**Question 1: Define the purpose of data analysis for the chosen dataset.**

**Answer:**The primary objective of analyzing the cinema sales dataset is to identify patterns and insights related to the sales of tickets, price, occupancy, and performance of various cinemas and films over time. It is directed towards predicting crucial measures such as total sales and ticket utilization using machine learning models. It will also help identify trends over quarters or months, the efficiency of cinemas, and the classification of show performance.

**Question 2: Identify and Justify the type of programming used for data analysis.**

**Answer:**

Python is used for this data analysis due to its strong ecosystem of data science libraries and tools. It supports:

* **Data handling:** with pandas and numpy
* **Visualization:** using matplotlib, seaborn, and plotly
* **Machine Learning:** via scikit-learn
* **Statistical Testing:** with scipy and stats models, Python's readability and community support make it ideal for rapid development and reproducibility in academic data analysis projects.

**Question 3: Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset.**

**Answer:**

This project applies three main types of machine learning algorithms:

* **Regression:** To predict continuous variables such as total\_sales or ticket\_use based on features like ticket price, cinema code, and occupancy.
* **Classification:** To classify cinema performances into categories like high/low ticket usage or sales based on thresholds (e.g., occupancy percentage > 70% is “High”).
* **Clustering:** To discover natural groupings among cinema shows based on multiple attributes like ticket price, show time, occupancy, and sales.

These algorithms help understand the data better and generate actionable insights or future predictions.

**Question 4: Identify and Justify the independent and dependent variables for the chosen dataset.**

**Answer:**

Dependent Variables (targets for prediction):

* Option 1: total\_sales – used in regression to predict revenue.
* Option 2: ticket\_use – useful for classifying show performance (high vs. low usage).
* Option 3: occu\_perc – also viable for classification or regression.

Independent Variables (predictors):

* film\_code – may represent the content type or popularity
* cinema\_code – different cinemas may perform differently
* tickets\_sold, tickets\_out, ticket\_price, show\_time – key numerical features influencing revenue or occupancy
* capacity – affects maximum occupancy
* month, quarter, day – may capture seasonal patterns

We selected these variables based on their logical and statistical relationship to the target values.

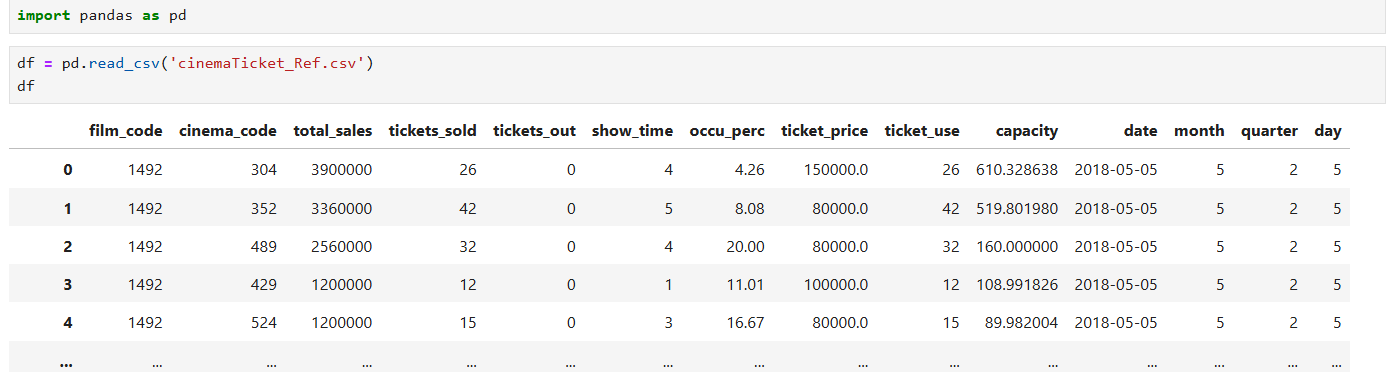
**Question 5: Justify why you want to perform the descriptive analysis for the chosen dataset.**

