1. 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.

• ATTRIBUTE INFORMATION:

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 proposal 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 tasks:

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

2. Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get a 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. 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.

1. 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.

ATTRIBUTE INFORMATION: 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 tasks: 1. Data warehouse: • Read the CSV file. 2. Data exploration: • Check the datatypes of each attribute. • Check for null values in the attributes. 3. Data preprocessing & visualisation: • Drop the null values. • Convert the ‘Funding’ features to a numerical value. • Plot box plot for funds in million. • Get the lower fence from the box plot. • Check number of outliers greater than upper fence. • Drop the values that are greater than upper fence. • Plot the box plot after dropping the values. • Check frequency of the OperatingState features classes. • Plot a distribution plot for Funds in million. • Plot distribution plots for companies still operating and companies that closed. 4. Statistical analysis: • Is there any significant difference between Funds raised by companies that are still operating vs companies that closed down? Write the null hypothesis and alternative hypothesis. Test for significance and conclusion • Make a copy of the original data frame. • Check frequency distribution of Result variable. • Calculate percentage of winners that are still operating and percentage of contestants that are still operating • Write your hypothesis comparing the proportion of companies that are operating between winners and contestants: Write the null hypothesis and alternative hypothesis. Test for significance and conclusion • Check distribution of the Event variable. • Select only the Event that has disrupt keyword from 2013 onwards. • Write and perform your hypothesis along with significance test comparing the funds raised by companies across NY, SF and EU events from 2013 onwards. • Plot the distribution plot comparing the 3 city events. 5. Write your observations on improvements or suggestions on quality, quantity, variety, velocity, veracity etc. on the data points collected to perform a better data analysis.