Exploratory Data Analysis of Gender Wage Gap

FINAL REPORT

Zeynep Elabiad

INFO I-590 Data Visualization

Summer 2022

[1] Image: Freepik.com

Abstract

The gender wage gap is defined as the difference in median earnings between men's and women`s wages.[2] While issues such as gender inequality, and gender discrimination are at the forefront today, the gender wage gap, one of the reflections of this critical issue, has also started to be discussed. According to the U.S. Census, in 2020, women full-time workers in the U.S. earned 83 cents to each dollar earned by men. [3]

This project uses exploratory data analysis and visualizations to reveal the gender wage gap percentages by comparing gender development indexes and human development indexes from selected OECD members.

Introduction

While women's participation in the labor force has significantly increased in the last century, a sign of progress on the gender inequality front, much must be done to address the gender wage gap. This is a very important topic for me as a woman and, more importantly, for the global community. The United Nations estimates that closing the gender wage gap could increase global GDP by 35% on average. [4] Closing this gap could go a long way to decreasing global poverty. Few measures can have that type of impact, which is why addressing this gap is essential to the achieve United Nations Sustainable Development Goals (SDGs). In this project, I will examine and find important correlations between the gender gap and other factors such as the Gender Development Index and Human Development Index. These visualizations are important as they enlighten people, politicians, and decision-makers to take corrective actions. The more we can do to address the wage gap, the more we can grow and prosper as a society together.

[2] https://www.oecd-ilibrary.org/employment/gender-wage-gap/indicator/english_7cee77aa-en
[3] https://www.census.gov/library/stories/2022/03/what-is-the-gender-wage-gap-in-your-state.html
[4] https://www.un.org/en/un75/women_girls_closing_gender_gap

Existing Visualizations

There are existing visualizations related to the Gender Wage Gap using different visualization methods. One of the visualization examples is from the OECD Gender Wage Gap page below. The page provides charts and map visualizations that compare Gender Wage Gap percentages. The visualization lets you choose by country from OECD countries and by year using beautiful color combinations. [5]

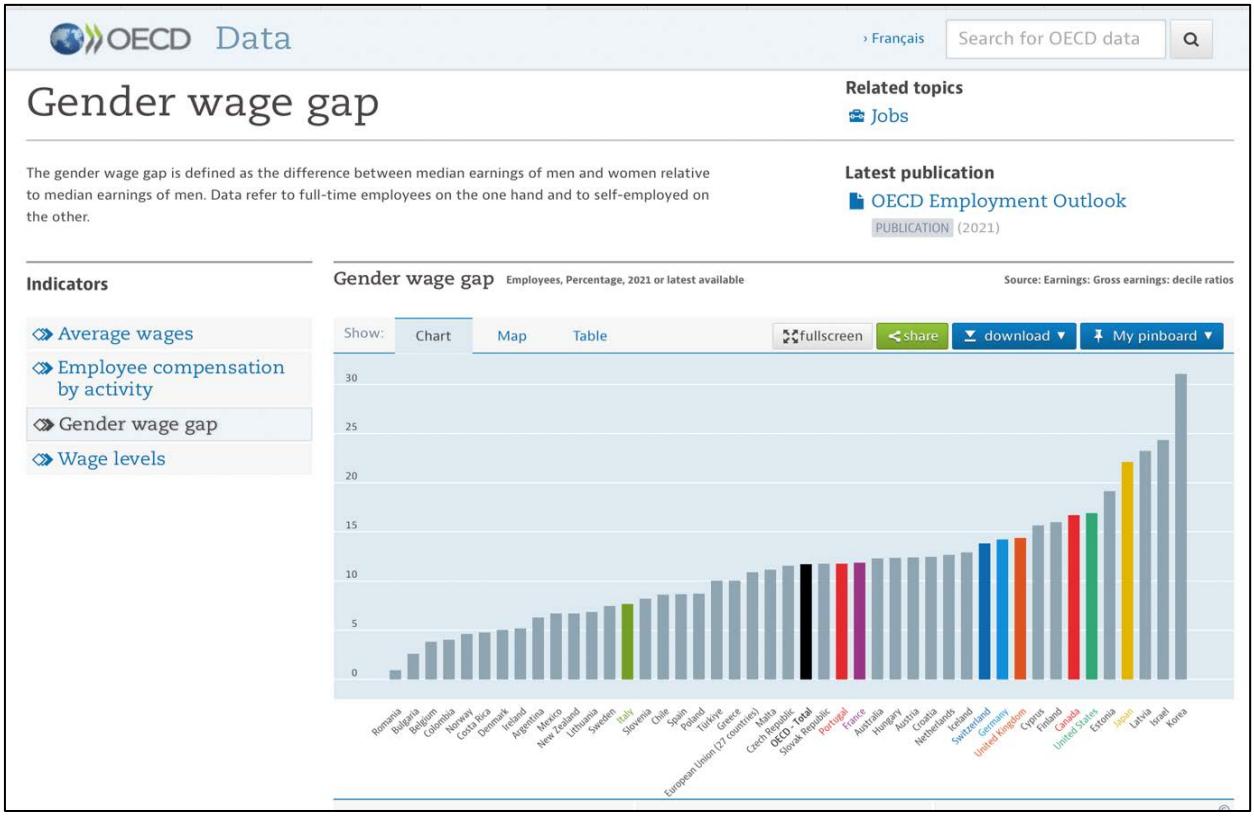

Figure 1. OECD Gender Wage Gap Chart of OECD Member Countries

[5] https://data.oecd.org/earnwage/gender-wage-gap.htm

Using a choropleth map, the US Census made another visualization to demonstrate the Gender Pay Gap. The graph shows the wage gap as the difference between full-time median annual earnings for men and women. According to the map Census, in 2019, the yearly national wage gap by gender in the U.S. was $10,150 (+/- $276). [6] The visualization used a blue color scale as color choice, but the State dividing lines are unclear, and States don`t have names on the map. Each state's gender pay gap data can be seen when you hover your mouse over it. Also, the darkest two blue colors representing the Wage gap almost look the same.

