Assignment Day 12

* Importing the dataset and cleaning the data by removing the null values.
* @author: Akash M Khyadi

"""

import pandas as pd

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

print(ds.head())

#Cleaning the data set.

#Removing null values

print(ds.isnull().sum())

cds=ds.dropna()

print(cds.head())

#Checking and removing duplicate values if available.

print(cds.duplicated().sum())

cds1=cds

#No Duplicates found.

* #Based on the given current data in order to find attrition causing variables.

#Converting Attrition from categorical to binomial values

cds["Attrition"].replace(to\_replace=("Yes","No"),value=(1,0),inplace=True)

print(cds.head())

* #Processing wilcoxon test to check if a particular variable

**H0= There is no significant attrition in the company caused by the age of the employees and the distance from home.**

**H1= There is significant attrition in the company caused by the age of the employees and the distance from home.**

from scipy.stats import wilcoxon

print("\n")

print(wilcoxon(cds.Age,cds.DistanceFromHome))

**Output**: Wilcoxon Result (statistic=2892.0, pvalue=0.0)

**H0 is unpredictable as there is no additional numeric after the first zero in p-value.**

* **Friedman test**

***H0 – There is no attrition in company relating to the employees Age, DistanceFromHome and their JobLevel.***

***H1 – There is an attrition in company relating to the employees Age, DistanceFromHome and their JobLevel.***

from scipy.stats import friedmanchisquare

print("\n")

print(friedmanchisquare(cds.Age,cds.DistanceFromHome,cds.JobLevel))

**Output:** FriedmanchisquareResult(statistic=7478.114178094286, pvalue=0.0)

**H0 is unpredictable as there is no additional numeric after the first zero in p-value.**

* **Mann Whitney test**

***H0 – There is no attrition in company relating to the employees JobLevel and NumCompaniesWorked.***

***H1 – There is a significant attrition in company relating to the employees JobLevel and NumCompaniesWorked.***

from scipy.stats import mannwhitneyu

print("\n")

print(mannwhitneyu(cds.JobLevel,cds.NumCompaniesWorked))

**Output:** Mannwhitneyu Result(statistic=9550590.0, pvalue=0.3301819742895579)

**H0 is accepted as the p-value is greater than 0.05.**

* **Krushkal Wallis Test**

***H0 – There is no attrition in company relating to the employees TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany and their YearsSinceLastPromotion.***

***H1 – There is a significant attrition in company relating to the employees TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany and their YearsSinceLastPromotion.***

from scipy.stats import kruskal

print("\n")

print(kruskal(cds.TotalWorkingYears,cds.TrainingTimesLastYear,cds.YearsAtCompany,cds.YearsSinceLastPromotion))

**Output:** KruskalResult (statistic=7162.787322762321, pvalue=0.0)

**H0 is unpredictable as there is no additional numeric after the first zero in p-value.**

* **ChiSquare Test**

***H0 - There is no dependency between the Attrition in the company with the Gender of the Employee.***

***H1 - There is dependency between the Attrition in the company with the Gender of the Employee.***

from scipy.stats import chi2\_contingency

print("\n")

chitab=pd.crosstab(cds1.Attrition,cds1.Gender)

print(chitab)

print(chi2\_contingency(chitab))

|  |  |  |
| --- | --- | --- |
| *Gender* ->  **Attrition** v | Female | Male |
|  |  |
| 0 | 1488 | 2189 |
| 1 | 268 | 437 |

**Output:** ChiSquare (Statistics: 1.3825823839528295,p = 0.23966176275638887)

**H0 is accepted as the p-value is grater than 0.05.**

* **One Sample t test**

***H0 – There is no significant difference of mean of employees Age again the population mean = 37.***

***H1 – There is a significant difference of mean of employees Age again the population mean = 37.***

from scipy.stats import ttest\_1samp

print("\n")

print(ttest\_1samp(cds.Age,37))

print("\n",cds['Age'].mean())

**Output:**

Ttest\_1sampResult (statistic=-0.48275870732389337,pvalue=0.6292911944360575)

**H0 is accepted as the p-value is greater than 0.05.**

* **Two Sample Paired t-test.**

***H0 – There is no significant difference between the mean of Education and the JobLevel of the employee.***

***H1 – There is a significant difference between the mean of Education and the JobLevel of the employee.***

from scipy.stats import ttest\_rel

print("\n")

print(ttest\_rel(cds.Education,cds.JobLevel))

**Output:**

Ttest\_relResult(statistic=38.13071729060291, pvalue=5.548448125213481e-275)

**H0 is rejected as the p-value is less than 0.05.**

* **Two Sample Independent t-test.**

***Unable to perform this test.***