CS 6603: AI, Ethics, and Society
Final Project Report

Cleo Zhang

yzhang3761@gatech.edu

1 STEP 1 – DATASET OVERVIEW

- Which dataset did you select? *IBM HR Analytics Employee Attrition & Performance.*
- Which regulated domain does your dataset belong to? *Employment.*
- How many observations are in the dataset? *1470.*
- How many variables are in the dataset? *35.*
- Which variables did you select as your dependent variables? *Attrition and Hourly rate.*
- How many and which variables in the dataset are associated with a legally recognized protected class? Which legal precedence/law (as discussed in the lectures) does each protected class fall under? *See Table 1.*

Table 1 — Protected class categories and associated variables.

Variables	**Protected classes**	**legal precedence**
Age	Age	*Age Discrimination in Employment Act of 1967.*
Gender	Sex	*Equal Pay Act of 1963; Civil Rights Act of 1964, 1991.*

2 STEP 2

2.1 Identification of the Members Associated with the Protected Classes and the Selected Dependent Variables

This section shows how the members associated with the selected dependent variables (Table 3) and the protected classes (Table 2) are discretized into numerical values. Those values will be used in the remaining calculations of this report.

Table 2 — The relationship between members and membership categories for each protected class.

Protected Class	Subgroups	Members
Age	18 - 25	0
	26 - 35	1
	36 - 45	2
	46 - 55	3
	55+	4
Gender	Female	0
	Male	1

Table 3 — The relationship between values and discrete categories/numerical values associated with the dependent variables.

Dependent Variables	Subgroups	Members
Attrition	No	*0*
	Yes	*1*

Dependent Variables	Subgroups	Members
Hourly rate	30 - 50	*0*
	51 - 70	*1*
	71 - 90	*2*
	91+	*3*

2.2 Frequency Tables and Histograms for Protected Class Variable vs. Dependent Variables

Below are the frequency table and histograms of each protected class and dependent variable combination.

- Age vs. Attrition

Table 4 — Frequency Table for Age vs. Attrition.

Age	Attrition	
	0 (No)	1 (Yes)
0 (18 – 25)	79	44
1 (26 – 35)	490	116
2 (36 – 45)	425	43
3 (46 – 55)	200	26
4 (55+)	39	8

```
Frequency Table for Age vs. Attrition:                ID
Attrition     No  Yes
Parsed_age
0             79   44
1            490  116
2            425   43
3            200   26
4             39    8

Text(0, 0.5, 'Count')
```

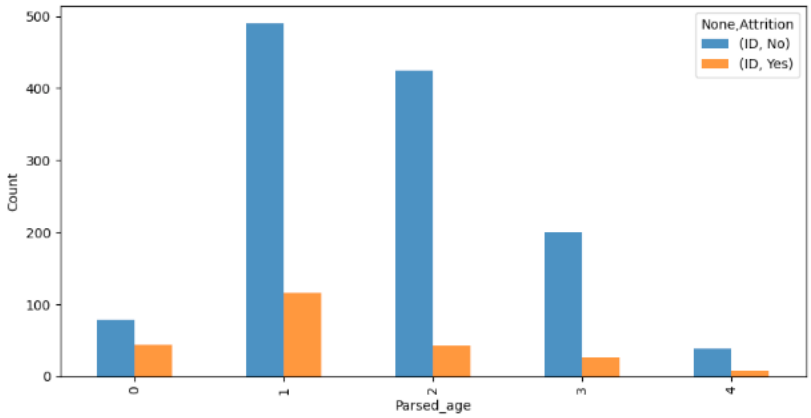

Figure 1 — Histograms for Age vs. Attrition.

- Age vs. Hourly Rate

Table 5 — Frequency Table for Age vs. Hourly Rate.

Age	**Hourly Rate**			
	0 (30 - 50)	1 (51 - 70)	2 (71 - 90)	3 (91+)
0 (18 – 25)	36	39	25	23
1 (26 – 35)	179	164	180	83
2 (36 – 45)	126	125	146	71
3 (46 – 55)	60	64	67	35
4 (55+)	8	16	16	7

```
Frequency Table for Age vs. Hourly Rate:                    ID
Parsed_hourly_rate    0    1    2   3
Parsed_age
0                    36   39   25  23
1                   179  164  180  83
2                   126  125  146  71
3                    60   64   67  35
4                     8   16   16   7

Text(0, 0.5, 'Count')
```

