Forecasting Family Physician Availability: Trends and Dynamics in New Brunswick and Canada

Marie-Line Forbes

July 23, 2024

Introduction

Background:

The availability of family physicians is crucial for maintaining public health across Canada. Understanding historical trends and dynamics is essential for addressing shortages effectively.

This research focuses on New Brunswick, analyzing 52 years of data on family physicians from 1971 to 2022 and forecasting future availability using advanced time series analysis and machine learning techniques (ARIMA), while comparing with other provinces.

Objectives of the Study :

- Conduct descriptive statistical analysis to understand historical trends in the supply and distribution of family physicians in New Brunswick.
- Develop and validate a predictive model to forecast the number of family physicians in New Brunswick beyond 2022.
- Scale the predictive model using High Performance Computing (HPC) to apply it to all provinces in Canada.
- Present the main findings, challenges faced, and future steps for further development.

Dataset Overview

- Supply, Distribution and Migration of Physicians in Canada,1971 to 2022 — Historical Data
- Provides 52 years of supply and distribution data for physicians in Canada by specialty, including demographic, education and migration information.
- The dataset compiles data from Canada’s provinces and territories, provided by the *Canadian Institute for Health Information.*
- Dataset Size: 101,697 rows and 60 columns

Canadian Institute for Health Information
Institut canadien d’information sur la santé

	Year	Jurisdiction	Health region	Specialty	Specialty sort	Physician-to–100,000 population ratio	Number of physicians	Number male	Number female	Number sex unknown	Average age
1	1971	Canada	Canada	All physicians	1	124.81085539	27411	24007	2182	1222	44.9
2	1971	Canada	Canada	All specialists	2	62.466897416	13719	12317	860	542	46.1
3	1971	Canada	Canada	Family medicine	3	62.343957973	13692	11690	1322	680	43.8
4	1971	Canada	Canada	__General practice	4	60.69110545	13329	11350	1311	668	43.8
5	1971	Canada	Canada	__Family medicine	6	1.6528525229	363	340	11	12	43.4
6	1971	Canada	Canada	Medical specialists	7	37.58759663	8255	7213	729	313	45.4
7	1971	Canada	Canada	Clinical specialists	8	34.505003909	7578	6626	658	294	45.5
8	1971	Canada	Canada	_Anesthesiology	9	5.4229954678	1191	1018	129	44	46.2
9	1971	Canada	Canada	_Dermatology	10	0.928875798	204	167	25	12	46.9
10	1971	Canada	Canada	_Diagnostic radiology	11	4.421266666	971	883	53	35	44.8
11	1971	Canada	Canada	_Internal medicine	15	7.9865105378	1754	1585	97	72	46.4
12	1971	Canada	Canada	__Cardiology	16	0.792276416	174	160	7	7	45.4

EDA & Descriptive Statistic Analysis

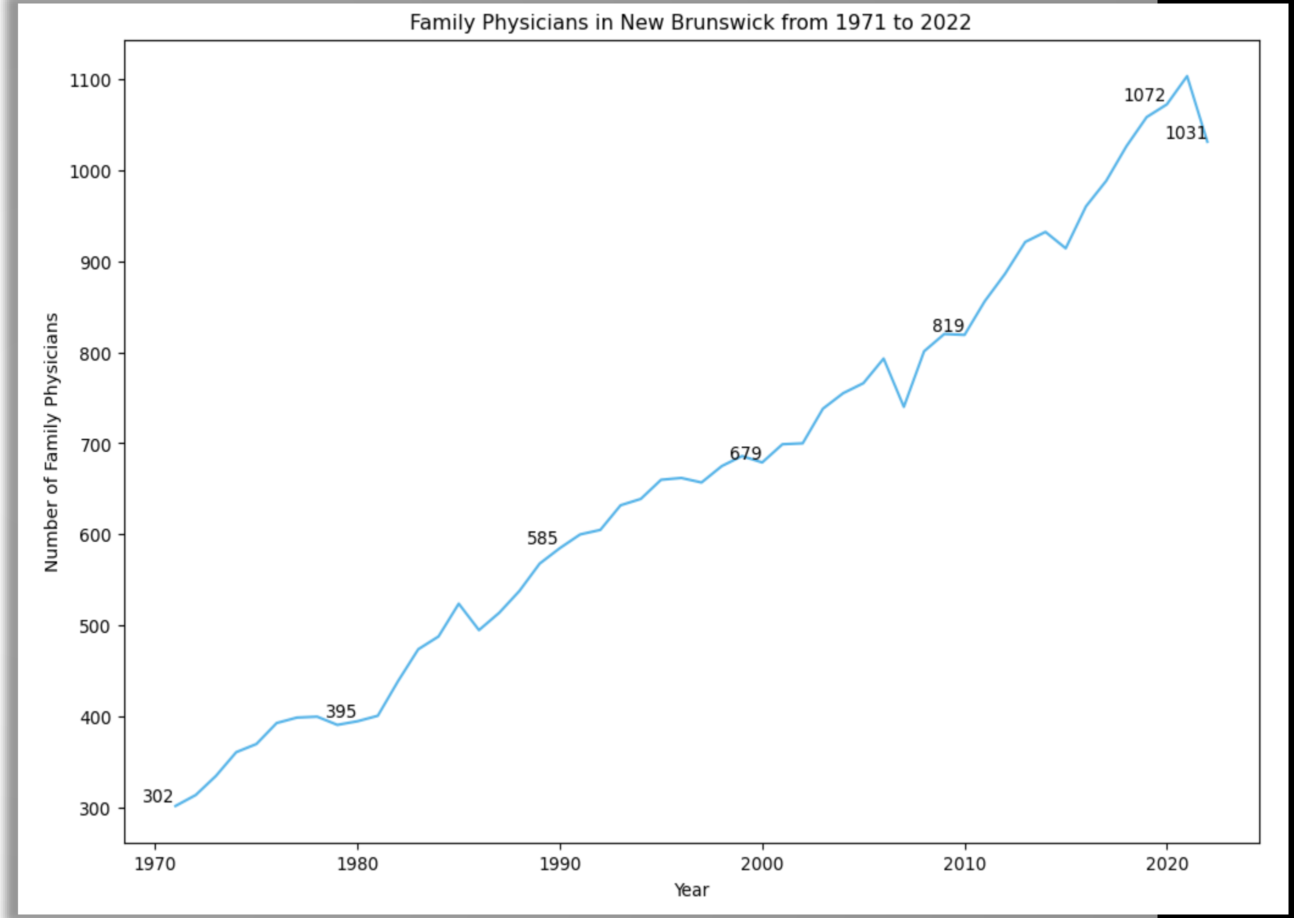
- Initial Data Cleaning and Processing
- Create the two main DataFrame for the research

```
# Create the dataframe for 'N.B.' jurisdiction and 'Family medicine' specialty
family_medicine_df_nb = df[
    (df['Specialty sort'] == 3) &
    (
        ((df['Jurisdiction'] == 'N.B.') & (df['Health region'] == 'N.B.'))
    )
]
```

Columns with missing values in family_medicine_df_nb and the count:
Number of physicians who returned from abroad 1
Number of physicians who moved abroad 1
Net migration between Canadian jurisdictions 30

```
# Create the dataframe for 'Canada' jurisdiction and 'Family medicine' specialty
family_medicine_df_canada = df[
    (df['Specialty sort'] == 3) &
    (
        ((df['Jurisdiction'] == 'Canada') & (df['Health region'] == 'Canada'))
    )
]
```

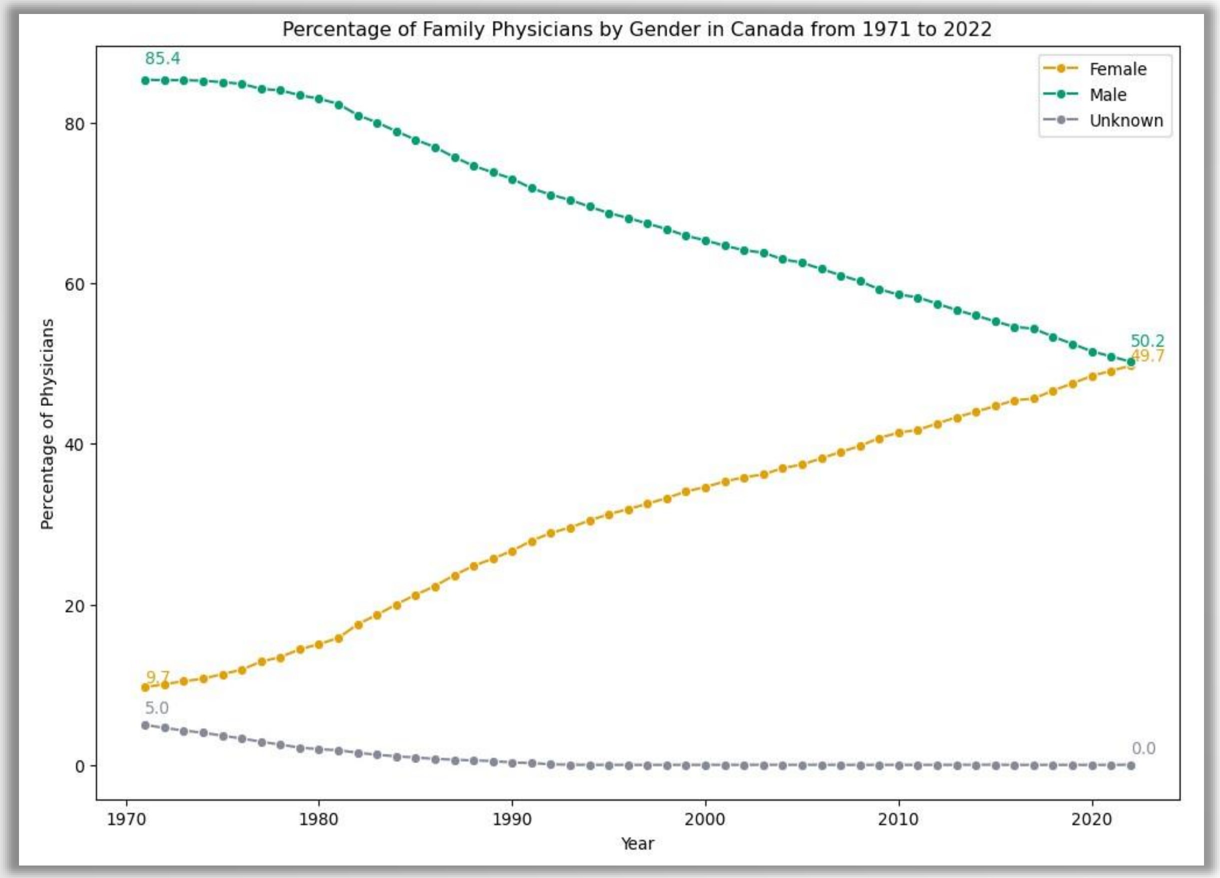
Columns with missing values in family_medicine_df_canada and the count:
Number of physicians who returned from abroad 1
Number of physicians who moved abroad 1
Net migration between Canadian jurisdictions 30

