Nina Hua (nhua2)
MSDS 622
Final Project

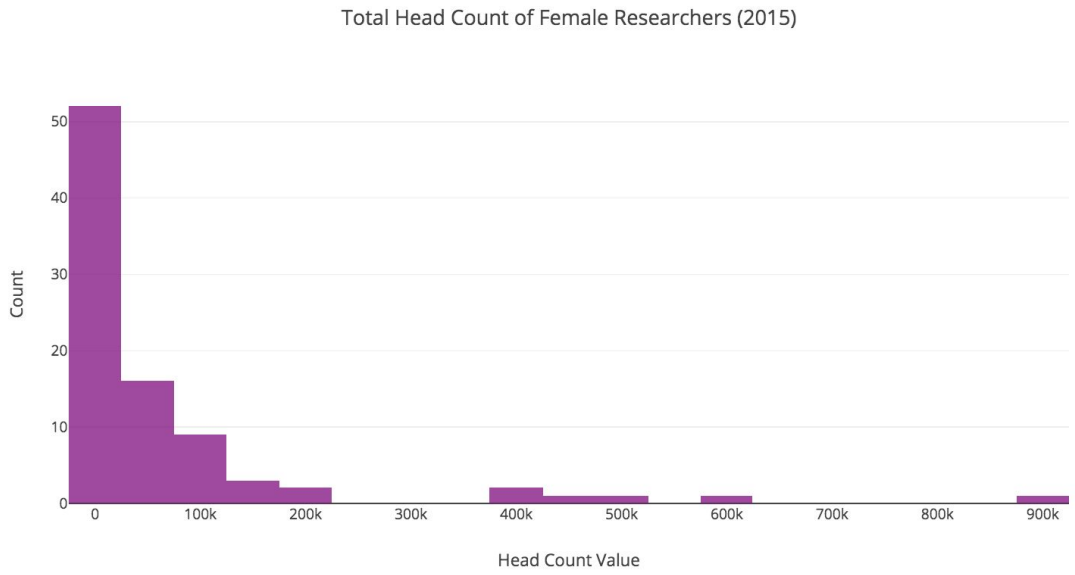
Introduction

For the final project, I wanted to explore the proportion of women who work in STEM fields around the world. Particularly, I wanted to answer if the proportion of women in STEM has increased over the years with the investment of education in girls. This stemmed from the fact that achieving gender equality and empowering women and girls is one of the United Nations' seventeen Sustainability Development Goals. As STEM fields are very male-oriented, I wanted to visualize if the number of women entering STEM fields around the world has changed over the last several years. My main data set consists of women who work in R&D, research, and as technicians from the UNESCO Institute for Statistics.

Summary of Data

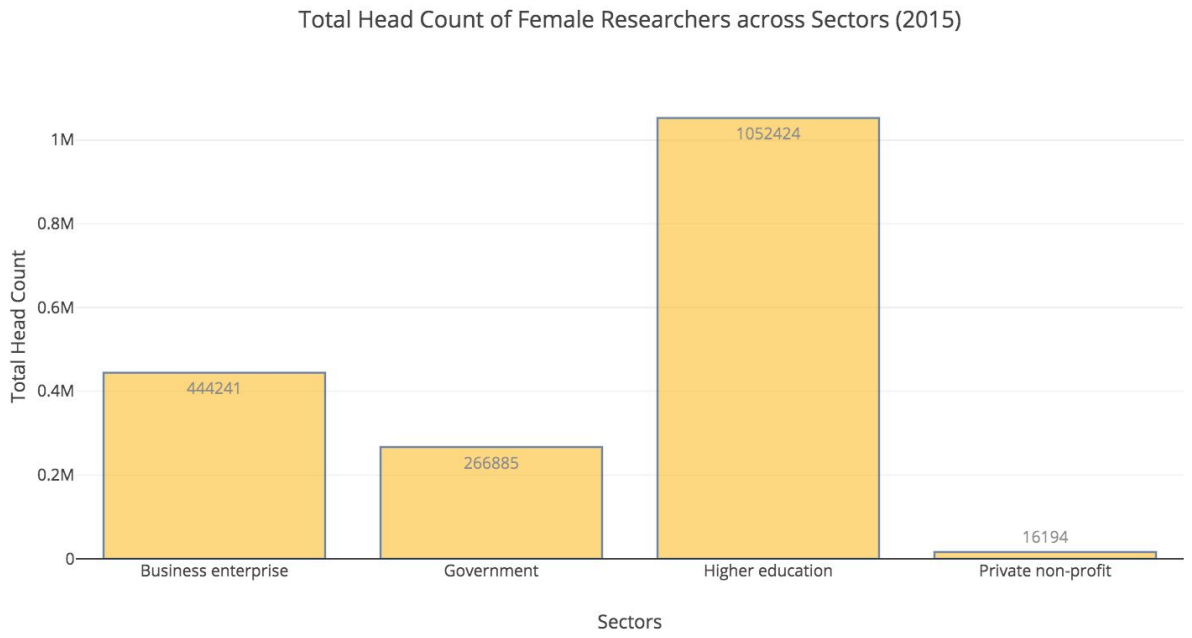
*With the exception of the connection map, bubble map, and treemap, all plots were created using Plotly, which implies that all plots are interactive.

1. What is the distribution of female researchers in 2015?



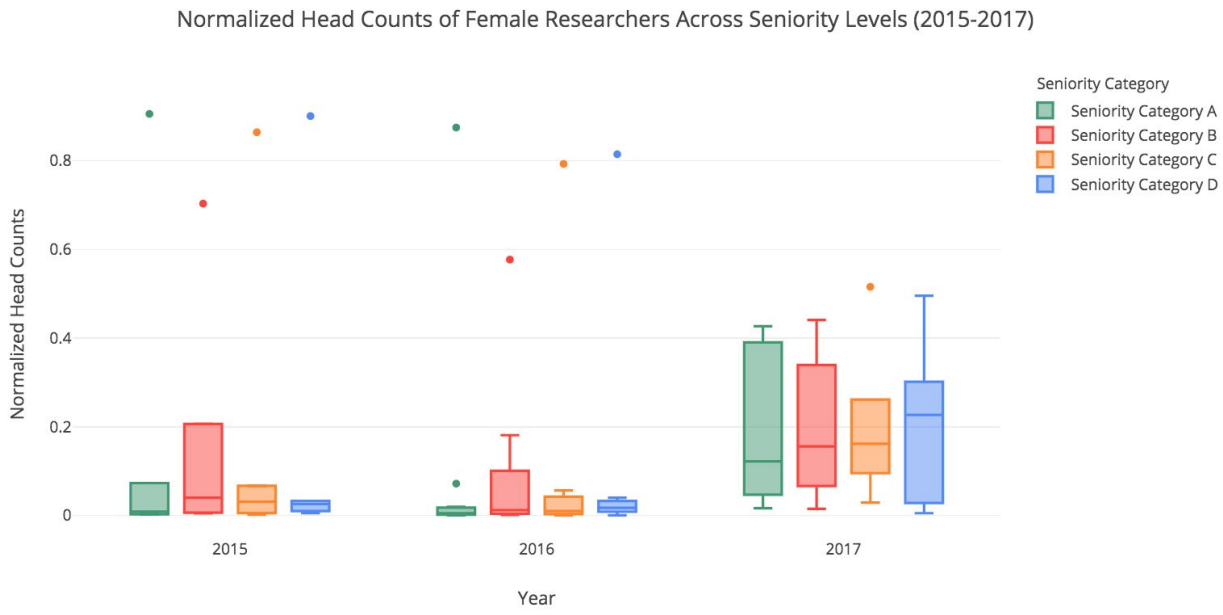
A majority of countries have reported having fewer than 100,000 female researchers (in headcount) in 2015.

