Project Report: Conducting Advanced Statistical Tests Using Python

Introduction

In this project, I conducted various advanced statistical tests to analyze the Attrition Data from IBM. The tests performed include Chi-Square test, independent t-test, paired t-test, ANOVA, Pearson correlation, Spearman correlation, Mann-Whitney U test, and Wilcoxon signed-rank test. The objective was to determine relationships and differences in the data based on various statistical methods.

Methodology

1. **Data Collection and Preparation**
    - Load the Attrition Data from IBM.
    - Create groups based on the 'EducationField' and 'MonthlyIncome'.
    - Generate variables for the paired t-test and Wilcoxon signed-rank test.
2. **Performing Statistical Tests**
    - Chi-Square test for independence between 'Age' and 'MonthlyIncome' above the median.
    - Independent t-test to compare 'MonthlyIncome' between 'Medical' and 'Marketing' education fields.
    - Paired t-test on generated before and after data.
    - ANOVA to compare 'MonthlyIncome' across 'Medical', 'Marketing', and 'Technical Degree' education fields.
    - Pearson and Spearman correlation tests between 'Age' and 'MonthlyIncome'.
    - Mann-Whitney U test comparing 'MonthlyIncome' between 'Medical' and 'Marketing' education fields.

- Wilcoxon signed-rank test on generated before and after data.

3. **Interpreting Results**

- For each test, interpret the results by comparing the test statistic to a critical value or by examining the p-value.
- If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis.
- Discuss the conclusions drawn from the tests and any limitations or assumptions made.

Data Collection and Preparation

In [19]:
```
import numpy as np
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
```

In [20]:
```
# Loading the data
df = pd.read_csv("C:\\Datasets\\IBM Company Data\\IBM.csv")
```

In [21]:
```
#checking if there are any null values
df.isnull().sum()
```

Out[21]:
```
Age                        0
Attrition                  0
Department                 0
DistanceFromHome           0
Education                  0
EducationField             0
EnvironmentSatisfaction    0
JobSatisfaction            0
MaritalStatus              0
MonthlyIncome              0
NumCompaniesWorked         0
WorkLifeBalance            0
YearsAtCompany             0
dtype: int64
```

In [22]:
```
# Basic statistical inference using describe()
df.describe()
```

Out[22]:

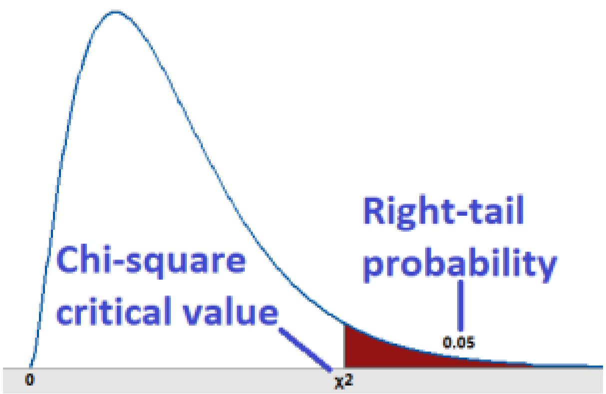
	Age	DistanceFromHome	Education	EnvironmentSatisfaction	JobSatisfaction	Mc
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	
mean	36.923810	9.192517	2.912925	2.721769	2.728571	
std	9.135373	8.106864	1.024165	1.093082	1.102846	
min	18.000000	1.000000	1.000000	1.000000	1.000000	
25%	30.000000	2.000000	2.000000	2.000000	2.000000	
50%	36.000000	7.000000	3.000000	3.000000	3.000000	
75%	43.000000	14.000000	4.000000	4.000000	4.000000	
max	60.000000	29.000000	5.000000	4.000000	4.000000	

Statistical Tests

In [23]:
```
df.columns
```

Out[23]:
```
Index(['Age', 'Attrition', 'Department', 'DistanceFromHome', 'Education',
       'EducationField', 'EnvironmentSatisfaction', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked',
       'WorkLifeBalance', 'YearsAtCompany'],
      dtype='object')
```

Chi-Square Test



Description: The Chi-Square test is used to test the independence of two categorical variables.

