Final Project Report – Data Science Lab

Contents

1.Introduction

The gender wage gap represents a global issue of substantial economic, social, and political importance. Despite narrowing trends over past decades, significant disparities persist in most countries. This project focuses on the Organization for Economic Co-operation and Development (OECD) countries, often perceived as leaders in promoting equality and exemplifying socio-economic progress.

Notwithstanding this, as observed in the preliminary analysis, even developed countries like Korea, the USA, the UK, Germany, and France exhibit a high wage gap. This discrepancy sparked the curiosity to investigate and explore the potential underlying factors contributing to this phenomenon.

The primary research question is: "What patterns or trends can be identified across various sectors (economy, education, government, jobs, society, health, innovation and tech, finance) that could potentially explain the disparities in gender wage gaps among OECD countries?"

2. Data Sources and Merging

The datasets for this project were gathered from the official OECD and World Bank databases (OECD website, World Bank Gender Statistics). These databases offer a wide range of data across various sectors such as economy, education, government, jobs, society, health, innovation and technology, finance, and more. For our investigation, we used data on gender/age wage gaps, education, population, health, and employment to explore the factors influencing the gender wage gap across different industries (Business, Government, Education and non-profit).

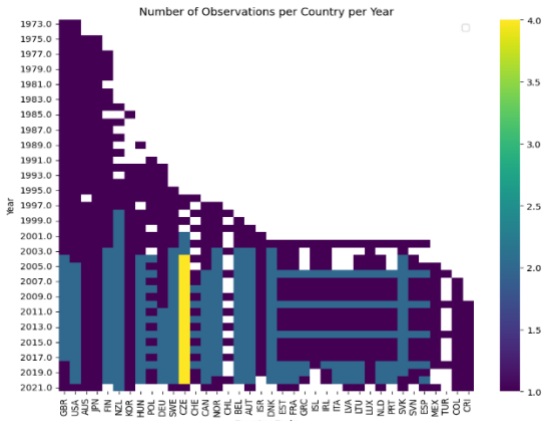
The merging process involved several steps to create a comprehensive and clean dataset for analysis:

1. **Loading data:** Five separate datasets ('extra.csv', 'education.csv', 'employment.csv', 'population.csv', and 'wage.csv') were loaded into separate pandas dataframes.
2. **Data cleaning:** The 'wage' dataset was cleaned to rename columns for consistency ('LOCATION' to 'Country Code', 'TIME' to 'Time') and to ensure the 'Time' column was of the correct data type (object to int64). The years before 2000 were excluded from the 'wage' dataset due to the high number of missing values.
3. **Handling missing values:** Any '..' entries in the merged dataset were replaced with NaN values for correct handling of missing data. The last two rows of the dataset, which were likely metadata or footnotes in the original data, were dropped.
4. **Importing additional data:** Additional data ('more.csv', 'more2.csv', 'more3.csv', 'more4.csv') was imported and melted into long format using pandas' melt function. Although the names of the files might be confusing, we only used them to gather and mine any additional features that were scattered across different websites,and might be useful for us. *[in the end , there were no benefits of using them, mostly nulls]*

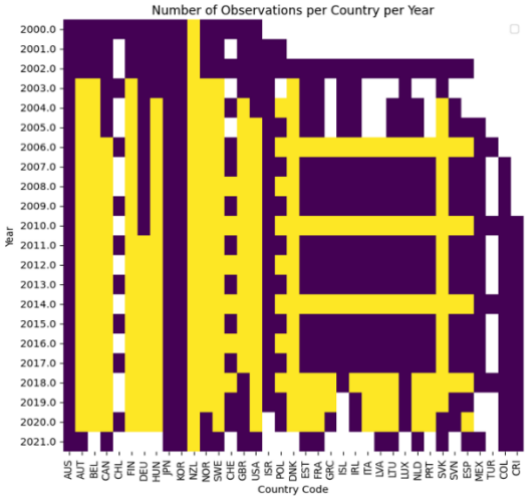
3. Data Pre-processing

The pre-processing phase was crucial to prepare our dataset for subsequent analysis. This involved handling missing values, cleaning, and filtering the data to ensure its suitability for further steps.

