increasing trend.

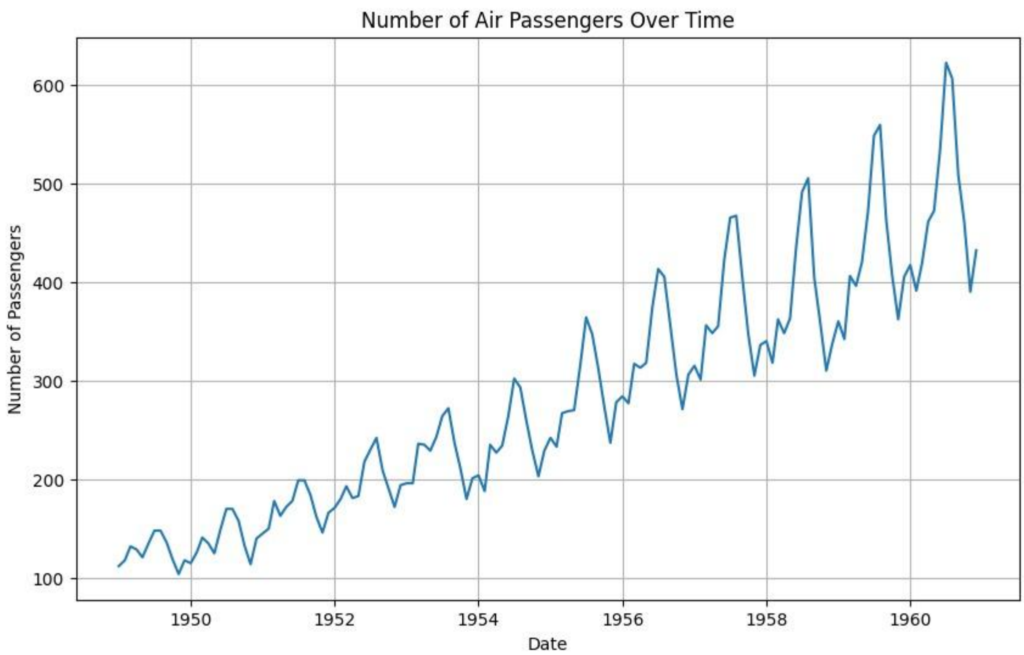
- **Wider Ranges:**

  The height of the boxes and whiskers increases over the years, indicating that passenger numbers became more variable.

- **More Outliers in Later Years:**

  The later years (1957–1960) show a broader spread, possibly hinting at new market conditions, routes, or external factors like economic expansion.

3.3.2 Line Plot: Number of Air Passengers Over Time



- **Upward Trend:**

  The number of air passengers steadily increased from 1949 to 1960, showing clear long-term growth in air travel.

- **Seasonal Pattern:**

  There's a **regular repeating wave pattern** each year — peaks and dips occur at consistent intervals, indicating **seasonality** (likely due to holidays or travel seasons).

- **Volatility Increases Over Time:**

  The size of fluctuations grows, suggesting increasing variability as more people begin to travel.