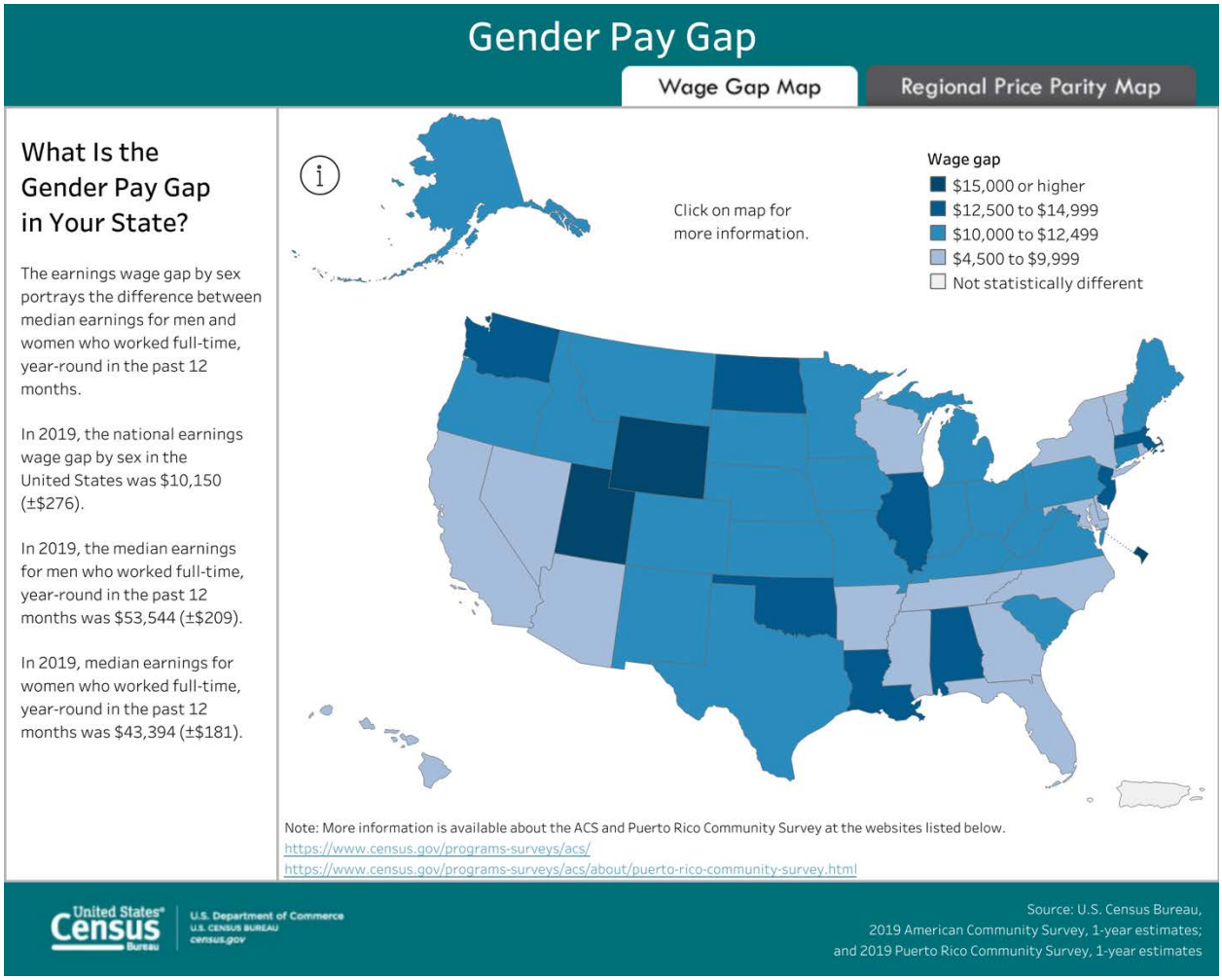

Figure 2. 2019 Annual wage gap by states in the U.S.

[6] https://www.census.gov/library/stories/2022/03/what-is-the-gender-wage-gap-in-your-state.html

Statista also shared a visualization titled “The Gender Pay Gap in Developed Nations Visualized,” using selected 15 OECD countries' data from OECD. [7] While South Korea has the highest gender wage gap, New Zealand has the lowest gender wage gap, according to the graph. The visualization shows bar graphs and map visualizations together, which is a creative idea. The graph is simple, and the message is clear. But the visualization does not show any yearly info and uses a red color choice which is not color-blind friendly.

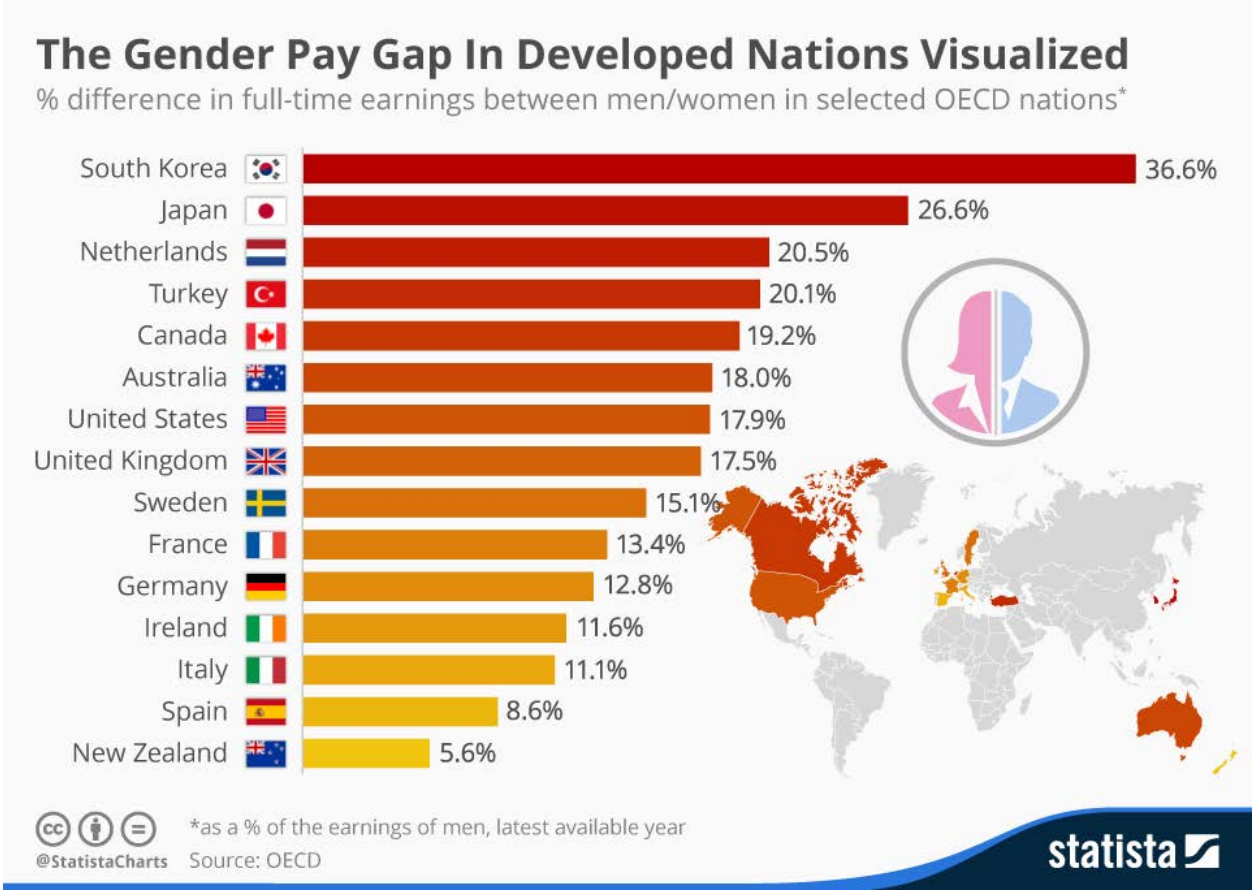

Figure 3. Gender Pay Gap in selected OECD nations by Statista

[7] https://www.statista.com/chart/4279/the-gender-pay-gap-in-developed-nations-visualised/

Objectives

With this project, I want to show which OECD member countries have the highest and the lowest Gender Wage Gap, whether the Gender Wage Gap has any correlation with the Gender Development Index (GDI) or the Human Development Index (HDI), and which OECD countries have the highest GDI and HDI ranking. Did the Gender Wage Gap increase or decrease over the years? Did GDI and HDI increase or decrease over the years?

