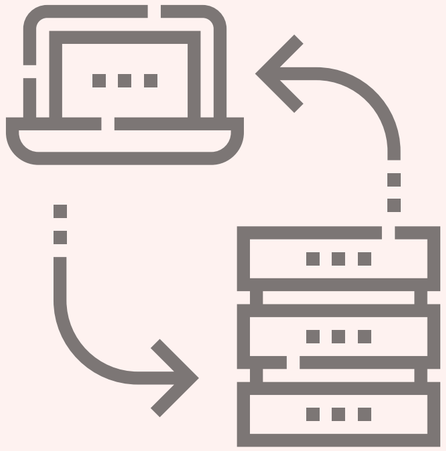
Exploring and predicting Post-Partum Depression (PPD) rates across the province in Canada

Introduction:

Maternal mental health is a critical global issue. Postpartum depression (PPD) is a common condition affecting women after childbirth,often accompanied by postpartum anxiety (PPA). These conditions can have severe consequences for both mothers and infants, including impaired parenting, disrupted family relationships,and long-term developmental challenges for children. Factors contributing to PPD include previous mental health issues, lack of social support,and stressful life events.

Objective of the study:

- Conduct analysis to understand the historical trends for the mothers who reported feeling consistent with PPD/anxiety and visualize the symptoms.
- Develop a model to predict the prevalence of Post Partum Depression / Anxiety across the provinces in Canada
- To align my predictive model with High Performance Computing (HPC).
- Develop a result / findings / challenges / future improvement on the model I used.



Dataset Overview:

1. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310041701 - Mean age of mothers
2. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2019041-eng.pdf - Maternal Mental Health
3. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009604 - Symptoms of PPA and PPD in Canada.xlsx

Mean age of mother at time of delivery (live births) 1 2 3 4 5 6 7 8	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed
Province	Years	Years	Years	Years	Year
Newfoundland and Labrador	29.7	29.9	30	30	30.
Prince Edward Island	30.3	30.5	30.7	30.9	31
Nova Scotia	30.2	30.2	30.6	30.6	3
New Brunswick	29.1	29.3	29.5	29.7	29.
Quebec	30.7	30.8	30.8	31	31
Ontario	31.6	31.8	31.9	32	32.
Manitoba	29.9	30.1	30.1	30.5	30.
Saskatchewan	29.4	29.5	29.5	29.7	29.
Alberta	30.7	30.8	30.9	31	31.
British Columbia	31.9	32	32.1	32.3	32.
Yukon	..	..	..	..	
Northwest Territories	29.3	29.6	29.7	30.1	30.
Nunavut	26.1	26.3	26.3	26.8	2

	Variable	PPA	PPA.2	PPA.3	PPD	PPD.2	PPD.3	Neither condition	Neither condition.2	Neither condition.3
0	0	Percent	95% confidence limits	0.0	Percent	95% confidence limits	0.0	Percent	95% confidence limits	0.0
1	Age category	n = 882*	0	0.0	n = 1131*	0	0.0	n = 5047	0	0.0
2	17–24	10.4	7.7	13.0	10.6	8.2	13.0	7.1	6.2	8.0
3	25–29	27.9	24.2	31.6	24.8	21.6	28.1	24	22.6	25.4
4	30–34	37	32.9	41.0	37.7	34.3	41.1	38.5	36.9	40.1

Exploratory Data Analysis (EDA):

Data Cleaning and processing on the datasets.

```python
import pandas as pd

# Load the data
df = pd.read_csv('Mean age of mothers to give birth.csv')
print(df.head())

#Display the shape of teh DataFrame
print("\nShape of the DataFrame:")
print(df.shape)

#Remove rows with missing values using dropna() function
df_rows_missing_values = df.dropna()
df_rows_missing_values.to_csv('birthrate_dataset_cleaned.csv', index=False)

#Load DataFrame after dropna()
df = pd.read_csv('birthrate_dataset_cleaned.csv')
df
```