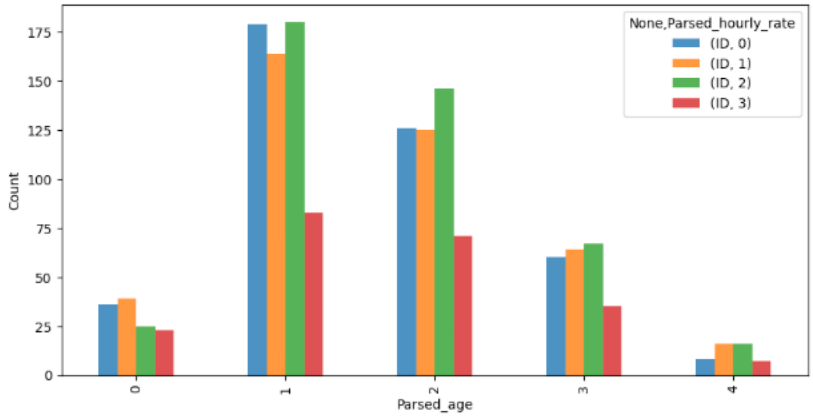

Figure 2 - Histograms for Age vs. Hourly Rate.

- Gender vs. Attrition

Table 6 — Frequency Table for Gender vs. Attrition.

Gender	**Attrition**	
	0 (No)	1 (Yes)
0 (Female)	501	87
1 (Male)	732	150

```
Frequency Table for Sex vs. Attrition:            ID
Attrition   No  Yes
Gender
Female     501   87
Male       732  150

Text(0, 0.5, 'Count')
```

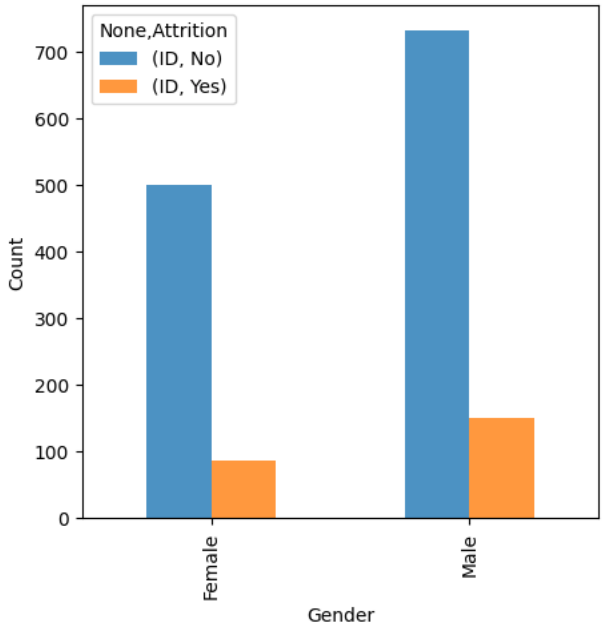

Figure 3 - Histograms for Gender vs. Attrition.

- Gender vs. Hourly Rate

Table 7 — Frequency Table for Gender vs. Hourly Rate.

Gender	**Hourly Rate**			
	0 (30 - 50)	1 (51 - 70)	2 (71 - 90)	3 (91+)
0 (Female)	163	163	175	92
1 (Male)	246	250	259	127

```
Frequency Table for Sex vs. Hourly Rate:
Parsed_hourly_rate    0    1    2    3
Gender
Female              163  158  175   92
Male                246  250  259  127

Text(0, 0.5, 'Count')
```

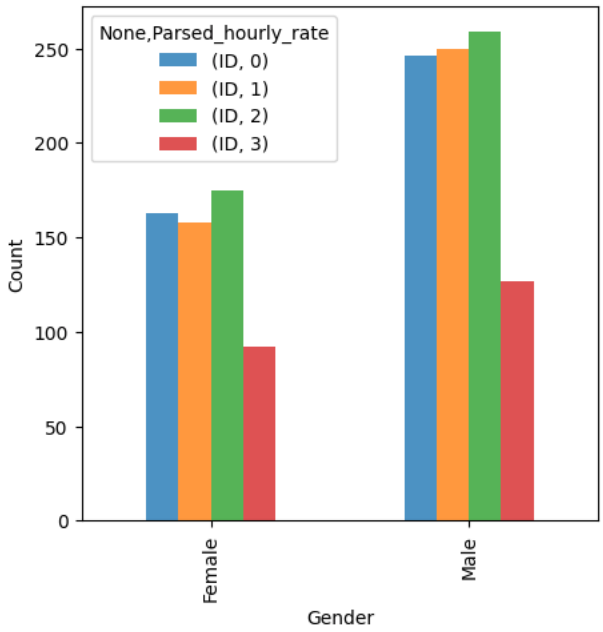

Figure 4 - Histograms for Gender vs. Hourly Rate.

