# Using Data Science to Explore Latent Cognitive Biases in Employer Compensation

VIRTEE PAREKH, Rochester Institute of Technology

**Abstract**

In data science projects, analysts usually have to work on data that is outside their field of domain. In this independent study, we apply techniques from data science to the domain of employer compensation, using standard and some novel techniques. This project is a study of cognitive biases that affect an employees' compensation. Exploratory data analysis is performed on salary information obtained from the datasets available on the United Kingdom and the United States government websites. Additional data is gathered by scraping websites providing relevant information and Twitter tweets are taken to perform sentiment analysis. We explore how factors like location of the company, composition of board of directors, size of the company, occupation etc. affect the compensation of both the genders. We explore Twitter tweets by performing sentiment analysis by using VADER, to observe what the social media has to say about the issue. Various reports and statistics regarding the wage gap issue are studied. Tools used to analyze and explore this data include Python, Plotly, R, Tableau, Seaborn etc.



DOI: 0000001.0000001

**ACM Reference format:**

Virtee Parekh. 2016. Using Data Science to Explore Latent Cognitive Biases in Employer Compensation. 1, 1, Article 1 (January 2016), 54 pages.

DOI: 0000001.0000001

## 1 INTRODUCTION

There has been persistent debate and discussion about the disparity in employer compensation of both the genders that are working in the same field. Due to gender discrimination and lack of equal representation of both the genders in the same field, the disparity has persisted. With increasing advancement in both technology and the society, the gender gap is decreasing. Unfortunately, the decrease in wage disparity is slow. According to The American Association of University Women (AAUW), a group working for equality and education for females, a women's median earnings are 80% of their male counterparts median earnings [15]. Asian women earn 85% of a white male's earnings and Hispanic or Latin women earn a mere 53% of a white male's earnings [15]. There are various factors that affect the compensation. Certain employers have a cognitive preference towards men and believe that women are not equally capable and deserving of the same pay. Racial discrimination might also occur in some places. Factors like location of the organization, sector of the organization, presence of women in the company's leadership team also act as biases while deciding employee compensation. The type of organization - whether it is a charitable organization, a private sector or a government job also decides the outcome. Certain jobs are male-dominated and unequal representation of women are also a hindering factor. Unequal education opportunities of women also contribute heavily to this issue. In this project, we explore the pay gap data provided by the governments of United Kingdom and United States. We see how factors like location, size of

the company, sector of the company, occupation, type of organization, composition of leadership team etc. affect the pay gap.

Section 2 lists out the goals of this independent study. In section 3, we study case studies and other reports that talk about the prevailing gender pay gap issue and explore the statistics. In the following section 4, we list out the various data sources used and how to access the data. From section 5 to 9, we perform exploratory data analysis on the data by using different visualization tools and techniques. In section 10, we perform sentiment analysis in gender pay gap related tweets. A short writeup on other concepts learned during the independent study is mentioned in section 11. We conclude by presenting the major insights gathered from the report and the conclusion.

## 2 GOALS OF THE INDEPENDENT STUDY

The goal of this independent study is to learn and master different data science skills that aid the student in becoming a better data scientist and were not fully studied till now. The following skills are expected to be mastered during the course of this study -

- Learn how to access data from the web - by scraping the website and by using the website's official APIs.
- Sentiment analysis of tweets in Python and study of VADER in detail and generation and use of word clouds.
- Study and implementation of different visualization charts and tools and to be able to use the visualizations for effective exploratory data analysis. The tools to be covered are Seaborn, Plotly, ggplot2 and Tableau.

## 3 LITERATURE SURVEY

In a world where mankind is making rapid technological advancements like building self-driving cars, exploring Mars and creating humanoids, it is unfortunate to see how little progress we have made in a humanitarian setting. According to the United Nations Development Programme [7], the global gender wage gap is 23% and will take nearly 100 years to balance the gap at the current rate.

The World Economic Forum(WEF) reports its findings on the global gender pay gap annually and in 2017 it reported that the progress towards balancing the employer compensation has stalled [34]. The WEF assigns each country with a pay gap score, where a score of 1 implies a total pay gap parity and 0 implies absolute disparity. It considers four criteria while assigning this score - gaps between the genders on salaries and participation, gaps in educational levels of both the genders, differences in women and men's health and the ratio of women to men in political positions. In 2017, WEF analyzed 144 countries for its report. The report [34] found that the wage gap had increased from 31.7% in 2016 to 32% in 2017. Iceland ranked highest with a score of 0.88 and Yemen ranked lowest with 0.52 The United States stands 49th with a score of 0.72 and the United Kingdom stood 15th with a score of 0.77. A subset of the report is shown in Table 1. Among regions of the world, Western Europe topped as the region with the lowest pay gap while the Middle East and North African region came last with the highest pay gap [34].

Table 1. 2017 Global Gender Gap Rankings

| Country | Rank | Score |
|---|---|---|
| Iceland | 1 | 0.878 |
| Switzerland | 21 | 0.755 |
| United Kingdom | 15 | 0.770 |
| Australia | 35 | 0.731 |
| United States | 49 | 0.718 |
| China | 100 | 0.674 |
| India | 108 | 0.669 |
| Japan | 114 | 0.657 |
| Yemen | 144 | 0.516 |

According to the US Census Bureau, the disparity among female and male employers was about 20% for 2018. Tampa, Florida has the narrowest pay gap and Seattle, Washington has the highest pay gap [46].

The American Association of University Women is a non-profit organization that works for the empowerment of women. They analyzed data obtained from the US Census Bureau, Bureau of Labor Statistics and American Community Survey for 2016. These agencies carry out massive surveys across the country and obtain data that consists of income details about on individual, household and business level [45]. They observed that the pay gap was maximum among Asian communities and minimum among African-American communities. However, when the salaries of all women were compared with white men, Asian women earned 87% of their white male counterparts and Hispanic women earned a meager 54% of their white male counterparts. They also discovered that women below 55 years of age faced a gap of 19-22% but those above 55 years of age faced a staggering gap of 26% [45].

Understanding the cause and nature of this inequality is a multifaceted job. One of the reasons that this wage gap can be attributed to is the fact that women and men are almost never equally represented in any occupation. Tasks involving intensive labor like mining, transportation, construction etc. are dominated by males and administrative or clerical, education, health care jobs hire more women [32]. Jobs dominated by women are usually pay less than the jobs dominated by men [26]. Even when the integration of the genders in the workplace takes place, it does not eliminate the gap. According to a research by New Scientist, women in STEM fields get paid 20% less than their male counterparts in the UK [33]. Another significant contributor to the pay gap is the ”motherhood penalty” [24] which stands for the undue repercussions a woman faces after joining work after childbirth. Taking maternity leaves is damaging to a woman’s career and employers are reluctant in hiring women with children. Employers are prejudiced in thinking that women with

children perform less ably than childless women or other men because having a child distracts a woman from focusing on her job exclusively. Research has shown that mothers earn about 7% less than childless women [31]. However, men enjoy what is termed as the ”fatherhood bonus” [24] where their salaries see an increase after having a child, as a man with a family is seen to be a stable asset. The gap can also be attributed to the fact that maybe men are better at negotiating salaries than women are [31]. Men can confidently ask for more than what they are being offered and bargain more effectively. Glassdoor's (a website that provides ratings and reviews for organizations work culture, environment, interview process and a job search portal) chief economist says that about 1/3 rd of the gap is because of gender discrimination at the workplace [27]. Some employers still believe that men outperform women and women do not deserve equal pay. They underplay a woman's performance by denying them their equal pay.

Closing the gender pay gap has beneficial effects on the country's economy. An International Monetary Fund study says that closing the gender pay gap would increase USA's GDP by 5%, Japan's by 9%, UAE's by 12% and India's by a whopping 27%. For every 0.1 decreases in the UN Gender Inequality Index, an index that deals with income inequality among genders, there is a 1% increase in the country's economic development [41]. Closing the gender pay gap is also advantageous for women that run a household themselves or help in running one. An Institute of Women's Policy Research study [39] shows that if the wage parity was achieved, the poverty rate of working women would cut back from 8.1% to 3.9% and that for single working mothers would drastically reduce to 15% from 28.7%. With a reduction in poverty, the US economy would generate an additional $500 billion that is about 3% of the 2012 US GDP [39]. Eliminating the pay gap also implies equal representation of both the genders in every occupation. The World Economic Forum studies have shown that companies that have more women in the top positions show an surge in net profit, more diverse views while problem-solving, better equity returns and payout ratios [50].

Laws have been in place to ensure that this disparity is eliminated. The Equal Pay Act [28] was passed in 1963 in the USA and it forbids discrimination on the grounds of gender for employer compensation. Consequently, in 1970, the United Kingdom also passed a similar Equal Pay Act. In January 2018, Iceland passed a strict law that makes it illegal for a company with more than 25 full-time employees to give unequal pay to men and women [38]. Companies must obtain government clearance that their salaries are based on skill set and education and not gender of the employees. Without this clearance, the companies may face a heavy fine. Iceland also has the world's smallest wage gap of just 8%. Despite having a law in the US, the gender wage gap still persists. This is due to the under-representation of women in male-dominated jobs. In Iceland, the parliament consists of nearly 50% women, making it easier for such laws to be passed [38]. The US Congress has less than 20% women. Also, the fine for violating the act in the US is not heavy and contains numerous loopholes [25]. Some states in the US like Alabama and Mississippi have no laws for equal pay. Strong equal pay laws exist in only a handful of 8 states like California, Oregon, Massachusetts, Illinois etc. In 2018, the United Kingdom has made it mandatory for organizations with more than 250 employees to submit their employee's salary data and gender pay details to the government [29]. Governments across the world, both federal and state must come up with more stringent equal pay laws. Legal action must be taken against offenders along with a heavy penalty.

On a company level, organizations must keep a close eye on their audits and ensure fair play. Having a good representation of both men and women at all levels is also a great start. Work schedules must be made flexible to allow for maternity leaves. Women on maternity leaves must be provided with better compensation. Their salaries post motherhood must not have a detrimental effect. Companies should not ask for any of their employee's salary history to avoid setting a baseline for their future salaries. Individually, women must negotiate their salaries and know what their skill set is worth [51]. Having a confident, calm composure and knowing your worth can help

go a long way in negotiating salaries. In the case of discrimination, the case should be reported to higher authorities or the police.

The next Equal Pay Day is on April 2, 2019. It symbolizes how far in the current year a woman must work to earn the same amount that men earned in the previous year [47]. It also helps in raising awareness about how wide the pay gap is.

Several companies have taken conscious efforts in trying to close the pay gap. In 2015, Salesforce CEO Mark Benioff examined all his employees' salaries and spent close to $3 million in bridging the gap. Tech giants like Apple, Intel and Adobe also declared that they have no gender pay gap as of 2018. Inarguably one of the most popular coffee chains, Starbucks also announced in 2018 that they had closed their gender pay gap [53].

## 4 DATA

### 4.1 Main Dataset

The government of United Kingdom made it mandatory for organizations with greater than 250 employees to report their gender pay gap. The data is publicly available on the website [14]. The organizations had to report the following details[5]:

- Mean gender pay gap
- Median gender pay gap
- Mean bonus gender pay gap
- Median bonus gender pay gap
- The proportion of men in the organization receiving a bonus payment
- The proportion of women in the organization receiving a bonus payment
- The proportion of men and women in each quartile pay band
- A statement on the website stating that the information is true

- An individual responsible for the statement
- Employer Size

The dataset also consists of the name of the organization, address, the date of the gender pay gap report submission and if the submission was after the deadline or not.

I added a target variable to determine which gender was paid more or less. If the median gender pay gap was within -5% to 5%, the record was labeled as equal pay. For median pay gaps greater than 5%, it implied that men were paid more and for median pay gaps less than -5%, it implied that women were paid more.

### 4.2 UK Charities Data

The data about organization types and charity data was taken from David Kane's blog [42]. He has modified the dataset to include attributes like organization type (whether the organization type is company, charity or a public sector), charity income and charity numbers if applicable, organization sub types (schools, universities, fire and police etc.) for public sector organizations and the female workforce in each organization. A subset of the data is shown below in Figure ??. 'CharityIncome' refers to the income of the charity, 'OrgType' refers to the organizational type - charity, company or public sector, 'OrgSubType' refers to the sub-type of the public sectors organizations - Fire and Police, NHS, Local Authority, Schools and Colleges, Universities and Building Societies, 'FemaleWorkforce' refers to the amount of female employees on the organization.