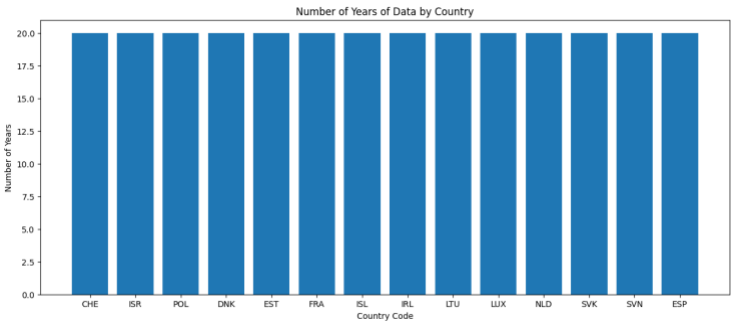
1. Handling Missing Values: We started by dropping columns with a null percentage above 50%. Columns with a high percentage of missing values can distort the analysis and lead to unreliable results. However, this step might have led to the loss of potentially important data that could have been relevant for our research. We then explored the missing values using a heatmap, which is an effective way to visualize the distribution of missing data. Based on this, we decided to drop the years before 2000 and the years 2021 and 22, due to a high number of missing values. This decision was made to ensure the reliability of our analysis, but it might introduce a time range bias, as it excludes certain years.



2. Country Filtering: Next, we sought to identify the common countries that had data across the years. This involved finding unique countries that had data for each year. We then kept countries with 5 or less missing years from 2000 to 2020. This decision ensures that the countries included in our analysis have a relatively complete dataset, which is crucial for the reliability of our analysis. However, this might introduce a selection bias, as it could potentially exclude countries with higher missingness that could have had different characteristics in terms of gender wage gap trends.



3. Data Imputation: After cleaning the data, we imputed the missing values using the K-nearest neighbors (KNN) method. KNN is a robust method for handling missing data, as it fills in missing values based on the mean of the k-nearest neighbors. This method is suitable for our case because it can handle multivariate data and it considers the relationships between features.



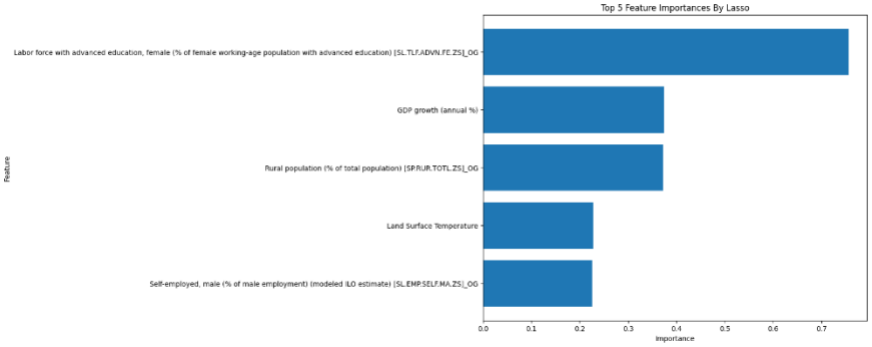
To find the optimal 'k', we experimented with different 'k' values. After careful consideration and comparison of the results, we found that **'k=1'** yielded the most reliable imputation results for our data. it's important to note that KNN can introduce their own biases and inaccuracies, as they are based on assumptions about the data.

4. Feature Selection

We used Lasso (Least Absolute Shrinkage and Selection Operator) regression for this task. We used the LassoCV function from the scikit-learn library to find the optimal amount of penalization.

This function performs cross-validation with a range of alpha values, where alpha is a parameter that controls the degree of penalization in Lasso. We set our alpha values in the logarithmic space between -4 and -0.5.

Lasso regression selected 75 variables and eliminated 52 variables, with an optimal alpha of 0.316228. These 75 variables are considered the most influential in predicting the gender wage gap and will be used in the subsequent machine learning modeling phase.



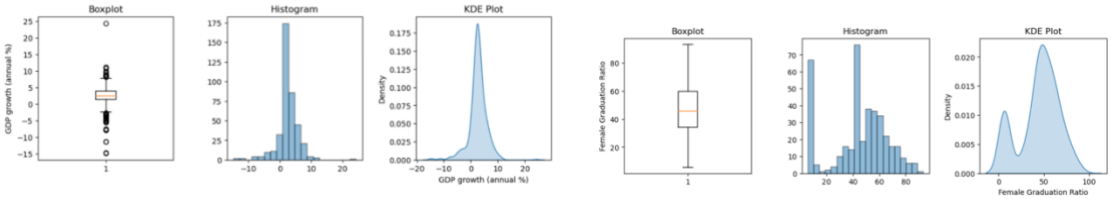
5.EDA Section

Here are some columns in the dataset:

1. **Time**: The year of the record.
2. **Country Code**: An abbreviation representing the country for which the data was collected.
3. **Female Senior Mgmt Share**: The share of senior management roles filled by women.
4. **Prevalence of overweight, male (% of male adults)**: The percentage of male adults who are overweight.
5. **Female Agriculture Grads Share**: The share of agriculture graduates who are female.
6. **Female share of graduates in Arts and Humanities programmes, tertiary (%)**: The percentage of graduates from tertiary Arts and Humanities programs who are female.
7. **Female Edu Grads Share**: The share of education graduates who are female.
8. **Value**: representing the gender wage gap.

