# Statistical tests in Attrition analysis.

1. Mann-Whitney test on YearsAtCompany between Attrition and Non-Attrition employees.

(For this analysis, the given .csv file was converted into .xlsx file and filtering was done to segregate between Attrition and Non-Attrition YearsAtCompany)

**Hypothesis:**

H0 -> There is no significant difference between YearsAtCompany for Attrition and NonAttrition

H1 -> These is significant difference between YearsAtCompany for Attrition and NonAttrition

*import pandas as pd*

*from scipy.stats import mannwhitneyu*

*# loading dataset*

*dataset = pd.* *read\_excel("general\_data.xlsx", sheet\_name=1)*

*# data treatment*

*dataset.isnull()*

*dataset.dropna(inplace=True)*

*# statistical test*

*stats,p = mannwhitneyu(dataset.* Attrition*, dataset.* NonAttrition*)*

*print(stats,p)*

*177883.5 1.274723628781325e-22*

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *p value* | *1.274723628781325e-22* | *Less than 0.05, thus NULL HYPOTHESIS IS REJECTED* |

1. Chi-square test on Attrition and Education.

**Hypothesis:**

H0 -> There is no dependency between Attrition and Education

H1 -> These is dependency between Attrition and Education

*import pandas as pd*

*from scipy.stats import chi2\_contingency*

*from sklearn.preprocessing import LabelEncoder*

*# loading dataset*

*dataset = pd.read\_csv("general\_data.csv")*

*# data treatment*

*dataset.isnull()*

*dataset.dropna(inplace=True)*

*# converting categorical string data into integers*

*number = LabelEncoder()*

*dataset["Attrition"] = number.fit\_transform(dataset["Attrition"].astype("str"))*

*# statistical test*

*chitable = pd.crosstab(dataset.Attrition, dataset.Education)*

*print(chitable)*

*stats, p, dof, expected = chi2\_contingency(chitable)*

*print(stats, p)*

*6.2735361781270615 0.17963050984273224*

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *p value* | *0.17963050984273224* | *Greater than 0.05, thus NULL HYPOTHESIS IS ACCEPTED* |

1. One-Sample t test on sample and population mean of Age.

**Hypothesis:**

H0 -> There is no significant difference of the mean of employee’s age against population mean = 37

H1 -> There is significant difference of the mean of employee’s age against population mean = 37

*import pandas as pd*

*from scipy.stats import ttest\_1samp*

*# loading dataset*

*dataset = pd.read\_csv("general\_data.csv")*

*# data treatment*

*dataset.isnull()*

*dataset.dropna(inplace=True)*

*# statistical test*

*stats, p = ttest\_1samp(dataset.Age, 37)*

*print(stats, p)*

*-0.48275870732389337 0.6292911944360575*

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *p value* | *0.6292911944360575* | *Greater than 0.05, thus NULL HYPOTHESIS IS ACCEPTED* |

1. Two-Sample independent t test on mean of PercentSalaryHike for Attrition and NonAttrition.

(For this analysis, the given .csv file was converted into .xlsx file and filtering was done to segregate between Attrition and Non-Attrition PercentSalaryHike)

**Hypothesis:**

H0 -> There is no significant difference in the mean of PercentSalaryHike between Attrition and NonAttrition employees

H1 -> There is significant difference in the mean of PercentSalaryHike between Attrition and NonAttrition employees

*import pandas as pd*

*from scipy.stats import ttest\_ind*

*# loading dataset*

*dataset = pd.read\_excel("general\_data.xlsx",sheet\_name=2)*

*# data treatment*

*dataset.isnull()*

*dataset.dropna(inplace=True)*

*# statistical test*

*stats,p = ttest\_ind(dataset.Attrition, dataset.NonAttrition)*

*print(stats,p)*

*1.2154526422906522 0.22439553878006924*

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *p value* | *0.22439553878006924* | *Greater than 0.05, thus NULL HYPOTHESIS IS ACCEPTED* |

1. One way ANOVA with PercentSalaryHike as dependent variable and Education as independent variable.

**Hypothesis:**

H0 -> There is no significant impact on PercentSalaryHike due to Education.

H1 -> There is significant impact on PercentSalaryHike due to Education.