Gender Wage Gap

The gender wage gap is the wage difference between men and women.[8] For example, if the gender wage gap is 30%, it means that for every dollar a man earns, a woman makes 70 cents. A real example can be seen in the 2019 South Korea Gender Wage Gap data from the OECD, in which its Gender Wage Gap percentage is 32.48. This is a very high score for an OECD country. Generally, the lower the percentage, the better and more equal the wage gap.

Human Development Index (HDI)

The Human Development Index (HDI) is an average index across key indicators in human development (Health, Education, and Gross National Per Capita Income).[9] According to the United Nations Development Program (UNDP), this index is a good way to compare countries with similar incomes but different human development outcomes and can be helpful in making policies. The HDI ratio values range from 0 to 1; the higher the score, the better the result.

Gender Development Index (GDI)

The United Nations Development Program defines GDI as a measure of gender inequality across three key human development measures: health, education, and command of economic resources[10]. For health, it measures life expectancy at birth for men and women; For education, it measures expected years of schooling for men and women; For the command of economic resources, it measures the estimated income at birth for both men and women. The GDI Index values also range from 0 to 1; the higher the score, the better the result. It is very similar to the HDI but focuses on Gender.

[8] https://www.britannica.com/topic/gender-wage-gap
[9] https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
[10] https://hdr.undp.org/gender-development-index#/indicies/GDI

Data & Methods

The first dataset I used is from the Organization for Economic Co-operation and Development (OECD). OECD is an important international organization that works with governments to find solutions to problems that countries face and share best practices for better lives. OECD also shares databases, statistics, maps, educational outputs, publications, and visualizations with people interested in various topics. One of their shared datasets is a dataset about the "Gender Wage Gap." This .csv dataset has 791 rows and seven columns: Location, Indicator, Subject, Measure, Frequency, Time, and Value. The time and value (percentage) columns are numerical, and the rest of the columns are categorical. There are 46 countries with gender gap percentage values. Some countries have gender gap values records from the 1970s, while others have only records after the 2000s. The latest year in the dataset is 2019.

The head of the gender wage gap dataset:

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value
0	AUS	WAGEGAP	EMPLOYEE	PC	A	1975	21.582734
1	AUS	WAGEGAP	EMPLOYEE	PC	A	1976	20.754717
2	AUS	WAGEGAP	EMPLOYEE	PC	A	1977	18.390805
3	AUS	WAGEGAP	EMPLOYEE	PC	A	1978	19.791667
4	AUS	WAGEGAP	EMPLOYEE	PC	A	1979	20.000000

```
RangeIndex: 791 entries, 0 to 790
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   LOCATION   791 non-null    object
 1   INDICATOR  791 non-null    object
 2   SUBJECT    791 non-null    object
 3   MEASURE    791 non-null    object
 4   FREQUENCY  791 non-null    object
 5   TIME       791 non-null    int64
 6   Value      791 non-null    float64
```

The second dataset I used is the Gender Development Index (GDI). The GDI is published by the UN Development Program (UNDP). GDI measures gender inequality in 3 dimensions which are

long and healthy life, the standard of living, and knowledge. The GDI ratio is calculated as the female Human Development Index (HDI) to the male HDI. The dataset includes 195 countries, with some countries not having any GDI values and others missing data for some years. Before 2010, GDI values were collected every five years from 1995 to 2010. After 2010, GDI values by country were collected every year. The dataset also has human development index by gender as an attribute by year. The dataset is available on the UN Development Program's website.

The head of the Gender Development Index dataset:

	iso3	country	hdicode	region	gdi_group_2019	gdi_1995	gdi_2000	gdi_2005	gdi_2010	gdi_2011	...	gni_pc_m_2010	gni_pc_m_2011	gni_pc_m_2012	gni_pc_m_2013	gni_pc_m_2014	gni_pc_m_2015	gni_pc_m_2016	gni_pc_m_2017
0	AFG	Afghanistan	Low	SA	5.0	NaN	0.322	0.519	0.595	0.609	...	3271.501421	3423.819414	3662.306275	3736.336223	3673.239176	3494.322823	3467.751600	3581.841147
1	AGO	Angola	Medium	SSA	4.0	NaN	NaN	NaN	NaN	NaN	...	8019.079329	7987.296132	8439.038862	8659.753034	8911.968376	8843.413491	8300.742016	7914.410262
2	ALB	Albania	High	ECA	2.0	0.938	0.936	0.942	0.961	0.958	...	13347.098260	13992.131660	13485.862740	14357.630790	14645.977210	14514.500590	15385.966020	15771.496410
3	AND	Andorra	Very High	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	ARE	United Arab Emirates	Very High	AS	3.0	0.952	NaN	NaN	NaN	0.957	...	67323.749300	68935.133390	70836.912400	74446.869010	77980.805240	82054.289990	83903.651850	84824.814320

5 rows × 148 columns

```
RangeIndex: 206 entries, 0 to 205
Columns: 148 entries, iso3 to gni_pc_m_2019
```

The final dataset I used is the Human Development Index (HDI). HDI is considered a new way to measure countries' economic health instead of GDP, and it factors the life expectancy, education, and standard of living. The UNDP compiles the Human Development Index data that comes from United Nations Agencies. This dataset is also available on the UN Development Program's website. The dataset has HDI data from 1990 to 2019. However, the data has only been collected annually after 2013. The latest available year is 2019, as well as for the GDI and Gender Wage Gap data.