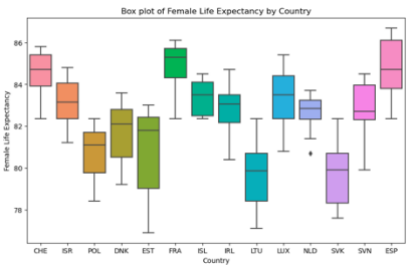
Outlier Detection and Visualization:

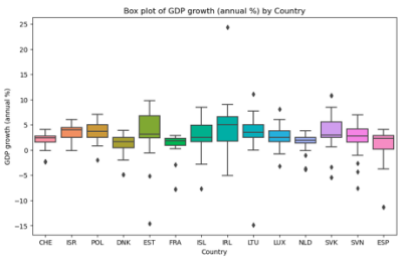
We conducted a comprehensive analysis of outliers in our dataset.
To visualize and assess the distribution of our data, we utilized various plotting methods including box plots, histograms, and Kernel Density Estimation (KDE) plots.
From our visual analysis, we found a diversity in our data distributions. For example, in one feature, the data was almost normally distributed but exhibited a slight skewness to the left.



Cross-country Outlier Analysis:

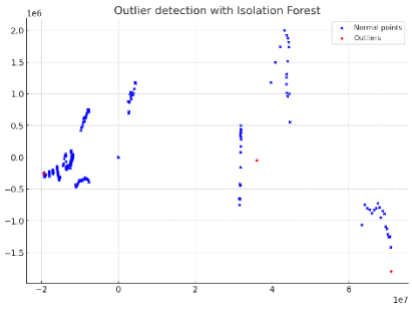
We generated comparative box plots. These box plots gave us the opportunity to observe country-specific variations and outliers. We noticed varying numbers of outliers among different features, with some having few outliers and others exhibiting many.





Multivariate Outlier Analysis:

In addition to univariate outlier analysis, we performed a multivariate outlier detection using the Isolation Forest algorithm. This method is particularly effective when dealing with high-dimensional data, as it considers the relationships among different features to identify anomalies.



We used PCA to reduce the data dimensionality to two, which is easily plottable. As seen in the scatter plot, the blue dots represent normal data points, while the red dots signify the outliers as detected by the Isolation Forest algorithm, and this method identified a minimal number of multivariate outliers.

Decision on Outliers:

After a careful analysis, we chose to retain the outliers in our dataset. This decision was made based on the understanding that these outliers might represent unique but important real-world scenarios or specific country situations. Removing these outliers could have introduced a selection bias.

Also, multivariate outlier detection identified a small number of outliers. This suggests that these points could represent important edge cases, which underscore the complexity of the gender wage gap.

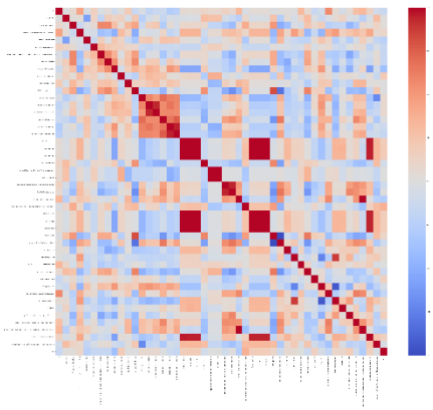
Feature Exploration – Collinearity

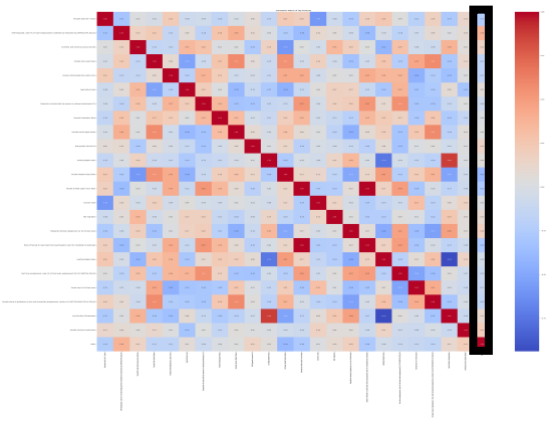
To check for collinearity, we calculated the correlation coefficient between each pair of variables in the dataset. We generated a heatmap to visualize the collinearity, and it became evident that several variables are highly correlated, some of which were different subjects under the same indicators.