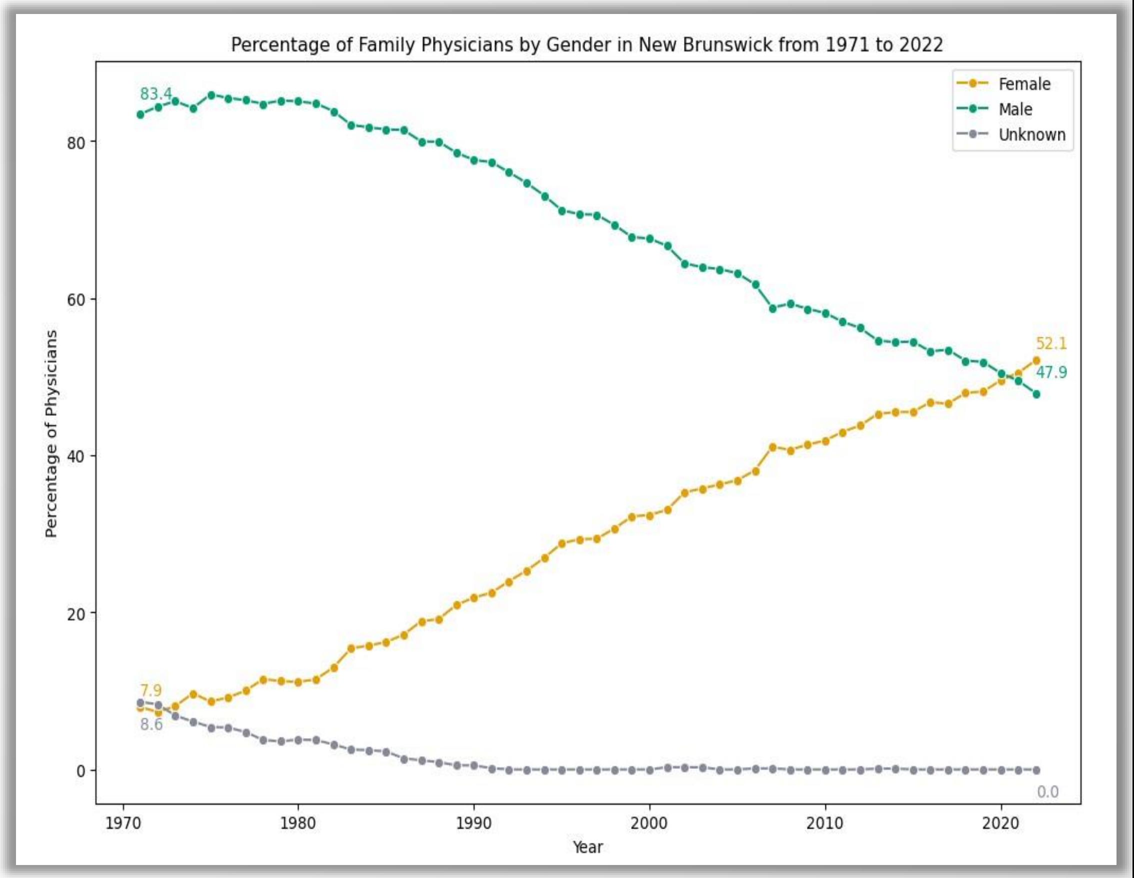


- Explore key trends and Insights from family_medicine_df_nb
  &
  family_medicine_df_canada

EDA & Descriptive Statistic Analysis

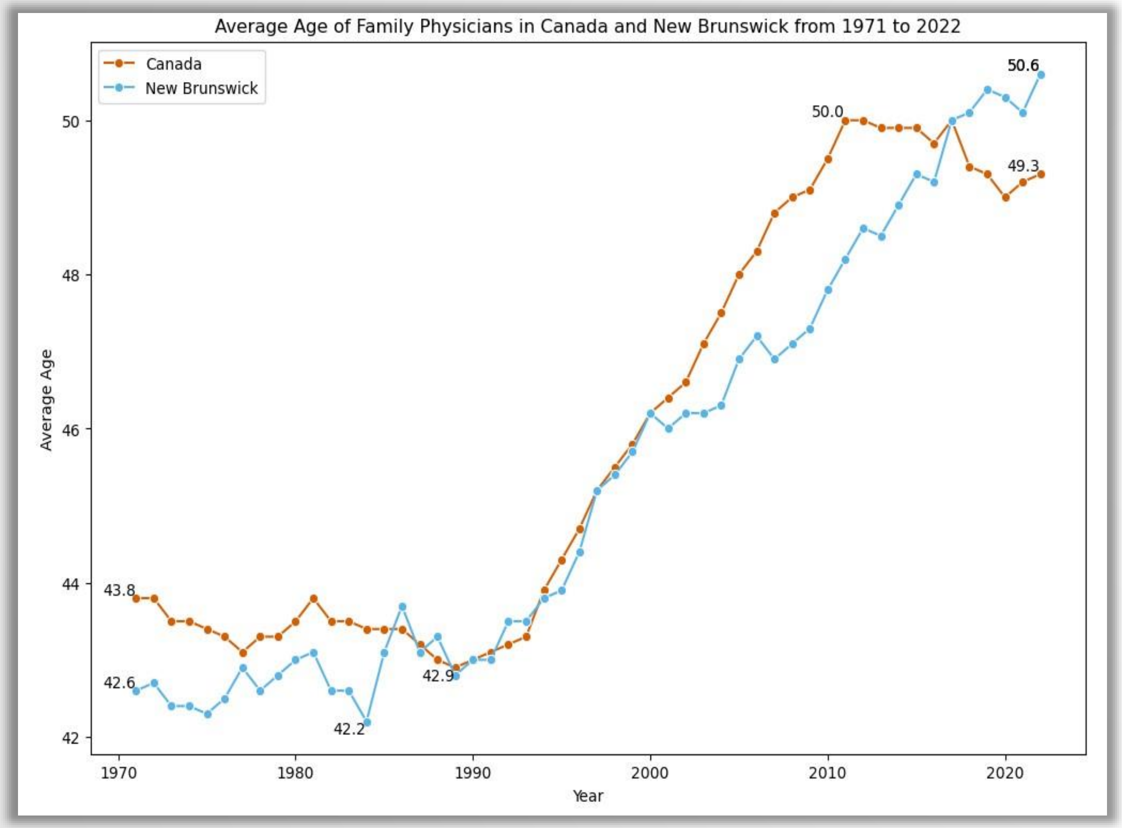
Percentage of Family Physicians by Gender in Canada & NB from 1971 to 2022

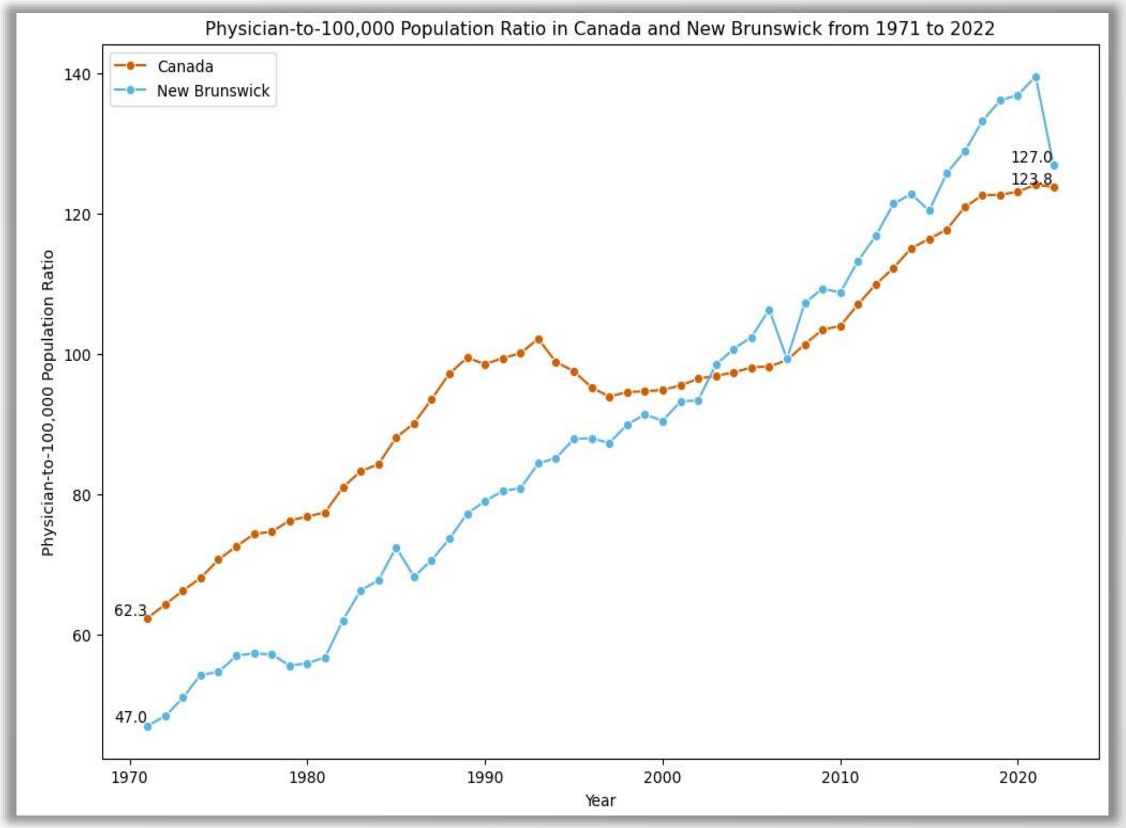




EDA & Descriptive Statistic Analysis

Average Age and Physician-to-100,000 population ratio of Family Physicians in Canada & NB from 1971 to 2022





Machine Learning Model : ARIMA

- ARIMA (AutoRegressive Integrated Moving Average) is a statistical model used for forecasting time series data by combining **autoregression**, **differencing**, and **moving average** techniques. It captures patterns and trends in past data to predict future values.

- Must define the 3 parameters in ARIMA model (p,d,q)

p = the order of the Autoregressive part of ARIMA

d = the degree of differencing involved

q = the order of the Moving Average part

```
# Test whether the series is stationary or non-stationary :  d parameter
#  Augmented Dickey-Fuller Test : significance level a, 0.05.
#H0 (Null hypothesis) = Time series non-stationary
#H1 (Alternative hypothesis) = Time series is stationary
adf_test = adfuller(family_medicine_df_nb['Number of physicians'])
# Output the results
print('ADF Statistic: %f' % adf_test[0])
print('p-value: %f' % adf_test[1])
print("Fail to reject the null hypothesis, the time series is non-stationary, p value greater than 0.05
")

# apply first differencing transformation
diff_data = family_medicine_df_nb['Number of physicians'].diff().dropna()
# Perform ADF test on the differenced series
adf_test_diff = adfuller(diff_data)
# Output the results
print('ADF Statistic (after differencing): %f' % adf_test_diff[0])
print('p-value: %f' % adf_test_diff[1])
print("After one differencing we achieve stationarity, d is equal to 1. ")
```

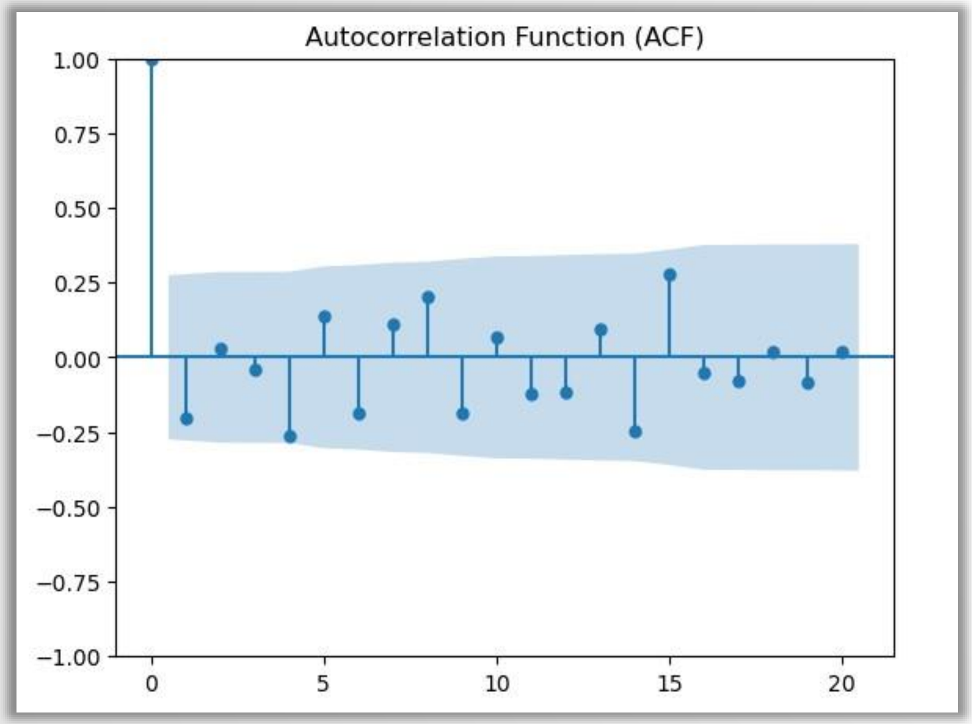
- Before applying ARIMA, we must ensure the time series is **stationary**, meaning the mean, variance, and autocorrelation remain constant over time.

