INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project – AMCAT Data Analysis

*By: **Matala Sushma***

About me

I'm Sushma Matala, currently pursuing a B-tech in Nadimpalli Satyanarayana Raju Institute of Technology of Computer Science specializing in Data Science. My academic pursuit is complemented by hands-on experiences, notably through my ongoing internship at Innomatics Research Labs, where I am actively engaging with real-world applications of my data science expertise.

- Linkedin link: https://www.linkedin.com/in/matalasushma/
- GitHub link: SushmaMatala (github.com)

INNOMATICS RESEARCH LABS

Objective:

The objective behind releasing the dataset by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO) is to provide researchers and analysts with valuable data to explore and understand various aspects of employment outcomes. This dataset serves as a resource for conducting in-depth analyses, uncovering trends, and deriving insights that can contribute to informed decision-making in the realm of employment and education.

My project focuses on delving into the AMCAT dataset to gain a comprehensive understanding of the determinants impacting the employability of engineering graduates. The primary aim is to extract actionable insights from the dataset, aiming to enhance the employability prospects for engineering graduates.

Exploratory Data Analysis

- **Data Cleaning Steps :**

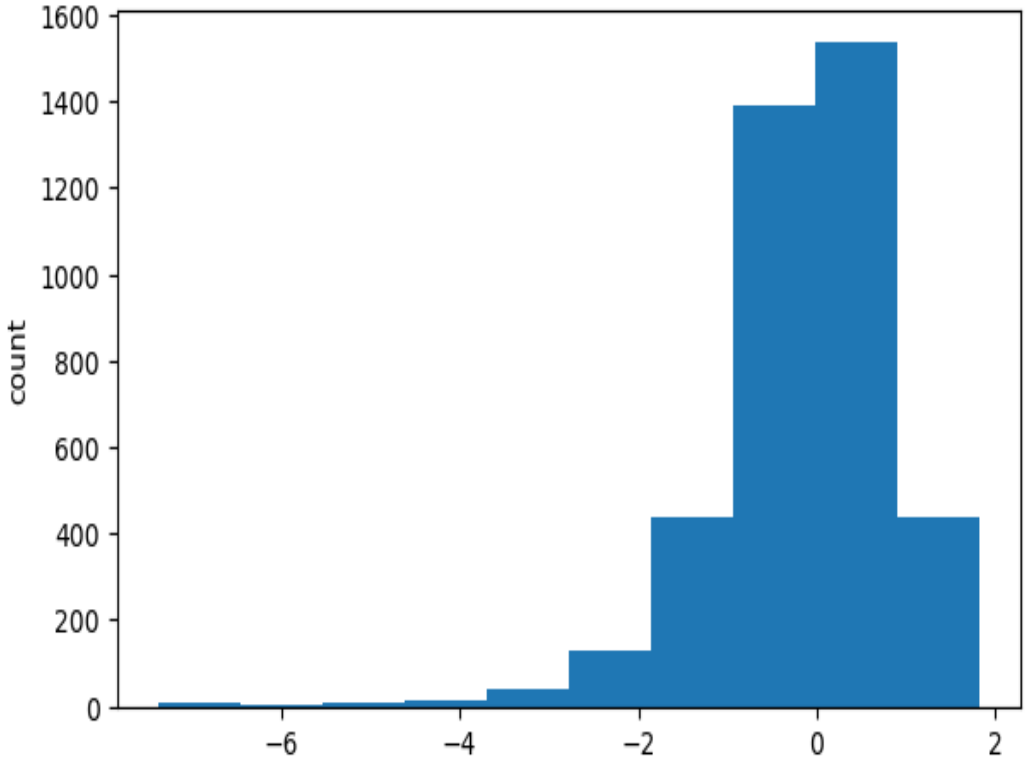
During the initial stages of our exploratory data analysis, our primary emphasis was on refining the dataset to uphold its quality and trustworthiness. This process entailed eliminating extraneous variables that held no significance in deriving meaningful insights for our analysis. Additionally, we meticulously managed missing data by employing suitable imputation techniques, thereby guaranteeing the integrity and comprehensiveness of our dataset. Moreover, we diligently detected and rectified outliers to safeguard against any potential distortion in our findings and conclusions.

