Carlos Ruiz, Paris Lee, Daniel Purrier, Neyda Morales

Professor Alexander Booth

Data Analytics Bootcamp (Project 1)

June 17, 2024

Diversity in Tech

Thesis

Companies at large have seen more of a push for diversity, equity, and inclusion in the workplace. Since the workforce is constantly changing, the analysis of this dataset allowed us to explore whether racial and gender diversity is reflected in the tech industry. After completing the analysis, the findings showed that the male population was about twice the size of the female population over a 5-year period. The percentage of Latino and Black people remained consistently below 10% while the percentage of Asian people was maintained at 20% and rose slightly. The findings also suggested that more women in a tech workplace could indicate more racial diversity within the company as well. Furthermore, this paper will discuss the implications of these findings, the limitations to the dataset, and the future work that could enhance this study.

Data Cleaning

The data chosen for this project is "Diversity in Tech Companies" from Kaggle Datasets (https://www.kaggle.com/datasets/jainaru/diversity-in-tech-companies). It provides a breakdown of employee demographics, focusing on racial and gender diversity in technology companies from 2014 to 2018. It offers percentages of female and male employees, as well as the percentage of different ethnic backgrounds within tech giants such as Google, Apple, Cisco, Amazon, and Microsoft.

```
# Examine data and datatypes
print(diversity_raw_df.shape)
print(diversity_raw_df.info())
print(diversity_raw_df.describe())
```

```
(94, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 94 entries, 0 to 93
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Year          94 non-null     int64
 1   Company       94 non-null     object
 2   Female %      94 non-null     int64
 3   Male %        94 non-null     int64
 4   % White       94 non-null     int64
 5   % Asian       94 non-null     object
 6   % Latino      94 non-null     object
 7   % Black       94 non-null     object
 8   % Multi       94 non-null     object
 9   % Other       93 non-null     object
 10  % Undeclared  94 non-null     object
dtypes: int64(4), object(7)
memory usage: 8.2+ KB
None
              Year   Female %     Male %    % White
count    94.000000  94.000000  94.000000  94.000000
mean   2016.106383  35.234043  64.744681  59.393617
std       1.432856   9.446426   9.464065   9.897559
min    2014.000000  16.000000  46.000000  37.000000
25%    2015.000000  29.000000  57.250000  53.000000
50%    2016.000000  33.000000  67.000000  60.000000
75%    2017.000000  42.750000  71.000000  66.500000
max    2018.000000  54.000000  84.000000  79.000000
```

The raw data is simple in nature as it contains 11 columns:

- **Year:** 5 years' worth of data, from 2014 to 2018 inclusive.
- **Company:** A total of 23 unique companies in the raw dataset. Most companies are represented in all years; however, some companies like Uber, Slack, AirBnB, Netflix and Yelp only have data for 1 or a few years.

- **Male and Female %:** Representation of 100% of the population, broken down by gender, at each company during the corresponding year.
- **% White, Asian, Latino, Black, Multi, Other and Undeclared:** Representation of 100% of the population, broken down by the racial form each employee filled out during hiring process.

To prepare the data for the purposes of this project's analysis, the following steps were taken:

- Removal of "Apple (excluding undeclared)" rows
- Conversion of "<1" on "% Other" column to 0
- Conversion of "-" on all columns to 0
- Conversion of NaN on "% Other" column to 0
- Conversion of all percentage columns to floats

Removal of Apple (excluding undeclared)

Apple offered 2 sets of rows on the first 3 years of data. One row was "Apple" which included all employees, and the second row was "Apple (excluding undeclared)". As the analysis of this project included undeclared, it was decided to remove "Apple (excluding undeclared)" to avoid Apple from being counted twice on all other percentages.

Conversion of "<1" on "% Other" column to 0

The data provided by the companies was a percentage rounded to the nearest full number. Since full numbers were used, any values less than 1 percent wouldn't impact the analysis; therefore, all data provided with "<1" were converted to zero.

Conversion of "-" on all columns to 0

It is difficult to understand what "-" represents, but the assumption was that the company didn't have any personnel that would meet the criteria, or no data was provided for the specific diversity. Converting "-" to zeros helped to retain the data of the other columns, and it was decided that the impact on the data would not make significant changes to the result of the analysis.

Conversion of NaN on "% Other" column to 0

One row had a NaN on the "% Other" column. Converting the NaN to zero helped retain the rest of the information provided in that row. Also, the change would not impact significantly on the results of the analysis.

Conversion of all percentage columns to floats

Most of the diversity columns were converted to float types to facilitate the analysis of the data.

Clean Dataset

Given the data preparation, the analysis was done on 11 columns and 91 rows. All percentage categories, including gender and diversity, were floats. Other interesting information included a normal distribution across all categories with no outliers.

```
print(df.shape)
print(df.info())
```

```
(91, 13)
<class 'pandas.core.frame.DataFrame'>
Index: 91 entries, 0 to 93
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Year                91 non-null     int64
 1   Company             91 non-null     object
 2   Female %            91 non-null     float64
 3   Male %              91 non-null     float64
 4   % White             91 non-null     float64
 5   % Asian             91 non-null     float64
 6   % Latino            91 non-null     float64
 7   % Black             91 non-null     float64
 8   % Multi             91 non-null     float64
 9   % Other             91 non-null     float64
 10  % Undeclared        91 non-null     float64
 11  Gender Total %      91 non-null     float64
 12  Diversity Total %   91 non-null     float64
dtypes: float64(11), int64(1), object(1)
memory usage: 10.0+ KB
None
```

To validate the data, aggregations of gender and diversity were added. The expectation was that they would add up to 100%; however, there were some rows that would not be exactly 100%. This is attributed to the rounding of the numbers provided by companies.

```
df.describe()
```

	Year	Female %	Male %	% White	% Asian	% Latino	% Black	% Multi	% Other	% Undeclared	Gender Total %	Diversity Total %
count	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000	91.000000
mean	2016.142857	35.406593	64.571429	59.208791	22.637363	7.263736	5.461538	2.329670	1.241758	0.428571	99.978022	98.571429
std	1.434274	9.553333	9.570955	9.905685	11.994922	4.057463	4.316139	3.283139	1.344459	1.795939	0.209657	8.306815
min	2014.000000	16.000000	46.000000	37.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	99.000000	54.000000
25%	2015.000000	29.000000	57.000000	52.500000	12.500000	4.000000	2.000000	0.000000	0.000000	0.000000	100.000000	100.000000
50%	2016.000000	36.000000	64.000000	60.000000	22.000000	6.000000	4.000000	1.000000	1.000000	0.000000	100.000000	100.000000
75%	2017.000000	43.000000	71.000000	66.000000	31.000000	9.000000	8.000000	3.000000	2.000000	0.000000	100.000000	100.000000
max	2018.000000	54.000000	84.000000	79.000000	45.000000	19.000000	21.000000	14.000000	5.000000	13.000000	101.000000	103.000000

Research Questions

1. *Has diversity in tech companies increased over time?*

To assess this question, we first made a line graph out of the data in the Pandas DataFrame, with "Years" on the x-axis and "Employment in Tech (%)" on the y-axis, using the column data in the "Year" column and the data in the columns of the various racial groups – "Asian", "Black", "Latino", "Multi", "Other", "Undeclared", and "White" – to create the plot (Figure 1).