3 STEP 3

3.1 Privileged/Unprivileged Groups and Fairness Metrics Identification

The privileged and unprivileged groups are identified in Table 8, and the selected fairness metrics are:

- Disparate Impact: an ideal outcome for this metric is between 0.8 and 1.25.
- Statistical Parity Difference: an ideal result for this metric is between -0.1 and 0.1.

Table 8 - Frequency Table for Gender vs. Hourly Rate.

Protected classes	Privileged	Unprivileged
Age	*18 – 25 (0)*	*26+ (1, 2, 3, 4)*
Gender	*Male (1)*	*Female (0)*

3.2 Age vs. Dependent Variables – Original Disparate Impact & Statistical Parity Difference

For Age vs. Attrition, this report uses the average Attrition values for each Age group. $\text{DisparateImpact}_{\text{AgeAttrition}}$ and $\text{StatisticalParityDifference}_{\text{AgeAttrition}}$ are out of the ideal range, so the age group of 18-25 is privileged for Attrition.

- $\text{DisparateImpact}_{\text{AgeAttrition}} = 1825Attrition_{avg} \div 2660Attrition_{avg} = 2.4967$
- $\text{StatisticalParityDifference}_{\text{AgeAttrition}} = 1825Attrition_{avg} - 2660Attrition_{avg} = 0.2144$

For Age vs. Hourly Rate, this report uses the average Hourly Rate values for each Age group. Even though both $\text{DisparateImpact}_{\text{Age}HourlyRate}$ and $\text{StatisticalParityDifference}_{\text{Age}HourlyRate}$ are within the ideal range, young people aged 18-25 are disadvantaged.

- $\text{DisparateImpact}_{\text{Age}HourlyRate} = 2660HourlyRate_{Avg} \div 1825HourlyRate_{Avg} = 0.9748$
- $\text{StatisticalParityDifference}_{\text{Age}HourlyRate} = 2660HourlyRate_{Avg} - 1825HourlyRate_{Avg} = -0.0332$

3.3 Gender vs. Dependent Variables - Original Disparate Impact & Statistical Parity Difference

For Gender vs. Attrition, we use the average of Attrition and Hourly Rate for each gender to calculate the Disparate Impact and Statistical Parity Difference. As shown below, even though both $\text{DisparateImpact}_{\text{SexAttrition}}$ and

$\text{StatisticalParityDifference}_{\text{SexAttrition}}$ are within the ideal range – the Male group is a bit more privileged than the Female group for Attrition.

- $\text{DisparateImpact}_{\text{SexAttrition}} = MaleAttrition_{avg} \div FemaleAttrition_{avg} = 1.1494$
- $\text{StatisticalParityDifference}_{\text{SexAttrition}} = MaleAttrition_{avg} - FemaleAttrition_{avg} = 0.0221$

Similarly, $\text{DisparateImpact}_{SexHourlyRate}$ and $\text{StatisticalParityDifference}_{\text{Sex}HourlyRate}$ are also within the reasonable range; however, the Female group is a bit more privileged than the Male group for the Hourly Rate.

- $\text{DisparateImpact}_{SexHourlyRate} = MaleHourlyRate_{Avg} \div FemaleHourlyRate_{Avg} = 0.9770$
- $\text{StatisticalParityDifference}_{\text{Sex}HourlyRate} = MaleHourlyRate_{Avg} - FemaleHourlyRate_{Avg} = -0.0306$

3.4 Transformation of the Original Dataset as a Function of Attrition

Based on the calculation from 3.2 and 3.3, the bias in Attrition by Age is the most significant. Therefore, I will focus on mitigating this bias by applying a pre-processing bias mitigation algorithm – Reweighting - to adjust the weights of the observations to equalize the outcomes for privileged and unprivileged age groups while preserving the overall distribution of the data. Specifically, I add a new column to the data table called "Weight," Then, we multiply the parsed Attrition value with the corresponding "Weight" value as an outcome.

For the rest of this assignment, I will use the parsed table (Figure 5) from Step 1 in the attached .ipynb file as the original dataset. This parsed dataset scratched out the unrelative columns, keeps the same data characteristics of the column we are interested in and works with a better calculation efficiency.

```
length of new_Df 1470
```

	Age	Gender	Attrition	Hourly_rate	ID
0	2	0	1	3	0925a8a4-3c67-4f04-910e-238a9e6c422b
1	3	1	0	1	9c5a4e99-b541-40d2-9651-a211900a2e9e
2	2	1	1	3	3f7886bb-20da-4285-b18b-a36941844952
3	1	0	0	1	7393fb8a-dd67-4ead-867e-4e434b5fefdc
4	1	1	0	0	18e5672c-2a7a-487a-9846-cd224a93196c

Figure 5 – Head of the Parsed Table.

