**GOVERNMENT POLYTECHNIC NAGAMANGALA**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Vth Semester Diploma

**Artificial Intelligence and Machine Learning (20CS51)**

**Assignment: 01**

**NAME: SINCHANA A.P**

**ROLL NO: 158CS22049**

AIML (20CS51)

ASSIGNMENT – WEEK 02

1. Download any two datasets from the internet and perform the following operations.

(a) Aggregate functions.

* HEAD

import pandas as pd

df=pd.read\_csv("/content/StudentsPerformance.csv")

df.head(5)

import pandas as pd

df=pd.read\_csv("/content/retail trade 1.csv")

df.head(5)

**OUTPUT**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score | |
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Series\_reference | Period | Data\_value | Suppressed | STATUS | UNITS | Magnitude | Subject | Group |
| 0 | RTTA.SF11CA | 1997.03 | 27070.4 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |
| 1 | RTTA.SF11CA | 1998.03 | 27748.4 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |
| 2 | RTTA.SF11CA | 1999.03 | 28216.6 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |
| 3 | RTTA.SF11CA | 2000.03 | 30204.1 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |
| 4 | RTTA.SF11CA | 2001.03 | 31641.6 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |
|  | Series\_reference | Period | Data\_value | Suppressed | STATUS | UNITS | Magnitude | Subject | Group |
| 0 | RTTA.SF11CA | 1997.03 | 27070.4 | NaN | F | Dollars | 6 | Retail Trade (ANZSIC06) - RTT | Sales and stocks by industry, in current and c... |

* **TAIL**

df.tail(5)

df.tail(5)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **gender** | **race/ethnicity** | **parental level of education** | **lunch** | **test preparation course** | **math score** | **reading score** | **writing score** |
| **995** | female | group E | master's degree | standard | completed | 88 | 99 |
| **996** | male | group C | high school | free/reduced | none | 62 | 55 |
| **997** | female | group C | high school | free/reduced | completed | 59 | 71 |
| **998** | female | group D | some college | standard | completed | 68 | 78 |
| **999** | female | group D | some college | free/reduced | none | 77 | 86 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Series\_ reference** | **Period** | **Data\_value** | **Suppressed** | **STATUS** | **UNITS** | **Magnitude** | **Subject** | **Group** | **Series\_title\_1** | **Series\_title\_2** | **Series\_title\_3** | **Series\_title\_4** |  |
| 24 | RTTA.SF1RNCA | 2020.03 | 9740.6 | NaN | F | Dollars | 6 | Retail Trade  (ANZSIC06) -  RTT | Sales by region in current prices (SAFC) | Core industries  total | Canterbury | Current | Actual |
| 25 | RTTA.SF1RNCA | 2021.03 | 9893.6 | NaN | F | Dollars | 6 | Retail Trade  (ANZSIC06) -  RTT | Sales by region in current prices (SAFC) | Core industries  total | Canterbury | Current | Actual |
| 26 | RTTA.SF1RNCA | 2022.03 | 11030.8 | NaN | F | Dollars | 6 | Retail Trade  (ANZSIC06) -  RTT | Sales by region in current prices (SAFC) | Core industries  total | Canterbury | Current | Actual |
| 27 | RTTA.SF1RNCA | 2023.03 | 11855.4 | NaN | F | Dollars | 6 | Retail Trade  (ANZSIC06) -  RTT | Sales by region in current prices (SAFC) | Core industries  total | Canterbury | Current | Actual |
| 28 | RTTA.SF1RNCA | 2024.03 | 12273.6 | NaN | F | Dollars | 6 | Retail Trade  (ANZSIC06) -  RTT | Sales by region in current prices (SAFC) | Core industries  total | Canterbury | Current | Actual |

* **SUM**

df.sum()

df.sum()

