AIML Online

Frequently Asked Questions in Problem Statement

Course: Applied Statistics

PART - A

** Direct or Self-explanatory questions are not covered in this FAQ.*

Q1 – Please refer the table below to answer below questions:

→ In this part you have to use the basic python skills and the mathematical formulas as per the requirement asked in the sub questions. You can use the markdown option to write theory based questions.

Q2 - An electrical manufacturing company conducts quality checks at specified periods on the products it manufactures. Historically, the failure rate for the manufactured item is 5%. Suppose a random sample of 10 manufactured items is selected. Answer the following questions.

→ This question and sub questions are related to the concept of probability, probability distributions & some libraries. Calculate the answers using the libraries only and not the manual calculations

Q3 - A car salesman sells on average 3 cars per week.

→ This question and sub questions are related to the concept of probability, probability distributions & some libraries. Calculate the answers using the libraries only and not the manual calculations

Q4 - Accuracy in understanding orders for a speech-based bot at a restaurant is important for the Company X which has designed, marketed and launched the product for a contactless delivery due to the COVID-19 pandemic. Recognition accuracy that measures the percentage of orders that are taken correctly is 86.8%. Suppose that you place an order with the bot and two friends of yours independently place orders with the same bot. Answer the following questions.

→ This question and sub questions are related to the concept of probability, probability distributions & some libraries. Calculate the answers using the libraries only and not the manual calculations

Q5 - Explain 1 real life industry scenario (other than the ones mentioned above) where you can use the concepts learnt in this module of Applied Statistics to get data driven business solutions.

→ This question is to be answered in the textual format and the answer should include the industry and how it can be used in the industry in 1-2 lines.

PART - B

Q1 - Read the data set, clean the data and prepare the final dataset to be used for analysis.

→ Read the given dataset.
This is a very important part of the project where the whole analysis will be done. In this question you need to perform data cleaning, checking null values and impute if there are any, check out all columns and replace or remove the unwanted characters if there are any (Replace () python function), data type conversion from object to integer except '**Team'** column

Q2 - Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get data driven insights on recommending which teams they can approach which will be a deal win for them. Also, as a data and statistics expert you have to develop a detailed performance report using this data.

→ This question to be performed as per the hint mentioned in the question
Hint: Use statistical techniques and visualization techniques to come up with useful metrics and reporting. Find out the best performing team, oldest team, team with highest goals, team with lowest performance etc. and many more. These are just random examples. Please use your best analytical approach to build this report. You can mix match columns to create new ones which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns. Use graphical interactive libraries to enable you to publish interactive plots in python.

Q3 - Please include any improvements or suggestions to the association management on quality, quantity, variety, velocity, veracity etc. on the data points collected by the association to perform a better data analysis in future. At-least 1 suggestion for each point.

→ You have to write about Improvements / Suggestions that can be applied to improve the data quality in 4-5 points. No codes required. For example, for 'quantity' you can suggest 'collecting more data' if the number of instances is less.

PART - C

Q1 - Read the CSV file.

→ Read the dataset given with the problem statement - CompanyX_EU.csv

Q2 - Data Exploration

→ In the sub questions check the features, their datatypes, check missing values

Q3 - Data preprocessing & visualization

→ In the sub questions you need to drop the null values then use the shared block of code to convert funding feature to numerical value. Plot the boxplot of funds in millions. Once you see the outliers get the upper fence of same feature and remove those by using suitable method considering the dataset is small.

Q4 A, B & C - Is there any significant difference between Funds raised by companies that are still operating vs companies that closed down?

a. Write the null hypothesis.
b. Write down alternative hypothesis.
c. Test for significance and conclusion.

Q4 D – Make a copy of the original data frame.

→ Here you have to make the copy of original data frame, save to new data frame and perform the remaining questions using this dataframe.

Q4 E - Check frequency distribution of Result variables.

→ Check the frequency distribution of result using appropriate python inbuilt function.

Q4 G, H – Write your hypothesis:

a. Comparing the proportion of companies that are operating between winners and contestants:
b. Test for significance and conclusion

Q4 I - Select only the Event that has 'disrupt' keyword from 2013 onwards.

→ In this question you need to show only those events which has disrupt keyword from the year 2013 onwards.