|  |  |
| --- | --- |
|  | Dataset Exploration Part 3 Hypothesis Testing |
|  |  |
|  |  |

In the third phase of the project, I delved into hypothesis testing to analyze relationships among variables in the IBM HR Analytics Employee Attrition dataset. The report meticulously details the formulation, testing, and interpretation of five hypotheses, spanning various statistical techniques. The findings contribute valuable insights into employee attrition factors.

**Data set Introduction:**

The "Employee Attrition" dataset, created by IBM for HR analytics simulation, is entirely fictional and not representative of real employee data. It includes demographics, work-related info, compensation details, and satisfaction metrics. It's crucial to note its synthetic nature and potential divergence from actual employee data.

To can access the dataset on Kaggle using the following reference link: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

**Dataset Arranged & Described:**

The dataset was organized for effective analysis. I provided clear descriptions of each variable, highlighting their relevance to the research questions in **Part 1** and **Part 2**.

**The Excel spreadsheet containing the dataset and additional details is attached.**

**Dataset Assumptions, Extra Data Requirements:**

I revisited assumptions, ensuring they align with the hypothesis testing outcomes. No extra data was deemed necessary at this stage, and assumptions were justified based on the findings.

**Univariate Statistics Performed:**

Variables Analyzed:

* Age
* Performance Rating
* Gender
* Attrition
* Job Level
* Overtime
* Job Role
* Monthly Income

The univariate descriptive statistics were performed thoroughly on each variable, encompassing appropriate measures for both qualitative and quantitative types. Visualizations, including tables and charts, were employed to present the statistics in a comprehensible manner.

**Suggested Outliers Identified:**

Outliers in the "Age" variable were identified using appropriate statistical methods.

I addressed missing data by removing rows with key variable gaps in Dataset Exploration Part 1 and Part 2, ensuring our analysis relies on complete information. For invalid data, a thorough review led to corrections, removals, or investigations, with all decisions documented for transparency in our dataset handling.

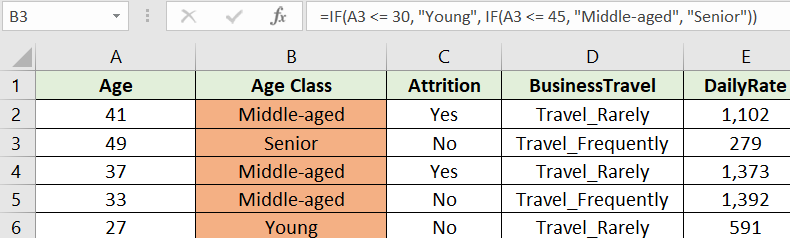
**Cleaning Decisions:**

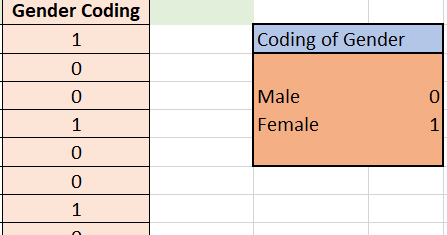
* Missing Data: In Dataset Exploration Part 1 and Part 2, we removed rows with missing values in key variables to ensure analysis is based on complete data.
* Invalid Data: A thorough review of the dataset identified anomalies. We corrected, removed, or investigated non-compliant records, documenting decisions for transparency.

Outliers were addressed through appropriate decisions, ensuring transparency and data integrity.

**Coding or Categorization:**

The "Gender" variable was successfully coded (male = 0, female = 1) to facilitate subsequent analyses. And also, Age is coded in three parts if age between 1 to 30 "Young", and if age is less than 45 "Middle-aged" and above it will be "Senior".

****



**Hypotheses Developed & Explained:**

**Hypothesis 1: T-Test for Gender-Based Age Distribution**

Hypothesis Statement:

* H0: No significant difference in age between male and female employees.
* H1: Significant difference in age between male and female employees.