The head of the Human Development Index dataset:

	iso3	country	hdicode	region	hdi_rank_2019	hdi_1990	hdi_1991	hdi_1992	hdi_1993	hdi_1994	...	gnipc_2010	gnipc_2011	gnipc_2012	gnipc_2013	gnipc_2014	gnipc_2015	gnipc_2016	gnipc_2017	gnipc_2018
0	AFG	Afghanistan	Low	SA	169.0	0.302	0.307	0.316	0.312	0.307	...	1917.394944	2013.614084	2164.641446	2229.906554	2214.41439	2128.161886	2134.866156	2229.657978	2217.175808
1	AGO	Angola	Medium	SSA	148.0	NaN	NaN	NaN	NaN	NaN	...	6913.160589	6887.003763	7282.049679	7478.856252	7704.36784	7652.152491	7189.031576	6861.580571	6360.551085
2	ALB	Albania	High	ECA	69.0	0.650	0.631	0.615	0.618	0.624	...	10774.721800	11237.447160	11365.140100	11806.357820	11951.26299	12273.472790	12753.307240	13071.095440	13636.864160
3	AND	Andorra	Very High	NaN	36.0	NaN	NaN	NaN	NaN	NaN	...	49261.522250	47366.246500	47347.415550	48486.415270	50567.86966	51779.832310	53245.151100	54371.344670	55253.539290
4	ARE	United Arab Emirates	Very High	AS	31.0	0.723	0.735	0.738	0.745	0.755	...	54911.286620	56152.974740	57447.350900	60007.280900	62499.79784	65528.562580	66881.303340	67667.529860	67195.144070

5 rows × 155 columns

```
RangeIndex: 206 entries, 0 to 205
Columns: 155 entries, iso3 to gnipc_2019
```

Due to data limitations and missing values, I decided to make my project scope smaller and chose only OECD member countries. Currently, there are 38 members of OECD. I included 18 of these 38 OECD countries' data between 2010-2019 in the project. These countries are Australia,

South Korea, Colombia, USA, Sweden, New Zealand, Mexico, Japan, Hungary, Finland, Canada, Israel, Slovakia, Norway, Germany, Denmark, Czech Republic, and Belgium.

I cleaned up rows of countries and years, self-employed data, which I don`t intend to use in the project, and dropped the unnecessary columns in python for each dataset as the data wrangling process. I reset the index and used the stack function to shape my datasets.

Gender Wage Gap Dataset

	location	year	gwg
0	AUS	2010	14.042934
1	AUS	2011	15.966387
2	AUS	2012	13.750000
3	AUS	2013	18.000000
4	AUS	2014	17.050691
...	...	...	...
175	USA	2015	18.882682
176	USA	2016	18.142077
177	USA	2017	18.172157
178	USA	2018	18.910586
179	USA	2019	18.470705

180 rows × 3 columns

HDI Dataset

	country	year	hdi
0	Australia	2010	0.930
1	Australia	2011	0.932
2	Australia	2012	0.937
3	Australia	2013	0.931
4	Australia	2014	0.933
...	...	...	...
175	United States	2015	0.921
176	United States	2016	0.922
177	United States	2017	0.924
178	United States	2018	0.925
179	United States	2019	0.926

180 rows × 3 columns

GDI Dataset

	country	year	gdi
0	Australia	2010	0.976
1	Australia	2011	0.976
2	Australia	2012	0.976
3	Australia	2013	0.975
4	Australia	2014	0.975
...	...	...	...
175	United States	2015	0.994
176	United States	2016	0.994
177	United States	2017	0.995
178	United States	2018	0.993
179	United States	2019	0.994

180 rows × 3 columns

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   location  180 non-null    object
 1   year      180 non-null    int64
 2   gwg       180 non-null    float64
```

```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   country  180 non-null    object
 1   year     180 non-null    object
 2   hdi      180 non-null    float64
```