### 4.3 Sector Information

The sector information consists of which sector the organization belongs to. Examples of sectors include accommodation and food services, administrative and support service activities, arts entertainment and recreation, manufacturing, agriculture forestry and fishing, human health

| | EmployerName | DiffMeanHourlyPercent | CharityNumber | CharityIncome | OrgType | OrgSubType | FemaleWorkforce |
|---|---|---|---|---|---|---|---|
| 85 | ABICARE SERVICES LIMITED | 10.3 | NaN | NaN | Company | NaN | 87.950 |
| 86 | Abingdon & Witney College | 7.5 | NaN | NaN | Public Sector | Schools and Colleges | 73.000 |
| 87 | ABINGDON FLOORING LIMITED | 11.3 | NaN | NaN | Company | NaN | 15.450 |
| 88 | Abingdon School | 23.0 | 1071298.0 | 25101000.0 | Charity | NaN | 54.875 |
| 89 | ABM FACILITY SERVICES SCOTLAND LIMITED | 3.0 | NaN | NaN | Company | NaN | 43.750 |
| 90 | ABM FACILITY SERVICES UK LIMITED | 5.5 | NaN | NaN | Company | NaN | 38.500 |

Fig. 1. Organization Type Information for UK companies

and social work activities, wholesale and retail trade, transportation and storage, professional scientific and technical activities, education, financial and insurance activities, real estate, mining and quarrying etc. These details are available on the dataset website [14]. I wrote a Python script to scrape each organization's data. This sector information was added to the original dataset. Each organization can belong to multiple sectors. For the sake of simplicity, I reduced the sectors to contain one value of the sector. A snapshot of a subset of the dataset is shown below.

| | EmployerName | Sector | SingleSector |
|---|---|---|---|
| 49 | A & P FALMOUTH LIMITED | Manufacturing | Manufacturing |
| 50 | A & S Restaurants Ltd | Accommodation and food service activities | Accommodation and food service activities |
| 51 | A E J MANAGEMENT LIMITED | Real estate activities, Administrative and sup... | Real estate activities |
| 52 | A GOMEZ LIMITED | Wholesale and retail trade; repair of motor ve... | Wholesale and retail trade |
| 53 | A J W AVIATION LTD | Transportation and storage | Transportation and storage |

Fig. 2. Sector Information for UK companies

### 4.4 Geographical Data

For the plotting of geographical maps explained in later sections, the latitude and longitude data are publicly available. The post codes were extracted from the address field of the original dataset. The area shape file for the choropleth map was taken from here [11].

### 4.5 Companies House API

The Companies House [3] website stores information about companies and makes it available to the public. Using the Companies House API, I extracted information about each companies active officers. These include directors, secretaries and other persons of significant control. I accessed their API to fetch names of these officers, per organization. A limitation to this API was that they allow only 600 requests within a 5 minute period. The Python script had to be modified so as to not return a 429 Too Many Requests HTTP error. Once I had a list of all active officers per company, I referred 2 datasets from data.world that provided me with name to gender mapping [22] [40]. Once I had access to all the genders of the officers, I calculated the majority gender in the directors group. A snapshot of a subset of the data is shown below in Figure 3. 'EmployerName' is the name of the organization. 'DiffMedianHourlyPercent' is the difference in median hourly wages between men and women, 'OrgType' is the type of organization, 'FemaleDir_Pct' is the percentage of female directors, 'MaleDir_Pct' is the percentage of male directors, 'Target' refers to which gender gets paid more or equal pay exists and 'Dir_Majority_Gender' refers to the majority gender composition in the directors group.

| | EmployerName | DiffMedianHourlyPercent | OrgType | FemaleDir_Pct | MaleDir_Pct | Target | Dir_Majority_Gender |
|---|---|---|---|---|---|---|---|
| 0 | "Bryanston School",Incorporated | 28.2 | Charity | 44.444444 | 55.555556 | Men | Male |
| 1 | 118 LIMITED | 2.8 | Company | 0.000000 | 100.000000 | EqualPay | Male |
| 2 | 123 EMPLOYEES LTD | 36.0 | NaN | 0.000000 | 100.000000 | Men | Male |
| 3 | 1610 LIMITED | -34.0 | Charity | 25.000000 | 75.000000 | Women | Male |
| 4 | 1879 EVENTS MANAGEMENT LIMITED | 8.1 | Company | 33.333333 | 66.666667 | Men | Male |

Fig. 3. Snapshot of UK data with Directors Information

To scrape the data from the Companies House API, it is necessary to create a account on their website and get access to a API key. This key grants the developer access to the data. We fetch

all the names of active officers and dump them into a csv file. The code for fetching this data is mentioned below. The code is taken from [49].

```
import requests

API_KEY = 'ff'  #get api key
BASE_URL = 'https://api.companieshouse.gov.uk/company/'

#get company name
def get_company_name(company_num):
   #send a HTTP request to fetch company name about company specified by company_num
    response = requests.get(BASE_URL + company_num, auth=(API_KEY, ''))

    #if HTTP request is successful
    if response.status_code == requests.codes.ok:
        data = response.json()
        company_name = data["company_name"]
        return company_name
    else:
        return 'No name listed'

#get officers information
def get_officers_info(company_num):
   #send a HTTP request to fetch officers info. about company specified by company_num
   response = requests.get(BASE_URL + company_num + '/officers'  , auth=(API_KEY, ''))

    #if HTTP request is successful
    if response.status_code == requests.codes.ok:
        return response.json()
    else:
        return 'CRN not listed with Companies House'


#get company name based on company number
company_name = get_comp_name(company_num)

#get officers information based on company number
officers_data = get_officers_info(company_num)

#write data to csv
```

## 5 VISUALIZATION USING CHOROPLETH MAPS

A choropleth map is a map with predefined regions shaded or colored according to the proportion of the variable that is observed in the map. One major goal of performing exploratory data analysis was to see how the gender pay gap varied across the country of the United Kingdom. Did urban areas fare better or worse than rural areas? Is there a particular pattern that is observed across the country? Why would certain regions fare much worse or better than others? To explore these areas, I developed a choropleth map.

In this scenario, we want to observe the variations in the gender pay gap across the country. The first map in Figure 4 shows areas where women are paid more than men. Darker the green shade, higher are the chances of women being paid more than men. The numbers on the colorbar scale represent the number of companies in the area that pay women more. To understand the patterns in the country I plotted some of the most populous cities as well as some rural towns on the map.

In the map, we notice that the urban areas of London, Reading, Birmingham, Nottingham, Liverpool, and Newcastle upon Tyne show the darkest shades of green. Intuitively, we can see that discrimination based on gender is less prevalent in urban areas due to modernization and appropriate education practices that encourage the empowerment of women. Areas like Oxford and Cambridge, which are mainly college towns show a relatively less light shade of green, indicating that fewer companies that pay women more. Rural areas like Plymouth, Swansea, Shrewsbury are extremely lighter in color. This can be attributed to the fact that there might not be equal opportunities for women in these areas or lack of awareness might have caused the lighter shade.

In 5, a choropleth map depicting areas with equal pay are shown. Again, we see that urban areas of the United Kingdom show more organizations that have equal pay rather than the rural ones. The area around London, Newcastle upon Twyne, Nottingham, Birmingham, and Manchester are said to be the most urban areas [48].

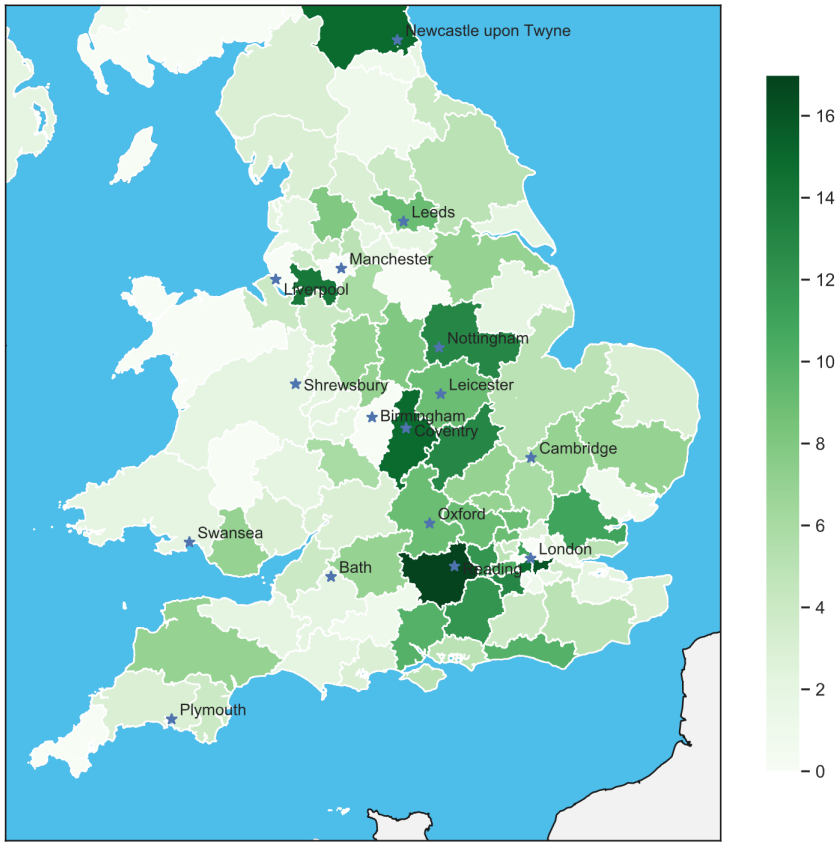


Fig. 4. Map showing regions where women are paid more, represented by darker shades of green

This implies that a woman looking for equal pay opportunities should look for jobs in the urban areas of the United Kingdom.

The visualization was created by using the Basemap [9] library in Python. The input dataframe to Basemap consists of the number of companies, per area, in which women are paid more. The central latitude and longitude of the area are also included as columns. Basemap can be initialized by specifying the resolution (crude, low, intermediate, high, full) of the map, the projection of the map, the latitude and longitude of the center of the region that is displayed, latitude and longitude of the lower left hand corner of the region and latitude and longitude of the upper right hand corner of the region. Then the color of the water bodies is set on the map, followed by color of the land

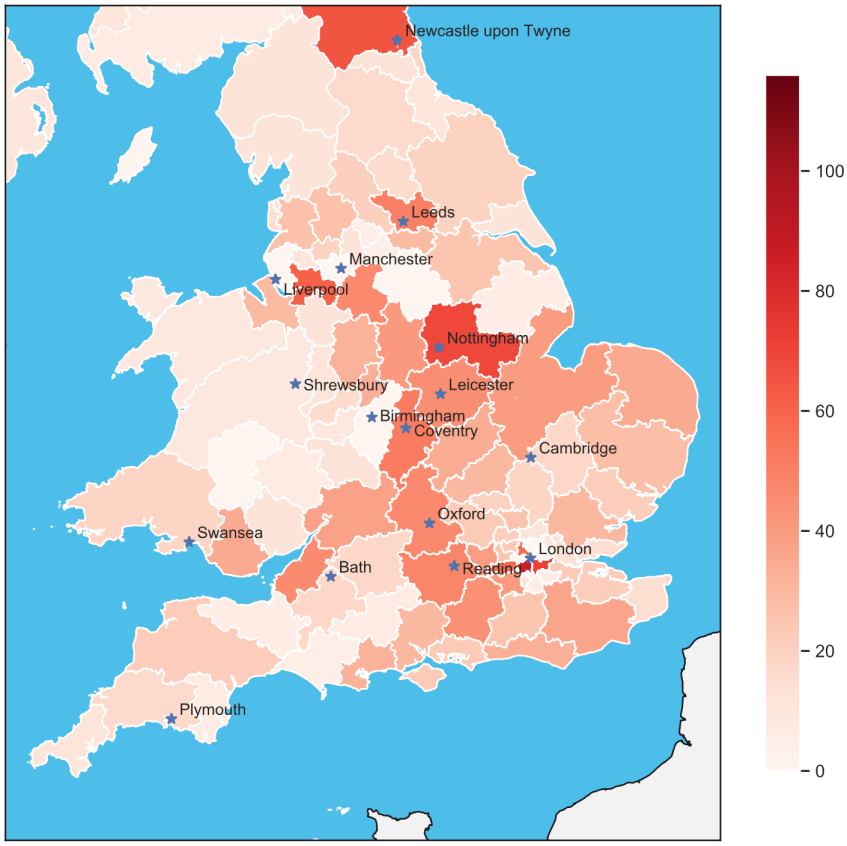


Fig. 5. UK Map for Equal Pay - Darker the shade of red, higher the chances of equal pay

masses. Lines are then drawn around the land masses. Shape files are files that contain information about the regions that are to be marked in a country. The file for UK is taken and added to the Basemap. A colormap is added to the Basemap. The colormap is responsible for coloring the regions of UK according to the number of regions where women stand a higher chance of being paid more. More companies where a women is paid more, darker is the color of that region. A colorbar is also added as an indicator of the above. The code [4] to generate a choropleth map is given below.

```
# code to generate a choropleth map

#plotting
import matplotlib.pyplot as plt
```

```
#drawing the map
from mpl_toolkits.basemap import Basemap
#heat map
from matplotlib.patches import Polygon
from matplotlib.collections import PatchCollection
from matplotlib.colors import Normalize
# colour map
import matplotlib.cm

fig, ax = plt.subplots(figsize=(10,20))

#defining the base map
m = Basemap(resolution='h', # c, l, i, h, f or None
            projection='merc',
            lat_0=54.5, lon_0=-4.36,
            llcrnrlon=-6., llcrnrlat= 49.5,urcrnrlon=2., urcrnrlat=55.2)

#set the colour of the seas and oceans on our map
m.drawmapboundary(fill_color='#46bcec')

#colour of land masses
m.fillcontinents(color='#f2f2f2',lake_color='#46bcec')

#draws lines around the land masses
m.drawcoastlines()

#shapefile for regions
m.readshapefile('Areas', 'areas')

df_poly = pd.DataFrame({
        'shapes': [Polygon(np.array(shape), True) for shape in m.areas],
        'area': [area['name'] for area in m.areas_info]
    })
df_poly = df_poly.merge(input_data, on='area', how='left')

#add colormap
cmap = plt.get_cmap('Greens')
pc = PatchCollection(df_poly.shapes, zorder=2)
norm = Normalize()
pc.set_facecolor(cmap(norm(df_poly['count'].fillna(0).values)))
ax.add_collection(pc)

#add a colorbar
mapper = matplotlib.cm.ScalarMappable(norm=norm, cmap=cmap)
mapper.set_array(df_poly['count'])
plt.colorbar(mapper, shrink=0.4)
```

## 6 EXPLORATORY DATA ANALYSIS USING SEABORN LIBRARY

For my next set of visualizations, I chose to use the Seaborn library in Python. Based on the Matplotlib library of Python, it provides a high-level interface of drawing informative statistical plots [13].

