AIML | MODULE PROJECT

Applied **Statistics**

TOTAL SCORE **60**

General Instructions:

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.
5. Any kind of Plagiarism will lead to 0 (zero) Marks.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and
2. '.html' (Jupyter Notebook > File > Download as > HTML)
5 Marks will be deducted if submission in any of the formats is missing.

Part A - 15 Marks

- **Please answer the following questions with all relevant assumptions, explanations and details.**

1. Please refer the table below to answer below questions: [2 Marks]

Planned to purchase Product A	Actually placed order for Product A - Yes	Actually placed order for Product A - No	Total
Yes	400	100	500
No	200	1300	1500
Total	600	1400	2000

1.A. Refer above table and find the joint probability of the people who planned to purchase and actually placed an order. [1 Mark]

1.B. Refer to the above table and find the joint probability of the people who planned to purchase and actually placed an order, given that people planned to purchase. [1 Mark]

2. 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. [4 Marks]

2.A. Probability that none of the items are defective? [1 Mark]

2.B. Probability that exactly one of the items is defective? [1 Mark]

2.C. Probability that two or fewer of the items are defective? [1 Mark]

2.D. Probability that three or more of the items are defective ? [1 Mark]

3. A car salesman sells on an average 3 cars per week. [3 Marks]

3.A. What is Probability that in a given week he will sell some cars? [1 Mark]

3.B. What is Probability that in a given week he will sell 2 or more but less than 5 cars? [1 Mark]

3.C. Plot the poisson distribution function for cumulative probability of cars sold per-week vs number of cars sold per week. [1 Mark]

4. 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. [3 Marks]

4.A. What is the probability that all three orders will be recognised correctly? [1 Mark]

4.B. What is the probability that none of the three orders will be recognised correctly? [1 Mark]

4.C. What is the probability that at least two of the three orders will be recognised correctly? [1 Mark]

5. 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 solution. [3 Marks]

Part B - 30 Marks

- **DOMAIN:** Sports
- **CONTEXT:** Company X manages the men's top professional basketball division of the American league system. The dataset contains information on all the teams that have participated in all the past tournaments. It has data about how many baskets each team scored, conceded, how many times they came within the first 2 positions, how many tournaments they have qualified, their best position in the past, etc.
- **DATA DESCRIPTION:** Basketball.csv - The data set contains information on all the teams so far participated in all the past tournaments.
- **DATA DICTIONARY:**
  1. **Team**: Team's name
  2. **Tournament**: Number of played tournaments.
  3. **Score**: Team's score so far.
  4. **PlayedGames**: Games played by the team so far.
  5. **WonGames**: Games won by the team so far.
  6. **DrawnGames**: Games drawn by the team so far.
  7. **LostGames**: Games lost by the team so far.
  8. **BasketScored**: Basket scored by the team so far.
  9. **BasketGiven**: Basket scored against the team so far.
  10. **TournamentChampion**: How many times the team was a champion of the tournaments so far.
  11. **Runner-up**: How many times the team was a runners-up of the tournaments so far.
  12. **TeamLaunch**: Year the team was launched on professional basketball.
  13. **HighestPositionHeld**: Highest position held by the team amongst all the tournaments played.

- **PROJECT OBJECTIVE:** Company's management wants to invest on proposals on managing some of the best teams in the league. The analytics department has been assigned with a task of creating a report on the performance shown by the teams. Some of the older teams are already in contract with competitors. Hence Company X wants to understand which teams they can approach which will be a deal win for them.
- **STEPS AND TASK [30 Marks]:**

1. Read the data set, clean the data and prepare final dataset to be used for analysis. [10 Marks]

2. 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. [10 Marks]

   Hint: Use statistical techniques and visualisation 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.

3. 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. [10 Marks]

Part C - 15 Marks

- **DOMAIN:** Startup ecosystem
- **CONTEXT:** Company X is a EU online publisher focusing on the startups industry. The company specifically reports on the business related to technology news, analysis of emerging trends and profiling of new tech businesses and products. Their event i.e. Startup Battlefield is the world's pre-eminent startup competition. Startup Battlefield features 15-30 top early stage startups pitching top judges in front of a vast live audience, present in person and online.
- **DATA DESCRIPTION:** CompanyX_EU.csv - Each row in the dataset is a Start-up company and the columns describe the company.
- **DATA DICTIONARY:**
    1. **Startup**: Name of the company
    2. **Product**: Actual product
    3. **Funding**: Funds raised by the company in USD
    4. **Event**: The event the company participated in
    5. **Result**: Described by Contestant, Finalist, Audience choice, Winner or Runner up
    6. **OperatingState**: Current status of the company, Operating ,Closed, Acquired or IPO

    *Dataset has been downloaded from the internet. All the credit for the dataset goes to the original creator of the data.

- **PROJECT OBJECTIVE:** Analyse the data of the various companies from the given dataset and perform the tasks that are specified in the below steps. Draw insights from the various attributes that are present in the dataset, plot distributions, state hypotheses and draw conclusions from the dataset.
- **STEPS AND TASK [15 Marks]:**

1. Read the CSV file.
2. **Data Exploration: [1 Mark]**
    A. Check the datatypes of each attribute.
    B. Check for null values in the attributes.
3. **Data preprocessing & visualisation: [4 Marks]**
    A. Drop the null values. [1 Mark]
    B. Convert the 'Funding' features to a numerical value.

    (Execute below code)

    *df1.loc[:,'Funds_in_million'] = df1['Funding'].apply(lambda x: float(x[1:-1])/1000 if x[-1] == 'K' else (float(x[1:-1])*1000 if x[-1] == 'B' else float(x[1:-1])))*

    C. Plot box plot for funds in million. [1 Mark]
    D. Check the number of outliers greater than the upper fence. [1 Mark]
    E. Check frequency of the OperatingState features classes. [1 Mark]
4. **Statistical Analysis: [10 Marks]**
    A. Is there any significant difference between Funds raised by companies that are still operating vs companies that closed down? [1 Mark]
    B. Write the null hypothesis and alternative hypothesis. [1 Mark]
    C. Test for significance and conclusion [1 Mark]
    D. Make a copy of the original data frame. [1 Mark]
    E. Check frequency distribution of Result variables. [1 Mark]
    F. Calculate percentage of winners that are still operating and percentage of contestants that are still operating [1 Mark]
    G. Write your hypothesis comparing the proportion of companies that are operating between winners and contestants: [2 Mark]
    H. Test for significance and conclusion [1 Mark]
    I. Select only the Event that has 'disrupt' keyword from 2013 onwards. [1 Mark]