```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   country  180 non-null    object
 1   year     180 non-null    object
 2   gdi      180 non-null    float64
```

	location	country	year	gwg	hdi	gdi
0	AUS	Australia	2010	14.042934	0.930	0.976
1	AUS	Australia	2011	15.966387	0.932	0.976
2	AUS	Australia	2012	13.750000	0.937	0.976
3	AUS	Australia	2013	18.000000	0.931	0.975
4	AUS	Australia	2014	17.050691	0.933	0.975
...	...	...	...	...	...	...
175	USA	United States	2015	18.882682	0.921	0.994
176	USA	United States	2016	18.142077	0.922	0.994
177	USA	United States	2017	18.172157	0.924	0.995
178	USA	United States	2018	18.910586	0.925	0.993
179	USA	United States	2019	18.470705	0.926	0.994

180 rows × 6 columns

I concatenated these three individual datasets to create a large dataset and dropped duplicated “Year” attributes; changed location and country column types to categories.

The dataset has six attributes and 180 rows. The attributes are location, country, year, gender wage gap percentage, and HDI and GDI ratios. The dataset includes 18 OECD member countries with data from between 2010-2019.

Descriptive statistics summary of the dataset:

	year	gwg	hdi	gdi
count	180.000000	180.000000	180.000000	180.000000
mean	2014.500000	14.551770	0.894444	0.977683
std	2.880293	8.023809	0.056708	0.015979
min	2010.000000	3.298900	0.729000	0.927000
25%	2012.000000	7.149332	0.887750	0.970000
50%	2014.500000	15.003014	0.916000	0.981000
75%	2017.000000	18.652107	0.931250	0.990000
max	2019.000000	39.605857	0.957000	1.006000

The Kernel Density Estimate helps us to visualize the distribution of the observation. I visualized the KDE of 2019`s Gender Wage Gap (percentage), HDI, and GDI data using Seaborn. As we can see from the graphs below, the gender wage gap graph has a right skew distribution, and the GDI and HDI graphs have a left skew distribution. In 2019, the average gender wage gap percentage of selected 18 OECD countries was around 14%, while worldwide, women only make 77 cents for every dollar men earn, making the gender wage gap 23%.

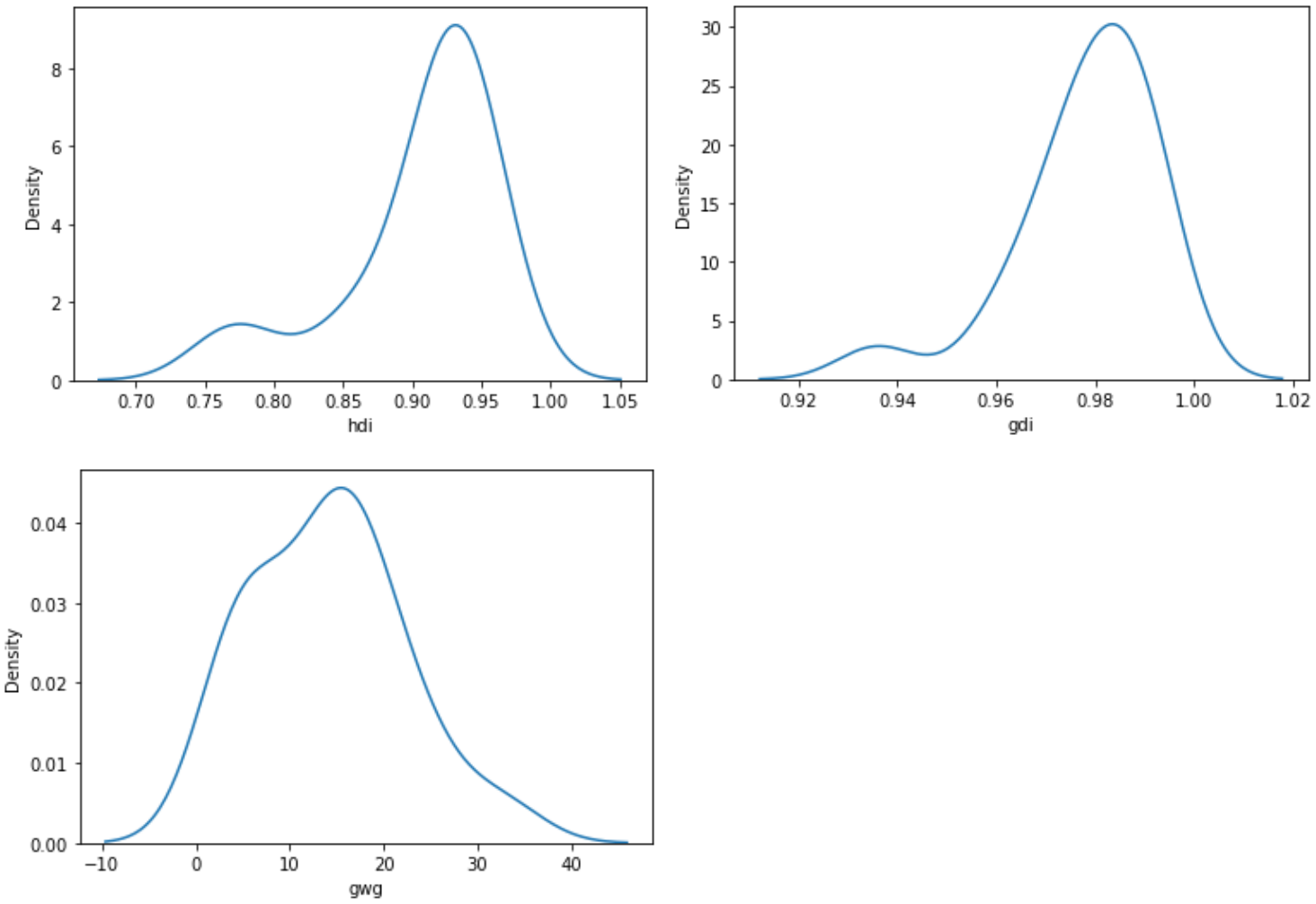

Figure 4. Kernel Density Plots of 2019 HDI/GDI/Gender Wage Gap

Plotly express is a Python built-in function of the plotly library. I created three interactive line charts with markers using Plotly Express to show how the selected OECD countries are doing on the human development index, gender development index, gender wage gap (percentage), and their rankings for the past ten years. Each country`s data can be seen in detail when you hover your mouse over it.

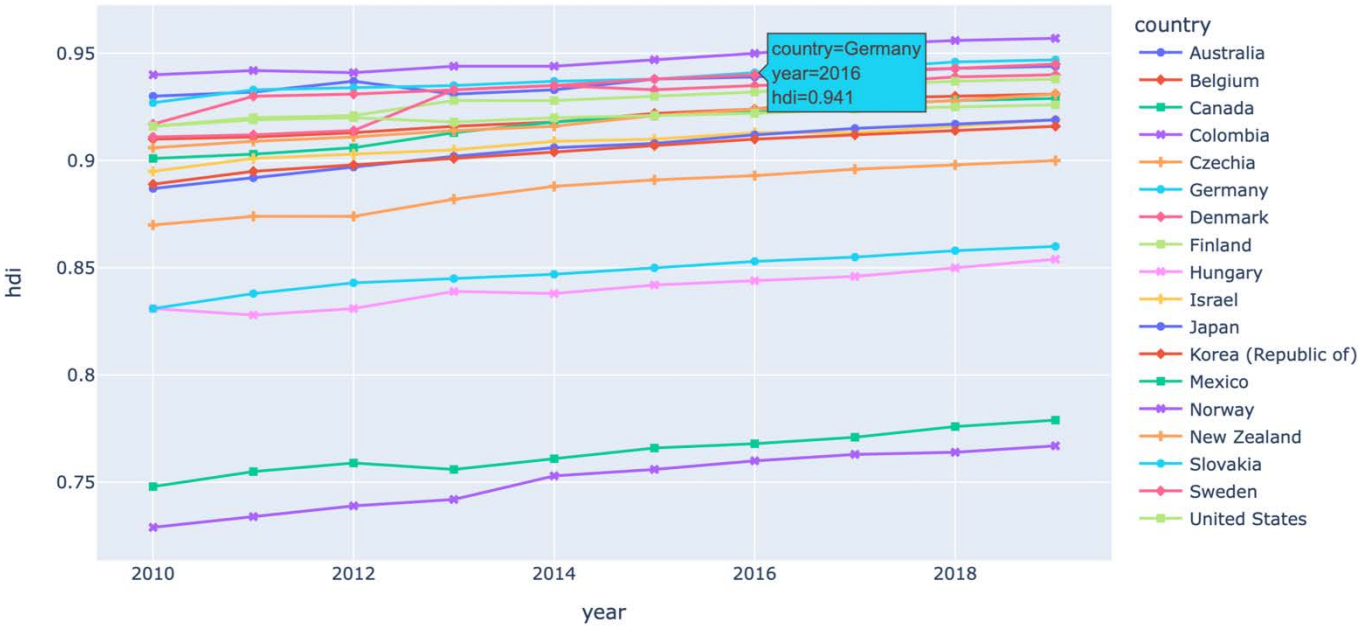

Figure 5. 10-year Human Development Index progression of 18 selected OECD countries

A high HDI means that the country has a high standard of living, with decent healthcare, education, and economic opportunities. [11] According to Wikipedia, the majority of OECD members are high-income economies with a very high Human Development Index (HDI). [12] The year data is on the X-axis, and the HDI ratio is on the Y-axis of Figure 5. Over the years, the developed countries' HDI ratio increased significantly. Norway, Germany, and Sweden have the highest HDI ratio, while Mexico (0.779) and Colombia (0.767) have the lowest HDI ratio of the selected OECD countries. The world average HDI ratio was around 0.72 in 2019. The selected OECD member countries have a higher HDI ratio than the average world HDI ratio.