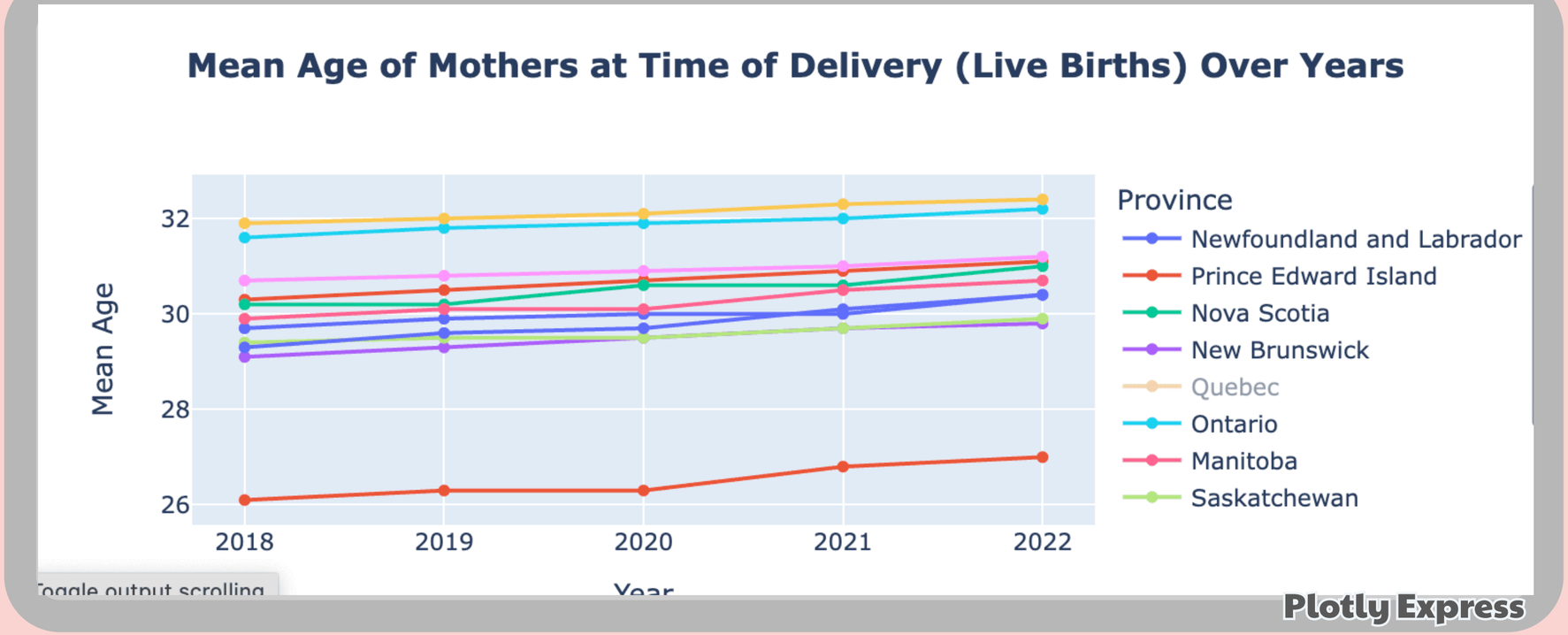
Mean age of mothers...csv

```python
import pandas as pd

# File path
file_path = 'Symptoms of PPA and PPD in Canada.xlsx'

# Define the columns range and number of rows
columns_range = 'A:J'
num_rows = 64

# Read the specific range from the Excel file into a DataFrame
df = pd.read_excel(file_path, sheet_name='Sheet1', usecols=columns_range

# Display the DataFrame info
df.info()
```

```python
# Define new column names
new_column_names = {
    'Unnamed: 2': 'PPA.2',
    'Unnamed: 3': 'PPA.3',
    'Unnamed: 5': 'PPD.2',
    'Unnamed: 6': 'PPD.3',
    'Unnamed: 8': 'Neither condition.2',
    'Unnamed: 9': 'Neither condition.3',
}

# Rename columns
df.rename(columns=new_column_names, inplace=True)

# Replace NaN values with 0
df.fillna(0, inplace=True)

# Display the DataFrame info to see updated column names
df.info()
```

