Tech Sector Employment Diversity in Silicon Valley
CICS 197R Final Portfolio Project

AIM: To understand workplace diversity in the technology sector through the lens of an observational study of employment statistics centered around the Silicon Valley tech companies.

OBJECTIVES:

- Observe workplace diversity in terms of:
  - Gender: Male vs. Female (sex ratio)
  - Race: White/Caucasian vs. Non-White (non-white ratio)
  - Job Category: White Collar Jobs vs. Blue Collar Jobs (blue-collar ratio)
- Draw bar graphs/pie charts of the factors:
  - Company Diversity
  - Racial Diversity
  - Gender Diversity
  - Job Role Diversity
- Deduce summary statistics on the figures of the aforementioned criteria:
  - Mean
  - Median
  - Standard Deviation
  - Quantile Ranges
- Draw histogram and plots of the factors' spread:
  - Sex Ratio
  - Non-White Ratio
  - Blue-Collar Ratio
- Generate and plot the model(s) to prove linear independence of the diversity factors.

THEORY:

1. **Sex Ratio:** In this context, the sex ratio is calculated as the ratio of number of female employees to the total number of employees. This can also be used to calculate the **inverted sex ratio** (the ratio of number of male employees to the total number of employees).

2. **Non-White Ratio:** In this context, the non-white ratio is calculated as the ratio of number of non-white employees (American-Indian/Alaskan Native, Asian, Black/African-American, Hispanic/Latino, Native Hawaiian/Pacific Islander, Multiracial) to the total number of employees. This can also be used to calculate the **white ratio** (the ratio of number of white employees to the total number of employees).

3. **Blue-Collar Ratio:** In this context, the blue-collar ratio is calculated as the ratio of number of blue-collar employees (Craft Workers, Laborers/Helpers, Operatives, Service Workers, Technicians) to the total number of employees. This can also be used to calculate the **white-collar ratio** (the ratio of number of white-collar employees to the total number of employees)

   While it's quite intuitive to imagine that the above 3 factors must be directly proportional or linearly related, this project aims to show that they are linearly independent of each other, especially relevant with the data from Silicon Valley.

DATASET METADATA:
Source: https://github.com/cirlabs/Silicon-Valley-Diversity-Data/blob/master/Reveal_EEO1_for_2016.csv (Data is available under the Open Database License)

Credits: “Reveal from The Center for Investigative Reporting." https://www.revealnews.org/svdiversity

Cleaned Working Data:

	company	race	gender	job_category	count
1:	23andMe	Hispanic/Latino	male	Executives	0
2:	23andMe	Hispanic/Latino	male	Managers	1
3:	23andMe	Hispanic/Latino	male	Professionals	7
4:	23andMe	Hispanic/Latino	male	Technicians	0
5:	23andMe	Hispanic/Latino	male	Sales Workers	0

4121:	Sanmina	Overall Totals	<NA>	Operatives	1660
4122:	Sanmina	Overall Totals	<NA>	Laborers/Helpers	4
4123:	Sanmina	Overall Totals	<NA>	Service Workers	57
4124:	Sanmina	Overall Totals	<NA>	Totals	5205
4125:	Sanmina	Overall Totals	<NA>	Managers	591

- **company** : the various companies centered around Silicon Valley
  25 levels: "23andMe", "Adobe", "Airbnb", "Apple", "Cisco", "eBay", "Facebook", "Google", "HP Inc.", "HPE", "Intel", "Intuit", "LinkedIn", "Lyft", "MobileIron", "NetApp", "Nvidia", "PayPal", "Pinterest", "Salesforce", "Sanmina", "Square", "Twitter", "Uber", "View"

- **race** : the race-wise distribution of employees
  8 levels: "American-Indian/Alaskan Native", "Asian", "Black/African-American", "Hispanic/Latino", "Native Hawaiian/Pacific Islander", "Overall Totals", "Multiracial", "White/Caucasian"

- **gender** : the gender-wise distribution of employees
  3 levels: “male”, “female”, <NA>

- **job_category** : the job type classifications of employees
  11 levels: "Administrative Support", "Craft Workers", "Executives", "Laborers/Helpers", "Managers", "Operatives", "Professionals", "Sales Workers", "Service Workers", "Technicians", "Totals"

Notes:

- The data is completely from the year 2016. It would be wise to mention this as the year column was removed during the cleaning of the data.

