BOTRUSH

OPTIMIZERS

-Team Members

Pushpalata(20232038)

Mahvish Fatma (20234098)

Round-1

**DATA CLEANING AND PREPROCESSING**

1. Checked **missing values** in the dataset.
2. Sorted it according to the Designation, as we had observed that **the Designation had a relation to the Constellation.** Stars belonging to a Constellation were grouped together and assigned similar Designation.
3. Then we **split the Bayern Designation into Bayer letter, Genetive Name and Identifier.** By splitting the Bayer letter into two columns – one which contained numerical values and other with the non-numerical values, we could interpolate the numerical values and hence estimate the Bayer letter of the missing value.
4. **By forward fill and interpolation,** we could fill the Genetive Name and Indicator of the Bayern Designation.
5. At the end, we regrouped the columns, **concatenated these columns** to get the Bayern Designation, dropped the unnecessary columns and resorted the data according to the index numbers.
6. We **split Approval Date into Approval Year and Approval Month** so as to plot the dataset and get better insights on it.

**NOTE:** Data Cleaning and Preprocessing was done in one Google colab file and the Questions were attempted on other ones. We are providing a drive link, and all the Google colab links are present in it.

**Questions Attempted**

1. Easy -

 Q3: Zodiac Star Distribution

Here we filtered the Zodaic signs 🡪 Counted the stars of each zodiac signs 🡪 plotted using matplotlib

1. Medium –

Q1: Heatmap of Approvals by Year and Month

Here intensity of color in each box represents no. of stars

Q2: Boxplot - Month vs Year

Sorted the months 🡪 plotted the boxplot using seaborn library

Q3: Scatter Plot - Year vs. Month

Plotted using scatter plot

1. Hard -

Q1: KMeans Clustering of Constellations

Selected categorical and numerical data 🡪 encoded the categorical data to numerical data 🡪 combined both categorical and numerical data 🡪 applied k means clustering using imported library 🡪plotted the data

Q2: PCA on Year-Month Data

Here we used PCA to reduce two variables (Approval Year, Approval Month)to 2 principle componenets .

Standardises the data using library 🡪 Selected no of components in PCA 🡪 Applied PCA 🡪 created plot using these 2 principle components as axis labels .

Round-2

**Questions Attempted**

Easy –

Q1) Approval Month Trend

Plotted no of approvals vs month year .

Medium –

Q2) Approval Density Plot (KDE)

Here x axis represent no of approvals per day and y axis represent no of days when these data of approvals was seen 🡪 approx 60 approvals were made on fewer days so this becomes anomaly

Hard –

Q1) Outlier Detection on Approval Dates

It is about finding data points that **stand out** or **don’t fit** with the rest of the data

Q2) Time-Series Anomaly Detection

Anomaly detection is used to identify **unusual data points** that deviate significantly from the rest of data — like a star approved way earlier or later than others