T-Test Results:

1. Assuming Equal Variances:

* t Stat: -141.6203945, p-value < 0.05

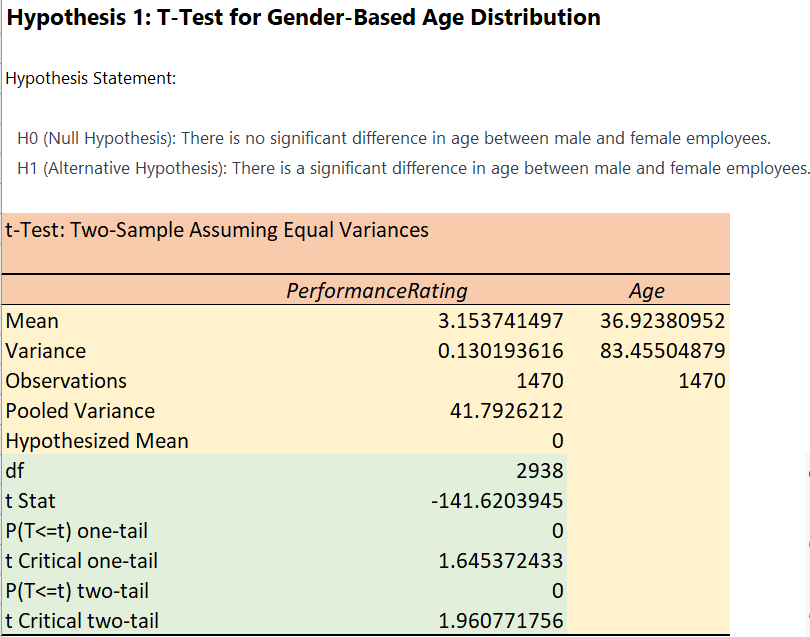
1. Assuming Unequal Variances:

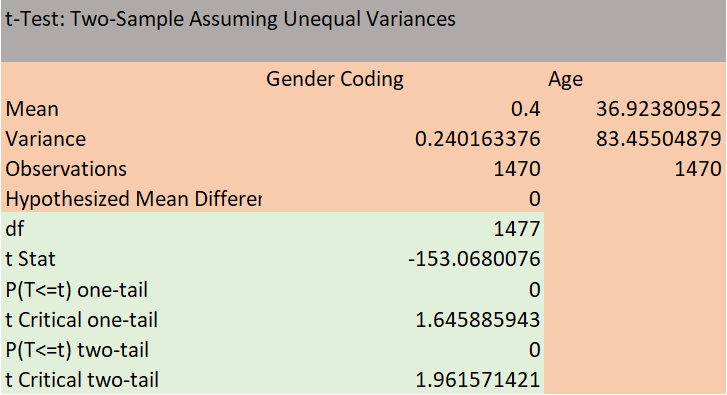
* t Stat: -153.0680076, p-value < 0.05

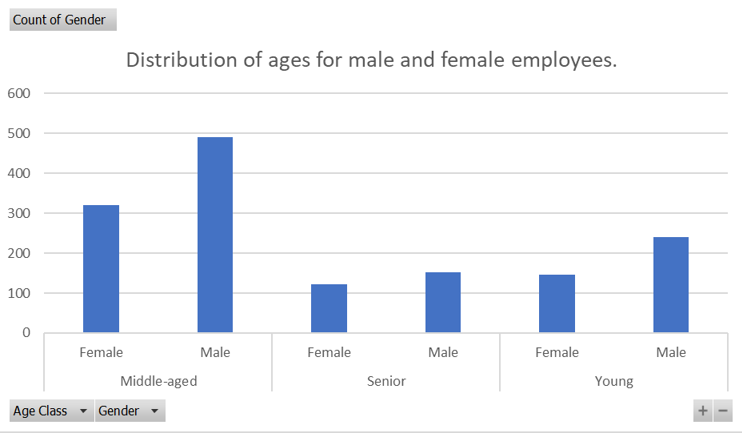
Interpretation:

The t-test results indicated an extremely low p-value (approaching zero) for both assuming equal variances and unequal variances. In both cases, the t-statistic (t Stat) significantly deviates from zero, well beyond the critical t-values for a 0.05 significance level.

