Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

Amazon Web Services

MLOps with AWS

Masterclass

aws

Machine Learning Operations with AWS

Day -7

Amazon Web Services

MLOps with AWS

Masterclass

Pantech e Learning

DIGITAL LEARNING SIMPLIFIED

Pandas

Analysing Data

```
Creating a new column:

  dataset["new column"] = 1


  dataset["new column"] = dataset["column1"]/4


deleting column:

  dataset.drop(["column1", "column3"], axis= 1, inplace= True)
deleting rows:

  dataset.drop([0,2], axis= 0, inplace= True)
```

Analysing Data

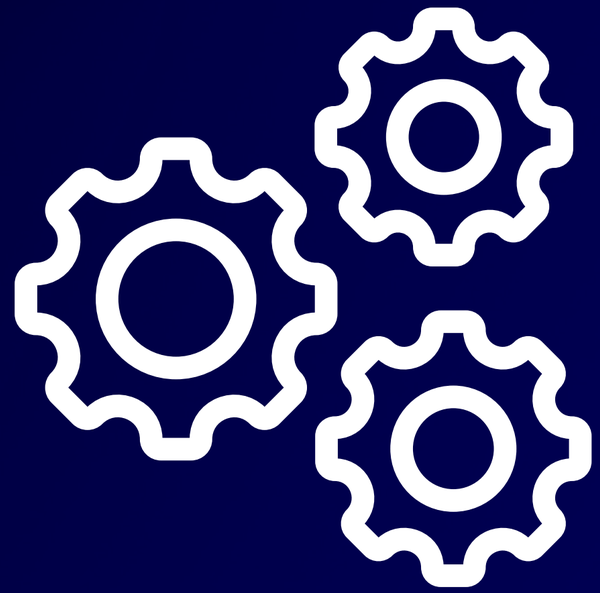
```
Rename columns:

  dataset.rename(columns={"Column1":"Column A","Column2":"Column B"})


combine two datasets:

  pd.concat([dataset, new dataset], axis=0, ignore_index= True)


Create new index:

  dataset.set_index("Name", inplace= True)
```

Analysing Data

```
dataset["Age"].agg(["mean", "median"])      ----->  used to apply a list of function

dataset.groupby("Pclass").Fare.median()     ----->  used to group the data

dataset[dataset["Pclass"].isin([3])]        ----->  used to filter based on specific element

pd.get_dummies(dataset["Embarked"])         ----->  used for encoding values

dataset.to_csv("Output.csv")                ----->  export and download dataframe into csv
```

Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

Data Preprocessing with Pandas



Data Preprocessing

- Data cleaning means fixing unwanted or improper data in your dataset
- This will improve the accuracy of data by removing or correcting inaccuracies, missing values, duplicates, and irrelevant data
- It can make the data more consistent, reducing the risk of errors in downstream processes

Check for null values

```
dataset.isnull().sum()
```

Handling Null Values

- Imputation
- Dropping

Imputation

```
dataset.fillna(X, inplace = True)
```

Imputation

```
x = dataset["Calories"].mean()

y = dataset["Calories"].median()

z = dataset["Calories"].mode()

dataset["Calories"].fillna(x, inplace = True)
```

Dropping

```
dataset.dropna(inplace = True)
```

Handling Duplicates

```
 Check for duplicate values:

print(dataset.duplicated())


Remove Duplicates:

dataset.drop_duplicates(inplace = True)
```

Correlation with Pandas

- It calculates the relationship between each column in your data set
- The Result is a table with a lot of numbers that represents how well the relationship is between two columns.

Correlation with Pandas

```
dataset.corr()
```

	index	wheel-base	length	horsepower	average-mileage	price
index	1.000000	0.013401	0.004828	-0.093809	0.176037	-0.197470
wheel-base	0.013401	1.000000	0.878381	0.463421	-0.547325	0.663085
length	0.004828	0.878381	1.000000	0.668555	-0.788429	0.788465
horsepower	-0.093809	0.463421	0.668555	1.000000	-0.808804	0.901707
average-mileage	0.176037	-0.547325	-0.788429	-0.808804	1.000000	-0.770217
price	-0.197470	0.663085	0.788465	0.901707	-0.770217	1.000000

AWS Glue

AWS Glue ETL

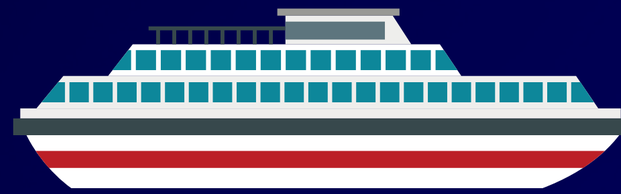
- ETL refers to three (3) processes that are commonly needed in Machine Learning processes: Extraction, Transformation, Loading.

- Extracting data from a source, transforming it in the right way for applications, and then loading it back to the data warehouse.

Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

Pandas Assignment

1. Collect Titanic dataset
2. What is the overall survival rate of passengers on the Titanic?
3. What was the gender distribution among the passengers on the Titanic?
4. Did the survival rate differ by gender? If so, how much?
5. What was the age distribution among the passengers on the Titanic?
6. Did the survival rate differ by the journey class? If so, how much?



Thank you