Exploratory Data Analysis

- StudentID: 22000741

- Name: Shinkook Cha

- 1st Major: Life Science

- 2nd Major: AI convergence

Brief summary of your proposed project idea: Factors that affect pay gap will be evaluated for racially divergent samples in order to determine whether the argument of racial pay gap is a statistically sound argument or not

1. Data overview

The table below is a summary of the descriptive statistics - data type, count, category/range, number of categories - on the overall dataset.

NA (i.e., Not Available) values have been considered and eliminated in the making of this table. Thus the difference in the sample size for each variables (e.g., Race and Income)

Although income and working hours, in first sight, could be thought as continuous type variables, due to their discrete characteristic, they have been categorized as discrete type variables.

	Race	**Income**	**Working Hours**	**Highest Degree**	**Sex**
Data Type	Categorical	Discrete	Discrete	Categorical	Categorical
Count	59599	38450	59597	46602	59599
Category/Range	1,2,3	1,2...99	-1,1,2...99	1,2,3,4,8,9	1,2
Number of Categories	3	NA	NA	6	2

[Table 1. Table of Descriptive Statistics]

2. Univariate Analysis

2.1 Race

The variable 'Race' consists of 3 categories: '1','2' and '3' and the respective count number (i.e., sample size) for each categories are: 48240, 8312 and 3047. In short, category '1' has the highest sample size, followed by category '2' and '3'.

2.2 Income

The variable 'Income', prior to data preprocessing, ranges from 1 to 99. However, in consideration of the fact that most sample's 'Income' ranges from 1 to 13, samples with 'Income' value higher than 90 or with NA values have been dropped. After an appropriate data preprocessing step, a simple univariate analysis was conducted using Python library Pandas' describe() function. The result of the univariate analysis is summarized in Table 1.

2.3 Working Hours

The variable 'Working Hours', prior data preprocessing, ranges from -1 to 99. After an appropriate data preprocessing step (i.e., elimination of negative numbers, NA values and excessively high values), a univariate analysis was carried out in a similar manner as the 'Income' variable. The results of the univariate analysis are summarized in Table 1.

	Income	**Working Hours**
count	36811	1123
mean	9.347	39.051
std*	3.409	13.150
min	1	1
25%	8	35
50%	11	40
75%	12	45
max	13	80

[Table 2. Univariate analysis result table for the variable 'Income' and 'Working Hours'] *std: standard deviation

2.4 Highest Degree

The variable 'Highest Degree' consists of 6 categories. Namely, '1', '2', '3', '4', '8' and '9'. 30556 samples are in category '1', followed by 3256, 8474, 4151, 30 and 135 samples for groups '2', '3', '4', '8' and '9' respectively.

2.5 Sex

Last but not least, the variable 'Sex' consists of 2 categories: '1' and '2'. The dataset consists of 26286 samples in sex category '1' and 33313 samples are in sex category '2'.

3. Multivariate analysis

For simplicity, from here on, I will denote the variable names as the following:

- Race: R
- Income: INC
- Working Hours: WH
- Highest Degree: HD
- Sex: Sex

3.1 Boxplot of Income by Race

After an appropriate data preprocessing step, the INC for each R was visualized using Python library Seaborn's boxplot() function. Furthermore, in order to determine whether the mean INC between R groups show statistically significant differences, the Welch's t-test was performed - Welch's t-test is a statistical test which is utilized to argue whether the mean of two sample groups, presumably sampled from non-normal distributions, shows statistical difference or not. As shown in the boxplot below, the p-value from Welch's t-test indicates that there is a statically significant difference in the mean INC between R groups.

