Project 1

Data Analyst – Case Study Case Guidelines:

Please go over the questions a couple of times to understand the asks.

Submit the scripts in a well-formatted PDF, so that our team can review and test your script.

The three tasks given below will test your grasp on Python (task one), SQL (task two) and data visualization (task three).

If you have any questions related to the case study, you can reach out to us at sagar.choudhari1007@gmail.com

Best of luck :)

Task One (Python):

1. Read the pupae data
2. Convert 'CO2_treatment' to a factor/categorical discrete variable. Inspect the levels of this factor variable.
3. Make a scatter plot of Frass vs. PupalWeight, with blue solid circles for a CO2 concentration of 280ppm and red for 400ppm. Also add a legend.
4. The problem with the above figure is that data for both temperature treatments is combined. Make two plots (either in a PDF, or two plots side by side), one with the 'ambient' temperature treatment, one with 'elevated'.
5. In the above plot, make sure that the X and Y axis ranges are the same for both plots.
6. Instead of making two separate plots, make one plot that uses different colors for the CO(2) treatments and different symbols for the 'ambient' and 'elevated' temperature treatments. Choose some nice symbols from the help page of the points function.
7. Add two legends to the above plot, one for the temperature treatment (showing different plotting symbols), and one for the CO2 treatments (showing different colours).
8. Generate the same plot as above but this time add a single legend that contains symbols and colours for each treatment combination (CO2 : T).

Task Two (SQL):

The table below indicates the search results of 'dog' and 'cat' on Twitter, position column represents each position the search result came in, and the rating column represents the rating allotted to the search result (1 to 5 rating, where 5 is high relevance and 1 is low relevance).

Write a query to compute a metric to measure the quality of the search results for each query.

QUERY	RESULT_ID	POSITION	RATING	NOTES
dog	1000	1	2	Picture of Snoop dog
dog	998	2	4	Dog walking
dog	342	3	1	Donkey
cat	123	1	4	Picture of cat
cat	435	2	2	Cat memes
cat	545	3	1	Burrito

Task Three (Tableau/Power BI):

Let's suppose you have been hired as a Data Analyst by Youtube, your first task is to create a dashboard for the latest trending YouTube videos across regions.This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, FR, RU, MX, KR, JP and IN regions regions (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India respectively respectively), with up to 200 listed trending videos per day. Each region's data is in a separate file and the data for each region includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.