According to the data seen in the line graph, we can easily deduce that the employment rate in the tech workforce for people belonging to the racial groups "Black," "Latino," "Multi," "Other," and "Undeclared" remained relatively unchanged from 2014 to 2018. We start to see significant movement within this time period when we assess the employment rates for Asian people and for White people. Interestingly enough, the employment rates for these two groups seem to move in an almost inverse fashion. From 2014 to 2016, the employment rates for both Asian and White people stayed relatively constant. From 2016 to 2017, however, the aforementioned inverse movement between these two groups is most observable, with a visually-estimated 5% increase in employment for the Asian population, and a visually-estimated 5% decrease in employment for White people in the tech workforce. From 2017 to 2018, the employment rates for these two groups remained constant. Figure 1 also allows us to clearly assess which racial group saw the most employment growth over the five-years this data was collected, and that was Asian people.

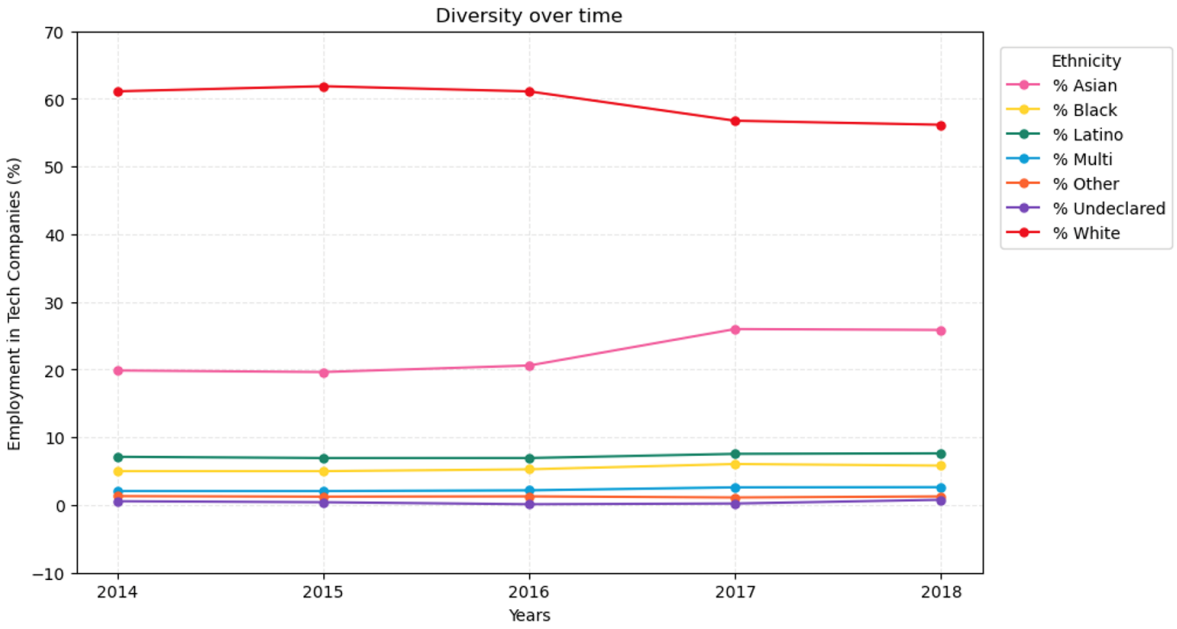


Figure 1. A line graph illustrating the employment rates of various racial groups in tech companies from 2014 to 2018.

```python
# Plot racial diversity over time
# 1. Get the data
pivot_df = df.pivot_table(index='Year', values=['% White', '% Asian', '% Latino', '% Black', '% Multi', '% Other', '% Undeclared'], aggfunc='mean')

# 2. Make the canvas
plt.figure(figsize=(10, 6))

# 3. Plot the data
for column in pivot_df.columns:
    plt.plot(pivot_df.index, pivot_df[column], label=column, marker='o', linestyle='-', markersize=5, color=diversity_colors[column])

plt.title('Diversity over time')
plt.xlabel('Years')
plt.ylabel('Percentage')
plt.xticks(pivot_df.index)
plt.yticks(range(-10, 71, 10))
plt.legend(title='Ethnicity', loc=(1.02, 0.6))
plt.grid(linestyle='--', alpha=0.5, color='lightgray')

# 4. Show/Save the plot
plt.savefig("./images/DiversityOverTime.png")
plt.show()
```

Figure 1.1. Code for the “Diversity over Time” graph written in Python and displayed using VS Code.

2. *What is the percentage of Males vs Females in the tech workplace over time?*

To answer this question, we created a copy of the cleaned data as displayed below. This prevented the risk of unintentionally changing anything in the original data set.

```
daniel_df = df.copy()
daniel_df
```

	Year	Company	Female %	Male %	% White	% Asian	% Latino	% Black	% Multi	% Other	% Undeclared	Gender Total %	Diversity Total %
0	2018	Yahoo!	37.0	63.0	45.0	44.0	4.0	2.0	2.0	3.0	0.0	100.0	100.0
1	2018	Google	31.0	69.0	53.0	36.0	4.0	3.0	4.0	0.0	0.0	100.0	100.0
2	2018	Apple	32.0	68.0	54.0	21.0	13.0	9.0	3.0	1.0	2.0	100.0	103.0
3	2018	Cisco	24.0	76.0	53.0	37.0	5.0	4.0	1.0	0.0	0.0	100.0	100.0
4	2018	eBay	40.0	60.0	50.0	39.0	6.0	3.0	1.0	1.0	0.0	100.0	100.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	2014	Groupon	47.0	53.0	71.0	15.0	5.0	4.0	0.0	4.0	0.0	100.0	99.0
90	2014	Amazon	37.0	63.0	60.0	13.0	9.0	15.0	0.0	3.0	0.0	100.0	100.0
91	2014	Salesforce	29.0	71.0	67.0	22.0	4.0	2.0	2.0	3.0	0.0	100.0	100.0
92	2014	Pandora	49.0	51.0	71.0	12.0	7.0	3.0	6.0	1.0	0.0	100.0	100.0
93	2014	Microsoft	29.0	71.0	61.0	29.0	5.0	4.0	1.0	1.0	0.0	100.0	101.0

As seen above, there are quite a few columns in the complete data set. To answer our research question, however, only 3 columns are needed. Those are the Year, Female %, and Male % columns. The “groupby” and “mean” functions were used to remove all unnecessary columns and to make a new data set with what was needed.

```
daniel_df.groupby(["Company","Year"])
average_male = daniel_df.groupby('Year')['Male %'].mean()

average_male_df = pd.DataFrame(average_male).reset_index()

average_male_df
```

