|  |  |
| --- | --- |
| Name of the Student | Bryan Menezes |
| Internship Project Topic | Classification Model for Drug trial |
| Industry Mentor | Dr. Pradheep |
| Institute Name | CSMU-Chhatrapati Shivaji Maharaj University |

TCS iON Remote internship

RIO-210

|  |  |  |
| --- | --- | --- |
| Date | Day | Hour Spent |
|  | DAY 4 | 6.5 |
| Activities done today—Industry project work:  As per the Day wise Schedule provided on Day 4 and Day 5 we are to do project work making sure that we reach a milestone by the Day 5 and give a milestone or Interim project report 1.  Using google Collab or Jupyter Notebook we set the upload the dataset and then we import the libraries we are going to use in the project for understanding  The libraries used:    And uploading the data set along with showing part of the dataset:    Now we do Exploratory data analysis:    Using df.info() we see all the Columns such as Age, Condition, Data, Drug, DrugID, EaseofUse, Effectiveness, Reviews Satisfaction, Sex, Side and UsefulCount.  Using df.shape we find out the number of Rows 362806 and Columns 12,  Now we do descriptive data analysis:    Using df.describe() we get to generate descriptive statistics of the data in a data frame. It shows a summary of count, mean, std (standard deviation), min value in each column, 25%, 50%,75% and max value in each column.  Using df[].nunique we can find the unique amount of Drugs which is 7093 and Side effects which is 1651 and using df[].min() and max() we can find out the starting which is 01-01-2008 and end date which is 9-30-2019 of the drug test.    Using count() function we have found the drug with no useful or useless count which is 71608 and number of drugs with 100 or more count which is 139.  Bar Plot diagrams    Using import matplotlib.pyplot as plt from the cell above in the start can use plt.rcParams to form a bar plot. The above bar plot shows the correlation between UsefulCount and Satisfaction.    The above bar plot is similar to the previous bar plot but rather than correlations with UsefulCount and Satisfaction it is UsefulCount and Effectiveness.    Using a bar graph to see the top 20 drugs used:  The common drug used is Cymbalta.    This bar plot shows the top 20 conditions  The top condition is other. The unlisted other condition is around 50000 and above.  The pain and high blood pressure is the 2nd and 3rd conditions.  Pie plot:  Using imported matplotlib.pyplot and pandas we have created an Pie plot.  The above pie plot is titled Gender distribution that show the female to male ratio along with another entry which we will have to deal with in data preprocessing stage.    The above pie plot show the different age groups present in the data. | | |
| df.isna().sum() is used to count the number of missing values in each column of the Data frame. Only the 43 values are missing in the Review column.    Using df=df.dropna() we have dropped all the rows with missing reviews values and using df.isna().sum() to again count the number of values with missing values that is now zero across the board.    Using this code we can figure out the number of unique values in each column.  Count plot:    Using matplotlib and seaborn libraries we have created a count plot to show which age group uses more drugs.  The graph show that the age group 45 -54 use the most drugs followed by 55-64 and 35-44. | | |