- We test for stationarity using the Augmented Dickey-Fuller Test, which helps determine the appropriate **d** parameter for differencing if the series is non-stationary.

```
ADF Statistic: 0.023263
p-value: 0.960414
Fail to reject the null hypothesis, the time series is non-stationary, p value greater than 0.05
ADF Statistic (after differencing): -7.763177
p-value: 0.000000
After one differencing we achieve stationarity, d is equal to 1.
```

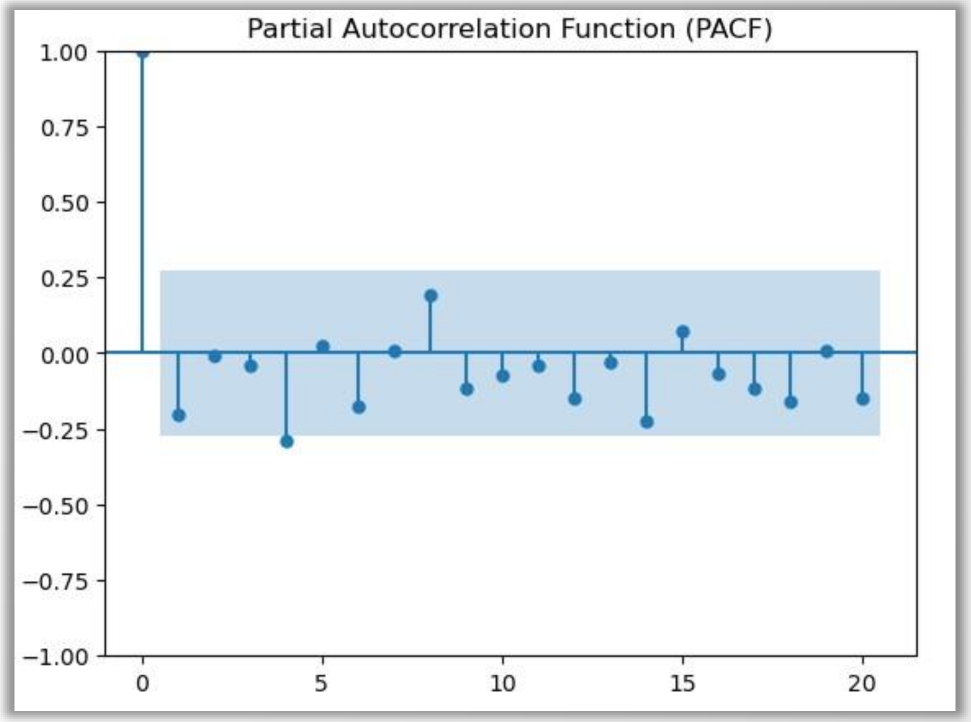
Machine Learning Model : ARIMA

ACF Plot to determine q parameter



q = 1

PACF Plot to determine p parameter



p = 1

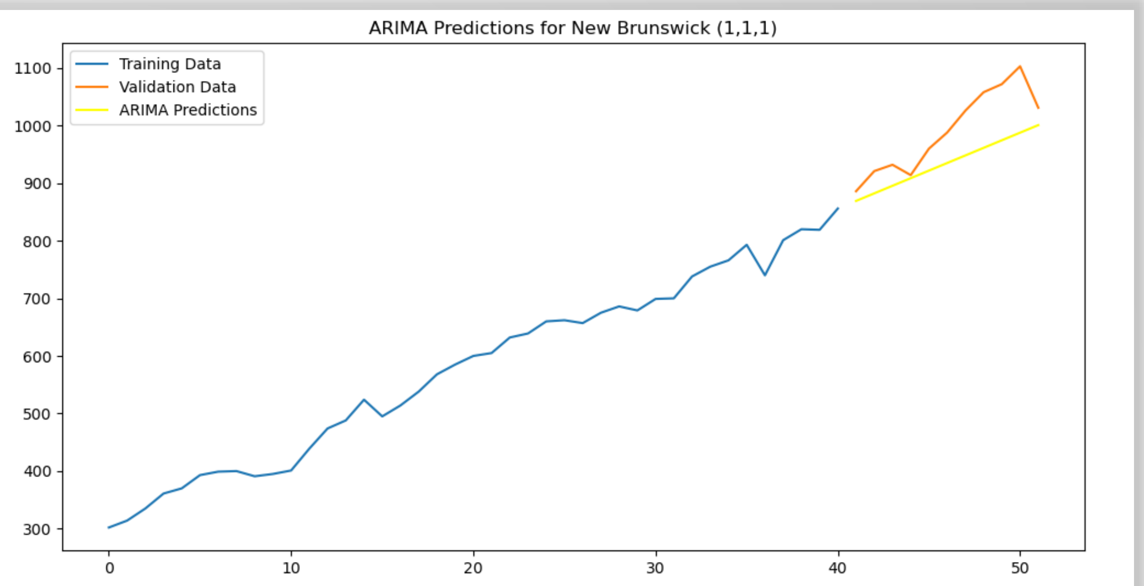
Machine Learning Model : ARIMA

Apply the parameters (1,1,1) , and create ARIMA model with train & validation data

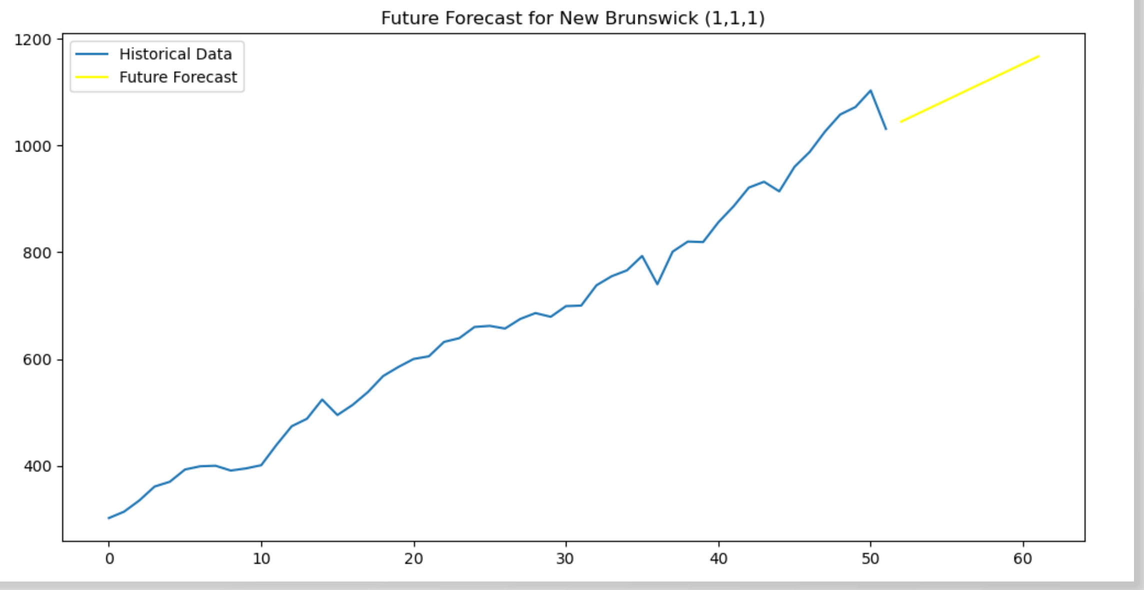
```
# Split data into train and validation sets
train_size = int(len(data) * 0.8)
train, validation = data[:train_size],
data[train_size:]

# Fit the ARIMA model
ARIMAmodel = ARIMA(train, order=(1, 1, 1))
ARIMAmodel = ARIMAmodel.fit()

# Make predictions on the validation set
y_pred =
ARIMAmodel.get_forecast(steps=len(validation))
y_pred_df = y_pred.conf_int(alpha=0.05)
y_pred_df["Predictions"] =
ARIMAmodel.predict(start=validation.index[0],
end=validation.index[-1])
y_pred_df.index = validation.index
y_pred_out = y_pred_df["Predictions"]
  return go(f, seed, [])
]
```