Equation: $\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right)$ where:

- O_i is the observed frequency.
- E_i is the expected frequency.

Assumptions:

- The observations are independent.
- The sample size is large enough.

Implementation of CHI SQUARE IN PYTHON

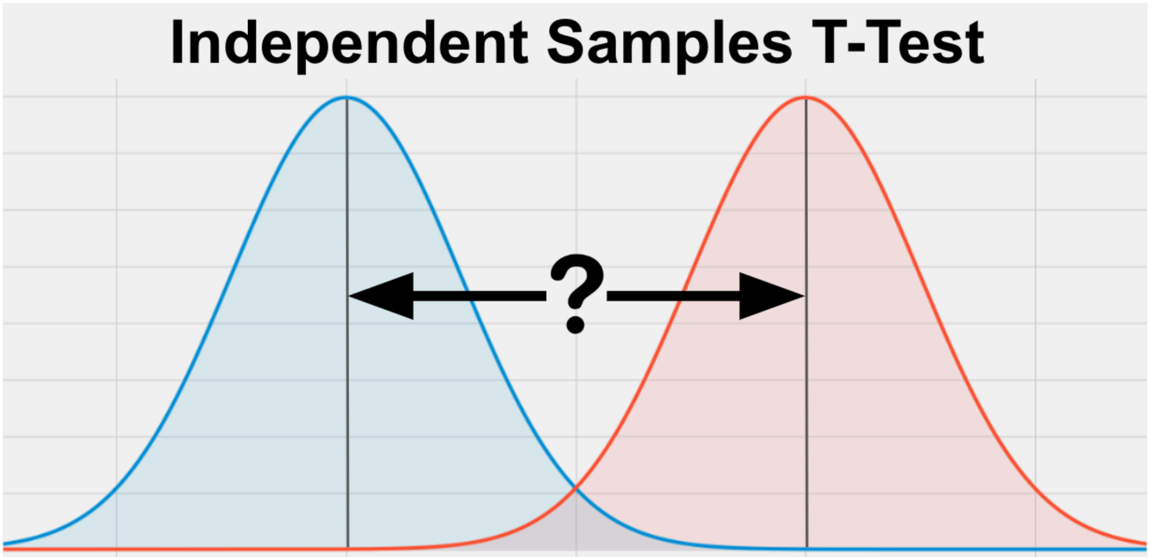
In [24]:
```python
# I'll be checking whether age ,education and monthly income are independent
contingency_table = pd.crosstab(df['Age'], df['MonthlyIncome'] > df['MonthlyIncome'
# I'll set the alpha to be 0.05
alpha = 0.05
# Perform Chi-Square test
chi2_stat, p_value, dof, ex = stats.chi2_contingency(contingency_table)
print(f"Chi-Square Test: chi2-statistic = {chi2_stat}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: The variables are not independent.")
else:
    print("Fail to reject the null hypothesis: The variables are independent.")
```

```
Chi-Square Test: chi2-statistic = 271.82908453974386, p-value = 2.0929211797239296
e-35
Reject the null hypothesis: The variables are not independent.
```

T TEST

Independent T-test



Description: The independent t-test compares the means of two independent groups to determine if there is statistical evidence that the associated population means are significantly different.

Equation: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$ where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means.
- S_1^2 and S_2^2 are the sample variances.
- N_1 and N_2 are the sample sizes.

Assumptions:

- The samples are independent.
- The populations are normally distributed.
- The variances of the populations are equal.