[11] https://www.investopedia.com/terms/h/human-development-index-hdi.asp
[12] https://en.wikipedia.org/wiki/OECD

Figure 6. 10-year Gender Development Index progression of 18 selected OECD countries

Another example of interactive visualization is the GDI ratio line chart. The year data is on the X-axis, and the GDI ratio is on the Y-axis in Figure 6. The Gender Development Index (GDI) is the ratio of female to male Human Development Index (HDI) values, and it captures only part of what human development implicates. It does not reflect on inequalities, poverty, or human security.[13] [14] As we can see from Figure 6, over the years, the selected OECD countries' GDI ratios are trending upward, but not as fast as the HDI ratios. The United States and Slovakia have the highest GDI ratio, and South Korea has the lowest GDI ratio (0.936), which is significantly lower than the selected OECD countries.

[13] https://resourcewatch.org/data/explore/soc002rw1-gender-development-index-gdi
[14] https://hdr.undp.org/gender-development-index#/indicies/GDI

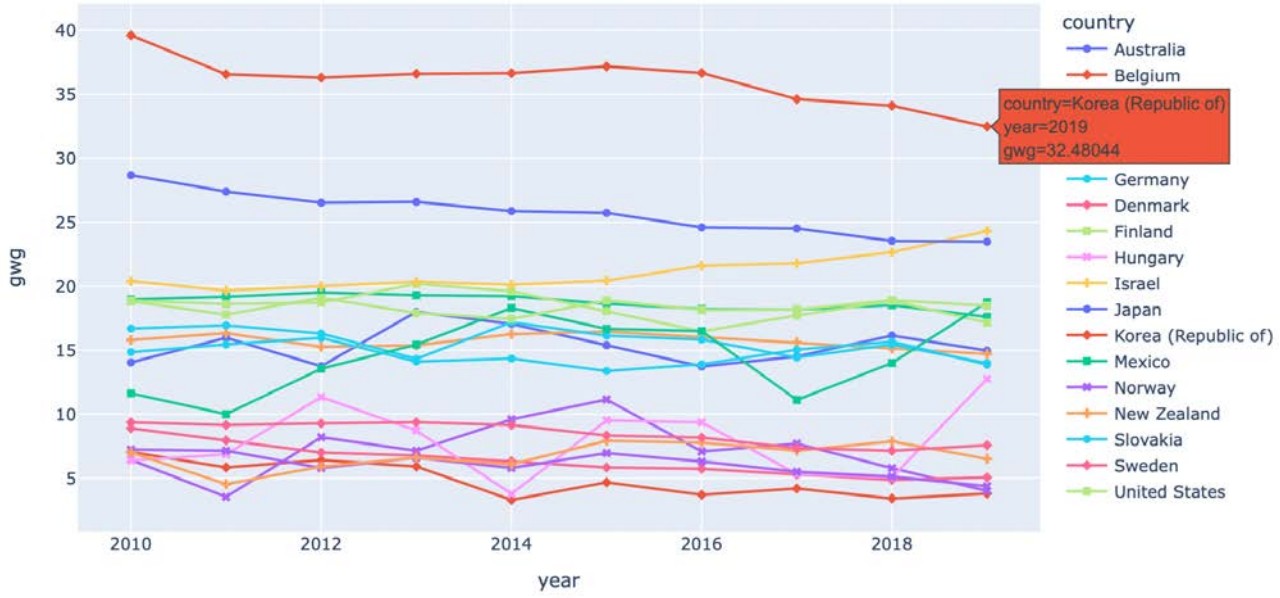

Figure 7. 10-year Gender Wage Gap progression of 18 selected OECD countries

UN Women is a UN organization delivering programs, policies, and standards that uphold women's human rights. According to UN Women, Worldwide, women only make 77 cents for every dollar men earn, and at the current rate of progress, there is no equal pay until 2069. [15]

I created a plot chart to show the gender wage gap in the selected OECD countries. The X-axis shows the year; the Y-axis shows the percentage of the gender wage gap. While Belgium, Colombia, and Norway have the lowest gender wage gap percentage, South Korea (32.48), Israel (24.3), and Japan (23.48) have the highest gender wage gap percentage, according to the 2019 data. The selected OECD countries have improved slightly in reducing the gender wage gap percentage, but the current gender wage gap is still significant, and there is a long way to go to achieve equality. It is interesting to see that countries like South Korea, Japan, and the United States have high HDI and GDI ratios and are economically well developed. However, they still have a significant high gender wage gap. It would be good to explore the reasons for this.

According to the OECD publication called "Closing the Gender Gap," if high childcare costs are not economically worthwhile for women to work full time or if the workplace culture penalizes women for interrupting their careers, if women bear the burden of unpaid household chores, childcare, it will be difficult for them to realize their full potential in paid work.[16] This issue can contribute to the gender wage gap.

[15] https://www.unwomen.org/en/news/in-focus/csw61/equal-pay

[16] OECD (2022), "Gender wage gap" (indicator), https://doi.org/10.1787/7cee77aa-en

