BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Title of the Project

Group Number-035 – Amit Deepak Dahore, Pravin Bhaskar Kadekodi , Gaurav Pandey,

Introduction to Data Science

M.Tech Data Science and Engineering – Cluster Batch 4

04-March-2021

Overview

- Objective- **Based on users' demography we want to recommend what genre type movies user may like.**
- Methodology- **CRISP – Created business statement and converted it to analytical problem. Collect and understand all the relevant features to solve the problem. Followed by data engineering, EDA, modeling.**

Dataset

- How many features- **33 features**
- Size of the dataset- **100 000**
- Multiple files- **3**
- What kind of data – numerical or character- **numerical, categorical, DateTime**
- Balanced or imbalanced – what is the distribution- **imbalanced**
- Distribution of Training set, validation set, testing set- **60-20-20 percentage**
- Missing data and Preprocessing challenges- **Release date has missing data, we removed null values and converted it to years. Also had release date DateTime format needs preprocessing. Occupation None seems incorrect data. Using exploratory data analysis, we tried to summarize the important characteristics. Age feature has some outliers.**

Feature Engineering Techniques

- Features removed- **"movie id","timestamp", "IMDb URL","video release date", "user id", "movie title"**
- Feature creation- **converted release date to years**
- Feature ranking- **computational limitation**
- Class imbalance treatment- **created synthetical data**
- Any other- **Using SMOTE technique to handle the imbalance class would have improved our solution. Used box plot to identify outliers**

Methodology

- The 3 classifiers used- **Decision tree, logistic regression**
- Ensemble pipeline
- Other models considered- **only Decision tree and logistic regression are considered**
- Hyper-parameter tuning- **We have tuned the parameters using grid search cross validation**. **criterion="entropy", max_depth=12**

Before starting with the assignment, we read and understood the business problem behind the assignment and converted that business problem into the analytical problem. We have identified features relevant to solve the objective. Once we have fixed our data variables and the problem-solving approach, we have filtered out irrelevant features and created some new features from the existing variables using feature engineering techniques. We performed in-depth data exploratory data analysis using various visualization methods for example boxplot, heat map etc., refer graph in following pages.

After completing pre-processing, we moved our focus towards predictive analytics. In predictive analysis we used two algorithms namely decision tree and logistic regression and calculated their accuracy. We tried to improve the accuracy of decision tree by using more optimal features for example we changed criteria from default value to entropy and maximum depth to 12 in decision tree.

Results

- Table for the evaluation metric for each ML technique used:
  Confusion matrix, given in below pages.

- Plot of the curves
  All the plots are in page below

- Conclusion
  For a user with given demographic information we can recommend the user a movie of genre with 37% accuracy.

BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Title of the Project

Group Number-035 – Amit Deepak Dahore, Pravin Bhaskar Kadekodi , Gaurav Pandey,

Introduction to Data Science

M.Tech Data Science and Engineering – Cluster Batch 4

04-March-2021

Answering questions given with the assignment.

Q-1. Write a Data Science Proposal for achieving the objective mentioned.

Answer 1:

Business problem - Based on users' demographic information we want to recommend a genre of movies to user. This is a generic problem.

Target audience – Recommendation engine will be using our algorithm.

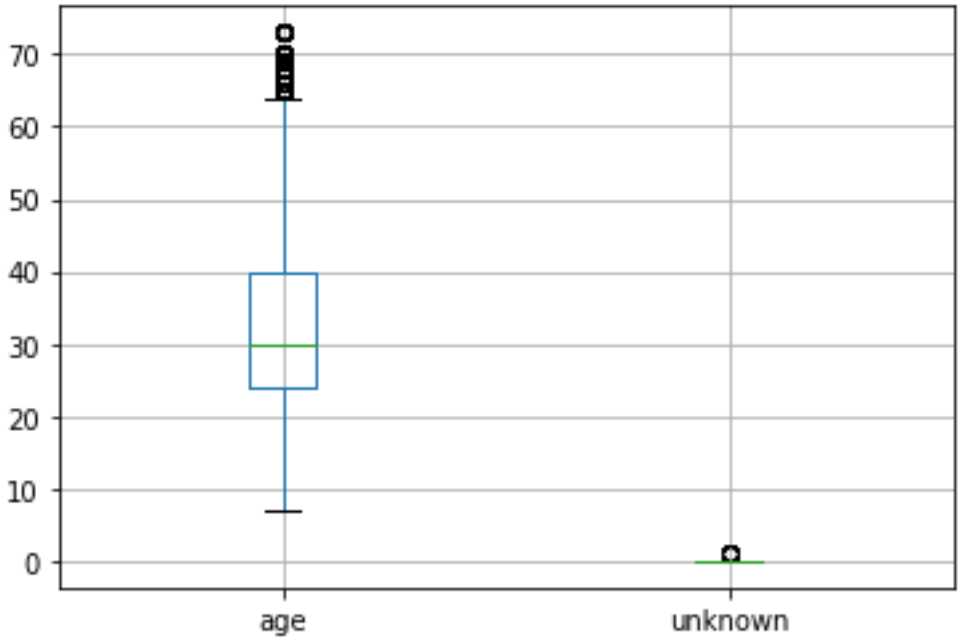
Evaluation method- F- score is used to evaluate.

Acceptable criteria- Accuracy of the solution should be greater than 50%.