2. How many female researchers are there in different sectors as reported by UNESCO?



UNESCO has defined four sectors of research: Business enterprise, Government, Higher Education, and Private non-profit. The headcounts were summed as an aggregate value, which showed that Higher education has the highest number of female researchers in 2015.

3. How many female researchers are there across seniority levels as reported by UNESCO?



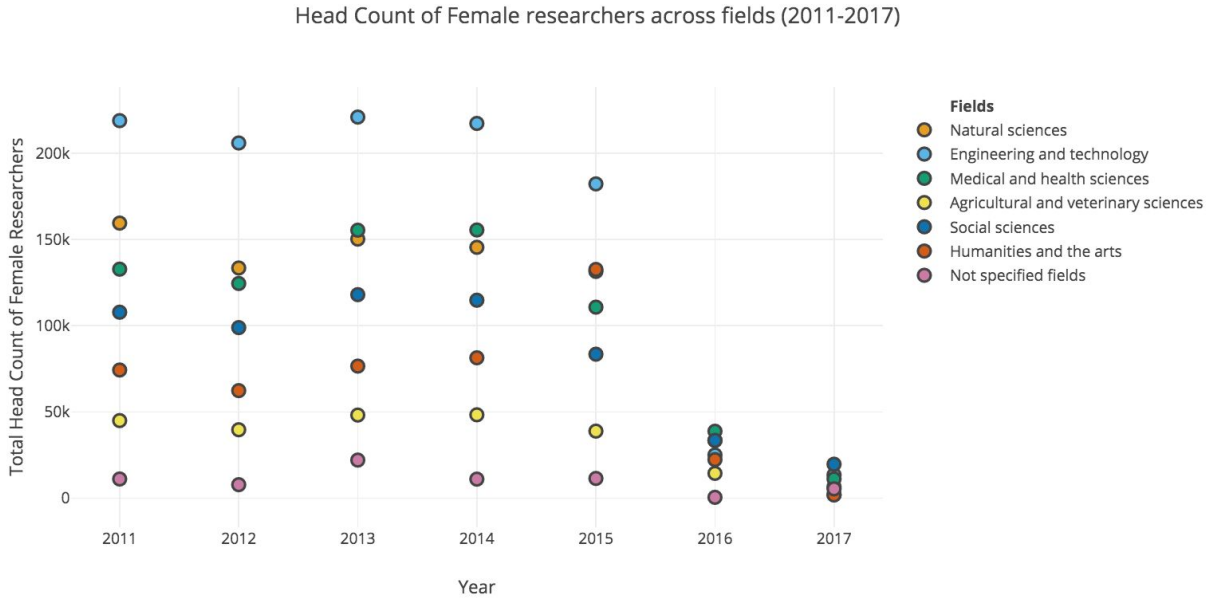
UNESCO has defined four categories of seniority:[1]

- Category A: The single highest grade/post at which research is normally conducted (e.g. Director of research, Full professor)

- Category B: Researchers working in positions not as senior as top position (A) but more senior than newly qualified doctoral graduates (ISCED level 8) (e.g. Senior researcher, Principal investigator, Associate professor)
- Category C: The first grade/post into which a newly qualified doctoral graduate would normally be recruited. (e.g. Researcher, Investigator, Assistant professor, Post-doctoral fellow)
- Category D: Either doctoral students at the ISCED level 8 who are engaged as researchers or researchers working in posts that do not normally require a doctorate degree (e.g. Ph.D. students, junior researchers)

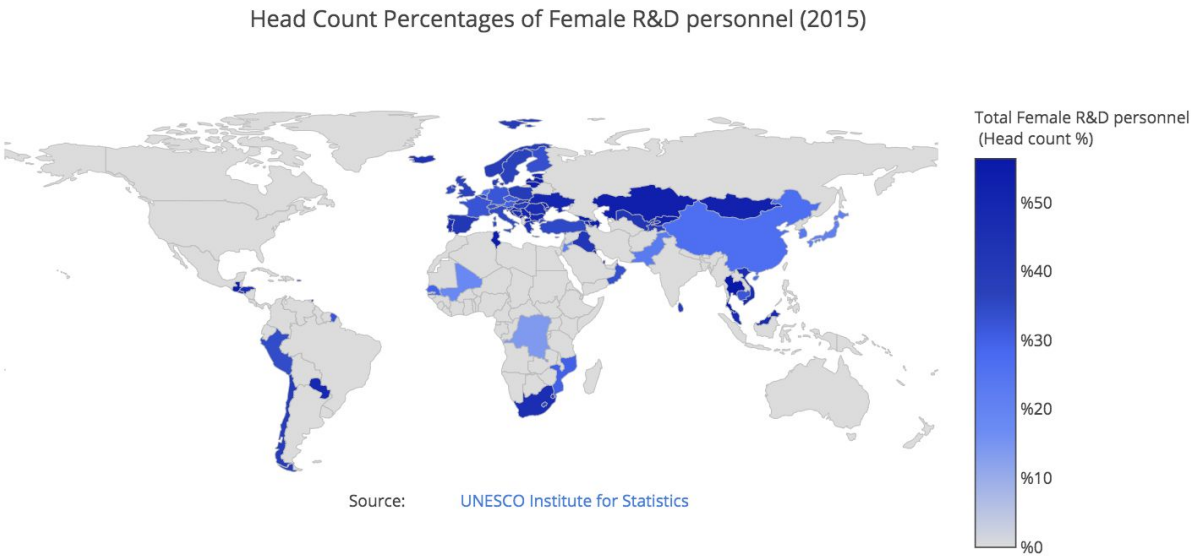
Headcount values were normalized to each category level per year. The highest normalized headcount values of female researchers were found in 2017.

4. What is the distribution of female researchers across different fields between 2011 and 2017?



UNESCO has defined seven fields of research: Natural sciences, Engineering and Technology, Medical and health sciences, Agricultural and veterinary sciences, Social sciences, Humanities and the arts, and unspecified fields. The headcount values were aggregated per year and field. Although the scatter plot shows a suspicious decreasing trend, there were fewer countries that have reported headcount values, so a trend should not be inferred from this visualization.