**OUTPUT**

|  |  |
| --- | --- |
| gender | femalefemalefemalemalemalefemalefemalemalemale... |
| ace/ethnicity | group Bgroup Cgroup Bgroup Agroup Cgroup Bgrou... |
| parental level of education | bachelor's degreesome collegemaster's degreeas... |
| Lunch | standardstandardstandardfree/reducedstandardst... |
| test preparation course | nonecompletednonenonenonenonecompletednonecomp... |
| math score | 66089 |
| reading score | 69169 |
| writing score | 68054 |

|  |  |
| --- | --- |
| Series\_reference | RTTA.SF11CARTTA.SF11CARTTA.SF11CARTTA.SF11CART... |
| Period | 58296.87 |
| Data\_value | 562476.9 |
| Suppressed | 0 |
| STATUS | FFFFFFFFFFFFFFFFFFFFFFFFFFFFF |
| UNITS | DollarsDollarsDollarsDollarsDollarsDollarsDoll... |
| Magnitude | 174 |
| Subject | Retail Trade (ANZSIC06) - RTTRetail Trade (ANZ... |
| Group | Sales and stocks by industry, in current and c... |
| Series\_title\_1 | Core industries totalCore industries totalCore... |
| Series\_title\_2 | Sales (operating income)Sales (operating incom... |
| Series\_title\_3 | CurrentCurrentCurrentCurrentCurrentCurrentCurr... |
| Series\_title\_4 | ActualActualActualActualActualActualActualActu... |

* **MIN**

df.min()

df.min()

**OUTPUT**

|  |  |
| --- | --- |
| gender | female |
| ace/ethnicity | group A |
| parental level of education | associate's degree |
| Lunch | Free/reduced |
| test preparation course | Completed |
| math score | 0 |
| reading score | 17 |
| writing score | 10 |

|  |  |
| --- | --- |
| Series\_reference | RTTA.SF11CA |
| Period | 1997.03 |
| Data\_value | 844.7 |
| Suppressed | NaN |
| STATUS | F |
| UNITS | Dollars |
| Magnitude | Retail Trade (ANZSTC06) - RTT |
| Subject | sales and stocks by industry, in current and c… |
| Group | core industries total |
| Series\_title\_1 | canterbury |
| Series\_title\_2 | current |
| Series\_title\_3 | actual |

* MAX

df.max()

df.max()

**OUTPUT**

|  |  |
| --- | --- |
| gender | Male |
| ace/ethnicity | Group E |
| parental level of education | Some high school |
| Lunch | Standard |
| test preparation course | None |
| math score | 100 |
| reading score | 100 |
| writing score | 100 |

|  |  |
| --- | --- |
| Series\_reference | RTTA.SF11CA |
| Period | 2024.03 |
| Data\_value | 43402.0 |
| Suppressed | NaN |
| STATUS | F |
| UNITS | Dollars |
| Magnitude | Retail Trade (ANZSTC06) - RTT |
| Subject | sales by region in current price (SAFC) |
| Group | core industries total |
| Series\_title\_1 | Taranaki |
| Series\_title\_2 | Defaulted, at September 2010 quarter prices |
| Series\_title\_3 | actual |

* COUNT

df.count()

df.count()

**OUTPUT**

|  |  |
| --- | --- |
| gender | Male |
| Race/ethnicity | 1000 |
| parental level of education | 1000 |
| Lunch | 1000 |
| test preparation course | 1000 |
| math score | 1000 |
| reading score | 1000 |
| writing score | 1000 |

|  |  |
| --- | --- |
| Series\_reference | 29 |
| Period | 29 |
| Data\_value | 29 |
| Suppressed | 29 |
| STATUS | 29 |
| UNITS | 29 |
| Magnitude | 29 |
| Subject | 29 |
| Group | 29 |
| Series\_title\_1 | 29 |
| Series\_title\_2 | 29 |
| Series\_title\_3 | 29 |

* Groupby

import pandas as pd

path='/content/StudentsPerformance.csv'

data=pd.read\_csv(path)

grouped = data.groupby('reading score')

print(grouped.agg({'math score':'sum','writing score':'sum'}))

import pandas as pd

data=pd.read\_csv("/content/retail trade 1.csv")

grouped\_data=data.groupby('Data\_value')['Group'].sum()

print(grouped\_data)