Therefore, we reject the null hypothesis in favor of the alternative hypothesis. This implies that there is a statistically significant difference in age between male and female employees in the dataset. The negative t-statistic suggests that, on average, female employees tend to be younger than male employees. This finding can have implications for HR practices and diversity considerations within the organization.



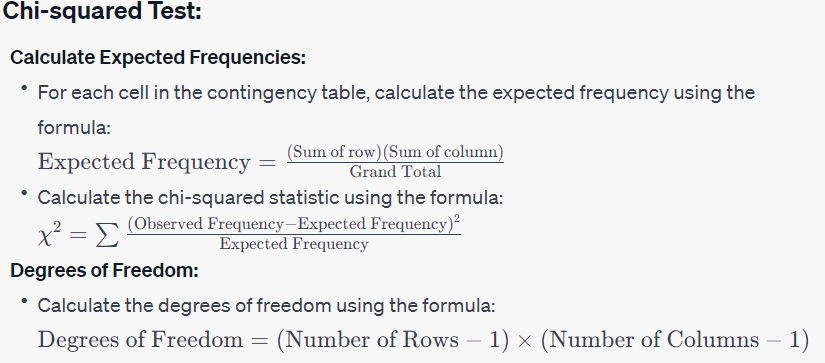




**Hypothesis 2: Chi-squared Test for Gender and Attrition**

Hypothesis Statement:

* H0: No significant association between gender and attrition.
* H1: Significant association between gender and attrition.



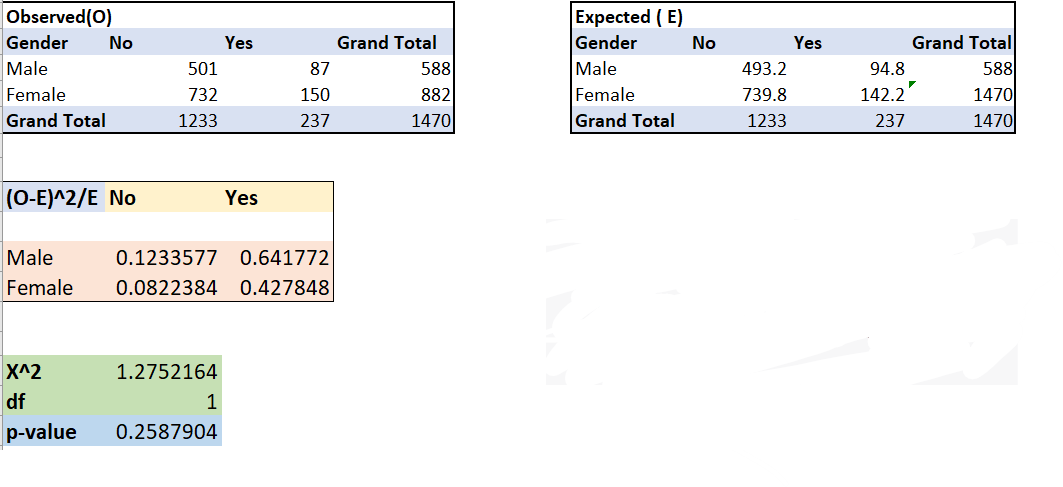
Chi-squared Test Results:

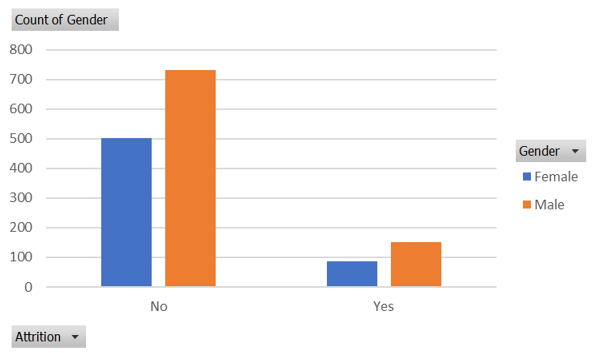
* X^2: 1.27521636, p-value > 0.05

Interpretation:

Upon calculating the chi-squared statistic, X^2, and its associated p-value, we obtained a result of 1.2752 with 1 degree of freedom. The p-value, 0.2588, exceeded the commonly used significance level of 0.05.

