Data Wrangling on Course Evaluation Parami Data

Presentation

By Htut Htet Naing

Agenda

- Exploration of Data
- Interpretation of Data
- Codes and Functions Used
- Conclusion

Exploration of Data

There are **11 rows × 22 columns.**

Rows Breakdown

Each row is a person describe by Gender (Male or Female).

Columns Breakdown

- 1 column each for Separate Data (Age, Attendance, and Gender) and TimeStamp
- 5 columns about Students Self Evaluation
- 3 columns about Skill and responsiveness of the instructor(s)
- 5 columns about Course Structure
- 3 additional columns about Student’s Personal Life to the course

Exploration of Data

4 integer (64) columns are :

- Age
- Attendance %
- On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?
- Would you recommend this course to your friend.

1 extreme columns :

- On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?

The rest are object data type columns.

Interpretation of Data

Average values of Int64 columns

- Age = 20.636364
- Attendance % = 97.000000
- hours per week spent = 34.636364 / 7 (Median)
- recommending course to a friend = 4.545455

Minimum and Maximum values of Int64 columns

- Age = 16.0000 / 36.000000
- Attendance % = 88.000000 / 100.000
- hours per week spent = 3 / 300
- recommending course to a friend = 3.000000/ 5.000000

Interpretation of Data

Finding Mean for Female and Male

Gender	Age	Attendance %	Would you recommend this course to your friend.	On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?
Female	19.000000	98.500000	4.250000	10.000000
Male	21.571429	96.142857	4.714286	48.714286

Codes and Functions Used

Cleaning the Columns

To drop unnecessary columns

- df = df.drop(columns = "Timestamp")

To Formattize the columns

```
[7]: df.columns = df.columns.str.strip()
     df.columns
```

```
[7]: Index(['Timestamp', 'Age', 'Attendance %', 'Gender',
       'Students Self Evaluation [I have been challenged to learn more than I expected.]',
       'Students Self Evaluation [I have put a great deal of effort in this course.]',
       'Students Self Evaluation [I consistently prepared for class.]',
       'Students Self Evaluation [I always perform well in the class. ]',
       'Students Self Evaluation [Compared to my peers, I am above average student. ]',
       'Skill and responsiveness of the instructor(s) [Teachings were clear and organized]',
       'Skill and responsiveness of the instructor(s) [Instructor(s) stimulated student interest]',
       'Skill and responsiveness of the instructor(s) [Instructor(s) effectively used time during class periods]',
       'Skill and responsiveness of the instructor(s) [Instructor(s) was available and helpful]',
       'Skill and responsiveness of the instructor(s) [The instructor(s) cared about the students, their progress, and successful course completion.]',
       'Course Structure [This class has increased my interest in data science field.]',
       'Course Structure [This course gave me confidence to do more advanced work in the subject.]',
       'Course Structure [I believe that what I am being asked to learn in this course is important.]',
       'Course Structure [This course was challenging.]',
       'Course Structure [This course helped me develop intellectual and critical thinking skills.]',
       'On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any
other course-related work?',
       'Would you recommend this course to your friend.',
       'Why did you choose this course?'],
      dtype='object')
```

Codes and Functions Used

To import data set variable

- df = pd.read_csv('Course Evaluation_Parami.csv')

To have an overview look of data

- df
- df.info()
- df.describe()

To check number of data types and data values

- dtypes_ct = df.dtypes.value_counts()
- df['Gender'].value_counts()

To filter columns of certain data types

- df.select_dtypes(include='int64')

Codes and Functions Used

To chech missing value or null or na

- df.isna().sum() - df.fillna

To get average value

- df['Age'].mean()
- df['On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?'].median()
- df.describe()

To groupby (with formatting column names):

- df.columns = df.columns.str.strip()
- df.groupby('Gender')[['Age', 'Attendance %', 'Would you recommend this course to your friend.', 'On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?']].mean()

Codes and Functions Used

To Count values of one column:

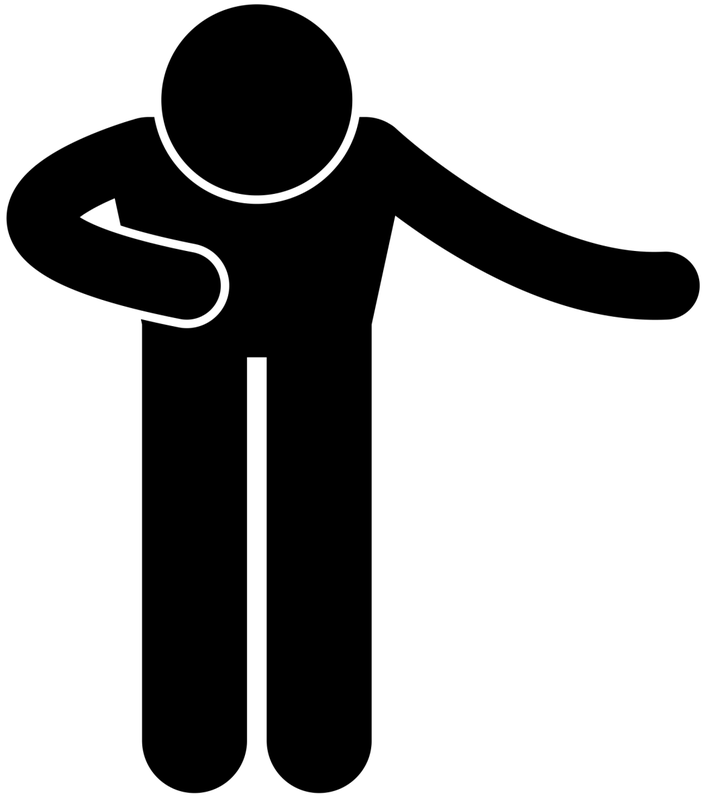
- df['Students Self Evaluation [I have been challenged to learn more than I expected.]'].value_counts()

```
ct_moretex = df['Students Self Evaluation [I have been challenged to learn more than I expected.]'].value_counts()

print(ct_moretex)

print(" ")

aaa = df.groupby(['Gender', 'Students Self Evaluation [I have been challenged to learn more than I expected.]']).size()
print(aaa)
```

```
Students Self Evaluation [I have been challenged to learn more than I expected.]
Strongly Agree       7
Agree                2
Strongly disagree    1
Neutral              1
Name: count, dtype: int64

Gender  Students Self Evaluation [I have been challenged to learn more than I expected.]
Female  Strongly Agree                                                                   4
Male    Agree                                                                           2
        Neutral                                                                         1
        Strongly Agree                                                                  3
        Strongly disagree                                                               1
dtype: int64
```

Conclusion

- Data are mostly categorial data
- Data overall represents Qualitative data
- No null data, No duplications
- Cleaning of column names is also necessary for longer column names
- Unnecessary data here is Time Stamp
- Grouping data into category is important in analyzing
- One outlier is 3 - 300 (On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, coding and any other course-related work?)



Thank You