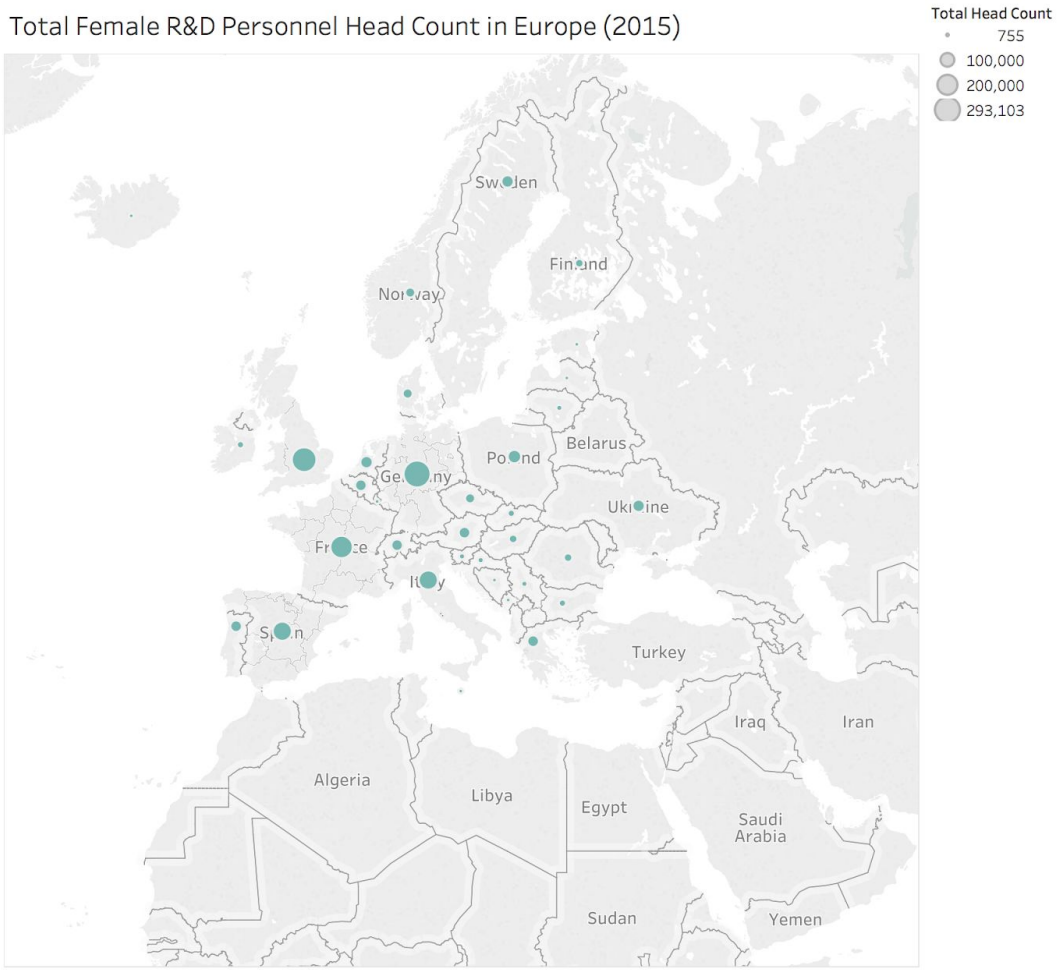
5. What is the global distribution of female Research and Development employees?



The top five countries with the highest headcount percentages of female R&D personnel in 2015 include:

- Thailand
- Tunisia
- Guatemala
- Latvia
- Armenia

6. What is the distribution of female Research and Development employees in Europe?



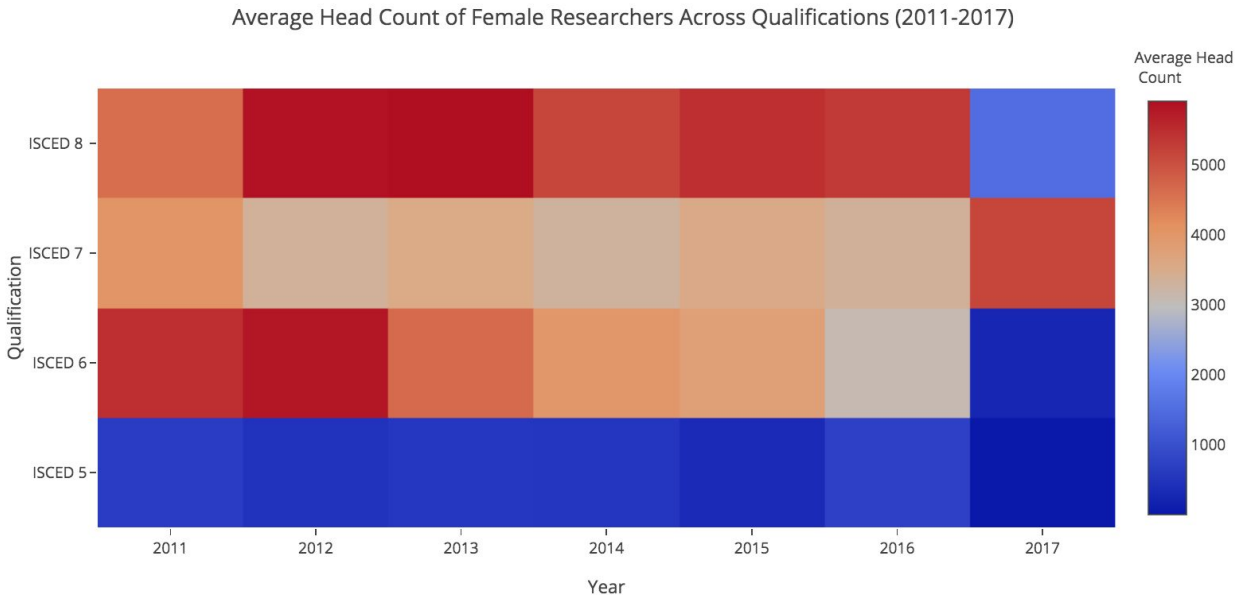
The choropleth map prior showed that European countries contained a great number of countries with female R&D personnel. This bubble map was used to further explore the headcount of personnel in 2015. Five countries with the greatest number of female R&D personnel include:

- Germany
- United Kingdom of Great Britain and Northern Ireland
- France
- Spain
- Italy

7. As ensuring gender equality is a Sustainability Development Goal, where are UN offices located in the world?

In order to create the connected graph above, I created my own data set that held the information centers of the United Nations where people can learn more about the sustainability goals.[2] Since there are fifty-nine offices around the globe, it was difficult to depict the offices as one connected graph. Random offices from five defined regions (Americas, Africa, Arab State, Asia and the Pacific, and Europe) were selected to be featured as connection points.