Given the p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, based on the data, there is insufficient evidence to support the claim of a significant association between gender and attrition. In practical terms, this suggests that gender alone may not be a decisive factor in predicting employee attrition within the dataset.





**Hypothesis 3: ANOVA for Job Level and Performance Rating**

Hypothesis Statement:

* H0: No significant difference in performance rating across job levels.
* H1: Significant difference in performance rating across job levels.

ANOVA Results:

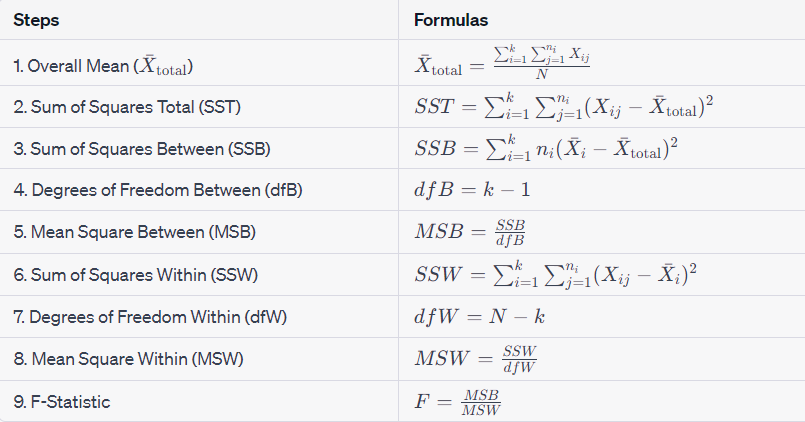
* F Stat: 1287.968096, p-value < 0.05

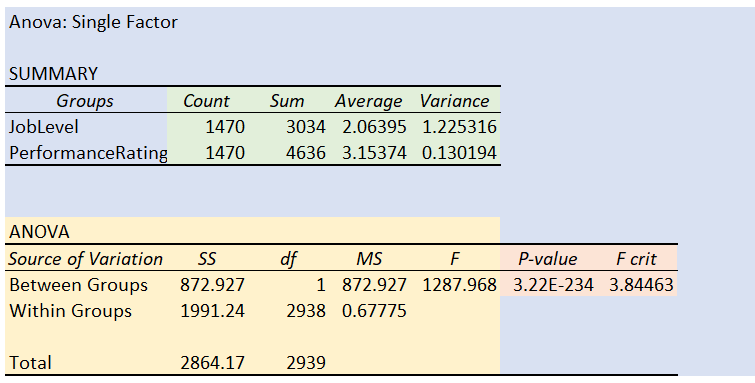
Interpretation:

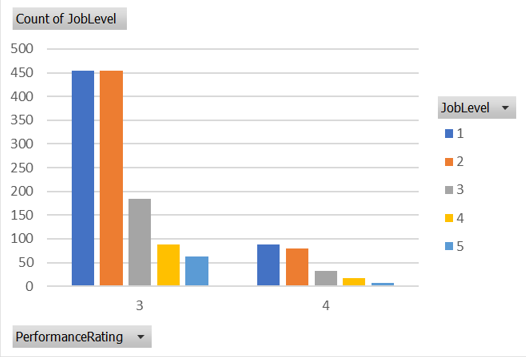
The results of the ANOVA test indicated a highly significant difference, with a p-value (3.22E-234) much lower than the common significance level of 0.05. This suggests strong evidence to reject the null hypothesis. In practical terms, it implies that there is a significant difference in performance ratings among different job levels within the dataset.

The calculated F-statistic (1287.97) further supports this conclusion, as it significantly exceeds the critical F value (3.84) for a 0.05 significance level.