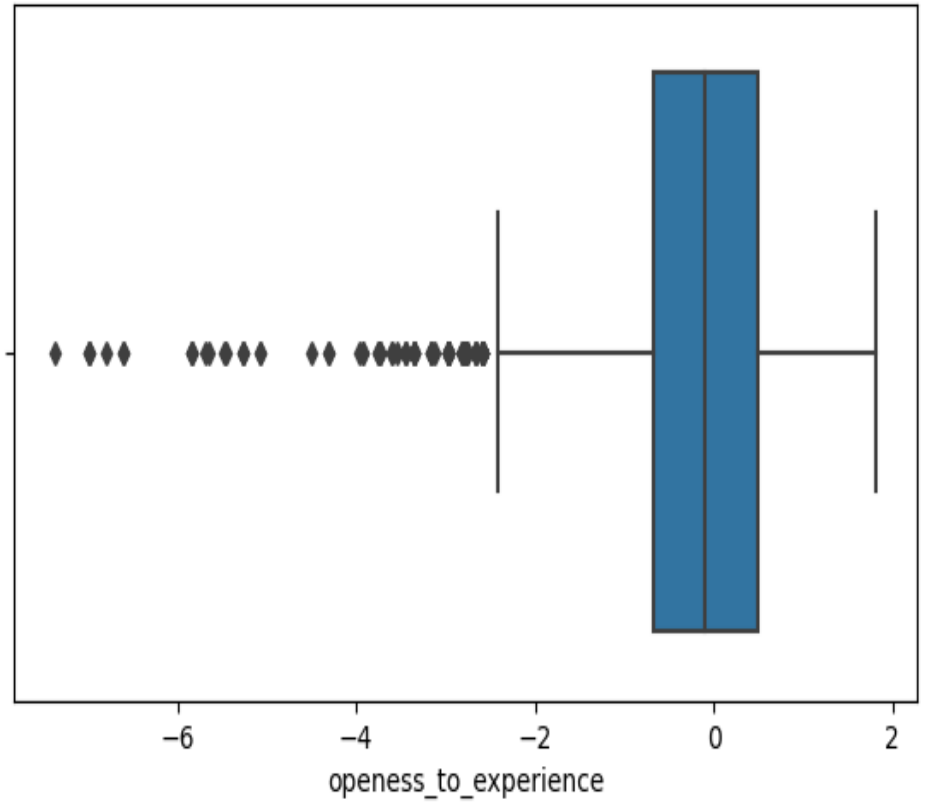
- **DataManipulation Steps :**

After completing the data cleaning phase, we advanced to data manipulation to elevate the dataset's quality, thereby enabling more profound analysis. This process encompassed transforming variables to streamline exploration and foster more robust insights. Additionally, we crafted new variables through amalgamation or segmentation of existing ones, unveiling latent patterns and correlations pivotal for deriving comprehensive conclusions from our analysis.

Univariate Analysis Steps :

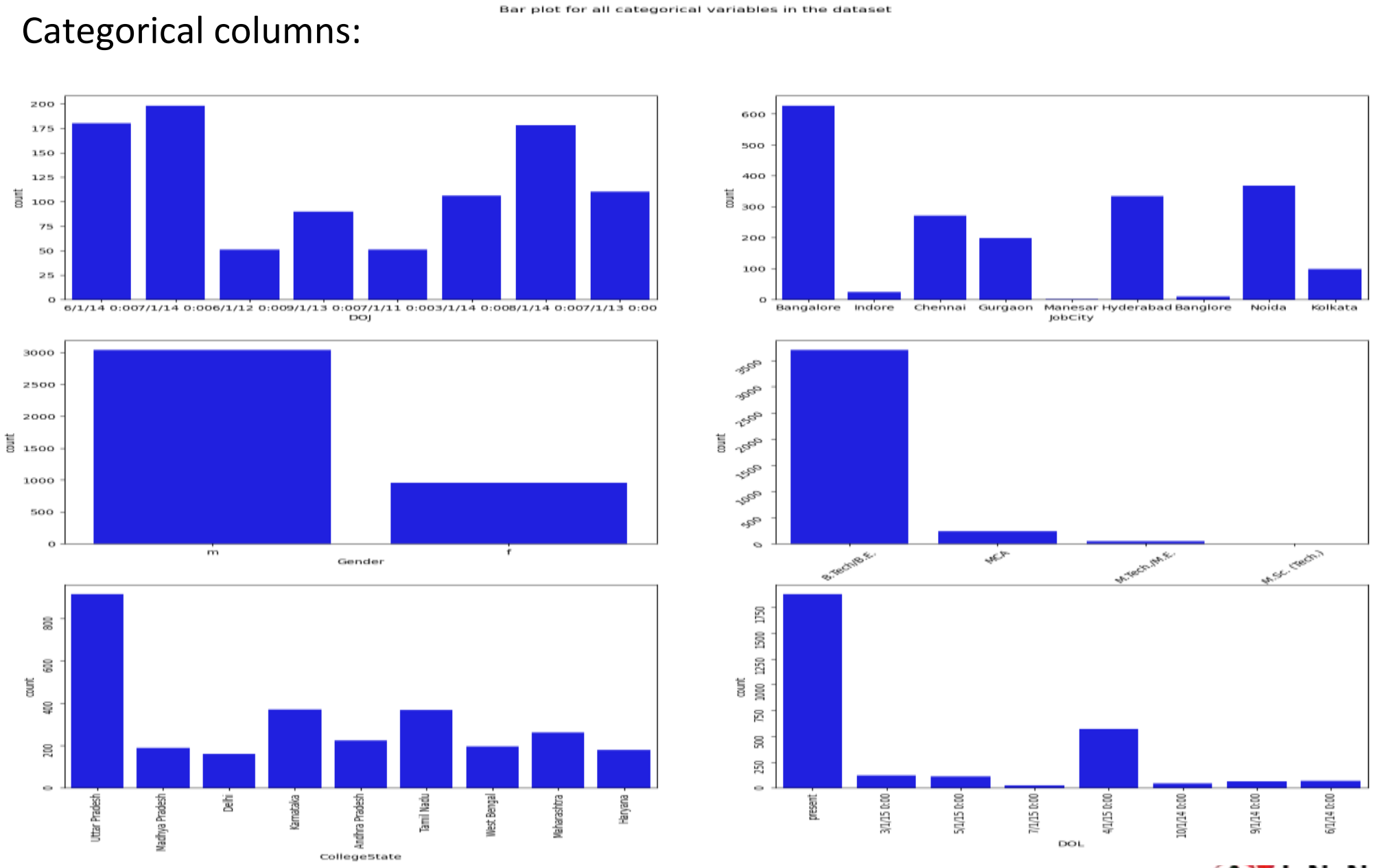
Numerical Columns:





INNOMATICS RESEARCH LABS

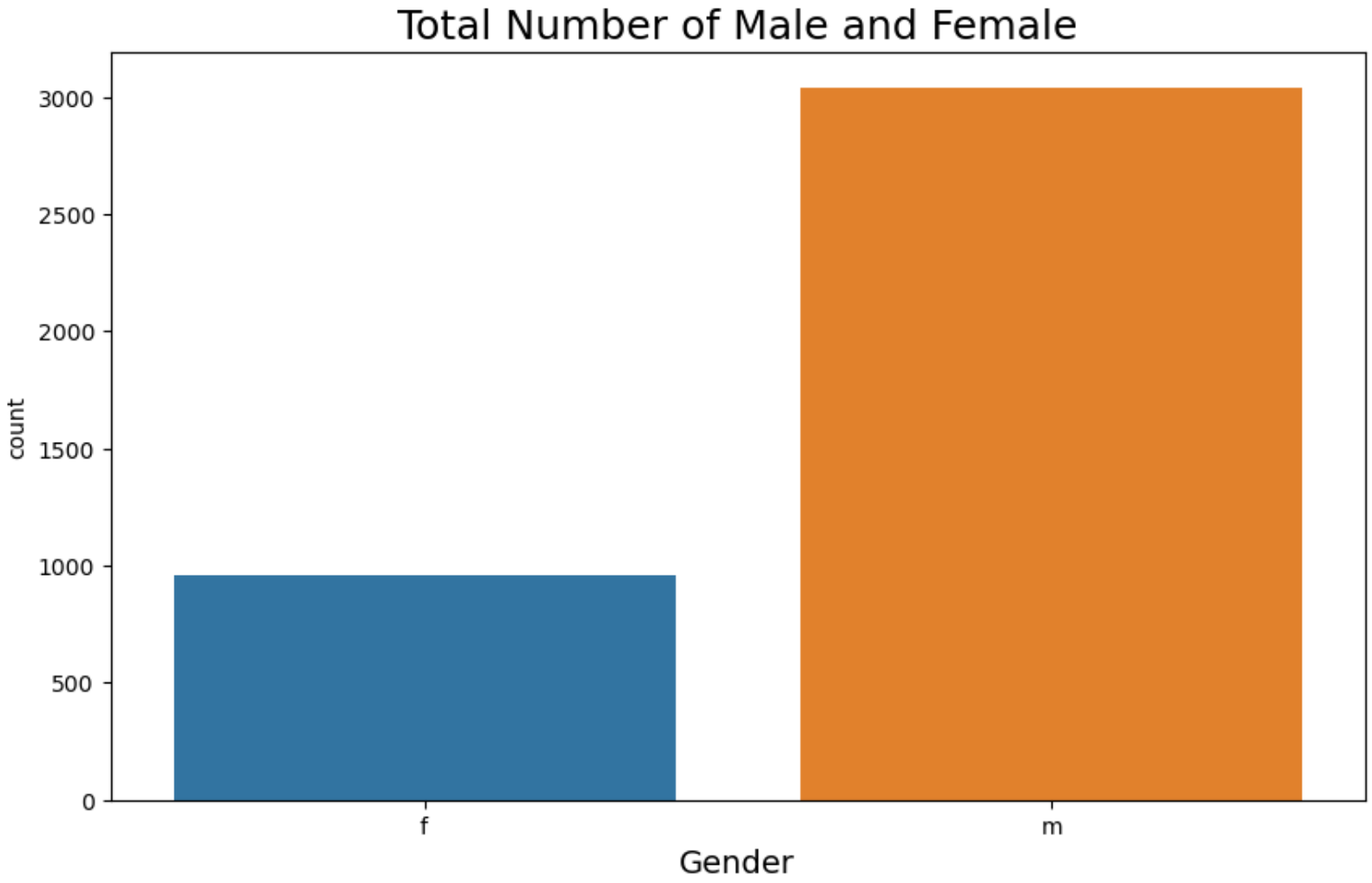
Categorical columns:



INNOMATICS RESEARCH LABS

Bivariate Analysis Steps:

Explored the relationship between two variables.



INNOMATICS RESEARCH LABS

Research Question:

After graduating with a degree in Computer Science Engineering, pursuing roles such as Programming Analyst, Software Engineer, Hardware Engineer, or Associate Engineer can yield a starting salary ranging from 2.5 to 3 lakhs for fresh graduates.

Using inferential statistics, I've demonstrated that the assertion suggesting fresh graduates can earn between 2.5 to 3 lakhs lacks support from the data.

Key Business Questions:

1. Can the salary expectations of fresh graduates, ranging between 2.5 to 3 lakhs, be validated through employment in roles such as Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer?

2. Is there a correlation between gender and salary levels?

3. Does gender influence the choice of specialization among fresh graduates, suggesting a dependence of specialization preference on gender?

Conclusion:

1. Mastery of English, logical reasoning, and quantitative aptitude significantly influences academic performance and the likelihood of securing lucrative salary packages.

2. Individuals with a more agreeable disposition often exhibit greater receptivity to new experiences, which in turn correlates with higher salary prospects.

3. Female graduates with advanced educational qualifications typically commence their careers with higher starting salaries compared to their male counterparts.

4. A compelling personality can mitigate the impact of a lower college GPA on potential salary earnings.

5. Students who completed their 10th and 12th grades under the CBSE and ICSE boards generally possess enhanced skill sets.

6. The predominant career path for graduates leads to roles as software engineers and developers, with an average annual salary of 307699.8 Dollars.

7. Graduates from Tier 1 educational institutions typically command higher salary packages in comparison to those from Tier 2 colleges.

THANK YOU

INNOMATICS
RESEARCH LABS

1.Ranking Top 10 Movies Of The Dataset

```
In [ ]: df.sort_values('vote_average', ascending=False).head(10)
```

```
Out[ ]:
        budget  genres  homepage    id  keywords  original_language  original_title  overview  popularity  production_companies  production_countries  release_date  revenue  runtime  spoken_languages  status  tagline  title  vote_average  vote_count
4391        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
3894        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4732        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4247        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4746        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4314        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4433        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
```

2.Sorting By The List Of Years Of The Dataset

```
In [ ]: df.sort_values('release_date', ascending=False).head(10)
```

```
Out[ ]:
        budget  genres  homepage    id  keywords  original_language  original_title  overview  popularity  production_companies  production_countries  release_date  revenue  runtime  spoken_languages  status  tagline  title  vote_average  vote_count
4732        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4746        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4314        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4433        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4391        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
3894        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4247        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
```

3.Sorting By The List Of Years Of The Dataset

```
In [ ]: df.sort_values('release_date', ascending=False).head(10)
```

```
Out[ ]:
        budget  genres  homepage    id  keywords  original_language  original_title  overview  popularity  production_companies  production_countries  release_date  revenue  runtime  spoken_languages  status  tagline  title  vote_average  vote_count
4732        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4746        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4314        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4433        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4488        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4391        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
3894        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
4247        0      NaN       NaN  9741       NaN                 fa             NaN       NaN       0.000                   NaN                   NaN           NaN        0      NaN               NaN     NaN      NaN    NaN           10.0           1
```

Find Change of Vote Vote Vote Vote Vote and Number of Vote Vote

```
In [ ]: df.shape
```

```
Out[ ]: (4803, 20)
```

Information about The Column Number of Rows Number of Columns Datatype Memory Usage

```
In [ ]: df.info()
```

```
Out[ ]:
```