	Year	Male %
0	2014	65.937500
1	2015	64.882353
2	2016	64.722222
3	2017	65.000000
4	2018	62.863636

```
average_female = daniel_df.groupby('Year')['Female %'].mean()

average_female_df = pd.DataFrame(average_female).reset_index()

average_female_df
```

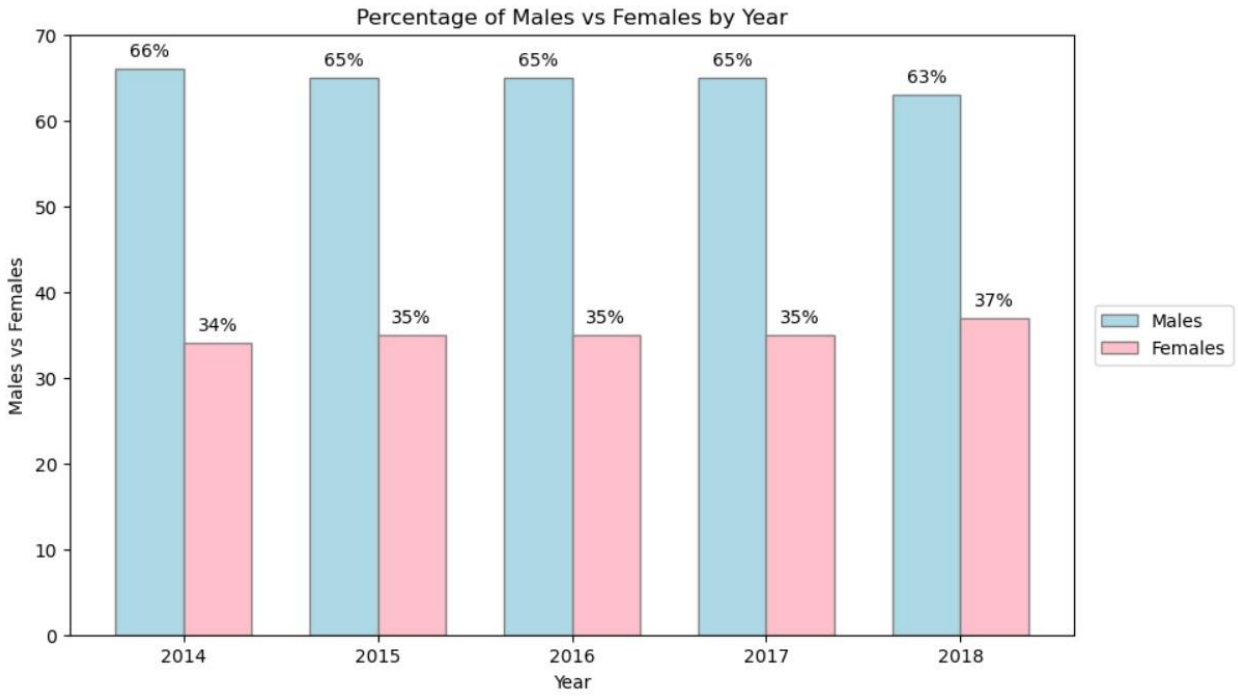
	Year	Female %
0	2014	34.000000
1	2015	35.058824
2	2016	35.222222
3	2017	35.000000
4	2018	37.181818

Finally, these two data sets were merged together and the values were rounded to clean any data that had large amounts of numbers after the decimal place. See the final data set below.

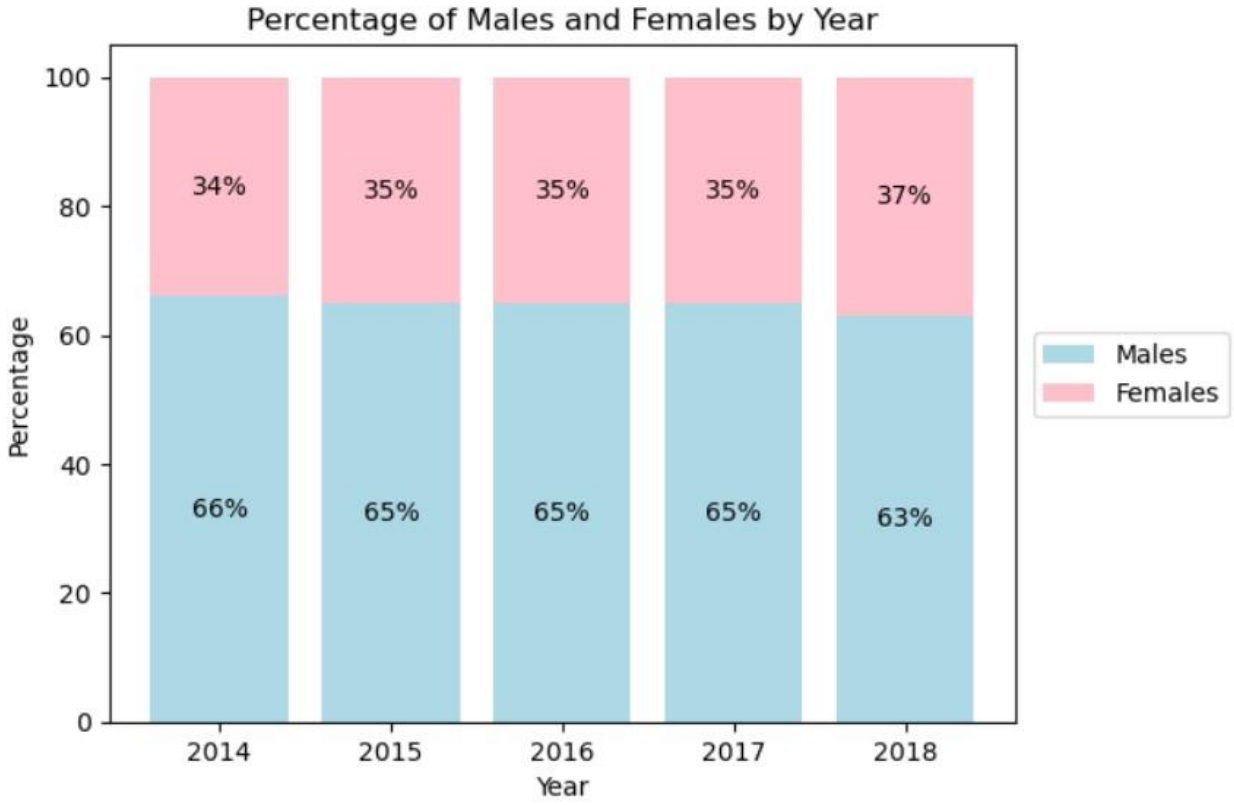
```
average_merge = pd.merge(average_male_df, average_female_df, on ='Year').round()
average_merge
```

	Year	Male %	Female %
0	2014	66.0	34.0
1	2015	65.0	35.0
2	2016	65.0	35.0
3	2017	65.0	35.0
4	2018	63.0	37.0

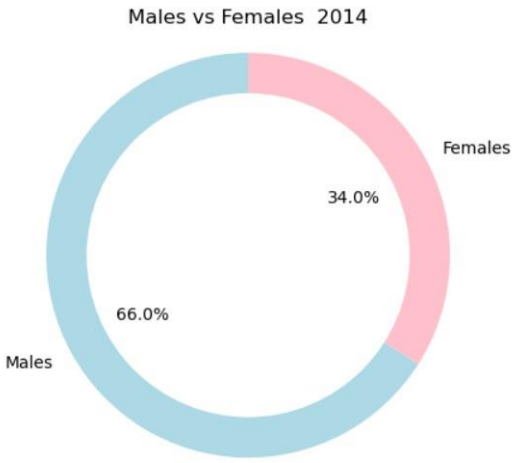
With this leaderboard created and rounded, the below visuals were constructed to further tell the story. The leaderboard was paired with a bar graph which *very* clearly showed that the male percentage versus female percentage populations at these companies was heavily in favor of the males.