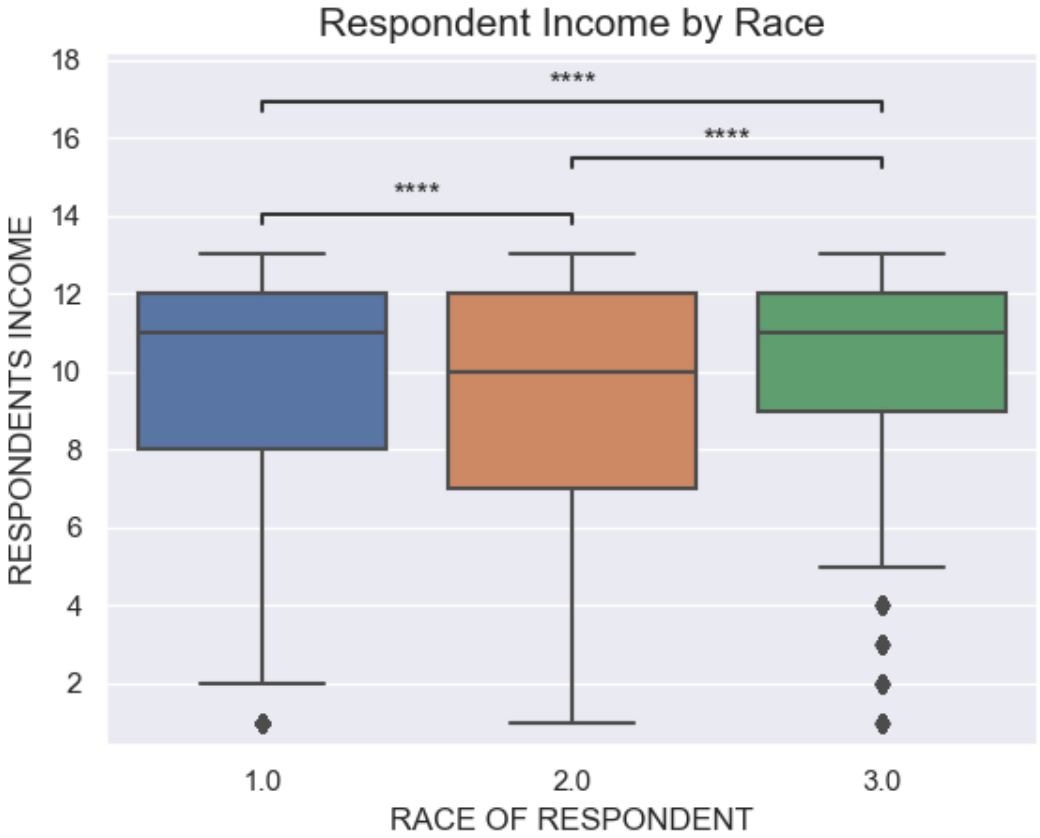

Fig.1 Boxplot of Income by Race

3.2 Boxplot of Weekly Working Hours by Race

One of the variables which shows a relatively high correlation to INC is WH (The correlation between the two variables are approximately 0.343). The fact that the mean INC for each race group is different could simply be due to the difference in WH. Thus, the WH by each race was visualized using a simple boxplot function and the Welch's t-test was performed in order to evaluate whether there is a significant difference in the mean WH between the R groups. As depicted in the graph below by the 'ns', short for 'not significant', the mean WH between each group was not significant.