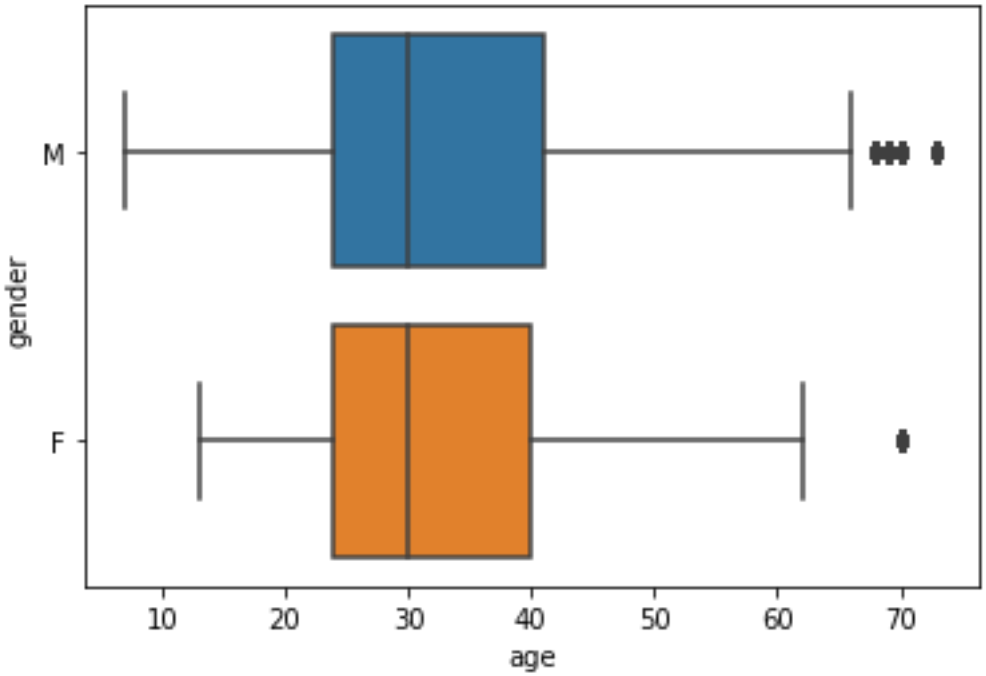
Q-2. Perform exploratory analysis on the data.
Q-3. Perform data wrangling / pre-processing.

Answer 2 and 3:

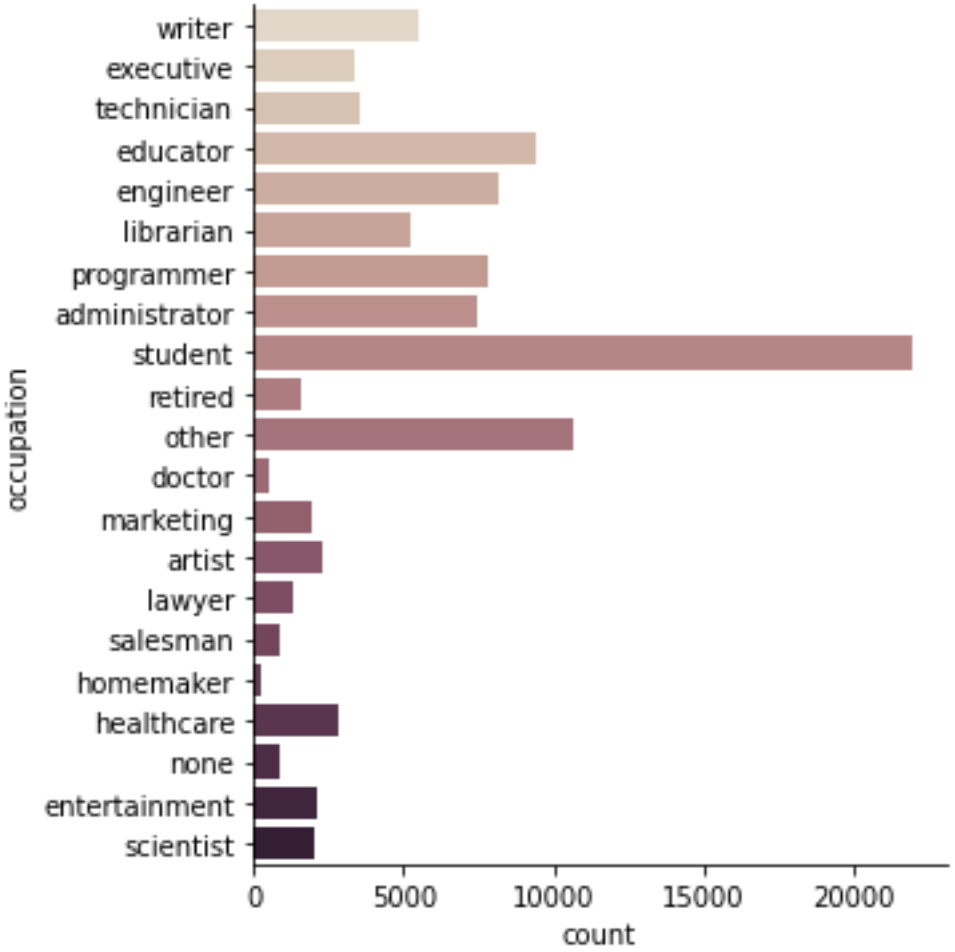
1) **Observation from below graph is that "Unknown" are zero and this column can be dropped.**



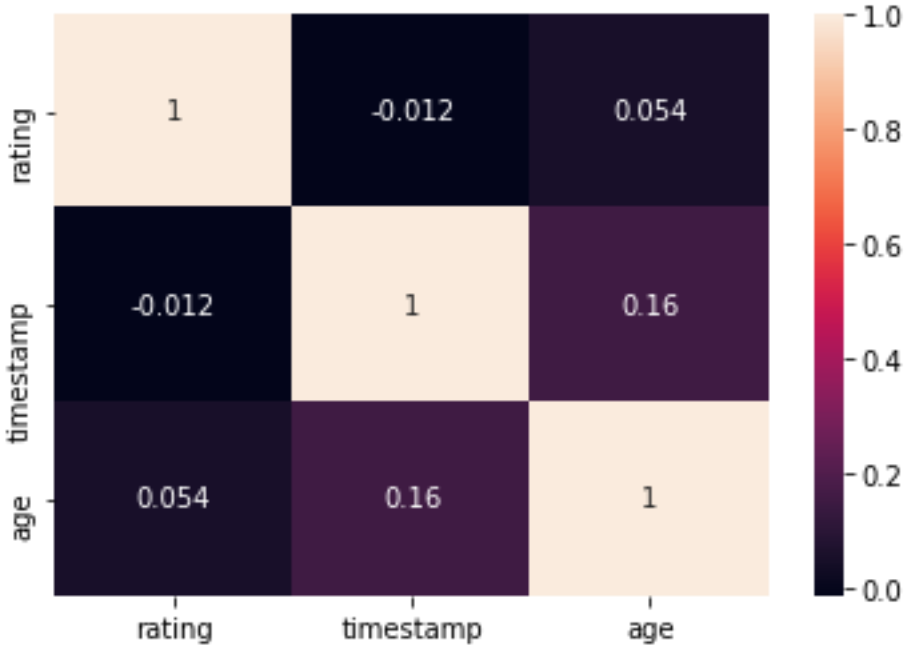
2) **Observation**: **We have found some outlier here. Lower age limit of Male is below Female. Median age seems same. Spread in Male age is more than Female.**