Visualizations for this dataset are restricted due to the fact that all the factors that are used to evaluate the gender pay gap are categorical. Visualizations where the features are categorical reduces the number of plots that can be generated. In this section, I have explored the dataset with the help of Seaborn library and found these to be the most insightful graphs.

### 6.1 Box and Whisker Plots

A box and whisker plot is a way of illustrating the distribution of data by mentioning five key statistics about the data - minimum, first quartile, median, third quartile, and the maximum. The median is shown by the horizontal line inside the box. The two horizontal edges of the box represent the first and third quartile (the first quartile being the lower edge). The two lines extending from the box are called its whiskers and the horizontal mark at the end of the whiskers represent the minimum and maximum value ( minimum value being the lower line). The dots represent the outliers of the data. Box and whisker plots enable us to compare the range and distribution of various features. The code to generate a box and whisker plot in Python's Seaborn library is given below. We need to specify the x axis, y axis and the target variable that we want to observe (optional) and the input dataframe. Seaborn also provides an option to save the image.

```
import seaborn as sns

#set the color theme
sns.set_palette("Paired", 5)

#box and whisker plot
```

```
sns.catplot(x='OrgType', y='FemaleWorkforce', hue='Target', data=data1, kind='box')
.savefig('workforce_org.svg')
```

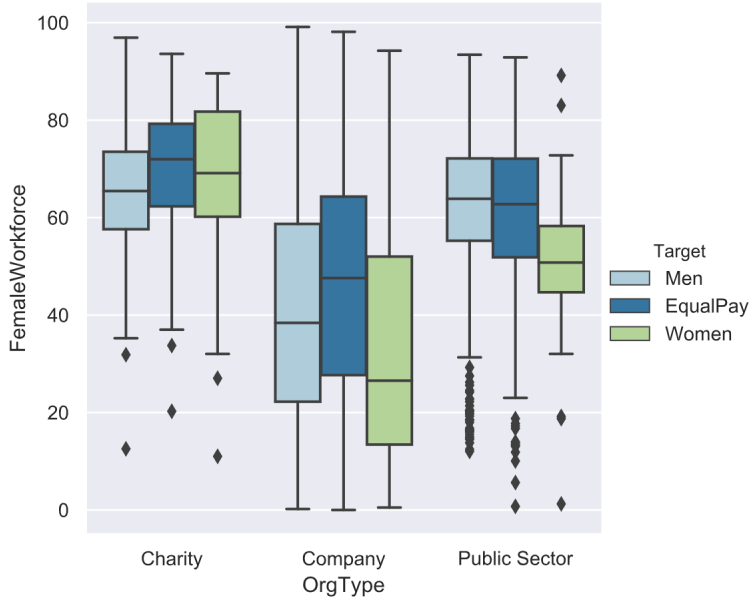


Fig. 6. Organization Type vs. Percent of Female Workforce shows that charities hire more women and tend to favor equal pay or paying women more

In Figure 6, the percent of the female workforce in each organization type - charity, company, and public sector are shown. The data is distributed in three classes - where men are paid more, where women are paid more and where both the genders are paid equally. It is quite obvious that charities tend to have a high percentage of women workforce and there are more charities that tend to pay more or equally. Charities are thus a better place to work for women. Companies, on the other hand, have a wide distribution for the number of women that are present in the workforce. Companies tend to prefer equal pay rather than favoring any one gender. Public Sector does seem to hire a higher percentage of women, but they tend to prefer either paying men more or both the genders equally.

### 6.2 Stacked Bar Charts

Stacked bar charts are a way of comparing totals against the parts. A stacked bar chart has segments of data stacked on top of each other and allows you to observe the composition of one category where it is split into multiple categories and the effect of each sub-category on the whole category [16]. A normalized stacked bar chart "percentage-of-the-whole of each group and are plotted by the percentage of each value to the total amount in each group" [16].

Creating a normalized stacked bar chart is not a easy thing to do. It requires some manipulation of the data first. We create a three lists - men_paid, women_paid, equal_pay - each contains a percent of companies that pay men more, women more and those that provide equal pay. These lists are then passed to the plot. Each list is plotted as a layer. The code [36] is given below.

```
import matplotlib.pyplot as plt
r = [0,1,2]
barWidth = 0.60
names = ('Charity','Company','Public Sector')

#plot bars for men paid more
plt.bar(r, men_paid, color='#3dbe98', edgecolor='white', width=barWidth, label='Men')

#plot bars for women paid more
plt.bar(r, women_paid, bottom=men_paid, color='#140ce7', edgecolor='white', width=barWidth

#plot bars for equal pay
plt.bar(r, equal_pay, bottom=[i+j for i,j in zip(men_paid, women_paid)], color='#f9bc86',

plt.xticks(r, names)
plt.xlabel("Organisation Type")
plt.ylabel("Percentage Paid More")
plt.legend(loc='upper left', bbox_to_anchor=(1,1), ncol=1)
```

In Figure 7, the percentage of genders being paid more are plotted for each organization type - charity, company, and public sector. We see that maximum equal pay practices are followed in

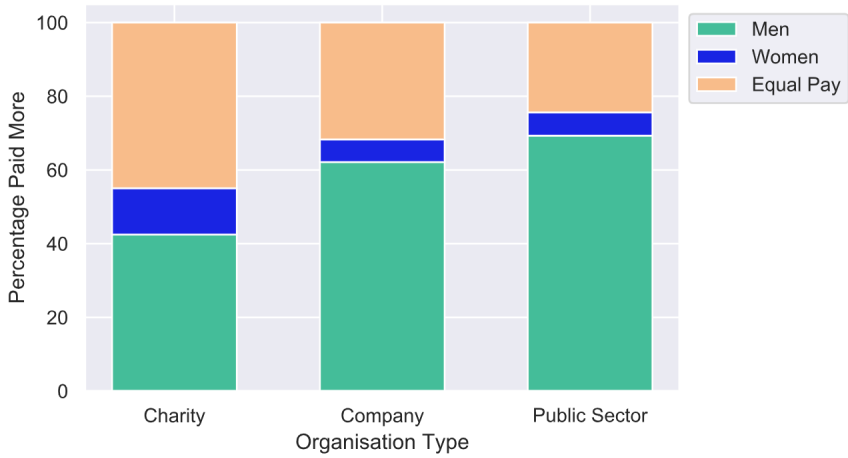


Fig. 7. Organization Type vs. Who Gets Paid More shows that charities favor equal pay the most and public sector companies fare the worst in this aspect

charities and the least in the public sector. Public sectors tend to have the highest percentage of organizations that pay men more than women. Charities have the highest percentage of organizations where women are paid more than men.

This graph supports the previous box plot. Charities tend to higher more women and are more inclined to equal pay or better pay than men and public sector organizations fare the worst in this case.

### 6.3 Bar Charts

Bar charts represent categories with the help of bars and the length of the bar indicates the quantity of the value being measured. Bar charts are easy to create in Seaborn. We have to provide the x axis, y axis, the target class if any and the input dataframe.

```
import seaborn as sns

#color
sns.set_palette("YlGnBu", 4)
p = sns.barplot(x="EmployerSize", y="Percentage", hue="Target", data=occupation_counts,
```

```
            order=['250 to 499', '500 to 999', '1000 to 4999', '5000 to 19,999', '20,000 o
_ = plt.setp(p.get_xticklabels(), rotation=90)  # Rotate labels
plt.legend(loc='upper left', bbox_to_anchor=(1,1), ncol=1)
plt.ylabel('Percentage Paid More')
plt.xlabel('Employer Size')
```

In Figure 8, the bar chart depicts the percentage of which gender gets paid more or equal pay for each different size of the employer. For example, half of the companies that have more than 20,000 employees tend to provide equal pay for both the genders. For all other company sizes, it is quite clear that only about a third of them that practice equal pay. It is evident that over half of them still tend to pay men more than women.

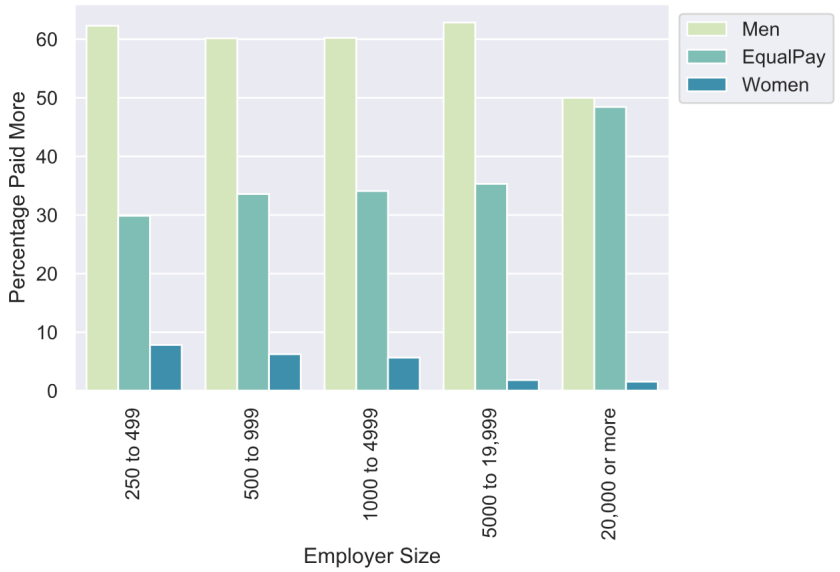


Fig. 8. Employer Size vs. Who Gets Paid More

## 7 USING PLOTLY TO ANALYSE COMPANIES HOUSE API

The Companies House API allows access to the names and numbers of each company's active officers. These include directors, secretaries and other persons of significant control. Analyzing the composition of the active officers can provide further insight into the gender pay gap of the

company. Does having more women in control directly imply a lower pay gap? Does having equal representation of both women and men cause a lower or higher pay gap?

To analyze this data, I decided to go with Plotly [12] library. Plotly is an outstanding visualization library to produce high-quality and interactive charts. Plotly allows you to apply layers on the graphs, zoom in or out, save the image, annotate each data point etc. The data for these plots is the one mentioned in Figure 3. The code for Figure 9 is given below.

```
#setting up plotly
from plotly.offline import init_notebook_mode, iplot
import plotly.graph_objs as go
import cufflinks as cf
import plotly.plotly as py

# Initialize plotly
cf.go_offline() # required to use plotly offline
cf.set_config_file(offline=False, world_readable=True)
init_notebook_mode(connected=True)

data.iplot(kind="scatter", theme="white",
x="DiffMedianHourlyPercent", y='FemaleDir_Pct',
categories='OrgType', mode='markers',
xTitle='Hourly Wage Gap', yTitle='% Female Directors')
```

Figure 9 shows the scatter plot of the hourly wage gap versus the percentage of female directors present in the company, for charities and companies. Plotly allows us to either view data only for charities or only for companies or both. This can be seen in Figure 9 and Figure 10. Figure 9 shows the relationship between hourly wage gap and number of female directors for both companies and charities, whereas Figure 10 shows the relationship only for charities data. Similarly, we can view the relationship for only companies data by disabling the charities button on the top right corner of the graph. On looking at the graph, we notice that there is no clear relationship between the hourly wage gap, the organization type and the percentage of female directors. We can, however, note that the blob of points tapers a little to the center where the percent of female directors is

high. The spread of the wage gap where the percent of female directors is low seems to be very wide as compared to the top. We can say that there is a high probability of a lower wage gap if the number of female directors in the company is higher than that of men.