I created two interactive visualizations to demonstrate the 10-year Progression of the Gender Wage Gap by country. The interactive visualizations have a play button to show progression by year between 2010-2019. The color scale on the right represents the percentage of the gender wage gap. Figure 8 shows the gender wage gap percentage as a choropleth map. The graph only shows the selected OECD countries. I would have preferred to have more data for the choropleth map. Figure 9 shows a scatter map that uses bubbles to represent the gender wage gap. The size and color of the bubbles changes depending on the size of the gender wage gap.

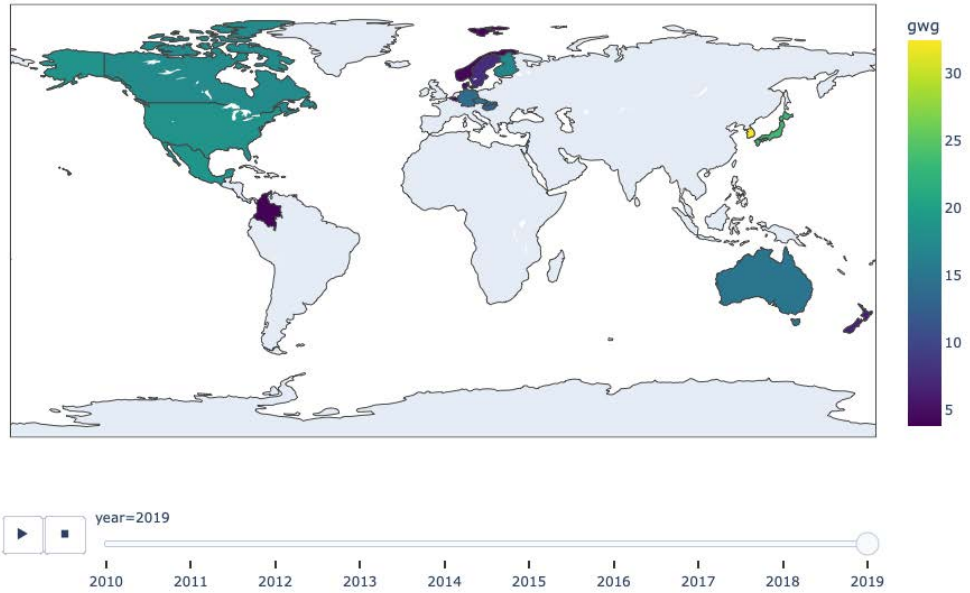

Figure 8. Interactive 10-year Gender Wage Gap Progression Choropleth Map

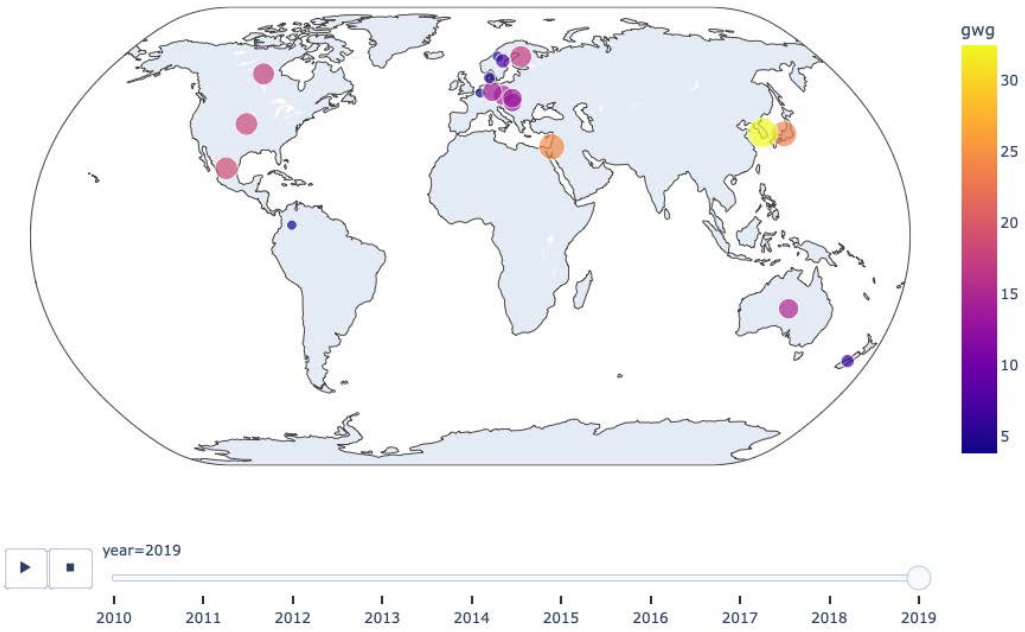

Figure 9. Interactive 10-year Gender Wage Gap Progression Scatter Map

Another visualization I created is an interactive five dimensions bubble chart. The X-axis shows the HDI ratio, the Y-axis shows the GDI ratio, the color code on the right represents countries, and the bubble size represents the size of the gender wage gap. The interactive visualization also has a play button showing progression by year from 2010 to 2019. As shown in Figure 10, the GDI and HDI ratios have positive correlations except for a few countries, such as South Korea, according to the 2019 data. We can`t deny the influence of economic development on the gender wage gap, but it is not the only factor that affects it. Other complex issues are at play, such as gender inequality, male domination, long working hours, sex discrimination, raising children, etc., that need to be addressed.

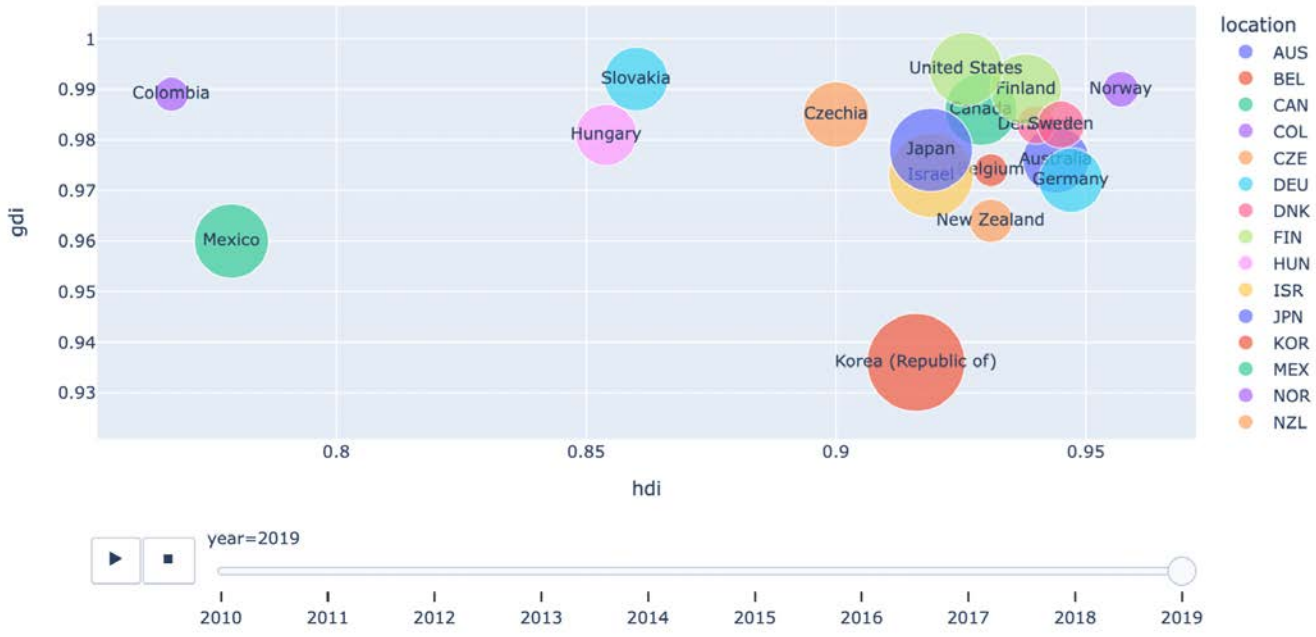

Figure 10. Interactive 10-year Country Progression Bubble Chart

Treemap is a type of visualization in which a dataset is hierarchically ordered with nodes, branches, and roots. Each rectangle is a tree branch, and smaller rectangles are sub-branches. Treemaps might cause size distortion, but still, they are easy to understand and display information efficiently. I created two treemaps to visualize the gender wage gap percentage. Each gender pay gap data can be seen when you hover your mouse over them.

Figure 11 demonstrates the gender wage gap by country. The countries are ordered from the highest to the lowest from left to right. For example, South Korea had the highest gender wage gap percentage in 2010, 2015, and 2016. South Korea has also shown improvement in the gender wage gap in recent years and had its lowest gender wage gap values in 2017, 2018, and 2019. The visualization provides a quick summary of the gender wage gap situation in each country.

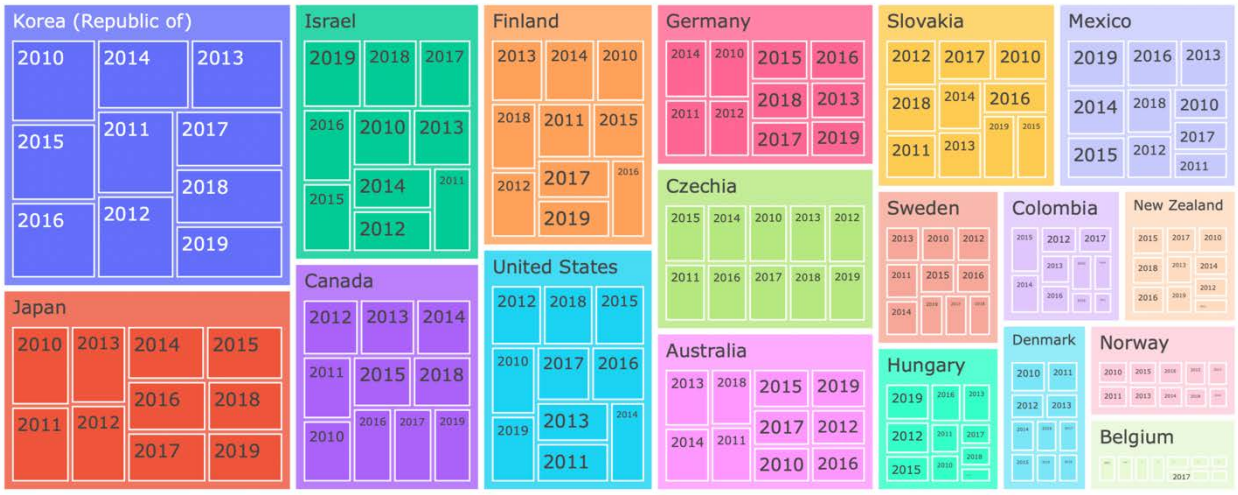

Figure 11. Gender Wage Gap Treemap by Country

Figure 12 demonstrates the gender wage gap by year. Each tree shows the country's yearly wage gap data from highest to lowest. For example, in 2015, while South Korea had the highest gender wage gap, Belgium had the lowest gender wage gap percentage among selected OECD countries. The visualization provides a quick summary of the gender wage gap data by year.