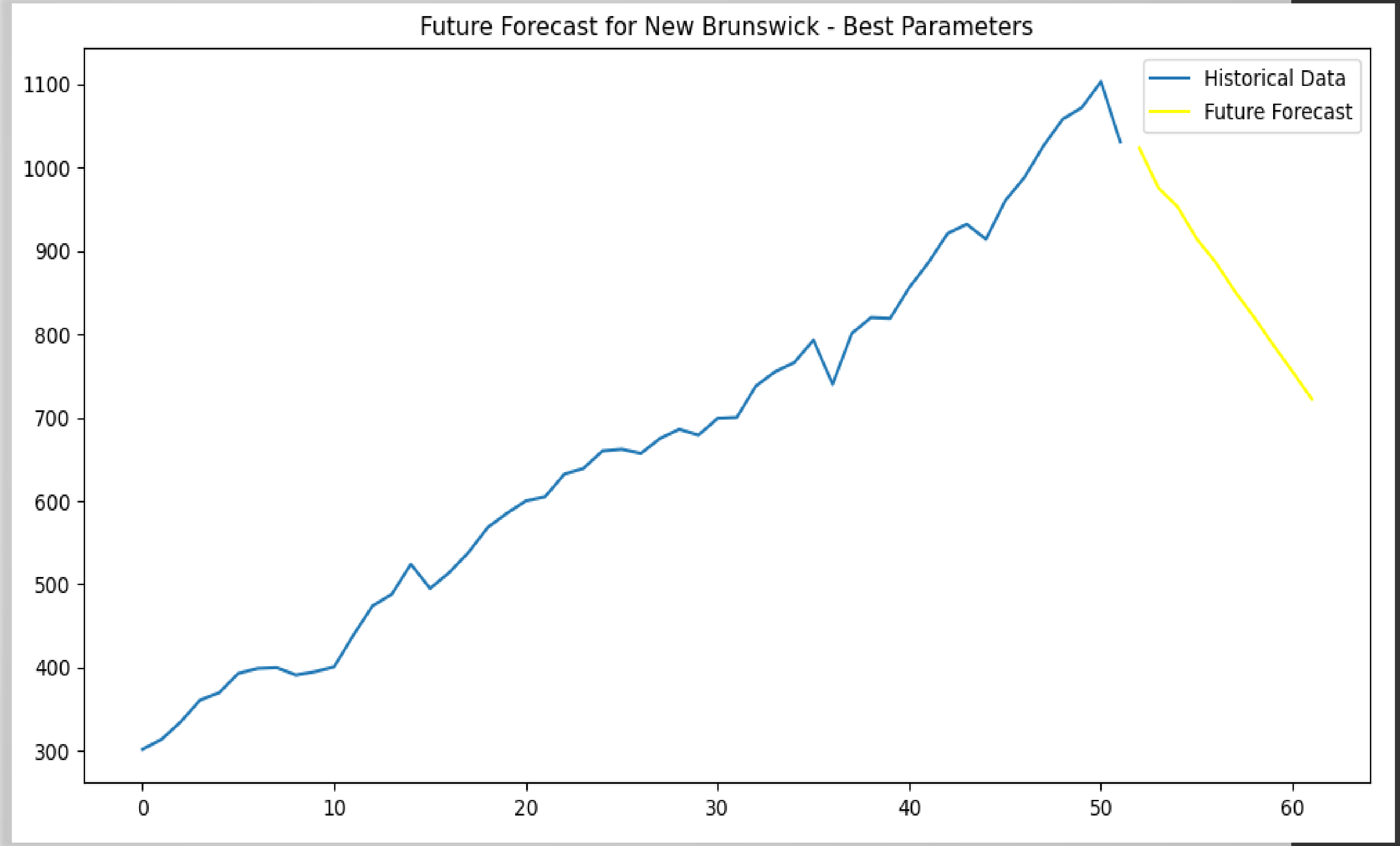
RMSE: 65.01411891897584



Machine Learning Model : ARIMA

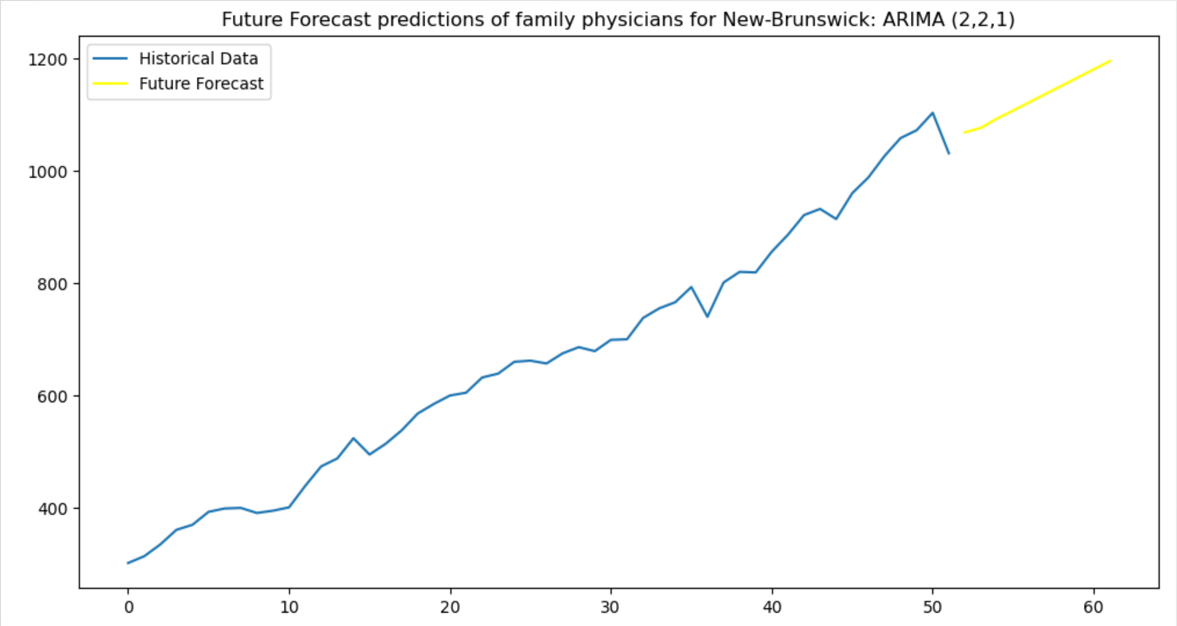
- Grid search to find best parameters (1,2,0)

```
ARIMA(0, 0, 0) RMSE=418.5866473822011
ARIMA(0, 0, 1) RMSE=411.02414157451085
ARIMA(0, 0, 2) RMSE=401.96683958361064
ARIMA(0, 1, 0) RMSE=150.90605144803294
ARIMA(0, 1, 1) RMSE=148.70074710593383
ARIMA(0, 1, 2) RMSE=164.34234230351169
ARIMA(0, 2, 0) RMSE=105.96526017350598
ARIMA(0, 2, 1) RMSE=60.83280452251429
ARIMA(0, 2, 2) RMSE=68.25384715986256
ARIMA(1, 0, 0) RMSE=158.3180055404624
ARIMA(1, 0, 1) RMSE=157.0581984611855
ARIMA(1, 0, 2) RMSE=180.8555734736873
ARIMA(1, 1, 0) RMSE=146.87128750057596
ARIMA(1, 1, 1) RMSE=65.01411891897584
ARIMA(1, 1, 2) RMSE=163.39170793306144
ARIMA(1, 2, 0) RMSE=27.919699027828877
ARIMA(1, 2, 1) RMSE=66.4029024046678
ARIMA(1, 2, 2) RMSE=70.3816806535317
ARIMA(2, 0, 0) RMSE=156.07154469830326
ARIMA(2, 0, 1) RMSE=158.0963276995034
ARIMA(2, 0, 2) RMSE=180.9027397885611
ARIMA(2, 1, 0) RMSE=129.86895342407857
ARIMA(2, 1, 1) RMSE=68.72812518451289
ARIMA(2, 1, 2) RMSE=70.74741014472218
ARIMA(2, 2, 0) RMSE=38.42435406527627
ARIMA(2, 2, 1) RMSE=62.89710100610015
ARIMA(2, 2, 2) RMSE=68.78093510038534
Best ARIMA(1, 2, 0) RMSE=27.919699027828877
```

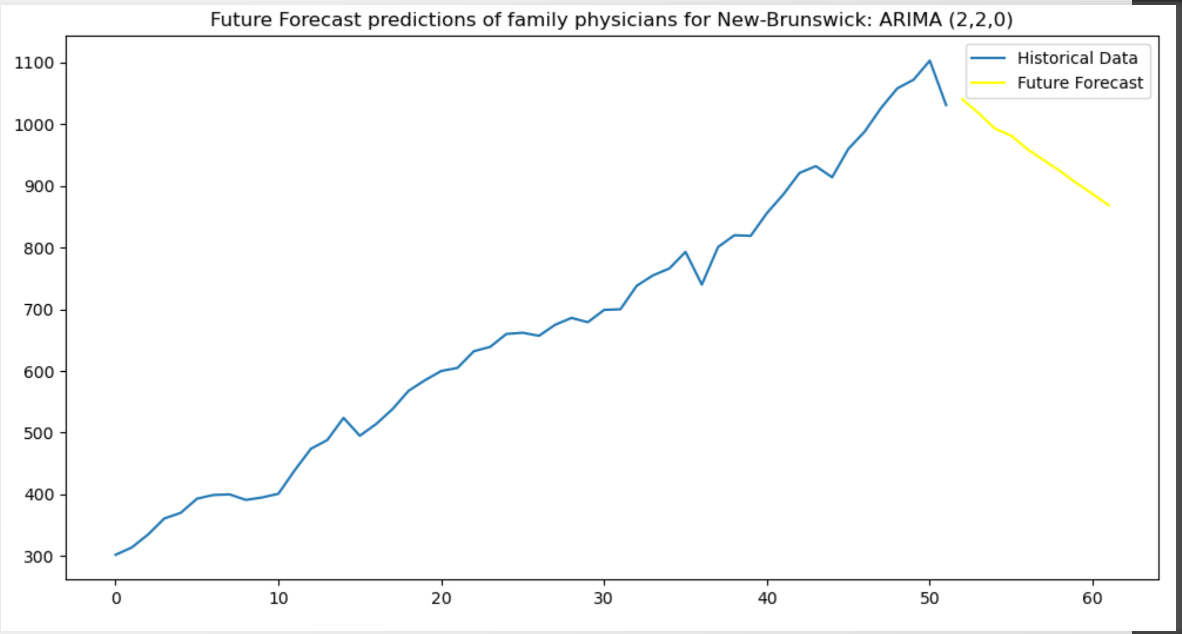


Machine Learning Model : ARIMA

- Testing other parameters from the Grid search



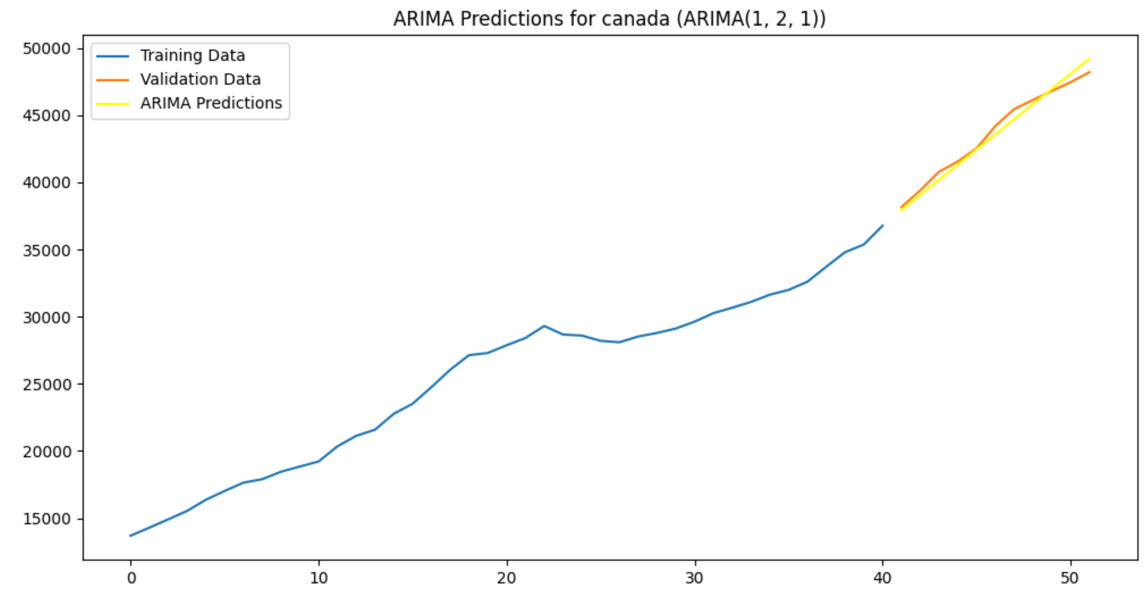
RMSE=62.89710100610015



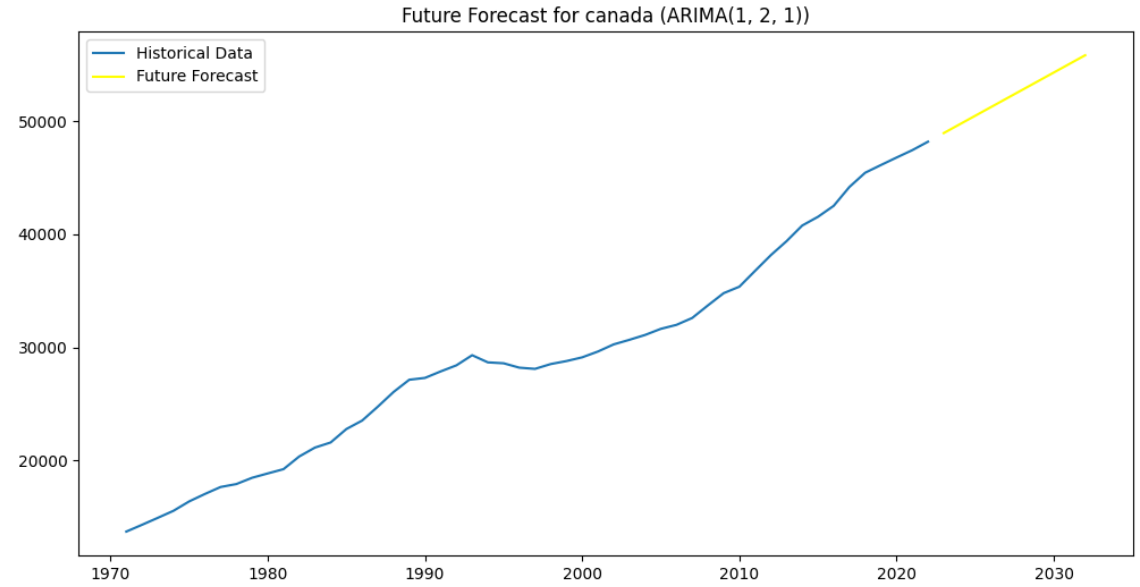
RMSE=38.42435406527627

HPC Implementation

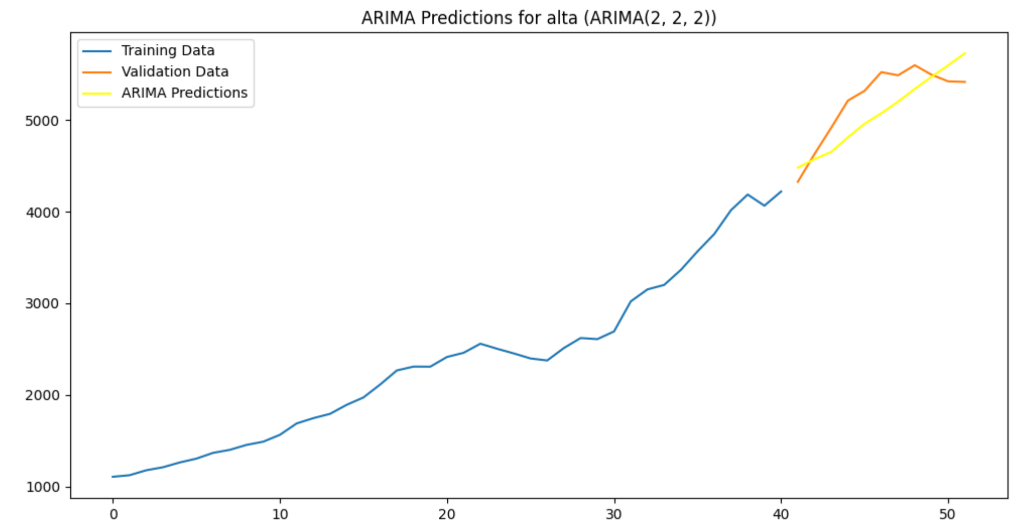
- Automation of the ARIMA Model including Grid Search for best parameters
- Loop through each provinces and territories
- Print Predictions and Future Forecast for each jurisdictions