**OUTPUT**

|  |  |  |
| --- | --- | --- |
| reading score | math score | writing score |
| 17 | 0 | 10 |
| 23 | 28 | 19 |
| 24 | 38 | 38 |
| 26 | 30 | 22 |
| 28 | 35 | 27 |
| … | … | … |
| 95 | 695 | 740 |
| 96 | 378 | 368 |
| 97 | 437 | 481 |
| 99 | 270 | 295 |
| 100 | 1607 | 1678 |

|  |  |
| --- | --- |
| Data\_value | group |
| 844.7 | Sales by region in current prices (SAFC) |
| 903..1 | Sales by region in current prices (SAFC) |
| 1437.5 | Sales by region in current prices (SAFC) |
| 1503.4 | Sales by region in current prices (SAFC) |
| 1690.9 | Sales by region in current prices (SAFC) |
| 1724.2 | Sales by region in current prices (SAFC) |
| 1784.9 | Sales by region in current prices (SAFC) |
| 2940.9 | Sales by region in current prices (SAFC) |
| 3094.9 | Sales by region in current prices (SAFC) |
| 3175.1 | Sales by region in current prices (SAFC) |
| 9740.6 | Sales by region in current prices (SAFC) |
| 9893.6 | Sales by region in current prices (SAFC) |
| 11030.8 | Sales by region in current prices (SAFC) |
| 11855.4 | Sales by region in current prices (SAFC) |
| 12273.6 | Sales by region in current prices (SAFC) |
| 27070.4 | sales and stock by industry, in current and c……. |
| 27748.4 | sales and stock by industry, in current and c……. |
| 28216.6 | sales and stock by industry, in current and c……. |
| 30204.1 | sales and stock by industry, in current and c……. |
| 31641.6 | sales and stock by industry, in current and c……. |
| 33853 | sales and stock by industry, in current and c……. |
| 35371.2 | sales and stock by industry, in current and c……. |
| 36284.6 | sales and stock by industry, in current and c……. |
| 37506.9 | sales and stock by industry, in current and c……. |
| 38184.9 | sales and stock by industry, in current and c……. |
| 38269.6 | sales and stock by industry, in current and c……. |
| 39423.6 | sales and stock by industry, in current and c……. |
| 41403.4 | sales and stock by industry, in current and c……. |
| 43402 | sales and stock by industry, in current and c……. |

(b) Use Map, Filter, Reduce, and Lambda Functions with Pandas data frames

* MAP

import pandas as pd

data=pd.read\_csv("/content/StudentsPerformance.csv")

grade\_map={9:'male',10:'female'}

data['gender\_level']=data['gender'].map(grade\_map)

print(data)

import pandas as pd

# read CSV file into Dataframe

d = pd.read\_csv('/content/retail-trade-survey-march-2024-quarter.zip')

d['Series\_title\_1'] = d['Series\_title\_1'].map(lambda x: x \* 1)

print(d)

**OUTPUT**

|  |  |
| --- | --- |
| gender | gender\_level |
| female | NaN |
| female | NaN |
| female | NaN |
| male | NaN |
| male | NaN |
| … | .. |
| female | NaN |
| male | NaN |
| female | NaN |
| female | NaN |
| female | NaN |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STATUS data value Suppressed STATUS UNITS \** |  |  |  |  |  |
| 0 RTTA.SF11CA 1997.03 27070.4 NaN F Dollars |  |  |  |  |  |
| 1 RTTA.SF11CA 1998.03 27748.4 NaN F Dollars |  |  |  |  |  |
| 2 RTTA.SF11CA 1999.03 28216.6 NaN F Dollars |  |  |  |  |  |
| 3 RTTA.SF11CA 2000.03 30204.1 NaN F Dollars |  |  |  |  |  |
| 4 RTTA.SF11CA 2001.03 31641.6 NaN F Dollars |  |  |  |  |  |
| ... ... ... ... ... ... ... |  |  |  |  |  |
| 26188 RTEQ.SNV 2023.03 1447.0 NaN F Index |  |  |  |  |  |
| 26189 RTEQ.SNV 2023.06 1482.0 NaN F Index |  |  |  |  |  |
| 26190 RTEQ.SNV 2023.09 1507.0 NaN F Index |  |  |  |  |  |
| 26191 RTEQ.SNV 2023.12 1520.0 NaN F Index |  |  |  |  |  |
| 26192 RTEQ.SNV 2024.03 1537.0 NaN F Index |  |  |  |  |  |