In [25]:
```
df.columns
```

Out[25]:
```
Index(['Age', 'Attrition', 'Department', 'DistanceFromHome', 'Education',
       'EducationField', 'EnvironmentSatisfaction', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked',
       'WorkLifeBalance', 'YearsAtCompany'],
      dtype='object')
```

In [26]:
```
df['EducationField'].value_counts().head()
```

Out[26]:
```
Life Sciences       606
Medical             464
Marketing           159
Technical Degree    132
Other                82
Name: EducationField, dtype: int64
```

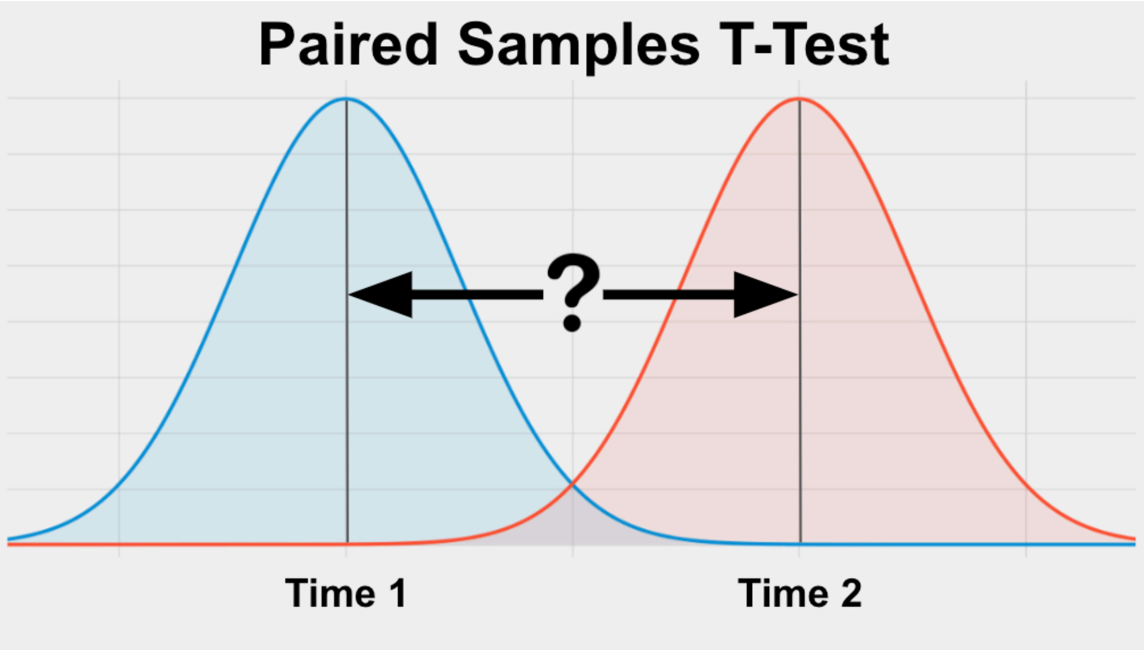
In [27]:
```
# I am  using EducationField and Monthly Income
group1 = df[df['EducationField'] == 'Medical']['MonthlyIncome']
group2 = df[df['EducationField'] == 'Marketing']['MonthlyIncome']
group3 = df[df['EducationField'] == 'Technical Degree']['MonthlyIncome']

# Perform independent t-test
t_stat, p_value = stats.ttest_ind(group1, group2)
print(f"Independent t-test: t-statistic = {t_stat}, p-value = {p_value}")

# Interpretation
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The means are significantly different.")
else:
    print("Fail to reject the null hypothesis: The means are not significantly diff
```

```
Independent t-test: t-statistic = -1.9261145657938827, p-value = 0.054546214645424
35
Fail to reject the null hypothesis: The means are not significantly different.
```

Paired t-test



Description: The paired t-test compares the means of two related groups to determine if there is a significant difference between the means.

Equation: $t = \frac{\bar{D}}{S_D/\sqrt{N}}$ where:

- $\bar{D}$ is the mean of the differences.
- S_D is the standard deviation of the differences.
- N is the number of pairs.