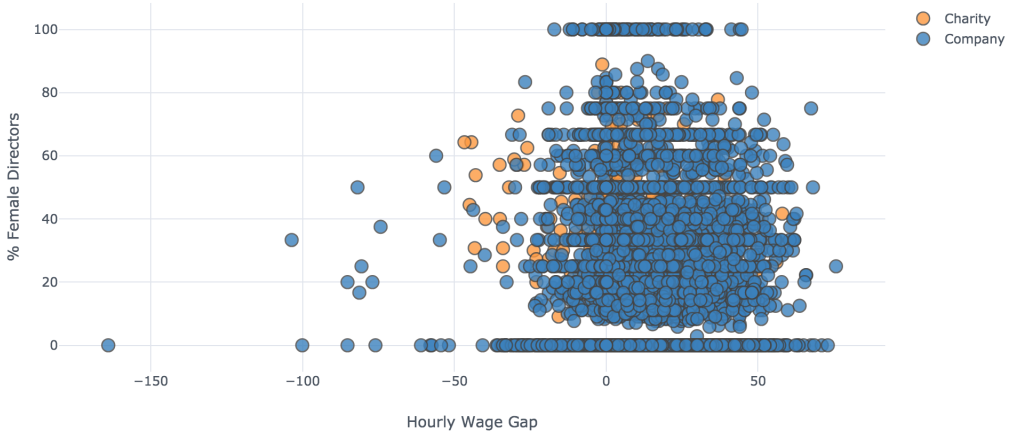


Fig. 9. Hourly Wage Gap vs. No. of Female Directors


Fig. 10. Hourly Wage Gap vs. Majority Gender of Directors for Charities shows that as no. of female directors increase, the wage gap converges to 0

In Figure 10, the companies option has been disabled in the graph. We see that as the number of female directors increase the hourly wage gap is closer to 0. Of course, there do exist outliers but

the natural trend is that as the number of female directors increase the points are clustered closer towards the 0 level.

In the Figure 11, we the effect of the majority gender representation on the wage gap. While there is not much difference in either group, we do see that if there is equal representation of the genders or a higher majority of women in the directors group, the wage gap tends to be relatively close to zero and the spread across the hourly wage is also lesser than if there is a male-dominated group of directors.

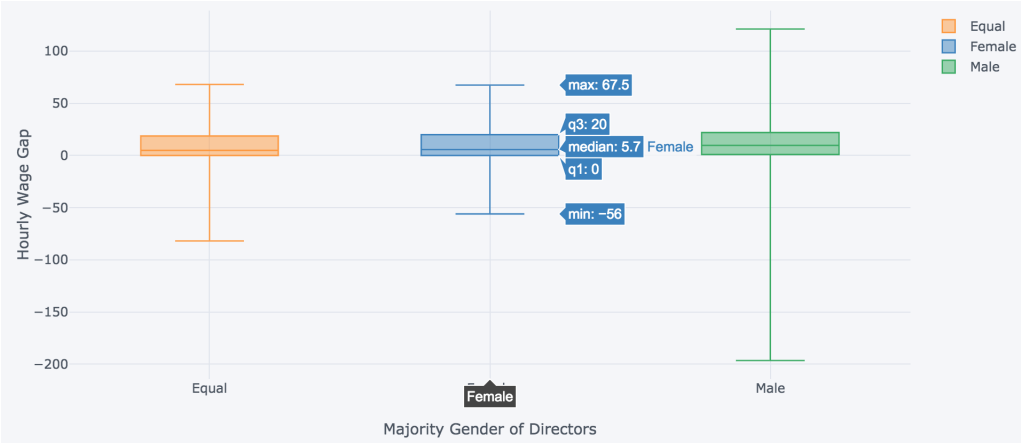


Fig. 11. Hourly Wage Gap vs. Majority Gender of Directors

The code for Figure 11 is given below.

```
#setting up plotly
from plotly.offline import init_notebook_mode, iplot
import plotly.graph_objs as go
import cufflinks as cf
import plotly.plotly as py

# Initialize plotly
cf.go_offline() # required to use plotly offline
cf.set_config_file(offline=False, world_readable=True)
init_notebook_mode(connected=True)

data[['Dir_Majority_Gender', 'DiffMedianHourlyPercent']].pivot
(columns='Dir_Majority_Gender', values='DiffMedianHourlyPercent').iplot(kind='box',
```

```
yTitle = 'Hourly Wage Gap', xTitle='Majority Gender of Directors')
```

## 8 ANALYZING 2017 US LABOR FORCE DATA

The Bureau of Labor Statistics in the United States Department of Labor, published a detailed data of median weekly earnings of men and women, as per the occupation in 2017 [10]. All occupations ranging from management occupations, legal, education, entertainment, farming, sales etc. are covered. A snapshot of the data is given below. All the numbers are in thousands.

| Occupation | 2017 | | | | | |
|---|---|---|---|---|---|---|
| | Total | | Men | | Women | |
| | Number of workers | Median weekly earnings | Number of workers | Median weekly earnings | Number of workers | Median weekly earnings |
| Total, full-time wage and salary workers | 113,272 | $860 | 62,980 | $941 | 50,291 | $770 |
| Management, professional, and related occupations | 47,207 | 1,224 | 22,815 | 1,442 | 24,393 | 1,052 |
| Management, business, and financial operations occupations | 19,414 | 1,327 | 10,415 | 1,526 | 8,999 | 1,134 |
| Management occupations | 13,169 | 1,392 | 7,568 | 1,573 | 5,600 | 1,173 |
| Chief executives | 1,136 | 2,296 | 823 | 2,415 | 313 | 1,920 |
| General and operations managers | 920 | 1,328 | 598 | 1,488 | 321 | 1,134 |
| Legislators | 14 | - | 5 | - | 9 | - |
| Advertising and promotions managers | 53 | 1,330 | 24 | - | 29 | - |
| Marketing and sales managers | 994 | 1,509 | 566 | 1,747 | 428 | 1,288 |
| Public relations and fundraising managers | 71 | 1,318 | 24 | - | 47 | - |
| Administrative services managers | 147 | 1,233 | 89 | 1,629 | 57 | 1,013 |
| Computer and information systems managers | 594 | 1,843 | 428 | 1,897 | 165 | 1,629 |
| Financial managers | 1,111 | 1,412 | 500 | 1,719 | 611 | 1,222 |

Fig. 12. 2017 Data of US Labor Force - Median Weekly Earnings by Detailed Occupation and Sex

To visualize this information, I wanted a graph that could convey the gaps in wages more easily than a bar chart or a line chart. I developed the visualization in R using the ggplot2 library. The ggplot2 library is an extremely efficient and easy library to use for visualizations. The visualizations are not interactive like Plotly, but they are easier to create and it's easy to create layer charts on each other so they can convey more information. The visualization is shown in Figure 13. The female median weekly earnings are marked with an orange dot and the male median weekly earnings are marked with a blue dot. The line between each pair of orange and blue dots represents the wage gap. This graph reveals some interesting information. We see that the wage gap is higher in professions like legal, management, health care, business and computers related occupations. Co-incidentally these are jobs requiring more sophisticated degrees. One would expect that jobs

having more educated employers would have a less of a pay gap. Unfortunately, that is not the case. These are also the jobs that have some of the highest median weekly earnings among all other occupations. Jobs like construction, food preparation and serving, administrative support occupations do not require fancy degrees and are also low paying occupations. However, these occupations have the lowest pay gaps. In fact, the construction occupation has an almost negligible wage gap. These do not have the most educated people and yet there is a minimal inequality. These are some points to ponder over as they reflect the thought process of the society.

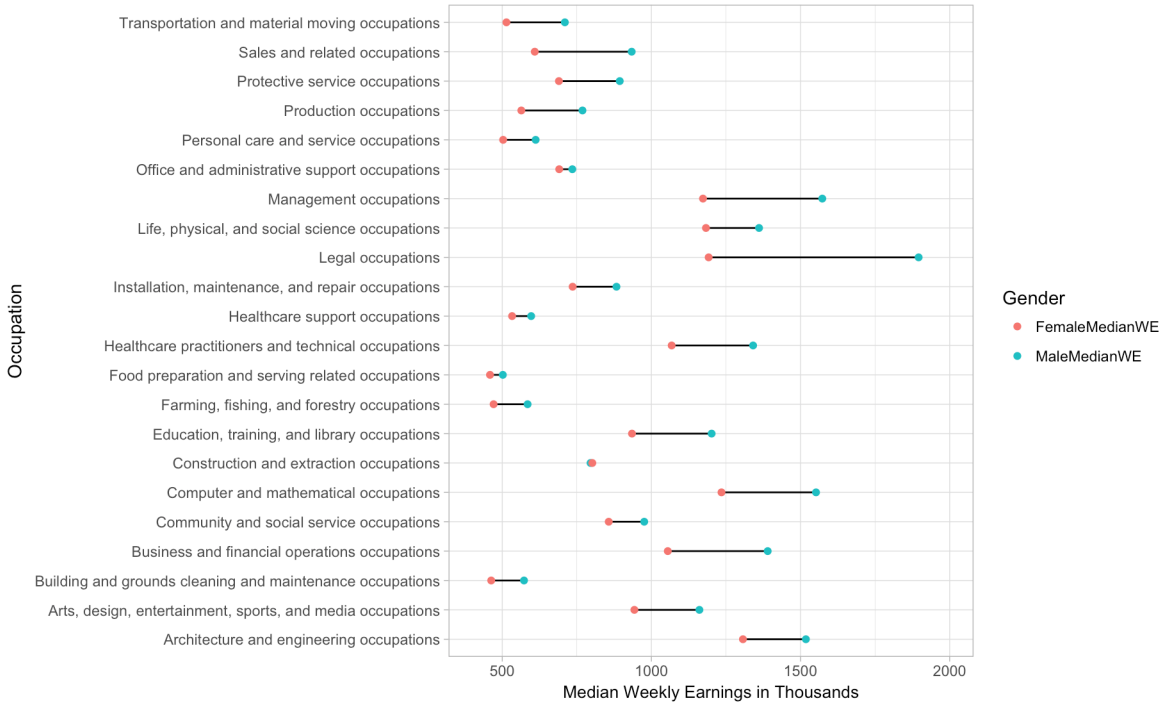


Fig. 13. Median Weekly Earnings by Occupation and Sex using ggplot2 in R

The code to generate the plot in Figure 13 is quite simple. An R editor like RStudio is required. ggplot2 has a function where we can add the dots, the line between them, set the theme of the background, adjust the size and font styles of the axes etc.

```
library(readr)   # read csv
library(ggplot2)  # data viz

data <- read_csv("2017_US_LaborForce_Data.csv")

# aes - aesthetics(x axis variable, y axis variable)
ggplot(data, aes(Median_Weekly_Earnings, Category)) +
  geom_point(aes(color = Gender)) +
  geom_line(aes(group = Category)) +
  theme_dark() +    #background theme
  coord_cartesian(xlim=c(400, 2000)) +  #x coord limits
  #setting the titles for the axes
  labs(title="2017 US Labor Force Statistics", x="Median Weekly
  Earnings in Thousands", y="Occupation") +
  #setting the font size and style for the elements on the axes
  theme(plot.title=element_text(size=13, face="bold"),
          axis.title.x=element_text(size=10),
          axis.title.y=element_text(size=11))
```

All of the plots that we have seen above require extensive knowledge of visualizing tools and various scripting languages. This is not the easiest task for laymen in programming. There is a need for a tool that even novices can use easily. One such free website is DataWrapper. This website and its interface is extremely easy and free to use for all. The first step is to upload the data store to the website. Once the data is uploaded, you can select the columns that you need for visualization. The website automatically detects the data type of the columns, which can be modified in case the website gets its wrong. There are numerous charts to select from like bar chart, split bars, range plot, stacked bars, lines, area chart, pie chart, donut chart, tables, scatter plots, bullet bars etc. One can manually set the range of the axes, add a title, set a description, add a link to the dataset etc. The colors used in the plot can also be modified. The rows in the plots can also be arranged as per the user's order. The graph is then ready to use. I used the range plot option to recreate the above chart that was created in R. The graph is similar to the one created in R and looks equally beautiful and insightful. The visualization is shown in Figure 14.