|  |
| --- |
| Series\_title\_1 Series\_title\_2 Series\_title\_3 \ |
| 0 Core industries total Sales (operating income) Current |
| 1 Core industries total Sales (operating income) Current |
| 2 Core industries total Sales (operating income) Current |
| 3 Core industries total Sales (operating income) Current |
| 4 Core industries total Sales (operating income) Current |
| ... ... ... ... |
| 26188 Food and beverage services NaN NaN |
| 26189 Food and beverage services NaN NaN |
| 26190 Food and beverage services NaN NaN |
| 26191 Food and beverage services NaN NaN |
| 26192 Food and beverage services NaN NaN |

* FILTER

filtered\_d=df[df['math score'].map(lambda x:x >=100)]

print(filtered\_d)

filtered\_d=df[df['Data\_value'].map(lambda x: x >= 100)]

print(filtered\_d)

**OUTPUT**

|  |  |  |
| --- | --- | --- |
| math score | reading score | writing score |
| 100 | 100 | 93 |
| 100 | 92 | 97 |
| 100 | 100 | 100 |
| 100 | 97 | 86 |
| 100 | 96 | 99 |
| 100 | 100 | 100 |
| 100 | 100 | 100 |

|  |
| --- |
| **Series\_reference Period Data\_value Suppressed STATUS UNITS \** |
| 0 RTTA.SF11CA 1997.03 27070.4 NaN F Dollars |
| 1 RTTA.SF11CA 1998.03 27748.4 NaN F Dollars |
| 2 RTTA.SF11CA 1999.03 28216.6 NaN F Dollars |
| 3 RTTA.SF11CA 2000.03 30204.1 NaN F Dollars |
| 4 RTTA.SF11CA 2001.03 31641.6 NaN F Dollars |
| ... ... ... ... ... ... ... |
| 26188 RTEQ.SNV 2023.03 1447.0 NaN F Index |
| 26189 RTEQ.SNV 2023.06 1482.0 NaN F Index |
| 26190 RTEQ.SNV 2023.09 1507.0 NaN F Index |
| 26191 RTEQ.SNV 2023.12 1520.0 NaN F Index |
| 26192 RTEQ.SNV 2024.03 1537.0 NaN F Index |

* **REDUCE**

from functools import reduce

if not filtered\_d['math score'].empty: # Check if 'age' column is empty

total = reduce(lambda x, y: x + y, filtered\_d['math score'])

print(total)

else:

print("The filtered DataFrame is empty, cannot calculate total age.")

from functools import reduce

if not filtered\_d['lunch'].empty:  # Check if 'age' column is empty

    total = reduce(lambda x, y: x + y, filtered\_d['lunch'])

    print(total)

else:

    print("The filtered DataFrame is empty, cannot calculate total age.")

**OUTPUT**

700

67660523.89999995

1. Visualize the data set (At least 6 different plots).

Dataset 1

import pandas as pd

# load the CSV file into DataFrame

d=pd.read\_csv('/content/StudentsPerformance.csv')

import matplotlib.pyplot as plt

import seaborn as sns

# Set plot size

plt.figure(figsize=(20, 15)

#1. Bar plot

plt.subplot(2, 3, 2)

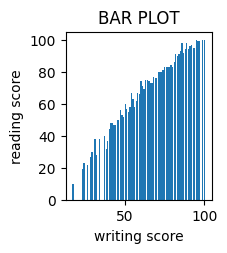
plt.bar(d['reading score'], d['writing score'])

plt.title("BAR PLOT")

plt.xlabel('writing score')

plt.ylabel('reading score')

0UTPUT



#2 HISTOGRAM

plt.hist(d['reading score'],bins=10,alpha=0.7,color='purple')