OBSERVATIONS & CONCLUSIONS:

i) Categorical Data

(All plots have been attached in the Plots/Charts folder)

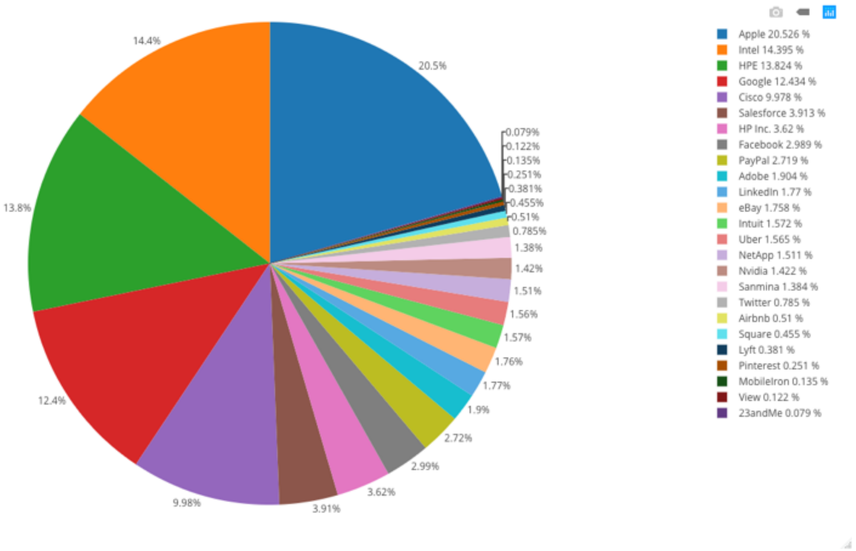


Figure 1: Company Diversity Pie Chart

```
      company  count percent
1     23andMe    594   0.079
2       Adobe  14324   1.904
3      Airbnb   3834   0.510
4       Apple 154384  20.526
5       Cisco  75052   9.978
6        eBay  13222   1.758
7    Facebook  22482   2.989
8      Google  93520  12.434
9     HP Inc.  27226   3.620
10        HPE 103978  13.824
11      Intel 108270  14.395
12     Intuit  11822   1.572
13   LinkedIn  13310   1.770
14       Lyft   2866   0.381
15 MobileIron   1012   0.135
16     NetApp  11362   1.511
17     Nvidia  10696   1.422
18     PayPal  20454   2.719
19  Pinterest   1888   0.251
20 Salesforce  29432   3.913
21    Sanmina  10410   1.384
22     Square   3422   0.455
23    Twitter   5904   0.785
24       Uber  11770   1.565
25       View    920   0.122
```

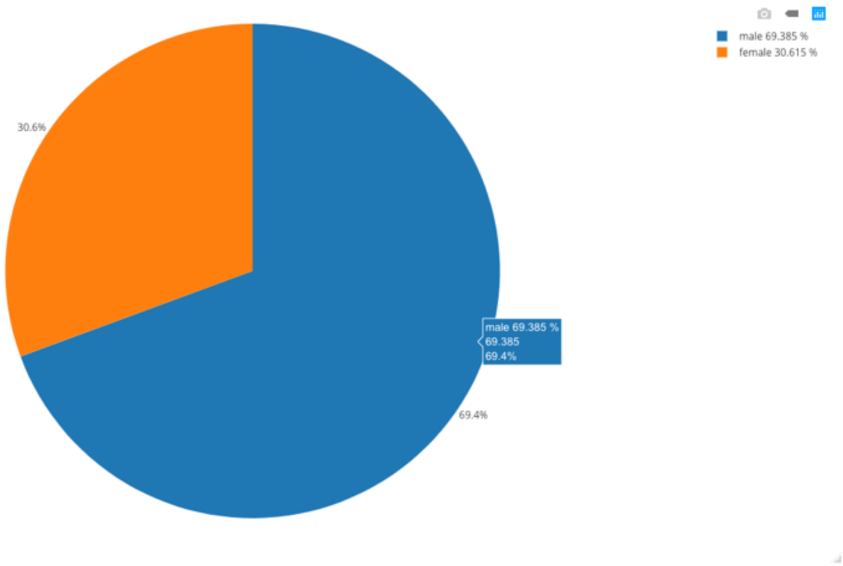


Figure 2: Gender Diversity Pie Chart

```
  gender  count percent
1 female 230270  30.615
2   male 521884  69.385
3  total 752154 100.000
```