Assumptions:

- The pairs are randomly sampled.
- The differences are normally distributed

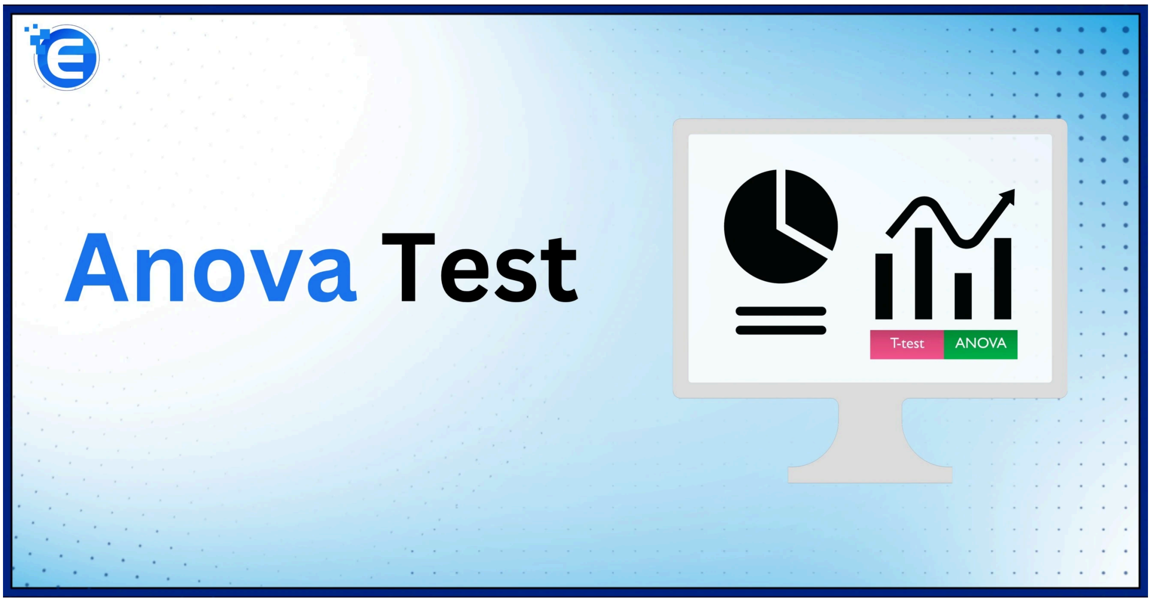
In [28]:
```
#I will Generate sample data for paired t-test ,since its a demonstration
# use numpy random variable generation ,create two variable with similar  length ar
np.random.seed(0)
before = np.random.normal(10, 1, 30)
after = before + np.random.normal(0.5, 1, 30)

# Perform paired t-test
t_stat, p_value = stats.ttest_rel(before, after)
print(f"Paired t-test: t-statistic = {t_stat}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: The means are significantly different.")
else:
    print("Fail to reject the null hypothesis: The means are not significantly diff
```

```
Paired t-test: t-statistic = -1.2609797789472506, p-value = 0.21736669382400253
Fail to reject the null hypothesis: The means are not significantly different.
```

ANOVA



Description: ANOVA tests the null hypothesis that the means of several groups are equal. It is used to compare the means of three or more groups.

Equation: $F = \frac{MS_{between}}{MS_{within}}$ where:

- $MS_{between}$ is the mean square between the groups.
- MS_{within} is the mean square within the groups.

Assumptions:

- The samples are independent.
- The populations are normally distributed.
- The variances of the populations are equal.

In [29]:
```
df.columns
```

Out[29]:
```
Index(['Age', 'Attrition', 'Department', 'DistanceFromHome', 'Education',
       'EducationField', 'EnvironmentSatisfaction', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked',
       'WorkLifeBalance', 'YearsAtCompany'],
      dtype='object')
```