Symptoms of PPA and PPD...xls

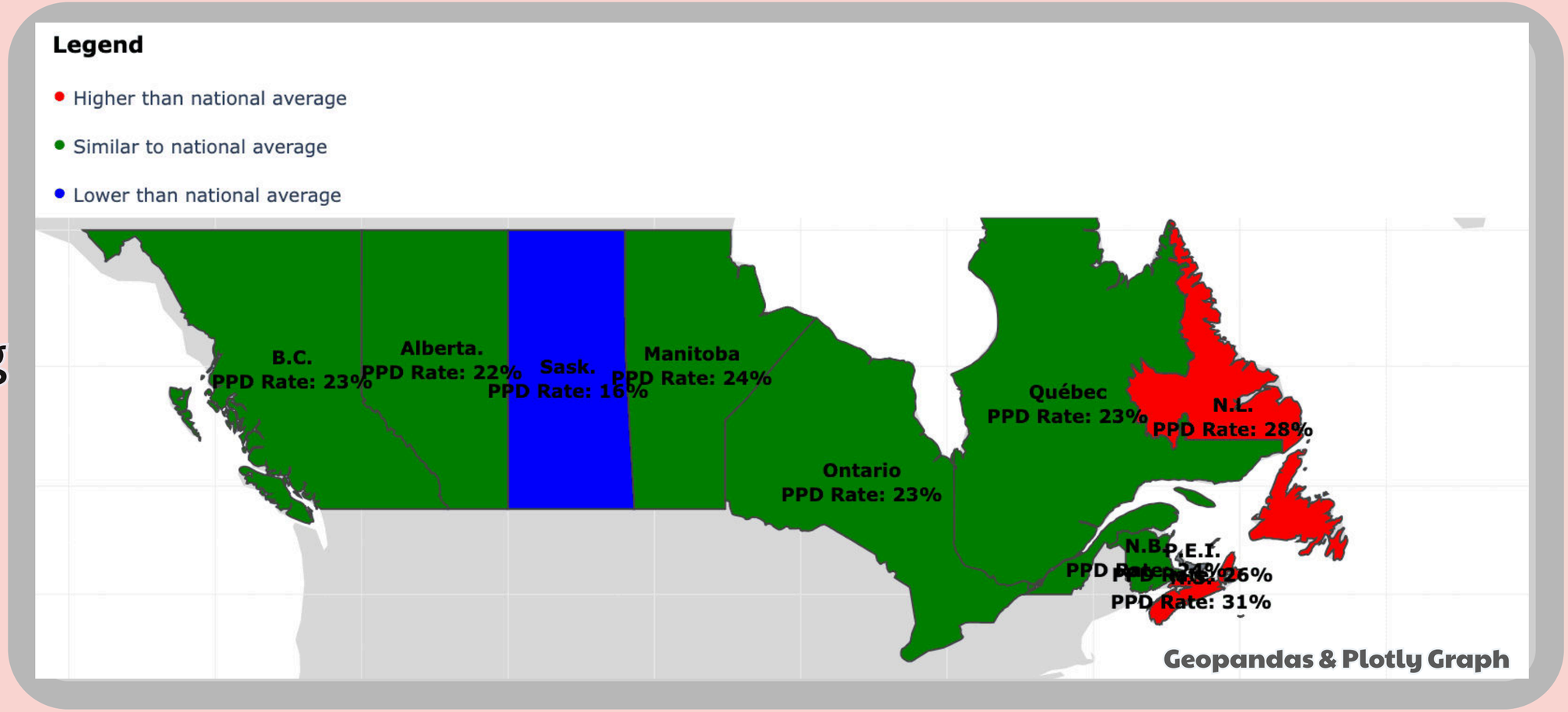
Exploratory Data Analysis (EDA):



Based on the data collected and above graph, how the mean age of mothers at the time of delivery has changed over time across different provinces in Canada. **Highest Mean Age:** Ontario consistently has the highest mean age, starting around 31.6 in 2018 and reaching approximately 32.2 by 2022. **Lowest Mean Age:** Nunavut is not included in this graph, but based on the data provided earlier, it typically has the lowest mean age. **Consistent Increase:** All provinces show an increase in the mean age of mothers from 2018 to 2022, indicating a trend towards older maternal age over these years.