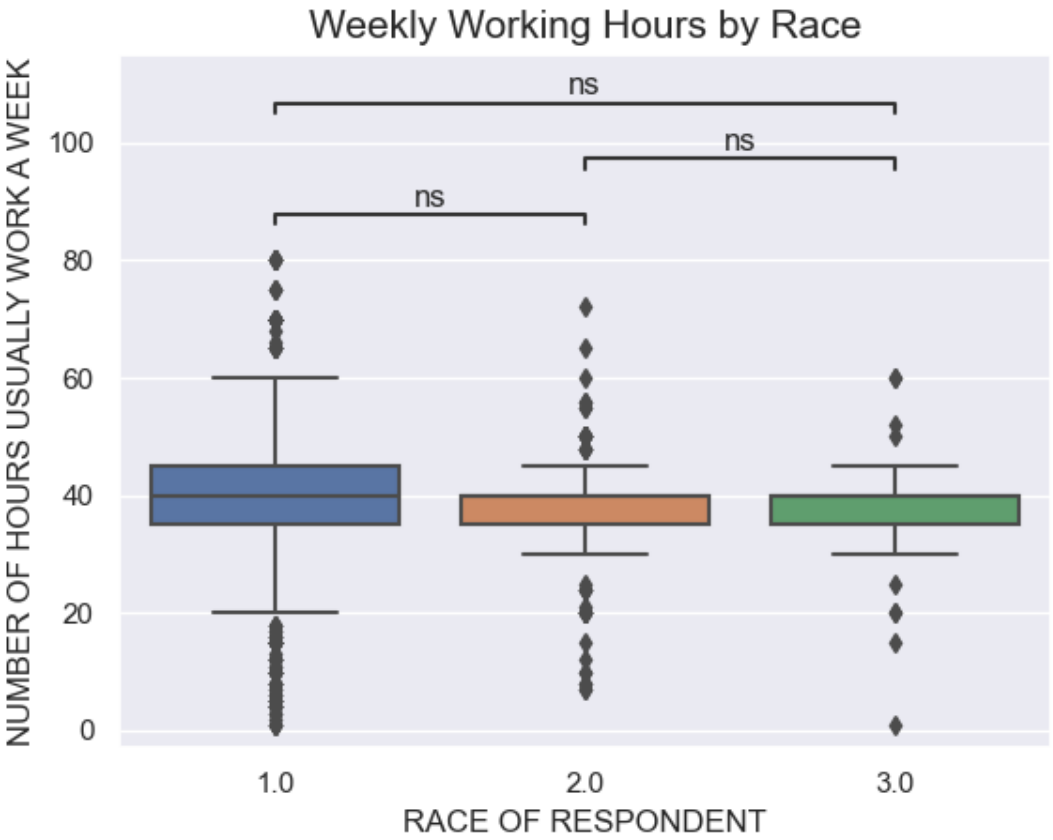

Fig.2 Boxplot of Weekly Working Hours by Race

3.3 Boxplot of Income over Highest Degree

It is common knowledge that the higher the education degree, the more likely one will receive a higher income. Thus, in order to validate this postulation, a simple boxplot was drawn for INC by HD; and again, Welch's t-test was performed in order to evaluate whether there is a significant difference in the mean INC between HD groups.