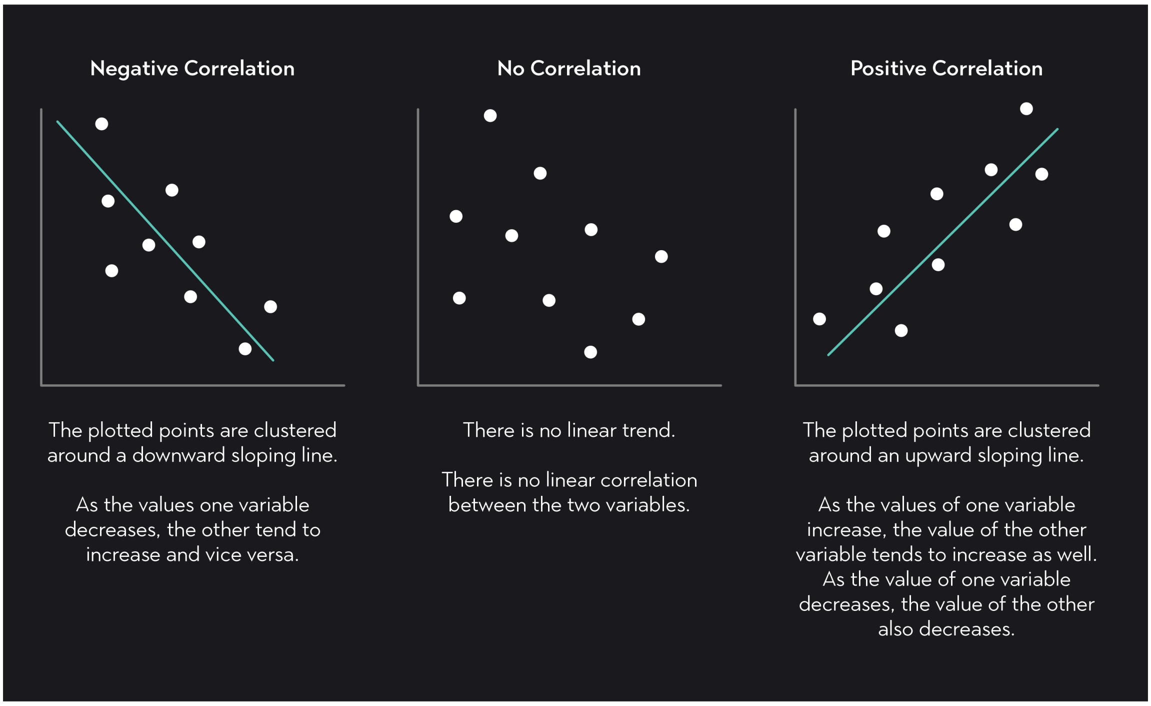
In [30]:
```
# Since its used to compare the means of three or more groups i will create the thr
group1 = df[df['EducationField'] == 'Medical']['MonthlyIncome']
group2 = df[df['EducationField'] == 'Marketing']['MonthlyIncome']
group3 = df[df['EducationField'] == 'Technical Degree']['MonthlyIncome']
# perfoming a one way Anova
'''f_stat, p_value = stats.f_oneway( df[df['EducationField'] == 'Medical']['Monthly
                                 df[df['EducationField'] == 'Marketing']['MonthlyIr
                                  df[df['EducationField'] == 'Technical Degree']['M


f_stat, p_value = stats.f_oneway(group1,group2,group3)
print(f"ANOVA: F-statistic = {f_stat}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant difference in means a
else:
    print("Fail to reject the null hypothesis: There is no significant difference i
```

```
ANOVA: F-statistic = 4.294905537862313, p-value = 0.013973938677770493
Reject the null hypothesis: There is a significant difference in means among the g
roups.
```

CORRELATION TESTS



Pearson Test

Description: The Pearson correlation measures the linear relationship between two continuous variables.

Equation: $r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$ where:

- X and Y are the variables.
- $\bar{X}$ and $\bar{Y}$ are the means of the variables.

Assumptions:

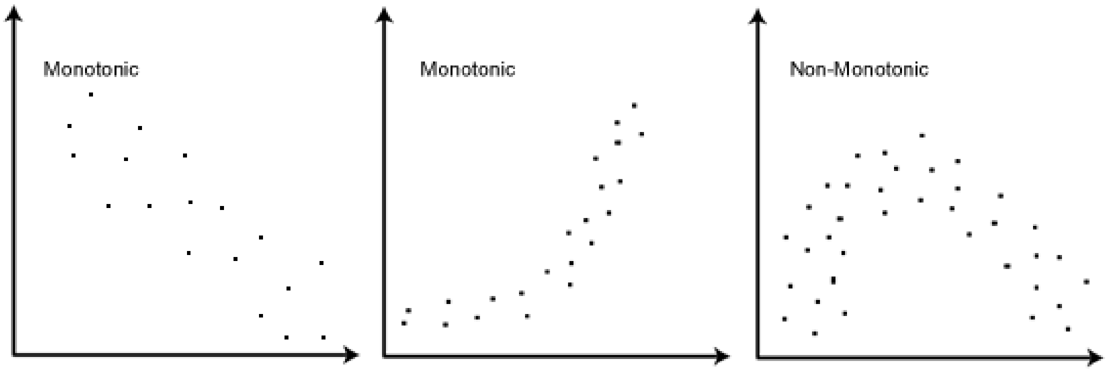
- The relationship is linear.
- The variables are normally distributed.
- Homoscedasticity (constant variance of errors).