8. What is the average number of female researchers with certain qualifications?

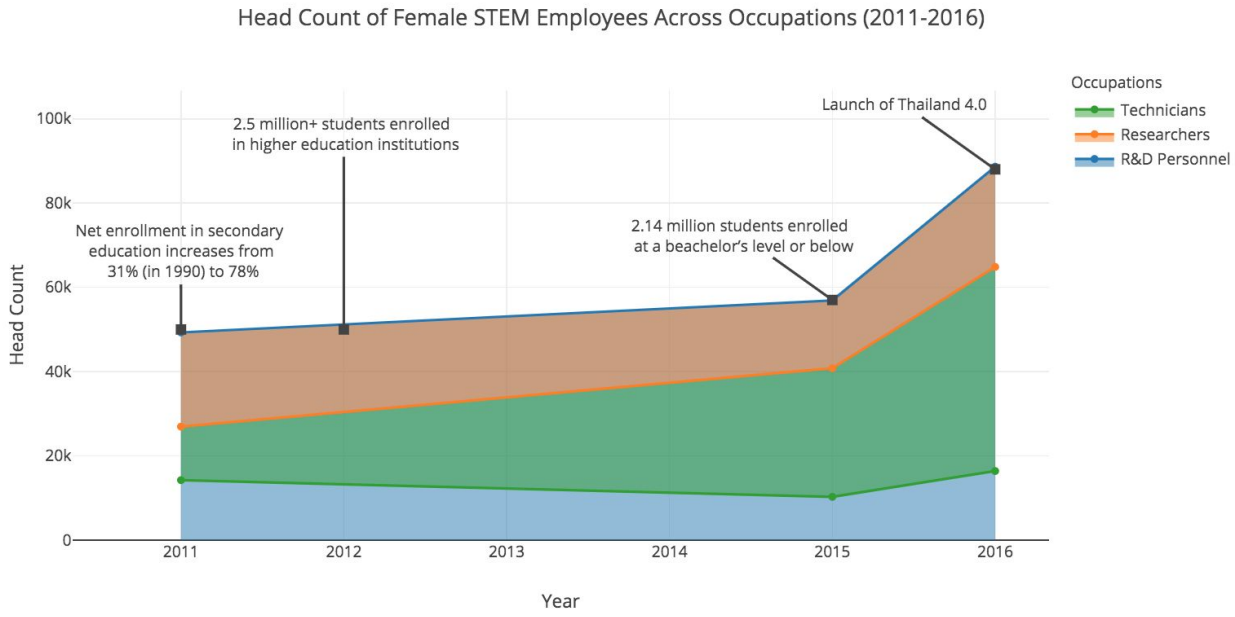


Qualifications are defined as the following:[3]

- ISCED 5: Short-cycle tertiary education
- ISCED 6: Bachelor's or equivalent level
- ISCED 7: Master's or equivalent level
- ISCED 8: Doctoral or equivalent level

The greatest average values of female researchers with a doctorate degree occurred between 2012 and 2013. The greatest average value of female researchers with a master's degree occurred was recorded in 2017. The greatest average value of female researchers with a bachelor degree occurred was recorded in 2012. However, the dataset does not include the same number of countries each year as some countries did not report their headcount values, so the average is not equally representative through time.

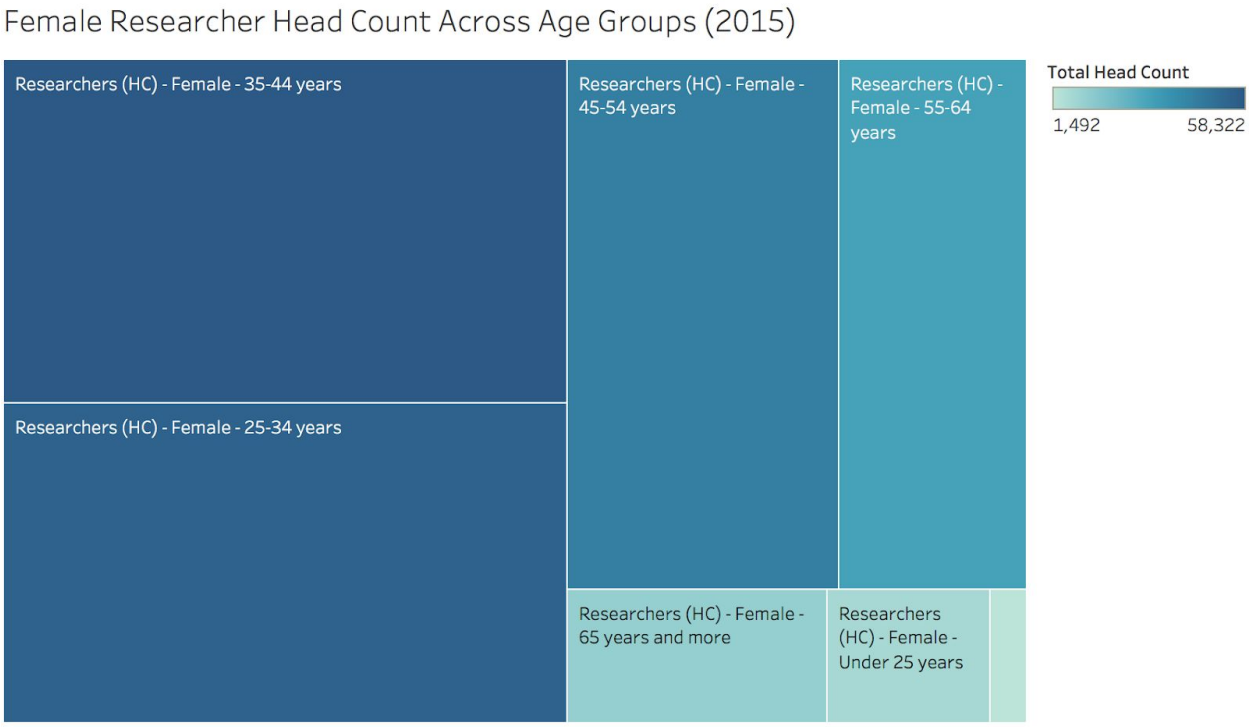
9. Since Thailand consistently shows an equal or greater proportion of women as R&D staff, researchers, and technicians, how do these proportions compare with one another over time?



The headcount values of female researchers and R&D personnel in Thailand increased from 2011 to 2016. However, the headcount values of female technicians decreased from 2011 to 2015 before increasing in 2016. Facts used to support the annotations can be found in the citation.[4,5]

10. What is the distribution of female researchers across age groups?

Nina Hua (nhua2)
MSDS 622
Final Project

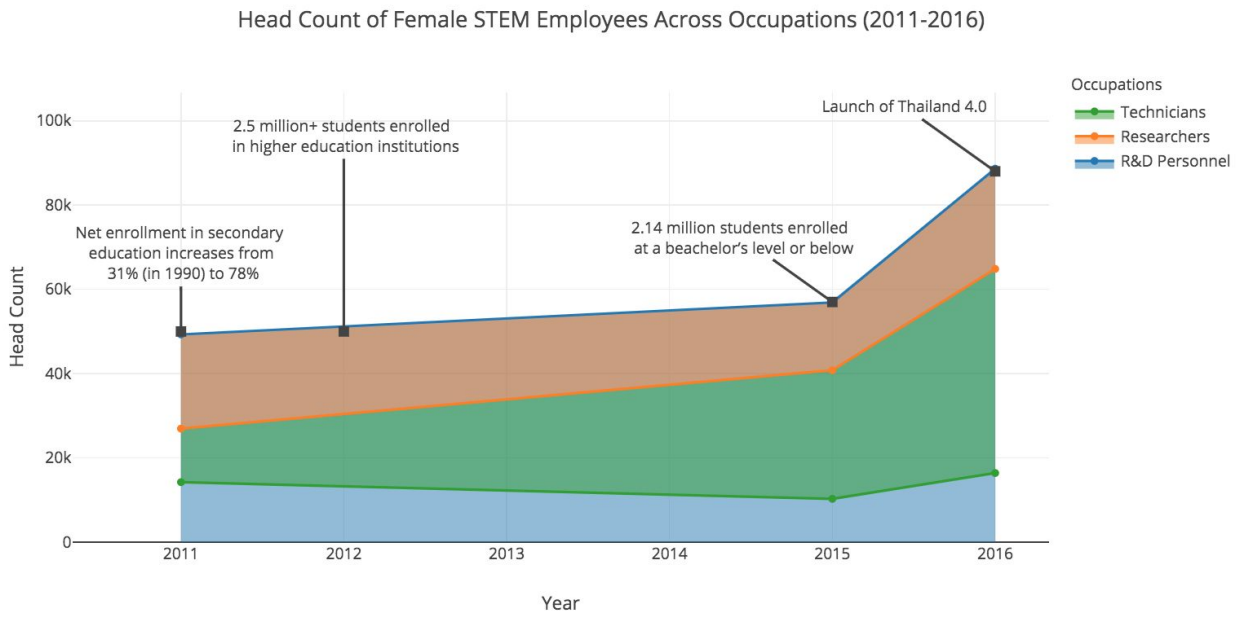


UNESCO has defined the following age groups:

- Not specified
- Under 25 years
- 25 - 34 years
- 35 - 44 years
- 45 - 54 years
- 55 - 64 years
- 65 years and more

This treemap displays the proportion of female researchers across different age groups. The largest proportion of female researchers are between the ages of thirty-five and forty-five years old.

Storyline



Thailand is an upper middle income Southeast Asian country that has shown the rewards of investing in education for all children, especially for girls.[6] In 1999, the National Education Act was enacted in order to ensure that all Thai children received access to education.[7] Through the years, Thailand has been devoted to creating a value-based economy that is driven by innovation, technology, and creativity with the adoption of the initiative Thailand 4.0.[5] Although Thailand boasts near gender equality in STEM fields, they have partnered with UNESCO to promote female STEM education. They are the first country in Asia-Pacific to pilot a policy toolkit under UNESCO's global STEM and Advancement, or SAGA. SAGA aims to analyze the impact of policies on gender disparities in STEM fields.[8] As a country that has evolved from a low-income economy to an upper-middle-income economy in less than a generation, it is worth believing that the continuous investment of educating all individuals is invaluable to a nation's success.[6]

Results/Summary/Conclusion

Overall, the world is seeing a surge of women in STEM, but there are countries and years where the proportion between genders is unequal. With the advent of programs such as SAGA and Girl Up's STEM for Social Good, there is confidence that the gap between the genders will decrease in the future.

Link to your github page with this analysis

https://github.com/nina-hua/data_visualization/tree/master/final

Citations

1. http://uis.unesco.org/node/458458
2. https://unic.un.org/aroundworld/unics/en/whereWeWork/africa/index.asp?regionCode=1
3. http://uis.unesco.org/en/glossary
4. https://wenr.wes.org/2018/02/education-in-thailand-2

5. https://thaiembdc.org/thailand-4-0-2/
6. http://www.worldbank.org/en/country/thailand/overview
7. https://borgenproject.org/facts-about-education-in-thailand/
8. https://thaiembdc.org/2017/09/07/thailand-promoting-female-stem-education-with-unesco/

barchart

May 12, 2019

1 Appendix

2 Bar Chart

Bar Chart of Female Researchers (head count %) across different sectors: - Business enterprise
- Government
- Higher education
- Private non-profit
- Not specified

```
In [1]: import pandas as pd
        import numpy as np
        import plotly.plotly as py
        import plotly.graph_objs as go
        from collections import Counter

In [2]: data = pd.read_csv('../data/section_researchers.csv')  # reading in data

In [4]: # filtering
        sectors = ['Researchers (HC) - Female - Business enterprise',
               'Researchers (HC) - Female - Government',
               'Researchers (HC) - Female - Higher education',
               'Researchers (HC) - Female - Private non-profit',
               'Researchers (HC) - Female - Not specified sectors']
        data_sectors = data[(data.Indicator.isin(sectors)) & (data.Time == 2015) & ~(data.Valu

In [6]: # sum headcounts per section across countries
        data_sector_agg = data_sectors.groupby(['Indicator'])['Indicator','Value'].sum().reset_

In [7]: # add sector name for cleaner plotting
        data_sector_agg['Sector'] = ['Business enterprise', 'Government', 'Higher education',

In [8]: # drop Not specified sector because so insignifcant compared to other sectors
        data_sector_agg = data_sector_agg.iloc[[0, 1, 2, 4],:]

        # round value
        data_sector_agg['Value'] = data_sector_agg.apply(lambda row: np.round(row['Value']), ax
```

In [11]:
```
x = data_sector_agg.Sector
y = data_sector_agg.Value

data_bar = [go.Bar(
            x=x,
            y=y,
            text=y,
            textposition = 'auto',
            marker=dict(
                color='rgb(255, 191, 0)',
                line=dict(
                    color='rgb(8,48,107)',
                    width=1.5),
            ),
            opacity=0.6
        )]

layout = go.Layout(
    title = go.layout.Title(
        text = 'Total Head Count of Female Researchers across Sectors (2015)'
    ),
    xaxis=dict(
        title='Sectors'
    ),
    yaxis=dict(
        title='Total Head Count'
    )
)
fig = go.Figure(data=data_bar, layout=layout)

py.iplot(fig, filename='Female Researchers HC by Sector')
```

Out[11]:
```
<plotly.tools.PlotlyDisplay object>
```

In [ ]:

boxplot

May 12, 2019

1 Boxplot

Boxplot of female researchers by seniority

```
In [1]: import pandas as pd
        import numpy as np
        import plotly.plotly as py
        import plotly.graph_objs as go
        from collections import Counter

In [2]: data = pd.read_csv('../data/seniority_researchers.csv')

In [5]: # filtering by seniority groups
        seniority = ['Researchers (HC) - Female - Category A ','Researchers (HC) - Female - Ca
         'Researchers (HC) - Female - Category C','Researchers (HC) - Female - Category D',
         'Researchers (HC) - Female - Not specified seniority levels']

        # only years with the data
        data_seniority_A = data[(data.Indicator == 'Researchers (HC) - Female - Category A ') &
        data_seniority_B = data[(data.Indicator == 'Researchers (HC) - Female - Category B ') &
        data_seniority_C = data[(data.Indicator == 'Researchers (HC) - Female - Category C') &
        data_seniority_D = data[(data.Indicator == 'Researchers (HC) - Female - Category D') &

In [6]: # find the sum to normalize the values
        group_a_sum = dict(data_seniority_A.groupby(['Time'])['Value'].sum())
        group_b_sum = dict(data_seniority_B.groupby(['Time'])['Value'].sum())
        group_c_sum = dict(data_seniority_C.groupby(['Time'])['Value'].sum())
        group_d_sum = dict(data_seniority_D.groupby(['Time'])['Value'].sum())

In [7]: def normalize_counts(row, sum_dictionary):
            """ Normalize head counts by year and seniority category """
            if row['Time'] == 2015:
                value = row['Value']/sum_dictionary[2015]
            elif row['Time'] == 2016:
                value = row['Value']/sum_dictionary[2016]
            elif row['Time'] == 2017:
                value = row['Value']/sum_dictionary[2017]
```

```
            else:
                value = np.nan
            return value
```

```
In [8]: # adding normalized values for boxplot
        data_seniority_A['normalized_val'] = data_seniority_A.apply(lambda row: np.round(normal
        data_seniority_B['normalized_val'] = data_seniority_B.apply(lambda row: np.round(normal
        data_seniority_C['normalized_val'] = data_seniority_C.apply(lambda row: np.round(normal
        data_seniority_D['normalized_val'] = data_seniority_D.apply(lambda row: np.round(normal
```

```
/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/ipykernel_launcher.py:2: Set


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html

/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/ipykernel_launcher.py:3: Set


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html

/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/ipykernel_launcher.py:4: Set


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html

/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/ipykernel_launcher.py:5: Set


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
```

```
In [10]: trace0 = go.Box(
             y=data_seniority_A.normalized_val,
             x=data_seniority_A.Time,
             name='Seniority Category A',
```

```
    marker=dict(
        color='#3D9970'
    )
)
trace1 = go.Box(
    y=data_seniority_B.normalized_val,
    x=data_seniority_B.Time,
    name='Seniority Category B',
    marker=dict(
        color='#FF4136'
    )
)
trace2 = go.Box(
    y=data_seniority_C.normalized_val,
    x=data_seniority_C.Time,
    name='Seniority Category C',
    marker=dict(
        color='#FF851B'
    )
)
trace3 = go.Box(
    y=data_seniority_D.normalized_val,
    x=data_seniority_D.Time,
    name='Seniority Category D',
    marker=dict(
        color='#4F86F7'
    )
)

data = [trace0, trace1, trace2, trace3]
layout = go.Layout(
    title = go.layout.Title(
        text = 'Normalized Head Counts of Female Researchers Across Seniority Levels
    ),
    xaxis=dict(
        title='Year',
        zeroline=False
    ),
    yaxis=dict(
        title='Normalized Head Counts',
        zeroline=False
    ),
    boxmode='group',
    annotations=[
        dict(
            x=1.11,
            y=1.05,
            align="left",
```

```
                    valign="top",
                    text='Seniority Category',
                    showarrow=False,
                    xref="paper",
                    yref="paper",
                    xanchor="center",
                    yanchor="top"
                )
            ]
        )
        fig = go.Figure(data=data, layout=layout)
        py.iplot(fig, filename='Normalized Head Counts of Female Researchers Across Seniority
```

Out[10]: <plotly.tools.PlotlyDisplay object>

In [ ]:

bubblemap_processing

May 12, 2019

1 Bubble Map

Bubble map of head counts of female R&D in Europe

```
In [1]: import pandas as pd
        import numpy as np
        import plotly.plotly as py
        import plotly.graph_objs as go

In [2]: data = pd.read_csv('../data/total_rd.csv')

In [3]: # filter data to include only European countries
        europe = ['Albania', 'Andorra', 'Austria', 'Belarus', 'Belgium', 'Bosnia and Herzegovi
                  'Channel Islands', 'Croatia', 'Czechia', 'Denmark', 'Estonia', 'Faeroe Islan
                  'Germany','Greece','Hungary','Iceland', 'Ireland','Isle of Man', 'Italy', 'La
                  'Lithuania', 'Luxembourg', 'Malta', 'Monaco', 'Montenegro', 'Netherlands', 'N
                  'Romania', 'Russian Federation','San Marino','Serbia','Slovakia','Slovenia',
                  'Switzerland', 'Ukraine', 'United Kingdom of Great Britain and Northern Irel
        data  = data[(data.Country.isin(europe)) & (data.Indicator == 'Total R&D personnel (HC

In [5]: data.sort_values('Value')

Out[5]:      INDICATOR                          Indicator LOCATION  \
        1674     21104  Total R&D personnel (HC) - Female      MLT
        1762     21104  Total R&D personnel (HC) - Female      MNE
        2871     21104  Total R&D personnel (HC) - Female      BIH
        1515     21104  Total R&D personnel (HC) - Female      LUX
        1491     21104  Total R&D personnel (HC) - Female      ISL
        1615     21104  Total R&D personnel (HC) - Female      EST
        1645     21104  Total R&D personnel (HC) - Female      LVA
        1722     21104  Total R&D personnel (HC) - Female      SVN
        1605     21104  Total R&D personnel (HC) - Female      HRV
        1652     21104  Total R&D personnel (HC) - Female      LTU
        1758     21104  Total R&D personnel (HC) - Female      SRB
        1540     21104  Total R&D personnel (HC) - Female      SVK
        1581     21104  Total R&D personnel (HC) - Female      BGR
        1495     21104  Total R&D personnel (HC) - Female      IRL
```

```
1710     21104  Total R&D personnel (HC) - Female      ROU
1489     21104  Total R&D personnel (HC) - Female      HUN
1473     21104  Total R&D personnel (HC) - Female      FIN
1525     21104  Total R&D personnel (HC) - Female      NOR
1464     21104  Total R&D personnel (HC) - Female      CZE
1468     21104  Total R&D personnel (HC) - Female      DNK
1456     21104  Total R&D personnel (HC) - Female      AUT
1459     21104  Total R&D personnel (HC) - Female      BEL
1484     21104  Total R&D personnel (HC) - Female      GRC
1550     21104  Total R&D personnel (HC) - Female      CHE
1535     21104  Total R&D personnel (HC) - Female      PRT
1520     21104  Total R&D personnel (HC) - Female      NLD
2829     21104  Total R&D personnel (HC) - Female      UKR
1548     21104  Total R&D personnel (HC) - Female      SWE
1530     21104  Total R&D personnel (HC) - Female      POL
1500     21104  Total R&D personnel (HC) - Female      ITA
1545     21104  Total R&D personnel (HC) - Female      ESP
1478     21104  Total R&D personnel (HC) - Female      FRA
1560     21104  Total R&D personnel (HC) - Female      GBR
1481     21104  Total R&D personnel (HC) - Female      DEU

                                                 Country  TIME  Time  \
1674                                               Malta  2015  2015
1762                                          Montenegro  2015  2015
2871                            Bosnia and Herzegovina  2015  2015
1515                                        Luxembourg  2015  2015
1491                                           Iceland  2015  2015
1615                                           Estonia  2015  2015
1645                                            Latvia  2015  2015
1722                                          Slovenia  2015  2015
1605                                           Croatia  2015  2015
1652                                         Lithuania  2015  2015
1758                                            Serbia  2015  2015
1540                                          Slovakia  2015  2015
1581                                          Bulgaria  2015  2015
1495                                           Ireland  2015  2015
1710                                           Romania  2015  2015
1489                                           Hungary  2015  2015
1473                                           Finland  2015  2015
1525                                            Norway  2015  2015
1464                                           Czechia  2015  2015
1468                                           Denmark  2015  2015
1456                                           Austria  2015  2015
1459                                           Belgium  2015  2015
1484                                            Greece  2015  2015
1550                                       Switzerland  2015  2015
1535                                          Portugal  2015  2015
1520                                       Netherlands  2015  2015
```

```
2829                                                Ukraine  2015  2015
1548                                                 Sweden  2015  2015
1530                                                 Poland  2015  2015
1500                                                  Italy  2015  2015
1545                                                  Spain  2015  2015
1478                                                 France  2015  2015
1560  United Kingdom of Great Britain and Northern I...  2015  2015
1481                                                Germany  2015  2015

           Value Flag Codes Flags
1674     755.000        NaN   NaN
1762    1194.000        NaN   NaN
2871    1425.000        NaN   NaN
1515    1706.000        NaN   NaN
1491    2476.000        NaN   NaN
1615    4698.000        NaN   NaN
1645    6025.000        NaN   NaN
1722    7282.000        NaN   NaN
1605    8789.000        NaN   NaN
1652   11447.000        NaN   NaN
1758   11695.000        NaN   NaN
1540   12242.000        NaN   NaN
1581   14167.000        NaN   NaN
1495   17148.000        NaN   NaN
1710   19913.000        NaN   NaN
1489   21918.000        NaN   NaN
1473   25390.000        NaN   NaN
1525   29888.000        NaN   NaN
1464   31121.710        NaN   NaN
1468   31740.000        NaN   NaN
1456   38084.000        NaN   NaN
1459   41298.000        NaN   NaN
1484   41606.000        NaN   NaN
1550   42222.821        NaN   NaN
1535   46367.000        NaN   NaN
1520   49689.000        NaN   NaN
2829   49782.000        NaN   NaN
1548   50638.000        NaN   NaN
1530   61613.000        NaN   NaN
1500  137339.000        NaN   NaN
1545  137822.000        NaN   NaN
1478  196026.340        NaN   NaN
1560  243246.000        NaN   NaN
1481  293103.000        NaN   NaN
```