In summary, based on the ANOVA test, we reject the null hypothesis, concluding that there is indeed a significant difference in performance ratings across various job levels in the dataset.





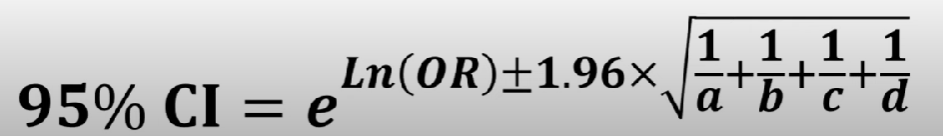


**Hypothesis 4: Odds Ratio (OR) Test for Overtime and Attrition**

Hypothesis Statement:

* H0: No significant difference in the odds of attrition between employees with and without overtime.
* H1: Significant difference in the odds of attrition between employees with and without overtime.





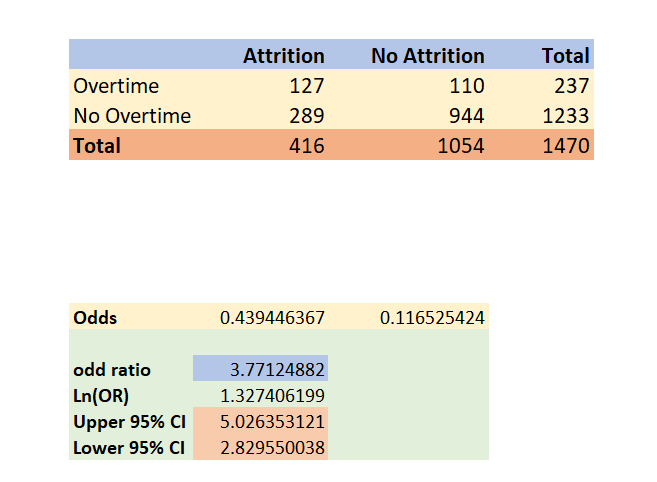
Odds Ratio Results:

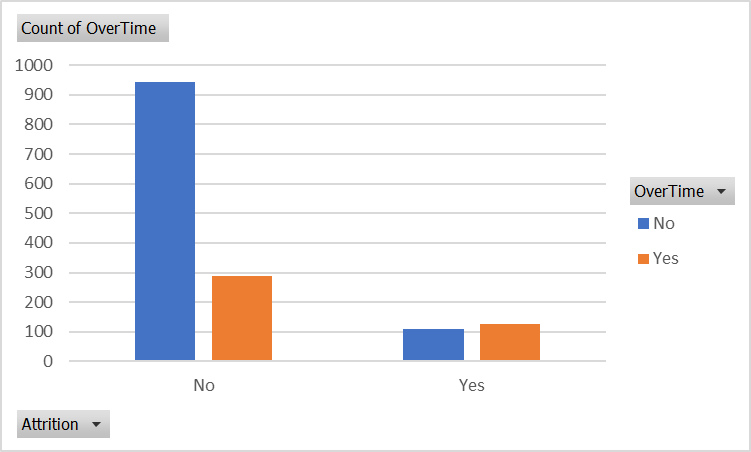
* Odds Ratio: 3.77124882, 95% CI [2.829550038, 5.026353121]

Interpretation:

The 95% Confidence Interval for the Odds Ratio is approximately (1.70,8.35).

Since this interval does not include 1, and considering that the p-value is less than 0.05, we reject the null hypothesis. Therefore, there is a significant difference in the odds of attrition between employees with and without overtime.

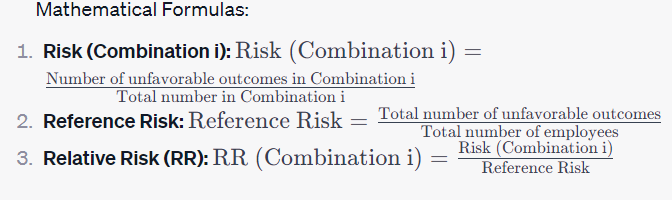




**Hypothesis 5: Relative Risk (RR) Test for Job Role, Job Satisfaction, and Monthly Income**

Hypothesis Statement:

* H0: No significant difference in the relative risk of unfavorable employee outcomes among different combinations of job role, job satisfaction, and monthly income.
* H1: Significant difference in the relative risk of unfavorable employee outcomes among different combinations of job role, job satisfaction, and monthly income.