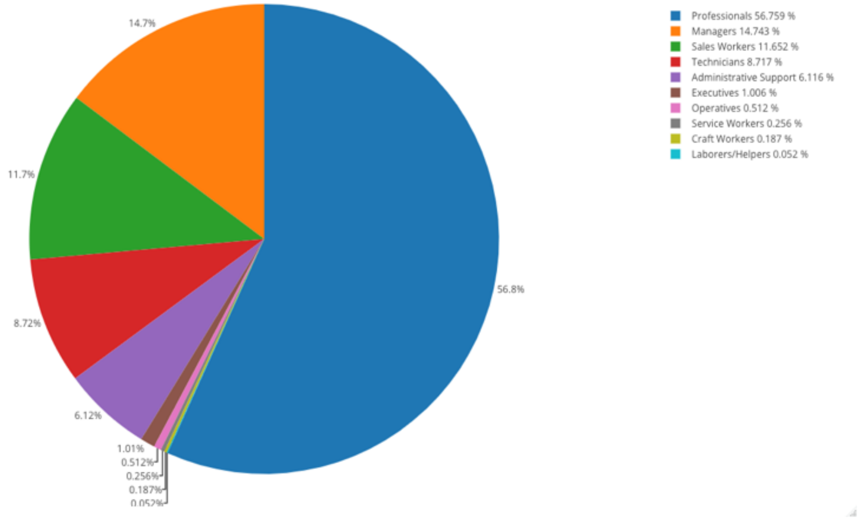


Figure 3: Racial Diversity Pie Chart

```
                                 race  count percent
1   American-Indian/Alaskan Native   2474   0.329
2                            Asian 203584  27.067
3           Black/African-American  38398   5.105
4                  Hispanic/Latino  54490   7.245
5                      Multiracial  12130   1.613
6 Native Hawaiian/Pacific Islander   2388   0.317
7                  White/Caucasian 438690  58.324
8                   Overall Totals 752154 100.000
```

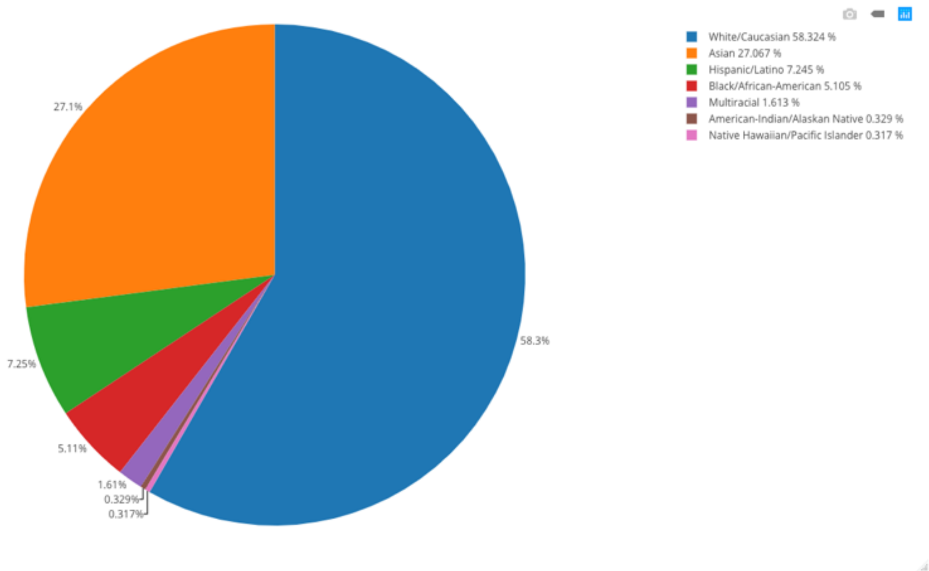
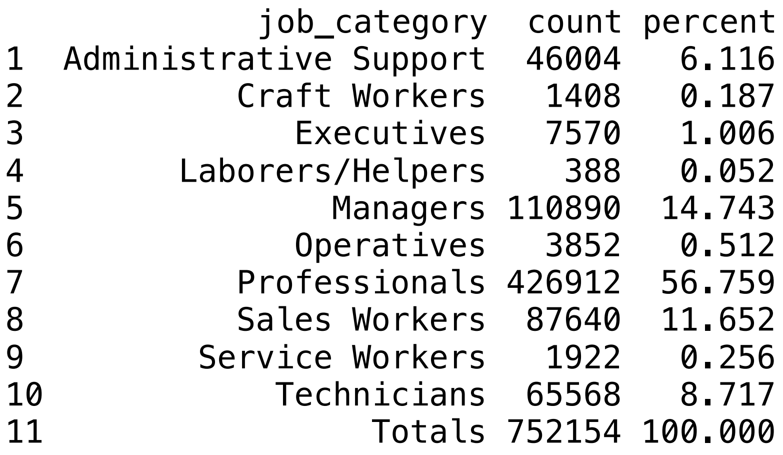