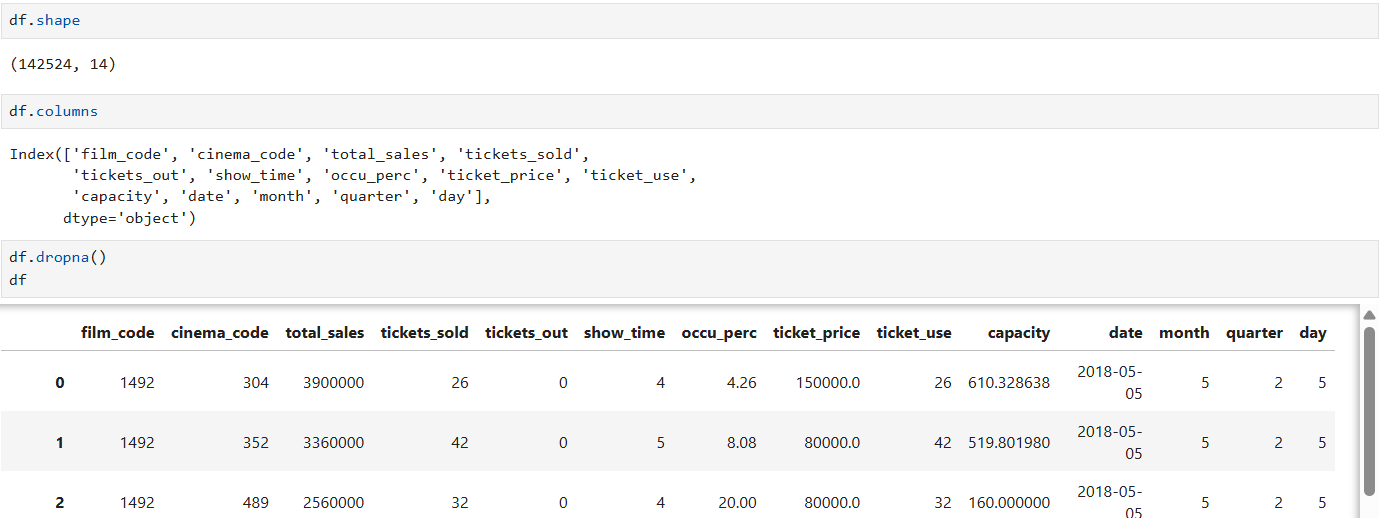
**Answer:**

Descriptive analysis assists in summarizing the key characteristics of the dataset in an easy-to-understand form. Through calculating statistical values like mean, median, standard deviation, and range, we can determine data distribution, identify outliers or anomalies, and gain insights into central tendencies. It also aids in more informed preprocessing decisions prior to implementing machine learning algorithms.

**Loading the selected data.**

To answer this question, we will start reading data and doing some processing of that data. Now I will share some of the screenshots of the data loading and processing.



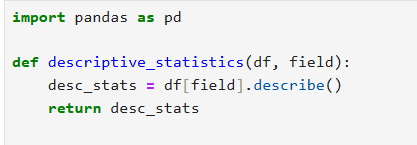


Now we are done with the initial analysis of the data after loading it, this data loading will help us for later analysis.

**Question 6: Create a script to develop a Python function for** **descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics.**

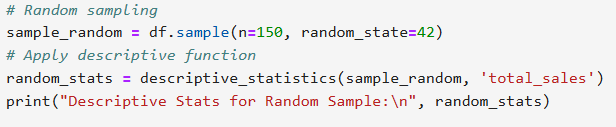
**Answer:**

Now we will write the script for the descriptive statistics.

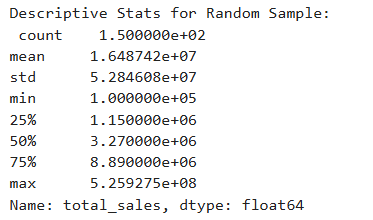


**Question 7: Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created.**

**Answer:**

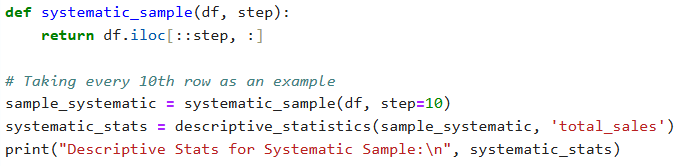
****

**This is the results of the script that we write.**

****

**Question 8: Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created].**

**Answer:**

****

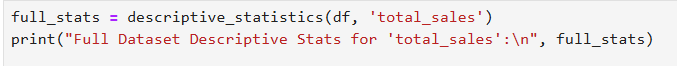
The output of the script is:

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Question 9: Create a detailed descriptive statistics report about the dependent variable of the chosen dataset.**

**Answer:**

****

The output will show the full dataset description.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Question 10: Visualize the dependent variable by the Graph/Chart of the following using Python Program:**

* Scatter plot
* Box Plot
* Histogram
* Heat Map

Hint: Use Matplot or Ski-learn library

**Answer:**

1. Script for the scatter plot

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A graph showing a scatter plot

AI-generated content may be incorrect.**

1. Script for the Box Plot

A screen shot of a computer code

AI-generated content may be incorrect.