In [31]:
```
# Since it measures the relationship between two continous variable ,I will check f
# Pearson correlation
pearson_corr, p_value = stats.pearsonr(df['Age'], df['MonthlyIncome'])
print(f"Pearson Correlation: correlation = {pearson_corr}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant correlation.")
else:
    print("Fail to reject the null hypothesis: There is no significant correlation.
```

```
Pearson Correlation: correlation = 0.4978545669265804, p-value = 6.669539203000309
5e-93
Reject the null hypothesis: There is a significant correlation.
```

SpearMans Rank Correlation



Description: The Spearman correlation measures the monotonic relationship between two continuous or ordinal variables.

Equation: $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$ where:

- d_i is the difference between the ranks of corresponding variables.
- n is the number of observations.

Assumptions:

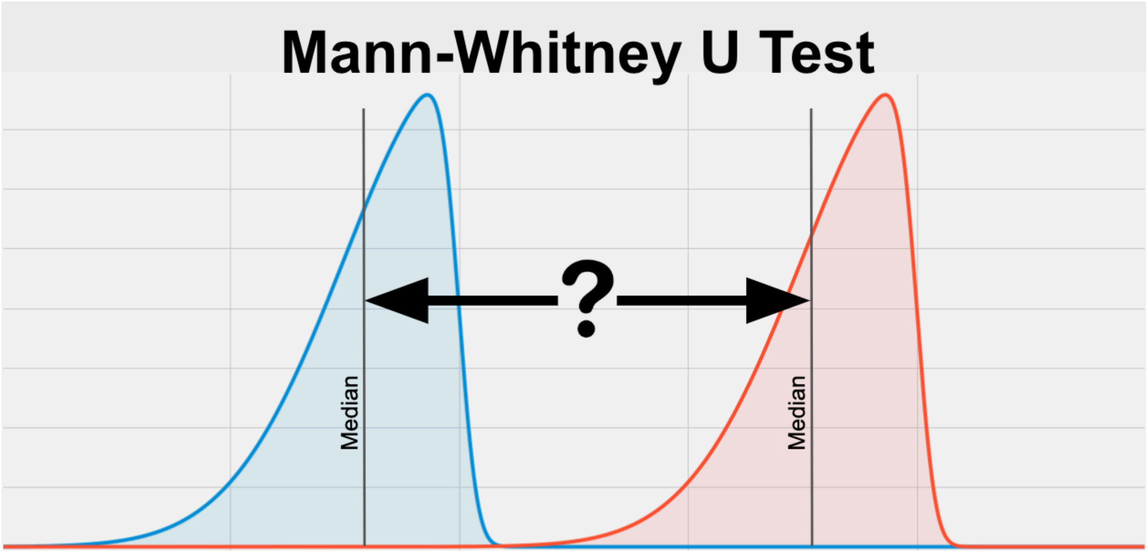
- The relationship is monotonic.

In [32]:
```
spearman_corr, p_value = stats.spearmanr(df['Age'], df['MonthlyIncome'])
print(f"Spearmans Correlation: correlation = {spearman_corr}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant correlation.")
else:
    print("Fail to reject the null hypothesis: There is no significant correlation.
```

```
Spearmans Correlation: correlation = 0.47190213023271405, p-value = 2.183456092645
1124e-82
Reject the null hypothesis: There is a significant correlation.
```

MANN - WHITNEY U TEST



Description: The Mann-Whitney U test compares the distributions of two independent groups to determine if they come from the same distribution.

