2024-06-23

#sic_lab #data_exploration

Merging dataframes

> Use students_list.csv files

```
df = pd.read_csv('data_studentlist.csv')
df1 = df.loc[7: ,['NAME','BLOODTYPE','WEIGHT','HEIGHT']]]
df2 = df.loc[:10,]
```

- Merge two dataframes [inner, left, right, outer join]

Concat data from different sources

- Merge data from `data.csv` and `data.json` files
    - How can we concat concat function
- hat are the similarities and difference in the two sources

Plot the following plots using dummy data in both matplotlib and seaborn

> Use any relevant dataset

- Bar plot
- Histogram (play with: density, bins, alpha)
- Boxplot [Multiple varables, vertical,]
- Lineplot [linestyle, markerstyles]
- Scatter plot

Plot sine wave over the linear space of 0-10 having 100 steps in betwee and check different

- markers [o, v, s, ^]
- color [red, green ,..]
- linestyle [-, -., ...]
- figure size of plt.figure

Using Axes:

> Use [plt.figure and fig.add_axes]

- Multiple plot in same axes [Sine and cos]
- Multiple plots in different axes
- Multiple plots using plt.subplots (draw 4 different plots in a 2x2 grid)

Plot using Pandas using iris flower dataset

> Use: [iris dataset, based on type]

- histogram
- barplot
- scatterplot
- scatter_matrix

Plot using Seaborn using mpg dataset:

> Use [x=weight, y= mpg , hue=origin if required]
> `dat = sns.load_dataset('mpg')`

- Histogram [kde, rug, bins, color,]
- KDEplot
- jointplot [kind]
- lmplot[hue, col, markers]
- barplot[(origin, mpg) estimator]
- countplot [hue]
- boxplot [multiple with x/y=origin/mpg, notch, palette, hue=cylinders]
- violinplot [(origin, mpg), (cylinders, horsepower)]
- stripplot, swarmplot, voilinplot + swarmplot overlap
- pairplot [hue=species]
- PairGrid
- FacetGrid
- heatmap of correlation

Feature selection and Engineering

Using the mpg dataset
`sns.load_dataset("mpg")`

- Based on correlation which feature are important / which can be skipped?
    - How do we get score of correlation between all the variables
- Using p-score
    - Use SelectKBest, f_regression from sklearn
- Feature normalization
    - StandardScalar
    - MinMaxScalar
- Adding new variables combining existing features
    - eg BMI
    - polynomial feature using sklearn ()
- Encodings

```
data = {
    "Rec-no": range(10),
    "Temperature": ["Hot", "Cold", "Very Hot",
                    "Warm", "Hot", "Warm",
                    "warm", "hot", "hot", "cold"],
    "Color": ["Red", "Yellow", "Blue",
              "Blue", "Red", "Yellow",
              "Red", "Yellow", "Yellow", "Yellow"],
    "Target": [1,1,1,0,1,0,1,0,1,1]
}
```

Is there any issue with the above data?

Explore the Different encodings we have learn about:

- One hot encoding
    - Using OHE from sklearn
    - pd.get_dummies from pandas
- Label Encoding
- Ordinal value encoding
- Binary Encoding
- Frequency encoding
    - Mean

what is the difference between

- fittransform and fit + transform
- transform vs predict