In [ ]:
```
data.
```

In [ ]:
```
# data.to_csv('../data/europe_bubble.csv', index=False)
```

1.0.1 This CSV file can be fed into Tableau to create the final bubble map. One would need to specify the Values as the columns and Country as the rows before clicking on 'Show Me' to select the bubble map option.

chloropleth

May 12, 2019

1 Chloropleth map

Creating chloropleth map of total Female R&D personnel (in head count percentages) in 2015.

In [1]:
```
import pandas as pd
import numpy as np
import plotly.plotly as py
import plotly.graph_objs as go
```

In [2]:
```
data = pd.read_csv('../data/total_rd.csv')
```

In [5]:
```
# filter out rows that are Total R&D personnel (HC) - % Female
hc_female = data[data['Indicator'] == 'Total R&D personnel (HC) - % Female']
```

In [7]:
```
hc_female_2015 = hc_female[hc_female['Time'] == 2015]
```

In [8]:
```
hc_female_2015.head()
```

Out[8]:

	INDICATOR	Indicator	LOCATION	Country	TIME	Time	Value	Flag Codes	Flags
1765	FPERSP_THC	Total R&D personnel (HC) - % Female	AUT	Austria	2015	2015	30.18443	NaN	NaN
1768	FPERSP_THC	Total R&D personnel (HC) - % Female	BEL	Belgium	2015	2015	36.36156	NaN	NaN
1773	FPERSP_THC	Total R&D personnel (HC) - % Female	CZE	Czechia	2015	2015	31.08191	NaN	NaN
1777	FPERSP_THC	Total R&D personnel (HC) - % Female	DNK	Denmark	2015	2015	37.13062	NaN	NaN
1782	FPERSP_THC	Total R&D personnel (HC) - % Female	FIN	Finland	2015	2015	33.36838	NaN	NaN

In [9]:
```
world = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/master/2014_worl
```

In [10]:
```
addition = set(world['CODE']) - set(hc_female_2015['LOCATION'])
```

In [11]:
```
world = world[world['CODE'].isin(addition)]
```

In [12]:
```
world.drop(columns="GDP (BILLIONS)", inplace=True)
```

In [13]:
```
world['INDICATOR'] = np.nan
world['Indicator'] = np.nan
world['TIME'] = 2015
world['Time'] = 2015
world['Value'] = 0
world['Flag Codes'] = np.nan
world['Flags'] = np.nan
world.rename(columns={'CODE':'LOCATION'}, inplace=True)
world.rename(columns={'COUNTRY':'Country'}, inplace=True)
```

In [14]:
```
# add countries so that they populate on the map
hc_female_2015 = pd.concat([hc_female_2015, world])
```

```
/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/ipykernel_launcher.py:2: Fu

Sorting because non-concatenation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.
```

In [16]:
```
fem_map = [go.Choropleth(
            locations = hc_female_2015['LOCATION'],
            z = hc_female_2015['Value'],
            text = hc_female_2015['Country'],
            colorscale = [
                [0, "rgb(5, 10, 172)"],
                [0.35, "rgb(40, 60, 190)"],
                [0.5, "rgb(70, 100, 245)"],
                [0.6, "rgb(90, 120, 245)"],
                [0.7, "rgb(106, 137, 247)"],
                [1, "rgb(220, 220, 220)"]
            ],
            autocolorscale = False,
            reversescale = True,
            marker = go.choropleth.Marker(
                line = go.choropleth.marker.Line(
                    color = 'rgb(180,180,180)',
                    width = 0.5
                )),
            colorbar = go.choropleth.ColorBar(
                tickprefix = '%',
                title = 'Total Female R&D personnel <br> (Head count %)'),
        )]
```

```
        layout = go.Layout(
            title = go.layout.Title(
                text = 'Head Count Percentages of Female R&D personnel (2015)'
            ),
            geo = go.layout.Geo(
                showframe = False,
                showcoastlines = False,
                projection = go.layout.geo.Projection(
                    type = 'equirectangular'
                )
            ),
            annotations = [go.layout.Annotation(
                x = 0.55,
                y = 0.1,
                xref = 'paper',
                yref = 'paper',
                text = 'Source: <a href="http://data.uis.unesco.org/">\
                    UNESCO Institute for Statistics</a>',
                showarrow = False
            )]
        )

        fig = go.Figure(data = fem_map, layout = layout)
        py.iplot(fig, filename = 'd3-world-map')
```

```
/Users/Nina/bin/anaconda3/envs/msds622/lib/python3.7/site-packages/IPython/core/display.py:689

Consider using IPython.display.IFrame instead
```

Out[16]: <plotly.tools.PlotlyDisplay object>

connected_graph_processing

May 12, 2019

1 Connected Graph Processing

```
In [1]: import pandas as pd
        import numpy as np

In [13]: data = pd.read_csv('../data/un_info_centers.csv')

In [16]: from geopy.geocoders import Nominatim
         latitude = []
         longitude = []
         geolocator = Nominatim(user_agent="find_location")

         for loc in data['Office'].values:
             location = geolocator.geocode(loc)
             latitude.append(location.latitude)
             longitude.append(location.longitude)

In [19]: data['latitude'] = latitude
         data['longitude'] = longitude

In [21]: data.to_csv('../data/un_office_complete.csv', index=False)
```