A stacked bar chart of this information was also made. However, as a group, we elected to not include this in the presentation as we favored how the information was depicted in the standard bar chart versus the stacked chart. Please see our stacked chart below.



While the bar charts did a great job of depicting the information, we took a deeper dive into the data to more clearly paint the picture of the changes in percentages over the years. Multiple donut charts were used to accomplish this. This first donut chart was for the year 2014. It shows the greatest discrepancy in the male versus female population for our data set with the male percentage at its highest of 66% and the female percentage at 34%.



The next donut chart shows the average percentage change for all years in the data set.

Additional coding was required to accomplish this and was necessary to make the next pie chart.

```
males_average = average_merge["Male %"].mean()
males_average
```

```
64.8
```

```
females_average = average_merge["Female %"].mean()
females_average
```

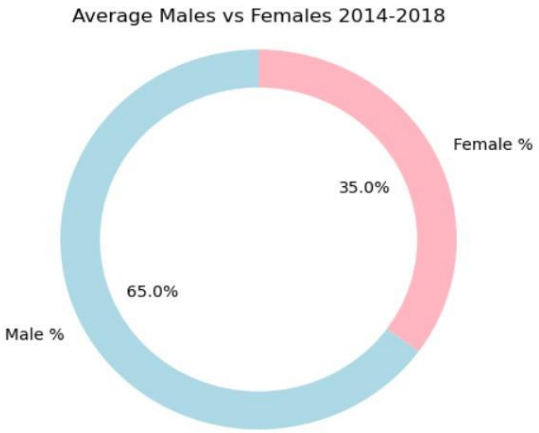
```
35.2
```

The final numbers for these averages were also rounded and placed into another leaderboard.

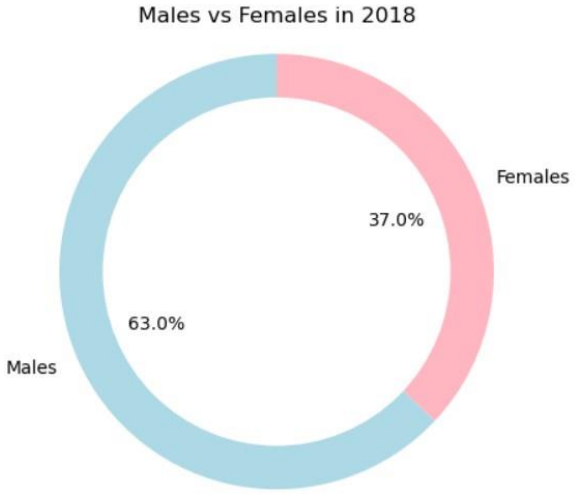
```
data = {
    "Male %": [males_average],
    "Female %": [females_average]
    }

total_df = pd.DataFrame(data).round()

total_df
```

	Male %	Female %
0	65.0	35.0

From the above leaderboard, our next donut chart was made.



As can been seen, the average amount of males versus females over the five years is higher than when it started in 2014, indicating an increase in the female population. Without these donut charts and coding, we would have never known that the average percentages over the five years were already higher than in the year 2014, where this all began. This increase is further reflected by the percentages in the final year (63% male and 37% female).



While these percentages are still heavily in favor of the males, we can see there was a shift in the percentages, and the result is an increase in the female population and a decrease in the male population over the years—thus answering our second research question.

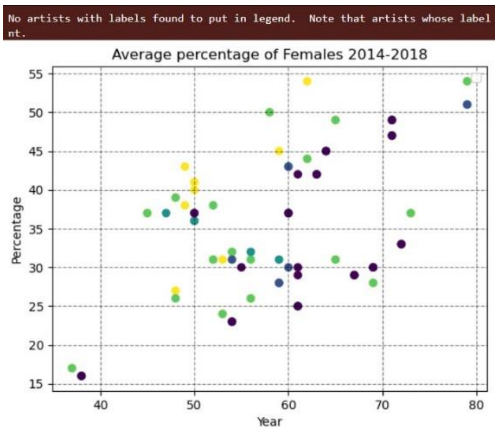
This analysis drove us to a sub-question: why is there such a discrepancy between the male and female populations in the tech industry? This led to some further research that brought us to the article, "The Gender Gap in Tech… Let's Talk About It". In this article, the author Juliette Carreiro, a tech writer, goes into detail about the many different gender obstacles faced in the tech industry. She also covers the issues faced in the tech industry across multiple different countries such as the United Kingdom, the United States, Spain, Portugal, and many others. While multiple countries are covered in the article, it is clear that each of them is affected by the same main issues. These issues are as follows:

- **Sexism and Gender Discrimination:** In this section, the author makes reference to both women and men being heavily influenced career-wise by what society teaches and expects. For example, society tends to perpetuate that jobs, like nurses, secretaries, receptionists and others are careers for women to pursue, while positions like tech and construction worker (among others) are for males.
- **Role models/ Leadership roles:** Women are lacking other women being in leadership roles in the industry, and as a result, the connection and motivation that comes with seeing others of the same sex doing what you are doing and leading others is not there. Juliette reports that just 40% of first-level managers are women, and that number continues to decline as the leadership roles gain more importance.
- **Harsh environment:** Women often report feeling isolated, not having their opinions heard, and experiences with mockery.

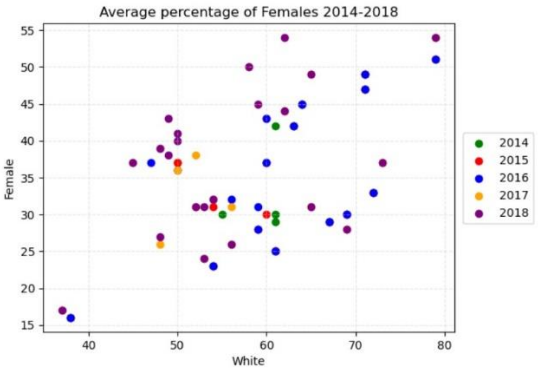
- **Not being informed:** A survey was done for men and women ages 18-28. 44% of women in this group report never receiving any type of information related to the careers in tech while only 33% of men report the same.

