

Coding Exercise – Python for Data Science

Dataset

There are three datasets that have been provided

- a) SalesData.xlsx
- b) imdb.csv
- c) diamonds.csv

The candidate will be utilizing them to answer the below given questions. Also provided along is Python and R answer template. Please fill in the attached template with respective functions codes.

Questions 1-6:

Utilize the sales data set. The sales data contains transactional sales information for each sales person. It also contains the date of sales, item sold, price of each item, sales amount, region and their corresponding manager information.

1. Find the least amount sale that was done for each item.
2. Compute the total sales for each year and region across all items
3. Create new column 'days_diff' with number of days difference between reference date passed and each order date
4. Create a dataframe with two columns: 'manager', 'list_of_salesmen'. Column 'manager' will contain the unique managers present and column 'list_of_salesmen' will contain an array of all salesmen under each manager.
5. For all regions find number of salesman and total sales. Return as a dataframe with three columns - Region, salesmen_count and total_sales
6. Create a dataframe with total sales as percentage for each manager. Dataframe to contain manager and percent_sales

Questions 7-10:

Utilize the imdb data set (duration is in seconds) The imdb data contains the rating and other information related to movies and episodes across a lot of genres and years.

7. Get the imdb rating for fifth movie of dataframe
8. Return titles of movies with shortest and longest run time

9. Sort the data frame by in the order of when they were released and have higher ratings, Hint :
release_date (earliest) and Imdb rating(highest to lowest)
10. Subset the dataframe with movies having the following parameters. duration between 30 minutes to 180 minutes.

Questions 11 – 15:

Utilize the diamonds data set. The diamonds data set contains the various dimensions and information for each diamond.

11. Count the duplicate rows of diamonds DataFrame.
12. Drop rows in case of missing values in carat and cut columns.
13. Subset the dataframe with only numeric columns.
14. Compute volume as (xyz) when depth is greater than 60. In case of depth less than 60 default volume to 8.
15. Impute missing price values with mean.