heatmap

May 12, 2019

1 Heat map

Heat map of mean amount of female researchers across different qualifications over the last couple of years

```
In [1]: import pandas as pd
        import numpy as np
        import plotly.plotly as py
        import plotly.graph_objs as go

In [2]: data = pd.read_csv('../data/qualification_researchers.csv')

In [5]: qualifications = ['Researchers (HC) - Female - ISCED 8 ', 'Researchers (HC) - Female -
                          'Researchers (HC) - Female - ISCED 6', 'Researchers (HC) - Female - IS
        data = data[(data.Indicator.isin(qualifications)) & ~(data.Value.isna())]

In [11]: # get summed Value information per year and indicator
         data_agg = data.groupby(['Time', 'Indicator']).mean().reset_index()[['Indicator','Time

In [23]: data_agg_5 = data_agg[data_agg.Indicator == 'Researchers (HC) - Female - ISCED 5'].Val
         data_agg_6 = data_agg[data_agg.Indicator == 'Researchers (HC) - Female - ISCED 6'].Val
         data_agg_7 = data_agg[data_agg.Indicator == 'Researchers (HC) - Female - ISCED 7 '].Va
         data_agg_8 = data_agg[data_agg.Indicator == 'Researchers (HC) - Female - ISCED 8 '].Va

In [50]: trace = go.Heatmap(z=[data_agg_5, data_agg_6, data_agg_7, data_agg_8],
                            x=['2011', '2012', '2013', '2014', '2015', '2016', '2017'],
                            y=['ISCED 5','ISCED 6','ISCED 7 ','ISCED 8 '])


         layout = go.Layout(
             title='Average Head Count of Female Researchers Across Qualifications (2011-2017)
             xaxis = dict(title='Year'),
             yaxis = dict(title='Qualification'), annotations=[
                 dict(
                     x=1.07,
                     y=1.10,
                     align = "left",
                     valign="top",
```

```
                    text='Average Head <br> Count',
                    showarrow=False,
                    xref="paper",
                    yref="paper",
                    xanchor="center",
                    yanchor="top"
                )
            ])

        fig = go.Figure(data=[trace], layout=layout)

        py.iplot(fig, filename='researchers_qualifications_heatmap')
```

Out[50]: <plotly.tools.PlotlyDisplay object>

In [ ]:

In [ ]:

histogram

May 12, 2019

1 Histogram

Creating a histogram of the total head counts of female researchers in 2015.

In [1]:
```
import pandas as pd
import numpy as np
import plotly.plotly as py
import plotly.graph_objs as go
from collections import Counter
```

In [2]:
```
# reading number of female researchers
data = pd.read_csv('../data/total_researchers.csv')
```

In [4]:
```
# filter data
# Time: 2015
# I
data_2015 = data[(data['Time'] == 2015) & (data['Indicator'] == 'Researchers (HC) - To
```

In [8]:
```
data_hist = [go.Histogram(x=data_2015['Value'],marker=dict(color='purple'), opacity=0.

layout = go.Layout(
    title = go.layout.Title(
        text = 'Total Head Count of Female Researchers (2015)'
    ),
    xaxis=dict(
        title='Head Count Value'
    ),
    yaxis=dict(
        title='Count'
    )
)
fig = go.Figure(data=data_hist, layout=layout)

py.iplot(fig, filename='Histogram: 2015 Female Researchers (HC)')
```

Out[8]:
```
<plotly.tools.PlotlyDisplay object>
```

In [ ]:

stacked_area_graph

May 12, 2019

1 Stacked Area Graph

Stacked Area Graph of headcounts of female R&D personnel, researchers, and technologists in Thailand across the last several years.

In [1]:
```
import pandas as pd
import numpy as np
import plotly.plotly as py
import plotly.graph_objs as go
```

In [2]:
```
rnd = pd.read_csv('../data/total_rd.csv')
research = pd.read_csv('../data/total_researchers.csv')
tech = pd.read_csv('../data/total_technicians.csv')
```

In [3]:
```
# rnd_thai = rnd[(rnd.Indicator == 'Total R&D personnel (HC) - % Female') & (rnd.Count
rnd_thai = rnd[(rnd.Indicator == 'Total R&D personnel (HC) - Female') & (rnd.Country ==
```

In [5]:
```
# research_thai = research[(research.Indicator == 'Researchers (HC) - % Female') & (re
research_thai = research[(research.Indicator == 'Researchers (HC) - Female') & (researc


# exclude 2014 to match the other data sources

research_thai = research_thai[research_thai['Time'] != 2014]
```

In [7]:
```
# tech_thai = tech[(tech.Indicator == 'Technicians (HC) - % Female') & (tech.Country ==
tech_thai = tech[(tech.Indicator == 'Technicians (HC) - Female') & (tech.Country == 'T
```

In [51]:
```
# Add original data
trace1 = go.Scatter(name = 'R&D Personnel',
    x=rnd_thai.Time,
    y=rnd_thai.Value,
    fill='tozeroy')
trace2 = go.Scatter(name = 'Researchers',
    x=research_thai.Time,
    y=research_thai.Value,
    fill='tonexty')
```

```
trace3 = go.Scatter(name = 'Technicians',
    x=tech_thai.Time,
    y=tech_thai.Value,
    fill='tonextx')

layout = go.Layout(
    title='Head Count of Female STEM Employees Across Occupations (2011-2016)',
    xaxis = dict(title='Year'),
    yaxis = dict(title='Head Count'),
    annotations=[
        dict(
            x=1.07,
            y=1.05,
            align = "left",
            valign="top",
            text='Occupations',
            showarrow=False,
            xref="paper",
            yref="paper",
            xanchor="center",
            yanchor="top"
        ),
        dict(
            x = 2011,
            y = 50000,
            xref = 'x',
            yref = 'y',
            text = 'Net enrollment in secondary <br> education increases from <br> 31
            showarrow=True,
            arrowhead = 7,
            ax=0,
            ay=-60
            ),
        dict(
            x = 2012,
            y = 50000,
            xref = 'x',
            yref = 'y',
            text = '2.5 million+ students enrolled <br> in higher education institutio
            showarrow=True,
            arrowhead = 7,
            ax=0,
            ay=-150
            ),
        dict(
            x = 2015,
            y = 57000,
            xref = 'x',
```

```
                    yref = 'y',
                    text = '2.14 million students enrolled <br> at a beachelors level or belo
                    showarrow=True,
                    arrowhead = 7,
                    ax=-70,
                    ay=-50
                    ),
                dict(
                    x = 2016,
                    y = 88000,
                    xref = 'x',
                    yref = 'y',
                    text = 'Launch of Thailand 4.0',
                    showarrow=True,
                    arrowhead = 7,
                    ax=-70,
                    ay=-50
                    )
            ])

        data = [trace1, trace2, trace3]

        fig = go.Figure(data=data, layout=layout)
```

In [52]:
```
py.iplot(fig, filename='stacked-area-thailand')
```

Out[52]:
```
<plotly.tools.PlotlyDisplay object>
```

In [ ]:

In [ ]:

treemapping

May 12, 2019

1 Treemap

Summed female researcher headcounts across different age groups in 2015

```
In [1]: import pandas as pd
        import numpy as np
        import plotly.plotly as py
        import plotly.graph_objs as go
        import squarify

In [2]: data = pd.read_csv('../data/age_researchers.csv')

In [5]: data = data[(data.Indicator.isin(['Researchers (HC) - Female - 65 years and more',
               'Researchers (HC) - Female - 55-64 years','Researchers (HC) - Female - 25-34 yea
                                          'Researchers (HC) - Female - Under 25 years ','Researc
                                         'Researchers (HC) - Female - Not specified age groups

In [6]: # sum the head counts per age group in 2015 for all countries
        data_agg = data.groupby('Indicator').sum().reset_index()[['Indicator', 'Value']]

In [8]: data_agg.to_csv('../data/treemap_data.csv', index=False)

In [ ]:

In [ ]:
```