Figure 6 shows the algorithm that transforms the original dataset.

```
# Identify the protected class: age 18-25
protected = orig_dataset[(orig_dataset['Age'] == 0)]
unprotected = orig_dataset[(orig_dataset['Age'] != 0)]

# Calculate the weights based on an inverse
unprotected_prob = orig_dataset.loc[unprotected.index, 'Attrition'].mean()
protected_prob = orig_dataset.loc[protected.index, 'Attrition'].mean()

orig_dataset.loc[unprotected.index, 'Weight'] = 1 / unprotected_prob
orig_dataset.loc[protected.index, 'Weight'] = 1 / protected_prob

# Come up with the transformed dataset
transformed_df_3 = orig_dataset.copy()
transformed_df_3['Attrition'] = transformed_df_3['Attrition'] * transformed_df_3['Weight']
```

Figure 6 – The Reweighing Algorithm and the transformed table (transformed_df_3).

3.5 Age –Disparate Impact & Statistical Parity Difference on Transformed Dataset

The Disparate Impact and Statistical Parity Differences for the transformed dataset are shown below:

- $\text{DisparateImpact}_{\text{AgeAttrition}} = 1825Attrition_{avg} \div 2660Attrition_{avg} \approx 1$
- $\text{StatisticalParityDifference}_{\text{AgeAttrition}} = 1825Attrition_{avg} - 2660Attrition_{avg} \approx 0$

Due to error accumulation, both Disparate Impact and Statistical Parity Differences are approximate rather than exact values.

4 STEP 4 – OPTION B

This step will take Attrition as my dependent variable and Age as my protected class (indicated in Table 9).

Table 9 - Frequency Table for privileged and unprivileged groups for Step 4.

Protected classes	Privileged	Unprivileged
Age	*18 – 25 (0)*	*26+ (1, 2, 3, 4)*

After examining the original dataset, I observed that the privileged age group is significantly smaller than the unprivileged age group. To address the existing bias, I will develop a new algorithm in this step that utilizes variant over-sampling (Figure 7).

```
# Copy new_df to transformed_df_3
transformed_df_4 = new_df.copy()

# Identify the privileged age group
privileged_df = transformed_df_4.loc[protected.index]
print("protected_df", len(protected_df))
# Identify the unprivileged age group
unprivileged_df = transformed_df_4.loc[unprotected.index]
print("unprotected_df", len(unprotected_df))

# Create another df using the original privileged_df which has
# 1. random duplicates of the original privileged_df to be included;
# 2. the size of oversampled_privileged_df = len(unprivileged_df) - len(privileged_df)
oversampled_privileged_df = resample(privileged_df,
                                     replace = True,
                                     n_samples = unprivileged_df.shape[0] - len(protected_df),
                                     random_state = 3)

# Assigned the Attrition values to 0 in the oversampled_privileged_df
# to not introduce more YESs (1)
oversampled_privileged_df['Attrition'] = 0

# cancat oversampled_privileged_df, unprivileged_df, privileged_df
# to create a balanced df and finish the transformation
df_balanced = pd.concat([oversampled_privileged_df, unprivileged_df, privileged_df])
```

Figure 7 – The Oversampling Algorithm and the transformed table (transformed_df_4).

The above algorithm aims to increase the size of sampling based on the unprivileged group size but not change the cumulative Attrition values of the Privileged

age group. Therefore, compared to the original dataset's Disparate Impact and Statistical Parity Difference, these two data will decrease accordingly.

Table 9 - Two fairness metrics associated with the original and transformed dataset.

	Original Testing	**Transformed Testing**
Disparate Impact	3.1272	≈ 0.2
Statistical Parity Difference	0.2616	≈ -0.1

While the current Statistical Parity Difference falls within a reasonable range, the Disparate Impact appears overly corrected. Table 10 illustrates the differences in the Disparate Impact and Statistical Parity Difference values for the current results vs. the transformed dataset discussed in Section 3 and the Original Testing dataset in Section 4.

Table 10 – Effect in each fairness metric after transforming the dataset in section 4.

	Original Testing	**Transformed Testing in Section 3**	**Transformed Testing in Section 4**
Disparate Impact	Positive change	Negative change	No Change
Statistical Parity Difference	Positive change	Negative change	No Change

This algorithm causes an overcorrection on Disparate Impact primarily because of the excessively oversampling of the privileged Age group. One possible solution is to dynamically adjust the sampling size according to the ratio of the privileged/unprivileged group size instead of simply matching up with the unprivileged group size.

5 STEP 5

I am a team of one.