Equation: $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ where:

- n_1 and n_2 are the sample sizes.
- R_1 is the sum of the ranks for group 1.

Assumptions:

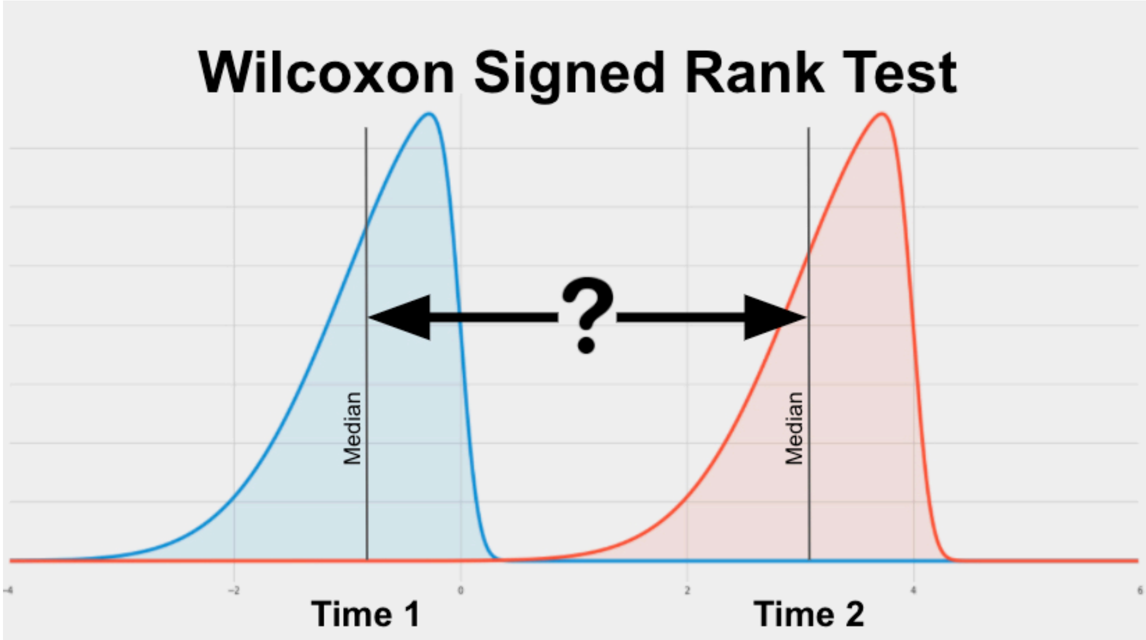
- The samples are independent.
- The distributions of the groups are similar in shape.

In [33]:
```
# It compares the distribution of two groups to check if the groups come from the s
u_stat, p_value = stats.mannwhitneyu(group1, group2)
print(f"Mann-Whitney U Test: U-statistic = {u_stat}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: The distributions are significantly differer
else:
    print("Fail to reject the null hypothesis: The distributions are not significar
```

```
Mann-Whitney U Test: U-statistic = 28603.5, p-value = 2.3429301656791918e-05
Reject the null hypothesis: The distributions are significantly different.
```

WILCOXON SIGNED -RANK TEST



Description: The Wilcoxon signed-rank test compares the distributions of two related groups to determine if they come from the same distribution.

Equation: $W = \sum T_i$ where:

- T_i is the signed rank of the differences.

Assumptions:

- The pairs are randomly sampled.
- The differences are symmetrically distributed.

In [34]:
```
# Wilcoxon signed-rank test
# i will use The random data I create with np.random
w_stat, p_value = stats.wilcoxon(before, after)
print(f"Wilcoxon Signed-Rank Test: W-statistic = {w_stat}, p-value = {p_value}")

# Interpretation
if p_value < alpha:
    print("Reject the null hypothesis: The distributions are significantly differer
else:
    print("Fail to reject the null hypothesis: The distributions are not significar
```