To refine our heatmap, we eliminated features that exhibited correlation coefficients greater than 0.75 or less than -0.75. Features from the same indicators but different subjects were also dropped due to high correlation.

Correlation with the Label:

We found that most features exhibit a low to moderate correlation with our label. However, it is crucial to note that correlation does not imply causation. Therefore, a low or moderate correlation does not inherently signify that these features are unimportant or that there is no relationship between them and the label. A feature could have a complex, non-linear relationship with the label that is not captured by a simple correlation coefficient. Considering this, we planned to utilize more sophisticated, non-linear models capable of capturing these complex relationships.





Insights from Correlation Analysis:

1. 'Self-employed, male (% of male employment) (modeled ILO estimate)' had a relatively high positive correlation (0.38) with the gender wage gap. This suggests that in countries with a higher proportion of self-employed men, the gender wage gap tends to be larger.
2. 'Female Services Grads Share' had a positive correlation (0.21) with the gender wage gap, indicating that as the share of female graduates in the services sector increases, the wage gap also tends to widen.
3. Conversely, 'Female Health Grads Share' had a relatively high negative correlation (-0.39) with the gender wage gap, suggesting that as the share of female graduates in health-related fields increases, the gender wage gap tends to decrease. This could be due to higher wages in the health sector or a more equitable pay structure in this sector.
4. The 'Ratio of female to male labor force participation rate (%) (modeled ILO estimate)' and 'Female to Male Labor Force Ratio' both displayed a negative correlation (-0.28) with the gender wage gap. This indicates that in countries where female labor force participation is closer to male labor force participation, the wage gap tends to be smaller.

6. Problems with the Data and Reevaluation of the Research Question

6.1 Problems with the Data

Our dataset did have certain limitations that may have affected the scope of our research. These limitations can be summarized as follows:

1. **Limited Data Size:** Our analysis was confined to the data available for a certain number of countries and years. This limitation could potentially restrict our ability to capture global trends and patterns in the gender wage gap fully.
2. **Loss of Important Features:** In the process of cleaning and filtering the data, we were compelled to drop some potentially important features from several sectors, it inevitably resulted in the loss of some information.
3. **Data Biases:** The extensive data pre-processing steps that we implemented could introduce certain biases. For example, our approach to handling missing values and outliers, as well as our decision to drop certain years and countries based on data availability, could all influence the results of our analysis.

6.2 Reevaluation of the Research Question

Given the data limitations and the steps we took in pre-processing, our revised question became more specific, focusing on the patterns and trends among the remaining features and the reduced set of countries and years. **Our revised question is:**
"What patterns or trends can be identified across the remaining sectors that could potentially explain the disparities in gender wage gaps among the selected countries from 2000 to 2020?" the revised research question provides a more realistic goal that aligns better with our pre-processed dataset, while still focusing on the essential theme of understanding the gender wage gap.

7. Data Biases

7.1 Selection Bias

Selection bias, arising from the non-random nature of our data, might lead to skewed findings due to overrepresentation or underrepresentation of certain countries. We mitigated this by

incorporating multiple datasets from reputable sources, the OECD and the World Bank, to provide a balanced view of the gender wage gap.

7.2 Time Range Bias

Focusing on data from 2000 onwards might cause us to overlook long-term trends in the gender wage gap. However, this decision was made to use more recent and reliable data, ensuring that our analysis is both robust and timely.

Our aim is to present an analysis that is both robust and timely, providing valuable insights into the current state of the gender wage gap.

7.3 Missing Data Bias

Missing data bias can distort the analysis if the missing data is not random. We addressed this by dropping countries with significant missing data and employing the K-nearest neighbors (KNN) algorithm. The KNN method imputes missing values based on the mean of the k-nearest neighbors, minimizing the impact of missing data bias and enhancing the validity of our findings.

8. Machine Learning Models

We employed three robust machine learning models: Decision Tree, Random Forest, and XGBoost. These models can handle a large number of features, complex interactions, and are known for their robustness against outliers and non-linear relationships.

1. **Decision Tree Regressor**: Decision trees are simple yet powerful models that split the data based on the feature that provides the most significant reduction in variance. Despite their simplicity, decision trees are prone to overfitting, especially when dealing with a large number of features.
2. **Random Forest Regressor**: Random Forest is an ensemble method that aggregates the predictions of many decision trees to make a final prediction. By doing so, it improves the model's generalization and reduces the risk of overfitting.
3. **XGBoost Regressor**: XGBoost, short for Extreme Gradient Boosting, is another ensemble method that builds weak prediction models sequentially, with each new model aiming to correct the errors of its predecessor. XGBoost is known for its high performance and flexibility, as it allows for extensive customization and fine-tuning.