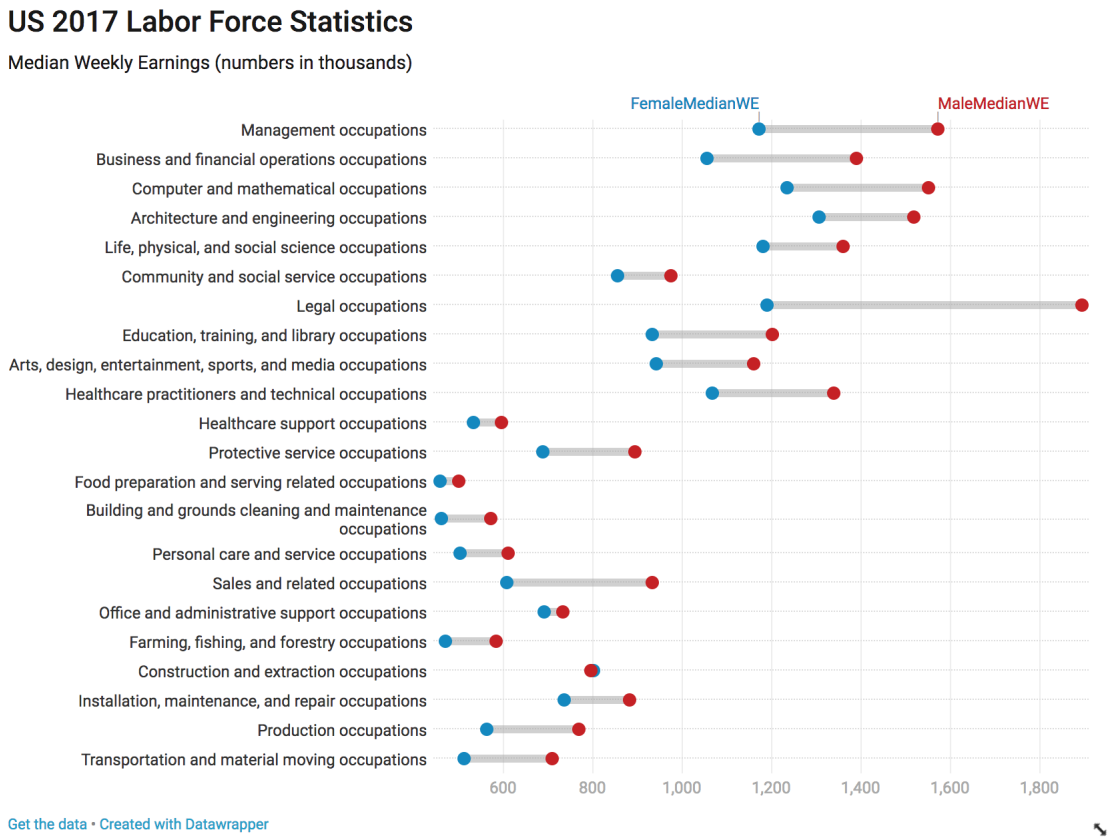


Fig. 14. Median Weekly Earnings by Occupation and Sex using DataWrapper

## 9 USING TABLEAU FOR VISUALIZATION

We analyze the public sector wage gap data by using Tableau. Tableau is an excellent and sophisticated tool for visualization. A free version, Tableau Public, is available for use and can be installed on a local machine. Data can be imported into Tableau through several ways - comma separated/Excel files, text file, json file, pdf file, spatial file or from a server. Tableau can automatically detect the data type of each of the columns or they can set manually. Multiple files can be aggregated in Tableau too. Beautiful and interactive visualizations can be created in Tableau without any prior knowledge of coding. For complex operations, scripting languages can be incorporated in Tableau as well [2]. It provides great support for mobile devices and visualizations can be adjusted

to various devices. It also has a very strong online community for support. The premium version can be pricey. Tableau does not support any data pre-processing tools and is strictly a visualization tool [2]. Tableau in this case is used to explore the wage gap details in the public sector data of United Kingdom.

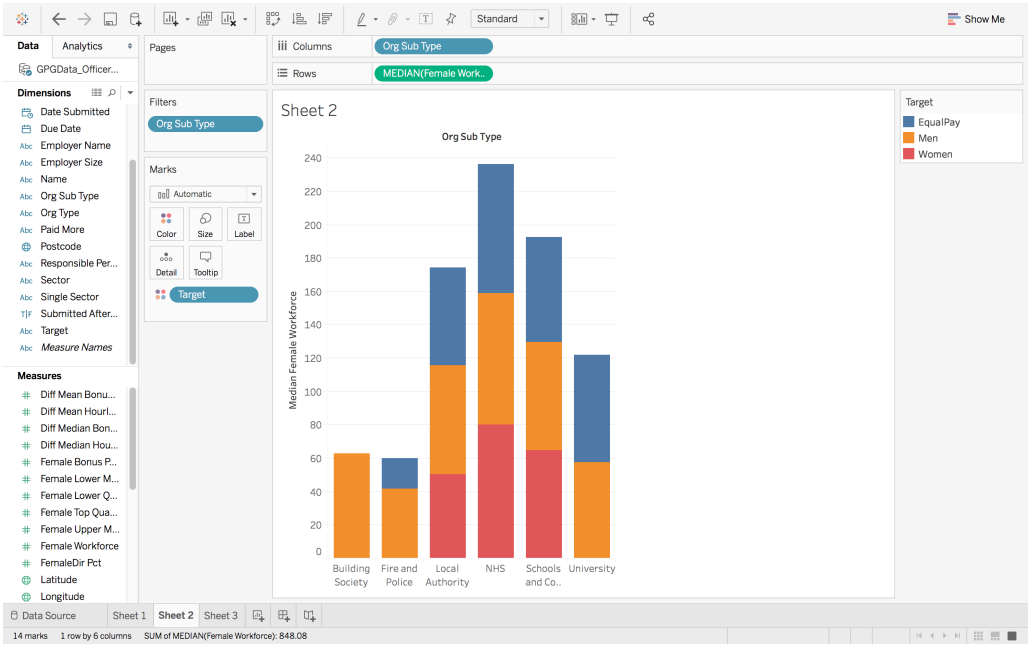


Fig. 15. Public Sector Analysis of the Wage Gap

We analyze the data only for public sector companies in United Kingdom. Public sector organizations include building societies, fire and police, local authorities, universities, schools and colleges and NHS (National Heath Service). The Tableau dashboard is shown in Figure 15. On the left hand side, the columns are mentioned in Dimensions and Measures. Dimensions (in blue) are usually fields that are discrete and cannot be aggregated over. Measures (in blue) are usually continuous values that can be aggregated over or mathematical operations can be performed on them. We simply drag the desired columns into the Columns and Rows section on the top of the sheet. The Show Me tab on the right most top corner shows all the possible visualizations for the given rows

and columns. Numerous visualizations can be created like bar charts, line plots, maps, tree maps, pie charts, text tables, heat maps, stacked bars, area charts, scatter plots, histograms, box and whisker plots, Gantt views etc. I added the OrgSubType as columns and FemaleWorkforce as row. Since Female Workforce is a measure, I chose the median as an aggregate measure. I also wanted to see how the Target, i.e. who gets paid more or equal pay with respect to the above factors. I dragged the Target column on the Color section on the Marks tab. This ensures that each Target value gets a unique color assigned to it, as shown in Figure 15. These colors are customizable. Appropriate marks according to the plot can be selected. Tooltip provides the option to add a tooltip that will display information on mouse hover actions. Size will change the size as per the target variable's value. Size is more useful while using scatter plots.

In the visualization Figure 15, we see that Building Societies and Fire and Police tend to have the smallest number of female workforce. NHS and Schools and Colleges tend to have the highest workforce of female employees. Women get a chance to be paid more in local authorities, NHS and schools and colleges. Men tend to be paid more in every sector. Equal pay is more prominent in all sectors except building societies where it is absent completely.

## 10 SENTIMENT ANALYSIS OF PAY GAP RELATED TWEETS

In this section, tweets pertaining to the pay gap are extracted from Twitter and analyzed to gain more insights into what the social media is talking about the issue.

### 10.1 Extracting Twitter tweets in Python

Twitter tweets can be extracted from Twitter by using the Tweepy [17] library in Python. Tweepy is an open source library that works with the Twitter API to provide streaming tweets. Below is a walk through on how to use the API to extract live streaming tweets.

The first step is to create an account on dev.twitter.com and obtain consumer keys and access tokens. These are needed by the Tweepy library to access the Twitter API. Tweepy uses OAuth, an authentication handler, that needs consumer keys and access tokens (per application), to allow access to Twitter. This is a modification from the BasicAuth that weepy used earlier. BaiscAuth required the user's user name and password for the purpose of authentication. OAuth is better because it does not require any such confidential information. In OAuth, explicit read/write permissions can be set on the consumer keys and access tokens as per the developer's requirement.

Tweepy provides a StreamListener object that is used to monitor and extract tweets in real-time. It is implemented in the StdOutistener class which is a simple listener that listens to the tweets and prints the output. An object Stream is created to collect the output of the listener and it has methods like filter() and track() which are used to filter the tweets by specific keywords. Other options include getting streams from a particular user, filter according to the hashtags etc. The code is adapted from [8] and a simple code snippet to extract tweets having the keyword "genderpaygap" is mentioned below.

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

class StdOutListener(StreamListener):
    def on_data(self, data):
        print(data)
        return True

if __name__ == '__main__':
    #create an object for streaming
    listener = StdOutListener()
    #provide the consumer and access keys obtained from Twitter
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, listener)
    #provide keywords for filtering
    stream.filter(track=['genderpaygap'])
```

A disadvantage of the Tweepy package is that the user cannot access tweets that are more than a week old. This is a major limitation for developers who are looking to collect a large amount of historical data. The typical workaround would be to run the code daily until the amount of data required is collected. This is a very time consuming method and not feasible for most applications. This hack would still not allow you to access data that is older than a week.

A developer came up with an idea to write a program that simulates how a Twitter feed and the scroll function works [6]. Once you start scrolling the Twitter page, as one keeps scrolling, older tweets keep getting displayed. Twitter uses a JSON application to do this. The developer writes a program that takes in search criteria like tweets from a certain user, keywords to be searched for, date ranges for the tweet etc. and generates the Twitter url for the above criteria. Using PyQuery, the program imitates a Twitter web page and receives the older tweets. The output of the program is the collection of tweets returned in a JSON file. This code was used to extract 10,446 gender pay gap related Twitter tweets from December 2017 to June 2018. The data collected consisted of date, the username of the Twitter account and the corresponding tweet. A subset of the data is shown below in Table 2.

### 10.2 Pre-processing the Tweets

The tweets have a lot of noise in them which will hamper the process of sentiment analysis. Examples of noise in tweets includes excess punctuation, hyperlinks, pictures or links to pictures etc. These do not provide any concrete evidence for sentiment analysis. This blog post [44] provides a detailed explanation on how to pre-process tweets in Python. The following steps were taken to pre-process the text.

- Removal of Usernames

Table 2. Twitter Dataset

| timestamp | username | tweet |
|---|---|---|
| 2018-04-30 23:00:00 | ShandyCruicks | Leading by example to close the #gender pay gap at @salesforce http://ow.ly/jpxf30jFfVh #gendergap #paygap #leadbyexample #raisethebar #salesforce #equality |
| 2018-04-30 21:18:00 | iamcelinepalmer | Third of British bosses do not believe gender #paygap is business issue https://buff.ly/2qzksxx @telebusiness |
| 2017-12-01 17:46:00 | ywtf | Women are 40% sole breadwinners yet #paygap & major #studentdebt burden hurt financial stability. We must do something: http:// bit.ly/2gcAtFC . #DeeperinDebt pic.twitter.com/aLAYQ7MY68 |
| 2018-04-28 07:12:00 | The_Pay_Gap_Bot | The median gender pay gap is 9.1%, or £7,274,462,533.85 so far this year. #paygap |

Several tweets contain usernames which are usually used to bring the tagged user's attention to the tweet, or the tweet might be about the username mentioned etc. In the following example the user is tweeting about an act that former President Barack Obama signed and thus he tags Barack Obama in the tweet. The username "@BarackObama" is of no use for sentiment analysis.

An example would be "#OTD in 2009, @BarackObama signed the Lilly Ledbetter Fair Pay Act into law. 9 years later, we continue to fight for #EqualPay for equal work. #PayGap #GenderPayGap #OurVoicesOurTime pic.twitter.com/rpWE39BGr7"

The regular expression for removing usernames is `r"@[^\s]+[\s]?"`

- Eliminating Hyperlinks and Links to Pictures

  Just as usernames provide no information, hyperlinks and images embedded in the tweets are also noisy data. This tweet "We already know adequate child care is part of the solution to the #PayGap http://bit.ly/2FwMhtu pic.twitter.com/WM15C4sbZj" illustrates that the urls are just a bunch of characters with no special meaning for natural language processing.

The regular expression for removing hyperlinks is `r"http.?://[^\s]+[\s]?"` and that for removing image links is `r'(pic\.twitter\.com/).{10}'`.

- Removal of Special Characters and Excess Punctuation

  Special characters like the symbol hashtag in tweets, emoticons and excess punctuation marks need to be eliminated as well. Even though some sentiment analysis tools like VADER which are discussed below, can handle some emoticons, it is still better to remove hashtags and excessive punctuation marks because those are not handled well by tools. To reduce a punctuation character like exclamations, use the regular expression `r'!{3,}'), "!"` and to remove special characters (`:, #, //, [, ], *, <, >, =, -, _, $, &, %`), simply replace them with a space.

### 10.3 WordCloud Analysis

A word cloud is a visual representation of textual data that consists of the most frequently occurring words in the text. Words are displayed and their size is directly proportional to their frequency in the text. word clouds help us to understand visually what is being said in the data and its emphasis. It helps in magnifying any underlying popular sentiments that are present in the data. The tweets collected above were used to create a word cloud and the following word cloud was generated.

On observing Figure 16, the words 'equality', 'genderequality' , 'british' etc. jump out to the viewer. This implies that these words are the most frequently occurring words in the tweets dataset. People are obviously talking about more equality when it comes to wage gaps. The mention of 'british' and 'bosses' can be attributed to an article published in The Telegraph [21] in 2017 that mentioned that a third of British bosses do not believe that the gender wage gap is an issue of immediate concern. The article caused quite a public outrage in social media as it was detrimental to the movement of eliminating the wage gap. On the top left corner, we see the word 'iceland'.


Fig. 16. Word Cloud that demonstrates what social media is talking about the pay gap

Iceland has the lowest wage gap in the world and recently passed a law that made discrimination in workplaces illegal and a mandatory rule for every company having greater than 25 employees to undergo a government-conducted audit to prove that discrimination on the basis of a wage gap does not exist [20]. In that corner we also see the word 'myth'. Maybe people are discussing about what pay gap statistics are myths or facts. In the center, we see that a tiny proportion of people are also talking about the 'timesup' movemnet and 'womenintech'. The pay gap issue has been quite prevalent in the tech industry too. According to this Business Insider article, the pay gap in the US is the worst in the Silicon Valley area [23]. The Times Up Movement [18], a movement against sexual harassment of women in workplaces, also finds its mention here. Workplace discrimination occurs in many forms and sexual harassment and disparity in wages are the biggest examples. Words like 'workplace', 'discrimination', 'feminism' etc. are expected to occur frequently. 'bbc' is also mentioned in the word cloud that refers to the BBC wage gap controversy where several wage disparities (around 9.3% wage gap difference [35]) in BBC employees were discovered. As a result,

salaries of multiple male journalists had to be reduced and the editor of the China BBC resigned in order to protest against the pay gap.

Python has an built-in wordcloud library to generate word clouds. It also has a set of stopwords that can be modified to add additional words. Stopwords are words like 'a', 'an', 'the' etc. which occur frequently but have no real significance to the text. Such words need to be eliminated so that other important and relevant words show up in the word cloud. Other typical words like 'man', 'women', 'paygap' were added so that we can see other words that are being talked about. The method WordCloud() also has abilities to set the background color, set sizes for the smallest(lowest frequency) word and the biggest(highest frequency) words. The code to generate a wordcloud is given below.