Figure 4: Job Diversity Pie Chart



```
               job_category  count percent
1  Administrative Support  46004   6.116
2           Craft Workers   1408   0.187
3              Executives   7570   1.006
4        Laborers/Helpers    388   0.052
5                Managers 110890  14.743
6              Operatives   3852   0.512
7           Professionals 426912  56.759
8           Sales Workers  87640  11.652
9         Service Workers   1922   0.256
10            Technicians  65568   8.717
11                 Totals 752154 100.000
```

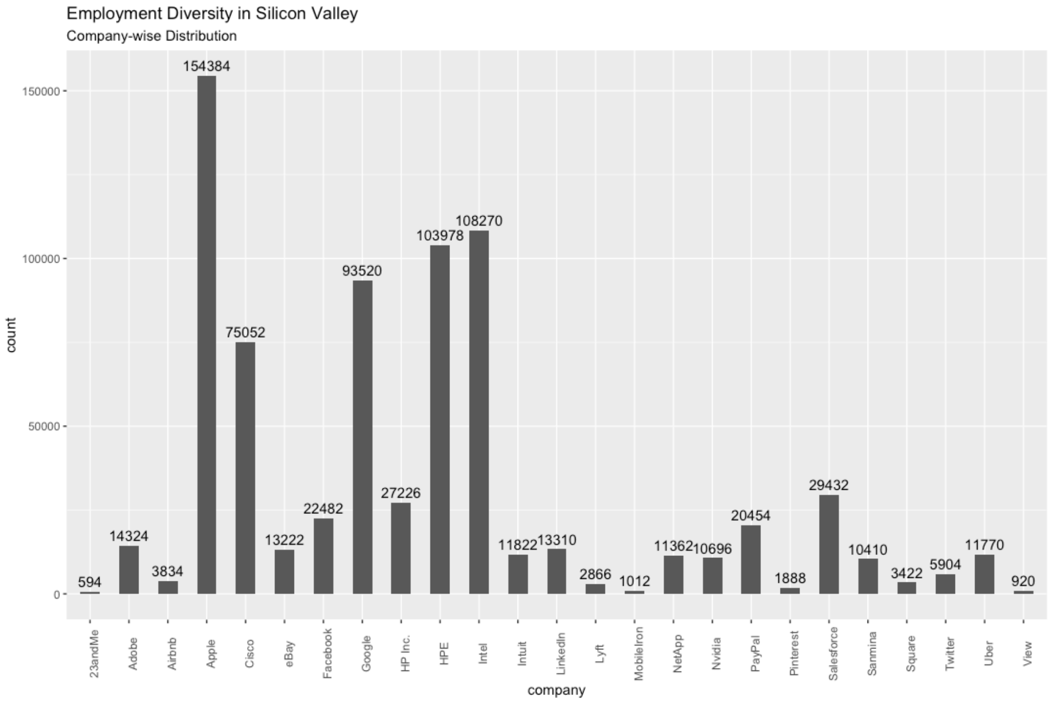


Figure 5: Bar Chart for Company Diversity