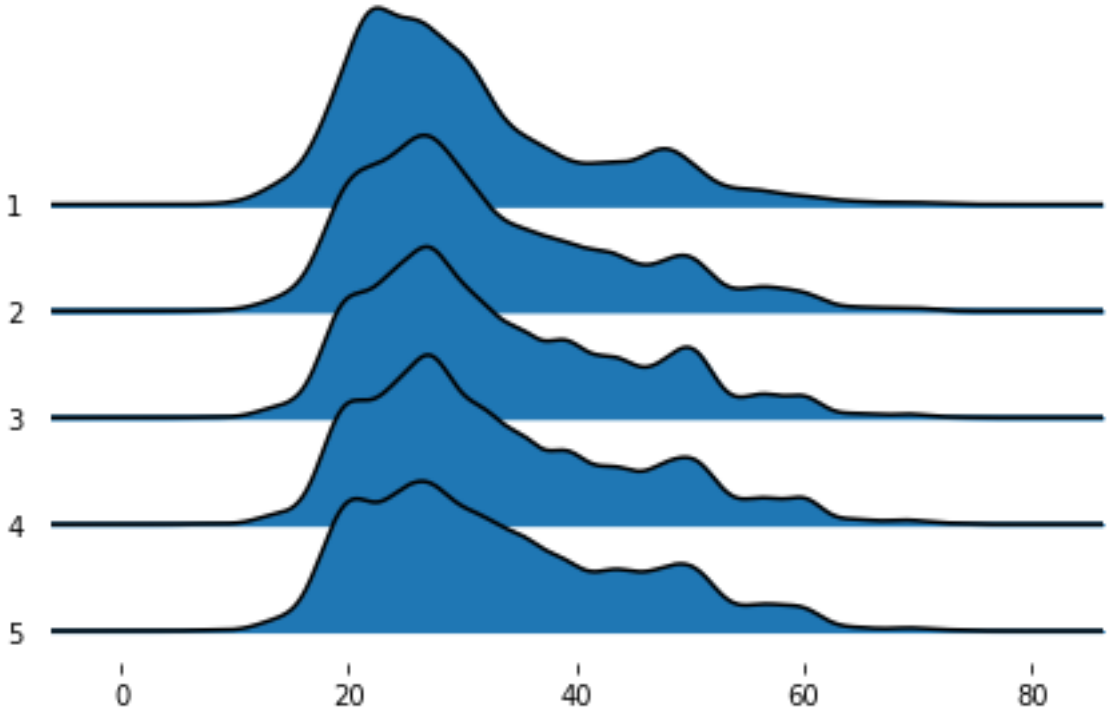
3) Observation: Student occupation is maximum, and some records are none. 'None' can be merged with 'other'.



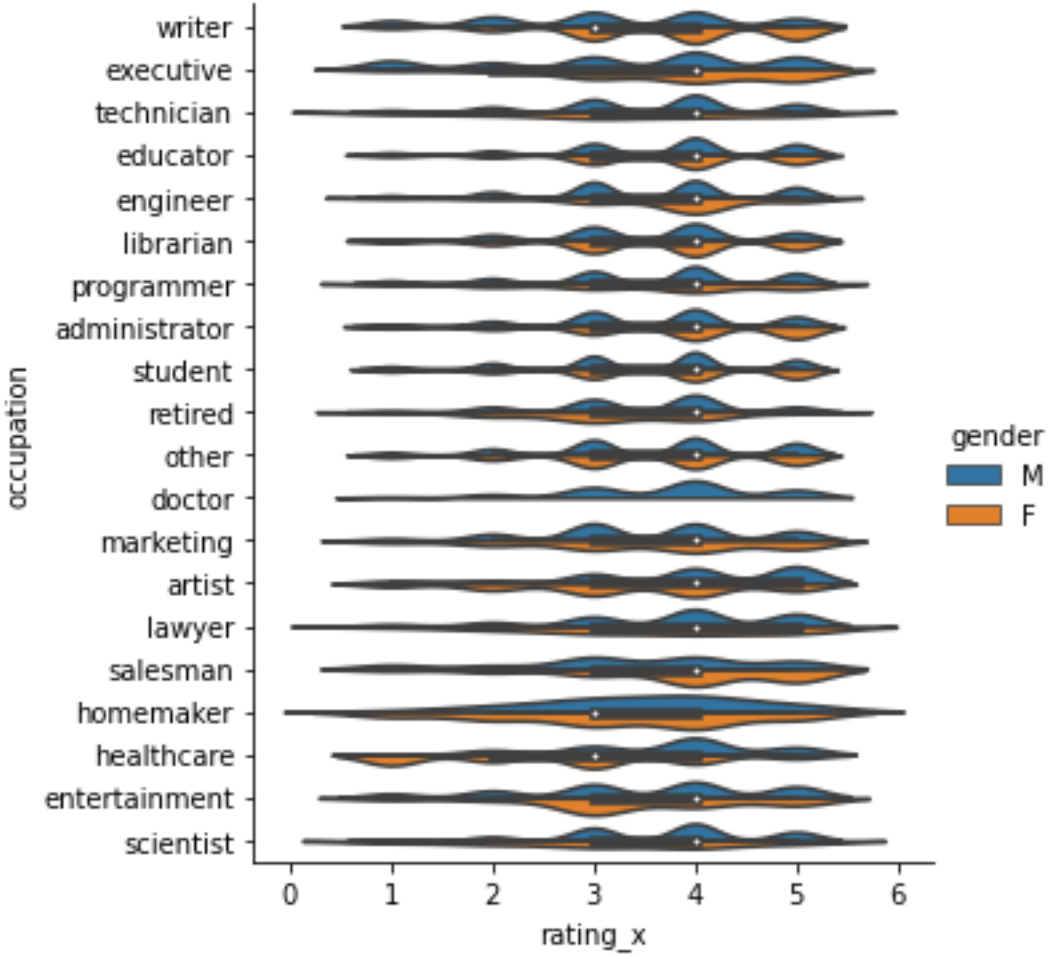
4) Observation: There is weak corelation between age, time-stamp and rating.