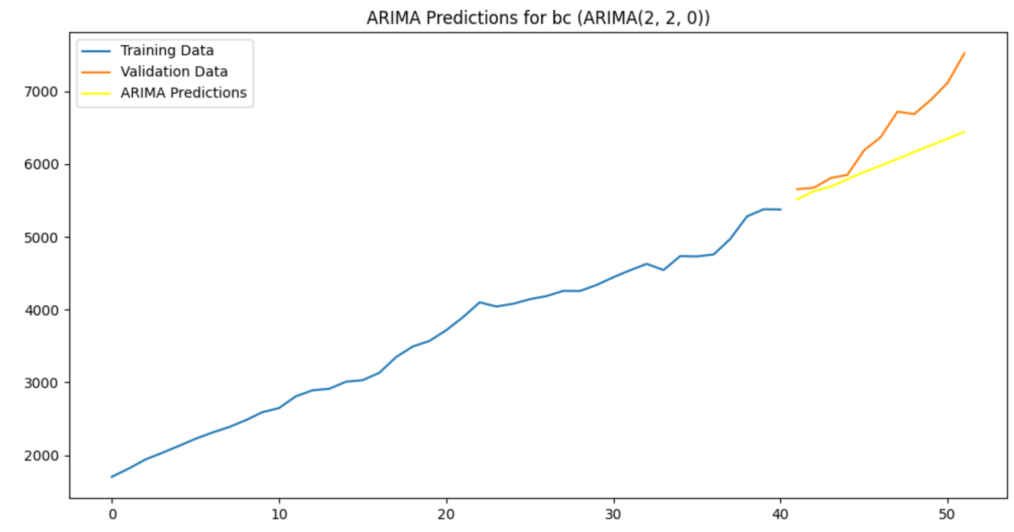


RMSE=532.2772364877885



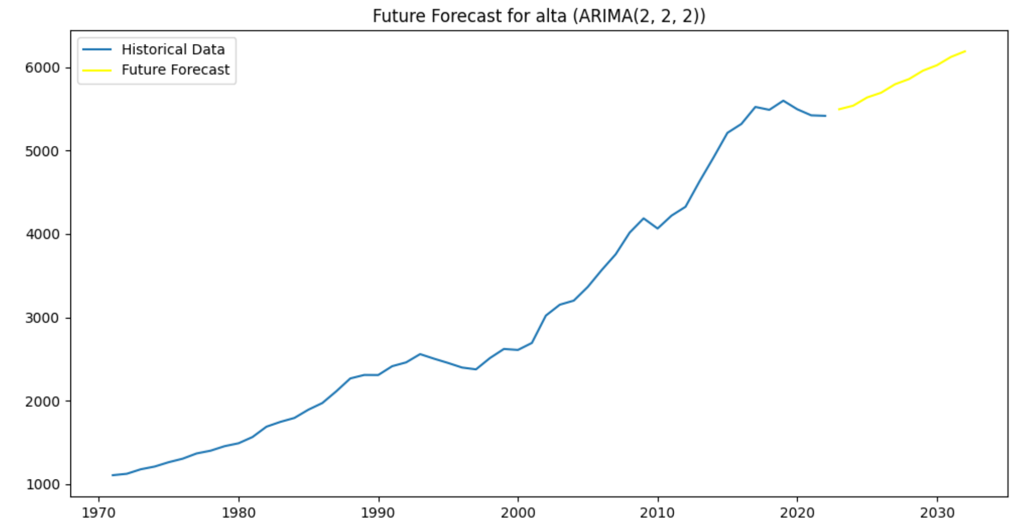
HPC Results

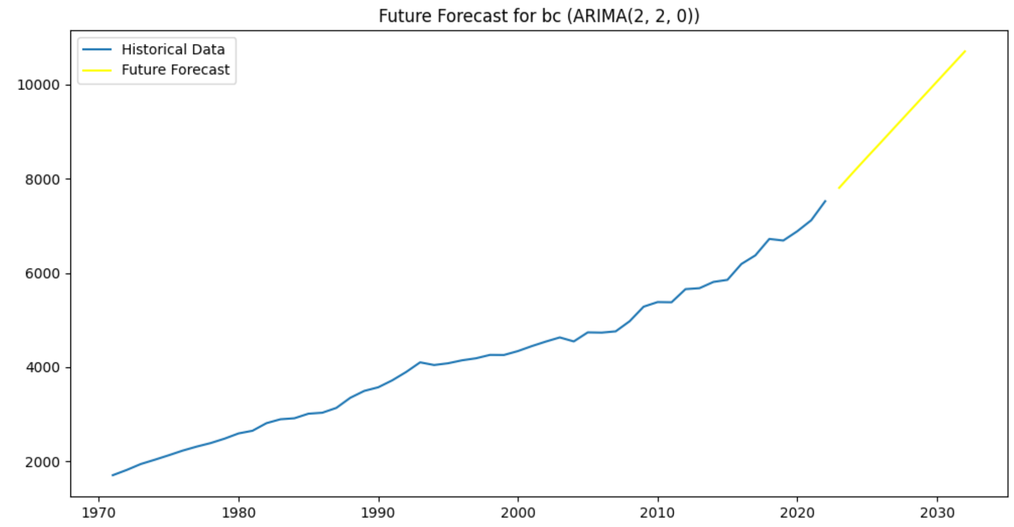




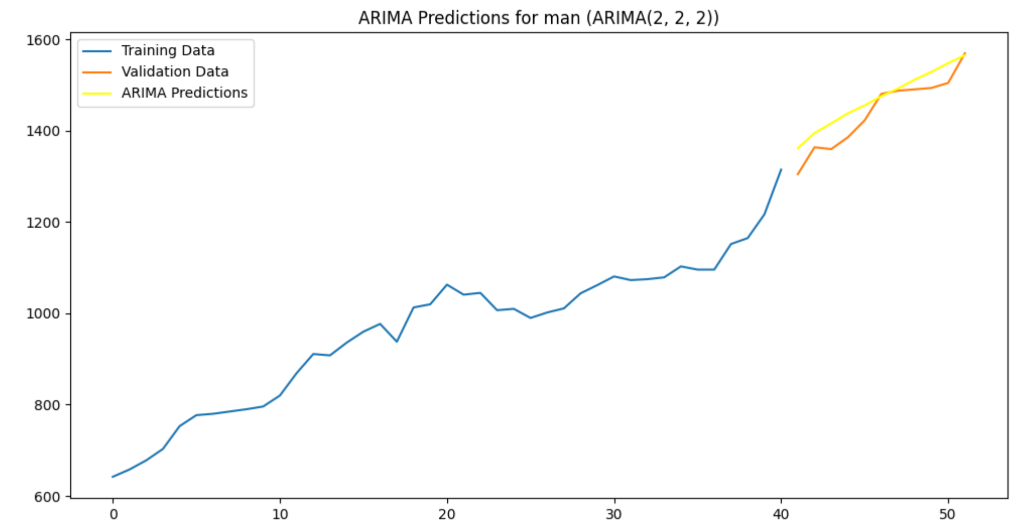
RMSE=280.9097975482264

RMSE=532.2616847112556

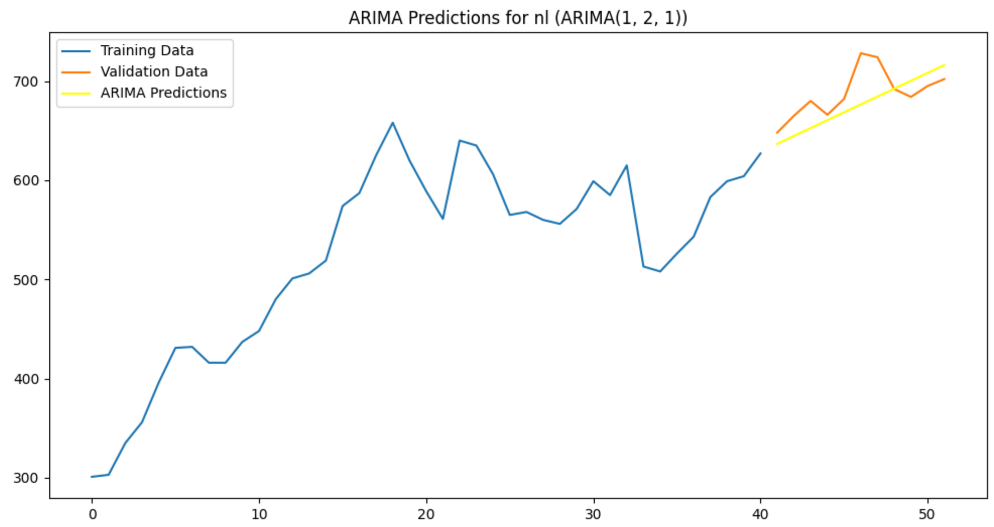




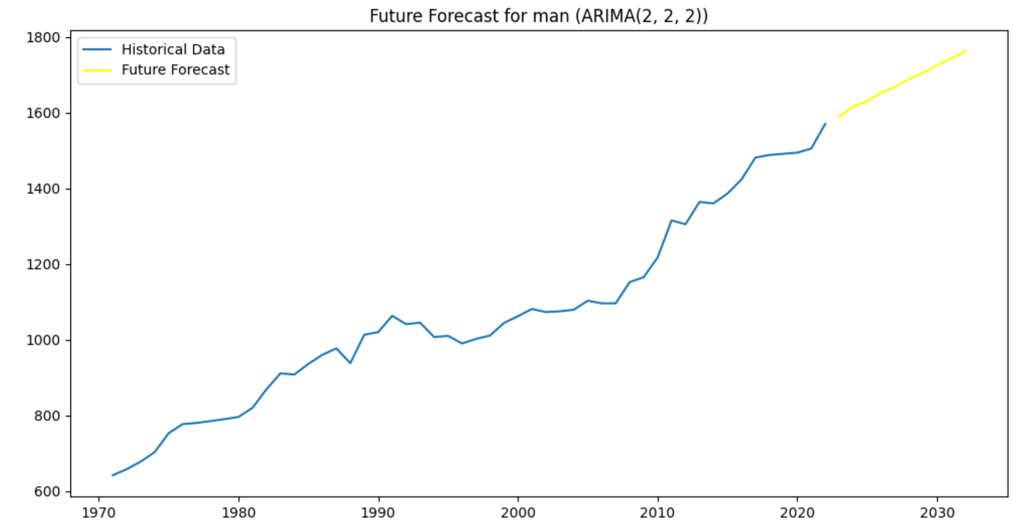
HPC Results

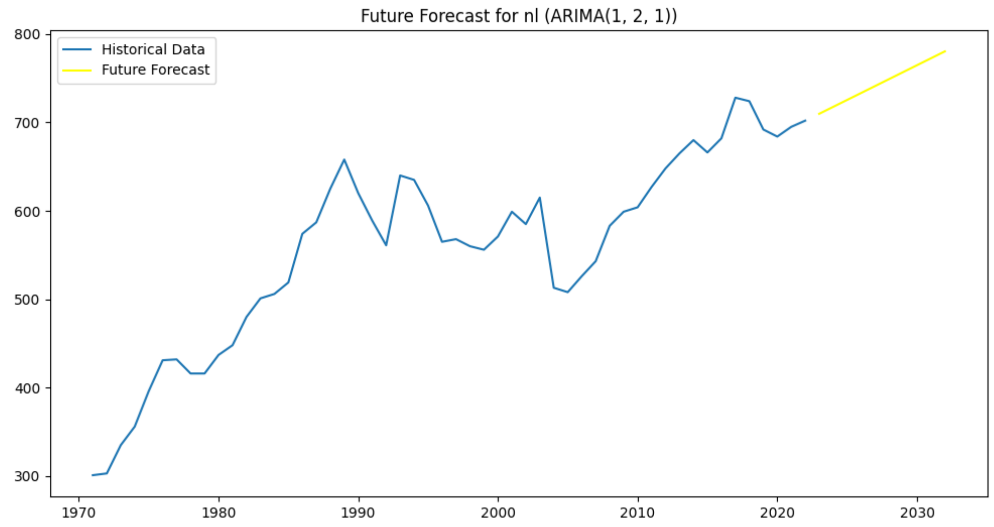


RMSE=36.782312015979485

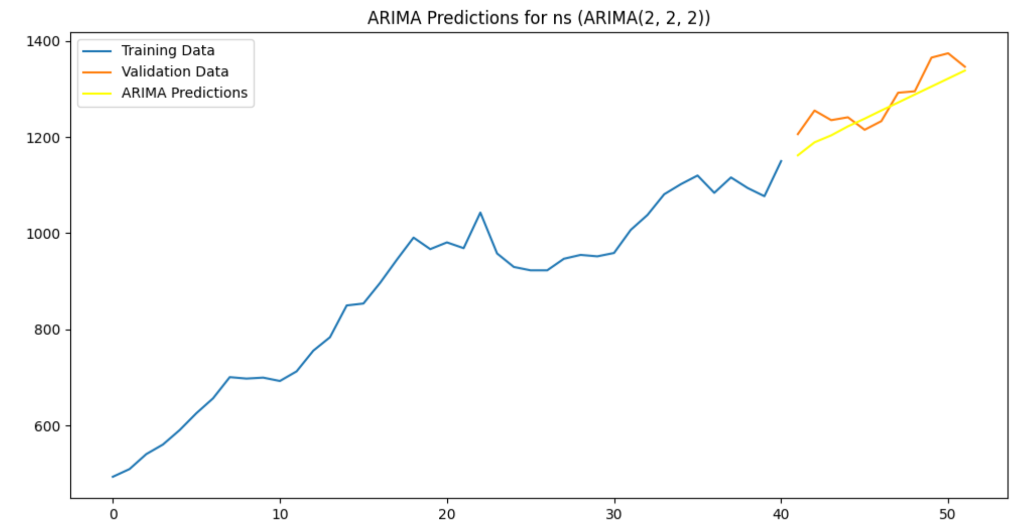


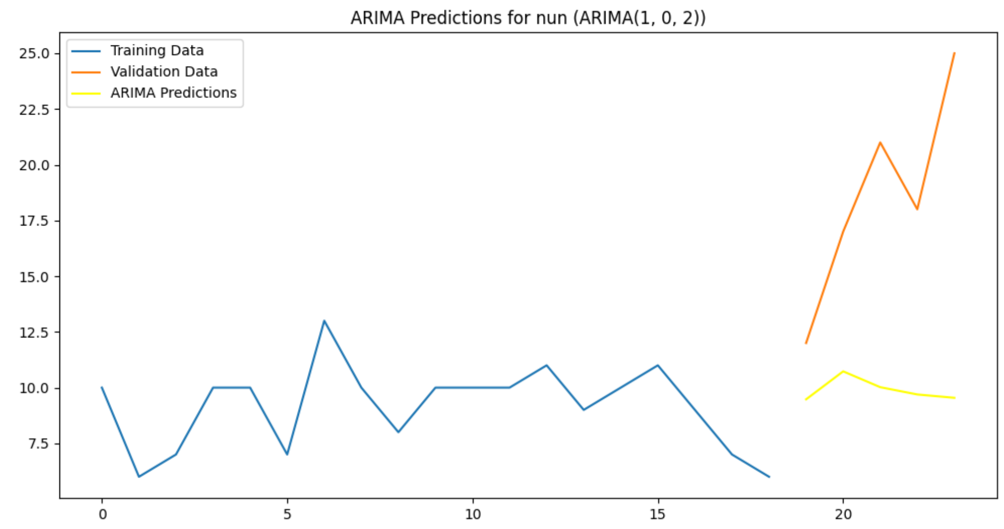
RMSE=24.063562944225193





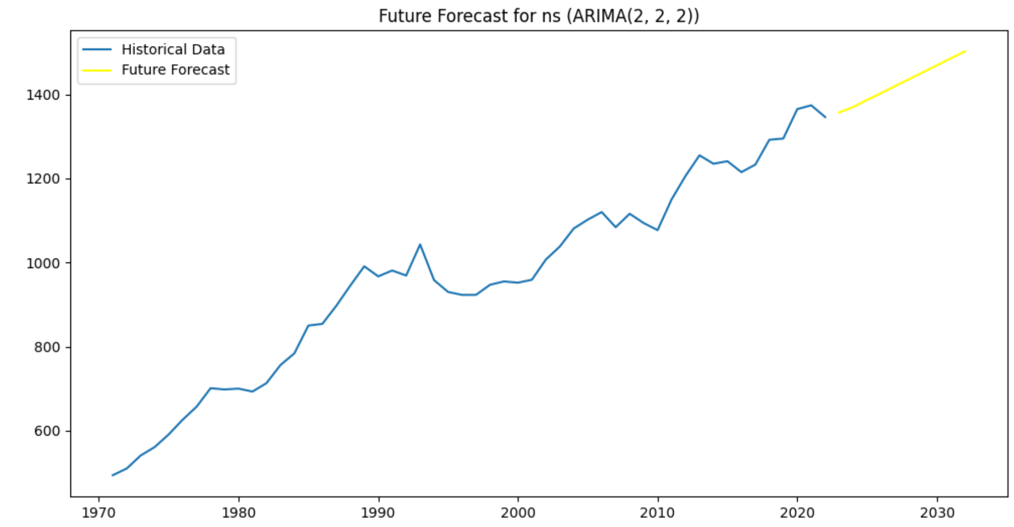
HPC Results

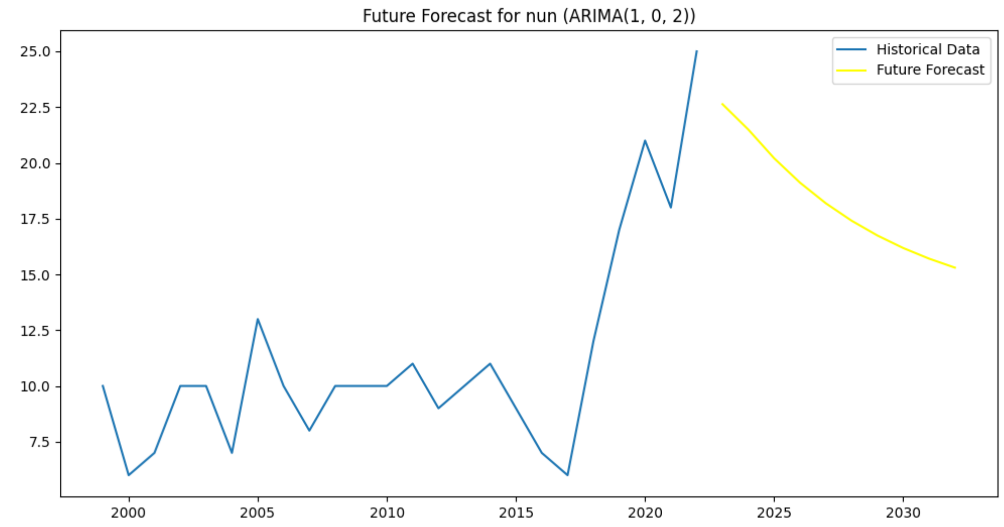




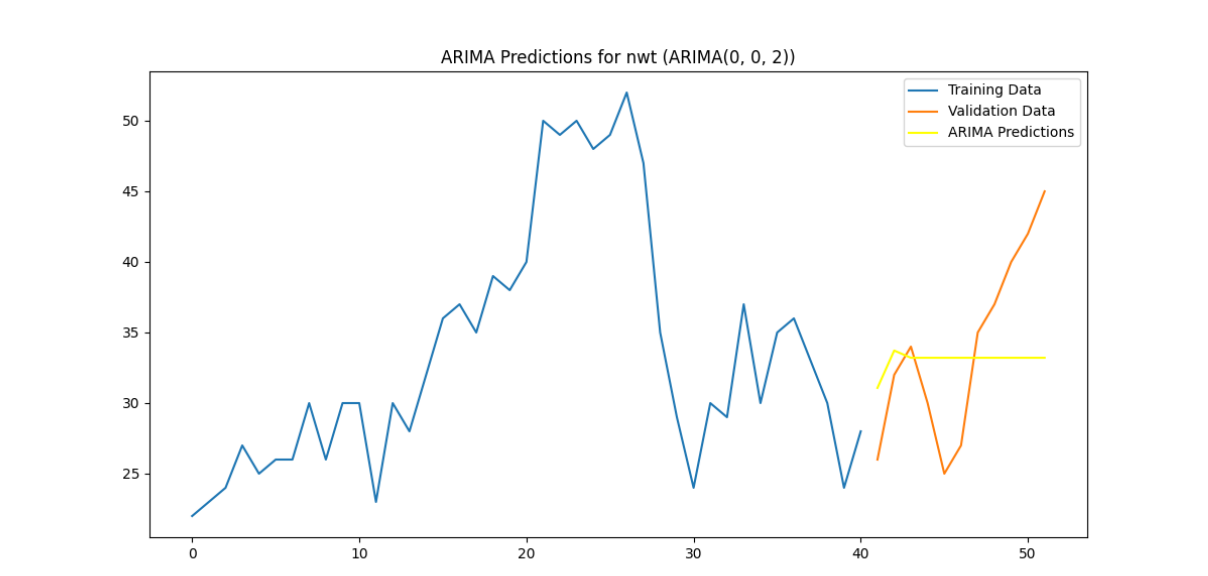
RMSE=37.63379879865523