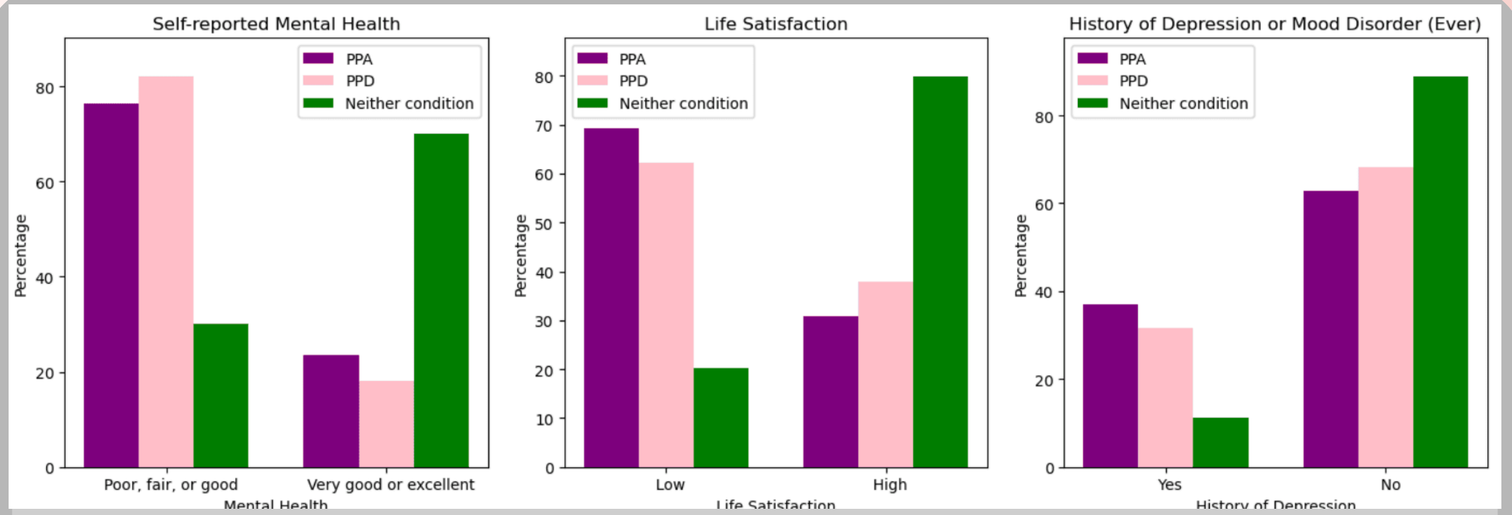
Exploratory Data Analysis (EDA):

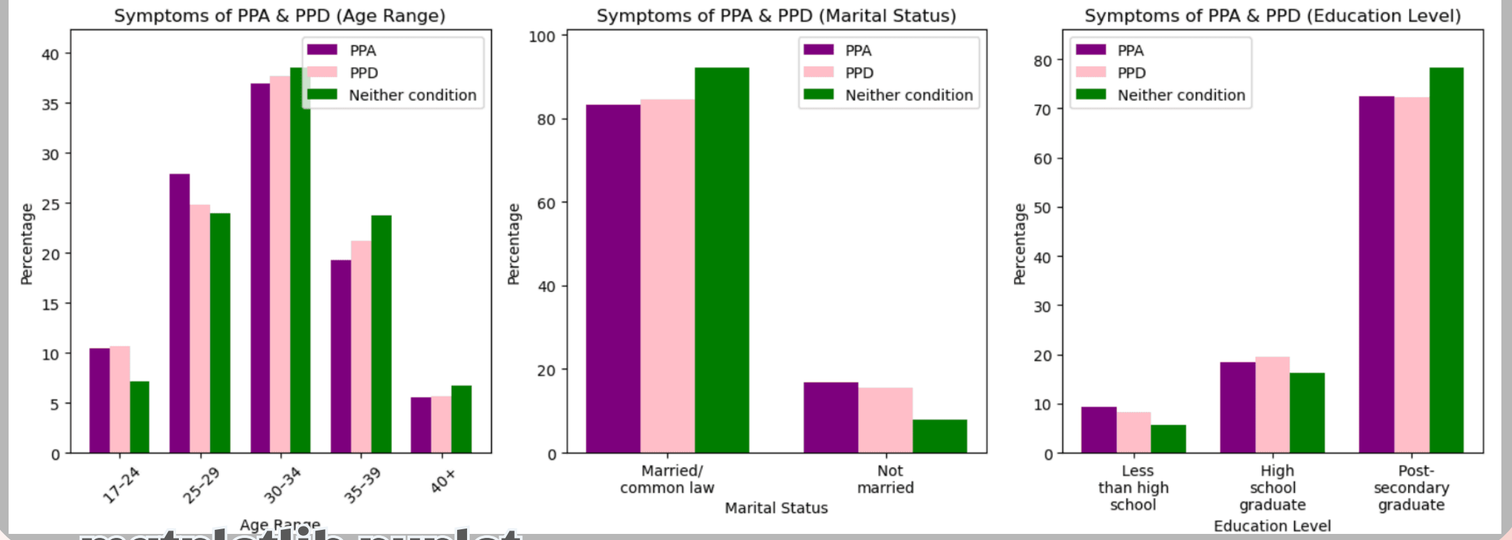
Percentage of mothers reporting feeling consistent with PPD or anxiety disorder per province.