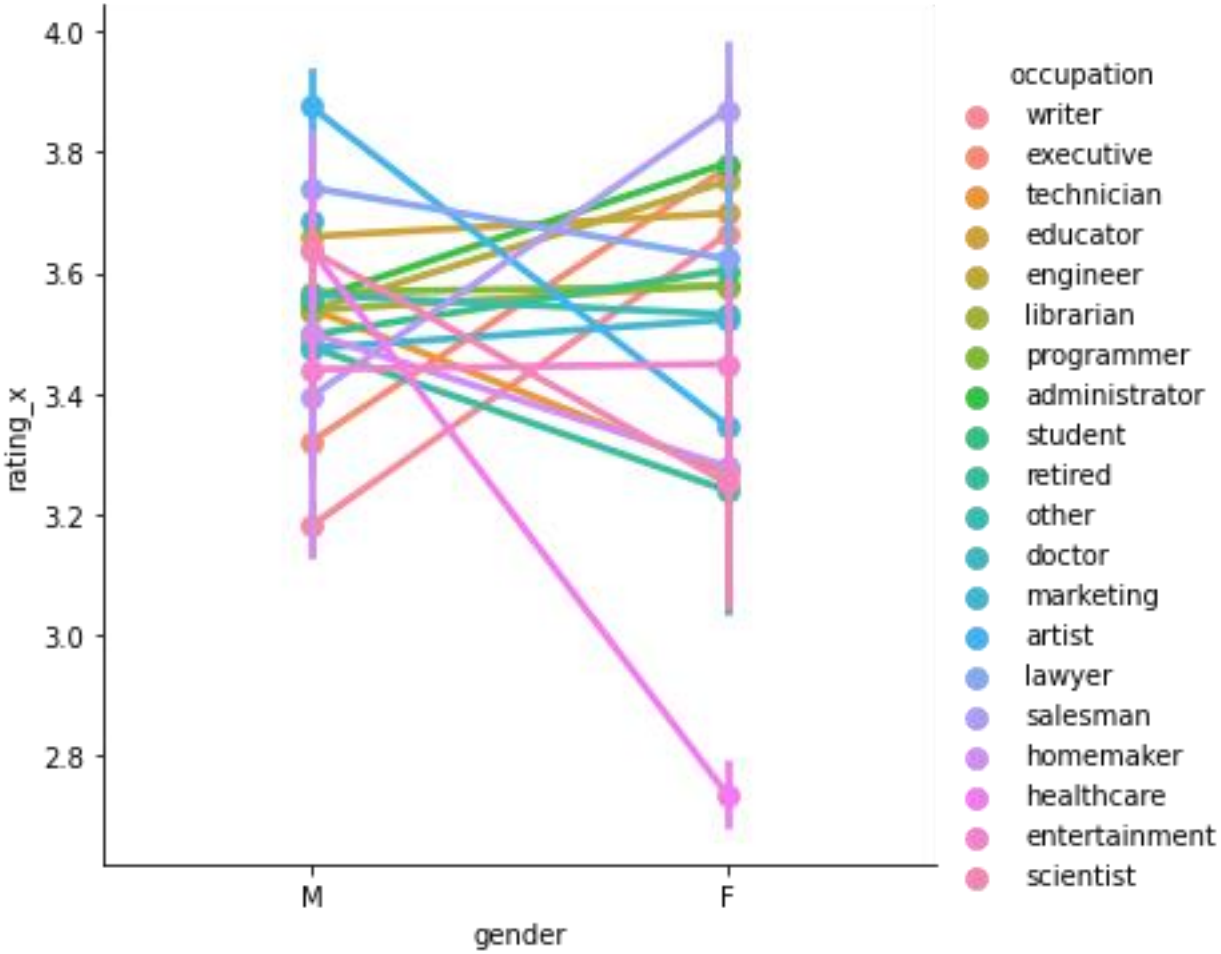
5) Observation: All ratings seems to have similar age distribution pattern.



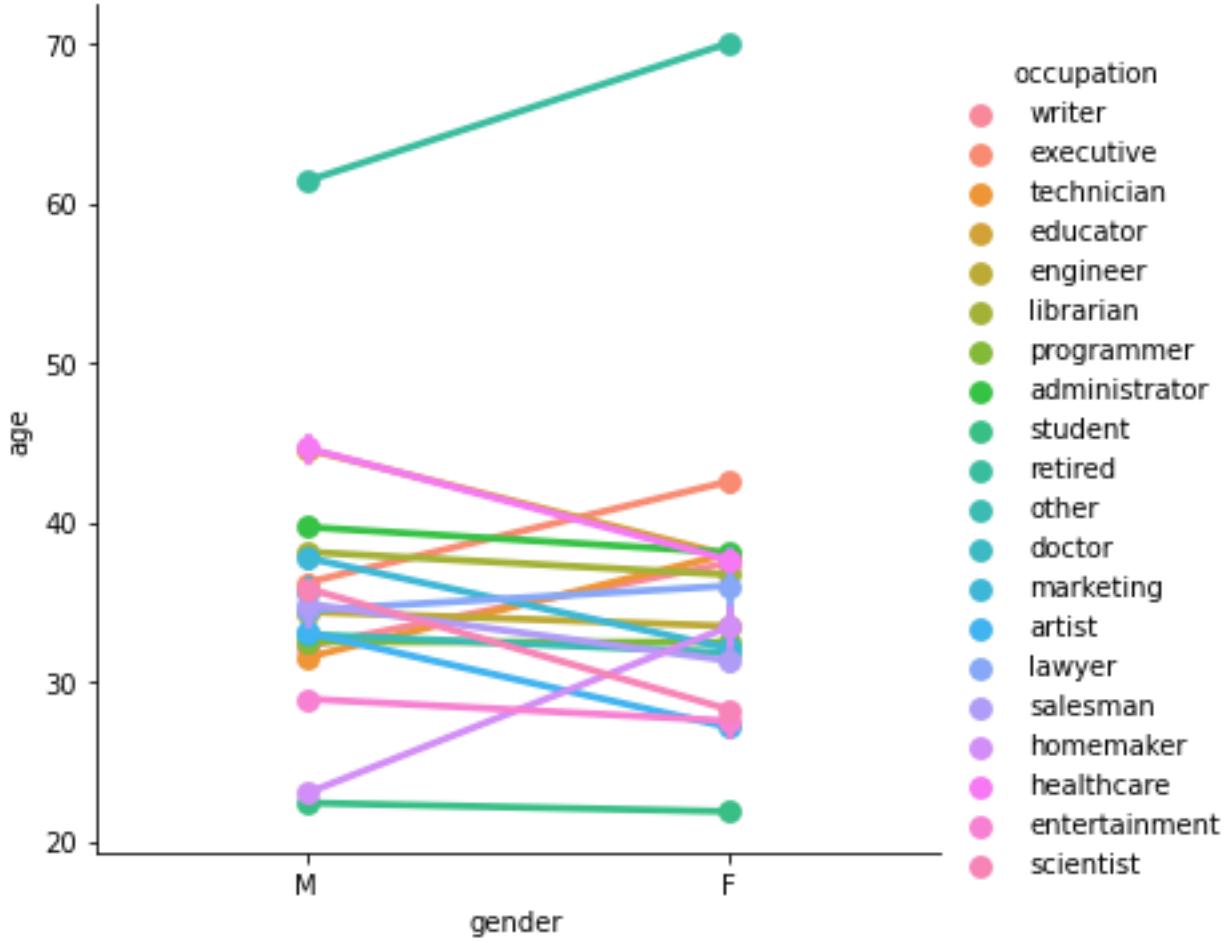
6) Observation: Occupation doctor has missing gender female, this might me incorrect data.