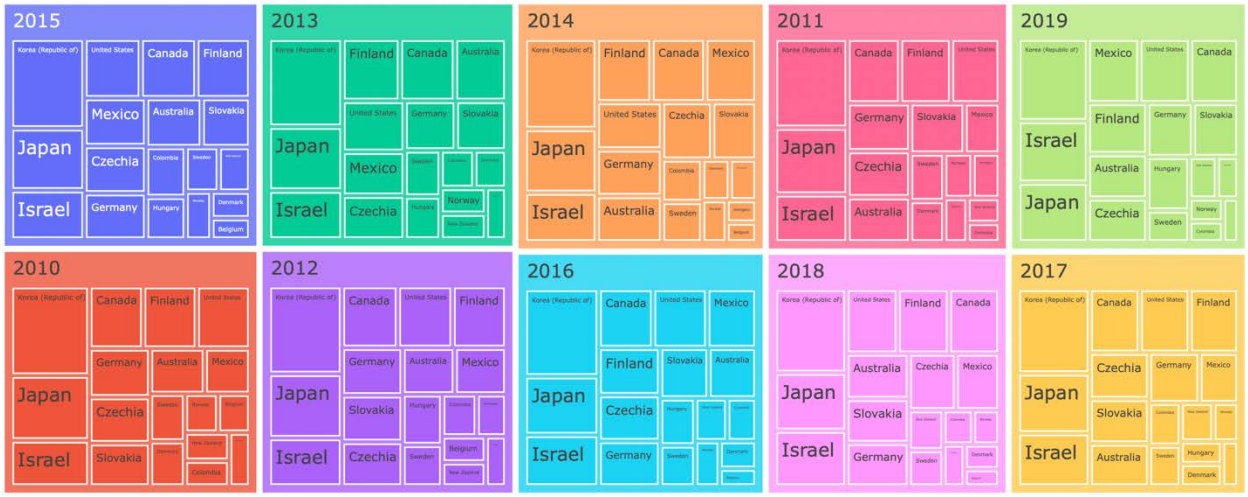

Figure 12. Gender Wage Gap Treemap by year

Discussion and Conclusion

I really enjoyed working on this project, as I am very passionate about the gender wage gap. The research and visualizations I produced from this project validated my own thoughts that much needs to be done to address this issue. As I mentioned earlier, at the current pace, the OECD countries will not achieve parity in pay until 2069. While there is undoubtedly progress, the situation is not improving fast enough. I did discover that HDI and GDI are correlated and have been improving over the years in the selected OECD countries but are not improving at the same pace. Economic development measures such as HDI and GDI are factors in the gender wage gap. Still, there are other factors such as gender inequality, male domination, long working hours, sex discrimination, pregnancy, raising children, etc.

Data Science and visualizations can play an essential role in bringing attention to this issue and forcing decision-makers to take action. If I had more time and data, I would have liked to explore the factors and correlations that influence the gender wage gap and study more countries. The project helped me learn about new visualization techniques. This was the first time I used interactive visualizations. I also learned and used python libraries that I have never used before, such as Plotly express, and improved my data ingestion and wrangling techniques. In the future, I would like to be able to use tools such as Stremlit to enhance the interactivity of the visualizations.

Appendix

Jupyter Notebook Script:

https://www.dropbox.com/sh/bxeg6l58h3p1dlt/AACNkn7UL9Czkovau-A63zhHa?dl=0

References:

OECD (2022), Gender wage gap (indicator). doi: 10.1787/7cee77aa-en

Wisniewski, Megan. In Puerto Rico, No Gap in Median Earnings Between Men and Women March 01, 2022, https://www.census.gov/library/stories/2022/03/what-is-the-gender-wage-gap-in-your-state.html

Women and Girls – Closing the Gender Gap
https://www.un.org/en/un75/women_girls_closing_gender_gap

Gender wage gap
https://data.oecd.org/earnwage/gender-wage-gap.htm

McCarthy, Niall. The Gender Pay Gap In Developed Nations Visualised
Jan 26, 2016, https://www.statista.com/chart/4279/the-gender-pay-gap-in-developed-nations-visualised/

Barnes Medora W. Gender wage gap
https://www.britannica.com/topic/gender-wage-gap

United Nations Development Programme (UNDP). Human Development Index (HDI)
https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

United Nations Development Programme (UNDP). Gender Development Index (GDI)
https://hdr.undp.org/gender-development-index#/indicies/GDI

The Investopedia Team. Human Development Index (HDI)
January 29, 2022, https://www.investopedia.com/terms/h/human-development-index-hdi.asp

OECD
https://en.wikipedia.org/wiki/OECD

Equal pay for work of equal value
https://www.unwomen.org/en/news/in-focus/csw61/equal-pay

Gender Development Index
https://resourcewatch.org/data/explore/soc002rw1-gender-development-index-gdi