Coding Exercises

There are three datasets that have been provided

- SalesData.xlsx
- imdb.csv
- 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 fucntions codes.

Pacakages for Python

```
pandas
```

Packages for R (Optional. You can solve the questions using the standard way as well.)

```
readxl
dplyr
lubridate
```

Questions 1 - 6 Utilize the sales data set.

`The sales data contians 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 generes 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 where released and have higer ratings, Hint : release_date (earliest) and Imdb rating(highest to lowest)
10. Subset the dataframe with movies having the following prameters.
    - duration between 30 mintues 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 ($x*y*z$) when depth is greater than 60. In case of depth less than 60 default volume to 8.
15. Impute missing price values with mean.

Bonus questions (Optional)

`The bonus questions utilize the same data sets and if answere need to be filled in the respetive template`

1. Generate a report that tracks the various Genere combinations for each type year on year. The result data frame should contain type, Genere_combo, year, avg_rating, min_rating, max_rating, total_run_time_mins
2. Is there a relation between the length of a movie title and the ratings?
3. Generate a report that captures the trend of the number of letters in movies titles over years. We expect a cross tab between the year of the video release and the quantile that length fall under. The results should contain year, min_length, max_length, num_videos_less_than25Percentile, num_videos_25_50Percentile , num_videos_50_75Percentile, num_videos_greaterthan75Precentile
4. In diamonds data set Using the volumne calculated above, create bins that have equal population within them. Generate a report that contains cross tab between bins and cut. Represent the number under each cell as a percentage of total.