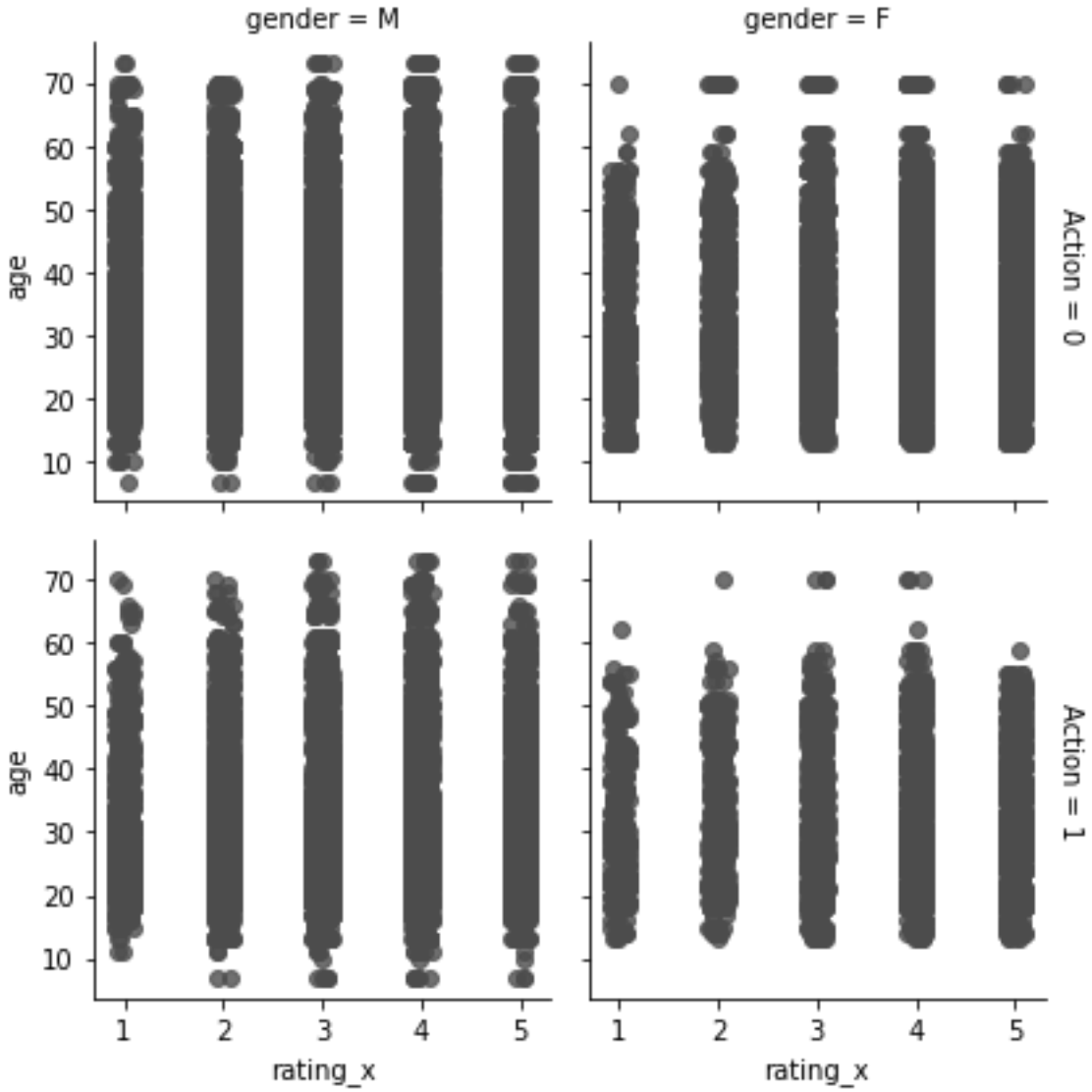
7) Observation: Number of Female healthcare data looks skewed.



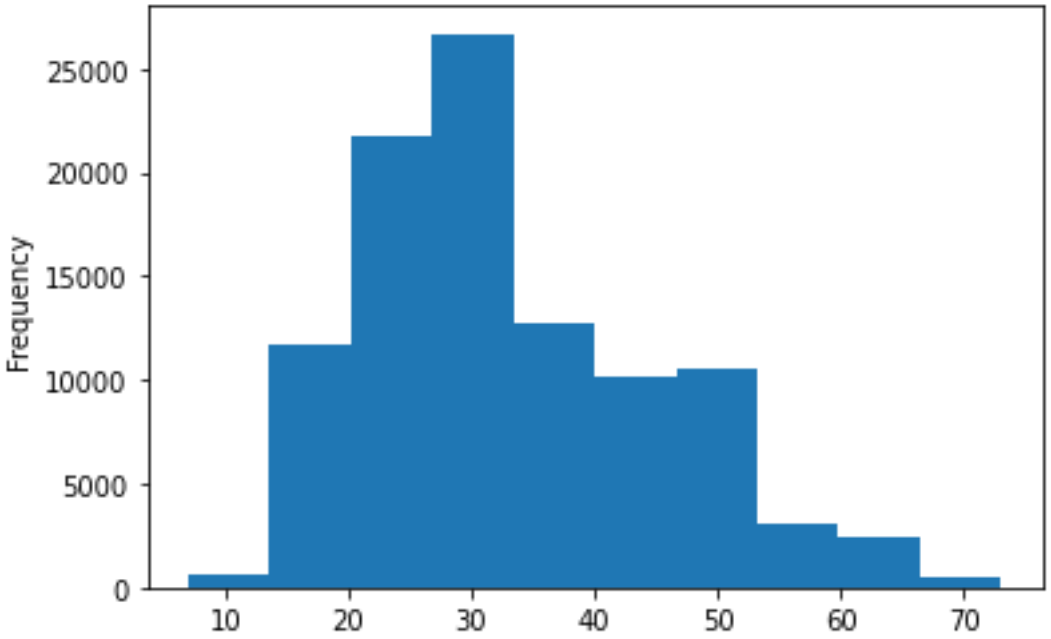
8) **Observation: Age of all retired occupation for male and female is above 60, which implies data is correct. Similarly, students age is near 20 which implies ages of students are correct.**