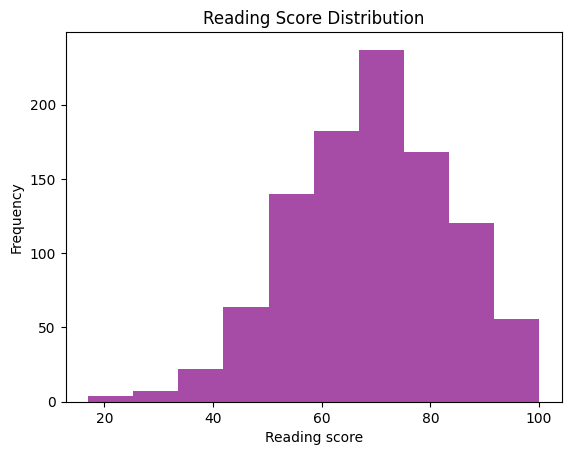
plt.title('Reading Score Distribution')

plt.xlabel('Reading score')

plt.ylabel('Frequency')

plt.show()

OUTPUT



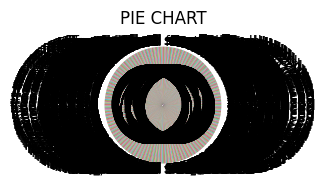
#3 PIE CHART

plt.subplot(2,3,6)

plt.pie(d['reading score'], labels=d['lunch'], autopct='%1.1f%%')

plt.title('PIE CHART')

OUTPUT



#4 Box plot

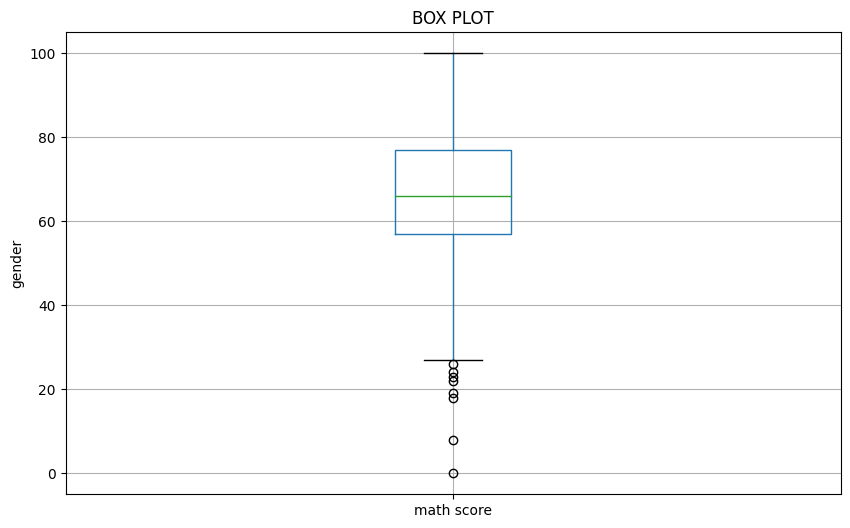
plt.figure(figsize=(10,6))

d.boxplot(column=['math score'])

plt.title("BOX PLOT")

plt.ylabel('gender')

OUTPUT



#5 scatter plot

plt.subplot(2,3,4)

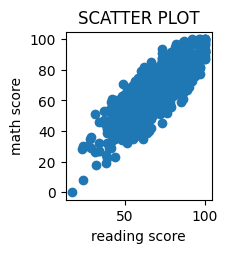
plt.scatter(d['reading score'],d['math score'])

plt.title("SCATTER PLOT")

plt.xlabel('reading score')

plt.ylabel('math score')

OUTPUT



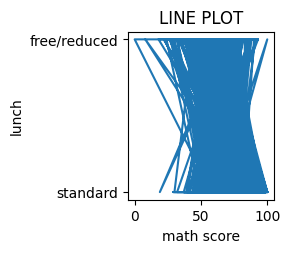
#6 line plot

from matplotlib import pyplot as plt

df['math score'].plot(kind='line', figsize=(8, 4), title='math score')

plt.gca().spines[['top', 'right']].set\_visible(False)

OUTPUT



Dataset 2

import pandas as pd

d=pd.read\_csv("/content/retail trade 1.csv")

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(20,15))

#1 Line plot

plt.subplot(2,3,1)