```
import matplotlib.pyplot as plt #display
from wordcloud import WordCloud, STOPWORDS  #wordcloud

#define stopwords
stopwords = set(STOPWORDS)
stopwords.update(['paygap', 'pay', 'gap', 'gender', 'genderpaygap',
'women', 'woman', 'man', 'men', 'female', 'male', 'take', 'equalpay'])

#generate wordcloud
wordcloud = WordCloud(stopwords=stopwords, background_color="white",
min_font_size=6, colormap='Dark2',max_font_size=35).generate(text)

#display the plot
plt.figure(figsize=(13,23))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

### 10.4 VADER for Sentiment Analysis

Valence Aware Dictionary for Sentiment Reasoning (VADER) [37] is an open source library for sentiment analysis. The tool is developed in Python but it can be imported to Java, JavaScript,

PHP and Scala. The developers of VADER have developed a ”gold standard sentiment lexicon” for ”micro-blog like contexts”. Given an input text, VADER provides the following scores as per the sentiment intensity - negative, neutral and positive. The range of these values ranges from -1 to +1 where 1 implies maximum intensity. A compound score (a one-dimensional measure of the sentiment) is also provided whose value ranges from -1 to +1 where -1 implies most negative sentiment and +1 implies most positive score. The values of the compound score can be interpreted as: values greater than +0.5 are positive, the values between -0.5 and +0.5 are neutral and values below -0.5 imply negative score. VADER looks up the sentiment of each constituent word in its sentiment lexicon and returns a score that is between -4 to +4. The scores of all the words are normalized to an output compound score that ranges from -1 to +1. The advantages of VADER are that work especially well for social media like texts ( which is highly useful because sentiment analysis is heavily used in analyzing customer reviews or feedbacks which are posted on social media platforms), it requires no training data unlike other machine learning approaches, and it is fast and efficient.

The developers of VADER first formed a lexicon of words gathered from various pre-existing lexicons. Other lexicons like emoticons and colloquial terms that are popular on social media were added to the word bank. To rate all of these words as per their sentiment intensity, VADER uses ”wisdom of the crowd” approach. It employs workers from Amazon Mechanical Turk (AMT), a crowd-sourcing website, to rate words according to their sentiment. A crowd-sourcing website like AMT employs workers to perform minor tasks for small amounts of money. The workers get paid $0.25 for every 25 words that they rate. Each worker rates batches of 25 words on a rating of -4 to +4 where -4 is most negative, 0 is neutral and +4 is most positive. Workers must score 80 percent and above on a English test which is used as a pre-screening round. They must also attend a mandatory training session where they are instructed on how to rate words. These rounds

ensure fair and consistent rating. If a worker's ratings deviate from the mean of other ratings in that batch of 25 words, that particular worker's ratings are discarded to avoid outliers. VADER uses five principles to assess the sentiment intensity -

- Punctuation marks increase the intensity of the sentiment. 'I hate you' is less negative than 'I hate you!'.
- Capitalization of words or all letters in a word also increases the sentiment intensity of the text like 'I hate you' is less negative than 'I HATE you'.
- Degree modifiers like 'very', 'extremely', 'moderately' etc. also affect the intensity accordingly.
- Conjunctions like 'but' are paid special attention and the text that follows the conjunction is given a higher priority and its sentiment is the dominant sentiment.
- A tri-gram that appears before a high intensity word is also given more weight. It can help to identify cases where the polarity of the text flips.

For ground truth, the developers selected 4000 tweets from Twitter, 309 customer product reviews, around 5000 snippets from New York Times editorials and around 10,000 movie reviews from Rotten Tomatoes. The developers used F1 score, precision, recall, correlation of the calculated sentiment versus the ground truth sentiment as evaluation metrics. VADER was compared to other pre-existing sentiment analysis tools like Linguistic Inquiry Word Count (LIWC), WordNet, SenticNet (SCN) etc. and it performs better than all of these, only worse than individual human raters. Thus, VADER is a reliable sentiment analysis tool that is open source, fast, efficient and requires no training or domain data.

VADER can be implemented in Python quite easily. A code snippet is provided to view its implementation.

```
#import for sentiment analysis
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

#object creation
analyser = SentimentIntensityAnalyzer()

#get sentiment intensity scores
score = analyser.polarity_scores("I am very happy!")

print(score)
#score - {'neg': 0.0, 'neu': 0.318, 'pos': 0.682, 'compound': 0.6468}
```

### 10.5 Sentiment Analysis on the Tweets

Sentiment analysis was performed on the tweets dataset by using VADER. VADER returns a compound score for each tweet. Based on the compound score, a sentiment of 'positive', 'negative' or 'neutral' is added to each tweet. If the score is above 0.5, the tag is positive, if the score is below -0.5, the tag is negative, else it is neutral. The distribution of tweets is shown in Figure 17. We see that over 50% of the tweets are neutral in sentiment. This can be attributed to the fact that mos of the tweets are links to other articles, solutions to the pay gap issue, or just facts about the issue rather than texts laden with explicit negative words.

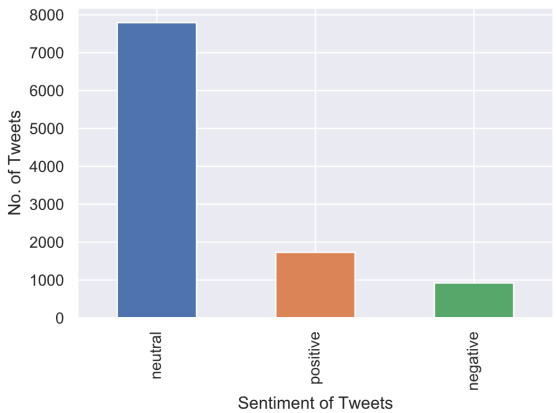


Fig. 17. Distribution of Tweets over 7 months

For a closer look at all the negative tweets, a word cloud was generated from all the tweets having a negative sentiment. The word cloud is shown below in Figure 18. Looking at the word cloud we see users talking about equality and equal pay. The word 'die' is one of the biggest and most frequent words. We also see 'bbc' mentioned where the users are talking about the BBC controversy about the pay gap in the corporation. We see all the expected words like 'discrimination', 'feminism', 'worst' etc.

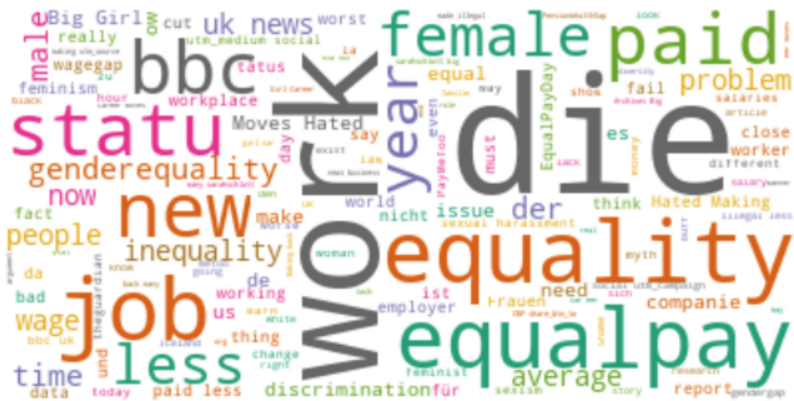


Fig. 18. Word Cloud Analysis of Negative Tweets

In Figure 19, the trend of tweets is plotted against the time. We see there are 2 distinctive spikes near January 2018 and April 2018. Iceland became one of the first countries to pass strict laws that made the wage gap illegal. Their stringent laws were widely applauded and seen as an inspiration to other countries to pass similar, effective laws. Thus we see a spike in social media where more people start talking about the wage gap. After January, there is a lull where the conversation seems to have died down, only to pick up pace near April 2018. Two major events occurred in April 2018. The first is the Equal Pay Day that occurred in April 2018. Equal Pay Day stands for the number of extra days a woman must work in a year, on average, to close the wage gap with her male counterparts. It is seen as a day to spread awareness about the issue. United Kingdom also made it mandatory for organizations to disclose company information about how the women and

men in the respective organization were getting paid. These events caused a major spike in the number of tweets that were generated and this period sees the maximum number of both, positive and negative tweets. After April, we again see a sharp decline. This is expected behavior as it is the trend with any topic on the social media. While the topic is trending, people discuss it extensively and as time passes all the debate and discussion lose steam, only to rise again when another trigger happens. The plot in 19 was generated using Plotly.

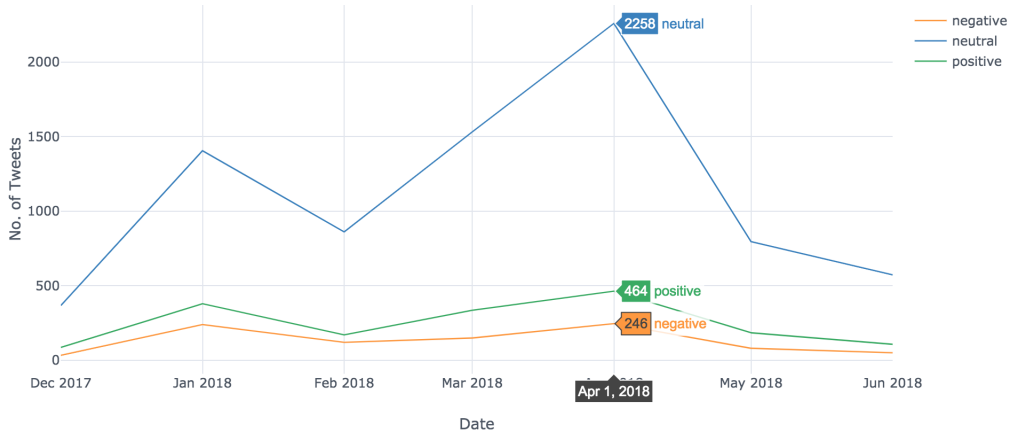


Fig. 19. Trends of Tweets over 7 months show sharp spikes in activity during certain events

## 11 OTHER CONCEPTS LEARNED IN THE INDEPENDENT STUDY

### 11.1 Dimensionality Reduction using t-SNE

Dimensionality reduction is a common practice that is carried out before developing a machine learning model. High dimensional data requires extensive computation time and storage. High dimensional data is also very sparse leading to a poorly built machine learning model [54]. It is also difficult to visualize the data. Reducing dimensionality allows removal of features that are collinear, irrelevant or redundant, making it easier for the machine learning model to understand

the data better. Data in higher dimensions also leads to over-fitting. Techniques for dimensionality reduction include principal component analysis, singular value decomposition, t-SNE etc.