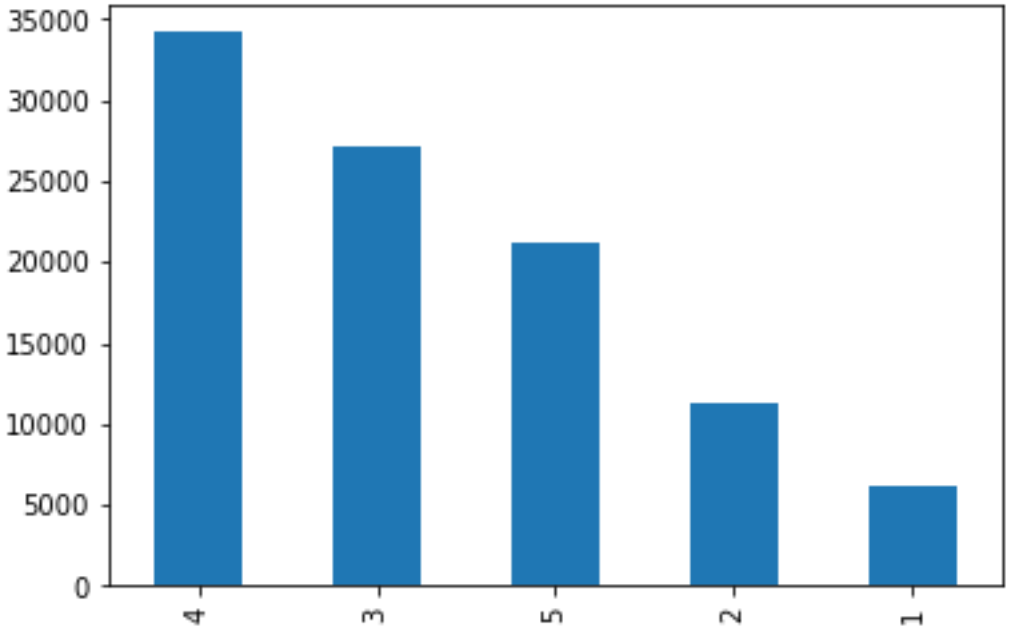
9) Observation: Rating, age seems to have no relation for Action movies



10) Observation: DataSet has maximum number of users from age group 25-30 and least for 5-10/65-70.



11) Observation: Rating of 4 has highest occurrence however rating 1 has least occurrence in data set



12) Observation: We wanted to test the hypothesis that is there any relation between gender and movie type. From chi-square test we have found that gender and genre move types are related.

```
import researchpy
crosstab, res = researchpy.crosstab(dataset['gender'], dataset['output'], test= "chi-square")

crosstab
```

Out[22]:

	output						
output	Action	Action_Adventure	Action_Adventure_Animation_Childrens_Fantasy	Action_Adventure_Animation_Horror_Sci-Fi	Action_Adventure_Childrens	Action_Adventure_Childrens_Fantasy	Action_Adventure_Child Fi
gender							
F	132	334	6	12	2	3	
M	747	1198	6	61	3	7	
All	879	1532	12	73	5	10	

3 rows × 217 columns

[23]:

```
res
```

Out[23]:

	Chi-square test	results
0	Pearson Chi-square (215.0) =	1375.8832
1	p-value =	0.0000
2	Cramer's V =	0.1173

13) Confusion Matrix: Average F-score for decision tree is 0.39.

	precision	recall	f1-score	support
Action	0.27	0.23	0.25	166
Action_Adventure	0.44	0.42	0.43	321
Action_Adventure_Animation_Childrens_Fantasy	0.00	0.00	0.00	1
Action_Adventure_Animation_Horror_Sci-Fi	1.00	1.00	1.00	10
Action_Adventure_Childrens	0.00	0.00	0.00	0
Action_Adventure_Childrens_Fantasy	0.00	0.00	0.00	2
Action_Adventure_Childrens_Fantasy_Sci-Fi	0.00	0.00	0.00	15
Action_Adventure_Childrens_Sci-Fi	0.00	0.00	0.00	2
Action_Adventure_Comedy	0.00	0.00	0.00	13
Action_Adventure_Comedy_Crime	0.14	0.32	0.20	59
Action_Adventure_Comedy_Horror	0.00	0.00	0.00	13
Action_Adventure_Comedy_Horror_Sci-Fi	0.00	0.00	0.00	19
Action_Adventure_Comedy_Musical_Thriller	0.00	0.00	0.00	18
Action_Adventure_Comedy_Romance	0.35	0.43	0.39	90
Action_Adventure_Comedy_Sci-Fi	0.81	0.95	0.88	60
Action_Adventure_Comedy_War	0.00	0.00	0.00	10
Action_Adventure_Crime	0.43	0.54	0.48	56
Action_Adventure_Crime_Drama	0.00	0.00	0.00	34
Action_Adventure_Crime_Thriller	0.00	0.00	0.00	3
Action_Adventure_Drama	0.31	0.17	0.22	24
Action_Adventure_Drama_Romance	0.00	0.00	0.00	19
Action_Adventure_Drama_Romance_Sci-Fi_War	0.39	0.66	0.49	73
Action_Adventure_Fantasy	0.52	1.00	0.68	26
Action_Adventure_Mystery	0.99	1.00	0.99	66
Action_Adventure_Mystery_Sci-Fi	0.00	0.00	0.00	3
Action_Adventure_Romance_Sci-Fi_War	0.75	0.84	0.79	204
Action_Adventure_Romance_War	0.26	0.83	0.40	29