Upon learning about all the challenges women face in the tech industry, it is now much easier to understand why it is such a male-dominated field. If we want to see an increase in women in this industry -and Juliette reports that such an increase would do well to improve the economy- we must do what is necessary to remove these extra challenges that women face when deciding to pursue a career in tech.

Finally, it is worth mentioning that a few more visuals did not make it into our presentation. We originally wanted to make a regression line that showed the relationship of women being hired versus the year throughout all companies. However, due to the data showing only percentages of women hired and not a specific number, we determined this could not be done. These charts are below.



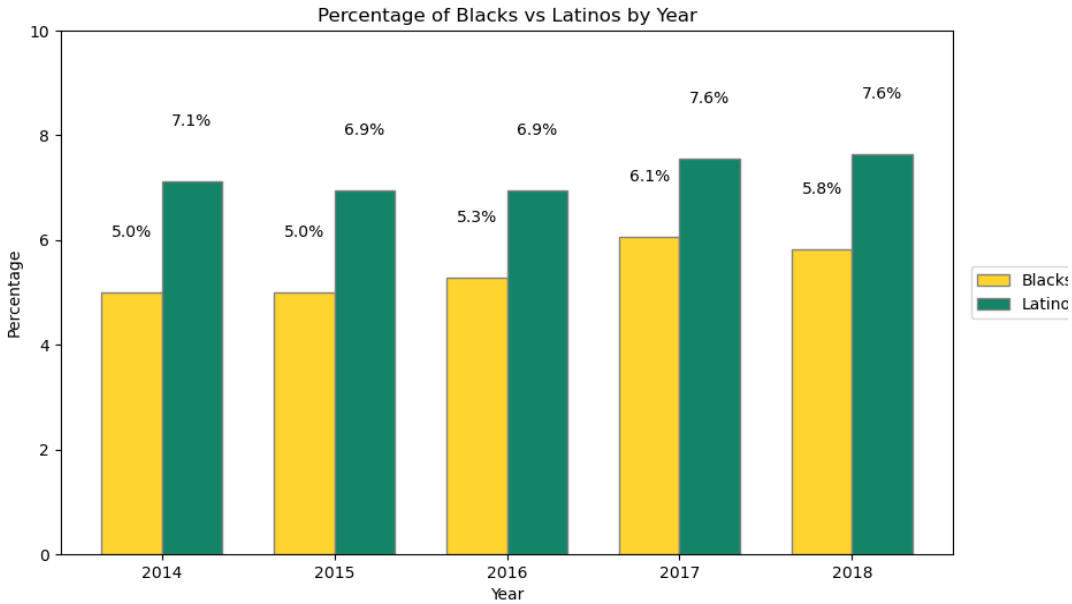
As you can also see, there is no legend on this chart. A good amount of time was spent trying to figure out how to do so, and eventually got it to display. Once again, these charts never made it into our presentation as ultimately they did not depict the information we wanted.




```
#plt.scatter(x, y, c = df15['Year'])
plt.scatter(df14['% White'], df14['Female %'], c = 'green')
plt.scatter(df15['% White'], df15['Female %'], c = 'red')
plt.scatter(df16['% White'], df16['Female %'], c = 'blue')
plt.scatter(df17['% White'], df17['Female %'], c = 'orange')
plt.scatter(df18['% White'], df18['Female %'], c = 'purple')
plt.xlabel('White')
plt.ylabel('Female')
plt.title('Average percentage of Females 2014-2018')
plt.grid(linestyle='--', alpha=0.5, color='lightgray')
plt.legend([2014, 2015, 2016, 2017, 2018], loc = (1.02, 0.35))
#plt.ylim(30, 90)
#plt.xlim(10,60)
plt.show()
```

3. How do Black people compare to Latino people over time?

After Asian people, the next major diversity groups are Latino people and Black people. Based on the table below, from 2014 – 2016, the Black population stayed around 5% whereas the Latino population remained at around 7%. In 2017, both populations saw a small increase, with the Black population rising by about 1% from 5% to 6%, and the Latino population rising by about 0.5% from 7% to 7.6%. Both groups remaining about the same in 2018. As noted, these numbers are below 10% across all the companies in this dataset.



Why does this matter? According to an article by Govt Tech, despite the efforts of big tech companies for inclusion, people of color struggle to break into a workforce that has historically been dominated by white and Asian men. In 2014, Google finally agreed to publicly report its racial and gender breakdown of its workforce. “Of nearly 50,000 employees at Google in 2014, 83% were men, 60% were white, and 30% were Asian. Just 2.9% were Latino, and 1.9% Black.” The following year, Google promised to invest $150 million to increase diversity in the company. However, as seen in the analysis of this dataset, little progress was made between 2014 - 2018. The lack of diversity directly impacts the representation that can exist at the executive level as well as in the world of entrepreneurship and venture capitalism. This leads to a sector of the economy, which has created billions and reshaped California, that is virtually inaccessible to Black and Latino people. Moreover, the percentage of Latino and Black graduates with degrees in computer science, engineering, and math has risen signaling to the stagnant progress across tech companies. The lack of access and support, essentially the lack of industry relationships, is a dominant obstacle that Black and Latino professionals face.

Linear Modeling

1. *Are companies that hire more Women more likely to hire Latino people?*
2. *Are companies that hire more Women more likely to hire Black people?*

Efforts to start answering these two questions started with the creation of scatter plots (Figures 2 and 3) with a line of best fit, the equation of the line, and an r-value. We used the data in the column “Female %” for the x-axis, and the data in the columns “% Latinos” and “% Black” for the y-axes for the respective plots. The r-values for the first and second regression lines were 0.126 and 0.129, respectively. Albeit weak, there is a positive correlation between the percentage of a tech company’s workforce and their propensity to hire Latino and Black people.

While no definitive conclusions can be made from this data alone, it does hopefully allow us to tentatively propose that more women in a tech workplace could indicate more racial diversity within the company as well.