The most popular method for dimensionality reduction is Principal Component Analysis (PCA). PCA reduces higher dimension data to lower dimensions while preserving the overall variance of the data. It captures the linearity of the data and focuses on preserving large pairwise distances [52]. It is fast and a simple algorithm to use. PCA uses a global covariance matrix and rotates it diagonally with the help of eigenvectors that now represent the data. However PCA is unable to handle non-linear data and this is the biggest advantage that t-SNE has over PCA.

t-SNE stands for t-Distributed Stochastic Neighbor Embedding. It captures non-linear structure of the data and uses the local relationship between the points to create a low dimensional mapping [52]. t-SNE creates a Gaussian probability distribution to capture the relationship between the neighboring points. It then recreates the same distribution in a lower dimensional space (two dimensional, by default) by using the Student t-distribution. t-SNE uses stochastic neighbors and it tries to preserve the local structure more than the global structure [43]. When trying to map data from higher dimensions to lower dimensions, there is less space in the lower dimensions, hence the points tend to get crowded. t-SNE avoids this crowding problem by using gradient descent. Although the disadvantage is that t-SNE might get stuck in a local minima leading to not well-defined clusters [43]. The advantages of t-SNE are that they help in visualizing complex data easily and get an intuition of how the data looks like. It helps to capture the non-linearity of the data. It avoids the crowding problem. However t-SNE does not use a learning function and it cannot be used for unseen data. It is non-deterministic, resulting in different results every time it is run [19]. The distances and densities in the clusters formed by t-SNE have no meaning to them. t-SNE also cannot deal with incomplete data. On the other hand, PCA is deterministic and can deal

with incomplete data [19]. t-SNE is also computationally expensive, hence PCA or SVD is first run to reduce the dimensions and then t-SNE is used for 2 dimensional mapping.

Implementation of t-SNE vs. SVD is explained with the help of the 20 Newsgroup dataset in Python [30]. The 20 Newsgroups dataset consists of 20000 news documents split evenly across 20 different newsgroups [1]. For the scope of this example, 6 different groups and their respective documents are taken. the newsgroups selected are 'alt.atheism', 'talk.religion.misc', 'comp.graphics', 'sci.space', 'talk.politics.misc' and 'rec.sport.baseball'. Python's Scikit Learn package has the dataset loaded already. The dataset can be loaded in the following manner. The training data belonging to the 6 newsgroups is selected and is converted into a numeric format so that it can be interpreted by algorithms.

```
from sklearn.datasets import fetch_20newsgroups
from sklearn.feature_extraction.text import TfidfVectorizer

# specify the newsgroups
categories = ['alt.atheism', 'talk.religion.misc', 'comp.graphics', 'sci.space',
'talk.politics.misc','rec.sport.baseball']

#fetch the training data
newsgroups = fetch_20newsgroups(subset="train", categories=categories)

#convert the text to numeric format
vectors = TfidfVectorizer().fit_transform(newsgroups.data)
```

Singular value decomposition is run to reduce the data to 2 dimensions. The visualization of the data in 2 dimensions is shown in Figure 20. We see that the data is not very well separated into clusters. The time taken by this algorithm is 0.25 seconds.

```
from sklearn.decomposition import TruncatedSVD

X_svd = TruncatedSVD(n_components=2, random_state=0).fit_transform(vectors)
```

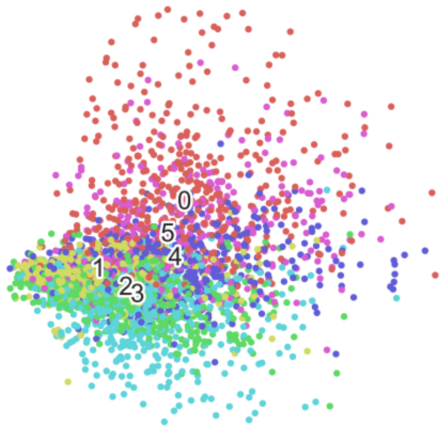


Fig. 20. Clusters formed by SVD are not very well-separated

We now implement SVD on the data to reduce it to 50 dimensions and then use t-SNE to convert it to 2 dimensions. The visualization is shown in Figure 21. We see that the clusters are well separated and we can see how the data looks like. The total time taken by the following code is 106.31 secs. If t-SNE was run directly on the data, it take 8 times longer to perform well-separated clusters that are slightly better than the results formed by SVD and t-SNE combination. The visualization is shown in Figure 22.

```
from sklearn.decomposition import TruncatedSVD
from sklearn.manifold import TSNE

X_svd = TruncatedSVD(n_components=50, random_state=0).fit_transform(vectors)

X_tsne = TSNE(n_components=2, perplexity=40, verbose=2).fit_transform(X_svd)
```

The comparison between the algorithms is shown in table Table 3. SVD followed by t-SNE seems to be a better choice as it is a quicker computation. It is a trade-off that must be decided by the developer according to his/her needs.


Fig. 21. SVD followed by t-SNE forms well-separated clusters in shorter time


Fig. 22. t-SNE forms relatively better well-separated clusters in a longer time

Table 3. Comparison of dimensionality reduction algorithms for 20 Newsgroups Data

| DR Algorithm | Time taken (secs) | Quality of Clusters |
|---|---|---|
| SVD | 0.25 | Clusters not well-separated |
| SVD followed by t-SNE | 106.31 | Less well-separated clusters |
| t-SNE | 821.22 | More well-separated clusters |

## 12 MAJOR INSIGHTS GATHERED FROM THE DATA

Listed below are some of the major insights concluded after a thorough analysis of the data -

- Urban areas in the UK are more likely to have an equal and fair compensation for all of its employees.
- In the UK, charitable organizations tend to hire more women and favor an equal pay system.
- While charities favor equal pay the most, public sector organizations tend to favor equal pay the least.
- The size of the company plays no direct role in the fairness of the compensation.
- Higher the number of female directors in the board, lesser are the chances of a large wage gap occurring.
- In the US, occupations like legal, computers, health care etc. tend to pay the most but also have a higher wage gap.

- In the US, jobs like construction, service, maintenance pay less but also have a minimum wage gap.
- Social media users tend to discuss a lot about the topic when a triggering event occurs, but the talks die down after a few days and the talk comes up only when another controversy occurs. Awareness must be raised on a regular level.

## 13 DATA SCIENCE TECHNIQUES LEARNED DURING THE INDEPENDENT STUDY

To summarize, the following data science techniques were studied and implemented throughout the course of this independent study -

- Techniques for scraping web data by using official APIs and manual scraping.
- Various steps involved in data cleaning, regular exressions and exploratory data analysis.
- Visualization tools - Seaborn, Plotly, ggplot2, Tableau, DataWrapper
- Sentiment Analysis and word cloud generation and analysis
- Dimenisonality reduction techniques like PCA and t-SNE and their implementation.

## 14 CONCLUSION

In this independent study, I have learned to use a set of data science tools to perform exploratory data analysis on the gender pay gap data in a proficient manner. For creating interactive visualizations, Tableau and Plotly are wonderful and efficient options. Plotly is free and allows the option of saving the image. The free version of Tableau is equally interactive but extremely easy to use and requires no prior knowledge of coding. Seaborn is also a great library to visualize data that is in the form of a dataframe. gglot2 in R is also a great library and an alternative to Seaborn. I would prefer ggplot2 to Seaborn as it was easy to layer charts on top of each other. Scraping data is slightly challenging task. If there is an API available like Twitter or the Companies House data, it is easier to follow.

Scraping a website does take some time to figure out the HTML tags. Care must be taken to follow the website's privacy guidelines. This project has been an extremely educational project where I had the chance to study and master a variety of skills and tools.

## 15 TIME SHEET

| Week No. | Date | Time In | Time Out | Total | Description |
|---|---|---|---|---|---|
| 1 | 08/24/18 | 13:00 | 15:00 | 2:00 | Meeting with Dr Kinsman |
| | 08/25/18 | 15:00 | 18:00 | 3:00 | Proposal writing |
| | 08/27/18 | 10:00 | 11:00 | 1:00 | Finalizing proposal |
| | 08/29/18 | 16:00 | 18:30 | 2:30 | Exploring the data |
| | 08/31/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 2 | 09/05/18 | 14:00 | 18:30 | 4:30 | Initial Cleaning + Web Scraping |
| | 09/05/18 | 22:30 | 23:45 | 1:15 | Web scraping for addresses and sectors |
| | 09/06/18 | 17:00 | 19:00 | 2:00 | Web scraping for board member info |
| | 09/07/18 | 9:00 | 11:30 | 2:30 | Getting charity data |
| | 09/07/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 3 | 09/09/18 | 14:00 | 17:00 | 3:00 | Gathering background info. + report |
| | 09/12/18 | 17:00 | 19:15 | 2:15 | Gathering background info. + report |
| | 09/13/18 | 22:00 | 0:30 | 2:30 | Gathering background info. + report |
| | 09/14/18 | 7:00 | 10:15 | 3:15 | Gathering background info. + report |
| | 09/14/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 4 | 09/20/18 | 20:00 | 22:30 | 2:30 | Visualizing-Basic EDA |
| | 09/21/18 | 10:00 | 11:45 | 1:45 | Visualizing-Basic EDA |

| Week No. | Date | Time In | Time Out | Total | Description |
|---|---|---|---|---|---|
| | 09/21/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 5 | 09/24/18 | 17:00 | 20:45 | 3:45 | Studying choropleth maps and Basemap |
| | 09/25/18 | 14:00 | 16:15 | 2:15 | Studying choropleth maps and Basemap |
| | 09/27/18 | 18:00 | 21:00 | 3:00 | Visualizing choropleth for UK Data |
| | 09/27/18 | 23:30 | 2:30 | 2:00 | Visualizing choropleth for UK Data |
| | 09/28/18 | 11:00 | 12:15 | 1:15 | End Member Analysis |
| | 09/28/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 6 | 10/01/18 | 10:00 | 12:30 | 2:30 | Adding a third ”EqualPay” Label + visualizing |
| | 10/01/18 | 15:00 | 16:45 | 1:45 | Adding a third ”EqualPay” Label + visualizing |
| | 10/04/18 | 21:00 | 1:00 | 4:00 | Scraping for Company House API |
| | 10/05/18 | 17:00 | 22:00 | 5:00 | code running for about 5 hours |
| | 10/06/18 | 16:00 | 17:00 | 1:00 | Meeting with Dr Kinsman |
| 7 | 10/08/18 | 18:30 | 20:45 | 2:15 | Scraping C.H. API again due to previous errors |
| | 10/11/18 | 7:00 | 9:30 | 2:30 | Scraping C.H. API again due to previous errors |
| | 10/11/18 | 11:30 | 3:00 | 3:30 | Cleaning and pre-processing data |
| | 10/12/18 | 10:00 | 12:30 | 2:30 | Pre-processing C.H. data and feature engineering |
| | 10/12/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 8 | 10/13/18 | 17:30 | 20:00 | 2:30 | Studying plotly |
| | 10/14/18 | 13:00 | 15:00 | 2:00 | Visualizations for plotly and C.H. API data |
| | 10/18/18 | 18:00 | 19:30 | 1:30 | Initial report writing |
| | 10/18/18 | 22:00 | 1:15 | 3:15 | Initial report writing |

| Week No. | Date | Time In | Time Out | Total | Description |
|---|---|---|---|---|---|
| | 10/19/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 9 | 10/21/18 | 13:00 | 17:00 | 4:00 | Collection and pre-processing of tweets |
| | 10/21/18 | 21:30 | 0:00 | 2:30 | Pre-processing and studying VADER |
| | 10/22/18 | 14:00 | 18:15 | 4:15 | Wordcloud generation |
| | 10/25/18 | 21:00 | 22:30 | 1:30 | Wordcloud generation |
| | 10/26/18 | 12:30 | 1:30 | 1:00 | Implementation of VADER + visualizations |
| | 10/26/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 10 | 10/28/18 | 15:00 | 17:30 | 2:30 | Finding and collecting 2017 US labor data |
| | 10/29/18 | 19:00 | 22:00 | 3:00 | Studying R and ggplot2 |
| | 10/31/18 | 17:00 | 20:00 | 3:00 | Implementing visualizations in ggplot2 |
| | 11/01/18 | 22:00 | 23:30 | 1:30 | Using DataWrapper for data visualization |
| | 11/02/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 11 | 11/05/18 | 14:00 | 17:00 | 3:00 | Study of t-SNE |
| | 11/06/18 | 15:00 | 18:00 | 3:00 | Study of t-SNE |
| | 11/07/18 | 17:00 | 19:30 | 2:30 | Implementing t-SNE on datasets |
| | 11/09/18 | 1:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| 12 | 11/15/18 | 14:30 | 18:00 | 3:30 | Tableau tutorial - Coursera |
| | 11/15/18 | 19:30 | 22:00 | 2:30 | Tableau tutorial - Coursera |
| | 11/16/18 | 16:00 | 19:00 | 3:00 | Tableau tutorial - other |
| | 11/16/18 | 21:00 | 0:30 | 3:30 | Implementing visualizations in Tableau for GPG data |
| 13 | 11/19/18 | 14:30 | 18:30 | 4:30 | Report Writing |

| Week No. | Date | Time In | Time Out | Total | Description |
|---|---|---|---|---|---|
| | 11/20/18 | 10:30 | 14:00 | 3:30 | Report Writing |
| | 11/21/18 | 15:00 | 20:00 | 5:00 | Report Writing |
| 14 | 11/25/18 | 16:30 | 19:00 | 3:00 | Report Writing |
| | 11/27/18 | 21:00 | 01:30 | 4:30 | Report Writing |
| | 11/28/18 | 17:00 | 20:00 | 3:00 | Report Writing |
| 15 | 12/14/18 | 13:30 | 15:00 | 1:30 | Meeting with Dr Kinsman |
| | 12/15/18 | 17:00 | 20:30 | 3:30 | Changes to report |
| | | | | **168:15** | Completed Independent Study |

## REFERENCES


[1] [n. d.]. 20 Newsgroups. ([n. d.]). http://qwone.com/~jason/20Newsgroups/ (Accessed on 12/01/2018).