Action_Adventure_Sci-Fi	0.37	0.30	0.33	349
Action_Adventure_Sci-Fi_Thriller	0.59	0.55	0.57	130
Action_Adventure_Sci-Fi_War	0.00	0.00	0.00	45
Action_Adventure_Thriller	0.78	0.67	0.72	275
Action_Adventure_Western	0.00	0.00	0.00	9
Action_Animation_Childrens_Sci-Fi_Thriller_War	0.00	0.00	0.00	4
Action_Childrens	0.00	0.00	0.00	4
Action_Comedy	0.24	0.22	0.23	49
Action_Comedy_Crime_Drama	1.00	1.00	1.00	1
Action_Comedy_Crime_Horror_Thriller	1.00	1.00	1.00	20
Action_Comedy_Drama	0.00	0.00	0.00	39
Action_Comedy_Musical	0.26	0.35	0.30	52
Action_Comedy_Musical_Sci-Fi	0.00	0.00	0.00	9
Action_Comedy_Sci-Fi_War	0.44	0.24	0.31	45
Action_Comedy_War	0.00	0.00	0.00	10
Action_Comedy_Western	0.54	0.33	0.41	88
Action_Crime	0.00	0.00	0.00	7
Action_Crime_Drama	0.82	0.60	0.69	126
Action_Crime_Mystery	0.00	0.00	0.00	11
Action_Crime_Romance	0.00	0.00	0.00	15
Action_Crime_Sci-Fi	0.00	0.00	0.00	18
Action_Crime_Thriller	0.00	0.00	0.00	63
Action_Drama	0.00	0.00	0.00	41
Action_Drama_Mystery	0.00	0.00	0.00	8
Action_Drama_Mystery_Romance_Thriller	0.00	0.00	0.00	14
Action_Drama_Romance	0.11	0.08	0.09	91
Action_Drama_Romance_War	0.00	0.00	0.00	22
Action_Drama_Thriller	0.00	0.00	0.00	87
Action_Drama_Thriller_War	0.91	0.96	0.93	52
Action_Drama_War	0.38	0.44	0.41	177
Action_Drama_Western	0.17	0.07	0.10	29
Action_Horror	0.46	0.73	0.56	67
Action_Horror_Sci-Fi	0.00	0.00	0.00	28
Action_Horror_Sci-Fi_Thriller	0.24	0.29	0.27	78
Action_Mystery_Romance_Thriller	0.84	0.98	0.90	52
Action_Mystery_Sci-Fi_Thriller	0.00	0.00	0.00	32

Action_Mystery_Thriller	0.27	0.50	0.35	8
Action_Romance	0.13	0.16	0.14	51
Action_Romance_Thriller	0.38	0.20	0.26	178
Action_Romance_War	0.00	0.00	0.00	32
Action_Sci-Fi	0.81	0.55	0.65	62
Action_Sci-Fi_Thriller	0.43	0.22	0.29	192
Action_Sci-Fi_Thriller_War	0.23	0.46	0.31	59
Action_Sci-Fi_War	0.86	0.99	0.92	85
Action_Thriller	0.40	0.41	0.40	660
Action_Thriller_War	0.00	0.00	0.00	20
Action_Western	0.93	0.96	0.95	27
Adventure	0.67	0.28	0.40	57
Adventure_Animation_Childrens_Comedy_Fantasy	0.00	0.00	0.00	20
Adventure_Animation_Childrens_Comedy_Musical	0.00	0.00	0.00	11
Adventure_Animation_Childrens_Musical	0.18	0.29	0.22	7
Adventure_Animation_Sci-Fi_Thriller	0.00	0.00	0.00	6
Adventure_Childrens	0.66	0.30	0.42	102
Adventure_Childrens_Comedy	0.49	0.79	0.60	82
Adventure_Childrens_Comedy_Fantasy_Romance_Sci-Fi	0.00	0.00	0.00	3
Adventure_Childrens_Drama	0.00	0.00	0.00	10
Adventure_Childrens_Drama_Musical	0.43	0.92	0.58	49
Adventure_Childrens_Fantasy	0.00	0.00	0.00	12
Adventure_Childrens_Fantasy_Sci-Fi	0.33	0.14	0.20	14
Adventure_Childrens_Musical	0.00	0.00	0.00	11
Adventure_Childrens_Romance	0.00	0.00	0.00	19
Adventure_Comedy	0.20	0.10	0.13	10
Adventure_Comedy_Drama	0.08	0.04	0.05	51
Adventure_Drama	0.00	0.00	0.00	30
Adventure_Drama_Western	0.11	0.09	0.09	47
Adventure_Romance	0.00	0.00	0.00	24
Adventure_Sci-Fi	0.00	0.00	0.00	17
Adventure_Sci-Fi_Thriller	0.38	0.71	0.50	7
Adventure_Thriller	1.00	1.00	1.00	22
Adventure_War	0.22	0.14	0.17	58
Animation	0.64	0.44	0.52	16
Animation_Childrens	0.24	0.07	0.11	86

Animation_Childrens_Comedy	0.00	0.00	0.00	92
Animation_Childrens_Comedy_Musical	0.15	0.21	0.18	42
Animation_Childrens_Comedy_Romance	0.00	0.00	0.00	8
Animation_Childrens_Musical	0.37	0.32	0.34	285
Animation_Childrens_Musical_Romance	0.00	0.00	0.00	12
Animation_Comedy	0.25	0.07	0.11	74
Animation_Comedy_Thriller	1.00	1.00	1.00	30
Animation_Sci-Fi	0.50	0.12	0.20	8
Childrens	0.00	0.00	0.00	2
Childrens_Comedy	0.29	0.21	0.24	234
Childrens_Comedy_Drama	0.15	0.04	0.06	51
Childrens_Comedy_Fantasy	1.00	0.20	0.33	15
Childrens_Comedy_Musical	0.48	0.42	0.45	52
Childrens_Comedy_Mystery	0.33	0.22	0.27	9
Childrens_Comedy_Western	0.00	0.00	0.00	3
Childrens_Drama	0.35	0.15	0.21	54
Childrens_Drama_Fantasy	0.00	0.00	0.00	10
Childrens_Drama_Fantasy_Sci-Fi	0.23	0.13	0.17	75
Childrens_Fantasy	0.00	0.00	0.00	4
Comedy	0.34	0.46	0.39	1990
Comedy_Crime	0.75	0.63	0.69	122
Comedy_Crime_Drama	0.00	0.00	0.00	1
Comedy_Crime_Drama_Mystery	0.00	0.00	0.00	18
Comedy_Crime_Fantasy	0.00	0.00	0.00	30
Comedy_Crime_Horror	0.00	0.00	0.00	8
Comedy_Crime_Mystery_Thriller	0.28	1.00	0.44	9
Comedy_Drama	0.42	0.22	0.29	475
Comedy_Drama_Musical	0.13	0.09	0.11	33
Comedy_Drama_Romance	0.34	0.10	0.15	147
Comedy_Drama_Thriller	0.00	0.00	0.00	4
Comedy_Drama_War	0.00	0.00	0.00	18
Comedy_Fantasy	0.00	0.00	0.00	5
Comedy_Fantasy_Romance_Sci-Fi	0.68	1.00	0.81	32
Comedy_Horror	0.00	0.00	0.00	100
Comedy_Musical	0.27	0.12	0.17	50
Comedy_Musical_Romance	0.60	0.52	0.56	91