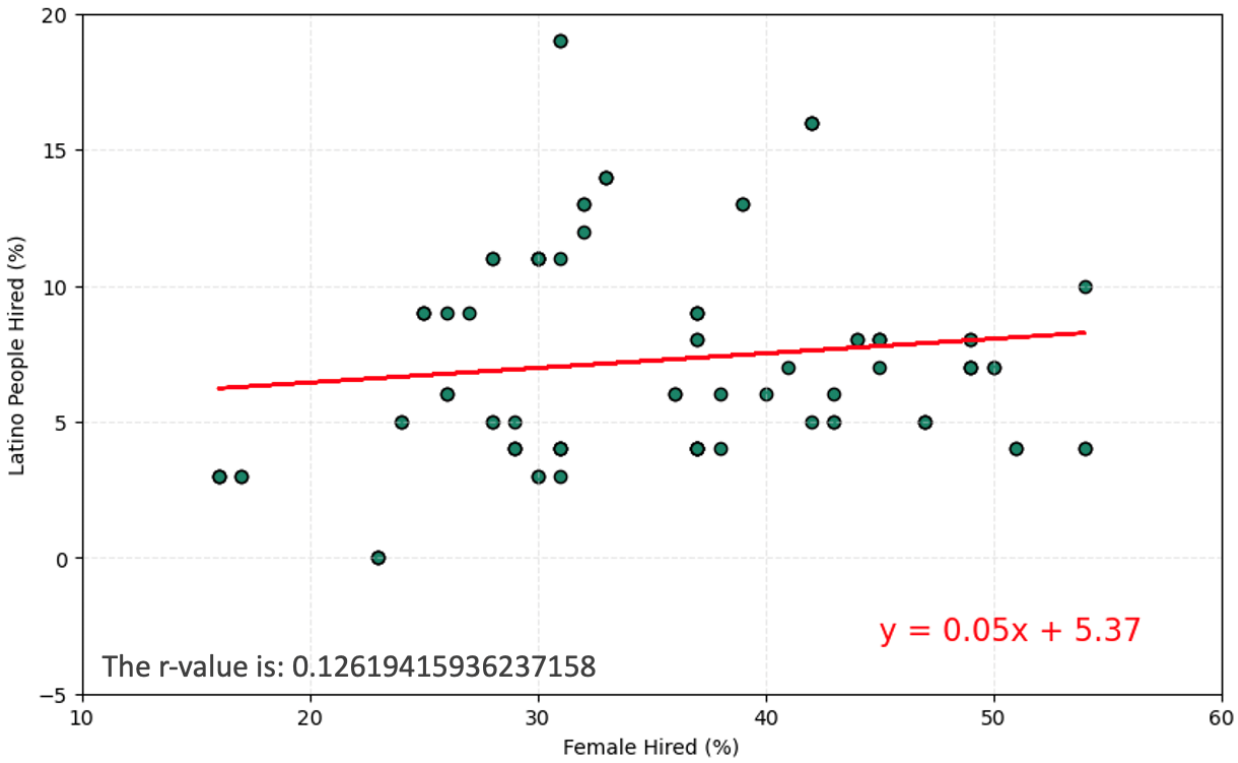


Figure 2. A scatter plot and a line of regression that assess the relationship between the percentage of a company's workforce that is women and the likelihood of said company to hire Latino people.

```python
# Define a function to create Linear Regression plots
def plot_linear_regression(gender, race, xlimits, ylimits, text_coordinates):

    # 1. Get the data
    x_values = df[f"{gender} %"]
    y_values = df[f"% {race}"]

    # Run regression on hemisphere weather data
    (slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)

    # Calculate the regression line "y values" from the slope and intercept
    regress_values = x_values * slope + intercept

    # Get the equation of the line
    line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(intercept,2))

    # 2. Make the canvas
    plt.figure(figsize=(10,6))

    # 3. Make the scatter plot
    # Create a scatter plot and plot the regression line
    plt.scatter(x_values, y_values, edgecolors='black', color=diversity_colors[f'% '+race])
    plt.plot(x_values,regress_values,"r")

    # Annotate the text for the line equation
    plt.annotate(line_eq, (text_coordinates), fontsize=15, color="red")
    plt.title(f'Percent {gender} vs {race} total people', fontsize=16, fontweight='bold', color='black')
    plt.xlabel(f'{gender} %')
    plt.ylabel(f'% {race}')
    plt.xlim(xlimits)
    plt.ylim(ylimits)
    plt.grid(linestyle='--', alpha=0.5, color='lightgray')

    print(f"The r-value is: {rvalue}")

    # 4. Save/Show the plot
    plt.savefig(f"images/{gender}_{race}_linregress.png")
    plt.show()
```

```python
# Linear regression on Percentage of Females vs. Percentage of Latinos
# 1. Get the data
gender = 'Female'
race = 'Latino'

plot_linear_regression(gender, race, (10,60), (-5,20), (45,-3))
```

Figures 2.1 and 2.2: Figure 2.1 establishes a function to create the linear regression plots. Figure 2.2 is the code for the first regression analysis written in Python and displayed in VS Code.

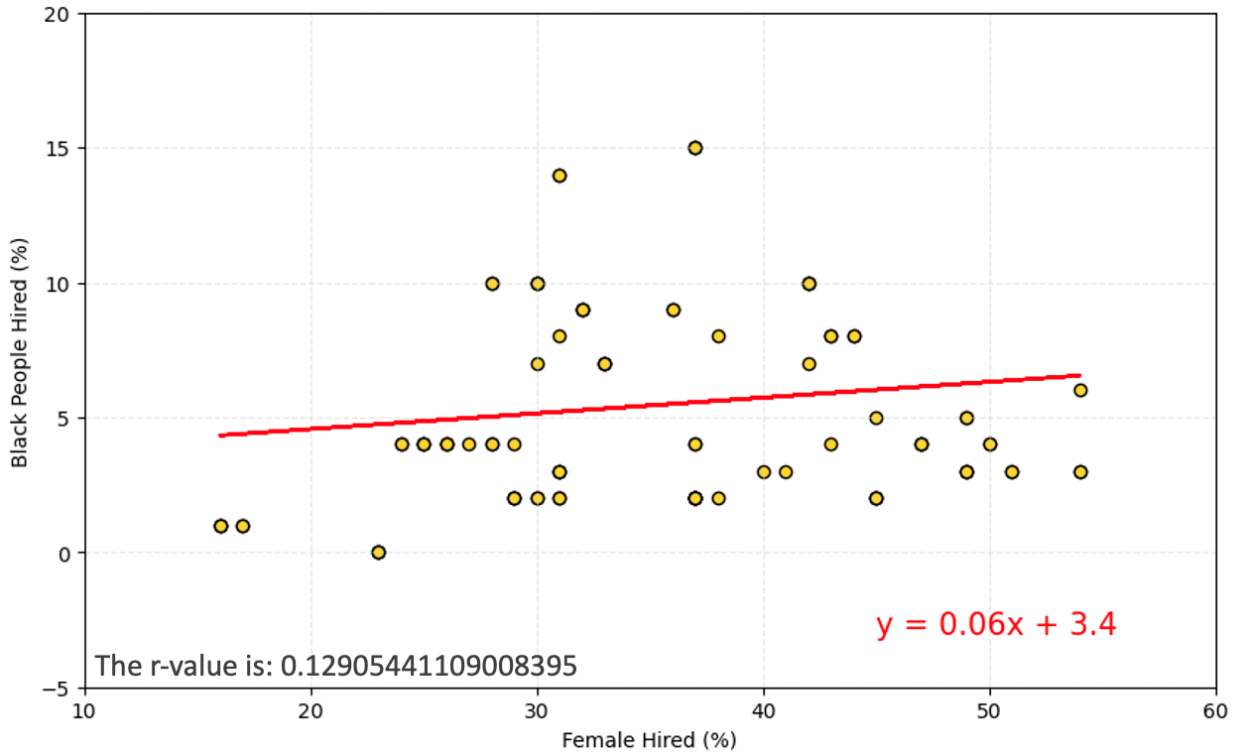


Figure 3. A scatter plot and a line of regression that assess the relationship between the percentage of a company's workforce that is women and the likelihood of said company to hire Black people.

```
# Linear regression on Percentage of Females vs. Percentage of Black
# 1. Get the data
gender = 'Female'
race = 'Black'

plot_linear_regression(gender, race, (10,60), (-5,20), (45,-3))
```