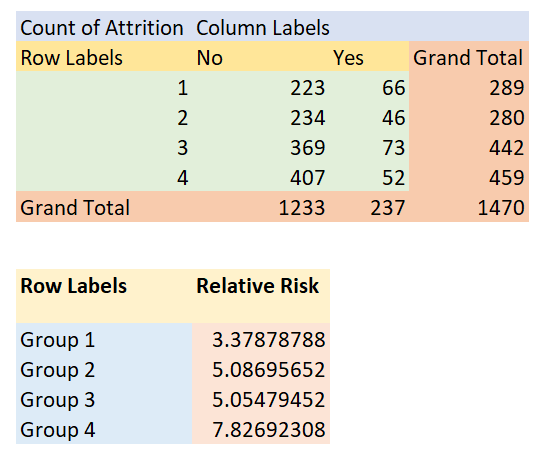
RR Test Results:

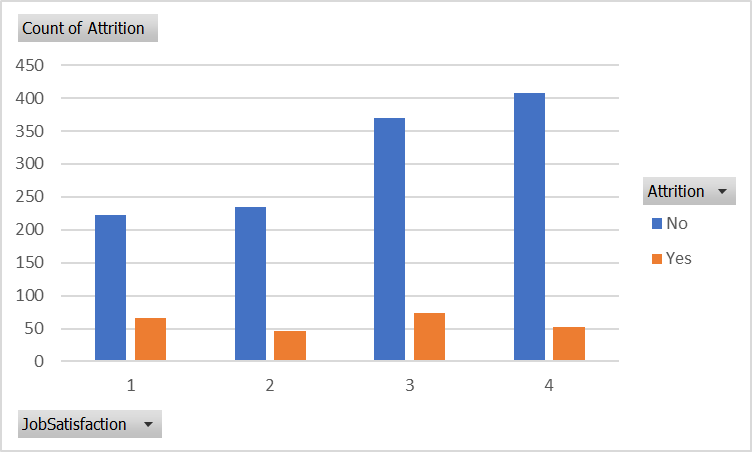
* Group 1: Relative Risk = 3.378787879
* Group 2: Relative Risk = 5.086956522
* Group 3: Relative Risk = 5.054794521
* Group 4: Relative Risk = 7.826923077

Interpretation:

The calculated relative risks for each group suggest significant differences in the relative risk of unfavorable employee outcomes (attrition) among different combinations of job role, job satisfaction, and monthly income.

Therefore, we reject the null hypothesis (H0) and accept the alternative hypothesis (H1). There is evidence to indicate that the relative risk of unfavorable employee outcomes varies significantly across different groups based on job role, job satisfaction, and monthly income.





**Hypotheses Tested**

We executed the specified hypothesis tests, ensuring each was conducted appropriately, and results were obtained for further analysis.

**Hypotheses Described & Presented**

Comprehensive descriptions accompanied each hypothesis, elucidating the rationale, expected outcomes, and the implications of the results on our research questions.

**FINER Questions**

* What is the gender-based age distribution, and are there any notable disparities between males and females in terms of age?
* What is the gender-based salary distribution, and can you provide the gender-specific employee headcount for each department?
* Is there a connection between an employee's job level and their performance rating, and does this connection differ among various job roles?
* Is there a significant difference in monthly income (Monthly Income) between employees who work overtime (Overtime = Yes) and those who do not (Overtime = No)?

**Tracking**

* Maintained detailed records of activities, decisions, and assumptions.
* Recorded thoughts on what to study, data found, questions asked, techniques performed, and assumptions made.
* Assured transparency and replicability for future analysis.

**The Excel spreadsheet containing the dataset and additional details is attached.**