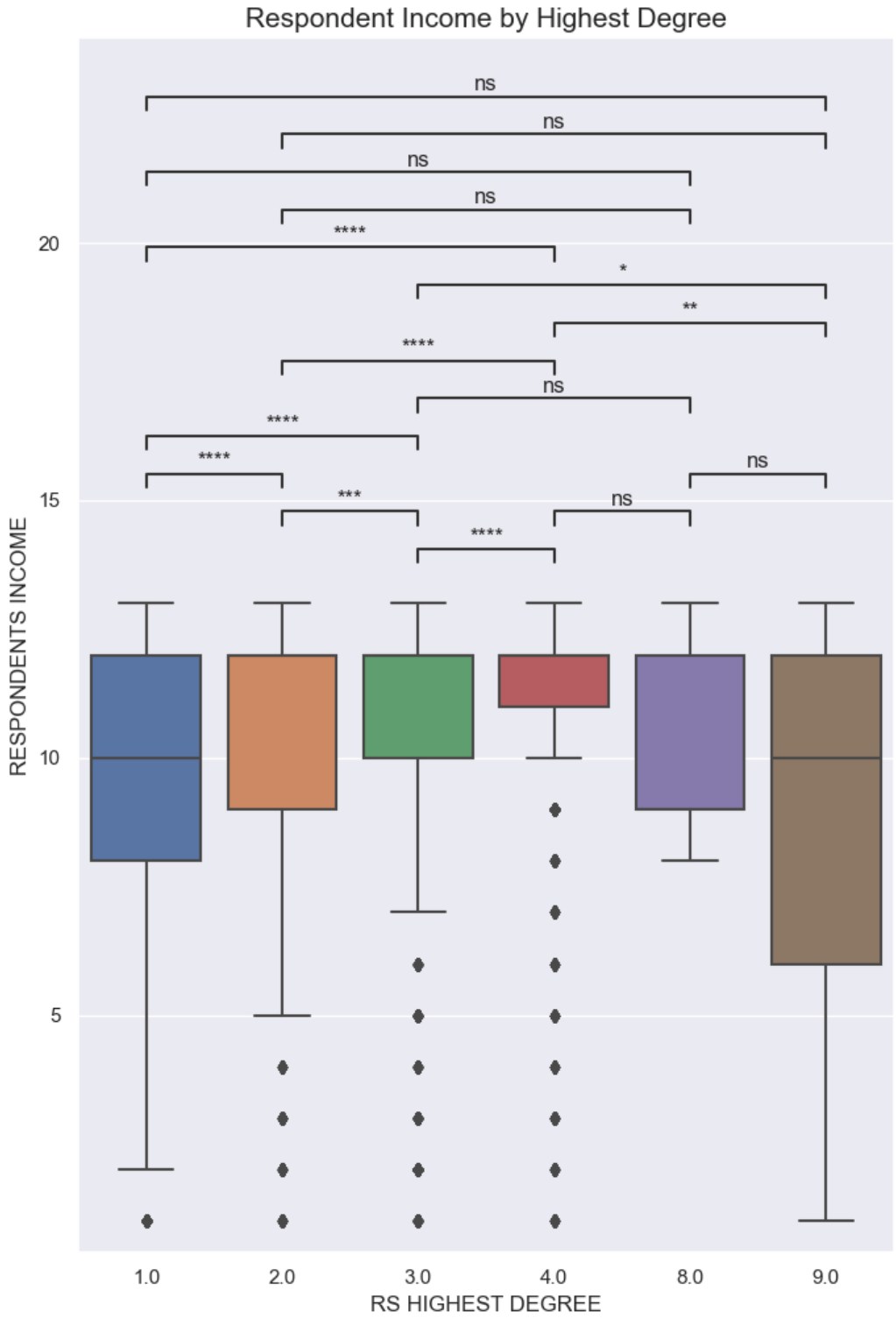

Fig.3 Boxplot of Income by Highest Degree

As shown in the diagram above, the mean INC shows significant difference between HD groups that correspond to '1', '2', '3' and '4'. However, as for HD groups '8' and '9', the mean INC showed little to no

significant difference between each other nor to other groups. This indicates that for HD '1', '2', '3', and '4', as HD increases, so does the the mean INC.

3.4 Violinplot of Highest Degree by Race

Fig.3 has shown that the mean INC of different HD groups shows statistically significant differences apart from groups '8' and '9'. Therefore, in 3.4, a violinplot of the HD will be visualized for each R groups on top of Welch's t-test to determine if there is a statistical difference in the mean HD between R groups. However, the reader should note a few presumptions that have been made prior to the analysis. The first presumption is the idea that a low numbered group corresponds to a relatively low HD and high numbered group corresponds to a relatively high HD. In other words, for instance, '1' would correspond to a highschool graduate and '4' would correspond to a respondent who holds a Ph.D degree. Due to the absence of a meta dataset, the mapping of each numbered category to the actual category was infeasible. Yet, from the fact that with increasing HD, INC increases, I believe that the inferred presumption that have been taken into account of, is not an unsound presumption. The second, which the reader should note, is that categories '8' and '9' have been eliminated prior to the analysis. The purpose of Welch's t-test in 3.4 is to validate whether there is a significant difference in the mean of HD between the R groups. If the mean HD between R group are different, based on the results from 3.3, we could further infer that the difference in mean INC is due to the difference in mean HD. However, if we included HD groups '8' and '9', we would not be able to argue as such, since HD groups '8'/'9' and '1'/'2' showed no significant difference in mean INC.