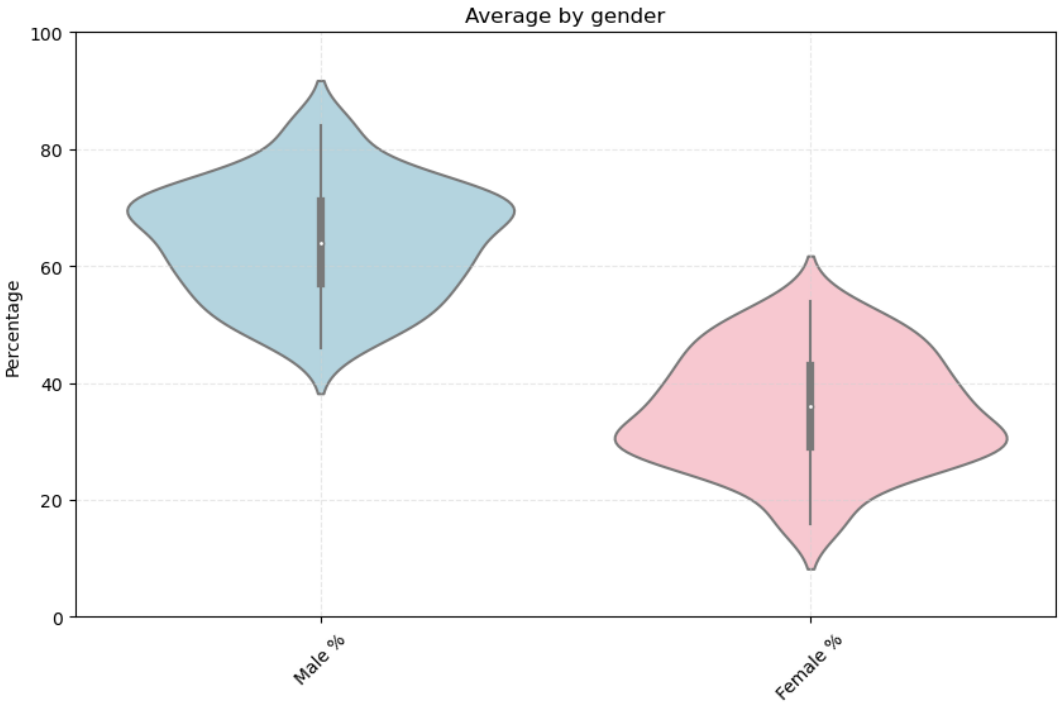
Figure 3.1. The code for the second regression analysis written in Python and displayed in VS Code.

Statistical Analysis

Gender and Diversity Statistical Analyses are provided as part of this project.

Gender Analysis

Gender average data shows a large difference between males and females during 2014 to 2018. However, a TTest was performed to prove this theory.



For Gender, the following full hypothesis was offered:

- Null hypothesis: There is no difference between the quantity of males vs females working in tech companies
- Alternative hypothesis: There is a difference between the prevalence of males vs females working in tech companies

The expectation is that there is a significant difference in the prevalence of males vs females working in tech companies.

Gender Analysis Results

A TTest was performed between both genders, and it provided the following result:

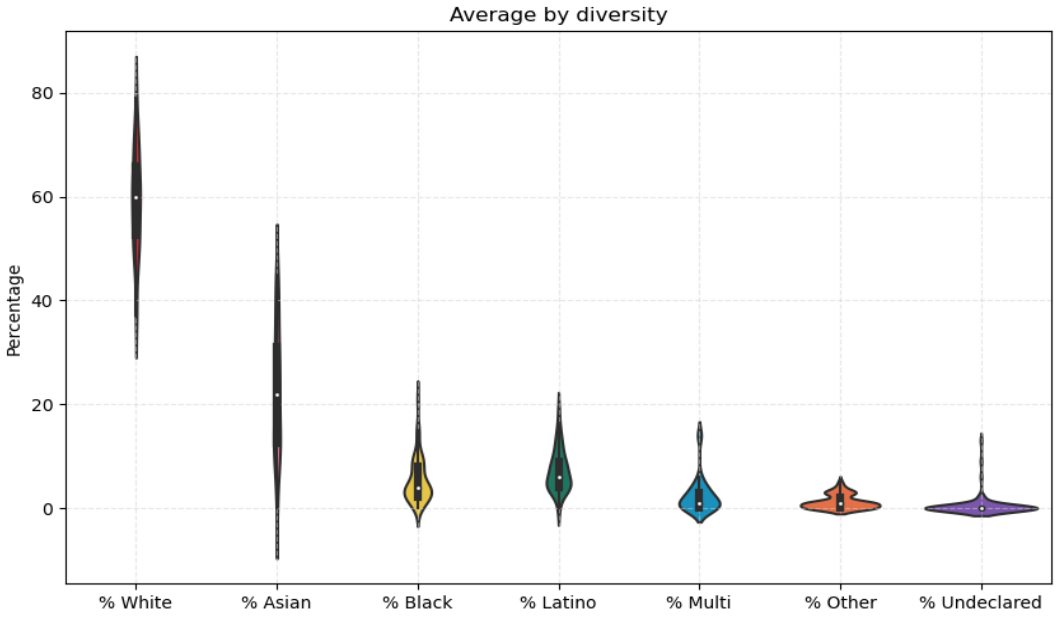
TtestResult(statistic=20.573574726611028, pvalue=3.78484545338512e-49, df=179.9993887209176)

The p-value is significantly small; therefore, we can validate that there is a difference between males vs females working on tech companies.

Based on the data, the null hypothesis is rejected, and the alternative hypothesis is accepted. The data supports the claim that there is a significant difference in the prevalence of males and female working in tech companies.

Diversity Analysis

Diversity average data shows a large difference between white people vs other diversities during 2014 to 2018. To analyze this data in more detail, ANOVA testing was done first, followed by conducting a TTest of all combinations of diversity.



For Diversity, the following full hypothesis is offered:

- Null hypothesis: There is no difference of averages between any diversity group
- Alternative hypothesis: is at least one difference between diversity groups

The expectation is that there is at least one difference in prevalences between diversity groups working in tech companies.

Diversity Analysis Results

ANOVA test was conducted between all diversity groups, and it provided the following results:

F_onewayResult(statistic=985.8325475532455, pvalue=2.47133054e-316)

The p-value demonstrates that there is at least one diversity group difference between all diversity groups.

Based on the data, the null hypothesis is rejected, and the alternative hypothesis is accepted. There is at least one diversity group that is different from the other groups. After performing TTests across all diversity groups, the data reflects that ALL groups are different from each other.