*import pandas as pd*

*dataset = pd.read\_excel("general\_data.xlsx", sheet\_name=0)*

*import statsmodels.api as sm*

*from statsmodels.formula.api import ols*

*model = ols('PercentSalaryHike~C(Education)', dataset).fit()*

*oneway = sm.stats.anova\_lm(model, typ=2)*

*print(oneway)*

*# User defined function to calculate eta square*

*def anova\_table(oneway):*

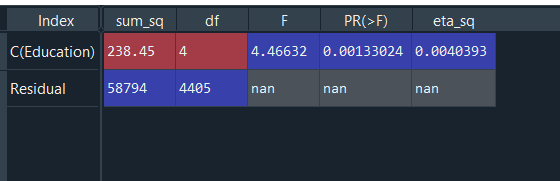
*oneway["eta\_sq"] = oneway[:-1]["sum\_sq"]/sum(oneway["sum\_sq"])*

*cols = ["sum\_sq", "df", "F", "PR(>F)", "eta\_sq"]*

*oneway = oneway[cols]*

*return oneway;*

*anova\_table(oneway)*

**

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *p value* | *0.0*0133024 | *L*esser *than 0.05, thus NULL HYPOTHESIS IS REJECTED* |

1. Two way ANOVA with PercentSalaryHike as dependent variable and Education, JobLevel as independent variables.

**Hypothesis:**

H0 -> There is no significant impact on PercentSalaryHike due to Education and JobLevel.

H1 -> There is significant impact on PercentSalaryHike due to Education and JobLevel.

*import pandas as pd*

*dataset = pd.read\_excel("general\_data.xlsx", sheet\_name=0)*

*import statsmodels.api as sm*

*from statsmodels.formula.api import ols*

*model = ols('PercentSalaryHike~C(Education)+C(JobLevel)', dataset).fit()*

*twoway = sm.stats.anova\_lm(model, typ=2)*

*print(twoway)*

*# User defined function to calculate eta square*

*def anova\_table(twoway):*

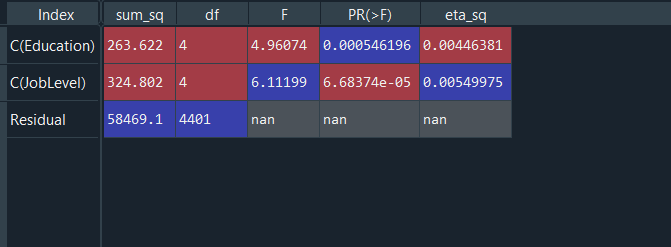
*twoway ["eta\_sq"] = twoway [:-1]["sum\_sq"]/sum(twoway ["sum\_sq"])*

*cols = ["sum\_sq", "df", "F", "PR(>F)", "eta\_sq"]*

*twoway = twoway [cols]*

*return twoway;*

*anova\_table(twoway)*

**

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *P(Education) value*  *P(JobLevel) value* | *0.0*0054  *0.00000683774* | *L*esser *than 0.05, thus NULL HYPOTHESIS IS REJECTED* |

1. Multi way ANCOVA with PercentSalaryHike as dependent variable and Education, JobLevel, YearsAtCompany as independent variables.

**Hypothesis:**

H0 -> There is no significant impact on PercentSalaryHike due to Education, JobLevel and YearsAtCompany.

H1 -> There is significant impact on PercentSalaryHike due to Education, JobLevel and YearsAtCompany.

*import pandas as pd*

*dataset = pd.read\_excel("general\_data.xlsx", sheet\_name=0)*

*import statsmodels.api as sm*

*from statsmodels.formula.api import ols*

*model = ols('PercentSalaryHike~C(Education)+C(JobLevel)+* *YearsAtCompany', dataset).fit()*

*ancova = sm.stats.anova\_lm(model, typ=2)*

*print(ancova)*

*# User defined function to calculate eta square*

*def anova\_table(ancova):*

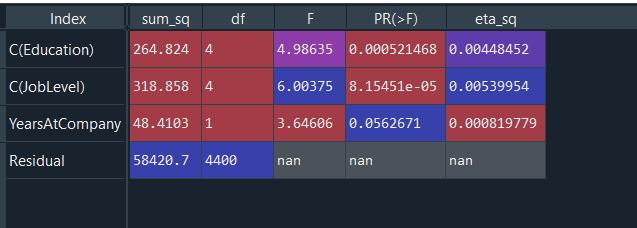
*ancova ["eta\_sq"] = ancova [:-1]["sum\_sq"]/sum(ancova ["sum\_sq"])*

*cols = ["sum\_sq", "df", "F", "PR(>F)", "eta\_sq"]*

*ancova = ancova [cols]*

*return ancova;*

*anova\_table(ancova)*

**

*Analysis:*

|  |  |  |
| --- | --- | --- |
| *Parameter* | *Value* | *Inference* |
| *P(Education) value*  *P(JobLevel) value*  *P(YearsAtCompany)* | *0.0*0054  *0.0000081545*  *0.056* | *L*esser *than 0.05 for except YearsAtCompany, thus NULL HYPOTHESIS IS REJECTED FOR EDUCTAION AND JOBLEVEL, BUT ACCEPTED FOR YEARSATCOMPANY* |