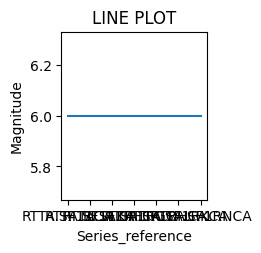
plt.plot(df['Series\_reference'],df['Magnitude'])

plt.title("LINE PLOT")

plt.xlabel("Series\_reference")

plt.ylabel("Magnitude")

OUTPUT



#2. Bar plot

plt.subplot(2, 3, 2)

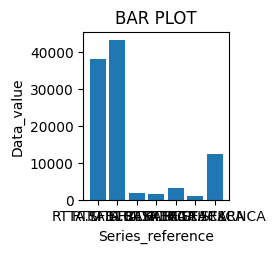
plt.bar(df['Series\_reference'], df['Data\_value'])

plt.title("BAR PLOT")

plt.xlabel('Series\_reference')

plt.ylabel('Data\_value')

OUTPUT



#3 Histogram

plt.hist(d['Period'],bins=10,alpha=0.7,color='green')

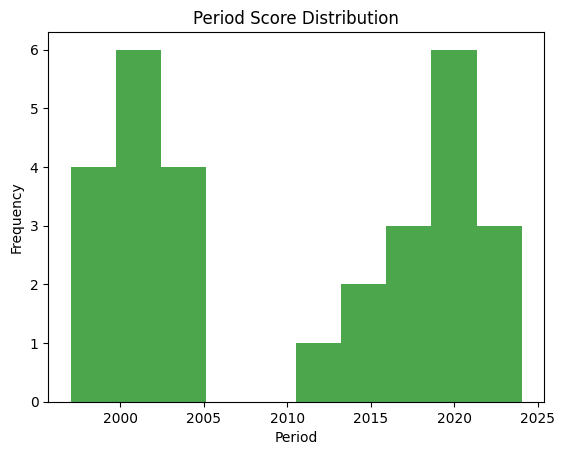
plt.title('Period Score Distribution')

plt.xlabel('Period')

plt.ylabel('Frequency')

plt.show()

OUTPUT



#4. Scatter plot

plt.subplot(2, 3, 4)

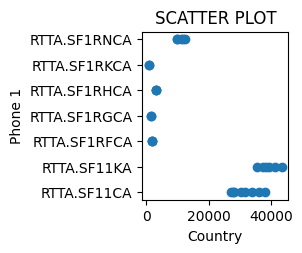
plt.scatter(df['Data\_value'], df['Series\_reference'])

plt.title("SCATTER PLOT")

plt.xlabel('Data\_value')

plt.ylabel('Series\_reference')

OUTPUT



#5. Box plot

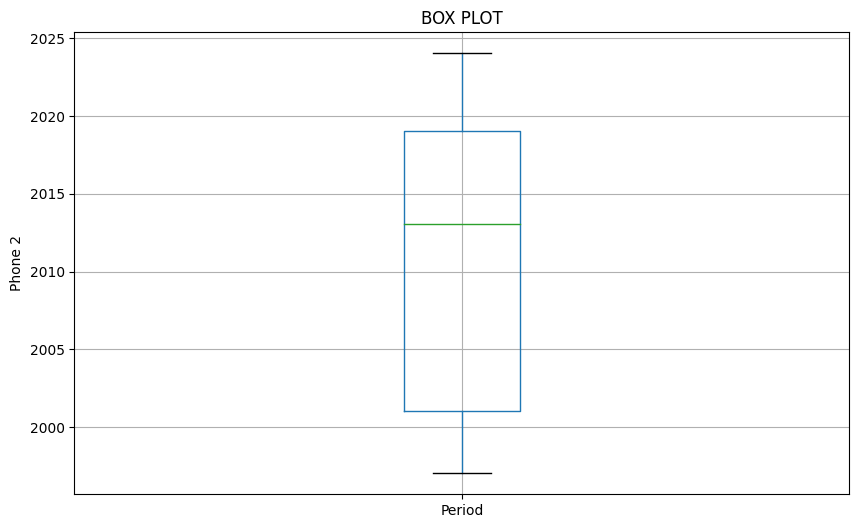
plt.figure(figsize=(10, 6))

d.boxplot(column=['Period'])

plt.title("BOX PLOT")

plt.ylabel('Phone 2')