A graph of a box plot

AI-generated content may be incorrect.

1. Script for the Histogram.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A graph with a bar graph

AI-generated content may be incorrect.

1. Script for of Heat Map.

A computer code with text

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**Question 11: Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable.**

**Note:** If you have more than one independent variable, then chose any one of the independent variables.

**Answer:**

We will:

* Use Pearson and Spearman correlation for numerical variables
* Use Chi-Square test for categorical variables

Let’s say:

* Dependent Variable: total\_sales
* Independent Variable: tickets\_sold (numeric)
* Optional Categorical: You could use something like cinema\_code or month if you want to run a chi-square test.

**Use Pearson and Spearman correlation for numerical variables**

A screenshot of a computer code

AI-generated content may be incorrect.

The output of this script show below:



**Use Chi-Square test for categorical variables**

A screenshot of a computer program

AI-generated content may be incorrect.

**Question 12: Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test.**

**Answer:**

Let’s check if the sample total\_sales has a mean significantly different from the assumed population mean. You can assume the population mean (pop\_mean) based on domain knowledge, or just use a fixed value like 3000000

**A screenshot of a computer code

AI-generated content may be incorrect.**

* The t-statistic is very high (115+), which indicates a large difference between your sample mean and the assumed population mean (3,000,000).
* The p-value is 0.0000, which is much less than 0.05, so we reject the null hypothesis.

**Question 13: Build, Train, Develop and Evaluate using Simple Regression for chosen dataset.**

**Answer:** We’ll choose one independent variable (e.g., ticket\_price) to predict the dependent variable (e.g., tickets\_sold).

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Question 14: Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression.**

**Answer:**

The Mean Squared Error (MSE) of 58,030.86 for your Multiple Linear Regression model indicates how well your model fits the data. A lower MSE value would mean a better fit, but this value depends on the scale and variance of your dataset.

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Question 15: Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree.**

**Answer:**

For doing that we have to face some issues like some of the columns had categorical values or missing data, which would hinder machine learning algorithms like Random Forest.

A screenshot of a computer program

AI-generated content may be incorrect.

**Random Forest Regressor**

A screenshot of a computer program

AI-generated content may be incorrect.

The R2 score of 0.998 is very high, which indicates that the model explains about 99.8% of the variance in the target variable, a fantastic result for most use cases.

A screenshot of a computer code

AI-generated content may be incorrect.

**Question 16: Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.**

**Answer:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A computer screen shot of a computer code

AI-generated content may be incorrect.**

**Question 17: Predict the dependent variable by using best-fit classifier.**

**Answer:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Question 18: Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset  
Answer:**

A screenshot of a computer program

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

**Question 19: Explain the strategy for improving the system after viewing the cluster diagram.  
Answer:**

To improve the system after viewing the cluster diagram, the following are key strategies:

* Segment Target Groups: Identify high and low-performing clusters, and then tailor strategies to leverage each. For example, spur low-performing areas with discounts or promotions, and reward high performers with loyalty schemes.
* Resource Optimization: Allocate resources (i.e., employees, advertising) based on the performance of the clusters. Focus efforts on low-performing areas or those with higher potential.
* Customer Personalization: Create tailored marketing or pricing strategies for every cluster. For example, offer special service to high-value customers and lower prices to price-conscious customers.
* Outlier Examination: Test outliers for error, fraud, or exceptional conditions requiring unique handling.
* Feature Refining: Adjust features based on insights gained through clusters. Create new features as needed to further differentiate groups.
* Continuous Tracking: Regularly update the clustering model as data shifts, adjusting strategies accordingly.

By using these findings, you can improve the performance of the system, reach key segments efficiently, and maximize resources.