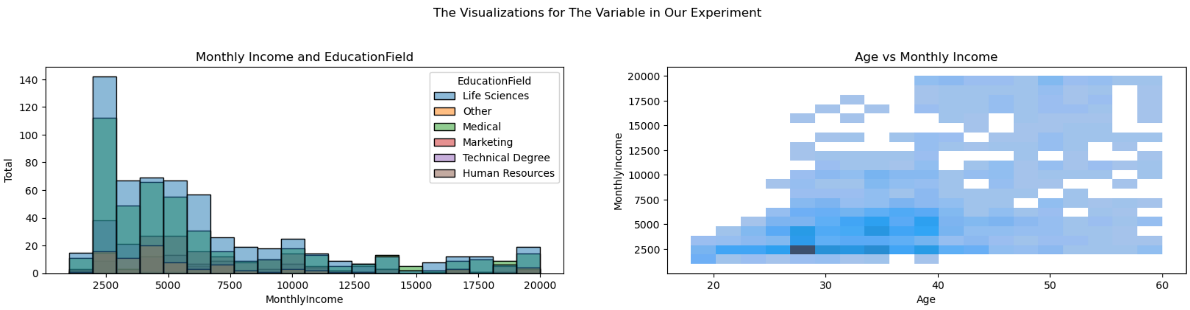
```
Wilcoxon Signed-Rank Test: W-statistic = 175.0, p-value = 0.24494642950594425
Fail to reject the null hypothesis: The distributions are not significantly differ
ent.
```

VISUALIZATIONS FOR ALL THAT WE HAVE COMPARED

In [35]:
```
import seaborn as sns
```

In [36]:
```
fig= plt.figure(figsize=(20,8))
fig.suptitle('The Visualizations for The Variable in Our Experiment')
plt.subplot(2,2,1)
sns.histplot(x='MonthlyIncome',hue='EducationField',data=df)
plt.ylabel("Total")
```

```python
plt.title('Monthly Income and EducationField')
plt.subplot(2,2,2)
sns.histplot(y='MonthlyIncome',x='Age',data=df)
plt.title('Age vs Monthly Income')
plt.show()
```



CONCLUSION AND RESULTS

Results

Chi-Square Test

Variables Tested: 'Age', 'MonthlyIncome' > median

Results: $[\chi^2 = 271.829, \quad p\text{-value} = 2.093 \times 10^{-35}]$

Conclusion: Reject the null hypothesis: The variables are not independent.

Independent t-test

Groups Tested: 'Medical' vs 'Marketing'

Results: $[t = -1.926, \quad p\text{-value} = 0.055]$

Conclusion: Fail to reject the null hypothesis: The means are not significantly different.

Paired t-test

Variables Tested: before and after

Results: $[t = -1.261, \quad p\text{-value} = 0.217]$

Conclusion: Fail to reject the null hypothesis: The means are not significantly different.

ANOVA

Groups Tested: 'Medical', 'Marketing', 'Technical Degree'

Results: $[F = 4.295, \quad p\text{-value} = 0.014]$

Conclusion: Reject the null hypothesis: There is a significant difference in means among the groups.

Pearson Correlation

Variables Tested: 'Age' vs 'MonthlyIncome'

Results: $[r = 0.498, \quad p\text{-value} = 6.670 \times 10^{-93}]$

Conclusion: Reject the null hypothesis: There is a significant correlation.

Spearman Correlation

Variables Tested: 'Age' vs 'MonthlyIncome'

Results: $[\rho = 0.472, \quad p\text{-value} = 2.183 \times 10^{-82}]$

Conclusion: Reject the null hypothesis: There is a significant correlation.

Mann-Whitney U Test

Groups Tested: 'Medical' vs 'Marketing'

Results: $[U = 28603.5, \quad p\text{-value} = 2.343 \times 10^{-5}]$

Conclusion: Reject the null hypothesis: The distributions are significantly different.

Wilcoxon Signed-Rank Test

Variables Tested: before and after

Results: $[W = 175.0, \quad p\text{-value} = 0.245]$

Conclusion: Fail to reject the null hypothesis: The distributions are not significantly different.

Conclusion

In this project, I conducted various advanced statistical tests on the Attrition Data from IBM to understand relationships and differences in the data. The results demonstrated the importance of selecting appropriate statistical tests based on data characteristics and assumptions. This project highlights the significance of statistical analysis in data-driven decision-making.