In Canada, almost one-quarter (23%) of mothers who recently gave birth reported feelings consistent with either post-partum depression or an anxiety disorder. The proportion of mothers reporting these feelings varied across provinces, ranging from 16% in Saskatchewan, 28% in Newfoundland Labrador (higher than the national average) to 31% in Nova Scotia.

Symptoms of Post Partum Depression / Anxiety





matplotlib.pyplot

Based on the survey the largest proportion of mother's with symptoms consistent with PPA or PPD were in the

- 25–29 and 30–35 age groups
- married or in a common-law relationship,
- had a post-secondary education.

A significant majority of mothers with symptoms of PPA (76.4%) and PPD (82.0%) rated their mental health as “poor,” “fair,” or “very good.” In contrast, only 30.0% of mothers without symptoms of either condition rated their mental health similarly.

Machine Learning Model: RandomForestRegressor

RandomForestRegressor : is a meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It can manage datasets with many features, which is often the case in medical datasets with numerous potential predictors (e.g., demographic data, medical history, psychological assessments).

```python
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train the RandomForestRegressor model with parallel computing
model = RandomForestRegressor(n_estimators=100, n_jobs=-1, random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}\n")
```

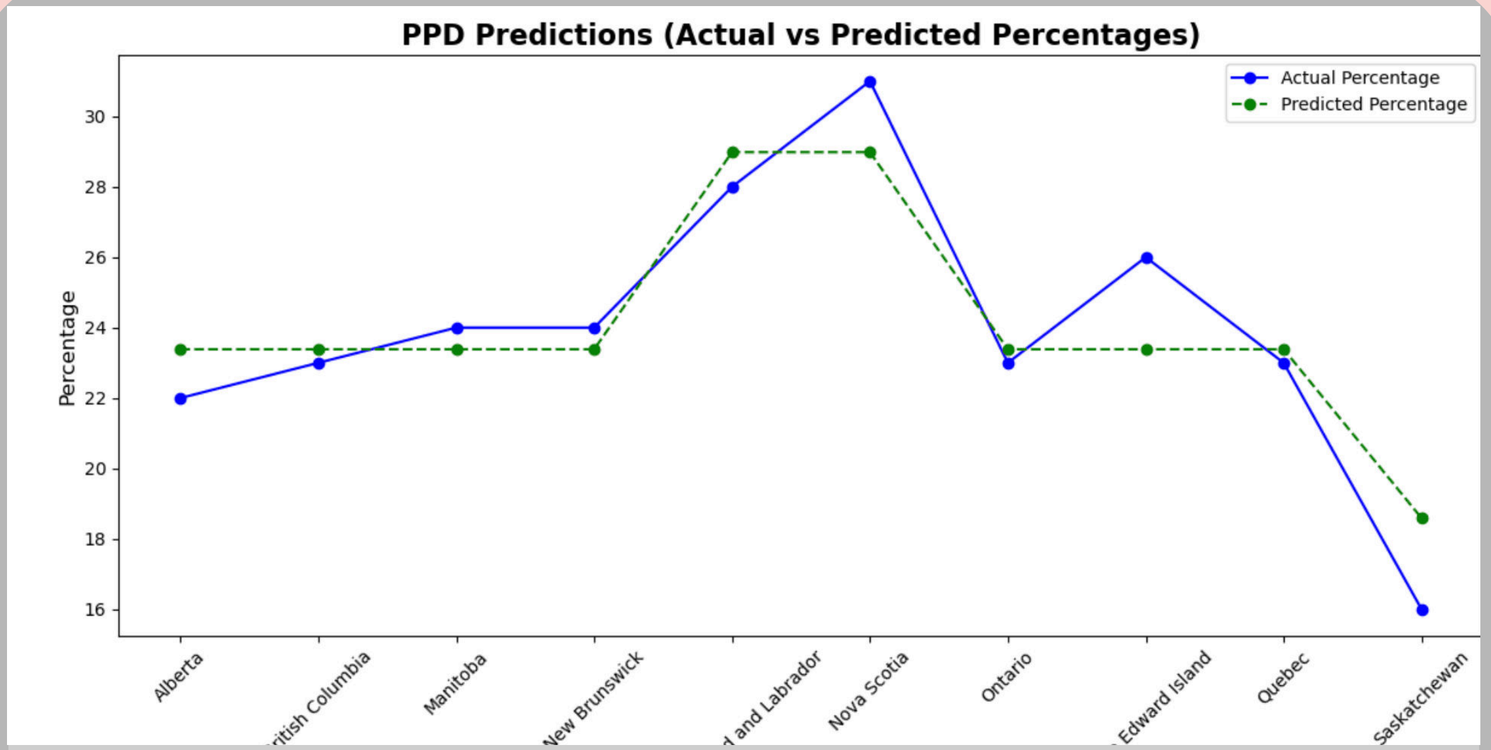
```python
# Perform Grid Search for hyperparameter tuning with parallel processing
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20],
    'min_samples_split': [2, 5]
}

# Use KFold with a sufficient number of splits to avoid small fold issues
kf = KFold(n_splits=5, shuffle=True, random_state=42)

grid_search = GridSearchCV(estimator=RandomForestRegressor(n_jobs=-1, random_state=42), param_grid=param_grid, cv=k
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_
print(f"Best Model Parameters: {grid_search.best_params_}")

# Cross-validation with parallel processing
scores = cross_val_score(best_model, X, y, cv=kf, n_jobs=-1)
print(f"Cross-Validation Scores: {scores}")

# Predict the percentage for each province
df['Predicted_Percentage'] = best_model.predict(X)