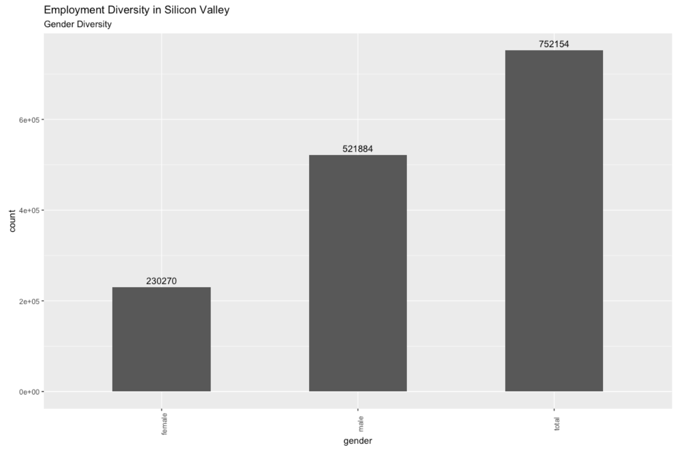


Figure 6: Bar Chart for Gender Diversity

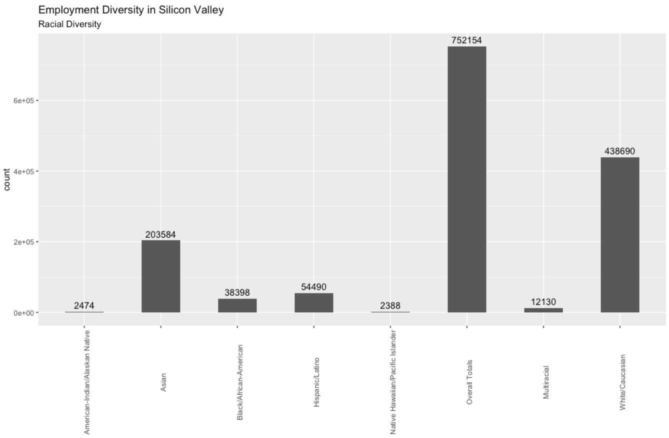


Figure 7: Bar Chart for Racial Diversity

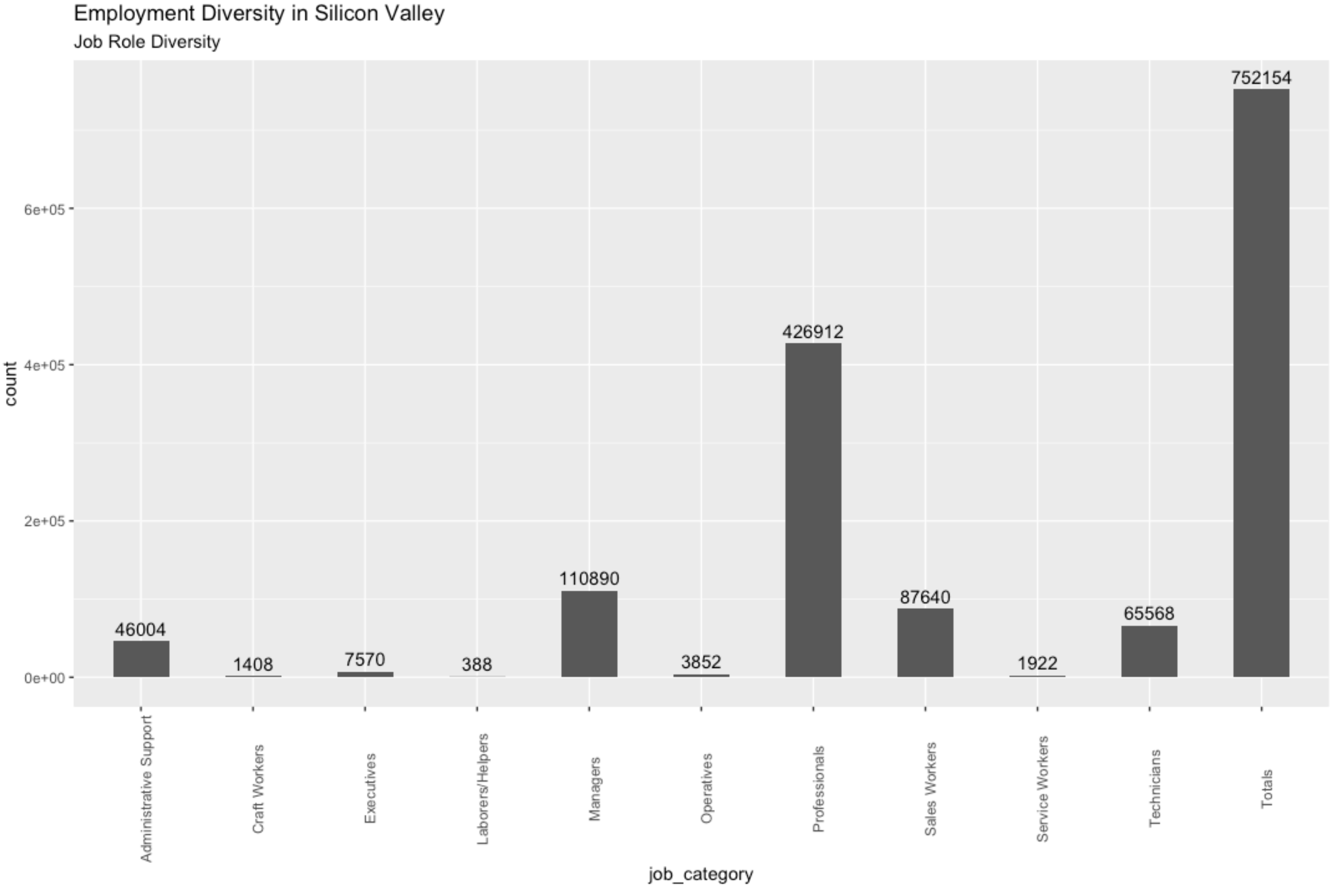


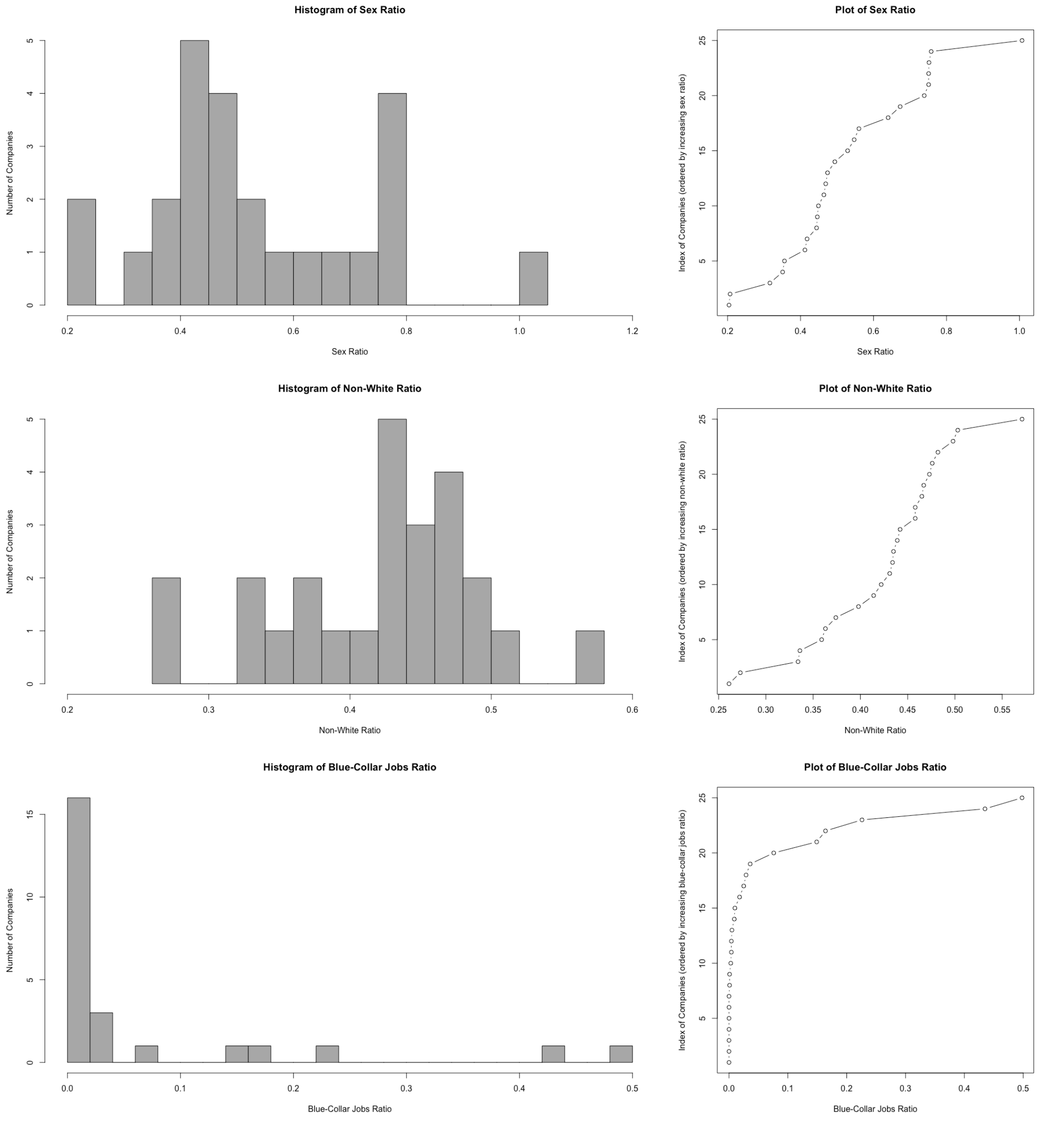
Figure 8: Bar Chart for Job Role Diversity