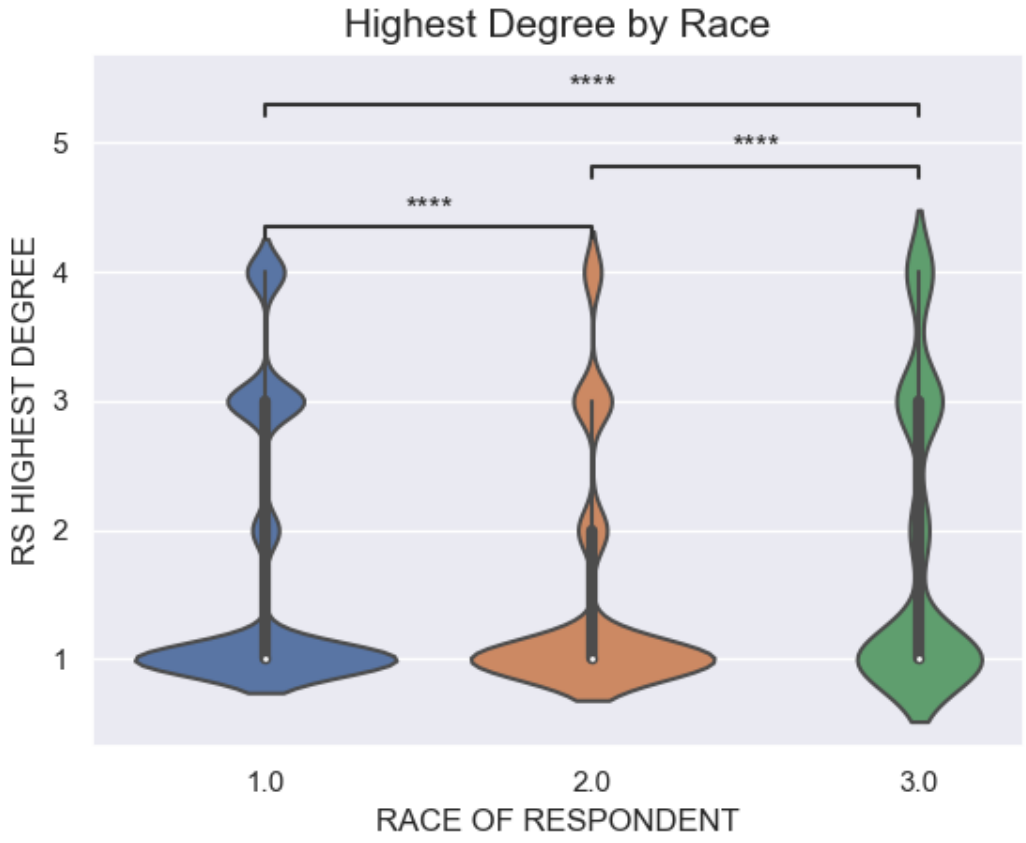

Fig.4 Violinplot of Highest Degree by Race

From the figure above, it is evident that the mean HD between R groups is significantly different and that HD '2' is significantly lower relative to the other two groups. Thus, from the results of 3.3, we can infer that the difference in mean INC is at least partially due to the difference in the HD between the R groups.

3.5 Boxplot of Income by Sex

Another factor that is known to influence INC is the respondent's Sex (i.e., Male/Female). In order to validate whether this statement's truthfulness, we, again, utilize boxplot to visualize the distribution of INC for each Sex and use Welch's t-test to determine whether the mean INC between the two Sex groups are significantly different.