print(df[['Province', 'Percentage', 'Predicted_Percentage']])
```

- Created a df using the historical data
- Defined the features (X=demographic) and target (y=PPD %).
- (train_test_split) data using an 80-20 split
- Created and trained a model on the training data using parallel computing (n_jobs=-1).
- Calculated the Mean Squared Error (MSE) of the predictions on the test set. `Mean Squared Error: 4.377376191043089`
- Used GridSearchCV with a 5-fold cross-validation. for Hyperparamter tuning.
- (KFold) to find the best model parameters. Trained theGridSearchCVobject to find the best estimator.
- Performed cross-validation using the best model and calculated the cross-validation scores. "[-0.0944428, 0.91537486, 0.31815751, 0.0, 0.0]"
- Used the best model to predict the percentage of mothers with PPD symptoms for each province in the full dataset.

RESULT:

The RandomForestRegressor model's predictions are reasonably accurate for some provinces but less accurate for others. This is reflected in both the line plot and the pie charts, where certain provinces show significant discrepancies.

Provinces where the model performs well (e.g., Newfoundland and Labrador and Nova Scotia) show close alignment between actual and predicted values. In contrast, provinces with larger deviations (e.g., Prince Edward Island and Saskatchewan) suggest that the model needs some improvement.




```
Best Model Parameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}
Cross-Validation Scores: [-0.0944428   0.91537486  0.31815751  0.          0.        ]
                   Province  Percentage  Predicted_Percentage
0  Newfoundland and Labrador          28             28.983071
1       Prince Edward Island          26             23.385690
2                Nova Scotia          31             28.983071
3              New Brunswick          24             23.385690
4                     Quebec          23             23.385690
5                    Ontario          23             23.385690
6                   Manitoba          24             23.385690
7               Saskatchewan          16             18.586786
8                    Alberta          22             23.385690
9           British Columbia          23             23.385690
```