ii) Statistical Summaries

Statistic	Mean	Median	Standard Deviation
Sex Ratio	0.52844	0.474	0.1917622
Non-White Ratio	0.42264	0.435	0.0721254
Blue-Collar Jobs Ratio	0.06772	0.005	0.1341428
Female Workforce	13484.7	6789.576	17670.84
Non-White Workforce	12537.99	5496.59	18067.25
Blue-Collar Workforce	2923.726	75.052	7713.895
Total Workforce	30086.16	11822	41788.9

Statistic	Quantile Distribution				
Sex Ratio	0%	25%	50%	75%	100%
	0.204	0.418	0.474	0.673	1.007
Non-White Ratio	0%	25%	50%	75%	100%
	0.261	0.374	0.435	0.467	0.571
Blue-Collar Jobs Ratio	0%	25%	50%	75%	100%
	0.000	0.000	0.005	0.036	0.498
Female Workforce	0%	25%	50%	75%	100%
	187.680	2214.072	6789.576	13803.608	68546.496
Non-White Workforce	0%	25%	50%	75%	100%
	222.156	1663.956	5496.590	9830.288	67774.576
Blue-Collar Workforce	0%	25%	50%	75%	100%
	0.000	0.000	75.052	562.050	34890.784
Total Workforce	0%	25%	50%	75%	100%
	594	3834	11822	27226	154384

iii) Discrete Distributions of Sex Ratio, Non-White Ratio and Blue-Collar Jobs Ratio



(All plots have been attached in the Plots/Charts folder)

iv) Linear Models to Prove Linear Independence

a) Sex Ratio ~ Non-White Ratio

```
# Residuals:
#   Min       1Q   Median       3Q      Max
# -0.32008 -0.11349 -0.06361  0.15951  0.46140
#
# Coefficients:
#                    Estimate Std. Error t value Pr(>|t|)
#   (Intercept)       0.6776     0.2355   2.878   0.0085 **
#   non_white_ratio  -0.3528     0.5495  -0.642   0.5271
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.1942 on 23 degrees of freedom
# Multiple R-squared:  0.01761,     Adjusted R-squared:  -0.0251
# F-statistic: 0.4123 on 1 and 23 DF,  p-value: 0.5271
```

b) Sex Ratio ~ Blue-Collar Jobs Ratio

```
# Residuals:
#   Min       1Q   Median       3Q      Max
# -0.34415 -0.12979 -0.03166  0.18819  0.45585
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
#   (Intercept)        0.55115    0.04284  12.864 5.45e-12 ***
#   blue_collar_ratio -0.33528    0.28976  -1.157    0.259
#
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.1904 on 23 degrees of freedom
# Multiple R-squared:  0.05501,     Adjusted R-squared:  0.01392
# F-statistic: 1.339 on 1 and 23 DF,  p-value: 0.2591
```

c) Non-White Ratio ~ Blue-Collar Jobs Ratio

```
# Residuals:
#   Min       1Q   Median       3Q      Max
# -0.15947 -0.05045  0.01755  0.04282  0.14655
#
# Coefficients:
#                      Estimate Std. Error t value Pr(>|t|)
#   (Intercept)        0.42445    0.01656  25.638   <2e-16 ***
#   blue_collar_ratio -0.02675    0.11197  -0.239    0.813
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.07359 on 23 degrees of freedom
# Multiple R-squared:  0.002475,    Adjusted R-squared:  -0.0409
# F-statistic: 0.05706 on 1 and 23 DF,  p-value: 0.8133
```

∴ From this entire exercise generating these linear models, we observe that there is no significant linear correlation between the factors, and therefore, we can conclude that these variables are linearly independent of each other's influence.