2022

Assignment-1

INTRODUCTION TO DATA SCIENCE

Course: Introduction to Data Science
Submitted To: Muhammad Sharjeel
Submitted By: Muhammad Faizan
Student Id: CIIT/SP20-BCS-101/LHR
Group: G2

The Hello Dataset

Data Normalization - Regex - Data Visualization

Changes

1. Cleaning the Symbols from HSSC-1 marks (Like: obtained_marks/total_marks format, grade in % percentage to obtained_marks format i.e 60% to 300)

Method:

```
marks = []
for i in range(len(data['HSSC-1'])):
    marks = data['HSSC-1'][i]
    if ('/' in str(marks)):
        obtained_marks, total_marks = marks.split('/')
        data['HSSC-1'][i] = obtained_marks
    if '%' in str(marks):
        marks, null = marks.split('%')
        obtained_marks = int((int(marks) / 100) * 510)
        data['HSSC-1'][i] = str(obtained_marks)
```

2. Cleaning the Symbols from HSSC-2 marks (Like: obtained_marks/total_marks format, 737 (total) format)

Method:

```
marks = []
for i in range(len(data['HSSC-2'])):
    marks = str(data['HSSC-2'][i])
    if '/' in marks:
        obtained_marks, total_marks = marks.split('/')
        marks = obtained_marks
    if '(' in marks:
        obtained_marks, null = marks.split('(')
        marks = obtained_marks
    data['HSSC-2'][i] = marks
```

3. Calculating Correct Marks of HSSC-2 out of 590

Method:

```
marks = []
for i in range(len(data['HSSC-2'])):
    marks = int(data['HSSC-2'][i])
    if(marks > 590):
        marks = marks - 590
    data['HSSC-2'][i] = str(marks)
```

4. Removing Discrepancies from CGPA (like: 2.84. to 2.84)

Method:

```
marks = []
for i in range(len(data['CGPA'])):
    marks = str(data['CGPA'][i])
    cgpa = marks.split('.')
    if(len(cgpa) > 1):
        marks = cgpa[0] + '.' + cgpa[1]
    else:
        marks = cgpa[0]
    data['CGPA'][i] = marks
```

5. Removing Kg Units from Weights (like: 60kg to 60)

Method:

```
weight = []
for i in range(len(data['Weight'])):
    weight = data['Weight'][i]
    if 'kg' in weight:
        w, null = weight.split('k')
        weight = w
    data['Weight'][i] = weight
```

6. Normalizing the Month Names (like: spelling mistakes, shortforms, unnecessary spaces, number for month, consistent upper and lower case)

Method:

```
months_upper = ['January', 'February', 'March', 'April', 'May',
'June', 'July', 'August', 'September', 'October', 'November',
'December']
months_lower = ['january', 'february', 'march', 'april', 'may',
'june', 'july', 'august', 'september', 'october', 'november',
'december']

for i in range(len(data['BirthMonth'])):
    month = data['BirthMonth'][i].strip()

    if '/' in month:
        date = month.split('/')
        month_no = int(date[1])
        month = months_upper[month_no - 1]
    if month.isdigit():
        month_no = int(month)
        month = months_upper[month_no - 1]
    if month == 'Feburary':
        month = 'February'

    for j in range(len(months_upper)):
        if (month in months_upper[j]):
            month = months_upper[j]
        if (month in months_lower[j]):
            month = months_upper[j]
    data['BirthMonth'][i] = str.title(month)
```

7. Normalizing the Color Names (like: unnecessary spaces, consistent upper and lower case)

Method:

```
color = []
for i in range(len(data['FavoriteColor'])):
    color = data['FavoriteColor'][i].strip()
    color = str.title(color)
    data['FavoriteColor'][i] = color
```