OUTPUT



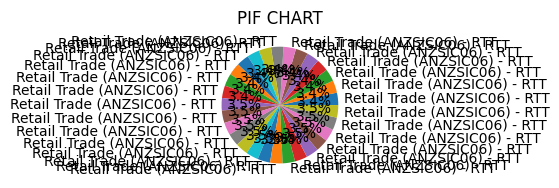
#6 pie chart

plt.subplot(2, 3, 6)

plt.pie(d['Period'], labels=d['Subject'], autopct='%1.1f%%')

plt.title('PIF CHART')

OUTPUT



**(D) How do you create a project plan and product backlog for an AI project?**

**b. Classification Project – ML / deep learning**

**Project Plan**

**Project Definition**

* Objective: Define the goal of the classification project (e.g., classify images into categories, predict customer churn).
* Scope: Determine what is included and excluded in the project.

**Stakeholders and Roles**

* Identify key stakeholders (e.g., project sponsor, data scientists, ML engineers, product managers).
* Define roles and responsibilities.

**Requirements Gathering**

* Collect and document functional and non-functional requirements.
* Determine data requirements (type, source, volume, quality).

**Project Milestones and Deliverables**

**Phase 1**: Data Collection and Preparation

**Data acquisition**

* Data cleaning and preprocessing

**Phase 2**: Model Development

* Exploratory Data Analysis (EDA)
* Feature engineering
* Model selection and training

**Phase 3:** Model Evaluation and Tuning

* Model validation and testing
* Hyperparameter tuning

**Phase 4**: Deployment

* Model integration
* Deployment to production

**Phase 5:** Monitoring and Maintenance

* Performance monitoring
* Model retraining as necessary

**Timeline**

* Develop a Gantt chart or timeline outlining tasks and milestones.

**Resource Allocation**

* Determine the necessary resources (e.g., hardware, software, personnel).
* Assign resources to tasks.

**Risk Management**

* Identify potential risks (e.g., data quality issues, model performance).
* Develop mitigation strategies.

**Budget**

* Estimate the project cost.
* Allocate budget for different phases.

**Product Backlog**

**User Stories**

* Write user stories to describe the desired features from an end-user perspective.
* Example: "As a data scientist, I want to preprocess the data so that it is clean and ready for modeling."

**Backlog Items**

* Break down the project into smaller tasks and features.
* Prioritize tasks based on their importance and dependencies.

**Tasks**

* **Data Collection and Preparation**
* Collect dataset
* Clean and preprocess data
* Perform EDA
* **Model Development**
* Select model architecture
* Train initial model
* Evaluate model performance
* **Model Evaluation and Tuning**
* Split data into training and testing sets
* Validate model performance
* Perform hyperparameter tuning
* **Deployment**
* Develop APIs for model serving
* Deploy model to production environment
* **Monitoring and Maintenance**
* Set up monitoring tools
* Schedule periodic model retraining

**Acceptance Criteria**

* Define acceptance criteria for each task to ensure quality and completeness.
* Example: "The data preprocessing task is complete when the dataset has no missing values and all features are normalized."

**Sprint Planning**

* Organize tasks into sprints (e.g., 2-week sprints).

Assign tasks to team members.

**Review and Adapt**

* Regularly review progress in sprint reviews.
* Adapt the backlog based on feedback and changing requirements.

**Example Product Backlog for Classification Project**

**Sprint 1**

* **Data Collection**
* Collect dataset from source
* Verify data quality
* **Data Preprocessing**
* Handle missing values
* Normalize data
* **Exploratory Data Analysis**
* Visualize data distributions
* Identify correlations

**Sprint 2**

* **Model Selection**
* Research potential model architectures
* Select initial model
* **Model Training**
* Train model on preprocessed data
* Evaluate initial performance

**Sprint 3**

* **Model Evaluation**
* Split data into training and test sets
* Validate model performance
* **Hyperparameter Tuning**
* Perform grid search for hyperparameters
* Re-evaluate model performance

**Sprint 4**

* **Model Deployment**
* Develop API for model serving
* Deploy model to production
* **Monitoring Setup**
* Set up performance monitoring tools
* Schedule regular model retraining