The models were trained on 80% of the data, with the remaining 20% held out for validation. We used Mean Absolute Error (MAE) as the metric to evaluate the performance of the models. MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It's a reliable metric for regression tasks as it provides a clear interpretation of the error in the original units of the target variable.

Cross-validation

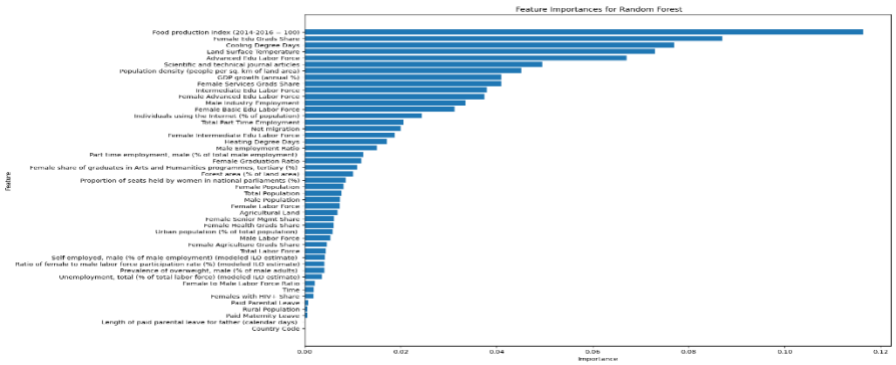
To further ensure the robustness and reliability of our models, we performed cross-validation with a range of folds.

```
Best CV for Decision Tree: 7
Best CV for Random Forest: 3
Best CV for XGBoost: 7
```

The MAE scores were as follows:

- Decision Tree MAE: 0.818
  [Graph in Plots File]
- Random Forest MAE: 0.661

- XGBoost MAE: 0.755
  [Graph in Plots File]



Our evaluation revealed that the Random Forest model achieved the lowest MAE, indicating that it was the most accurate of the three models. However, we observed that the model's predictions were less precise for some countries, particularly Switzerland, Poland, and Estonia. This inconsistency merits further analysis to identify potential causes and solutions.

Moreover, we extracted the feature importance from the Random Forest model to gain insights into the factors that significantly influence the model's predictions. The most important feature in our model was 'Females with HIV+ Share', followed by 'Female Intermediate Edu Labor Force' and 'Scientific and technical journal articles'. However, due to potential outlier concerns, we suggest further investigation into the 'Females with HIV+ Share' feature.

9. Feasibility Assessment

The (MAE) served as our performance metric, with the Random Forest model returning the lowest MAE. However, we observed varying predictive accuracy for different countries, highlighting inconsistencies in model performance.
During data preprocessing, we had to exclude numerous potentially important features due to missing values, which could have limited the predictive power of our models. Our decision to only include data from the year 2000 onwards may have excluded valuable historical context as well.

Given these considerations, the project, in its current form, cannot be deemed fully feasible. While our analysis provided valuable insights into the gender wage gap, limitations related to data completeness, variable model performance, and exclusion of important features necessitate further refinement of our approach.

Despite these limitations, our work is a crucial step towards understanding the gender wage gap, setting the groundwork for more comprehensive future research. This project underscores the complexities involved in analyzing such a multifaceted issue and emphasizes the need for continuous exploration and refinement in our modeling approach.

10. Future Work and Improvement

As we move forward, there are several areas that we believe warrant further investigation for improving our model's performance:

1. Country-Specific Analysis: One of our key observations was the inconsistent model performance across different countries. An interesting direction for future work could involve

focusing on these specific countries, such as Switzerland, Poland, and Estonia, which exhibited higher prediction errors. We could gather more granular, country-specific data and potentially build individual models for these countries to address their unique characteristics and trends in the gender wage gap.

2. Expanding the Temporal Scope: Our current model considers data from the year 2000 onwards. However, incorporating data from earlier years might provide valuable historical context and additional insights into the long-term trends and factors affecting the gender wage gap. This could enhance the predictive power of our model.

3. Alternative Modeling Techniques: There are advanced modeling techniques that could be explored. For instance, deep learning models, like neural networks, can capture complex non-linear relationships and could prove beneficial given the intricate nature of our data.

4. Domain-Specific Investigation: A deeper, domain-specific investigation into some well-performed features could provide better context and understanding