Comedy_Mystery	0.00	0.00	0.00	9
Comedy_Mystery_Romance	0.00	0.00	0.00	2
Comedy_Mystery_Romance_Thriller	0.00	0.00	0.00	4
Comedy_Mystery_Thriller	0.51	1.00	0.68	22
Comedy_Romance	0.28	0.26	0.27	973
Comedy_Romance_Thriller	0.07	0.02	0.03	61
Comedy_Romance_War	0.00	0.00	0.00	66
Comedy_Sci-Fi	0.42	0.43	0.43	136
Comedy_Thriller	0.45	0.79	0.57	38
Comedy_War	0.51	0.51	0.51	88
Comedy_Western	0.00	0.00	0.00	9
Crime	0.39	0.42	0.41	59
Crime_Drama	0.46	0.28	0.35	276
Crime_Drama_Film-Noir	0.00	0.00	0.00	40
Crime_Drama_Mystery	0.00	0.00	0.00	32
Crime_Drama_Mystery_Thriller	0.57	1.00	0.73	4
Crime_Drama_Romance	0.36	0.56	0.43	18
Crime_Drama_Romance_Thriller	0.29	0.15	0.20	47
Crime_Drama_Sci-Fi	0.00	0.00	0.00	41
Crime_Drama_Thriller	0.67	0.54	0.60	181
Crime_Film-Noir	0.00	0.00	0.00	2
Crime_Film-Noir_Mystery	0.00	0.00	0.00	8
Crime_Film-Noir_Mystery_Thriller	0.00	0.00	0.00	69
Crime_Film-Noir_Thriller	0.36	0.32	0.34	28
Crime_Horror_Mystery_Thriller	0.00	0.00	0.00	44
Crime_Thriller	0.46	0.36	0.40	172
Documentary	0.71	0.13	0.22	116
Documentary_Drama	0.00	0.00	0.00	11
Documentary_War	0.00	0.00	0.00	11
Drama	0.29	0.59	0.39	2668
Drama_Fantasy_Thriller	0.00	0.00	0.00	20
Drama_Horror	0.00	0.00	0.00	69
Drama_Musical	0.45	0.53	0.49	79
Drama_Musical_War	0.13	0.22	0.16	23
Drama_Mystery	0.35	0.33	0.34	123
Drama_Mystery_Romance	0.00	0.00	0.00	9

Drama_Mystery_Sci-Fi_Thriller	0.60	0.91	0.72	46
Drama_Mystery_Thriller	0.00	0.00	0.00	17
Drama_Romance	0.39	0.34	0.36	937
Drama_Romance_Thriller	0.00	0.00	0.00	20
Drama_Romance_War	0.74	0.74	0.74	200
Drama_Romance_War_Western	0.00	0.00	0.00	17
Drama_Sci-Fi	0.86	0.50	0.63	206
Drama_Sci-Fi_Thriller	0.00	0.00	0.00	22
Drama_Thriller	0.50	0.55	0.53	536
Drama_Thriller_War	0.00	0.00	0.00	27
Drama_War	0.39	0.18	0.25	405
Drama_Western	0.00	0.00	0.00	4
Fantasy	0.00	0.00	0.00	1
Film-Noir	0.33	0.25	0.29	12
Film-Noir_Mystery	0.25	0.06	0.10	34
Film-Noir_Mystery_Thriller	0.25	0.19	0.21	27
Film-Noir_Romance_Thriller	0.00	0.00	0.00	8
Film-Noir_Sci-Fi	0.26	0.41	0.32	64
Film-Noir_Sci-Fi_Thriller	0.00	0.00	0.00	1
Film-Noir_Thriller	0.40	0.35	0.37	46
Horror	0.40	0.28	0.33	303
Horror_Mystery_Thriller	1.00	1.00	1.00	13
Horror_Romance	0.10	0.15	0.12	20
Horror_Romance_Thriller	0.62	0.86	0.72	44
Horror_Sci-Fi	0.50	0.15	0.23	20
Horror_Sci-Fi_Thriller	0.00	0.00	0.00	13
Horror_Thriller	0.57	0.65	0.61	139
Musical	0.67	0.69	0.68	62
Musical_Romance	0.42	0.32	0.36	90
Mystery	0.49	0.58	0.53	43
Mystery_Romance_Thriller	0.60	0.75	0.67	8
Mystery_Sci-Fi	0.00	0.00	0.00	1
Mystery_Thriller	0.60	0.50	0.55	285
Romance	0.62	0.28	0.38	58
Romance_Thriller	0.00	0.00	0.00	23
Romance_War	0.50	0.29	0.36	7

Sci-Fi	0.37	0.23	0.28	122
Sci-Fi_Thriller	0.16	0.12	0.14	48
Sci-Fi_War	0.48	0.40	0.44	40
Thriller	0.41	0.07	0.12	208
War	0.00	0.00	0.00	4
Western	0.09	0.10	0.10	116
accuracy			0.39	19676
macro avg	0.26	0.26	0.24	19676
weighted avg	0.37	0.39	0.36	19676

Q-4. Apply any 2 features engineering technique.

Answer 4:

a) **We have applied Release date from DateTime format to years.**
b) **Converted all the categorical variables into vectors using One-hot-encoding.**

Q-5. Plot top 10 features.

Answer 5: Refer Answer 2 and 3.

Q-6. Identification of the performance parameters to be improved, for the given problem statement.

Answer: F-score to be improved.

Q-7. Design Machine Learning models – Logistic regression and Decision tree to predict.

Answer: It is done in code, please refer .ipynb file attached.

Q-8. Compare the performance of selected feature engineering techniques.

Answer: Accuracy without performing feature engineering is 0.14 and after feature engineering is 0.38.

Q-9. Compare the performance of the 2 classifiers – Logistic regression and Decision tree to predict.

Answer: F-score of logistic regression is 0.14 and F-score for decision tree is 0.39.

Q-10. Present the conclusions/results in the format shared.

Answer: Refer page-1 of this document.