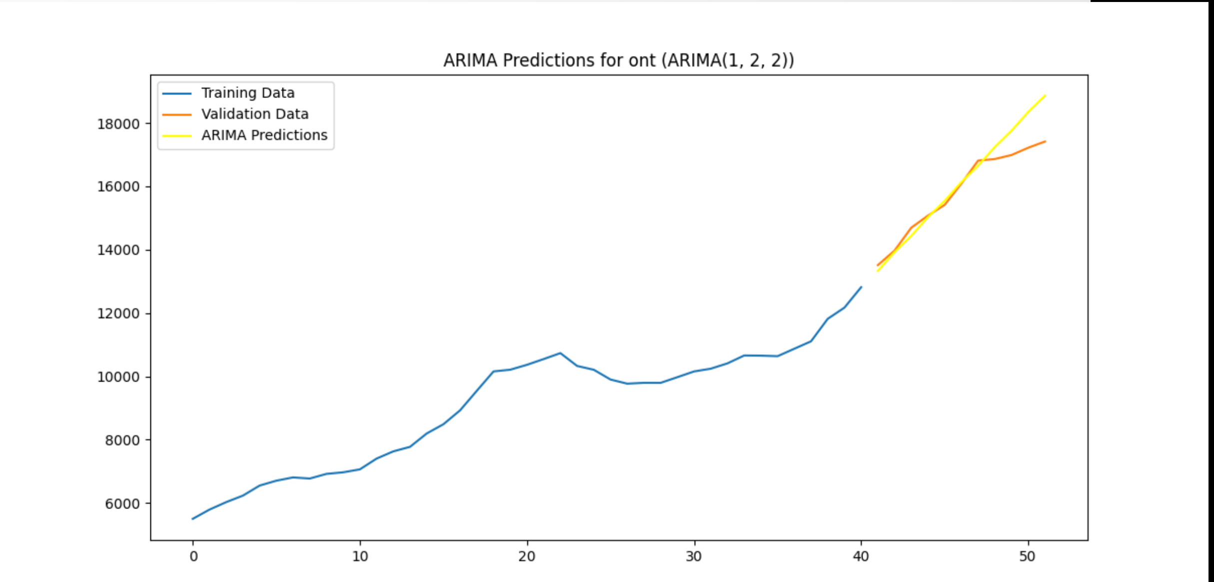
RMSE=9.738591823198291





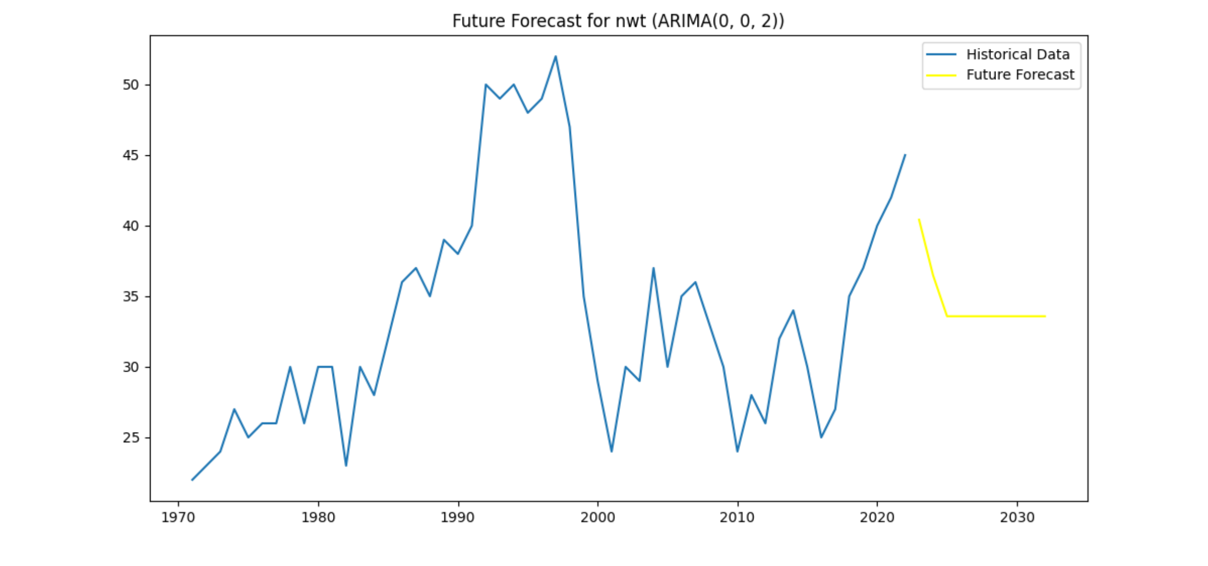
HPC Results

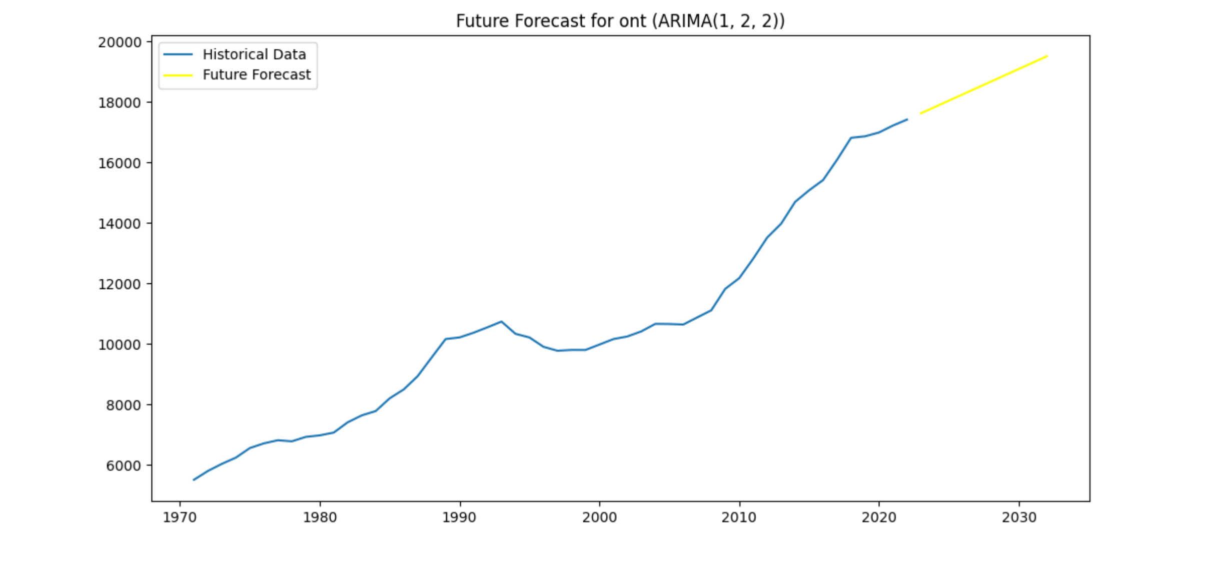




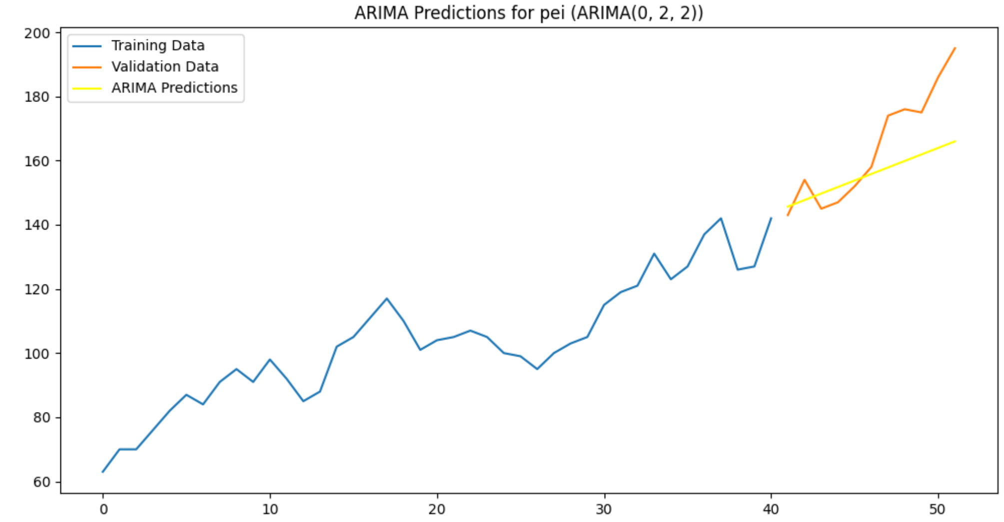
RMSE=6.2198140350867215

RMSE=621.45062287329

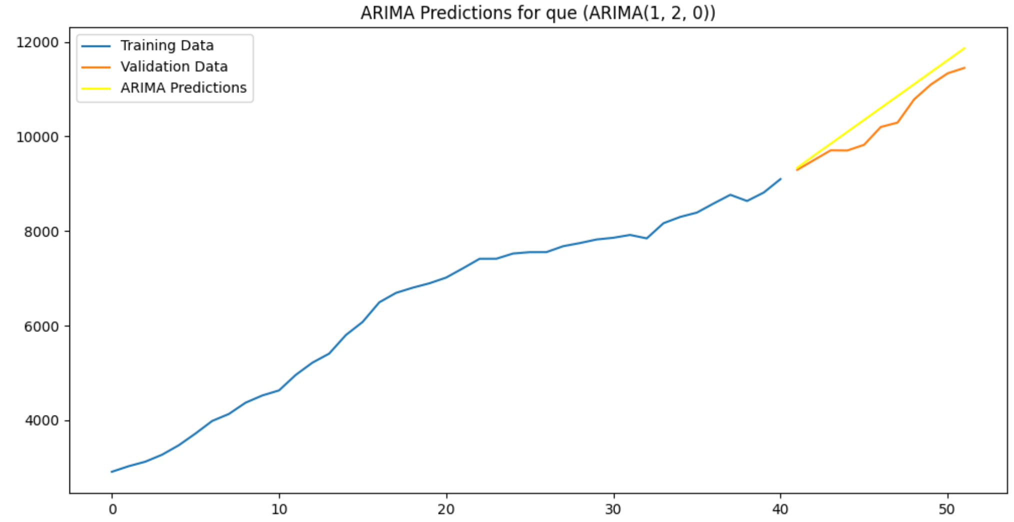




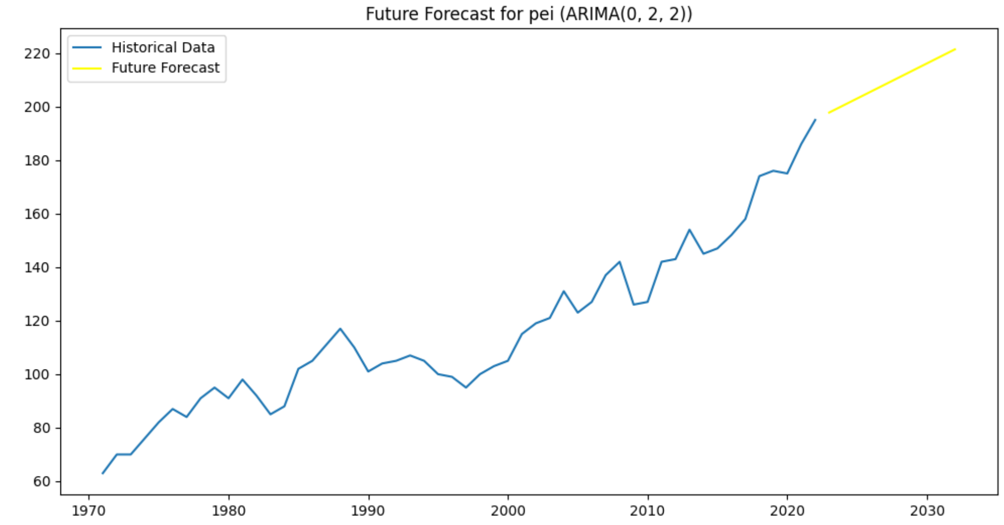
HPC Results

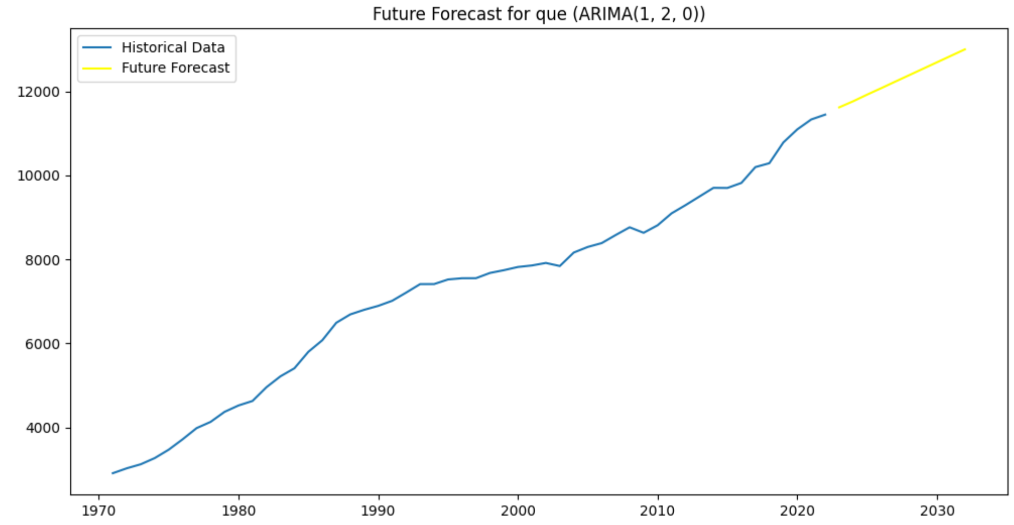


RMSE=13.884844989677397

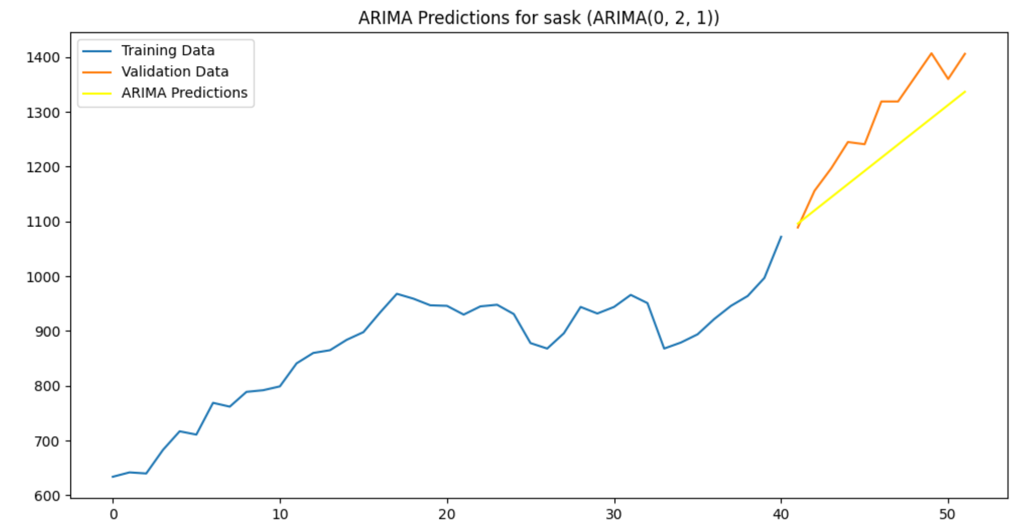


RMSE=351.8651314240781

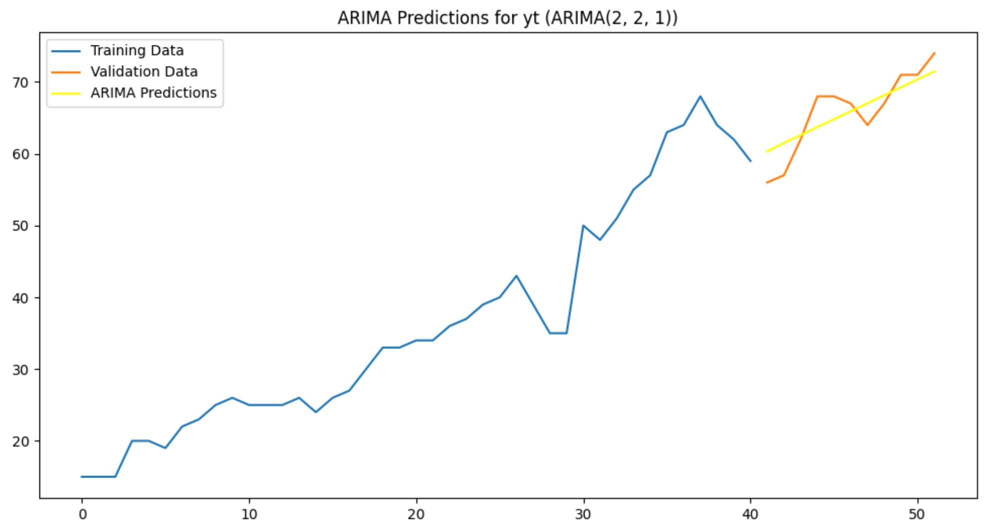




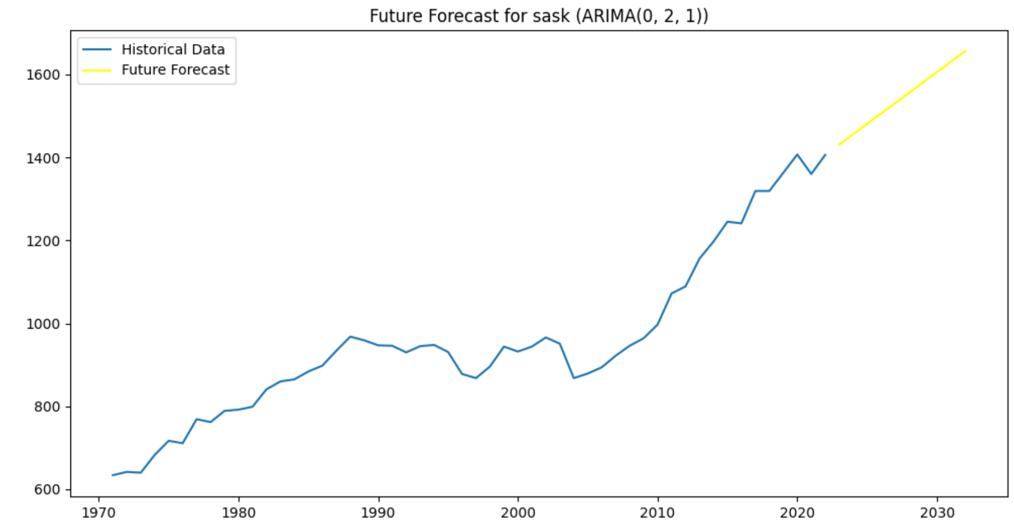
HPC Results

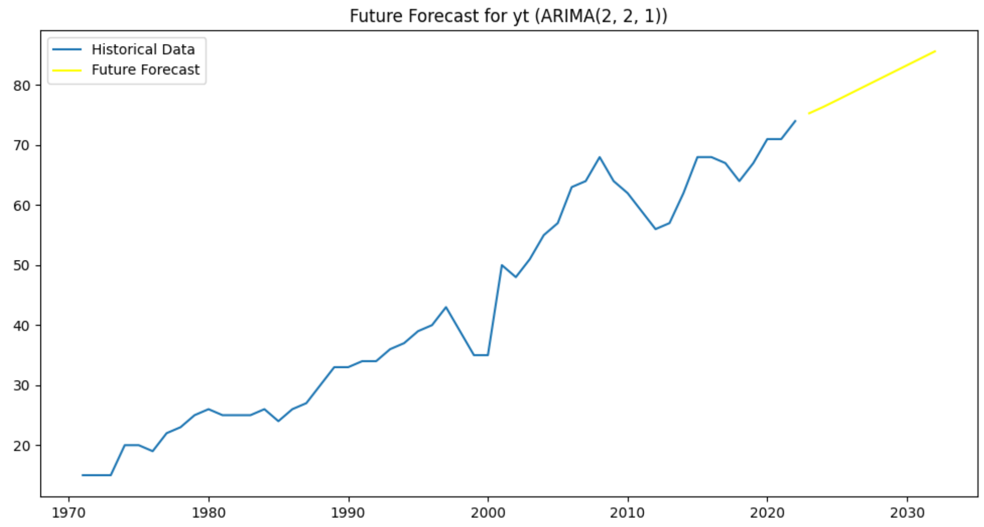


RMSE=73.75198732514221



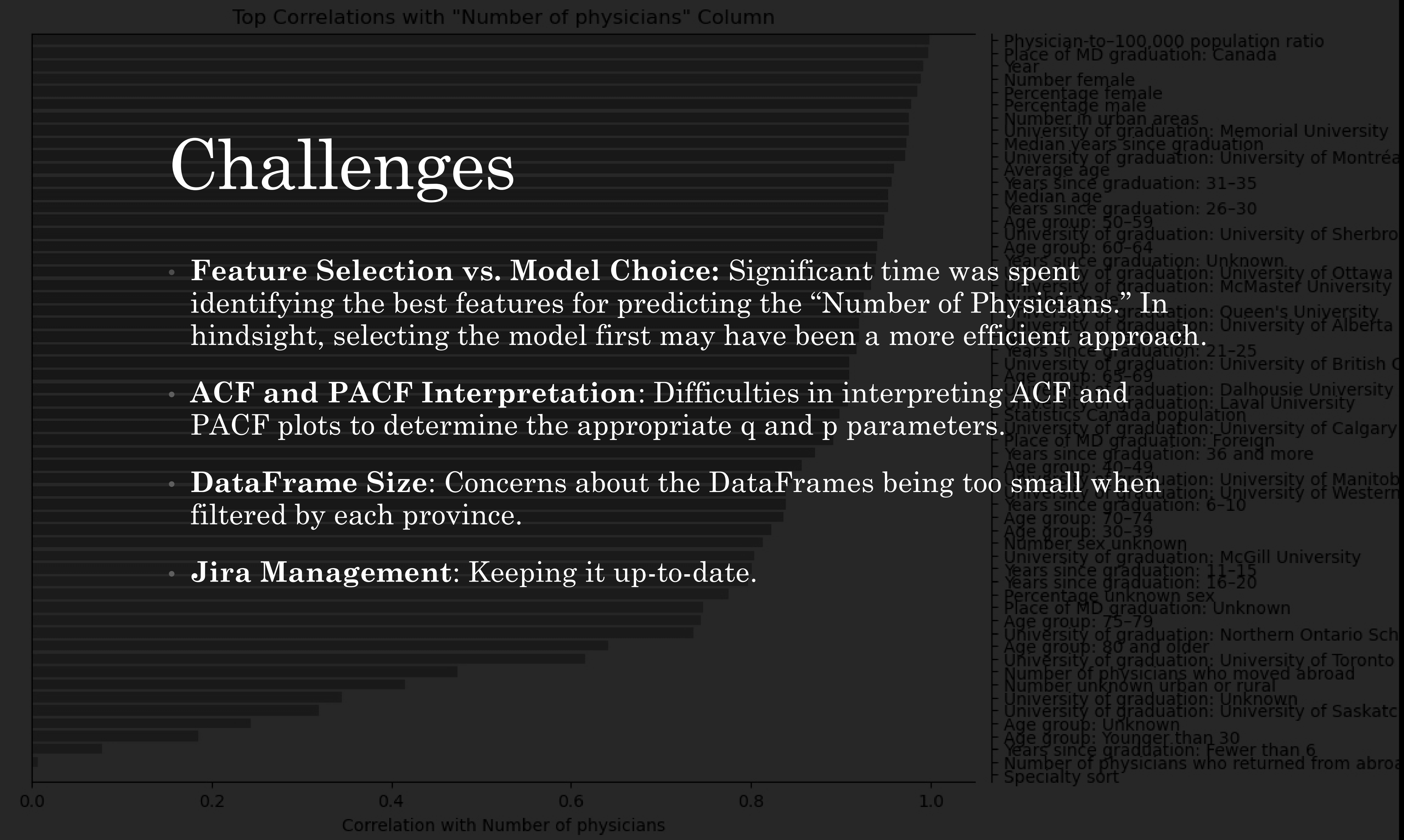
RMSE=2.8500608228991284





Challenges

- **Feature Selection vs. Model Choice:** Significant time was spent identifying the best features for predicting the “Number of Physicians.” In hindsight, selecting the model first may have been a more efficient approach.
- **ACF and PACF Interpretation**: Difficulties in interpreting ACF and PACF plots to determine the appropriate q and p parameters.
- **DataFrame Size**: Concerns about the DataFrames being too small when filtered by each province.
- **Jira Management**: Keeping it up-to-date.



Conclusion and Future Work

- **Effective Model for Short-Term Forecasting**: The current model is suitable for forecasting and predicting short-term values.
- **Challenges in Parameter Optimization:** Finding the best parameters involves a trial-and-error process. While automating this process may not always yield the most accurate results, manual testing can be more precise. Nevertheless, the automated approach performs reasonably well for many individual samples.
- **Expansion Opportunities**: extending the forecasting model to incorporate other specialties from the dataset
- **Model Limitations:** The model has limitations as it assumes that future trends will mirror past patterns.
- **Explore Advanced Techniques**: Investigate alternative time series models and machine learning methods to improve accuracy.