Diversity ttest

White vs Asian: TtestResult(statistic=22.426127199064165, pvalue=3.4974324485005776e-53, df=173.78741107942048)
White vs Black: TtestResult(statistic=47.45100949423269, pvalue=6.374334017920352e-81, df=122.98495278778121)
White vs Latino: TtestResult(statistic=46.29135565132039, pvalue=4.13282193180125e-78, df=119.3735241720709)
White vs Multi: TtestResult(statistic=51.99438750290772, pvalue=4.870119400677852e-79, df=109.537654889487)
White vs Other: TtestResult(statistic=55.316345330041564, pvalue=3.910652332452989e-73, df=93.31475419432466)
White vs Undeclared: TtestResult(statistic=55.69860802269435, pvalue=7.557267383459079e-75, df=95.9104106591678)
Asian vs Black: TtestResult(statistic=12.852924987195221, pvalue=7.644253536628321e-24, df=112.92177336927801)
Asian vs Latino: TtestResult(statistic=11.581759018419703, pvalue=8.827328477772503e-21, df=110.3300092458146)
Asian vs Multi: TtestResult(statistic=15.577446894105972, pvalue=6.992841235175281e-29, df=103.40989499464588)
Asian vs Other: TtestResult(statistic=16.909733828894062, pvalue=5.166342130749162e-30, df=92.26101951932218)
Asian vs Undeclared: TtestResult(statistic=17.467632811843643, pvalue=2.8276886801936257e-31, df=94.03313383302206)
Black vs Latino: TtestResult(statistic=-2.9021422691618346, pvalue=0.004171105579260181, df=179.31680824799332)
Black vs Multi: TtestResult(statistic=5.509231046627633, pvalue=1.3317013546634914e-07, df=168.02727152668047)
Black vs Other: TtestResult(statistic=8.904425131520638, pvalue=1.4956780083145585e-14, df=107.30242991242221)
Black vs Undeclared: TtestResult(statistic=10.270102429424778, pvalue=3.7853392676111054e-18, df=120.25779691951077)
Latino vs Multi: TtestResult(statistic=9.017914867341968, pvalue=3.608912970795619e-16, df=172.49074872237733)
Latino vs Other: TtestResult(statistic=13.439520313134684, pvalue=6.477281620974978e-25, df=109.52785001293975)
Latino vs Undeclared: TtestResult(statistic=14.69482678067052, pvalue=6.301597651996616e-29, df=123.96164151053163)
Multi vs Other: TtestResult(statistic=2.92523538481525, pvalue=0.004120717814921395, df=119.35924965926462)
Multi vs Undeclared: TtestResult(statistic=4.846109485972371, pvalue=3.3103904655276187e-06, df=139.43501659992066)
Other vs Undeclared: TtestResult(statistic=3.457792920018384, pvalue=0.0006911859728477218, df=166.76559172686433)

Bias and Limitations

The dataset consisted of a relatively small list of companies giving us a small dataset to analyze. The number of years included in the dataset was small with none being the most recent years leading up to 2024. The raw values were not presented in this dataset. We only had the percentage values. Company size was not included, and it was another factor we wanted to take into consideration for this analysis. None of the data was intersectional meaning we could not determine the percentage of females that were of Latino race, or the percentage of males that were of Black race. There were also values that were Undeclared which portrays the employees that choose not to answer questions regarding gender and race.

Future Work

Regarding further exploration and analysis of diversity in the tech workforce, we have several propositions that would provide a more thorough and clearer picture of the data, and hopefully increase the quality of subsequent analyses of the data.

- Pulling together data that spans a longer period of time would allow for the observation of more trends in the data, possible correlations of workforce trends with social events and social movements and allow for a look at the current employment trends.
- Incorporation of the sizes of the companies within the dataset would allow for us to extrapolate some of the raw data that we did not have in the original dataset, as we just had percentages. This would allow for the meaningful demonstration of the raw number of men and women at each company, as well as that for each racial category. It would also allow us to more accurately manipulate the data mathematically. The size of the company could also give context to some of the previously mentioned points, such as that

of reports of women not feeling respected, valued, and heard in their workplaces. The number of women at a company out of the total number of people that comprise that company's workforce could be quite emotionally evocative, particularly when paring it with a personal call to action.

- Gathering data that is more gender inclusive would allow us to be able to analyze not just racial diversity within tech workspaces, but also gender diversity. As it stands, our dataset encompasses just the gender binary, but it is unrealistic to assume that the only people in tech workplaces are those that fit neatly into the gender binary. A dataset that encompassed more gender options would more accurately reflect the world at large, as that is what any workforce population would ideally reflect: the population of the world at large.
- Capturing employee satisfaction data for each respective company would also give context to whether minority employees are happy in their respective companies. This has tremendous value for several reasons: 1. It would allow us to begin to assess whether minority employees are regarded fairly and evenly in their respective workplaces. 2. It would allow us to gauge the validity of various companies' diversity, equity, and inclusion efforts. Is the DEI statement of one company simply just a statement, or does it transcend words and manifest as meaningful, purposeful action and policy change? 3. It would allow us to track employee satisfaction over time, which would also help us to, in some way, qualitatively measure the diversity in the workplace over time, and possibly discern why the trends appear to be the way they are.

Call to Action

The purpose of this project was to assess the diversity in the tech workforce over the five-year period from 2014 to 2018 to see if the tech workforce diversity reflected the diversity of the world at large. Given our analysis, we proposed two main calls to action that could help to diversify the tech workplace.

The first call to action is tackling unconscious bias within the workplace and on individual teams. This will help to make for a healthier and more welcoming workplace. It will also help to decrease, and hopefully eliminate, the proportion of minority employees that leave their workplaces, or tech altogether, because of discrimination and hostility, thus helping to retain these employees, which will hopefully further promote more diversity in the future.

The second call to action would be to introduce tech into more schools in the form of workshops, camps, clubs, classes, and afterschool programs that are centered around technology. This will allow for an earlier introduction to the world of tech for kids, and it will also help to close the aforementioned information gap between men and women who are introduced to tech in some way, shape, or form. This will help to diversify not only the current workforce, but the future workforce as well.

Works Cited

Jainaru. **"Diversity in Tech Companies."** Kaggle, www.kaggle.com/datasets/jainaru/diversity-in-tech-companies/data. Accessed 10 June 2024.

"The Diversity Gap in Silicon Valley." Lee & Low Books Blog, 12 Mar. 2015, blog.leeandlow.com/2015/03/12/the-diversity-gap-in-silicon-valley/. Accessed 10 June 2024.

"Set Color for Each Violin in Violin Plot." Stack Overflow, stackoverflow.com/questions/34058188/set-color-for-each-violin-in-violin-plot. Accessed 10 June 2024.

Juliette Carreiro **"The Gender Gap in Tech: Let's Talk About It."** Ironhack Blog, www.ironhack.com/gb/blog/the-gender-gap-in-tech-let-s-talk-about-it. Accessed 10 June 2024.

Lardinois, Frederic. **"Black, Latino People Are Being Left Out of the Tech Workforce."** GovTech, www.govtech.com/workforce/black-latino-people-are-being-left-out-of-the-tech-workforce.html. Accessed 10 June 2024.