[2] [n. d.]. Advantages and Disadvantages of Tableau - AbsentData. https://www.absentdata.com/advantages-and-disadvantages-of-tableau/. ([n. d.]). (Accessed on 11/28/2018).

[3] [n. d.]. Companies House API. ([n. d.]). https://developer.companieshouse.gov.uk/api/docs/index.html (Accessed on 09/14/2018).

[4] [n. d.]. Creating Attractive and Informative Map Visualisations in Python with Basemap - Data Dependence. http://www.datadependence.com/2016/06/creating-map-visualisations-in-python/. ([n. d.]). (Accessed on 11/28/2018).

[5] [n. d.]. The Gender Pay Gap Explained. ([n. d.]). https://gender-pay-gap.service.gov.uk/public/assets/pdf/gender-pay-gap-explained.pdf (Accessed on 09/14/2018).

[6] [n. d.]. GitHub - Jefferson-Henrique/GetOldTweets-python: A project written in Python to get old tweets, it bypass some limitations of Twitter Official API. https://github.com/Jefferson-Henrique/GetOldTweets-python. ([n. d.]). (Accessed on 11/28/2018).

[7] [n. d.]. International Women's Day — UNDP. http://www.undp.org/content/undp/en/home/news-centre/speeches/2018/international-womens-day.html. ([n. d.]). (Accessed on 11/28/2018).

[8] [n. d.]. Introduction to tweepy, Twitter for Python - Python Central. https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/. ([n. d.]). (Accessed on 11/28/2018).

[9] [n. d.]. matplotlib basemap toolkit fi!? Basemap Matplotlib Toolkit 1.1.0 documentation. https://matplotlib.org/basemap/api/basemap_api.html. ([n. d.]). (Accessed on 11/28/2018).

[10] [n. d.]. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. https://www.bls.gov/cps/cpsaat39.htm. ([n. d.]). (Accessed on 11/28/2018).

[11] [n. d.]. ODL Studio v1.4.1 - released 3rd November 2017. ([n. d.]). https://www.opendoorlogistics.com/downloads/ (Accessed on 09/14/2018).

[12] [n. d.]. plotly. ([n. d.]). https://plot.ly/python/ (Accessed on 09/14/2018).

[13] [n. d.]. seaborn: statistical data visualization. ([n. d.]). https://seaborn.pydata.org/ (Accessed on 09/14/2018).

[14] [n. d.]. Search gender pay gap data. ([n. d.]). https://gender-pay-gap.service.gov.uk/ (Accessed on 09/14/2018).

[15] [n. d.]. The Simple Truth about the Gender Pay Gap. ([n. d.]). https://www.aauw.org/research/the-simple-truth-about-the-gender-pay-gap/ (Accessed on 11/28/2018).

[16] [n. d.]. Stacked Bar Graph. ([n. d.]). https://datavizcatalogue.com/methods/stacked_bar_graph.html (Accessed on 09/14/2018).

[17] [n. d.]. Streaming With Tweepy fi!? tweepy 3.6.0 documentation. http://docs.tweepy.org/en/v3.6.0/streaming_how_to.html. ([n. d.]). (Accessed on 11/28/2018).

[18] [n. d.]. Time's Up. ([n. d.]). https://www.timesupnow.com/home (Accessed on 12/01/2018).

[19] 2016. Are there cases where PCA is more suitable than t-SNE? (Dec 2016). https://stats.stackexchange.com/questions/238538/are-there-cases-where-pca-is-more-suitable-than-t-sne/249520 (Accessed on 12/01/2018).

[20] 2018. New Law In Iceland Aims At Reducing Country's Gender Pay Gap. (Jan 2018). https://www.npr.org/2018/01/05/576082449/new-law-in-iceland-aims-at-reducing-countrys-gender-pay-gap (Accessed on 09/14/2018).

[21] Ashley Armstrong. 2017. Third of British bosses do not believe gender pay gap is business issue. (Apr 2017). https://www.telegraph.co.uk/business/2017/04/06/third-british-bosses-do-not-believe-gender-pay-gap-business/ (Accessed on 12/01/2018).

[22] arunbabu. 2017. gender by names - dataset by arunbabu. (Jul 2017). https://data.world/arunbabu/gender-by-names (Accessed on 09/14/2018).

[23] Prachi Bhardwaj. 2018. Silicon Valley's pay gap between women and men in tech is so wide it increases the national average. (May 2018). https://www.businessinsider.com/silicon-valley-pay-gap-lower-national-average-2018-5

(Accessed on 09/14/2018).

[24] Michelle Budig. 2014. The Fatherhood Bonus and The Motherhood Penalty: Parenthood and the Gender Gap in Pay fi?! Third Way. (Sep 2014). https://www.thirdway.org/report/the-fatherhood-bonus-and-the-motherhood-penalty-parenthood-and-the-gender-gap-in-pay

[25] Shawn M. Carter. 2018. In the US, unlike Iceland, itfis still OK to pay women lessfi!?herefis why. https://www.cnbc.com/2018/01/09/in-the-us-unlike-iceland-its-still-ok-to-pay-women-less-heres-why.html. (January 2018). (Accessed on 11/28/2018).

[26] National Women's Law Center. 2018. The Wage Gap: The Who, How, Why, and What To Do — NWLC. https://nwlc.org/resources/the-wage-gap-the-who-how-why-and-what-to-do/. (October 2018). (Accessed on 11/28/2018).

[27] M Chamberlain. 2016. Demystifying the Gender Pay Gap: Evidence from Glassdoor Salary Data. (2016). (Accessed on 11/28/2018).

[28] U.S. Equal Employment Opportunity Commission. [n. d.]. The Equal Pay Act of 1963 (EPA). https://www.eeoc.gov/laws/statutes/epa.cfm. ([n. d.]). (Accessed on 11/28/2018).

[29] Kristy Dorsey. 2018. UK gender pay gap reporting - what we know so far - Business Insider. https://www.insider.co.uk/news/gender-pay-gap-reporting-uk-11615242. (April 2018). (Accessed on 11/28/2018).

[30] Alexander Fabisch. 2014. t-SNE in scikit learn. (Jun 2014). http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html (Accessed on 12/01/2018).

[31] MADELINE FARBER. 2017. Wage Gap: 3 Big Reasons It Still Exists on Equal Pay Day — Fortune. http://fortune.com/2017/04/03/equal-pay-day-2017-wage-gap/. (April 2017). (Accessed on 11/28/2018).

[32] Jane Farrell and Sarah Jane Glynn. 2013. What Causes the Gender Wage Gap? - Center for American Progress. https://www.americanprogress.org/issues/economy/news/2013/04/09/59658/what-causes-the-gender-wage-gap/. (April 2013). (Accessed on 11/28/2018).

[33] Nic Fleming. 2018. How the gender pay gap permeates science and engineering — New Scientist. https://www.newscientist.com/article/mg23731670-100-how-the-gender-pay-gap-permeates-science-and-engineering/. (February 2018). (Accessed on 11/28/2018).

[34] World Economic Forum. 2017. The Global Gender Gap Report. http://www3.weforum.org/docs/WEF_GGGR_2017.pdf. (2017). (Accessed on 11/28/2018).

[35] Kimiko De Freytas-tamura. 2018. BBC, Criticized Over Pay Gap, Cuts Salaries of Some Male Journalists. (Jan 2018). https://www.nytimes.com/2018/01/26/business/media/bbc-pay-gap.html (Accessed on 12/01/2018).

[36] Olivier Gaudard. 2017. Percent stacked barplot. (Dec 2017). https://python-graph-gallery.com/13-percent-stacked-barplot/ (Accessed on 12/10/2018).

[37] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.*

[38] Sarah Gray. 2018. Paying Women Less Than Men Is Illegal. (Jan 2018). http://fortune.com/2018/01/02/illegal-to-pay-men-more-than-women-iceland/ (Accessed on 11/28/2018).

[39] Heidi Hartmann, Jeffrey Hayes, and Jennifer Clark. 2014. How equal pay for working women would reduce poverty and grow the American economy. *Washington, DC.: Institute for Womenfis Policy Research, Briefing paper (IWPR# C411). Retrieved on January* 10 (2014), 2016.

[40] howarder. 2018. Gender by Name - dataset by howarder. (May 2018). https://data.world/howarder/gender-by-name (Accessed on 09/14/2018).

[41] Maria Jovanovifj. 2017. Picture This – Girl Power – Finance & Development, March 2017. https://www.imf.org/external/pubs/ft/fandd/2017/03/picture.htm. (March 2017). (Accessed on 11/28/2018).

[42] David Kane. 2018. David Kane: An analysis of the gender pay gap in charities. (Apr 2018). https://www.civilsociety.co.uk/voices/david-kane-an-analysis-of-the-gender-pay-gap-in-charities.html (Accessed on 09/14/2018).

[43] Keita Kurita. 2018. Paper Dissected: "Visualizing Data using t-SNE" Explained. (Sep 2018). http://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-t-sne-explained/ (Accessed on 12/01/2018).

[44] Marrrcin. 2017. Sentiment analysis of tweets with Python, NLTK, word2vec and scikit-learn. (May 2017). https://zablo.net/blog/post/twitter-sentiment-analysis-python-scikit-word2vec-nltk-xgboost (Accessed on 12/10/2018).

[45] Kevin Miller. 2018. The Simple Truth about the Gender Pay Gap (Fall 2018). *AAUW. org* (2018). (Accessed on 11/28/2018).

[46] Moneywatch. 2017. Global gender gap grows for first time in 11 years - CBS News. https://www.cbsnews.com/news/global-gender-gap-grows-for-first-time-in-11-years/. (November 2017). (Accessed on 11/28/2018).

[47] NCPE. [n. d.]. Equal Pay Day. https://www.pay-equity.org/day.html. ([n. d.]). (Accessed on 11/28/2018).

[48] Tim Pateman. 2010. *Rural and urban areas: comparing lives using rural/urban classifications.* Office for National Statistics. (Accessed on 09/14/2018).

[49] Paygaphack. 2018. paygaphack/mentors-repo. (Jun 2018). https://github.com/paygaphack/mentors-repo/tree/master/companies-house-api (Accessed on 11/28/2018).

[50] Laura Tyson and Saadia Zahadi. 2014. Why everyone benefits from closing the gender gap. In *World Economic Forum.*

[51] Alanna Vagianos. 2015. 10 Things We Need To Do To Close The Wage Gap HuffPost. https://www.huffingtonpost.com/2015/04/14/things-we-need-to-do-to-close-wage-gap_n_7056322.html. (April 2015). (Accessed on 11/28/2018).

[52] Laurens van der Maaten. 2018. Do's and Don'ts of using t-SNE to Understand Vision Models. (Jun 2018). http://deeplearning.csail.mit.edu/slide_cvpr2018/laurens_cvpr18tutorial.pdf (Accessed on 12/01/2018).

[53] Danielle Wiener-Bronner. 2018. Starbucks achieves pay equity in the United States. https://money.cnn.com/2018/03/21/news/companies/starbucks-pay-equity/index.html. (March 2018). (Accessed on 11/28/2018).

[54] Chuan Xu. 2018. Why is Dimensionality Reduction so Important? (Jun 2018). https://medium.com/@cxu24/why-dimensionality-reduction-is-important-dd60b5611543 (Accessed on 12/01/2018).

Received May 2018