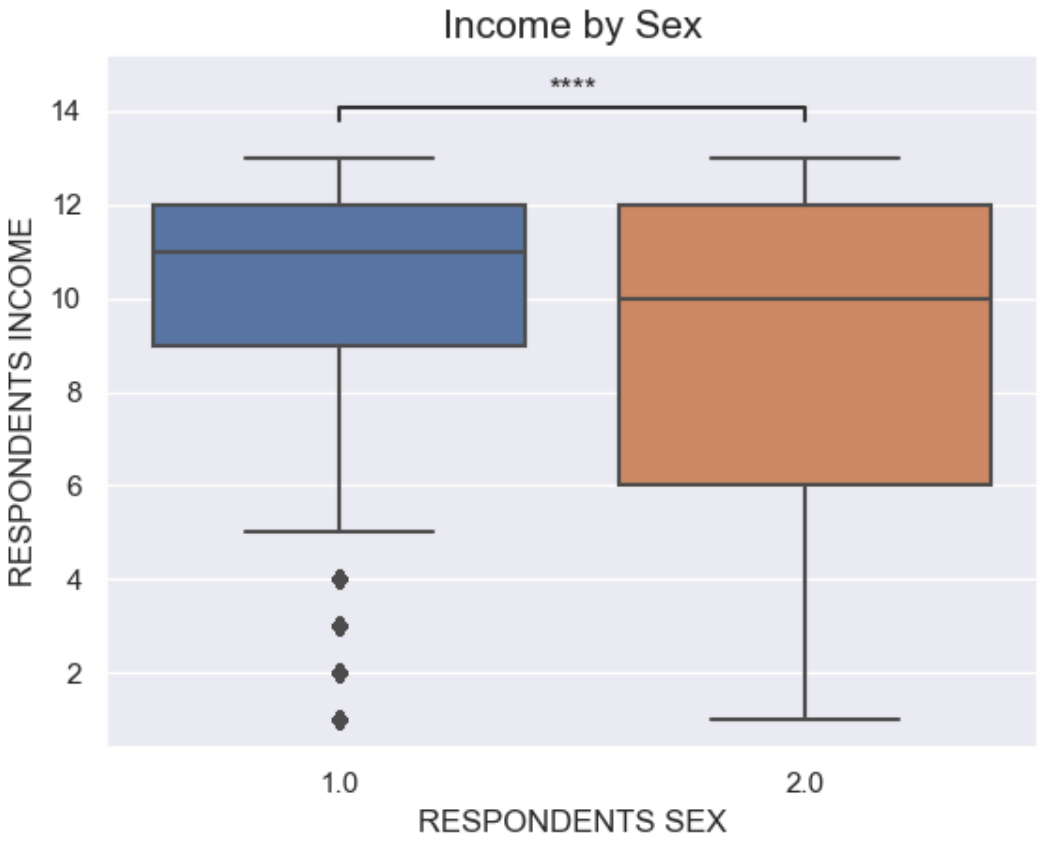

Fig.5 Boxplot of Income by Sex

As shown in the diagram above, the p-value indicates that the mean income of Sex group '1' is significantly higher than Sex group '2'. We will now evaluate the Sex distribution in each R group.

3.6 Frequency Distribution Table of Sex and Race

	Race 1	**Race 2**	**Race 3**
Sex 1	21688	3164	1434
Sex 2	26552	5148	1613
Sex 1/2 Ratio	0.816	0.614	0.889

[Table 3. Frequency Distribution Table of Sex and Race]

As shown in the Frequency Distribution Table above, through simple eye-balling, we see that the the Sex 1/2 ratio in R group '2', is at least 0.2 lower than R '1' and '3'. In other words, this tells us that R group '2' has a higher number of Sex '2' respondents who earn significantly lesser than Sex '1' respondents, as shown in 3.5. This indicates that the difference in pay gap between R groups is partially due to the unequal distribution of males and females in each R groups.

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea: Based on the findings of the exploratory data analysis (EDA), the apparent racial pay gap could be further analyzed for multiple factors other than the ones that have been explored here. In this report we have considered WH, HD, and Sex and its effect on the pay gap between different R groups. WH analysis has shown that the difference in pay gap is not due to the difference in working hours. HD analysis, on the other hand, has pointed out that the difference in pay gap is due to the difference in HD. R groups that have shown high mean INC had a significantly higher mean HD, and vice versa. Sex, on the other hand, has shown that the unequal sex distribution in each R group has led to the difference in INC between R groups, pointing out the need to control features other than the feature of interest prior to analysis. Due to the ambivalence of multivariate analysis results, it is yet early to conclude whether the dataset supports or does not support the idea of racial pay gap. Therefore, a further analysis is necessary on multiple features that are known to directly/indirectly affect one's income.

In addition, a further in-depth analysis on the features analyzed in this report could be beneficial. Identifying whether the apparent racial pay gap is simply the reflection of sex/gender pay gap or in fact, truly due to HD, is only dependent on the in-depth analysis results for HD and Sex. If, through exploring the distribution of HD in males and females, we could prove that the difference in pay gap in males/females is also due to the difference in HD, we could argue that racial pay gap is a by-product of the pay-gap due to difference in HD.